Leandro Marcio Moreira
organizador

Ciências genômicas:
fundamentos e aplicações

SBG Sociedade Brasileira de Genética

Leandro Marcio Moreira
organizador

Ciências genômicas:
fundamentos e aplicações

Sociedade Brasileira de Genética

Apoio

CAPES

PROJETO BIGA
Bioinformática, Genômica e Associados
Processo 3385/2013, edital 051/2013

Sociedade Brasileira de Genética
Rua Capitão Adélmio Norberto da Silva, 736
Alto da Boa Vista | CEP 14025-670 | Ribeirão Preto, SP
Tels.: (16) 3621-8540 | contato@sbg.org.br | www.sbg.org.br

Ciências genômicas : fundamentos e aplicações / Leandro Marcio Moreira (organizador). – Ribeirão Preto: Sociedade Brasileira de Genética, 2015.
403 p: il

Idioma: Português
ISBN 978-85-89265-22-5

1. Genômica estrutural. 2. Genômica funcional. 3. Genômica comparativa. 4. Biologia sintética. 5. Biologia de sistemas. 6. DNA barcoding. 7. Metagenômica. 8. Epigenômica. 9. Metabolômica. 10. Filogenômica. I. Moreira, Leandro Marcio, org. II. Título.

Capa, projeto gráfico e diagramação

editora cubo
soluções para o universo acadêmico

A minha esposa Edmara, e ao meu filho, Ângelo.
Aos meus Mestres, amigos e contemporâneos de formação.
Com carinho.

Leandro Marcio Moreira

Os autores

Alessandro de Mello Varani. Biomédico. Especialista em Bioinformática (LNCC) e Doutor em Biotecnologia pela USP. Pesquisador III do Departamento de Tecnologia da FCAV/UNESP-Jaboticabal. Trabalha com Bioinformática, Genômica e Evolução de Micro-organismos e Plantas.

Camila Carrião M. Garcia. Graduada em Química e Doutora em Bioquímica pelo IQ-USP. Atualmente é professor efetivo de Bioquímica e Biologia Molecular lotado no Departamento de Ciências Biológicas da Universidade Federal de Ouro Preto. Trabalha mecanismos de danos e reparo de DNA e suas implicações no Câncer e Envelhecimento.

Claudio de Oliveira. Biólogo, Mestre e Doutor em Biologia/Genética pela USP. Atualmente é professor titular de Biologia Celular, lotado no Departamento de Morfologia, Instituto de Biociências da Universidade Estadual Paulista (UNESP). Trabalha desenvolvendo estudos em sistemática, biodiversidade, conservação e evolução.

Denise Dagnino. Bióloga, Especialista em Biologia de Cianobactérias. Doutora em Ciências Naturais e Matemática pela Universidade de Leiden na Holanda. Professora Associada de Biologia no Laboratório de Biotecnologia da Universidade Estadual do Norte Fluminense. Trabalha com fisiologia de micro-organismos.

Diego Bonatto. Biólogo pela UFRGS. Especialista em Biofísica, Mestre e Doutor em Biologia Celular e Molecular UFRGS. Atualmente é professor adjunto de Biologia Molecular e Bioinformática, lotado no Departamento de Biologia Molecular e Biotecnologia da Universidade Federal do Rio Grande do Sul. Trabalha com ferramentas de biologia de sistemas aplicadas a mecanismos biológicos complexos, como envelhecimento e desenvolvimento. É membro afiliado da Academia Brasileira de Ciências (2011-2015).

Francisco Prosdócimi. Biólogo, Mestre em Genética e Doutor em Bioinformática pela UFMG. Atualmente é professor adjunto de Bioinformática no Instituto de Bioquímica Médica da Universidade Federal do Rio de Janeiro. Trabalha com montagem e anotação de genomas animais, genômica e transcriptômica comparativa, filogenômica e genética de populações.

Gonçalo Castelo-Branco. Licenciado em Bioquímica pela Universidade de Coimbra, Portugal, Doutor em Bioquímica Medica, pelo Instituto Karolinska, Estocolmo, Suécia. Atualmente é Professor Associado e Investigador Principal no Departamento de Bioquímica e Biofísica do Instituto Karolinska, Estocolmo, Suécia. Trabalha com estudos de epigenética em células progenitoras e células-mãe, com um foco em desenvolvimento neural e doenças neurológicas.

Helder I. Nakaya. Obteve seu grau de Bacharel em Ciências Biológicas pela Universidade de São Paulo em 2002 e seu doutorado em Bioquímica e Biologia Molecular pelo Departamento de Bioquímica do Instituto de Química da USP em 2007. O doutorado foi voltado para a área de Biologia Molecular e ioinformática. Se tornou professor assistente do Departamento de Patologia da Emory University em 2011, onde utiliza a biologia de sistemas para prever e entender a resposta imune a diferentes vacinas. Foi contratado como docente do Departamento de Análises Clínicas e Toxicológicas do Instituto de Ciências Farmacêuticas da USP em 2013.

Jan Schripsema. Farmacêutico, Especialista em Fitoquímica, Biotecnologia Vegetal e Metabolômica. Doutor em Ciências Naturais e Matemática pela Universidade de Leiden na Holanda. Professor Titular de Química de Produtos Naturais no Laboratório de Ciências Químicas da Universidade Estadual do Norte Fluminense. Trabalha com estudos de metabolômica em plantas medicinais e alimentos.

Jeronimo Conceição Ruiz. Químico pela UFSCar, Especialista em Bioinformática e Biologia Computacional e Doutor em Bioquímica pela FMRP/USP com pós-doutorado no Wellcome Trust Sanger Institute, Inglaterra. É pesquisador adjunto da FIOCRUZ (CPqRR, Belo Horizonte, MG). Atua primariamente na integração computacional de dados biológicos (redes de interação de proteínas, desordem estrutural proteica, imunoinformática, genômica e transcriptômica comparativa) em patógenos de interesse em saúde pública.

Juliana Lopes Rangel Fietto. Farmacêutica. Doutora em Bioquímica pela USP. Professora associada da Universidade Federal de Viçosa. Tem experiência na área de Biotecnologia e Bioquímica, com ênfase em biologia molecular, produção de proteínas recombinantese e aplicações biotecnológicas, bioquímica de enzimas e biologia celular, voltados para doenças infecciosas humanas e animais.

Julio Cezar Franco de Oliveira. Farmacêutico-Bioquímico. Doutor em Bioquímica pela USP. Professor Adjunto da Universidade Federal de São Paulo, UNIFESP - Diadema. Trabalha em Bioquímica com ênfase em Genômica Funcional.

Laila Alves Nahum. Bióloga. Bacharel em Microbiologia pela UFMG, Mestre em Bioquímica e Doutora em Genética pela USP com pós-doutorados no Marine Biological Laboratory e Louisiana State University, Estados Unidos. É pesquisadora efetiva na FIOCRUZ (CPqRR, Belo Horizonte, MG). Trabalha com filogenômica e evolução molecular de famílias de proteínas de diversos organismos atuando na interface entre Ciência e Educação.

Leandro Marcio Moreira. Biólogo e Especialista em Biologia Molecular pela USJT, Mestre e Doutor em Bioquímica pelo IQ-USP. Atualmente é professor efetivo de Bioquímica e Biologia Molecular lotado no Departamento de Ciências Biológicas da Universidade Federal de Ouro Preto. Trabalha com estudos de genômica comparativa e funcional de micro-organismos, bioprospecção e metagenômica de ambientes de canga e desenvolvimento de ferramentas didáticas em baixo custo para o ensino de ciências moleculares.

Leandro Xavier Neves. Bacharel em Nutrição. Mestre e doutorando em Biotecnologia pela Universidade Federal de Ouro Preto, área de concentração Genômica e Proteômica. Tem atuado na caracterização de proteínas da interface parasito-hospedeiro, em infecções causadas por helmintos e protozoários, e métodos de purificação, depleção e fracionamento proteico.

Luciana Principal Antunes. Bióloga. Bacharel em Ciências Biológicas pela UNIFESP e doutoranda em Ciências Biológicas (Bioquímica) pelo IQ- USP. Desenvolve projeto nas áreas de metagenômica e metatranscritômica de micro-organismos da compostagem do Zoológico de São Paulo. Atualmente também atua como Perita Criminal na Superintendência da Polícia Técnico-Científica de São Paulo.

Luciano Antonio Digiampietri. Bacharel e Doutor em Ciência da Computação pela UNICAMP. Atualmente é professor-doutor nos cursos de bacharelado e mestrado em Sistemas de Informação na Universidade de São Paulo. Trabalha com montagem e anotação de genomas e metagenomas; gerenciamento de experimentos científicos; mineração de dados; e análise de redes sociais.

Luciano Gomes Fietto. Farmacêutico. Doutor em Bioquímica e Biologia Molecular pelo Instituto de Química da USP. Professor Associado do Departamento de Bioquímica e Biologia Molecular da Universidade Federal de Viçosa (UFV). Trabalha com melhoramento de leveduras para produção de Bioetanol e no estudo da resposta a estresses em plantas e leveduras.

Luiz Henrique Garcia Pereira. Biólogo e Doutor em Ciências Biológicas (Zoologia) pela UNESP. Atualmente é Professor Adjunto de Biologia Molecular e Genética, lotado no Instituto Latino-Americano de Ciências da Vida e da Natureza da Universidade Federal de Integração Latino-Americana (UNILA). Trabalha com estudos em genética animal nas áreas de biodiversidade, conservação, sistemática e evolução.

Marcia Regina Soares da Silva. Física. Mestre em Física pelo Centro Brasileiro de Pesquisas Físicas. Doutora em Ciências Biológicas (Biofísica) pela UFRJ. Pós-doutora pelo LNLS. Professor adjunto do Instituto de Química da UFRJ. Tem experiência na área de Biofísica, com ênfase em Biofísica Molecular, atuando principalmente com proteínas e peptídeos, espectrometria de massas e proteômica.

Márcia Rogéria de Almeida. Bióloga (PUC-MG), Mestre em Bioquímica e Imunologia pela UFMG, Doutora em Bioquímica pela UFRGS/Plum Island Animal Disease Center-USDA-EUA. Professora titular da Universidade Federal de Viçosa. Têm experiência em Bioquímica Aplicada e Medicina Veterinária Preventiva, com ênfase em biologia molecular de vírus animais, atuando principalmente nos seguintes temas: infectologia molecular, diagnóstico, vacinas recombinantes e antivirais.

Mateus Schreiner Garcez Lopes. Biólogo, Especialista em Gestão de Projetos, Doutor em Biotecnologia com ênfase em Engenharia Metabólica e cursando MBA em Agronegócios. Possui experiência internacional em P&D em biotecnologia e atualmente trabalha na Inovação Corporativa da Braskem S.A. com novos negócios em biotecnologia.

Paulo Adriano Zaini. Biólogo e Doutor em Bioquímica pela USP. Atualmente é pesquisador nível pós-doutorado no Departamento de Bioquímica da Universidade de São Paulo. Trabalha com estudos de genômica comparativa e funcional de micro-organismos, com ênfase em bactérias fitopatogênicas.

Paulo de Paiva Amaral. Bacharel em Ciências Biológicas pela Universidade de Brasília, Mestre em Bioquímica pela Universidade de São Paulo e Doutor em Genética Molecular pela Universidade de Queensland, Austrália. Atua como Pesquisador Associado no Wellcome Trust / Cancer Research UK Gurdon Institute, Universidade de Cambridge, Reino Unido. Sua pesquisa abrange a caracterização das funções, evolução e mecanismos de ação de RNAs regulatórios em vertebrados, com foco na regulação da cromatina.

Renata Guerra de Sá Cota. Farmacêutica. Doutora em Bioquímica pela Universidade USP. Pós-doutora na área de Parasitologia Molecular pela USP. Professora Associada da Universidade Federal de Ouro Preto. Tem experiência nas áreas de Bioquímica e Biologia Molecular, com ênfase em Biologia Molecular de Parasitos.

Talles Eduardo Ferreira Maciel. Bioquímico. Mestre e Doutor em Bioquímica Agrícola pela UFV. Atualmente é pesquisador nível pós-doutorado no Departamento de Zootecnia da Universidade Federal de Viçosa; trabalhando essencialmente com Bioinformática. Trabalha com análise de dados oriundos de plataformas de sequenciamento de nova geração (tratamento inicial dos dados, montagem de genomas e transcriptomas, predição gênica e anotação), análise de expressão gênica diferencial, estudos de genômica comparativa e metagenômica.

William de Castro Borges. Farmacêutico. Doutor em Bioquímica pela USP. Pós-doutor pelo Centre of Excellence in Mass Spectrometry - University of York / UK (2005 - 2007). Professor Adjunto do Departamento de Ciências Biológicas da UFOP. Trabalha com Bioquímica dando ênfase na área de Proteômica.

Agradecimentos

A todos os professores e pesquisadores que fizeram parte desta conquista, ora como membro de autoria em um dos capítulos, ora como orientadores em tomada de decisão.

À CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) e ao Projeto BIGA - Biologia Computacional (Coordenado pelo Prof. Dr. João Carlos Setubal) por fomentarem este sonho.

À SBG (Sociedade Brasileira de Genética), em especial ao editor de livros Prof. Dr. Élgion Lúcio Silva Loreto, pelo apoio técnico nesta empreitada.

Aos alunos dos cursos de graduação e pós-graduação que, de uma forma ou de outra, me mostraram a importância de produzir algo dentro desta linha de conhecimento.

Aos professores/pesquisadores Fernando de Castro Reinach, Paulo Lee Ho, Carlos Frederico Martins Menck e Bayardo Baptista Torres por redigirem respectivamente os textos presentes no prefácio, forewords, orelha de capa e capa, respectivamente.

A todos que de alguma forma participaram da construção e caracterização desta obra.

Sumário

Prefácio 1

Fernando Reinach

O livro que você têm em mãos demonstra de maneira cabal o enorme progresso da Genômica nas últimas décadas. É impressionante o quanto a tecnologia mudou desde meados do século XX quando a estrutura do DNA foi descoberta.

Hoje somos capazes de sequênciar genomas inteiros em dias, mapear padrões de expressão de RNAs, caracterizar a expressão de proteínas e até determinar o genoma de comunidades de microorganismos. Além disso a tecnologia necessária para alterar genes individuais ou conjuntos de genes em diversos tipos de organismos é hoje uma atividade quase rotineira. Todas essas tecnologias estão muito bem explicadas nos diversos capítulos desse livro, com exemplos de aplicações e ilustrações de ótima qualidade.

A simples existência desse livro é uma demonstração que a Genômica é hoje uma área do conhecimento muito bem estabelecida em que o arcabouço teórico e experimental está suficientemente desenvolvido para permitir o surgimento de novas tecnologias e produtos. Essa é uma área do conhecimento em que os avanços estão deixando de ser revolucionários para serem incrementais. Mas isso não quer dizer que já sabemos o que está por ser descoberto. Olhando pela perspectiva do genoma e seus constituintes o que sabemos é impressionante. Mas se olharmos sob a perspectiva do fenótipo nossa ignorância é quase total.

Sugiro ao leitor, antes e depois de ler esse livro, que faça o seguinte exercício mental. Imagine o olho de um vertebrado. Depois tente listar o que ainda falta para podermos descrever a formação desse olho durante a embriogênese com uma sequência de ativações e desativações de genes presentes do genoma no desse animal. Basta tentar fazer um exercício dessa natureza para perceber que apesar de termos técnicas altamente sofisticados para analisar e interpretar o genoma, ainda estamos longe de entender o que está codificado no DNA que compõe esse genoma.

O estudo da Genômica só poderá ser considerado terminado quando, de posse da sequência de DNA de um dado organismos, formos capazes de derivar e descrever a forma do seu corpo, seu desenvolvimento embrionário, parte de seu comportamento, e seu ciclo de vida. Ainda estamos muito, muito longe essa façanha.

É por isso que a leitura desse livro impressiona. Se por um lado ela demonstra o quanto esta ciência progrediu nos últimos 60 anos, por outro lado deixa claro o quanto ainda temos pela frente.

Prefácio 2

Dr. Paulo Lee Ho

Bolsista de Produtividade em Pesquisa do CNPq - Nível 1A - CA BI - Biotecnologia
Pesquisador Científico IV e Diretor Técnico da Divisão de
Desenvolvimento Tecnológico e Produção do Instituto Butantan

A obra "Ciências Genômicas: fundamentos e aplicações", organizada pelo Prof. Dr. Leandro Márcio Moreira, contém diversos capítulos escritos por vários especialistas reconhecidos em suas áreas de atuação, incluindo o próprio Prof. Leandro. Tem o suporte da Capes (Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior), da Sociedade Brasileira de Genética e da Universidade Federal de Ouro Preto. Portanto, ela já se apresenta com vários selos de qualidade, que é confirmada com sua leitura.

Organizar e escrever uma obra técnica-científica, como o presente livro, é um desafio ingrato e de grande risco. O livro tem um propósito diferente de um artigo científico e busca revisar o conhecimento atual, de forma a educar um leitor fora da área e ao mesmo tempo organizar esse conhecimento e discuti-lo criticamente, permitindo àqueles que atuam em ciências genômicas, uma visão orgânica e atual dos mesmos. Por ser uma área técnico-científica, novos conhecimentos são gerados a cada dia. Assim, escrever um livro com tal conteúdo significa que ele já está desatualizado quando é finalizado. Desta maneira, este livro é muito diferente de um livro de romance, de crônicas, entre outros, que são obras perenes. Ainda assim, um livro técnico-científico é necessário e de grande importância, pois em função dos avanços diários no conhecimento científico, da complexidade e do grande volume de informações, facilita o acompanhamento no desenvolvimento das pesquisas e os avanços teóricos e técnicos da área, mesmo para os mais experimentados no assunto.

Conheci o Prof. Leandro quando foi meu aluno. Ele já era um aluno que se destacava dos demais não somente quanto à facilidade de aprendizado, mas também pelo "brilho nos olhos", de querer fazer mais, de querer melhorar sempre, de querer aprender constantemente, características imprescindíveis a um cientista. Outras qualidades que me chamaram a atenção foram sua honestidade, sinceridade e a transparência das suas ideias. De fato, esta obra é produto de um profissional preocupado em educar e passar adiante os conhecimentos, de uma pessoa que pensa grande. O País precisa de talentos e de profissionais ousados como o Prof. Leandro, que aceitou o desafio de organizar este livro com todos os riscos mencionados acima. A leitura do livro é apenas o início da aventura do conhecimento. Se ela for bem-sucedida, irá envolver as mentes de jovens

estudantes com perguntas, curiosidades, desejo de conhecer mais e mais esta área, de experimentar, de entender os fatos da natureza, de criar conhecimento, de contribuir para o bem-estar da humanidade com os novos conhecimentos científicos. E é isto o que importa, pois estes jovens irão construir o futuro deste País e irão inspirar outros jovens estudantes e assim, sucessivamente. Parabéns ao Prof. Leandro, aos autores dos capítulos, à Editora e aos patrocinadores pelo excelente trabalho.

1

História e importância da genômica

Luciano Gomes Fietto
Márcia Rogéria de Almeida Lamëgo

Introdução

Várias descobertas podem ser consideradas como o início da Ciência Genômica (Singer and Berg 2004). Na Tabela 1 está representada uma linha do tempo, onde descobertas no estudo da Genética e da Biologia Molecular estão listadas como fundamentais para o desenvolvimento da Genômica como nós a conhecemos hoje. Muitos leitores sentirão falta de algumas descobertas importantes ou mesmo não considerarão algumas destas apresentadas como importantes para o desenvolvimento desta importante e moderna Ciência. Este fato é esperado já que apesar de descobertas importantes serem atribuídas a uma pessoa ou a um grupo específico, certamente os achados fundamentais estão baseados em uma bibliografia vasta e coerente. Desse modo podemos considerar a Genômica uma ciência moderna que surgiu e cresceu pela convergência de outras ciências, principalmente a Bioquímica, a Estatística e a Ciência da Computação.

O objetivo central deste capítulo é posicionar o leitor num contexto histórico que o permita compreender melhor a importância destas descobertas no atual desenvolvimento da Ciência Genômica. Ao mesmo tempo, esperamos que esta proposta possa ser amplamente utilizada por docentes e alunos dos mais variados cursos de graduação e pós-graduação como uma ferramenta fundamental ao desenvolvimento de disciplinas relacionadas. Após uma breve descrição e cronologia das principais descobertas faremos uma abordagem detalhada de algumas experimentações que consideramos fundamentais e deveriam ser discutidas com todos os alunos formandos em ciências biológicas, uma forma de resgatar o conhecimento histórico e o desenvolvimento do pensamento científico.

Cronologia das descobertas

Desde que Gregor Mendel realizou experimento com ervilhas e definiu os primeiros conceitos na genética, em 1858, se passou mais de um século até que os primeiros genomas viessem a ser sequenciados. Após o primeiro genoma ser decifrado, em pouco mais de uma década foram sequenciados os genomas completos de vários organismos procariotas e eucariotas, chegando até o genoma humano.

Simplificando a linha do tempo podemos considerar o nascimento da Era Genômica quando o genoma de 5.375 mil pares de bases do Bacteriófago ϕx174 foi determinado (Sanger, Air et al. 1977). Outra revolucionária técnica descrita por Sanger conhecida como *shotgun*, que consiste no isolamento de pedaços randômicos do DNA que depois de sequenciados são agrupados em pedaços maiores até a montagem completa de uma longa sequência de DNA contínua, possibilitou o sequenciamento do Bacteriófago λ que possui um genoma de 48.502 mil pares de bases, consideravelmente maior que o do fago ϕx174 (Sanger, Coulson et al. 1980). Após Sanger demonstrar que o sequenciamento de genomas era possível uma série de outros genomas virais foi sequenciada, entre eles o do Vírus *Vaccinia*, de 187.000 mil pares de bases (Paoletti, Weinberg et al. 1984), e o do *Cytomegalovírus*, de 192.000 mil pares de bases (Chee, Bankier et al. 1990).

Após o sucesso do sequenciamento destes genomas virais um grupo de cientistas Europeus, liderados pelo belga Andre Goffeau, iniciou em 1989 uma empreitada para sequenciar o genoma da levedura *Saccharomyces cerevisiae*. Inicialmente o grupo de cientistas reunia 74 laboratórios organizados em consórcios. Os cientistas supunham possuir em torno de 12.5 milhões de pares de bases (Dujon 1993; Goffeau 1994; Goffeau, Barrell et al. 1996). Esta associação contribuiu com mais de 55% para a identificação completa do genoma da levedura, juntamente com dois grupos norte-americanos.

A ideia de consórcios para sequenciamento de genomas complexos, idealizada por Goffeau, se espalhou pelo mundo e vários consórcios foram formados em diferentes países. O projeto de sequenciamento do genoma da levedura possibilitou o aprimoramento das técnicas de montagem e sequenciamento do genoma, inicialmente descritas por Sanger, tornando mais dinâmico este processo.

Em 1990 o *US Department of Energy* e o *National Institute of Health* (NIH), ambos nos Estados Unidos, propuseram o Projeto Genoma Humano. A proposta se baseou em três frentes principais: o mapa genético, o mapa físico e, o mais ousado, a sequência completa de nucleotídeos (Marshall 1995; Marshall 1995; Venter, Adams et al. 1998). Durante os anos que desenrolavam o projeto da levedura *S. cerevisiae* e do Genoma Humano, vários outros genomas começaram a ser sequenciados e o primeiro genoma de um organismo não viral a ser completado foi o da bactéria *Haemophilus influenzae* (Fleischmann, Adams et al. 1995). Dentre outros, também foi sequenciado os genomas de importantes organismos modelo, como o da bactéria *Escherichia coli* (Daniels, Plunkett et al. 1992; Venter 1995) e do nematódeo *Caenorhabditis elegans* (Horvitz and Sulston 1990; Barstead, Kleiman et al. 1991; Sulston, Du et al. 1992; Williams, Schrank et al. 1992), ambos em 1992.

O genoma da bactéria *H. influenzae* foi sequenciado pelo grupo do pesquisador Craig Venter, que então trabalhava no *Institute for Genomic Research* (TIGR). Na época a estratégia adotada por Venter permitiu terminar o sequenciamento em tempo recorde. Eles dividiram o genoma em pedaços menores de 40.000 mil pares de bases e depois estes pedaços eram divididos e sequenciados por *shotgun*. Isto facilitou muito

Tabela 1. Alguns dos principais avanços conceituais e tecnológicos ocorridos desde o surgimento da ciência da genética.

Marcos Históricos na Genética			
1858	Charles Darwin "A teoria da evolução através da seleção natural"	1980	Frederick Sanger e colaboradores "É publicado o primeiro genoma completo o do bacteriófago φX174"
1865	Gregor Mendel "O conceito de hereditariedade"	1982	GenBank "O maior banco de dados de sequencias de acesso livre é estabelecido"
1928	Frederick Griffith "O DNA é o princípio transformante"	1983	Kary Mullis "A reação em cadeia da polimerase é descrita e revolucionou as metodologias de clonagem e de sequenciamento"
1941	George Beadle e Edward Lawrie Tatum "Um gene uma enzima"	1986	Applied Biosystems "O primeiro sequenciador automático entra no mercado"
1950	Erwin Chargaff "O número de Adeninas é igual ao número de timinas e o número de guaninas é igual ao número de citosinas"	1990	É anunciado nos Estados Unidos o projeto do Genoma Humano e neste mesmo ano Altshul e colaboradores descrevem o aplicativo computacional BLAST
1952	Alfred Hershey e Martha Chase "Trabalhando com fagos determinam definitivamente que o DNA é o material genético"	1995	Craig Venter "O genoma da bactéria *Haemophilus influenza* é publicado e é o primeiro genoma completo de um organismo de vida livre"
1953	James Watson e Francis Crick "A estrutura do DNA é desvendada"	1996	O genoma da levedura *Saccharomyces cerevisiae* é publicado
1957	Francis Crick "O dogma central da Biologia Molecular é proposto"	1997	O genoma da bactéria *Escherichia coli* é publicado
1970	Werner Arber e Hamilton Smith "A primeira enzima de restrição é isolada"	1998	O genoma do verme *Caenorhabditis elegans* é publicado
1971	Ray Tomlinson "O primeiro aplicativo da Internet é criado"	2000	O genoma da planta modelo Arabidopsis é publicado
1972	Paul Berg "A primeira molécula de DNA recombinante é criada"	2000	O genoma da mosca *Drosophila* é publicado
1973	Hebert Boyer e Stanley Cohen "A bactéria *Escherichia coli* é transformada com uma molécula de DNA recombinante carregando o gene de resistência ao antibiótico Kanamicina"	2001	O genoma humano é publicado
1974	Vint Cerf e Robert Kahn "É criado o conceito de internet e os protocolos para transferência *on line* de dados"	2002	Os genomas do camundongo e do peixe Fugu e um rascunho do genoma do arroz são publicados
1975	Erwin Southern "Técnica *southern* de transferência e imobilização de DNA em uma membrana foi desenvolvida"	2012	...?
1977	Frederick Sanger e Walter Gilbert "A técnica de sequenciamento de DNA por terminação de cadeia é desenvolvida"		

a montagem e acelerou a finalização do genoma de 1,8 milhões de pares de bases. Além do método desenvolvido no TIGR, os softwares escolhidos para montagem das sequencias foram decisivos para o sucesso do projeto (Sutton, White et al. 1995). A nova estratégia de sequenciamento do TIGR levou a publicação, nos anos subsequentes, de uma série de genomas completos de microorganismos. Entre estes genomas estavam o da bactéria *Helicobacter pylori* (Tomb, White et al. 1997) e o da Archaea *Methanococcus jannaschii* (Bult, White et al. 1996).

Em 1997, quase dez anos após o início, o genoma da levedura *S. cerevisiae* foi publicado e consistiu no primeiro genoma eucarioto sequenciado. Neste mesmo ano foi publicado o genoma da *E. Coli* K-12 (Blattner, Plunkett et al. 1997). Estes genomas foram um marco para a Ciência devido à importância destes organismos modelos para as áreas de Biotecnologia, Bioquímica, Biologia Molecular e Genética de organismos procariotos e eucariotos. Em setembro daquele ano 13 genomas já haviam sido sequenciados e outros quatro grandes projetos estavam em andamento: o do nematódeo *C. elegans* (1998), o da mosca-da-fruta *Drosophila melanogaster* (Adams, Celniker et al. 2000), o do camundongo *Mus musculus* (Waterston, Lindblad-Toh et al. 2002) e o humano *Homo sapiens* (Watson and Cook-Deegan 1991; Marshall 1995; Marshall 1995; Venter, Adams et al. 2001).

O Projeto Genoma Humano teve sua conclusão anunciada em 26 de junho de 2000. A imprensa mundial saudou o anúncio com grande empolgação, porém houve uma compreensão inadequada, pois a população ficou com a informação de que toda esta etapa estava vencida, quando sequer foi iniciada a totalidade de identificação de genes humanos em todos os cromossomos. O volume de interpretações corresponde ao de um texto de 800 volumes semelhantes ao de uma Bíblia, só que não se sabe em que idioma está escrito (http://www.bioetica.ufrgs.br/genoma.htm). As pesquisas concluíram que o genoma humano é formado por aproximadamente três bilhões de pares de bases distribuídos em 24 cromossomos. Para espanto de todos, apenas 3% do nosso genoma foi categorizado como passível de ser transcrito em moléculas de RNA, podendo então ser convertidos em proteínas. Esses dados aproximaram o ser humano de outros animais quanto à quantidade de genes funcionais, além de mostrar a semelhança de vários genes com os de outras espécies (bactérias, vírus, vermes, moscas, camundongos e chipanzés).

Trazendo esta cronologia genômica no contexto Brasil, o primeiro projeto genoma foi iniciado pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) em 1998 e, assim como idealizado por Goffeau, teve o formato de uma rede de pesquisa ligada virtualmente para a troca de dados, compartilhamento e divisão de tarefas. Na época o projeto proposto foi o sequenciamento da bactéria fitopatogênica *Xylella fastidiosa* (Simpson, Reinach et al. 2000), que causa uma doença conhecida como Clorose Variegada dos Citros (Amarelinho) em plantas do gênero *Citrus*. A rede paulista para sequenciamento de *X. fastidiosa* foi batizada de ONSA (sigla em inglês para Organização para Sequenciamento e Análise de Nucleotídeos), além de fazer uma analogia clara ao nosso genuíno felino, e ao TIGR nos Estados unidos. A Coordenação do projeto ficou com o britânico Andrew Simpson do Instituto Ludwig de Pesquisa sobre o Câncer e contou com a participação de 30 unidades de pesquisa no estado de São Paulo. A iniciativa da FAPESP foi um sucesso e o genoma de *X. fastidiosa* foi publicado dois anos depois, no ano 2000 e foi capa da revista *Nature* como o primeiro

genoma de um fitopatógeno a ser sequenciado (2010) – ver: http://www.lbi.ic.unicamp.br/xf/.

O Projeto Genoma *Xylella* contribuiu de maneira significativa para o desenvolvimento da Bioinformática no Brasil, além de contribuir para equipar importantes centros de pesquisas de ponta nas áreas de Bioquímica e Biologia Molecular. O custo do Projeto *Xylella* foi de aproximadamente 13 milhões de dólares, investidos não apenas em equipamentos de ultima geração para sequenciamento e análise funcional de genes mas também no suporte e capacitação/treinamento em todos os níveis acadêmicos, graduação, mestrado, doutorado e pós-doutorado. Este foi, sem sombra de dúvidas, um dos maiores benefícios para a Ciência, muitas vezes ocultados nas discussões sobre fomento em pesquisa avançada. Vários pesquisadores que trabalharam no Projeto Genoma *Xylella* são professores e ou pesquisadores dos mais importantes centros de pesquisa do Brasil e do exterior.

O sucesso da rede ONSA (ver: http://watson.fapesp.br/genoma.htm) levou o Ministério da Ciência e Tecnologia, por meio do CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) a criar a Rede Nacional do Projeto Genoma Brasileiro, que integra 25 centros de pesquisa distribuídos por todas as regiões brasileiras em um projeto genoma (ver: http://www.papociencia.ufsc.br/bio3.htm). O primeiro desafio da Rede Brasileira foi o sequenciamento de uma bactéria isolada no Rio Negro, Amazônia, *Chromobacterium violaceum* que foi concluído em 2001 e teve investimentos da ordem de 10 milhões de reais (2003).

Os projetos genomas transformaram-se num empreendimento multidisciplinar, que envolve recursos humanos, técnicos e financeiros de grande porte, primeiro com a iniciativa de instituições públicas espalhadas pelo mundo e depois com empresas que vendem os serviços de sequenciamento. Para um registro histórico em junho de 2012 estavam depositados no Genbank os genomas de 3.136 vírus, 12.018 procariotos e 1.830 eucariotos.

É importante lembrar que por mais novas e avançadas que sejam as técnicas de sequenciamento, elas só nos permitem saber a ordem dos nucleotídeos no genoma do organismo em questão. Para solucionar as questões que permeiam os problemas genéticos teve início a Era da Pós-Genômica, visando interpretar a função da sequência obtida nestes projetos. Neste sentido seguem-se os estudos da Mutassômica (Capítulo 7), Transcriptômica (Capítulo 8), Proteômica (Capítulo 9), Metabolômica (Capítulo 10) e mais recentemente a Fluxômica e Integrômica que se associam para formar a Biologia de Sistemas (Capítulo 13).

Bibliografias

(1998). "Genome sequence of the nematode C. elegans: a platform for investigating biology." Science **282**(5396): 2012-2018.

(2003). "The complete genome sequence of Chromobacterium violaceum reveals remarkable and exploitable bacterial adaptability." Proc Natl Acad Sci U S A **100**(20): 11660-11665.

(2010). "Brazil's biotech boom." Nature **466**(7304): 295.

Adams, M. D., S. E. Celniker, et al. (2000). "The genome sequence of Drosophila melanogaster." Science **287**(5461): 2185-2195.

BARSTEAD, R. J., L. KLEIMAN, et al. (1991). "Cloning, sequencing, and mapping of an alpha-actinin gene from the nematode Caenorhabditis elegans." Cell Motil Cytoskeleton **20**(1): 69-78.

BLATTNER, F. R., G. PLUNKETT, 3rd, et al. (1997). "The complete genome sequence of Escherichia coli K-12." Science **277**(5331): 1453-1462.

BULT, C. J., O. WHITE, et al. (1996). "Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii." Science **273**(5278): 1058-1073.

CHEE, M. S., A. T. BANKIER, et al. (1990). "Analysis of the protein-coding content of the sequence of human cytomegalovirus strain AD169." Curr Top Microbiol Immunol **154**: 125-169.

DANIELS, D. L., G. PLUNKETT, 3rd, et al. (1992). "Analysis of the Escherichia coli genome: DNA sequence of the region from 84.5 to 86.5 minutes." Science **257**(5071): 771-778.

DUJON, B. (1993). "Mapping and sequencing the nuclear genome of the yeast Saccharomyces cerevisiae: strategies and results of the European enterprise." Cold Spring Harb Symp Quant Biol **58**: 357-366.

FLEISCHMANN, R. D., M. D. ADAMS, et al. (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd." Science **269**(5223): 496-512.

GOFFEAU, A. (1994). "Yeast. Genes in search of functions." Nature **369**(6476): 101-102.

GOFFEAU, A., B. G. BARRELL, et al. (1996). "Life with 6000 genes." Science **274**(5287): 546, 563-547.

HORVITZ, H. R. and J. E. SULSTON (1990). ""Joy of the worm"." Genetics **126**(2): 287-292.

MARSHALL, E. (1995). "Human genome project. Emphasis turns from mapping to large-scale sequencing." Science **268**(5215): 1270-1271.

MARSHALL, E. (1995). "A strategy for sequencing the genome 5 years early." Science **267**(5199): 783-784.

PAOLETTI, E., R. L. Weinberg, et al. (1984). "Genetically engineered poxviruses: a novel approach to the construction of live vaccines." Vaccine **2**(3): 204-208.

RONAGHI, M. (2001). "Pyrosequencing sheds light on DNA sequencing." Genome Res **11**(1): 3-11.

SANGER, F., G. M. Air, et al. (1977). "Nucleotide sequence of bacteriophage phi X174 DNA." Nature **265**(5596): 687-695.

SANGER, F., A. R. COULSON, et al. (1980). "Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing." J Mol Biol **143**(2): 161-178.

SIMPSON, A. J., F. C. REINACH, et al. (2000). "The genome sequence of the plant pathogen Xylella fastidiosa. The Xylella fastidiosa Consortium of the Organization for Nucleotide Sequencing and Analysis." Nature **406**(6792): 151-159.

SINGER, M. and P. BERG (2004). "George Beadle: from genes to proteins." Nat Rev Genet **5**(12): 949-954.

SULSTON, J., Z. Du, et al. (1992). "The C. elegans genome sequencing project: a beginning." Nature **356**(6364): 37-41.

TOMB, J. F., O. WHITE, et al. (1997). "The complete genome sequence of the gastric pathogen Helicobacter pylori." Nature **388**(6642): 539-547.

VENTER, J. C. (1995). "E. coli sequencing." Science **267**(5198): 601.

VENTER, J. C., M. D. ADAMS, et al. (2001). "The sequence of the human genome." Science **291**(5507): 1304-1351.

VENTER, J. C., M. D. ADAMS, et al. (1998). "Shotgun sequencing of the human genome." Science **280**(5369): 1540-1542.

WATERSTON, R. H., K. LINDBLAD-TOH, et al. (2002). "Initial sequencing and comparative analysis of the mouse genome." Nature **420**(6915): 520-562.

WATSON, J. D. and R. M. COOK-DEEGAN (1991). "Origins of the Human Genome Project." FASEB J **5**(1): 8-11.

WILLIAMS, B. D., B. SCHRANK, et al. (1992). "A genetic mapping system in Caenorhabditis elegans based on polymorphic sequence-tagged sites." Genetics **131**(3): 609-624.

2

Sequenciando genomas

Juliana Lopes Rangel Fietto
Talles Eduardo Ferreira Maciel

Introdução

O sequenciamento genômico é uma técnica que permite identificar, na ordem correta, a sequência de nucleotídeos de uma molécula de DNA ou RNA, visando conhecer a informação genética contida nesta estrutura. As metodologias responsáveis por tal façanha fornecem, para cada uma das bases determinadas, uma informação referente a sua qualidade (confiabilidade).

Desde o desenvolvimento das primeiras metodologias de sequenciamento (no final da década de 70) até as tecnologias atuais, denominadas de "Sequenciamento de Nova Geração" (New Generation Sequencing - NGS); passamos da escala de sequenciamento manual de poucos kilobases para o seqüenciamento maciço e paralelo de genomas inteiros e em curto período de tempo.

Neste capítulo iremos discutir algumas das metodologias de sequenciamento mais utilizadas, focando em seus princípios, peculiaridades, aplicações, vantagens e desvantagens. Além disto, será apresentado, sucintamente, tecnologias ainda em desenvolvimento, classificadas como de terceira geração.

De forma geral, o sequenciamento é feito a partir de moléculas de DNA advindas diretamente do DNA genômico (aquele que contém a maior parte da informação genética dos organismos) ou de outras moléculas de DNA celular como: DNA mitocondrial, DNA cloroplastídico, DNA plasmidial, dentre outros.

O sequenciamento, seguido de uma boa **montagem** das sequencias obtidas, permite obter informações referentes a: expressão gênica diferencial, estrutura e função dos genes, diversidade genética, presença de elementos móveis no genoma, presença de genes adquiridos por transferência lateral, relações evolutivas, além de permitir a construção de mapas metabólicos dentre outras (Nierman et al., 2000).

Não é possível através do sequenciamento do DNA genômico, obter informação referente a quais genes estão sendo expressos no momento do ensaio e, em que nível

de expressão, estes genes se encontram. Este tipo de informação é importante por vários motivos: para saber se um determinado gene é importante numa situação específica ou para saber se o mesmo sofre algum tipo de regulação da expressão ao nível transcricional. Para conseguir estas informações, seria necessário sequenciar os RNA mensageiros (RNAm), que são RNAs que podem ou não ser traduzidos em proteínas funcionais nas células. Como não é possível sequenciar diretamente fragmentos de RNA; torna-se necessário isolar os mesmos e transcrevê-los de forma reversa, através do uso de uma enzima específica (transcriptase reversa) em cDNA, que corresponde a parte codificante dos RNAs mensageiros). A retrotranscrição também é válida para situações em que se queira sequenciar outros tipos de RNAs.

As metodologias de sequenciamento descritas podem ser utilizadas para sequenciar fragmentos de cDNA, como é o caso do item 4.1 (ESTs) e 4.2 ORESTES, que serão descritas nas próximas seções; ou de DNA genômico.

Uma característica comum à maioria das tecnologias de sequenciamento atuais é a limitação do tamanho dos fragmentos de DNA sequenciados, ou seja, de forma geral os sequenciadores ainda são incapazes de sequenciar fragmentos de DNA longos. Esta realidade nos obriga a fragmentar moléculas grandes de DNA, como DNA genômico; ou em outros casos nos limita a isolar apenas alguns fragmentos de interesse a serem sequenciados.

Iniciaremos agora a descrição das tecnologias de sequenciamento mais utilizadas mundialmente. Estas serão separadas em dois grandes grupos: as tecnologias de pequena escala, que proporcionaram os primeiros sequenciamentos de DNAs; e as novas tecnologias de sequenciamento em larga escala. Embora estas sejam mais robustas em diversos aspectos, o sequenciamento em pequena escala é continuado por motivos que serão discutidos à medida que as tecnologias forem sendo discutidas.

Metodologias de sequenciamento em pequena escala

Entre 1800 e 1900, as proteínas foram consideradas as moléculas mais importantes dentre os constituintes celulares. No entanto, a primeira sequencia protéica só foi sequenciada em 1953. Neste mesmo ano, Watson e Crick propuseram o modelo de dupla hélice do DNA, iniciando uma nova era no estudo do DNA (Watson e Crick, 1953).

Apesar dos avanços, era muito difícil até o começo da década de 70, obter a sequencia de nucleotídeos de um fragmento de DNA, por menor que fosse. Este problema foi resolvido com o surgimento em 1977 de duas tecnologias: uma desenvolvida por Alan Maxam e Walter Gilbert (baseada em hidrólise química) e outra por Frederick Sanger e cols. (baseada em reações enzimáticas), que permitiram determinar a sequencia de nucleotídeos de fragmentos maiores de DNA. Estas metodologias revolucionoram as pesquisas científicas e se difundiram rapidamente pelo mundo, sendo a base da Genômica (Sanger et al., 1977).

Sequenciamento químico de Maxam-Gilbert

Após divulgada, esta metodologia foi amplamente utilizada por proporcionar a obtenção da sequência de nucleotídeos de fragmentos maiores de DNA.

A técnica desenvolvida por eles utiliza marcação do DNA alvo a ser sequenciado com fósforo radioativo (P^{32}). O P^{32} é inicialmente ligado ao dATP formando P^{32}-dATP que é incorporado, pela enzima polinucleotídeo kinase, ao DNA a ser sequenciado. Esta incorporação pode ser tanto na extremidade 5' quanto na extremidade 3', ficando a critério do executor da técnica. Neste método, o rompimento das pontes de hidrogênio da fita dupla de DNA ocorre pela adição de dimetil sulfato e aquecimento a 90º C (Figura 1).

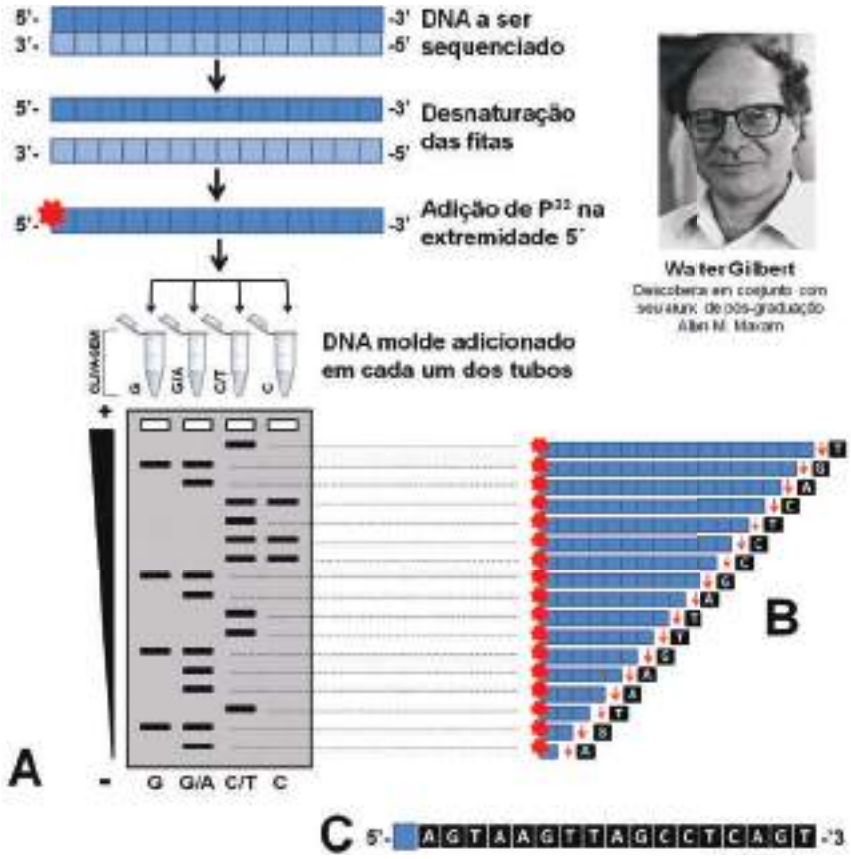


Figura 1. Metodologia de sequenciamento proposta por Maxam-Gilbert. Em **(A)** adiciona-se um fosfato radioativo numa das extremidades após a separação da dupla fita do DNA (ver detalhes na Figura 2). Em seguida, o DNA marcado é colocado em quatro tubos, onde ocorre a clivagem do mesmo, através da utilização de compostos químicos, em posições específicas (antes dos "G"s, antes de G ou A, antes de C ou T e antes dos "C"s). Para identificar a sequencia de nucleotídeos do DNA, aplica-se o produto das quatro reações em canaletas diferentes do gel. Após a separação, o perfil de bandas obtidas deve ser lido de baixo para cima, uma a uma representando os nucleotídeos. Observe que quando uma banda aparece em G/A e ao mesmo tempo só em G, significa que o nucleotídeo da respectiva posição é o G. Caso a banda seja observada apenas em G/A, então o nucleotídeo da respectiva posição é o A. A mesma lógica segue para C/T e T. **(B)** Representação esquemática evidenciando o último nucleotídeo de cada um dos fragmentos do gel. **(C)** Sequencia de nucleotídeos do DNA sequenciado.

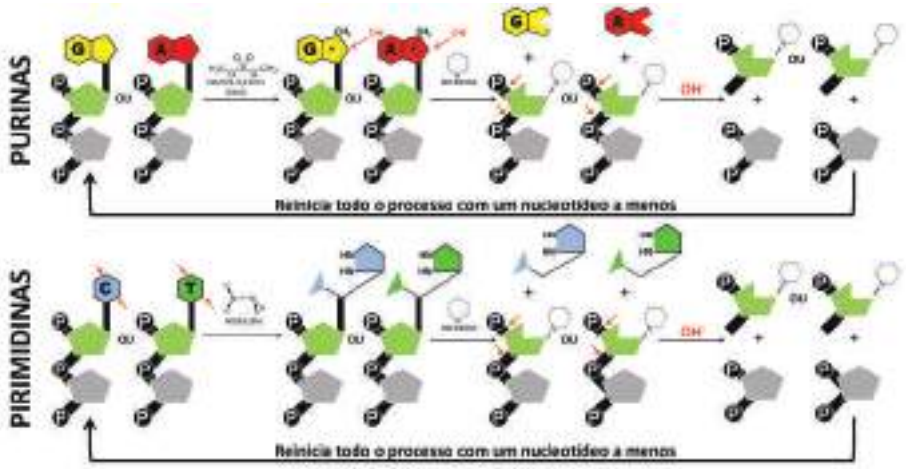


Figura 2. Processo de incorporação de DMS ou hidrazina pelas bases nitrogenadas durante o processo se sequenciamento proposto por Maxam-Gilbert. Observe que os processos são análogos para as bases púricas e pirimídicas (painéis superior e inferior), com formação de um intermediário de bases púricas ligados a DMS ou bases pirimídicas ligadas a hidrazina. Estes intermediários quando tratados com piperidina promovem uma modificação na estrutura da pentose, com consequente liberação da base nitrogenada modificada. Finalmente, as ligações fosfodiesteres são rompidas (setas vermelhas), liberando consequentemente o nucleoísídeo terminal, permitindo que o ciclo possa ser reiniciado.

Tabela 1. Compostos químicos utilizados na técnica de Maxam-Gilbert e especificidades pelas bases nitrogenadas.

Especificidade de bases	Modificador de bases	Remover de bases	Químico para clivar a fita
G	Dimetil sulfato	Piperidina	Piperidina
A + G	Ácido	Ácido	Piperidina
C + T	Hidrazina	Piperidina	Piperidina
C	Hidrazina + álcali	Piperidina	Piperidina
A > C	Álcali	Piperidina	Piperidina

O princípio básico desta técnica consiste na clivagem do DNA alvo marcado, através da utilização de compostos químicos, em posições específicas (antes dos "G"s, antes de "A" ou "G", antes de "C" ou "T" e antes dos "C"s). Acima encontra-se uma tabela com os compostos químicos utilizados nas diferentes etapas deste método.

A posição a ser quebrada depende do composto químico que é adicionado, num só tipo, a um dos quatro tubos contendo o DNA molde a ser sequenciado (Maxam e Gilbert, 1977). Como resultado, tem-se após a fragmentação um conjunto de fragmentos de diferentes tamanhos em cada um dos quatro tubos. As bandas geradas após a "corrida" destes fragmentos em gel de poliacrilamida podem ser visualizadas após a impressão de uma chapa radiográfica. A determinação da sequencia de nucleotídeos é obtida "lendo-se" de baixo para cima, um a um, os nucleotídeos representados pelas bandas do gel (Figura 1).

O método de Sanger, discutido em seguida, com seus aprimoramentos fez com que o método de Maxam-Gilbert não fosse utilizado por muito tempo em virtude de suas desvantagens.

Método de Sanger

Assim como a metodologia proposta por Maxam e Gilbert, a técnica de sequenciamento desenvolvida por Sanger, em 1977, também utiliza marcação radioativa.

A diferença é que a primeira marcava diretamente o DNA a ser sequenciado enquanto a de Sanger, marcava os fragmentos de DNA sintetizados a partir da fita molde. A síntese de novos fragmentos de DNA a partir da fita molde só foi possível graças ao desenvolvimento da técnica de PCR (Reação em Cadeia da Polimerase) (Mullis et al., 1986), que consiste na síntese *in vitro* de uma fita de DNA complementar a um DNA molde, utilizando os seguintes componentes básicos da replicação celular:

- Cópias do DNA molde que deverá ser sequenciado, apresentando relativo grau de pureza.
- Enzima DNA polimerase capaz de produzir cópias relativamente fiéis do DNA molde.
- Um DNA iniciador (primer) que propicia o início da extensão pela DNA polimerase.
- Os desoxinucleotídeos que são as unidades básicas para construção da fita complementar ao DNA molde. São eles: dATP, dCTP, dGTP e dTTP.
- Solução tampão, contendo o co-fator magnésio (Mg), necessário para que a enzima DNA polimerase desempenhe sua atividade.

Por fim, é necessário ainda a presença de didesoxinucleotídeos (ddATP, ddCTP, ddGTP e ddTTP), que atuam como terminadores da síntese de DNA. A chance dos desoxi ou didesoxinucleotídeos serem incorporados numa determinada posição da cadeia de DNA nascente é a mesma, uma vez que a DNA polimerase não consegue distinguir estes dois nucleotídeos pelo fato da diferença entre eles ser apenas a ausência do grupo OH na posição 3´ (Figura 3).

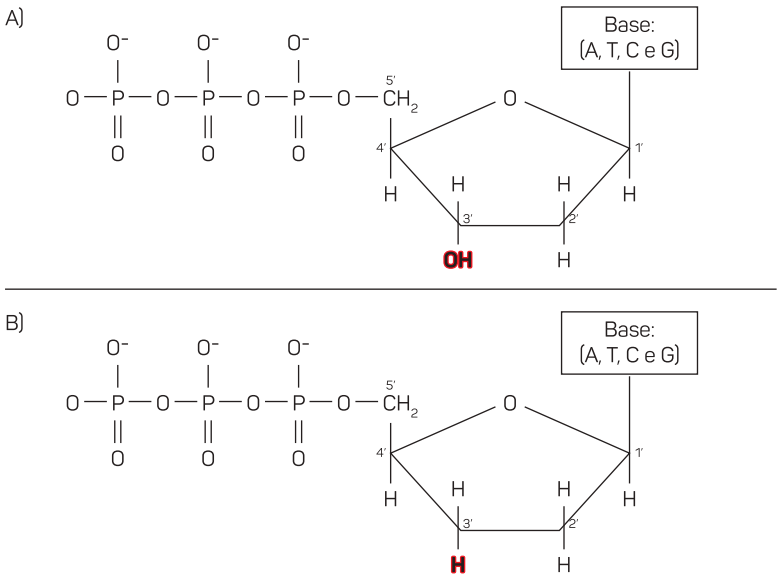


Figura 3. Diferença entre desoxinucleotídeo e didesoxinucleotídeo. Em (A), temos um desoxinucleotídeo contendo três grupos fosfato, uma ribose com uma hidroxila na posição 3' (vermelho) e uma das quatro bases nitrogenadas. Em (B), temos o didesoxinucleotídeo evidenciando a ausência da hidroxila na posição 3' da ribose (vermelho).

No entanto, esta diferença é suficiente para bloquear a síntese da cadeia de DNA nascente. A explicação é simples: se um desoxinucleotídeo (que é o substrato normal da DNA polimerase) é adicionado, a síntese da cadeia de DNA continua, pois haverá após sua incorporação na molécula de DNA nascente a presença de uma hidroxila livre na posição 3′, onde deverá ser ligado o próximo desoxinucleotídeo. Por outro lado, se um didesoxinucleotídeo for adicionado à cadeia nascente de DNA, a síntese da mesma será interrompida neste ponto, pois a ausência do grupo OH na posição 3′ impede a entrada de um novo nucleotídeo (por isso este método é também conhecido como "terminador de cadeia" ou "didesoxi").

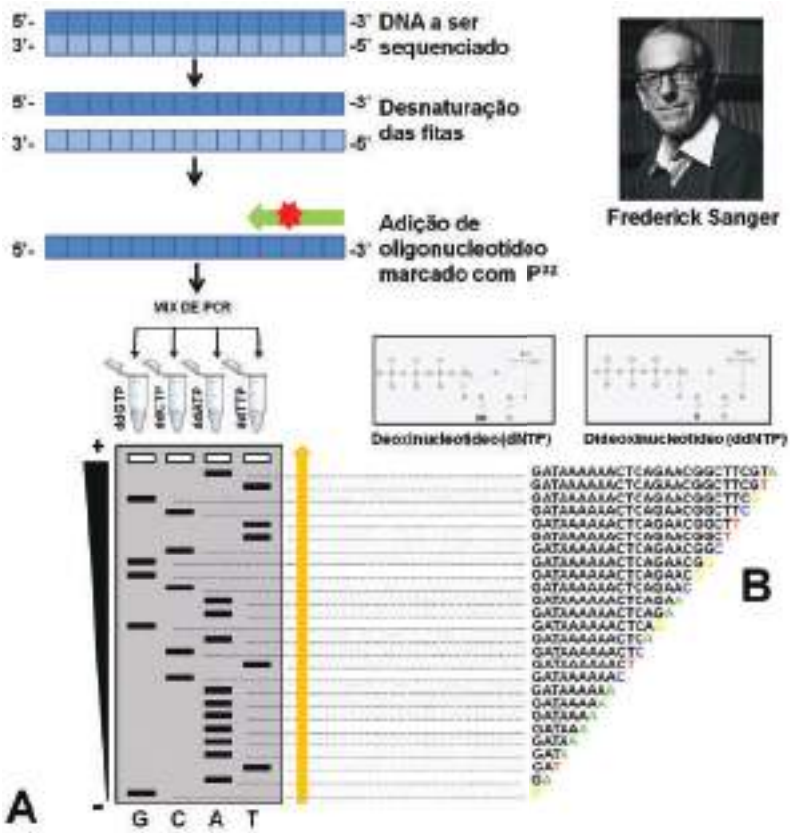


Figura 4. Sequenciamento pelo método de Sanger. Com base no processo de sequenciamento proposto por Maxam-Gilbert, Sanger reduziu os trabalhos e facilitou o processo, tornando a dinâmica de sequenciamento mais rápida. Para isso incorporou os reagentes padrões para uma técnica de PCR, com a adição em tubos isolados dos respectivos didesoxinucleotídeo (ddNTP). Cada vez que um ciclo de PCR era finalizado um fragmento de DNA de tamanho distinto poderia ser formado, e esta variação dependia da incorporação de um ddNTP ou desoxinucleotídeo (dNTP). Se acaso o ddNTP fosse incorporado, por falta da hidroxila na posição 3 da pentose, o processo de extensão da cadeia crescente era interrompido (comparar estruturas). Após inúmeros ciclos de PCR, amostras dos respectivos tubos, contendo especificamente seus respectivos ddNTPs são aplicadas em canaletas individualizadas do gel de acrilamida. Feita a eletroforese, o perfil de bandas, lidas de baixo pra cima, determina a correta sequencia da molécula de DNA em sua fita complementar, logo o reverso complementar poderia ser definido.

O princípio da técnica consiste em marcar radioativamente alguns dos desoxinucleotídeos livres em solução ou o primeiro desoxinucleotídeo do primer com P^{32} ou S^{35}. Após incorporação na cadeia de DNA nascente, estes átomos marcados emitem radiação que é utilizada para impressão de uma chapa radiográfica, permitido desta forma, visualizar os fragmentos resultantes da amplificação.

A técnica se desenvolve da seguinte maneira: primeiro o DNA fita dupla é desnaturado e utilizado para montar quatro reações independentes contendo os mesmos reagentes, com exceção dos didesoxinucleotídeo, que são adicionados separadamente (um determinado tipo em cada reação) (Figura 4). Após um determinado tempo de reação, considerando que nada dirige a entrada de desoxi ou didesoxinucleotídeos na cadeia de DNA nascente e que os mesmos são colocados em excesso na reação, será produzido um conjunto de fragmentos complementar ao DNA molde com tamanhos variados. Sendo o tamanho de cada fragmento dependente da posição onde o didesoxinucleotídeo terminador foi adicionado. Se pensarmos que existem na mistura muitas moléculas do mesmo DNA molde, compreenderemos que todas as posições do DNA molde, em algum momento, terá um dNTP, hora um ddNTP complementar. Assim, teremos amplicons (produto da PCR) terminando em diferentes posições do DNA molde.

O produto heterogêneo de cada uma das quatro reações é aplicado em canaletas diferentes do gel, que frequentemente, tem a poliacrilamida como matriz. Devido ao alto poder de resolução (separação dos fragmentos) deste gel, é possível separar e visualizar fragmentos que diferem entre si por apenas um nucleotídeo. As bandas produzidas são visualizadas numa chapa radiográfica após sua impressão. Assim como no método anterior, a análise da ordem das bandas na chapa radiográfica começa pelo final do gel, permitindo determinar a sequencia de nucleotídeos da fita de DNA recém- sintetizada (Figura 4).

Esta técnica permitiu inicialmente separar de 200 a 300 nucleotídeos por corrida, sendo considerada uma revolução na época em que foi descoberta.

Aprimoramento do método de Sanger

Método semi-automatizado

A ciência não pára e está sempre buscando novas descobertas, que na maioria das vezes surgem para melhorar a vida de todos nós. Não foi diferente com a metodologia de sequenciamento proposta por Sanger. Classificada como manual por não utilizar o computador em nenhuma de suas etapas, esta metodologia foi aperfeiçoada originando o método semi-automatizado, que é a base de muitas metodologias de sequenciamento atuais. A idéia de automatizar o sequenciamento foi proposta por Lloyd M.Smith, Mike Hunkapiller e Tim Hunkapiller na universidade privada do estado da Califórnia.

O princípio do método proposto por Sanger permaneceu o mesmo. No entanto, a técnica foi aprimorada ficando mais simples, rápida e segura por não utilizar compostos radioativos prejudiciais a saúde humana. Mas que mudança foi esta que trouxe tantas melhorias a técnica, fazendo com que dominasse as três décadas

seguintes? A principal modificação foi a adição aos didesoxinucleotídeos, de corantes capazes de emitir fluorescência quando excitados em comprimento de onda específico.

No início, Smith mostrou-se pessimista quanto a exequibilidade do método, temendo que a quantidade de corantes adicionados aos didesoxinucleotídeos fosse insuficiente para ser detectada pelo computador. No entanto este problema foi rapidamente resolvido pela utilização de corantes especiais, que emitem luz ao serem atravessados por feixe de raios laser. O método aprimorado utiliza fluoróforos diferentes para cada um dos quatro tipos de didesoxinucleotídeos, que ao serem excitados, emitem luz característica do didesoxinucleotídeo incorporado.

Utilizaremos aqui o mesmo raciocínio apresentado no método inicial de Sanger: se pensarmos que existem na reação muitas moléculas do mesmo DNA molde, compreenderemos que todas as posições deste DNA terão em algum momento, hora um dNTP, hora um ddNTP incorporado pela DNA polimerase durante a PCR. Assim, teremos amplicons terminando em diferentes posições do DNA molde.

Como conseqüência da incorporação dos didesoxinucleotídeos marcados com fluorescência, as quatro reações passaram a ocorrer num tubo único e seu conteúdo podia agora ser aplicado numa única canaleta do gel. Este fato fez com que o número de amostras analisadas por corrida fosse quatro vezes maior, considerando que no método radioativo eram necessárias quatro canaletas do gel para obter o mesmo resultado que o novo método conseguia em uma canaleta (Prober et al., 1987; Macbeath et al., 2001).

Podemos citar o ABI 377 como exemplo de sequenciador que utiliza este método. Este sequenciador detecta a fluorescência emitida pelos didesoxinucleotídeos e a decodifica para determinar a sequencia de nucleotídeos do fragmento de interesse, sendo este método considerado semi-automatizado, pois o produto das PCRs precisa ser aplicado pelo analista. Este sequenciador possibilitou sequenciar 48 fragmentos de DNA num intervalo de 5 a 6 horas.

Método automatizado

Nos anos 90 os géis (de difícil manuseio) foram substituídos por finíssimos capilares preenchidos com gel onde os fragmentos de DNA são separados em altíssima velocidade. Os sequenciadores baseados neste sistema são, aproximadamente, duas vezes mais rápidos do que os semi-automatizados. As amostras são aplicadas, através de um sistema de eletroinjeção diretamente nos capilares, diminuindo consideravelmente o trabalho do analista. Para termos uma idéia do nível de automação dos sequenciadores de capilares atuais, 15 minutos de intervenção humana a cada 24 horas é suficiente para produzir aproximadamente meio milhão de pares de bases.

Após a eletroinjeção, os fragmentos começam a migrar e encontram, num determinado ponto, um feixe de raios laser que excita os fluoróforos presentes na extremidade 3′ de cada fragmento fazendo com que estes emitam fluorescência característica de um dos quatro tipos de fluoróforos. Um detector registra esta fluorescência e a transmite para um computador que possui um software capaz de converter fluorescência em picos coloridos, sendo utilizado uma única cor para cada um dos quatro tipos de nucleotídeos (verde para adenina, preto para guanina, azul para

citosina e vermelho para timina). Este procedimento é efetuado para cada fragmento no gel. No final do processo, o software gera um cromatograma que corresponde a sequencia de DNA complementar ao DNA molde utilizado (Figura 5).

O sequenciador MegaBace® é capaz de sequenciar 96 fragmentos de DNA num intervalo de 1 a 3 horas. O sistema continuou sendo aperfeiçoado, sendo possível hoje sequenciar até 384 amostras em uma única corrida. Na Figura 6, temos exemplos de sequenciadores que utilizam capilares.

O método de sequenciamento automatizado foi utilizado para sequenciar o genoma humano (Figura 7). Nas próximas seções serão discutidas as principais metodologias, baseadas no método automatizado, utilizadas para sequenciar genomas completos ou trechos destes (seção 3) e regiões de interesse do exoma (parte codificante do genoma) (seção 4).

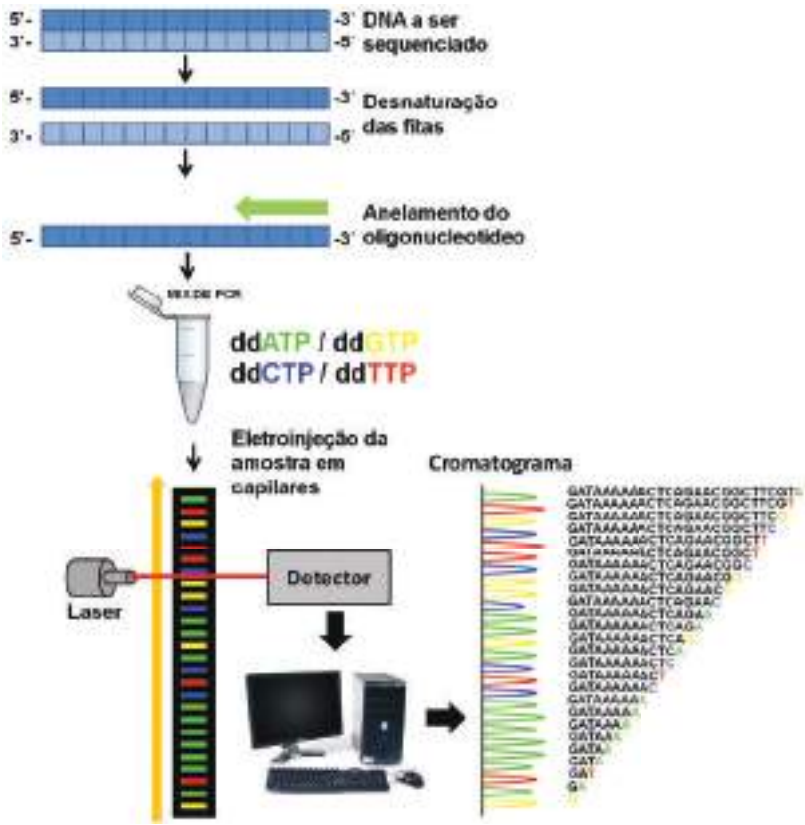


Figura 5. Automatização do método de Sanger. Após a desnaturação da dupla fita, didesoxinucleotídeos marcados com compostos florescentes são incorporados à cadeia nascente de DNA sintetizada pela DNA polimerase. Através de um sistema de eletroinjeção, os fragmentos de DNA recém-sintetizados começam a migrar e encontram, num determinado ponto, um feixe de raios laser que excita os fluoróforos fazendo com que estes emitam fluorescência característica de um dos quatro tipos de nucleotídeos. Um detector registra a intensidade e comprimento de onda desta fluorescência e a transmite a um computador que possui um software capaz de converter fluorescência em picos coloridos (cromatograma) que são decodificados na sequencia de nucleotídeos do fragmento.

Figura 6. Modelos de sequenciadores automáticos que fazem uso da técnica de Sanger.

Estratégias de sequenciamento de DNA

A técnica de sequenciamento automatizada, descrita anteriormente, permite sequenciar com qualidade, aproximadamente 700 nucleotídeos consecutivos de um fragmento. Assim, quando o objetivo é o sequenciamento de genomas, seja de organismos simples como bactérias ou organismos complexos como o homem, torna-se necessário: "picotar" o DNA em fragmentos menores, sequenciar os pedacinhos obtidos e depois sobrepô-los em busca do genoma completo. As técnicas de fragmentação são várias, dentre as quais destacamos: uso de enzimas de restrição de corte freqüente, como *Alu1*; e quebra aleatória por fragmentação mecânica do genoma a ser sequenciado (shotgun). A última é mais utilizada e será o foco deste capítulo.

Shotgun

Consiste em "atirar no escuro" (do inglês, *shotgun*), ou seja, bombardear aleatoriamente o genoma a ser sequenciado com partículas que promovem sua fragmentação (Figura 7A e 7B). Após este passo, obtêm-se a biblioteca genômica pela inserção de cada fragmento (inserto) num vetor apropriado (processo conhecido como clonagem) (Figura 7C). Fragmentos clonados pequenos (no máximo 1400 pb) podem ser completamente sequenciados somente com o uso de primers que anelam no vetor. O processo é diferente para grandes fragmentos clonados devido a limitada processividade da DNA polimerase e limitada resolução do gel de acrilamida. Assim, temos como resultado, no caso de fragmentos grandes, duas sequencias (uma para cada extremidade) ainda ligadas ao vetor (Figura 7D).

A técnica descrita até este ponto, conhecida como shotgun de genoma inteiro (WGS) (Figura 7, painel da esquerda), foi utilizada pela empresa Celera Genomics para sequenciar o genoma humano. Esta empresa utilizou uma estratégia extremamente elegante para sequenciar o genoma humano que foi a clonagem de fragmentos de tamanhos diferentes (2 mil, 10 mil e 130 mil pares de bases) em vetores apropriados, produzindo três bibliotecas que tiveram as extremidades de seus insertos sequenciados. Durante a montagem, uma determinada sequencia de DNA (chamada comumente de "read") vai encontrar região de sobreposição com outro read qualquer da biblioteca. Se considerarmos que os mesmos pertencerem à mesma biblioteca, entenderemos que eles tem o mesmo tamanho de vetor. Assim, a sequencia correspondente a extremidade do primeiro *read* também irá se brepor a extremidade do outro *read* (Figura 7E). O conhecimento da distância entre extremidades de insertos grandes

direciona a montagem de regiões repetitivas além de proporcionar ligar e ordenar os conjuntos de reads alinhados, denominados de "contigs", formando assim sequencias ainda maiores denomidas como "scaffolds" (Figura 7E).

O consórcio público responsável pelo sequenciamento do genoma humano também utilizou o shotgun durante o sequenciamento. No entanto a metodologia utilizada pelo consórcio foi um pouco diferente, sendo conhecido como shotgun hierárquico

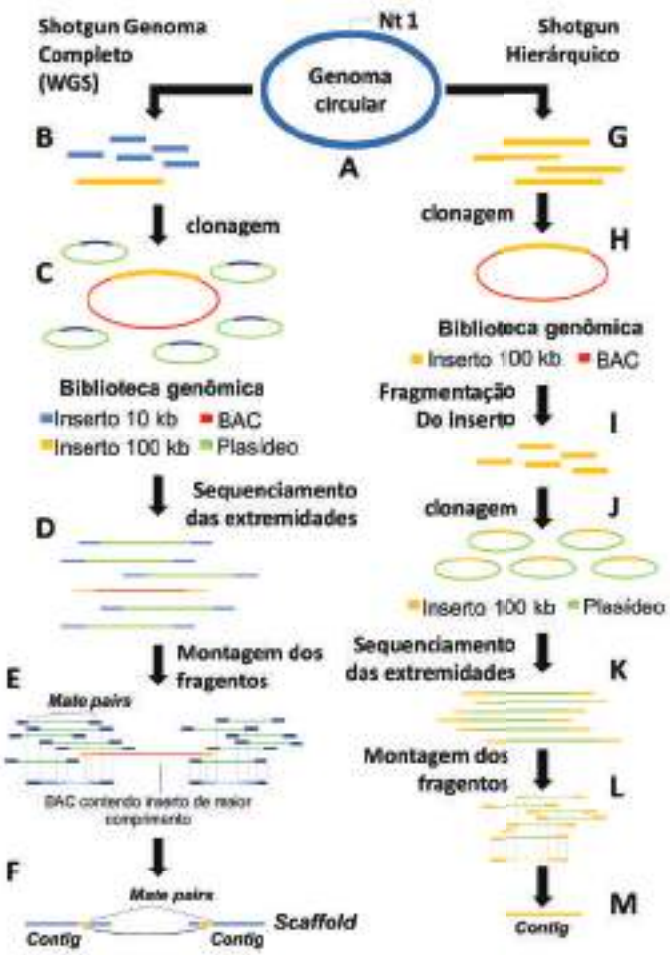


Figura 7. Exemplo da técnica de *shotgun* usada em um genoma inteiro - WGS (painel da esquerda - B a F) e a técnica de *shotgun* hierárquico (painel da direita - G a M). No WGS o DNA total do organismo, representado no modelo como um genoma circular bacteriano, é fragmentado, clonado e sequenciado. Fragmentos grandes (em amarelo) são essenciais ao processo de montagem (E) por permitir identificar e ligar *contigs* adjacentes. Na técnica de *shotgun* hierárquico (também conhecido como sub-biblioteca) também há fragmentação do material genético, só que os fragmentos gerados são grandes, e por isso precisam ser primeiro clonados em BACs. Posteriormente, este inserto é novamente clivado em fragmentos menores e subclonados em plasmídeos. A partir deste ponto a técnica é idêntica ao WGS. No primeiro tipo de *shotgun* temos o sequenciamento do genoma total do organismo, ao passo que no *shotgun* hierárquico, sequencia-se apenas uma região (fragmento) de interesse.

ou “clone por clone” (Figura 7: painel da direita). Nesta metodologia, os fragmentos obtidos da clivagem são muito grandes (cerca de 150 mil pb) e por isso precisam ser clonados em BACs (cromossomo artificial de bactéria) (Figura 7H). Posteriormente, os insertos clonados são clivados e subclonados em plasmídeos (Figura 7I e 7J). Após o sequenciamento, obtêm-se as sequencias correspondentes às extremidade dos fragmentos ligadas ao vetor (Figura 7K).

Neste ponto, vocês devem estar se perguntando: e para montar estes milhares de fragmentos sequenciados? Para conseguir tal proeza foi necessário criar programas de bioinformática que tem como objetivo montar a maior sequencia possível (Figura 7F) utilizando pequenas regiões de sobreposição entre os fragmentos sequenciados, como se fosse um quebra cabeças (Figura 7E).

Uma diferença entre estas duas estratégias é que a primeira utiliza, de uma só vez, todas as bases sequenciadas para montar o genoma. Já a segunda utiliza as 150 mil bases sequenciadas para montar o fragmento inicial inserido no BAC.

O WGS é uma técnica mais simples por utilizar menos etapas (uma só clonagem), sendo utilizada com sucesso no sequenciamento do primeiro genoma bacteriano (Haemophilus influenzae) e posteriormente, no sequenciamento de diversos outros genomas, destacando o humano (Venter et al., 2001).

A técnica de shotgun hierárquico é mais laboriosa, sendo apropriada para montagem de regiões repetitivas do genoma e para montagem de genomas grandes e complexos.

Primer Walking

O próprio nome (“andamento” do primer) remete o princípio dá técnica que consiste em sequenciar o DNA clonado maior em várias etapas, já que a limitação de cerca de 700 pb por sequenciamento não permite o sequenciamento completo de fragmentos maiores do que aproximadamente 1400 pb com alta confiança. Assim utiliza-se a estratégia de sequenciar o início das extremidades usando primers que anelam no vetor e dar continuidade ao sequenciamento a partir de novos primers desenhados para o fim das sequencias primariamente obtidas. Assim, primers que anelam ao vetor são utilizados permitindo o sequenciamento de uma das extremidades do fragmento de interesse. Um novo primer capaz de anelar ao fragmento sequenciado é desenhado iniciando o sequenciamento de uma região mais distante da extremidade do fragmento. Esse processo é repetido várias vezes até que toda a extensão dos fragmentos seja sequenciada (Figura 8).

Os primers devem ser cuidadosamente desenhados, pois a última região sequenciada deve se sobrepor aos fragmentos sequenciado anteriormente em aproximadamente 100 pb.

Estratégias de sequenciamento de RNA

Outras abordagens surgiram para sequenciar somente os genes expressos. Quando falamos em genes expressos devemos logo pensar nos RNAs que estão sendo expressos num determinado momento do desenvolvimento celular.

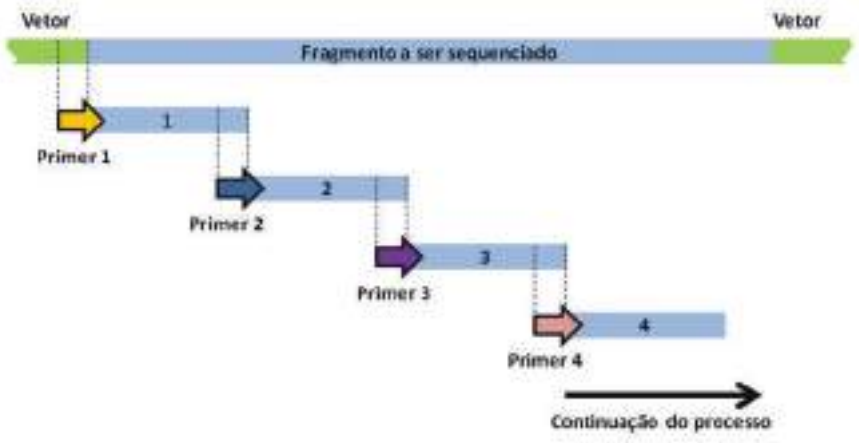


Figura 8. Modelo esquemático da técnica de *Primer Walking*. Um oligonucleotídeo inicial é usado no processo, determinando a sequencia a jusante após uso da técnica de sequenciamento. Como existe uma limitação da ação da polimerase e devido a baixa resolução dos géis de sequenciamento, haverá um momento em que os nucleotídeos não serão mais determinados, reduzindo a eficiência do processo. A partir desta nova sequencia determinada (1), desenha-se um novo oligonucleotído nas proximidades da posição 3′e tem-se inicio a um novo processo de sequenciamento, gerando o fragmento 2, que servirá para molde do desenho de um novo oligonucleotídeo, e assim por diante continua-se o processo. Repare que é como se fossem dados pequenos passos para se conhecer a sequencia completa, daí a definição *primer walking*.

O transcriptoma é o conjunto de transcritos e suas quantidades num estágio específico do desenvolvimento ou condição fisiológica (CAPÍTULO 8). A descoberta de novos transcritos, assim como a quantificação destes de fundamental importância para entender os fenômenos biológicos na célula. O sequenciamento do RNA tem permitido mensurações mais precisas do nível destes transcritos (Wang et al., 2009).

Na introdução, foi visto que não é possível sequenciar diretamente o RNA. Este problema é resolvido com a abordagem apresentada na Figura 9, na qual temos inicialmente três mRNAs processados (sem íntrons e com a cauda poli-A) (a), servindo de molde para síntese, pela transcriptase reversa, de uma fita de cDNA complementar, utilizando primers apropriados (b). Em seguida temos a degradação das fitas de mRNAs inicial e síntese de uma segunda fita de DNA, que em conjunto com a primeira fita, origina o DNA fita dupla referente a parte codificadora (c).

Neste ponto, você pode estar se perguntando, porque não existe nenhuma técnica que permite sequenciar o RNA? A principal razão se deve a sua instabilidade fora da célula.

É necessário para perfeita execução destas técnicas, que não haja contaminação do material a ser sequenciado com DNA genômico e que os mRNAs estejam em boa qualidade. Estas metodologias permitem estudar todo o transcriptoma de uma determinada espécie sem precisar sequenciar completamente todos os genes que estão sendo expressos. Assim, estas abordagens têm como vantagem rapidez na obtenção dos dados e redução de custos.

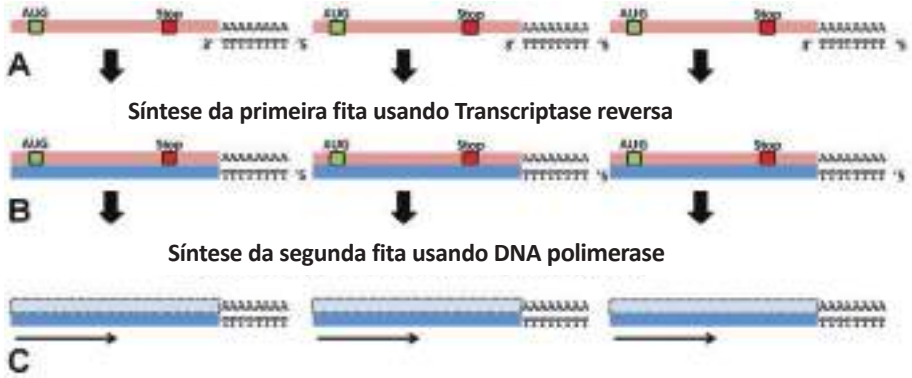


Figura 9. Técnica que permite obter a informação genética contida nos RNAs. Em (A) temos três RNA processados (em vermelho), contendo os códons de inicio (AUG em verde) e final da tradução (Stop em vermelho), seguido da calda de poliadenilação (poli-A). Em (B), os RNAs servem de molde para síntese da fita de cDNA complementares (mediante uso de transcriptase reversa), sintetizados a partir de um oligonucleotídeo iniciador poli-T, que por consequência ancora na região poli-A. Em (C) ocorre a degradação das fitas de RNAs molde e síntese de uma nova fita de DNA complementar à fita recém sintetizada.

Técnicas que utilizam estas abordagens são de grande importância por permitir: descobrir genes novos, identificar polimorfismos (SNPs), descobrir mutações, construir mapas genômicos e estudar a expressão gênica em condições distintas, sendo esta última aplicação discutida na próxima seção.

Os diversos projetos de sequenciamento de transcriptomas têm evidenciado, com altíssima frequência, um mecanismo que ocorre no interior das células, conhecido como splicing. Este mecanismo é responsável pelo processamento (retirada dos íntrons e "escolha" de éxons) dos transcritos primários para obtenção do mRNA maduro. Consequentemente, é possível obter diferentes tipos mRNAs maduros a partir de um mesmo pre-mRNA e estes mRNA maduros podem ter diferentes funções em virtude dos éxons "escolhidos".

Discutiremos agora as duas principais metodologias destinadas a este propósito.

Sequenciamento de ESTs

O próximo passo após a retrotranscrição consiste em clonar os cDNA em vetores apropriados e sequenciar suas extremidades, obtendo fragmentos que geralmente variam de 200 a 500 nucleotídeos.

Produção de bibliotecas de cDNA

Sequencias curtas correspondendo a parte dos cDNAs são conhecidas como EST (*Expressed Sequence Tags*) e em português significa "Etiquetas de Sequencias Expressas". A terminologia "etiqueta" é uma analogia as etiquetas encontradas nos produtos comercializados no cotidiano, as quais, por si só, permitem inferir sobre as características dos produtos.

Dependendo do objetivo do pesquisador, pode ser sequenciada apenas uma das extremidades do inserto (cDNA). Assim, podemos ter as 5' EST que geralmente, correspondem a região codificadora da proteína. Esta região tende a ser conservada entre espécies próximas evolutivamente, facilitando a identificação do gene por homologia. O sequenciamento da região final do inserto produz as 3' EST, que comumente correspondem a região não codificadora 3' UTR (untranslated regions) dos mRNAs. Geralmente esta região não é conservada entre espécies (Figura 10).

Apesar do pequeno tamanho, as ESTs permitem (na maioria das vezes) identificar os genes que as originaram (e consequentemente suas funções) utilizando programas que efetuam busca por identidade/similaridade, tal como o programa denominado BLAST (Basic Local Alignment Tool).

O conjunto de ESTs de um mesmo transcrito pode se sobrepor em regiões com alta identidade, gerando uma sequência maior representativa do cDNA que as originaram.

Esta abordagem é muito interessante para estudar diversos fenômenos biológicos através da comparação de bibliotecas de ESTs de duas condições distintas. Este tipo de abordagem permite inferir sobre adaptações biológicas que se correlacionam com diferenças na expressão dos genes. Por exemplo, se um dado transcrito de cDNA aparece múltiplas vezes numa biblioteca de ESTs é porque o mesmo se acumulou naquela determinada situação, sendo provavelmente importante para o organismo naquele momento celular. O contrário pode ser pensado para transcritos pouco expressos, ou seja, estes devem ser menos importantes para a mesma situação celular.

Como aplicação, tem-se, por exemplo, a busca por marcadores de condições celulares ou biológicas específicas, estudo da diferença de expressão gênica entre um tecido normal e um tecido tumoral em busca de marcadores de tumorigênese que possam ser utilizados no diagnóstico de um determinado tumor. Bibliotecas de ESTs podem ser utilizadas ainda para comparar tecidos tumorais com graus diferenciados de um determinado tumor e procurar assim por marcadores de prognóstico de câncer.

Orestes (Open Reading Frame ESTs)

Esta técnica foi desenvolvida no Brasil (Dias Neto et al., 2000) e tem como objetivo o sequenciamento de regiões internas dos genes, onde se concentra a informação referente a região codificadora das proteínas. Surgiu devido a limitação do tamanho das ESTs e do fato destas conterem, em sua maioria, sequencia relativa as extremidades não codificadoras dos RNAm, regiões estas que não trazem informações que possam ser relacionadas as possíveis funções dos transcritos.

A eficiência em conseguir a função do gene por esta técnica é devido a maior identidade em regiões internas de genes homólogos. Nesta técnica após a produção dos cDNAs, usa-se primers degenerados aleatórios (misturas de variados oligonucleotídeos) num passo de amplificação por PCR antes da clonagem e sequenciamento dos mesmos (Dias Neto et al., 1997; Dias Neto et al., 2000; Fietto et al., 2002). A utilização destes primers é necessária no caso de RNAm de bactérias, pois estes organismos não possuem poliA na porção 3′.

Uma comparação das bibliotecas de ESTs construídas com oligosdT e com oligos randômicos (ORESTES) pode ser vista na Figura 10.

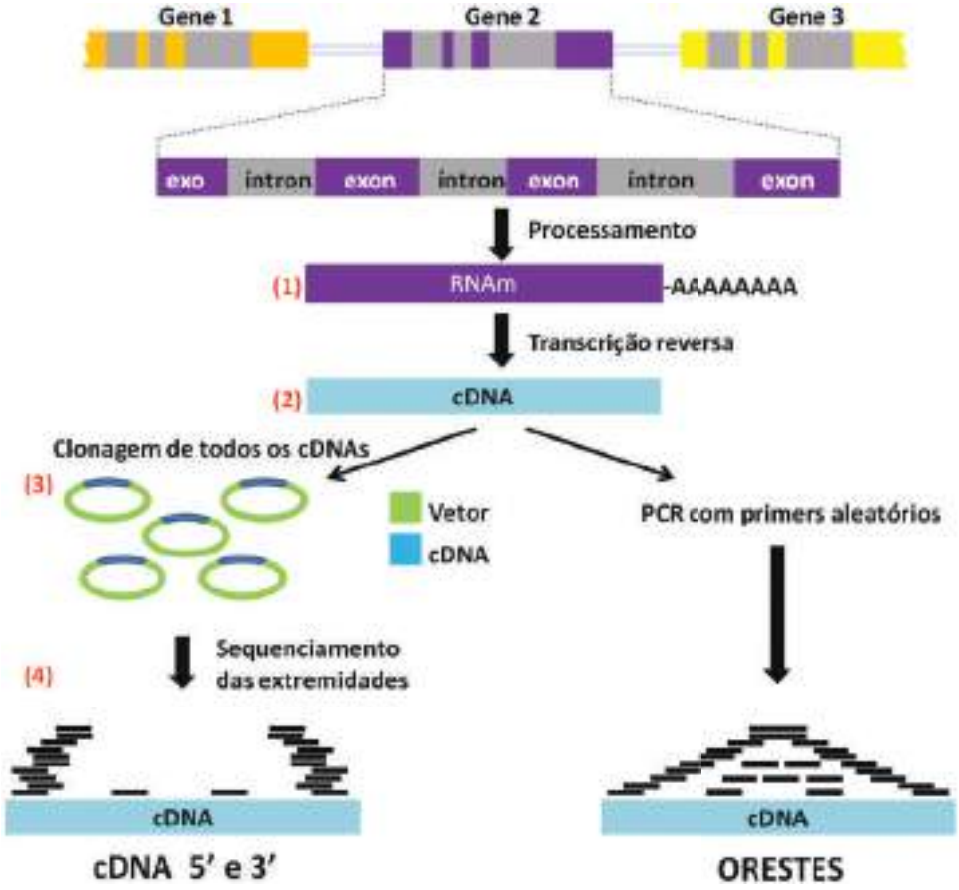


Figura 10. Representação esquemática das técnicas de EST e ORESTES. Neste esquema usa-se um gene hipotético eucariótico para comparar o sequenciamento de bibliotecas convencionais de cDNA (à esquerda) em relação ao uso da técnica ORESTES (à direita). Em ambas as técnicas o cDNA é produzido como produto final. A técnica de EST faz uso de um primer oligodT que ancora na região de poli-A, presente nos transcritos eucarióticos, especificamente na extremidade 3′. Isto permite que se construa um cDNA fazendo uso de sua transcrição reversa. O próximo passo é a clonagem seguida de sequenciamento. Já na técnica ORESTES antes da clonagem é feito um passo de PCR com oligonucleotídeos randômicos. Os produtos do PCR são então clonados e sequenciados. Esta pequena diferença gera sequencias, predominantemente, no centro dos insertos clonados no método ORESTES.

Esta metodologia tem como vantagem a normalização da população de genes expressos, permitindo que genes raros ou poucos expressos sejam amostrados.

Em muitas das vezes, utiliza-se sequencias oriundas desta abordagem juntamente com ESTs, com a finalidade de obter contigs maiores que permitem uma melhor identificação (anotação) do transcrito.

Sequenciamento de nova geração (Next Generation Sequencing – NGS)

Após a publicação do draft do genoma humano (Venter et al., 2001), houve um avanço nas tecnologias de sequenciamento culminando no surgimento dos "sequenciadores de segunda geração". Atualmente existem diversas tecnologias voltadas para o sequenciamento do DNA em larga escala, sendo a Roche a primeira empresa a desenvolver esta estratégia que se baseia na tecnologia de pirosequenciamento. A partir deste período, outros métodos foram desenvolvidos, sendo os mais importantes: o método Polony (Shendure et al., 2005) utilizado no sequenciador SOLID (Applied Biosystems) e o método de amplificação em ponte (BENNETT et al., 2005) utilizado no sequenciador Genome Analyser (Illumina). Com isso, a ênfase passou do sequenciamento de pequenos fragmentos de DNA ao estudo de genomas inteiros.

Plataforma 454

O princípio desta tecnologia foi proposto por Hyman em 1988 (1988), mas somente em 2005 tivemos o primeiro sequenciador de segunda geração disponibilizado no mercado pela empresa Roche (sequenciador 454 GS20) (Figura 11).

Esta tecnologia dispensa a clonagem e tem baixo custo (comparado a outros métodos existentes), e o sistema de sequenciamento é cerca de 100 vezes mais rápido quando comparado ao método de sequenciamento padrão de Sanger. A eficiência e rapidez da técnica foram comprovadas pelo ressequenciamento do genoma da bactéria *Mycoplasma genitalium* (508.069 bases) com 96% de cobertura e 99.96% de precisão num único processamento de quatro horas (Margulies et al., 2005).

Este método pode ser dividido em três etapas: a) preparo da amostra, b) PCR em emulsão e c) sequenciamento. Na primeira etapa, o DNA é fragmentado aleatoriamente por nebulização, sendo selecionados os fragmentos com tamanho adequado. Em

Figura 11. Modelo de Sequenciador (GS FLX®) que faz uso da plataforma 454.

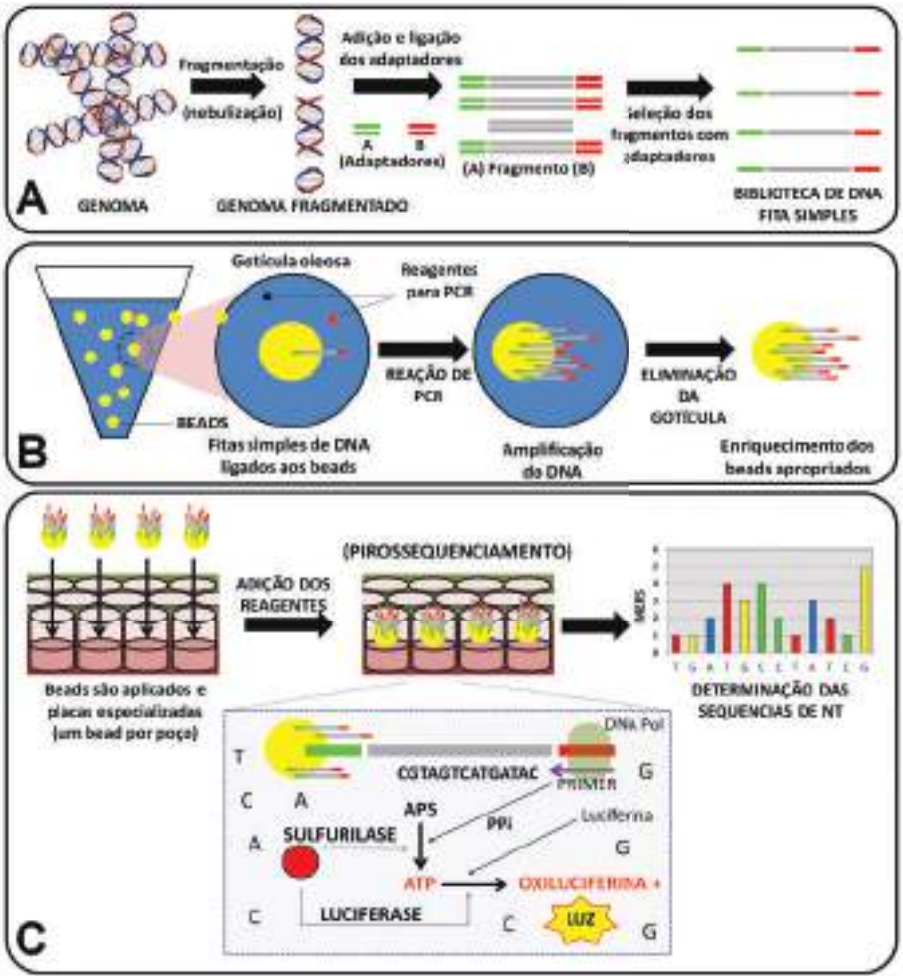


Figura 12. Esquema do processo de sequenciamento usando uma plataforma 454. O sequenciamento é dividido em três etapas: (A) preparo da amostra, (B) PCR em emulsão e (C) sequenciamento. Em (A) o DNA é fragmentado aleatoriamente e ligado a adaptadores A (verde) e B (vermelho) em suas extremidades, permitindo que fragmentos ligados a estes adaptadores possam ser separados, caracterizando assim a biblioteca. Em (B) os fragmentos da biblioteca são ligados às microesferas magnéticas por meio do pareamento do adaptador B, por intermédio de sequências curtas complementares a este adaptador presentes na superfície da microesfera. Apenas um único tipo de fragmento se liga a uma determinada microesfera. As microesferas são então capturadas individualmente em gotículas oleosas, onde a PCR em emulsão deverá ocorrer. Milhares de cópias do fragmento alvo são então produzidas nessa fase, no interior destas gotículas. Então (C) as microesferas ligadas às sequências alvo de fita simples são capturadas individualmente em poços no suporte de sequenciamento. Em seguida, são fornecidos os reagentes para a reação de pirossequenciamento. Cada nucleotídeo incorporado, em cada um dos poços da placa de sequenciamento, liberará um pirofosfato que será convertido em luz e, consequentemente, registrado na forma de pirograma. Posteriormente, estes pirogramas são decodificados por softwares especializados, resultando numa sequencia de nucleotídeos única representando cada *bead* (microesfera) na placa.

seguida, liga-se dois adaptadores (A e B) às extremidades dos fragmentos selecionados. Em "b", os fragmentos obtidos são ligados à microesferas magnéticas por meio do pareamento do adaptador B com sequências curtas complementares presentes na superfície da microesfera, onde ocorrerá a amplificação deste fragmento. O adaptador A, por sua vez, servirá de molde para anelamento do primer responsável pelo início da amplificação. Desta forma, estas microesferas agem como reatores de amplificação individual produzindo milhares de cópias de um único molde. Na última etapa, estas microesferas são adicionadas a uma placa de modo que cada orifício da placa receba uma única microesfera. Posteriormente são adicionados os reagentes necessários para amplificação do DNA. Pirofosfato inorgânico (PPi) é liberado a cada incorporação de um nucleotídeo complementar a fita molde. Este PPi livre é convertido, pela enzima ATP sulfurilase, em ATP, que por sua vez fornece energia para oxidar a luciferina à oxiluciferina. Como consequencia desta reação, temos a emissão de luz. Desta forma, cada incorporação de nucleotídeo à cadeia de DNA nascente emitirá luz, imediatamente convertida em pirogramas (Morozova e Marra, 2008). A interpretação destes "pirogramas" permite identificar a sequências de nucleotídeos do DNA molde (Figura 12).

Resumidamente, podemos dizer que esta técnica baseia-se na detecção de fótons de luz produzidos em quantidade proporcional ao número de nucleotídeos incorporados a cadeia de DNA nascente (Ronaghi, 2001; Mardis, 2008). No entanto, a quantidade de luz emitida após a incorporação do quarto homopolímero não é mais linear ao número de nucleotídeos incorporados, sendo este um problema desta técnica. Outra limitação consiste na geração de quimeras que são sequencias de DNA provenientes da ligação de dois fragmentos distintos.

Plataforma Illumina®

O desenvolvimento da plataforma Illumina® ocorreu devido ao trabalho conjunto de quatro companhias: Solexa, Lynx Therapeutics, Manteia Predictive Medicine e Illumina, baseando-se na metodologia proposta por Turcatti (2008) e colaboradores (Shendure e Ji, 2008). Esta parceria resultou no desenvolvimento do sequenciador Illumina Genome Analyser (Figura 13).

O princípio desta metodologia é similar ao método proposto por Sanger, pois temos em ambas a síntese de uma fita complementar ao DNA alvo utilizando DNA polimerase e nucleotídeos terminadores marcados com diferentes fluoróforos. A fluorescência emitida após a incorporação de cada nucleotídeo é registrada como imagem e no final, através de uma decodificação destas imagens, tem se a sequencia de interesse.

Como toda técnica de sequenciamento de segunda geração, é preciso primeiro preparar as bibliotecas contendo o DNA a ser sequenciado. Assim, o DNA é clivado, os fragmentos de tamanho apropriado são selecionados e ligados à adaptadores em ambas as extremidades.

Estas bibliotecas podem ser de dois tipos: paired-end, que proporciona o sequenciamento nas duas extremidades de reads com tamanho entre 200 e 500 nucleotídeos; e mate-pair que também sequencia as duas extremidades de reads maiores (de 2000 a 5000 nucleotídeos). Neste caso, é necessário um passo extra,

Figura 13. Modelo de um sequenciador que faz uso da técnica Illumina, Illumina Genome Analyzer.

pois a eficiência de ligação destes reads grandes na *flowcell* (placa de vidro) é baixa. O truque é ligar na placa de vidro apenas as extremidades destes fragmentos maiores. O kit desenvolvido pela Illumna® contorna, de maneira simples, este problema. Esta segunda biblioteca direciona a montagem de regiões repetitivas além de proporcionar a ligação e ordenação de contigs.

Adaptadores são ligados às extremidades dos fragmentos de tamanho apropriado (obtidos pela fragmentação randômica do DNA a ser sequenciado) durante a preparação da biblioteca. Estes fragmentos são fixados na placa distantes o suficiente para que, após a amplificação, exista somente um tipo de fragmento dentro do cluster, que geralmente é formado por mais de um milhão de cópias do mesmo fragmento. Os adaptadores têm a função de imobilizar os fragmentos fita simples, pela hibridização a primers complementares, numa placa de vidro onde acontecerá todo o processo. Após a fixação, o adaptador da extremidade livre liga-se ao primer complementar adjacente na placa de vidro. A alta densidade de primers ligados a flowcell facilita esta ligação que tem a função de iniciar a extensão da fita de DNA nascente, pela DNA polimerase, ao encontrar um OH livre na extremidade do primer livre. Na primeira etapa são fornecidos apenas nucleotídeos não marcados que proporcionará a extensão de uma fita de DNA complementar ao DNA molde fixado na placa. A extremidade desta fita recém-sintetizada se anela ao primer na extremidade da fita molde formando uma estrutura em ponte que dá nome ao processo de amplificação (amplificação em ponte). Posteriormente ocorre uma elevação de temperatura no suporte sólido desnaturando e linearizando as duas fitas que encontram outros dois primers reiniciando o processo (segundo ciclo). A clonagem *in vitro* é finalizada após 35 ciclos, resultado em milhares de clusters (cada um representando um fragmento a ser sequenciado) (Fedurco et al., 2006; Turcatti et al., 2008). Na segunda etapa, após os 35 ciclos de sequenciamento, são adicionados, ao longo de toda a extensão da placa, uma solução contendo os quatro

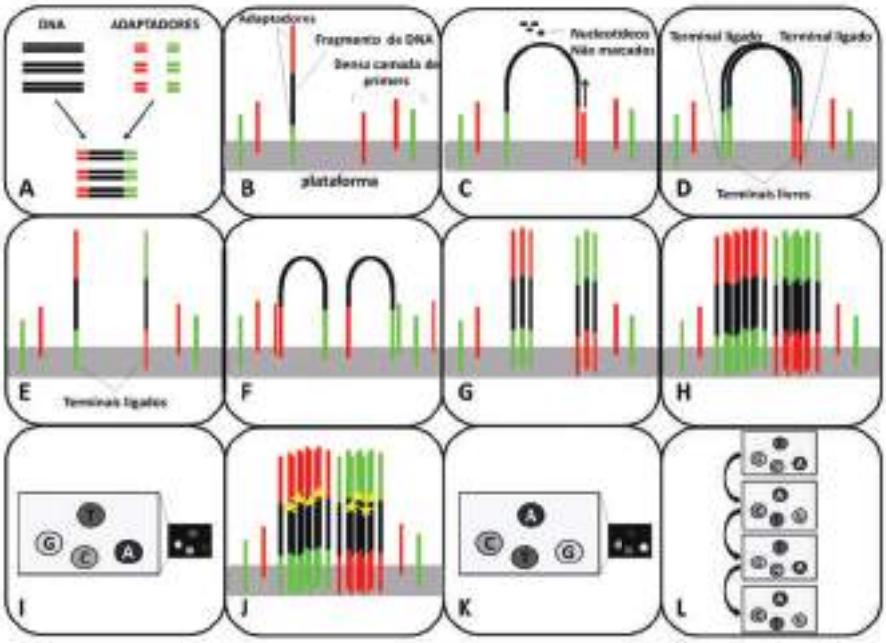


Figura 14. Princípio da técnica Illumina. Em (A) tem-se a fragmentação do DNA a ser sequenciado (através de um processo de nebulização), com posterior seleção dos fragmentos de tamanho apropriado e ligação de adaptadores em ambas as extremidades. Em (B), estes fragmentos são colocados em uma placa de vidro (*flowcell*) densamente povoada por adaptadores complementares aos adaptadores contidos nas extremidades dos fragmentos, de maneira que os fragmentos possam então se ligarem à placa. Em (C), ocorre a incorporação de nucleotídeos não marcados com fluorescência até que toda a extensão do fragmento seja amplificada. Em (D) tem-se a formação da estrutura em ponte, que dá nome ao processo de amplificação (amplificação em ponte) evidenciando dois adaptadores presos a placa e outros dois livres. Em (E) ocorre a desnaturação do duplex. Em (F) os adaptadores livres se ligam a adaptadores complementares na placa, iniciando um novo ciclo. Em (G), temos o cluster sendo formado, o qual provavelmente conterá mais de um milhão de cópias do mesmo fragmento. Em (H) adiciona-se os quatro tipos de didesoxinucleotídeos terminadores reversíveis contendo fluoróforos, junto com a enzima DNA polimerase, que fará a incorporação do didesoxinucleotídeo apropriado. A incidência de um feixe de raios laser excita os fluoróforos proporcionando emissão de luz que difere em função da base incorporada. Em seguida, efetua-se uma etapa de lavagem para remoção do grupo bloqueador presente na extremidade 3' junto com o fluoróforo; fato este que permitirá a incorporação do segundo nucleotídeo. Estes ciclos se repetem até que toda a extensão do DNA seja polimerizada. Em "I" ocorre o registro da imagem correspondendo a incorporação do primeiro didesoxinucleotídeo. "J" e "K" representam sucessivos ciclos de incorporação de didesoxinucleotídeos marcados, incidência de raios laser, emissão de luz e registro da imagem. Por fim, em "L", as imagens registradas em cada ciclo são decodificação para determinar a sequencia de bases de cada cluster na placa.

tipos de nucleotídeos terminadores marcados com fluorescência. Quando necessário estes nucleotídeos, contendo os fluoróforos, são adicionados as respectivas cadeias de DNA nascente em cada um dos clusters. Posteriormente ocorre a excitação do fluoróforo por feixes de raios laser fazendo com que os nucleotídeos emitam uma fluorescência que tem sua intensidade proporcional ao número de representantes (fragmentos) dos clusters. O número elevado de fragmentos no clusters (acima de um milhão) é necessário para produzir intensidade suficiente que permita detectar com exatidão uma determinada base no sequenciamento. Uma imagem, contendo a cor da fluorescência, é capturada para cada posição dos clusters na placa de vidro. Para termos uma idéia da quantidade de imagens geradas, cada corrida pode conter até 400 milhões de clusters. Em seguida, a extremidade 3' é desbloqueada com consequente remoção dos reagentes em excesso e do fluoróforo do nucleotídeo incorporado no ciclo anterior, permitindo o inicio de um novo ciclo. Este processo se repete até que todas as bases de um determinado fragmento sejam determinadas (Figura 14).

Assim, temos um sistema de amplificação totalmente diferente das outras duas metodologias de sequenciamento de segunda geração (454 discutida na seção anterior e SOLiD que será discutido na próxima seção) que são baseadas em PCR em emulsão.

A *flowcell* pode ser dividida em regiões (chamadas de canais ou linhas). No entanto, dependendo do objetivo do pesquisador, é possível sequenciar várias amostras por região (multiplex). Para isto, é necessário adicionar um pequeno adaptador que difere para cada uma das amostras a serem sequenciadas (Figura 15). O número de amostras a serem sequenciadas, em uma única corrida, está diretamente relacionado com a cobertura desejada.

Devido ao pequeno tamanho dos reads, esta plataforma era mais utilizada inicialmente para análise de expressão gênica diferencial, sequenciamento de pequenos RNAs e estudo envolvendo interação proteína-DNA (Chip-seq). O aperfeiçoamento desta tecnologia tem possibilitado sua utilização para "sequenciamento de novo" de genomas e ressequenciamento de genomas.

Figura 15. Sequenciamento através da técnica de multiplex. A adição de adaptadores distintos à extremidade dos fragmentos a serem sequenciados possibilita sequenciar várias amostras por linha.

Plataforma SOLiD®

A metodologia do sequenciador SOLiD® teve inicio no trabalho de Mckernan et al. (2006), sendo implementada pela empresa Applied Biosystems. O primeiro sequenciador foi liberado comercialmente no final de 2007, mas somente em 2008 e 2009 começaram a ser amplamente utilizados (Figura 16).

Em 2011, começou a ser comercializada a versão 5500 xl, que tem enfrentado alguns problemas que serão discutidos posteriormente. Este sequenciador pode ser utilizado para sequenciamento de genoma, ressequenciamento de regiões de interesse, experimentos envolvendo imunoprecipitação de cromatina, análise de expressão gênica e análise de pequenos RNAs, sendo as duas últimas aplicações muito utilizadas em virtude do tamanho dos fragmentos sequenciados.

Primeiro é preciso preparar a biblioteca a ser sequenciada, que pode ser de *tags* únicas ou de *tags* duplas (mate-pair). Na primeira os fragmentos são ligados diretamente aos adaptadores universais P1e P2. O segundo tipo de bibliotecas (mate-pair) gera fragmentos maiores, que variam de 1 a 10 kb e também tem os adaptadores P1 e P2 ligados nas extremidades. A diferença é que a segunda possui um adaptador interno que tem a função de unir os dois fragmentos a serem sequenciados (Figura 17). As bibliotecas mate-pair são ideais para sequenciamento de regiões mais longínquas do genoma.

Os fragmentos contendo os adaptadores são ligados, por meio do pareamento do adaptador P1, à uma sequencia curta complementar presentes na superfície da *bead*. Apenas um tipo de fragmento se liga a uma determinada *bead* que posteriormente são "capturadas" individualmente em gotículas onde a PCR em emulsão ocorre. Milhares de cópias do mesmo fragmento são produzidas nessa fase.

Os passos anteriores são muito parecidos com as etapas de preparo da amostra da PCR em emulsão, apresentadas para o sequenciador da Roche. A diferença neste método consiste em aplicar as *beads* geradas diretamente numa placa de vidro, processo este que se assemelha à etapa correspondente do Illumina®. Este é um dos motivos do SOLiD e illumina produzirem número parecidos de leitura que na maioria das vezes é muito superior ao número de leituras geradas pelo 454. Outro fator que justifica esta discrepância no número de reads gerados, comparado ao 454, se deve

Figura 16. Modelos de sequenciados que faze uso da plataforma SOLiD.

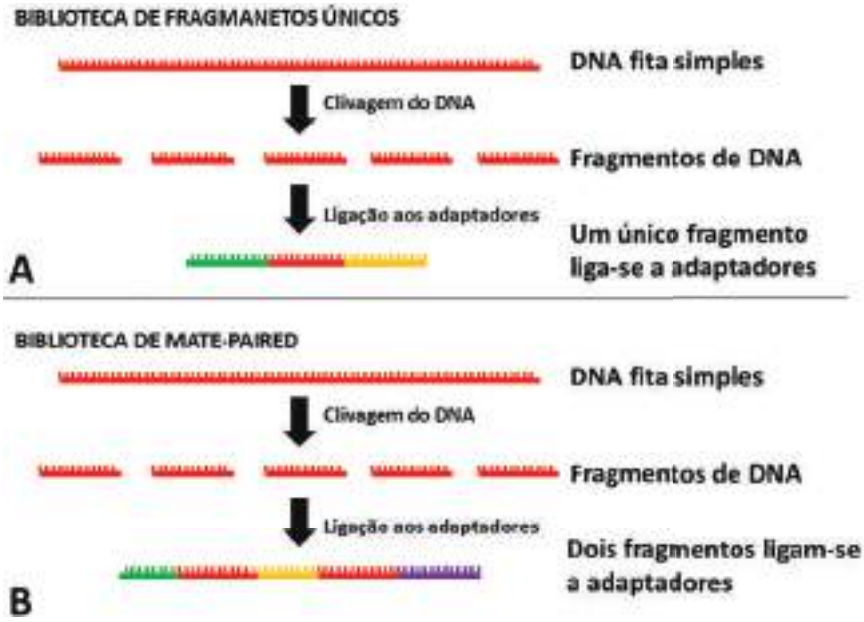


Figura 17. Tipos de bibliotecas sequenciáveis pela técnica de SOLiD: as de *tags* únicas (A) e as de *tags* duplas (mate-pair) (B). Em ambas, é preciso primeiro fragmentar o DNA a ser sequenciado e selecionar os fragmentos de tamanho apropriados. Em "A", os adaptadores universais P1e P2 são ligados diretamente na extremidade dos fragmentos. Em "B", os mesmos adaptadores são ligados nas extremidades. A diferença é que a segunda possui um adaptador interno que tem a função de unir dois fragmentos a serem sequenciados.

ao tamanho das *bead*s, que no SOLiD tem um micrômetro ao passo que no 454, as microesferas tem 20 micrômetros. Em outras palavras, podemos dizer que o número maior de reads é devido a uma melhor ocupação da lâmina, uma vez que as *beads* são menores e não precisam se alocar no interior de poços, proporcionando maior densidade de *bead*s na placa de vidro.

A plataforma SOLiD® utiliza um método de sequenciamento diferente das outras duas metodologias descritas anteriormente, sendo utilizada a enzima DNA ligase, ao contrário dos outros dois métodos citados anteriormente que baseiam-se na DNA polimerase.

O princípio desta técnica consiste na utilização de uma sonda com oito nucleotídeos, dos quais apenas os dois iniciais são informativos. Os três nucleotídeos seguintes podem ser qualquer dos quatro tipos de nucleotídeos (degenerados em todas as posições). E os três últimos são iguais (universais) capazes de parear com qualquer outro nucleotídeo (geralmente são inosina ou algum análogo da inosina) (Figura 18). No total, são utilizadas 1024 tipos de sondas durante a etapa de sequenciamento por esta plataforma.

Os primeiros kits de sequenciamento utilizavam sondas contendo nucleotídeos informativos na posição central.

O sequenciamento do DNA alvo ocorre em função da hibridização de sondas fluorescentes em cinco etapas distintas. Na primeira etapa, a ligação de um primer

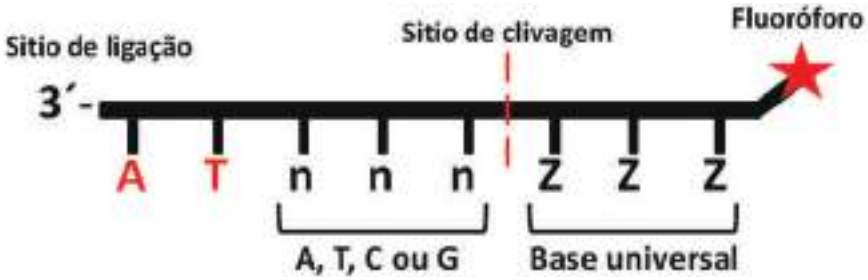


Figura 18. Representação esquemática de uma sonda do pool , evidenciando seus oito nucleotídeos, sendo os dois primeiros os nucleotídeos informativos; os três seguintes podem ser um dos quatro tipos de nucleotídeos e o três últimos são nucleotídeos universais (capazes de se parear com qualquer nucleotídeos). Esta figura evidencia ainda o sítio de clivagem na sonda, o fluoróforo e o sítio que é reconhecido pela DNA ligase.

de tamanho n (oligo n) ao adaptador P1 cria a condição necessária para ligação da primeira sonda ao complexo formato por: DNA alvo a ser sequenciado, adaptador P1 e a *bead*. O tamanho do primer é importantíssimo para o processo, pois determina a posição onde a primeira sonda deve se ligar ao DNA alvo. A DNA ligase é responsável pela ligação covalente do último nucleotídeo do primer com o primeiro nucleotídeo informativo da sonda. Em seguida, tem-se a excitação do fluoróforo da primeira sonda incorporada e detecção da fluorescência emitida, que é convertida, pela utilização de softwares específicos, em cor. Posteriormente ocorre a clivagem num ponto específico da sonda liberando as três bases universais e o fluoróforo. Esta clivagem finaliza o primeiro ciclo e permite a inserção de uma nova sonda por deixar um fosfato livre na extremidade 5'. No segundo ciclo, uma segunda sonda é ligada ao DNA molde com posterior excitação do fluoróforo e clivagem no ponto específico, liberando as bases universais e o fluoróforo. Este ciclo é repetido inúmeras vezes (lembrando que o *pool* de sondas contem dinucleotídeos informativos hápitos a hibridizar com qualquer sequencia encontrada no DNA alvo), até que o DNA molde seja todo coberto por sondas. Os passos descritos até este ponto constituem a primeira etapa do sequenciamento, e como resultado se tem todo o DNA molde coberto por dinucleotídeos informativos separados por três nucleotídeos não informativos. A segunda etapa é iniciada com a desnaturação da sequencia de DNA dupla fita originada e hibridização de um segundo primer de tamanho n - 1. Todas os passos referentes a primeira etapa são repetidos e no final temos um DNA molde todo coberto por dinucleotídeos informativos. O segundo primer é um nucleotídeo menor que o primeiro utilizado e este fato faz com que os dinucleotídeos informativos sejam incorporados uma posição antes dos dinucleotídeos informativos incorporados na primeira etapa. Assim, teremos na segunda etapa um dos nucleotídeos do dinucleotídeo informativo incorporado numa nova posição e o outro incorporado numa posição que já era conhecida da primeira etapa. Desta forma, cada nucleotídeo é sequenciado duas vezes, sendo esta uma das vantagens deste método. Este processo é repetido para as etapas três, quatro e cinco que utilizam, respectivamente, primers de tamanho (n - 2), (n - 3) e (n - 4). Após estas cinco etapas, temos como resultado um sistema de cores que representa a informação contida ao longo de toda extensão do DNA alvo (Figura 19).

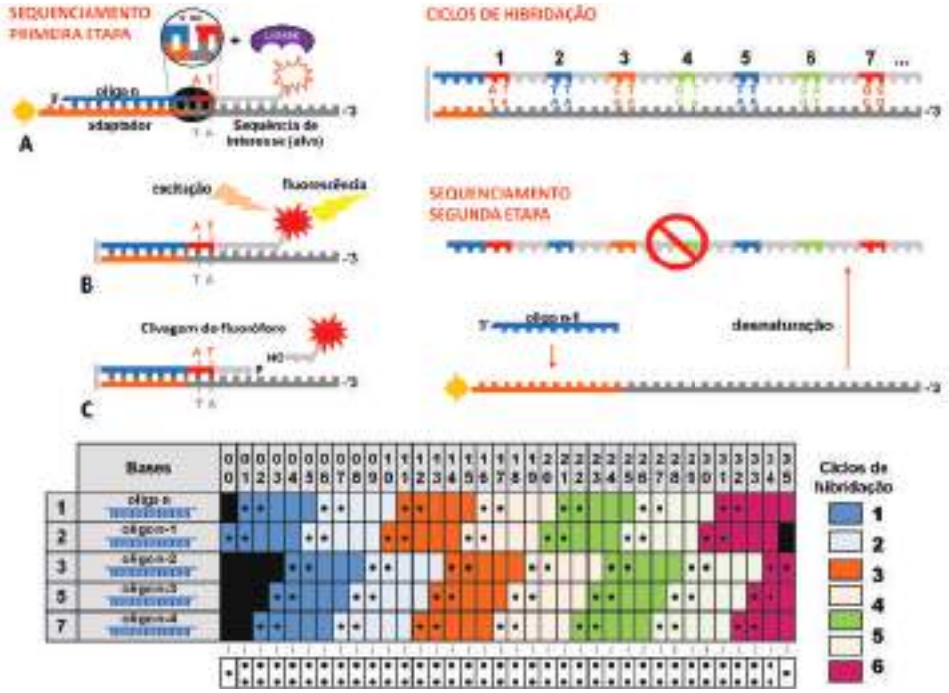


Figura 19. Princípio da técnica SOLiD. Cinco etapas distintas, baseando-se em hibridização de sondas fluorescentes, são necessárias para o sequenciamento do DNA molde. Na primeira (em "A"), temos a ligação da primeira sonda (contendo, neste exemplo, o dinucleotídeo informativo AT) ao complexo formato por: DNA alvo a ser sequenciado (cor cinza), adaptador P1 (cor laranja) e *bead* (microesfera amarela). O adaptador funciona como molde para hibridização de um primer de tamanho n (cor azul). Em seguida a DNA ligase faz a ligação covalente do último nucleotídeo do primer que contem um fosfato livre na extremidade 5' (adenina) com o primeiro nucleotídeo informativo da sonda que contêm um OH livre (timina), conforme pode ser visto na região ampliada. Em "B", temos a excitação do fluoróforo que emite fluorescência dependendo do dinucleotídeo incorporado. Esta fluorescência é registrada e convertida, pela utilização de softwares, em cores. Em "C", temos a clivagem num ponto específico da sonda liberando as três bases universais e o fluoróforo. Uma nova sonda é adicionada com posterior excitação do fluoróforo e clivagem no ponto específico, liberando as bases universais e o fluoróforo. D) este passo é repetido inúmeras vezes, até que toda a extensão do DNA molde seja coberto por dinucleotídeos. E) A sequencia de DNA dupla fita é desnaturada e inicia-se a segunda etapa de sequenciamento com um primer de tamanho n - 1. Este processo é repetido para as etapas três, quatro e cinco que utilizam, respectivamente, primers de tamanho (n - 2), (n - 3) e (n - 4). Após estas cinco etapas, temos como resultado um sistema de cores que representa a informação contida ao longo da extensão do DNA alvo.

O próximo passo consiste em decodificar as cores obtidas nas cinco etapas descritas anteriormente em sequencias de nucleotídeos. Para aprender como este processo é feito é preciso primeiro entender que a combinação dos quatro nucleotídeos resulta em 16 dinucleotídeos possíveis (Figura 20). No entanto, o método utiliza apenas quatro tipos de fluoróforos (quatro cores). Assim, teremos uma cor representando quatro dinucleotídeos. Como saber então qual dinucleotídeo uma determinada cor representa? A resposta vem do conhecimento do nucleotídeo pertencente a posição zero (último nucleotídeo do primer de tamanho n -1), que geralmente é

uma adenina. Sabendo que a primeira base é uma adenina e considerando que cada base é sequenciada duas vezes, temos que a segunda base deste nucleotídeo tem que ser igual a primeira base do dinucleotídeo seguinte. Seguindo este raciocínio é possível seguir um caminho e chegar na sequencia de nucleotídeos final do DNA molde. Desta forma é possível conhecer qual dinucleotídeo uma determinada cor representa (Figura 20).

Esta é a tecnologia de segunda geração que proporciona maior acurácia, devido ao fato de cada base ser sequenciada duas vezes. Como o erro é, na maioria das vezes, aleatório; a probabilidade de ocorrer dois erros na mesma posição, durante o sequenciamento, é quase zero. Por esta razão esta plataforma é a mais indicada para estudos de polimorfismos (SNPs), os quais são confundidos com erro de sequenciamento em outras plataformas.

O tamanho dos fragmentos gerados são de no máximo 75 pb podendo ser gerado até 1 bilhão de reads por corrida. A versão 5500 xl, lançada recentemente, alterou o princípio da PCR, passando a amplificar as moléculas na própria lâmina (como acontece na plataforma illumina) e não mais utilizando PCR em emulsão. A diferença principal para a plataforma illumina é a movimentação das moléculas ao longo da placa após a amplificação em decorrência de sua desnaturação. Como resultado, houve uma diminuição dos custos e tempo de preparo das amostras, sendo de 2 a 4 vezes mais produtivo.

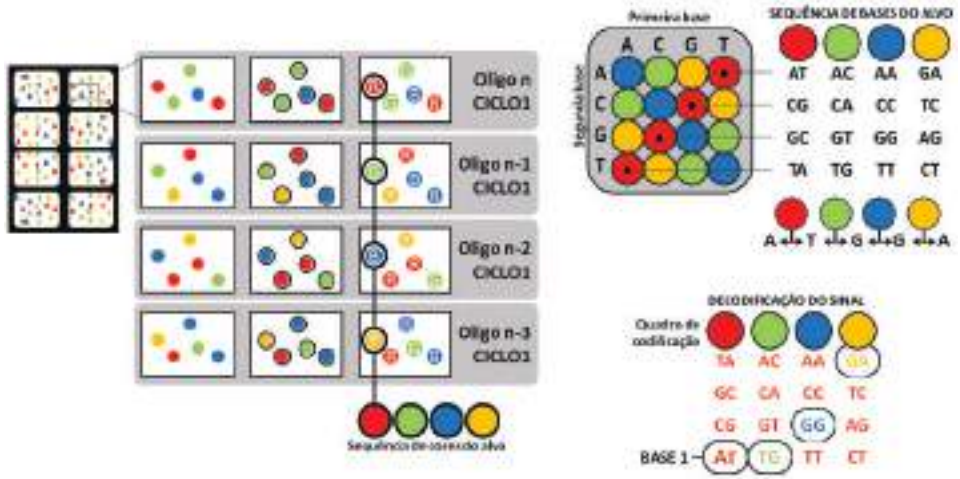


Figura 20. Decodificando cor em sequencia de nucleotídeos no método SOLiD. Combinando os quatro nucleotídeos temos como resultado 16 dinucleotídeos, que são representados por quatro cores distintas (vermelho, verde, azul e amarelo). Consequentemente, é preciso utilizar a mesma cor para representar quatro dinucleotídeos. A decodificação da cor em sequencia é efetuada pelo conhecimento da base referente a posição 0 do sequenciamento (última posição do primer de tamanho n - 1), que neste exemplo é uma adenina. Desta forma é possível conhecer o nucleotídeo referente a uma determinada cor. Neste exemplo, podemos dizer que o vermelho corresponde ao dinucleotídeo AT. Como cada base é sequenciada duas vezes, temos que a segunda base do dinucleotídeo descoberto (timina) tem que ser igual a primeira base do dinucleotídeo seguinte (timina). Seguindo este raciocínio, é possível seguir um "caminho" e descobrir a sequencia de nucleotídeos do DNA sequenciado, que neste exemplo é ATGGA.

Figura 21. Novos sequenciadores da Applied Biosystem.

Mesmo com todos os esforços, a tendência é que SOLiD seja cada vez menos utilizado com o lançamento, pela Applied Biosystem dos sequenciadores Ion PGM (Personal Genome Machine) e Ion Proton visto que estes novos sequenciadores possuem metodologias muito mais simples e baratas, diminuindo o custo de sequenciamento. (Figura 21).

Conforme visto nas seções anteriores, o sequenciador 454 mede a luz emitida como consequência da liberação do pirofosfato. Os dois sequenciadores mencionados anteriormente baseiam-se numa abordagem similar a plataforma 454®; a diferença é que neste caso, mede-se o hidrogênio liberado. Para termos uma idéia do poder deste tipo de sequenciador, a promessa de sequenciar o genoma humano em um dia por mil dólares tornou-se realidade.

Cronologia e evolução do sequenciamento

Até meados da década de 70 o DNA era a biomolécula mais difícil de ser analisada, sendo sua análise efetuada através de métodos indiretos, o que dificultava a identificação das sequencias que por consequência ocorriam após um período longo de tempo. Este problema foi resolvido com o surgimento, nesta mesma década, de duas metodologias independentes: uma proposta por Maxam e Gilbert e outra proposta por Sanger. Como resultado imediato, este avanço permitiu clonar e sequenciar genes específicos. Conforme visto nas seções anteriores, os diversos aprimoramentos do método de Sanger resultaram no desenvolvimento dos sequenciadores automatizados, suplantando a metodologia proposta por Maxam e Gilbert. O primeiro sequenciador automatizado utilizou a metodologia de Sanger et al. (1977) modificada por Edwards et al. (1990), liderando a pesquisa genômica por aproximadamente 30 anos, sendo responsável pelo sequenciamento de diversas espécies. O primeiro genoma sequenciamento foi de um vírus, o do fago ϕX174 em 1975. Em 1995, 20 anos depois, foi sequenciado o primeiro genoma de um organismo de vida livre, o da bactéria *Haemophilus influenzae* (Fleischmann et al., 1995), com 1,8 Mpb.

Os cinco anos seguintes foram marcados pela publicação do sequenciamento do genoma de mais de 50 outras espécies, destacando-se: as bactérias *Mycobacterium tuberculosis* (Cole et al., 1998) por causar a tuberculose e *Escherichia coli* que é um dos principais patógenos humanos (Blattner et al., 1997); e o eucarioto *Plasmodium falciparum* (Gardner, Hall, et al., 2002; Gardner, Shallom, et al., 2002; Hall et al., 2002; Hyman et al., 2002), causador da malária. Neste intervalo foram publicados ainda genomas mais complexos, como: da levedura, da mosca-das-frutas, de *Caenorhabditis elegans* e *Arabidopsis thaliana* (Brown, 2003). Finalmente, tivemos o sequenciamento de genomas de mamíferos, como o humano (Lander et al., 2001), do rato (Waterston et al., 2002) e do chimpanzé (Mikkelsen et al., 2005).

O crescente número de sequenciadores de segunda geração vendidos tem causado um crescimento no número de genomas sequenciados numa proporção inimaginável. Em Dezembro de 2015, foram registrados 36.052 projetos de sequenciamento finalizados (7.747 concluídos e 28.305 classificados como *draft* permanentes), 33.571 projetos em andamento e 1.856 projetos de sequenciamento de regiões de interesse do genoma (http://genomesonline.org). O número total de projetos (71.479) evidencia que estamos vivendo uma revolução genômica.

A enorme quantidade de dados genômicos, oriundos dos diversos projetos de sequenciamento, culminou no surgimento de inúmeros bancos de dados biológicos, destacando-se o NCBI (National Center for Biotechnology Information) (http://www.ncbi.nlm.nih.gov/), o EBI (European Bioinformatics Institute) (http://www.ebi.ac.uk/) e o DDBJ (DNA Data Bank of Japan) (http://www.ddbj.nig.ac.jp/).

O programa brasileiro de sequenciamento de genomas nasceu devido a iniciativa da Fundação de Amparo à Pesquisa do estado de São Paulo e resultou no sequenciamento do genoma da bactéria Xylella fastidiosa (primeiro fitopatógeno a ser completamente sequenciado no mundo), constituindo num dos principais feitos da ciência nacional. Os investimentos em infraestrutura e capacitação de inúmeros pesquisadores foram essenciais para o surgimento, em seguida, dos três outros importantes projetos: SUCEST (Sugar Cane EST), Xanthomonas campestris (segundo fitopatógeno) e do Câncer Humano (HCGP). Este último foi estimado em 12 milhões e teve a participação de entidades públicas (FAPESP) e privadas (Instituto Ludwig de Pesquisas sobre o Cancer - ILPC).

As tecnologias de sequenciamento de terceira geração, ainda em fase de desenvolvimento, prometem novamente revolucionar a ciência genômica por diminuir ainda mais o custo do sequenciamento. Ainda há controvérsias em sua definição, mas alguns autores consideram como de terceira geração as técnicas que não se baseiam em fluorescência ou luz. Outros consideram como de terceira geração as técnicas que prometem sequenciar fragmentos muito maiores do que os fragmentos sequenciados atualmente.

Sanger × tecnologias de sequenciamento de segunda geração

A técnica de sequenciamento proposta por Sanger no final da década de 70 foi aperfeiçoada e como resultado, o método passou de manual (caro e oneroso) à automatizado (rápido e com pouca intervenção humana). Uma das mudanças na

Tabela 2. Comparação do método de Sanger automatizado com as principais plataformas de sequenciamento de segunda geração.

	Sanger	454	SOLiD	Illumina
Tamanho dos reads	Longos	Médios - Longos	Curtos - Médios	Curtos - Médios
Bases por corrida	Muito pouco	Poucas	Muitas	Muitas
Custo	Muito alto	Médio	Muito Baixo	Muito Baixo
Vantagem	Reads longos	Reads longos	Muitos reads e estudo de SNPs	Muito reads
Desvantagem	Trabalhoso e muito caro	Erros em regiões homopoliméricas	Reads curtos*	Reads curtos *

* Este problema está sendo resolvido com o aperfeiçoamento dos novos sequenciadores.

técnica original foi a substituição dos compostos radioativos danosos a saúde humana por fluoróforos, sendo esta a base das plataformas Illumina e Applied Biosystems. De modo geral, as plataformas de sequenciamento de segunda geração têm suprido muitas das limitações do método de Sanger automatizado, tais como: custo, tempo e quantidade de dados produzidos em uma única corrida. O desenvolvimento dos sequenciadores de segunda geração causou uma revolução na genômica, uma vez que um único sequenciador fazia o trabalho de 30 mil outros baseados na metodologia de Sanger (Mardis, 2008; Cullum et al., 2011).

Como resultado da automação das metodologias de sequenciamento, houve uma significativa diminuição dos custos global, permitindo que laboratórios menores, espalhados pelo mundo, começassem a planejar e desenvolver seus próprios projetos genomas. Este fato, aliado ao número crescente de sequenciadores vendidos, tem resultado numa liberação exponencial do número de fragmentos sequenciados.

As principais diferenças entre os sequenciadores discutidos nesta seção são apresentadas na Tabela 2.

Sequenciamento do genoma humano

O sequenciamento do genoma humano foi efetuado, de forma independente, por um consórcio público e pela empresa privada Celera Genomics, tendo como objetivo a identificação, na ordem correta, dos 3 bilhões de nucleotídeos que compõe o genoma. O consórcio público foi iniciado em 1989 com a participação de aproximadamente 5000 cientistas distribuídos em centenas de laboratórios pelo mundo e tinha a previsão de término para 2005. Estima-se que o projeto público tenha utilizado 600 sequenciadores, contra 300 (todos situados no mesmo prédio) utilizados pela empresa privada, coordenada pelo pesquisador Craig Venter. Esta empresa tinha a intenção de patentear vários dos genes sequenciados e de cobrar pelo acesso às informações geradas, tendo sua entrada antecipado a finalização do projeto para 2003.

Após mais de uma década de trabalho, o médico-geneticista Francis Collins que na época era diretor do Instituto Nacional de Pesquisa do Genoma Humano

e Craig Venter, presidente da Celera Genomics (empresa que ele mesmo fundou) anunciaram em 2001, o primeiro rascunho do genoma humano. O consórcio público, mais cuidadoso e preocupado com as análises efetuadas, utilizou a metodologia de shotgun hierárquico, sequenciando um cromossomo por vez. Este fato tornava a montagem mais segura e confiável. O sequenciamento efetuado pela empresa Celera era mais ambicioso e arriscado, pois sequenciou o genoma completo de uma só vez, através da técnica "shotgun de genoma inteiro" (WGS). Esta empresa utilizou uma estratégia extremamente elegante que foi a clonagem de fragmentos de tamanhos diferentes (2 mil, 10 mil e 130 mil pares de bases) em vetores apropriados, produzindo três bibliotecas que tiveram as extremidades de seus insertos sequenciados.

Somente em abril de 2003 o projeto terminou, dois anos antes da data prevista oficialmente, mas apenas em 2007 foi publicada a primeira sequencia completa de um organismo diploide que era do próprio Craig Venter.

Foram obtidos como resultado do consórcio público, aproximadamente, 30 milhões de fragmentos medindo cerca de 800 pb que correspondem a cerca de 24 bilhões de bases (uma cobertura de oito vezes se considerarmos que o genoma humano tem 3 bilhões de nucleotídeos).

Este feito revolucionou o estudo de diversas áreas, principalmente a ciência biomédica, por permitir descobrir a causa de inúmeras doenças.

Montagem de genomas

O ideal seria sequenciar o genoma inteiro ou o maior tamanho de fragmento possível. No entanto, a maioria das técnicas de sequenciamento apresentadas anteriormente utiliza a enzima DNA polimerase para incorporação de nucleotídeos a cadeia de DNA nascente, sendo o tamanho dos fragmentos gerados, uma limitação destas técnicas, uma vez que a processividade desta enzima é limitada. Como conseqüência são gerados fragmentos pequenos, se comparados ao tamanho do genoma, medindo no máximo 800 pares de bases.

Neste ponto, surge um grande problema que é colocar estes milhares de fragmentos obtidos, na ordem correta após o sequenciamento, processo intitulado montagem do genoma. Como montar então genomas com milhões ou até mesmo bilhões de pares de bases? Este processo é efetuado devido ao desenvolvimento de vários programas de montagem (*assemblers*) que alinham os reads (fragmentos) gerados baseando-se em regiões de sobreposição entre eles para produzir sequências únicas denominadas contigs (Figura 22).

Um passo anterior à montagem consiste na retirada de regiões que não fazem parte do genoma do organismo sequenciado, tais como vetores e adaptadores. O ideal é que após a ordenação destes fragmentos, obtenhamos uma sequencia única para genomas circulares, ou várias sequencias contiguas, representando o número total de cromossomos da espécie.

Um problema que persiste, mesmo com o crescente investimento no refinamento dos programas de montagem, é a montagem de regiões repetidas do genoma, devido a dificuldade de ordenar corretamente estas regiões. Outro problema inerente aos sequenciadores atuais é a incorporação errônea de nucleotídeos, pela DNA polimerase, à cadeia de DNA nascente. Como resultado, temos dentre os milhares de fragmentos

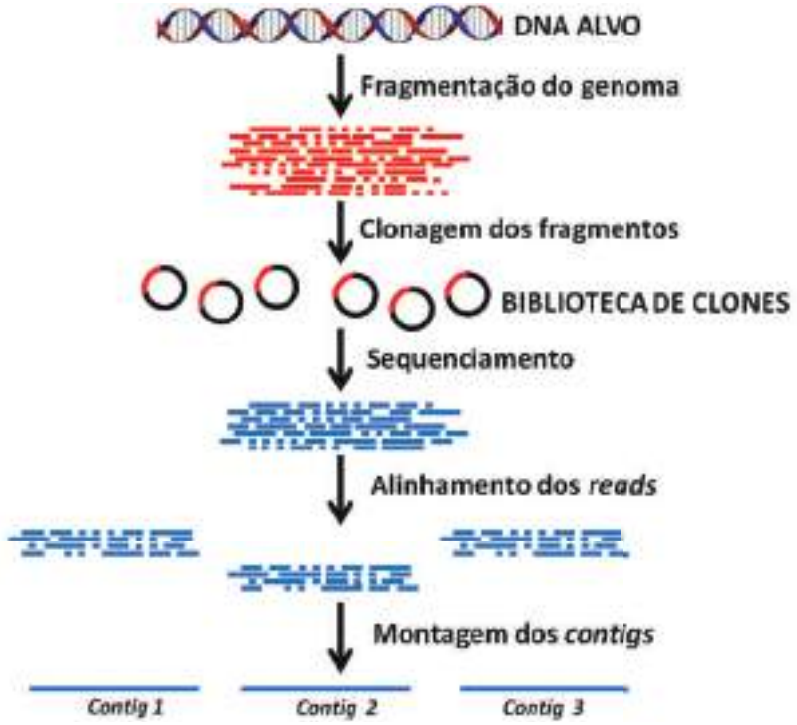


Figura 22. Geração de sequencias *contigs* através do alinhamento de *reads* contendo regiões de sobreposição durante a etapa de montagem de um genoma.

amplificados alguns contendo determinados nucleotídeos incorporados de forma equivocada pela polimerase. Para contornar o problema referente a incorporação errônea de nucleotídeos a cadeia de DNA, torna-se necessário que cada posição de base do cromossomo seja representada várias vezes, fazendo com que o número de bases sequenciadas seja dez vezes ou mais o tamanho do genoma original (Bouck et al., 1998).

Outra forma de resolver os problemas de montagem apresentados anteriormente foi associar um valor de qualidade (Base-Calling) a cada nucleotídeo sequenciado, fato este que permite eliminar ou mascarar sequencias (ou nucleotídeos) com baixo valor de qualidade.

Predição gênica

Após sequenciar e "montar" o genoma de um determinado organismo é preciso efetuar uma varredura neste em busca das sequencias de nucleotídeos correspondentes a cada um de seus genes (predição gênica) ou de outras regiões de interesse. Existem, para esta finalidade, diversos programas de bioinformática, cada um com suas peculiaridades metodológicas. O princípio básico desta metodologia consiste em fazer com que o programa reconheça nucleotídeos que são característicos de um determinado tipo de elemento gênico. Desta forma é possível identificar: regiões promotoras, junção dos

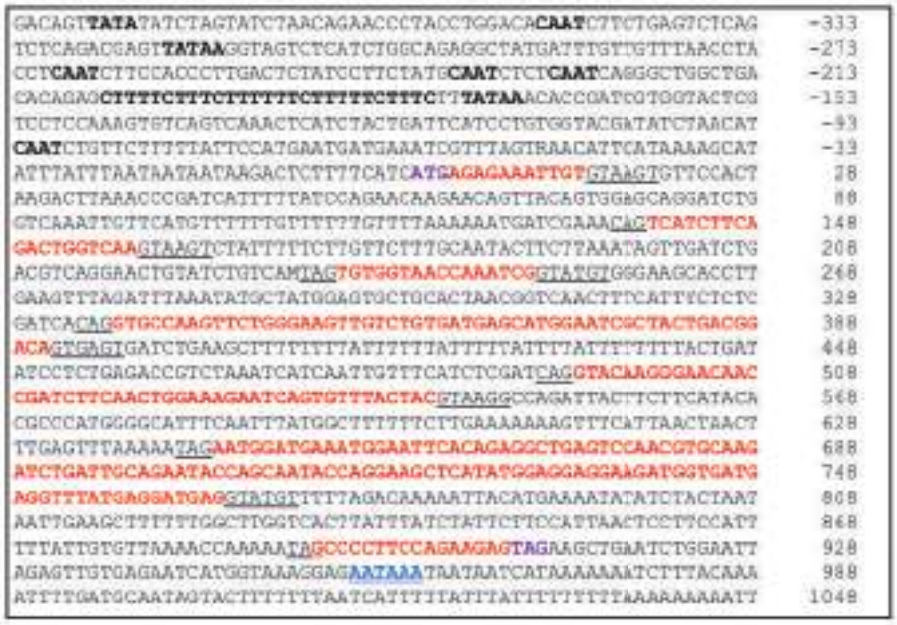


Figura 23. Sítios gênicos conservados. A sequencia de nucleotídeos é numerada começando pela primeira base do códon de iniciação ATG (+1) (em roxo), e os números na margem direita representam a posição do último nucleotídeo de cada linha em relação ao códon de iniciação. Os éxons estão indicados em vermelho (7 no total) e as regiões entre os éxons são os íntrons (6 no total). Sítios 5' e 3' conservados, flanqueando os exóns estão sublinhados em todos os íntrons. Domínios conservados envolvidos na transcrição de eucariotos (ex: TATAA, CAAT e motivos CT) estão em negrito na região 5' não codificante. Sublinhado azul a 3' do códon de terminação (TAG) mostra um potencial sítio de poliadenilação. Os códons de início e final da tradução estão em roxo.

éxons com os íntrons, os códons de início e parada da tradução e consequentemente onde começam as regiões 5' e 3' UTR (regiões não traduzidas "Untranslated Regions"). Desta forma, os programas são capazes de predizer a sequencia do gene, identificando o conjunto gênico de um organismo após seu sequenciamento. Os sítios conservados mencionados acima podem ser encontrados no gene hipotético representado na Figura 23.

Programas específicos podem ser utilizados para identificação de outras regiões do genoma, tais como: marcadores moleculares, regiões repetitivas, elementos transponíveis, dentre outras. Para termos uma noção, aproximadamente 98,5 % do DNA humano não codifica proteína.

Anotação gênica

O processo de anotação consiste em descobrir os diversos elementos presentes no genoma e atribuir, a eles, o máximo de informações biológicas possível (Stein, 2001). Na grande maioria das vezes estamos interessados em descobrir os genes e suas respectivas funções (ou seja, seu provável produto) a fim de entender os fenômenos biológicos que acontecem nos organismos.

Como visto na seção anterior, a predição de ORFs é uma das maneiras de obtermos o conjunto gênico de um determinado organismo após o sequenciamento de seu genoma. Outra maneira seria sequenciar o transcriptoma desta espécie. O princípio é o mesmo nestes dois casos, sendo a anotação feita através da comparação das ORFs ou dos transcritos com genes homólogos previamente anotados disponíveis em banco de dados públicos.

A anotação pode ser automática ou manual. A primeira é efetuada através de programas de bioinformática capazes de anotar o conjunto gênico de um organismo de uma só vez. A segunda é muito mais demorada, pois é efetuada, pelo anotador, para cada gene separadamente. Neste caso dizemos que houve uma curadoria, ou seja, o processo foi feito de maneira mais cuidadosa. É evidente que a segunda forma é a mais indicada por ser mais confiável. Na prática, ocorre os dois tipos de anotação, sendo a anotação automática mais comum devido a grande quantidade de dados biológicos gerados e devido ao tempo gasto para efetuar a anotação manual.

UniProtKB/Swiss-Prot (http://www.uniprot.org/uniprot/) (Magrane, 2011) é o principal banco de dados biológico destinado ao armazenamento de sequencias protéicas revisadas (anotadas manualmente).

Diversos programas de bioinformática destinados a este propósito surgiram em consequência da grande quantidade de dados genômicos gerados diariamente. Dentre estes, temos o Glimmer, o RBSfinder, o tRNAscan, o GeneMark e o BLAST2GO (Conesa et al., 2005).

O princípio básico destes programas baseia-se em comparar a sequencia de cada ORF ou transcrito, através de programas de alinhamento, com os diversos tipos de bancos de dados disponíveis, tais como: KEGG, GenBank, COG, Interpro, PSORT e outros. Dentre os principais programas de alinhamento destacam-se o ClustalW (Thompson et al., 1994).

Desta forma, é possível identificar em genomas recém sequenciados: os genes, as regiões promotoras, as seqüências repetitivas, os RNAs estáveis, os RNAs não codificantes (tRNA, rRNA, snRNA), além de inúmeros outros elementos.

Grande parte das sequencias depositadas nos bancos de dados públicos estão erroneamente anotadas e mesmo assim, estas sequencias servem de base para a anotação de ORFs e transcritos. A propagação do erro constitui no maior problema que a anotação enfrenta hoje.

Mapas genômicos

As informações genéticas contidas no genoma podem ser representadas por diversos tipos de mapas genômicos, dentre os quais destacamos: mapas genéticos, físicos, de restrição, de bandeamento cromossômico, de ligação genética, dentre inúmeros outros. Estes podem ser utilizados para representar as informações genéticas contidas em mitocôndrias, cloroplastos e plasmídeos.

O foco desta sessão são os mapas físicos que consistem na representação da ordem dos genes no cromossomo, levando em consideração a distância relativa entre eles. Esta representação permite conhecer a localização de um gene no cromossomo assim como suas regiões adjacentes, permitindo estudar as relações entre genes e

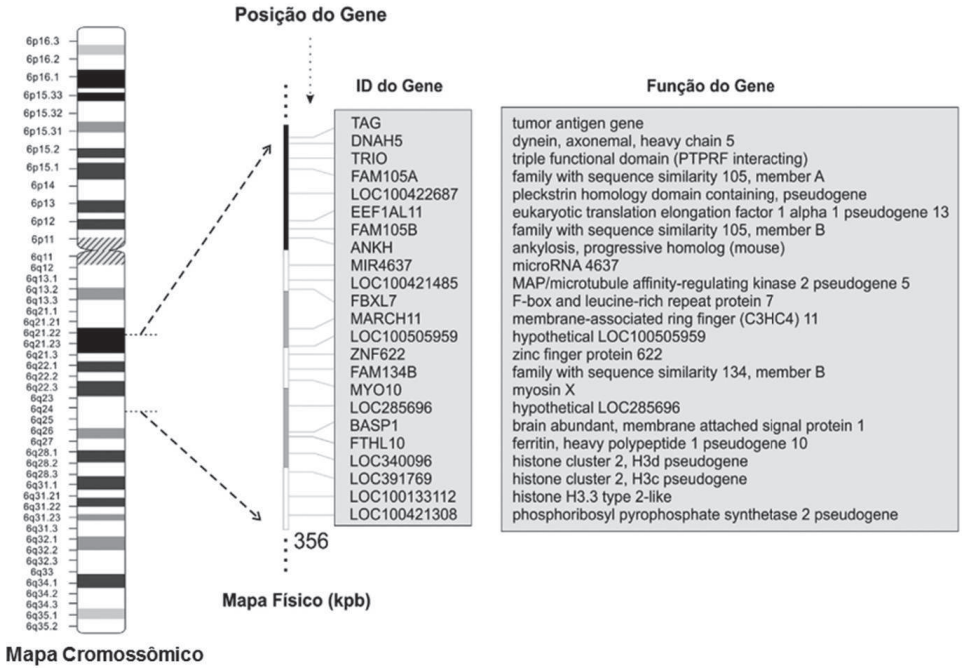


Figura 24. Mapa de bandeamento do cromossomo 4 humano e mapa físico evidenciando posição, identificação e função de alguns genes.

espécies. Este tipo de mapa nos permite ainda: identificar regiões do genoma com maior probabilidade de ocorrer recombinação, construção de linhagens genéticas e descoberta de funções de genes.

Os mapas físicos (Figura 24) utilizam o número de pares de bases (pb) como unidade para mensurar a distância entre os genes e podem trazer inúmeras informações, tais como: número de genes, suas respectivas posições, orientações, nomenclatura e função. A Figura 24 exemplifica um mapa de bandeamento cromossômico. A partir deste foi construído um mapa físico contendo 23 genes de um total de 1.444 existentes no cromossomo 4 humano. A resolução dos mapas físicos (quantidade de informação contida) varia em função do que se quer mostrar com a representação.

Genomas incompletos (draft): problemas e soluções

A maioria dos genomas sequenciados e montados até o momento é representada por *draft* (que ainda não foram finalizados), podendo este fato ser devido a uma ou mais das seguintes causas:

- baixa cobertura proporcionada pelos reads sequenciados,
- configuração inadequada dos softwares de montagem,
- alto percentual de reads com baixa qualidade,
- genomas com alto percentual de regiões repetitivas,
- genomas com alto percentual de elementos móveis (transposons e retrotransposons),

- presença de reads/contigs quiméricos,
- posicionamento errado dos *mate-pairs*,
- erros causados pelo sequenciamento,
- presença de vetores e adaptadores nos *reads*,
- contaminação do material genético gerando reads que não pertence ao organismo de interesse.
- dificuldade em sequenciar regiões complexas do genoma, tais como centrômeros, telômeros e compressões.

Em decorrência de um ou mais destes problemas, podemos ter uma montagem mau sucedida, resultando em genomas com gaps, mismatches, indels ou polimorfismo. Neste caso temos o numero de contigs maior do que o número de cromossomos da espécie.

Existe uma relação inversa entre o tamanho dos reads gerados e o grau de cobertura necessária para uma montagem confiável do genoma. Lander & Waterman (1988) mostraram, através de análises estatísticas, que uma cobertura de 8 a 10 vezes é suficiente para montagem de genomas menos complexos. Para reads muito pequenos, tais como os gerados pela plataforma SOLiD®, esta cobertura deve ser bem maior. Para termos uma idéia, a montagem do genoma humano utilizando esta abordagem necessitaria de uma cobertura de aproximadamente 30 vezes.

Genomas incompletos trazem prejuízos incalculáveis a comunidade científica, pois limitam muito as análises e mesmo as possíveis de serem efetuadas são prejudicadas devido a incerteza dos resultados encontrados. A dúvida geralmente consiste em saber se um gene (ou uma proteína ou uma determinada via), não foi encontrado numa determinada espécie por não existir nela ou porque seu genoma está incompleto?

Uma montagem bem sucedida requer inúmeros cuidados que devem começar no desenho racional do experimento (identificar a cobertura necessária), eliminar os contaminantes da biblioteca (caso existam), tratar adequadamente os reads antes da montagem para retirada de vetores, adaptadores e regiões de baixa complexidade e por fim, configurar corretamente o software utilizado em função dos dados gerados e do objetivo proposto.

Bibliografias

ALTSCHUL, S. F.; GISH, W.; MILLER, W. et al. Basic local alignment search tool. **J Mol Biol,** v. 215, p. 403-410, 1990.

BAUM, B. R. PHYLIP: Phylogeny Inference Package. Version 3.2. (Software review). **Quart. Rev. Biol,** v. 64, p. 539-541, 1989.

BLATTNER, F. R.; PLUNKETT, G., 3RD; BLOCH, C. A. et al. The complete genome sequence of Escherichia coli K-12. **Science,** v. 277, p. 1453-1462, 1997.

BOUCK, J.; MILLER, W.; GORRELL, J. H. et al. Analysis of the quality and utility of random shotgun sequencing at low redundancies. **Genome Research,** v. 8, p. 1074-1084, 1998.

BROWN, T. A. Clonagem Gênica e Análise de DNA: Uma Introdução. Artmed Editora S.A. Porto Alegre, RS. 375p., 2003.

COLE, S. T.; BROSCH, R.; PARKHILL, J. et al. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. **Nature,** v. 393, p. 537-544, 1998.

CONESA, A.; GOTZ, S.; GARCIA-GOMEZ, J. M. et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. **Bioinformatics,** v. 21, p. 3674-3676, 2005.

CULLUM, R.; ALDER, O.; HOODLESS, P. A. The next generation: using new sequencing technologies to analyse gene regulation. **Respirology,** v. 16, p. 210-222, 2011.

DIAS NETO, E.; CORREA, R. G.; VERJOVSKI-ALMEIDA, S. et al. Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. **Proc Natl Acad Sci U S A,** v. 97, p. 3491-3496, 2000.

DIAS NETO, E.; HARROP, R.; CORREA-OLIVEIRA, R. et al. Minilibraries constructed from cDNA generated by arbitrarily primed RT-PCR: an alternative to normalized libraries for the generation of ESTs from nanogram quantities of mRNA. **Gene,** v. 186, p. 135-142, 1997.

EDWARDS, A.; VOSS, H.; RICE, P. et al. Automated DNA sequencing of the human HPRT locus. **Genomics,** v. 6, p. 593-608, 1990.

FEDURCO, M.; ROMIEU, A.; WILLIAMS, S. et al. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. **Nucleic Acids Res,** v. 34, p. e22, 2006.

FIETTO, J. L.; DEMARCO, R.; VERJOVSKI-ALMEIDA, S. Use of degenerate primers and touchdown PCR for construction of cDNA libraries. **Biotechniques,** v. 32, p. 1404-1408, 1410-1401, 2002.

FLEISCHMANN, R. D.; ADAMS, M. D.; WHITE, O. et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. **Science,** v. 269, p. 496-512, 1995.

GARDNER, M. J.; HALL, N.; FUNG, E. et al. Genome sequence of the human malaria parasite Plasmodium falciparum. **Nature,** v. 419, p. 498-511, 2002.

GARDNER, M. J.; SHALLOM, S. J.; CARLTON, J. M. et al. Sequence of Plasmodium falciparum chromosomes 2, 10, 11 and 14. **Nature,** v. 419, p. 531-534, 2002.

HALL, N.; PAIN, A.; BERRIMAN, M. et al. Sequence of Plasmodium falciparum chromosomes 1, 3-9 and 13. **Nature,** v. 419, p. 527-531, 2002.

HIGGINS, D. G.; SHARP, P. M. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. **Gene,** v. 73, p. 237-244, 1988.

HYMAN, E. D. A new method of sequencing DNA. **Anal Biochem,** v. 174, p. 423-436, 1988.

HYMAN, R. W.; FUNG, E.; CONWAY, A. et al. Sequence of Plasmodium falciparum chromosome 12. **Nature,** v. 419, p. 534-537, 2002.

KUMAR, S.; NEI, M.; DUDLEY, J. et al. MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. **Brief Bioinform,** v. 9, p. 299-306, 2008.

KURTZ, S.; PHILLIPPY, A.; DELCHER, A. L. et al. Versatile and open software for comparing large genomes. **Genome Biol,** v. 5, p. R12, 2004.

LANDER, E. S.; LINTON, L. M.; BIRREN, B. et al. Initial sequencing and analysis of the human genome. **Nature,** v. 409, p. 860-921, 2001.

LANDER, E. S.; WATERMAN, M. S. Genomic mapping by fingerprinting random clones: a mathematical analysis. **Genomics,** v. 2, p. 231-239, 1988.

MACBEATH, J. R.; HARVEY, S. S.; OLDROYD, N. J. Automated fluorescent DNA sequencing on the ABI PRISM 377. **Methods Mol Biol,** v. 167, p. 119-152, 2001.

MAGRANE, M. UniProt Knowledgebase: a hub of integrated protein data. **Database (Oxford),** v. 2011, p. bar009, 2011.

MARDIS, E. R. Next-generation DNA sequencing methods. **Annual Review of Genomics and Human Genetics,** v. 9, p. 387-402, 2008.

MARGULIES, M.; EGHOLM, M.; ALTMAN, W. E. et al. Genome sequencing in microfabricated high-density picolitre reactors. **Nature,** v. 437, p. 376-380, 2005.

MAXAM, A. M.; GILBERT, W. A new method for sequencing DNA. **Proc Natl Acad Sci U S A,** v. 74, p. 560-564, 1977.

MCKERNAN, K.; BLANCHARD, A.; KOTLER, L. et al. Reagents, methods, and libraries for bead-based sequencing. **US patent application 20080003571**, 2006.

MIKKELSEN, T.; HILLER, L. W.; EICHLER, E. E. et al. Initial sequence of the chimpanzee genome and comparison with the human genome. **Nature,** v. 437, p. 69-87, 2005.

MORIYA, Y.; ITOH, M.; OKUDA, S. et al. KAAS: an automatic genome annotation and pathway reconstruction server. **Nucleic Acids Res,** v. 35, p. 182-185, 2007.

MOROZOVA, O.; MARRA, M. A. Applications of next-generation sequencing technologies in functional genomics. **Genomics,** v. 92, p. 255-264, 2008.

MULLIS, K.; FALOONA, F.; SCHARF, S. et al. Specific Enzymatic Amplification of DNA In Vitro: The Polymerase Chain Reaction. **Cold Spring Harbor Symposium in Quantitative Biology,** v. 51, p. 263-273, 1986.

NIERMAN, W. C.; EISEN, J. A.; FLEISCHMANN, R. D. et al. Genome data: what do we learn? **Curr Opin Struct Biol,** v. 10, p. 343-348, 2000.

PROBER, J. M.; TRAINOR, G. L.; DAM, R. J. et al. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. **Science,** v. 238, p. 336-341, 1987.

RONAGHI, M. Pyrosequencing sheds light on DNA sequencing. **Genome Research,** v. 11, p. 3-11, 2001.

SALZBERG, S. L.; DELCHER, A. L.; KASIF, S. et al. Microbial gene identification using interpolated Markov models. **Nucleic Acids Res,** v. 26, p. 544-548, 1998.

SANGER, F.; NICKLEN, S.; COULSON, A. R. DNA sequencing with chain-terminating inhibitors. **Proc Natl Acad Sci U S A,** v. 74, p. 5463-5467, 1977.

SHENDURE, J.; JI, H. Next-generation DNA sequencing. **Nat Biotechnol,** v. 26, p. 1135-1145, 2008.

SHENDURE, J.; PORRECA, G. J.; REPPAS, N. B. et al. Accurate multiplex polony sequencing of an evolved bacterial genome. **Science,** v. 309, p. 1728-1732, 2005.

STEIN, L. Genome annotation: from sequence to biology. **Nat Rev Genet,** v. 2, p. 493-503, 2001.

THOMPSON, J. D.; HIGGINS, D. G.; GIBSON, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. **Nucleic Acids Research,** v. 22, p. 4673-4680, 1994.

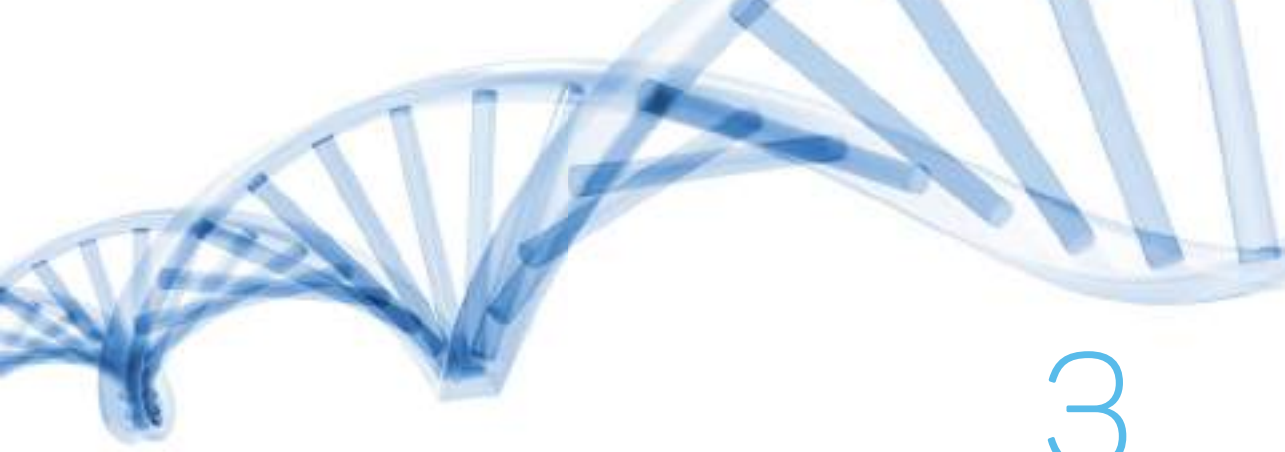
TURCATTI, G.; ROMIEU, A.; FEDURCO, M. et al. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. **Nucleic Acids Res,** v. 36, p. e25, 2008.

VENTER, J. C.; ADAMS, M. D.; MYERS, E. W. et al. The sequence of the human genome. **Science,** v. 291, p. 1304-1351, 2001.

WANG, Z.; GERSTEIN, M.; SNYDER, M. RNA-Seq: a revolutionary tool for transcriptomics. **Nat Rev Genet,** v. 10, p. 57-63, 2009.

WATERSTON, R. H.; LINDBLAD-TOH, K.; BIRNEY, E. et al. Initial sequencing and comparative analysis of the mouse genome. **Nature,** v. 420, p. 520-562, 2002.

WATSON, J. D.; CRICK, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. **Nature,** v. 171, p. 737-738, 1953.



3

Construindo bancos de dados biológicos

Luciano Digiampietri
Jerônimo Conceição Ruiz

Introdução

A bioinformática contemporânea engloba uma crescente e ampla gama de atividades relacionadas ao gerenciamento, análise e visualização da informação biotecnológica. Dentre a pletora de programas e abordagens de bioinformática aplicadas às analises de proteômica, genômica, microarranjos, metabolômica e biologia de sistemas, o desenvolvimento e utilização de banco de dados tem crescido em importância e mostrado sua fundamentalidade não só como essência dos grandes repositórios de dados biológicos como o DDBJ (DNA Data Bank of Japan), EMBL Nucleotide Sequence DB (European Molecular Biology Laboratory) e GenBank (National Center for Biotechnology Information), mas também no cotidiano científico dos laboratórios de pesquisa envolvidos em projetos que empreguem alguma abordagem de sequenciamento de nova geração.

Nosso objetivo nesse capítulo é trazer de uma maneira prática a essencialidade e aplicabilidade dos bancos de dados para o cotidiano das análises em bioinformática. Para tanto, a estratégia adotada integra a utilização de duas ferramentas de código aberto e independentes de plataforma: a linguagem de programação PERL (*Practical Extraction and Report Language*) e o Sistema de Gerenciamento de Banco de Dados (SGBD) MySQL. Pela sua consistência e segurança, o enfoque será dado em um tipo particular de banco de dados, o banco de dados relacional. Nos exemplos práticos utilizados, o MySQL representa o "Sistema Gerenciador de Banco de Dados Relacional" ou SGBDR que viabiliza a mineração das informações contidas nos bancos de dados de maneira extremamente eficiente e frequentemente através de comandos simples

criados através de uma linguagem de programação especial conhecida como SQL ou "*Structured Query Language*".

Tipos comuns de bancos de dados

Tendo como objetivo informar o leitor sobre as diferentes interpretações possíveis considerando o entendimento biológico e computacional de um DB, algumas considerações que levam em conta o *background* do leitor precisam ser feitas.

Assim sendo, no contexto dos leitores com *background* em bioinformática e área afins a definição de SGBC fica voltada para o entendimento de uma coleção de programas e recursos usados para armazenar, gerenciar e possibilitar a extração otimizada de informações em DBs relacionais. Cada SGBD pode gerenciar diversos Bancos de Dados (BD) locais ou distribuídos, sendo que em cada BD pode existir dezenas de centenas de tabelas (relações) extremamente estruturadas e relacionadas entre si (BDs relacionais).

Vale ainda destacar a existência SGBDs de distruição gratuita como MySQL (citado anteriormente) e Postgres e comericais como Oracle, SqlServer e DB2 e que portais como o NCBI, EMBL, DDBJ usam para gerenciar e disponibilizar suas informações aos usuários um tecnologia mista envolvendo SGBD e *flat-file*.

Arquivos texto (Flat text files)

Nesse tipo particular de banco de dados toda informação é armazenada em um esquema simples através de arquivos texto comumente denominada "flat files". Esses textos podem representar qualquer informação e são criados em um formato comum de leitura como os arquivos .txt criados por editores de texto como o TextPad (http://www.textpad.com/download/) ou Kate (http://kate-editor.org/). Um conjunto de documentos como esse no disco rígido do computador, representa um exemplo de banco de dados não realcional, contudo é de grande ajuda se algum tipo de estrutura de ordenamento é imposta à esses arquivos. Em termos de anotação genômica por exemplo, esses arquivos poderiam representar as sequencias de genes codificadores de proteínas, de RNAs (tRNAs e rRNAs) e outros elementos estruturais como repetições e regiões promotoras além é claro dos "contigs" ou "contig" (*Contig é o termo generico utilizado para descrever uma sequência consenso gerada computacionalmente a partir de fragmentos de sequências menores relacionados uns aos outros em função da sobreposição de suas sequências*) que representam a sequencia completa do genoma. A utilização desse tipo de estruturação ajuda o pesquisador que analisa os dados a rapidamente encontrar a informação de interesse e posteriormente aprofundar a pesquisa em uma subseção particular. Essa estruturação também ajuda o processo de extração de relatórios ou "parsing" dos dados. Se considerarmos um banco de dados texto formados apenas por sequencias no formato FASTA, várias informações extremamente úteis poderiam ser extraídas, vejamos o exemplo à seguir:

As duas sequências presentes na Figura 1, (114794895 e 212375028), são apresentadas no formato Fasta (Pearson) (http://www.ebi.ac.uk/help/formats.

```
>gi|114794895|pdb|2HAQ|A Chain A, Cyclophilin A From Leishmania Donovani
HHHHHHEPEVTAKVYFDVMIDSEPLGRITIGLFGKDAPLTTENFRQLCTGEHGFGYKDSI

>gi|212375028|pdb|3EOV|B Chain B, Cyclophilin From Leishmania Donovani
HHHHHHEPEVTAKVYFDVMIDSEPLGRITIGLFGKDAPLTTENFRQLCTGEHGFGYKDSI
```

Figura 1. Exemplos de sequências no formato FASTA.

html#fasta). Esse formato é amplamente utilizado em bionformática e possui várias características que podem ser exploradas.

Com um olhar criterioso poderíamos identificar vários padrões estruturais do texto que descrevem as duas sequencias que poderiam ser utilizados na extração de informações. São eles:

a. O formato Fasta contem uma única linha de cabeçalho para cada sequencia descrita no arquivo que se inicia com o símbolo “>” (maior);
b. Logo após o símbolo de maior (>) vários blocos de informação são separados pelo sinal “|”(pipe). O primeiro bloco de informação representa o gi, ou número de identificação da proteína no NCBI (http://www.ncbi.nlm.nih.gov/Sitemap/sequenceIDs.html), o segundo bloco de informação representa um identificador único do banco de dados PDB (http://www.rcsb.org/pdb/staticHelp.do?p=help/advancedsearch/pdbIDs.html), o terceiro e último bloco representa uma breve descrição da proteína anotada e o organismo de origem seguido pela inserção de uma nova linha. Apesar de serem apresentados apenas três blocos informativos, teoricamente não existe um limite para a quantidade de blocos, contudo apenas uma linha pode ser utilizada. A limitação fica relacionada a dificuldade humana de leitura de uma linha muito longa; e
c. As linhas restantes representam a sequencia propriamente dita. Rotineiramente são representados 60 caracteres por linha mas esse número pode variar até 80 caracteres por linha.

Apesar de o exemplo apresentar apenas um trecho da sequencia de proteína, o padrão identificado é consistente para um arquivo multi-FASTA e assim um dos pontos centrais nesse arquivo é que dentro dele os dados são separados pelo emprego de uma estrutura definida. Nenhuma informação sobre a convenção de nomes utilizada na nomenclatura das proteínas, sua descrição ou sequência é necessária para atribuição dos dados nesses três padrões. Além disso, cada pedaço de informação do arquivo está explicitamente relacionado pela sua localização no arquivo.

Tendo em mente essa abordagem, vários formatos muito mais elaborados como o formato EMBL (http://www.ebi.ac.uk/embl/Documentation/User_manual/usrman.html) e o formato GenBank (http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html) são utilizados em muitos aplicativos de bioinformática.

Arquivos XML (Extensible Markup Language)

O formato XML representa um formato bastante empregado em vários aplicativos de bioinformática e adiciona o conceito de sintaxe a um texto estruturado. Similarmente ao que acontece na linguagem escrita (língua portuguesa por exemplo) onde o emprego de uma pontuação e gramática correta ajudam o entendimento da mensagem escrita, a sintaxe define a ordem dos elementos de linguagem a serem utilizados no texto.

A descrição do arquivo Fasta descrito anteriormente poderia ser feita pela utilização da linguagem XML, contudo a idealização dessa formatação de arquivo foi feita primariamente para ser "machine readable", ou seja, feita para ser interpretada pelo computador, o que torna difícil o entendimento e leitura humana desse tipo de arquivo.

Apesar disso, a sintaxe XML básica é composta por um pequeno número de elementos estruturais. O primeiro elemento básico de um arquivo XML é a declaração que define a versão de XML utilizada e o tipo codificação. Veja o exemplo abaixo:

<? XML version="1.0" enconding="UTF-8"?>

Adicionalmente, vários outros elementos estruturais podem ser utilizados tais como: *names* (nomes), *attributes* (atributos) e *values* (valores). A informação textual pode ser integrada em cada um desses elementos e genericamente poderiam ser escritos da seguinte forma:

<*nome* **atributo_1**="valor_1"
atributo_2="valor_2"

atributo_n="valor_n">
Algum texto o elemento nome
</*nome*>

No contexto da bioinformática que utiliza uma grande e heterogênea variedade de formatos de dados tais como fasta, embl, genbank, csv, resultados de alinhamento (.aln), filogenia, etc, o formato XML é muito interessante uma vez que formatos pré-definidos facilitam a validação e consistência dos dados bem como sua análise. Várias propostas de modelos estão disponíveis para bioinformática entre eles podemos citar: AGAVE (Architecture for Genomica Annotation, Visualization and Exchange | http://www.agavexml.org/), BIOML (BIOpolymer Markup Language | http://www.thegpm.org/bioml/), BSML (Bioinformatic Sequence Markup Language | http://xml.coverpages.org/bsml.html), HSAML (Multiple sequence alignment format with guide tree data anda quality scores | http://www.ebi.ac.uk/goldman-srv/hsaml/) e NCBI (http://www.ncbi.nlm.nih.gov/IEB/ToolBox/XML/).

Como exemplo utilizaremos a proteina com número de acesso GI:7387828 que foi salva do NCBI no formato XML (http://www.ncbi.nlm.nih.gov/protein/7387828) (Figura 2). Devido a extensão do arquivo, e para efeito de maior clareza das informações, apenas parte das informações sobre esta proteína foram apresentada.

```
<!DOCTYPE Bioseq-set SYSTEM "http://www.ncbi.nlm.nih.gov/dtd/NCBI_Seqset.dtd" PUBLIC "-//NCBI//NCBI Seqset/EN">
<Bioseq-set>
  <Bioseq-set_seq-set>
    <Seq-entry>
      <Seq-entry_seq>
        <Bioseq>
          <Bioseq_id>
            <Seq-id>
              <Seq-id_swissprot>
                <Textseq-id>
                  <Textseq-id_name>LEA3_WHEAT</Textseq-id_name>
                  <Textseq-id_accession>Q03968</Textseq-id_accession>
                  <Textseq-id_release>reviewed</Textseq-id_release>
                  <Textseq-id_version>1</Textseq-id_version>
                </Textseq-id>
              </Seq-id_swissprot>
            </Seq-id>
            <Seq-id>
              <Seq-id_gi>7387828</Seq-id_gi>
            </Seq-id>
          </Bioseq_id>
          <Bioseq_descr>
            <Seq-descr>
              <Seqdesc>
                <Seqdesc_title>RecName: Full=Late embryogenesis abundant protein, group 3; Short=LEA; AltName:
Full=PMA2005</Seqdesc_title>
              </Seqdesc>
              ...
    <Title>
                <Title_E>
                  <Title_E_name>Sequence and regulation of a late embryogenesis abundant group 3 protein of maize.</Title_E_name>
                </Title_E>
              </Title>
              ...
        <Bioseq>
      <Seq-entry_seq>
    <Seq-entry>
  <Bioseq-set_seq-set>
<Bioseq-set>
```

Figura 2. Proteína com número de acesso GI:7387828 salva a partir do NCBI no formato XML.

Assim como os bancos de dados gerenciados por um SGBD, os arquivos XML possuem estrutura extremamente flexível. Isto é, não existem regras rigorosas para determinar quais campos e informações farão parte destes arquivos e assim a padronização final fica dependente do problema a ser resolvido, do conhecimento e criatividade de quem esta gerando tais arquivos. Arquivos em formatos XML podem ser complexos para a interpretação humana, porém, são facilmente interpretados por inúmeros programas de computador, inclusive alguns SGBDs, navegadores WEB e bibliotecas de diversas linguagens de programação como PERL (http://search.cpan.org/~msergeant/XML-Parser-2.36/Parser.pm). Por este motivo XML é um formato muito utilizado para migrar (transferir) dados de um determinado sistema informatizado para outro.

Relevância da linguagem Perl

A linguagem de programação Perl (*Practical Extraction and Report Language | http://www.perl.com*) foi originalmente concebida para processar textos de maneira fácil e otimizada. Esta é uma das principais características que popularizou esta linguagem nas ferramentas de bioinformática nas quais, comumente, é necessário processar textos (sequências de caracteres) utilizando poderosos mecanismos de expressões regulares.

Há três tipos principais de variáveis na linguagem Perl: escalares (simples), arranjos (ou vetores) e hashes. As variáveis não precisam ser declaradas e o tipo do conteúdo de uma variável é atribuído dinamicamente, ou seja, uma variável pode receber valores numéricos e efetuar cálculos sobre esses valores e em seguida concatenar uma sequência de caracteres a esta mesma variável. A Figura 3 apresenta um código simples escrito em Perl que utiliza os três principais tipos de variáveis da linguagem. Na linha 2 uma variável escalar chamada *$simples* recebe o valor 5, que é multiplicado por 4 na linha 3 e o seu resultado é impresso na linha 4. Na linha 5 essa variável tem um texto concatenado ao seu valor e o novo valor da variável *$simples* é impresso na linha 6. Na linha 7 um arranjo chamado *@arranjo* recebe três valores. O laço das linhas 8 a 10 imprime cada um dos valores do arranjo. A expressão *$#arranjo* indica o índice do último elemento na variável *@arranjo* e neste caso *$#arranjo* vale 2, pois há três elementos (das posições 0 até 2). Na linha 11 uma variável do tipo hash, chamada *%hash* é declarada nas linhas seguintes duas chaves e seus respectivos valores são inseridos

```
1  #!/usr/bin/perl -w
2  $simples = 5;
3  $simples *= 4;
4  print "$simples\n";
5  $simples = $simples . ", este foi o resultado da multiplicação.";
6  print "$simples\n";
7  @arranjo = ('a', 'b', 'c');
8  for ($i=0;$i<=$#arranjo;$i++){
9    print "$arranjo[$i]\n";
10 }
11 my %hash;
12 $hash{'h1'} = "elemento h1";
13 $hash{'h2'} = "elemento h2";
14 verificaChave('h1');
15 verificaChave('h3');
16 exit 0;
17 sub verificaChave{
18   $chave = shift;
19     if (defined($hash{$chave})){
20       print "Chave $chave definida: $hash{$chave}\n";
21     }else{
22       print "Chave $chave não definida.\n";
23   }
24 }
25
```

Figura 3. Exemplo de Variáveis em Perl.

neste hash. Uma função (ou subrotina) é criada da linha 18 a 25 para verificar se uma dada chave existe ou não no hash. Note que tipos escalares, arranjos e hashs, quando declarados, utilizam respectivamente os seguintes caracteres especiais $, @ e % (linhas 2, 7 e 11). Porém, para acessar o conteúdo individual de um escalar de um arranho ou hash utilliza-se o caractere $ (linhas 9, 12, 13, 20 e 21 do código).

O resultado da execução do programa anterior (Figura 3) pode ser visto na Figura 4.

Um pequeno exemplo de um programa escrito em Perl para processar uma sequência de caracteres é apresentado na Figura 5. Neste exemplo deseja-se analisar a sequência de aminoácidos de uma proteína alfa-amilase da *Haloarcula hispanica* (sequência retirada do GenBank). Na primeira linha do código há uma indicação de como esse programa deve ser processado: pelo programa perl, localizado em /usr/bin e com o parâmetro -w indicando que os avisos (*warnings*) devem ser exibidos caso ocorram operações ilegais. Na segunda linha é declarada uma variável chamada $sequencia que recebe como (valor) parâmetro as informações da alfa-amilase (tanto

```
20
20, este foi o resultado da multiplicacao.
a
b
c
Chave h1 definida: elemento h1
Chave h3 não definida.
```

Figura 4. Resultado da Execução do Programa na Figura 3 do Capítulo 2.

```
1  #!/usr/bin/perl -w
2  my $sequencia = "
3  >CAI64586 alpha-amylase [Haloarcula hispanica].
4  MNRPRITGSKQASRRTVLKGIGVLGAAVFGTAASVGSSAAVGDSAVYQYYHTDWTEITAT
5  LETVAQQGYDAIQVPPAQRSRLDRSHQNGVTDPPLGYQPVDLTDFNSVFGTEDEYEAMVQ
6  EAHNQDLDVVADAVINHMAANDDFRDAPGITFADLPRFSERDFHPKDDINYDNPESVEDD
7  WLVGLKDLKQESAYVRGELQAYVQKYADLGVDGIRWDAAKHVPESFFADYANQWADDLDL
8  WTVGEVLDGDIGVCQGYADTGMSVTDYPLYYTMKEEAFHSDGNMQALDGAGMVNQSPFQA
9  FTFVSNHDSGPPDYEKLAYAYILTYEGYPRVYSNRISVDDDDIRNLLWIRNNLASGGSQT
10 RHVDQDLYVYEREGNLLVGLNRASGQRSKWVPTSWTNQTLNDYSGNAGNISTNGDSWVQI
11 TVPATSWVCYAPE";
12 $sequencia =~s/.+\.//;
13 $sequencia =~s/\n//g;
14 if ($sequencia =~ /DDD(...)/){
15   print "$1\n";
16 }
17 exit 0;
```

Figura 5. Exemplo de Uso de Expressão Regular em Perl.

```
1  create database livro;
2  use livro;
3  create table `ResultadosBlast` (
4    `QueryAccession` varchar(63) NOT NULL,
5    `Accession` varchar(63) NOT NULL,
6    `Description` varchar(1023) NOT NULL,
7    `Genus` varchar(511) NOT NULL,
8    `MaxScore` int(11) NOT NULL,
9    `TotalScore` int(11) NOT NULL,
0    `QueryCoverage` int(11) NOT NULL,
10   `EValue` double NOT NULL,
11   primary key (`QueryAccession`,`Accession`)
12 );
13
```

Figura 6. Comandos MySQL para criação de banco de dados e tabela.

o cabeçalho quanto a sequência propriamente dita). As linhas 12, 13 e 14 possuem diferentes expressões regulares. A expressão da linha 12 faz com que todo conteúdo (.+) anterior ao ponto final (\.) seja substituído por nada (ou seja, removido), com esta expressão o cabeçalho da sequência é removido. A expressão regular da linha 13 faz que todos os "finais de linha" (\n) sejam removidos, ou seja, as diversas linhas da sequência serão concatenadas em uma única linha. Por fim, a expressão regular da linha 14, que está dentro de um desvio condicional (**if**), verifica se existe nesta sequência uma subsequência iniciada por três aminoácidos D seguidos por três aminoácidos quaisquer: **DDD(...).** Os parênteses nessa expressão servem para indicar que o conteúdo entre eles deverá ser armazenado automaticamente na variável especial do perl ($1) para posterior utilização. Na linha 12 é solicitado que o conteúdo de $1 seja impresso seguido de uma instrução de "fim de linha" (\n). O resultado da execução desse programa será a impressão na tela dos aminoácidos **DIR**, pois, como visto na linha 9 do código, os três primeiros aminoácidos posteriores ao padrão de três letras D são os aminoácidos **DIR**. Vale a pena lembrar que dependendo da base de dados utilizada, os aminoácidos podem ser apresentados por letras minúsculas ou maiúsculas, como é o caso das sequências do GenBank onde são representadas por letras maiúsculas.

O SGBD MySQL

De acordo com a Seção 2 deste capítulo, MySQL é um dos mais populares sistemas gerenciadores de bancos de dados (SGBD) gratuitos.

Nesta seção será criado um banco de dados extremamente simples dentro do MySQL, que será utilizado nos exemplos do uso de Perl juntamente com MySQL.

A Figura 6 apresenta um conjunto de comandos para a criação de um banco de dados e de uma tabela. Esses comandos devem ser executados de dentro do MySQL.

```
use livros;
show tables;
+-----------------+
| Tables_in_livro |
+-----------------+
| ResultadosBlast |
+-----------------+

desc ResultadosBlast ;
+----------------+---------------+------+-----+---------+-------+
| Field          | Type          | Null | Key | Default | Extra |
+----------------+---------------+------+-----+---------+-------+
| QueryAccession | varchar(63)   | NO   | PRI | NULL    |       |
| Accession      | varchar(63)   | NO   | PRI | NULL    |       |
| Description    | varchar(1023) | NO   |     | NULL    |       |
| Genus          | varchar(511)  | NO   |     | NULL    |       |
| MaxScore       | int(11)       | NO   |     | NULL    |       |
| TotalScore     | int(11)       | NO   |     | NULL    |       |
| QueryCoverage  | int(11)       | NO   |     | NULL    |       |
| EValue         | double        | NO   |     | NULL    |       |
+----------------+---------------+------+-----+---------+-------+
```

Figura 7. Comandos MySQL: show tables e desc.

Na linha número 1 (um) desta figura é possível observar a criação de um novo banco de dados chamado livro (dentro de um SGBD vários bancos de dados podem ser criados). Na segunda linha a instrução é que utilizaremos agora este banco para a criação de uma nova tabela. As linhas de 4 a 13 descrevem a estrutura da tabela que será criada (ResultadosBlast). Esta tabela deverá albergar um resumo dos resultados de busca por similaridade de sequências produzidas pela ferramenta BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi). As linhas de 5 a 11 descrevem os campos da tabela e a linha 12 informa que a chave primária da tabela será composta pelos campos QueryAccession e Accession.

A Figura 7 apresenta dois comandos usados para consultar informações gerais de um banco de dados. Os resultados destes comandos, quando utilizados dentro do nosso banco de dados "livro" são: **show tables;** que exibe a lista de tabelas dos banco de dados e **desc <nome_da_tabela>;** que exibe a descrição de uma tabela, isto é, nome dos campos, tipos, etc.

Acessando banco de dados relacionais utilizando Perl

Em Perl, assim como na maioria das linguagens de programação, existem bibliotecas (ou módulos) para facilitar o acesso aos bancos de dados. Neste capítulo será utilizada a biblioteca DBI[2] (*Database Interface* | http://dbi.perl.org/) para a comunicação com o SGBD MySQL. Para utilizar (ou importar) bibliotecas em Perl, basta tê-las instaladas em seu sistema e utilizar o comando *"use"* seguido do nome da biblioteca.

A Figura 8 apresenta um programa Perl que abre um arquivo do tipo CSV (*Comma Separated Values*) que contém a tabela dos *hits* de alinhamentos produzidos pelo programa BLAST e salvo previamente no computador (arquivo Alignment-HitTable.

```
1  #!/usr/bin/perl -w
2  use DBI;
3  if ($#ARGV<1){
4   die "Modo de uso: perl -w parserBlastTabela.pl <nome do arquivo blast> <query accession id>\nExemplo: perl -w parserBlastTabela.pl Alignment-HitTable.csv CAI64586.1";
5  }
6  my $arquivo = $ARGV[0];
7  my $queryAccession = $ARGV[1];
8  my $DATABASE = 'livro';
9  my $TABLE = 'ResultadosBlast';
10 my $CONFIG   = ";mysql_read_default_file=/home/usuario/.my.cnf";
11 my $dbh = DBI->connect("DBI:mysql:$DATABASE" . $CONFIG, , ) or die "Não foi possivel conectar no bando de dados '$DATABASE'\n";
12 my $sth;
13 open(IN, "< $arquivo") or die "Nao foi possivel abrir arquivo de entrada: '$arquivo'\n";
14 my @campos;
15 my $SQL_INSERT;
16 my $genero;
17 while (<IN>){
18  chomp;
19  @campos = split("\t", $_);
20  $campos[4]=~s/\%//;
21  if ($campos[1] =~ /\[/){
22                  $genero = $campos[1];
23                  $genero =~s/\].*//;
24                  $genero =~s/.*\[//;
25                  $genero =~s/ .*//;
26                  $SQL_INSERT = "INSERT INTO $DATABASE.$TABLE VALUES ('$queryAccession', '$campos[0]', '$campos[1]', '$genero', $campos[2], $campos[3], $campos[4],
27 $campos[5]);";
28                  $sth = $dbh->prepare($SQL_INSERT);
29                  if ($sth->execute){
30                   print "Inserção realizada com sucesso.\n";
31                  }else{
32                   print "Não foi possivel inserir o registro.\n";
33                  }
34  }
35 }
36 close IN;
37 exit 0;
```

Figura 8. Programa para Ler Resultados do BLAST e Armazená-los no Banco de Dados.

csv). Para gerar este arquivo, foi feito um alinhamento de aminoácidos da sequência da alfa-amilase CAI64586.1 contra o banco *nr (non redundante database)* do GenBank. Na Figura 3.6, observamos na linha 2 o uso da biblioteca DBI. A condição da linha 3 verifica se menos que dois parâmetros foram passados como entrada para o programa, em caso afirmativo a execução será cancelada, pois o programa espera receber dois parâmetros: o nome do arquivo de entrada e a identificação da sequência que foi usada como sequência de consulta do alinhamento. Nas linhas 6 e 7 as variáveis do programa recebem o valor dos parâmetros de entrada. Das linhas 8 a 10 há a definição de algumas variáveis para acesso ao banco de dados, como o nome do banco (livro), o nome da tabela (ResultadosBlast) e por fim parâmetros de configuração que indicam como a autenticação será feita no banco de dados. Neste caso utilizamos o arquivo *.my.cnf (O arquivo .my.cnf contém as informações necessários para o estabelecimento de uma conexão com o banco de dados, por exemplo, nome do usuário e senha. Usando este arquivo o nome do usuário e a senha não ficam expostos no código do programa)* que foi criado no diretório exemplo */home/usuario*. Na linha 11 há existe conexão com o banco de dados e é importante observar que um dos parâmetros indica que o SGBD será o MySQL. Na linha 13 o arquivo de entrada é aberto para leitura. O laço que se estende da linha 17 até a linha 35 será executado enquanto houver linhas a serem lidas no arquivo de entrada e cada uma dessas linhas do arquivo será jogada numa variável padrão chamada *$_*. Das linhas 19 a 26 alguns processamentos são feitos para separar cada linha de entrada em campos e para tentar identificar o gênero de cada uma das descrições dos resultados do BLAST. Na linha 27 foi definida uma variável com o comando SQL para inserção dos resultados BLAST no banco de dados. Nessa linha

é necessário dizer em que tabela será feita a inserção e definir os valores que serão inseridos em cada campo da tabela, separados por vírgula. Na linha 28 a variável $sth recebe o comando pré-processado pela biblioteca DBI e na linha 29 esse comando será executado. Se não houver erro na execução do comando a mensagem "Inserção realizada com sucesso." será exibida na tela. Na linha 36 o arquivo de entrada é fechado.

BioPerl

BioPerl (http://www.bioperl.org) consiste de uma biblioteca desenvolvida para facilitar o processamento de dados biológicos. Esta biblioteca apresenta diversas funcionalidades relacionadas a bioinformática, ferramentas para copiar sequências e outras informações de bancos de dados importantes, como o GenBank, e tratamento de diferentes tipos de arquivos, como FASTA ou resultados da ferramenta BLAST.

O pacote chamado **bioperl** encontra-se disponível no repositório de pacotes para Linux, podendo ser facilmente instalado (http://www.bioperl.org/wiki/Installing_BioPerl). Para utilizar o pacote BioPerl em um programa Perl basta inserir a linha "use Bio::Perl;" no início do código e utilizar os métodos e funções desejadas. A Figura 9 apresenta um código que utiliza funções da biblioteca bioperl para baixar sequências do GenBank no formato FASTA, tendo como parâmetro o *accession* da sequência. A linha 2 deste código informa que o código utilizará a biblioteca BioPerl. Na linha 3 é declarada a variável *database* com o valor *genbank*. A linha 4 contém um arranjo com a identificação das sequências que serão baixadas, neste exemplo há apenas duas sequências no arranjo. Na linha 5 é definida uma variável *format* que indica que o formato utilizado será o FASTA. Das linhas 6 a 9 há um laço iterativo que para cada um dos identificadores de sequência do arranjo *accessions* fará o seguinte: utilizará a função *get_sequence* da biblioteca BioPerl para receber e guardar na variável *sequence* a sequência desejada (linha 7) e utilizará a função *write_sequence* para imprimir essa sequência no formato FASTA.

O resultado da execução desse código pode ser visto na Figura 10.

```
1 #!/usr/bin/perl -w
2 use Bio::Perl;
3 $database="genbank";
4 @accessions = ("CAI64586.1","BAD06002.1");
5 $format="fasta";
6 for (my $i=0;$i<=$#accessions;$i++){
7   $sequence = get_sequence($database, $accessions[$i]);
8   write_sequence(">-", $format, $sequence);
9 }
```

Figura 9. Exemplo de Código Utilizando BioPerl.

```
>CAI64586 alpha-amylase [Haloarcula hispanica].
MNRPRITGSKQASRRTVLKGIGVLGAAVFGTAASVGSSAAVGDSAVYQYYHTDWTEITAT
LETVAQQGYDAIQVPPAQRSRLDRSHQNGVTDPPLGYQPVDLTDFNSVFGTEDEYEAMVQ
EAHNQDLDVVADAVINHMAANDDFRDAPGITFADLPRFSERDFHPKDDINYDNPESVEDD
WLVGLKDLKQESAYVRGELQAYVQKYADLGVDGIRWDAAKHVPESFFADYANQWADDLDL
WTVGEVLDGDIGVCQGYADTGMSVTDYPLYYTMKEEAFHSDGNMQALDGAGMVNQSPFQA
FTFVSNHDSGPPDYEKLAYAYILTYEGYPRVYSNRISVDDDDIRNLLWIRNNLASGGSQT
RHVDQDLYVYEREGNLLVGLNRASGQRSKWVPTSWTNQTLNDYSGNAGNISTNGDSWVQI
TVPATSWVCYAPE
>BAD06002 alpha-amylase [Aspergillus awamori].
MMVAWWSLFLYGLQVAAPALAATPADWRSQSIYFLLTDRFARTDGSTTATCNTADQKYCG
GTWQGIIDKLDYIQGMGFTAIWITPVTAQLPQTTAYGDAYHGYWQQDIYSLNENYGTADD
LKALSSALHERGMYLMVDVVANHMGYDGAGSSVDYSVFKPFSSQDYFHPFCFIQNYEDQT
QVEDCWLGDNTVSLPDLDTTKDVVKNEWFDWVGSLVSNYSIDGLRIDTVKHVQKDFWPGY
NKAAGVYCIGEVLDGDPAYTCPYQNVMDGVLNYPIYYPLLNAFKSTSGSMDDLYNMINTV
KSDCPDSTLLGTFVENHDNPRFASYTNDIALAKNVAAFIILNDGIPIIYAGQEQHYAGGN
DPANREATWLSGYPTDSELYKLIASANAIRNYAISKDTGFVTYKNWPIYKDDTTIAMRKG
TDGSQIVTILSNKGASGDSYTLSLSGAGYTAGQQLTEVIGCTTVTVGSDGNVPVPMAGGL
PRVLYPTEKLAGSKICSSS
```

Figura 10. Saída Produzida pelo Código da Figura 9.

Interação na Web

Enquanto grande parte do processamento relacionado a bioinformática é feito em servidores potentes, a consulta, a anotação e o refinamento dos resultados costumam ser tarefas feitas por usuários que interagem com os dados produzidos via navegador Web.

Conceitos básicos de HTML

A maioria dos sites na Internet está escrita em HTML (*HyperText Markup Language*), uma linguagem de marcação que define os elementos básicos de um site, por exemplo, título, tabelas e figuras. A linguagem HTML é interpretada pelos navegadores Web permitindo uma fácil visualização de textos, imagens, sons e vídeos, além de possibilitar a navegação entre diferentes sites. Uma página (ou site) HTML é um conjunto de marcações (tags) e conteúdos. A Figura 11 representa uma página HTML extremamente simples. Nesta figura podemos observar algumas marcações como o cabeçalho (head) que possui um título (title) cujo conteúdo é "Olá mundo" e o corpo da página HTML delimitada pelas marcações body. Cada marcação é iniciada pelo seu nome dentro de sinais de menor e maior (<html>) e encerrada pela marcação com o mesmo nome porém com um barra (</html>), alternativamente, marcações que não possuam marcações internas podem ser iniciadas e encerradas com uma única marcação: <nome_da_marcação ... />. A marcação META tem objetivo, nesta página, de informar qual será a codificação e conteúdo da página. Neste caso o conteúdo é texto/html e a codificação dos caracteres é a UTF-8 (codificação que permite que acentos sejam utilizados sem a necessidade de marcações especiais).

```
<html>
 <head>
  <META http-equiv="Content-Type" content="text/html; charset=UTF-8"/>
  <title>Olá mundo!</title>
 </head>
 <body>
  <h2>Olá mundo!</h2>
  Uma página HTML extremamente simples.
 </body>
</html>
```

Figura 11. Página HTML simple.

Diversas tecnologias e linguagens de programação foram desenvolvidas para complementar a experiência na navegação Web possibilitando um nível de interação muito maior com o usuário, bem como permitir que diferentes tipos de processamento sejam feitos no navegador Web ou no servidor. Por exemplo, javascript, php e CGI.

Uso de CGI

CGI (Common Gateway Interface | http://tools.ietf.org/html/rfc3875) é um padrão que permite aos servidores Web compartilhar a responsabilidade da geração de páginas Web com scripts CGI. Em outras palavras, o servidor Web, ao atender uma requisição de um usuário para a visualização de uma página CGI, solicita que o script correspondente seja executado e a execução deste programa produzirá a página solicitada pelo usuário.

Para se utilizar CGI é necessário configurar seu servidor Web de forma que ele identifique alguns diretórios do computador como autorizados a executar scripts CGI. Um diretório tipicamente utilizado para este fim é o /var/www/cgi-bin.

CGI permite que diferentes linguagens de programação sejam usadas para a produção de páginas Web e interação com o usuário. Neste capítulo descreveremos apenas o uso de CGI juntamente com a linguagem Perl.

Scripts CGI escritos em Perl

Para se fazer um script CGI escrito em Perl, basta desenvolver o programa Perl de forma a imprimir todo o conteúdo HTML desejado, com uma única restrição: a primeira linha a ser impressa deverá informar o tipo de conteúdo que está sendo produzido, conforme código da Figura 12. Este código apresenta um script perl para gerar uma página HTML simples. A primeira linha do código informa que este código é um script perl e deverá ser interpretado pelo comando /usr/bin/perl; a segunda linha informa o tipo de conteúdo que está sendo produzido, sem esta linha ocorrerá o seguinte erro no servidor Web: “Premature end of script headers: ola_mundo.cgi”, considerando que nosso script chama-se ola_mundo.cgi.

```
#!/usr/bin/perl
print "Content-Type: text/html; charset=UTF-8\r\n\r\n";
print '<html>';
print '<head>';
print '<title>Olá mundo!</title>';
print '</head>';
print '<body>';
print '<h2>Olá mundo!</h2>';
print 'Um programa CGI extremamente simples.';
print '</body>';
print '</html>';
```

Figura 12. Script Perl para a geração de HTML simples.

```
#!/usr/bin/perl
use CGI;
my $meuCGI = new CGI;
print $meuCGI->header(-charset=>"UTF-8");
print $meuCGI->start_html(-title=>"Olá mundo!");
print '<h2>Olá mundo!</h2>';
print 'Um programa CGI extremamente simples que usa a biblioteca CGI.';
print $meuCGI->end_html;
```

Figura 13. Script Perl para a geração de HTML usando a biblioteca CGI.

A linguagem de programação Perl, assim como diversas outras linguagens, possui uma biblioteca para facilitar o desenvolvimento de scripts CGI. Com esta biblioteca o usuário não precisa se lembrar dos comandos ou tags específicas para a construção de uma página HTML e, mesmo usando a biblioteca, ainda poderá imprimir comandos HTML da mesma forma que faria sem o uso desta biblioteca. A Figura 13 apresenta o código equivalente ao da Figura FF2, porém utilizando A biblioteca CGI.

Recebendo e retornando informações básicas

Há diversos tipos de campos HTML feitos para o envio/recebimento de informações. Nesta seção será apresentado um pequeno exemplo que utiliza a maioria desses campos. Neste exemplo, o usuário será convidado a preencher um conjunto de campos de um formulário Web gerado e cujos dados serão processados por um script CGI. A Figura BB1 apresenta este script. A biblioteca CGI é usada para produzir o

conteúdo HTML dos diferentes campos do formulário (*textfield*, *textarea*, *checkbox*, *radio_group*, e *submit*), bem como para recuperar o valor desses campos. A condição da linha 7 do código verifica se o campo chamado *botao* não está definido. Se não estiver definido, o código das linhas 8 a 13 produzirá o formulário HTML, caso já esteja definido, o código das linhas 15 a 25 produzirá uma página HTML com o resultado do processamento das informações passadas pelo usuário. A Figura 14 apresenta o formulário a ser preenchido pelo usuário e a Figura 15 apresenta o resultado do processamento desse formulário.

No material suplementar deste livro você encontrará um script CGI Perl que produz um formulário HTML. Esse formulário será capaz de receber um arquivo contendo uma lista de identificadores de sequência, copiar esse arquivo localmente, buscar no GenBank cada uma das sequências e imprimir seus dados no formato FASTA.

Permitindo interação Web com um banco de dados

É possível utilizar informações recebidas por scripts CGI Perl para consultar ou atualizar o banco de dados. Para isto, basta combinarmos o recebimento de informações de formulários HTML com o uso de bancos de dados dentro de um script Perl.

No material suplementar deste livro você encontrará um script CGI Perl que produz um formulário HTML que permite ao usuário escolher qual termo procurar no banco de dados.

```
1  #!/usr/bin/perl
2  use CGI;
3  my $meuCGI = new CGI;
4  print $meuCGI->header(-charset=>"UTF-8");
5  print $meuCGI->start_html(-title=>"Formulágio Perl CGI");
6  print $meuCGI->start_multipart_form;
7  if (!defined($meuCGI->param('botao'))){
8   print '<h2>Olá, favor preencher o formulário e pressionar "Ok".</h2>';
9   print "Nome: ", $meuCGI->textfield(-name => 'nome', -size => 25, -maxlength => 25);
10  print "<br>Endereço: ", $meuCGI->textarea(-name => 'endereco', -value => 'Preencha aqui seu endereço.', -cols => 50, -rows => 4);
11  print "<br>Você é professor: ", $meuCGI->checkbox(-name => 'ehProfessor', -label => 'sim', -checked => 0);
12  print "<br>Sexo: ", $meuCGI->radio_group(-name => 'sexo', -values  => ['feminino', 'masculino'], -default => 'feminino', -columns => 1, -rows => 2);
13  print "<center>", $meuCGI->submit(-name => 'botao', -value  => 'Ok');
14 }else{
15  my $textoProfessor = "";
16  if ($meuCGI->param('ehProfessor')  eq 'on'){
17   if ($meuCGI->param('sexo') eq 'feminino'){
18                    $textoProfessor = "professora ";
19   }else{
20    $textoProfessor = "professor ";
21   }
22  }
23  print '<h2>Formulário preenchido corretamente.</h2>';
24  print "Obrigado $textoProfessor", $meuCGI->param('nome'), ".";
25  print "<br>Endereco: ", $meuCGI->param('endereco');
26 }
27 print $meuCGI->end_form;
28 print $meuCGI->end_html;
```

Figura 14. Script Perl para a geração e processamento de um formulário HTML.

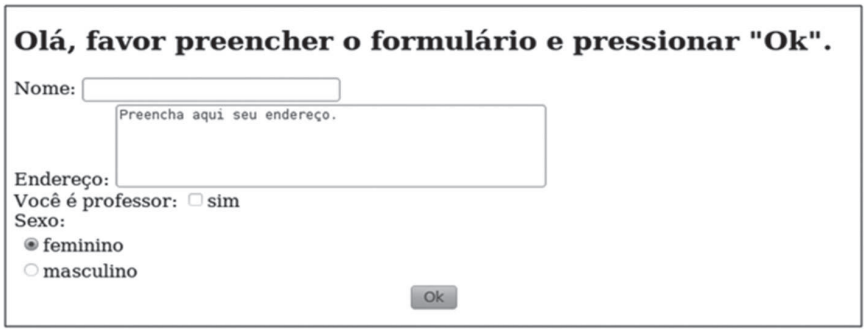


Figura 15. Formulário HTML produzido pelo script da Figura 14.

Figura 16. Página HTML com o resultado do processamento dos dados.

Bibliografias

ALTSCHUL, S; GISH, W; MILLER, W; MYERS, E; LIPMAN, D. *Basic local alignment search tool*. Journal of Molecular Biology. 1990. 215 (3): 403–410.

CHROMATIC. *Modern Perl*. Onyx Neon Press, pp. 2004, 2012.

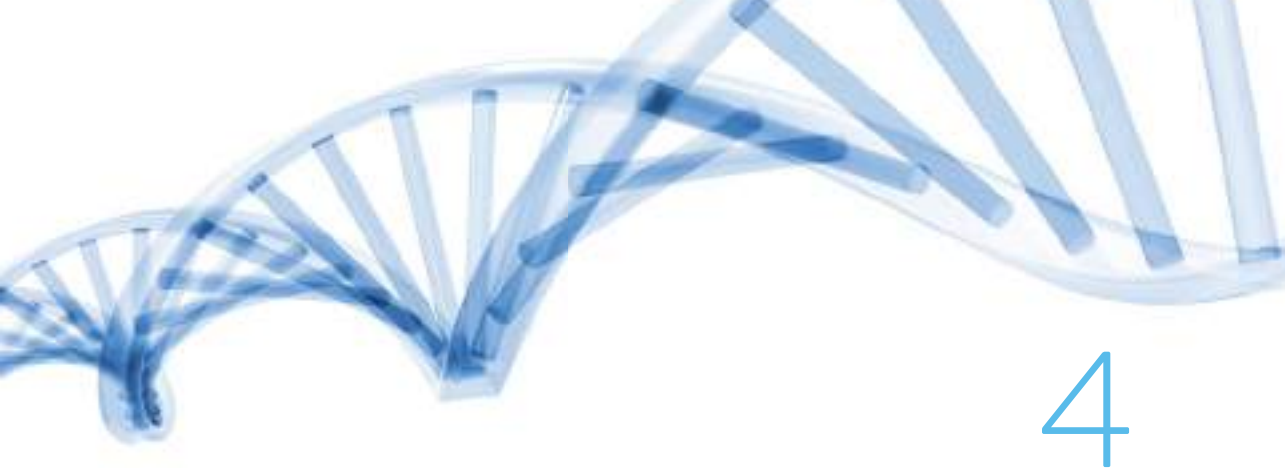
BIOPERL. QUICK START, 2013. Disponível em http://www.bioperl.org/wiki/Quick_start . Acessado em 24/02/2013

MUSCIANO, C.; KENNEDY, B. *HTML & XHTML: The Definitive Guide, 6th Edition*. O'Reilly Media, pp. 680, 2009.

MySQL. MySQL 5.6 Reference Manual, 2013. Disponível em http://dev.mysql.com/doc/refman/5.6/en/ . Acessado em 24/02/2013

POE, C.. *Beginning Perl*. Wrox, pp. 744, 2012.

WEINMAN , W.E.. *The CGI Book*. New Riders Pub, pp. 304, 1996.



4

Genômica comparativa

Francisco Prosdócimi
Leandro Marcio Moreira

Introdução: por que comparar genomas?

Desde que os primeiros genomas começaram a ser produzidos na década de 90 (Capítulo 2), os pesquisadores perceberam que a informação de um único genoma era difícil de ser analisada separadamente. A compreensão da biologia e a busca por explicações do tipo "por que" nos estudos da vida sempre tiveram suas respostas na análise comparativa e evolutiva das informações dispostas. Um exemplo que retrata bem esta perspectiva tem relação com a evolução do sistema nervoso em primatas (Figura 1). Jamais descobriríamos que uma das grandes características da evolução humana foi um enorme desenvolvimento do córtex cerebral a partir de nossos ancestrais primatas, se não tivéssemos meios de comparar os crânios para percebermos as novidades evolutivas que ocorreram ao longo da nossa evolução e da evolução em diversos grupos animais. Essas novidades evolutivas percebidas na comparação entre **clados** é conhecida em biologia com o nome de **apomorfia**.

Atualmente, todo o tipo de informação biológica que tende a responder alguma pergunta com relação à motivação de um organismo possuir determinada característica baseia-se, essencialmente, na observação da função dessa característica/órgão/sequência de DNA e a presença ou ausência dessa mesma característica em organismos de outros grupos animais.

Nesse mesmo contexto, vale notar a tentativa de alguns cientistas em realizar uma **analogia** da genômica com a anatomia. Portanto, comparar presença ou ausência de órgãos em organismos ou presença ou ausência de genes em genomas retratam, de forma mais geral, o mesmo tipo de pensamento científico que utilizamos para descobrir a evolução do cérebro no desenvolvimento dos primatas.

O conhecimento da anatomia foi de extrema importância para a compreensão geral da organização dos organismos biológicos. **Lineu**, ainda no século XVI, foi capaz de

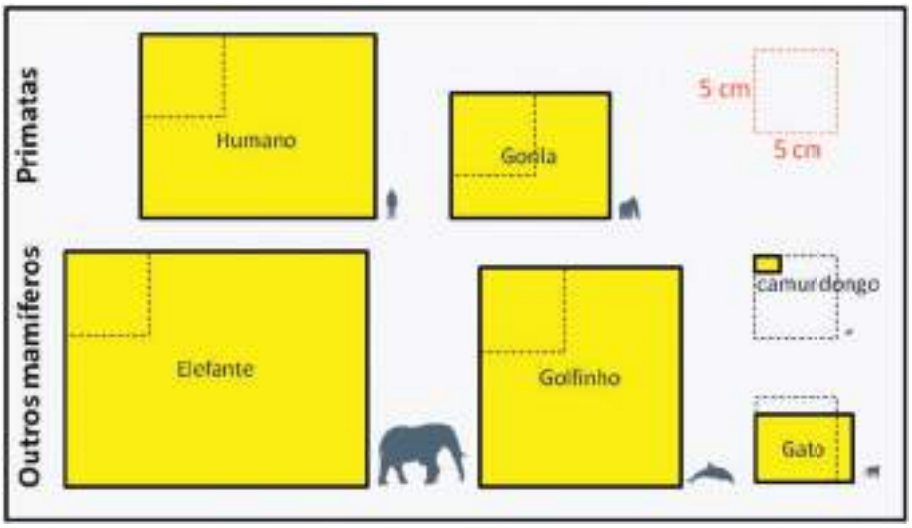


Figura 1. Correlação e evolução no tamanho de cérebros entre diferentes vertebrados. Em amarelo destacam-se os tamanhos totais aproximados dos cérebros de cada organismo em destaque. A referência de tamanho pode ser tomada com base no quadrado pontilhado que possui aresta de 5 cm. Para tornar a análise mais interessante, procure comparar as dimensões do cérebro entre si, mas também com as dimensões corporais relativas das espécies em questão (modelo sombreado posicionado ao lado). Observe que nos primatas as dimensões cerebrais são evidenciadas.

realizar categorizações bastante detalhadas da relação de parentesco entre os mais diversos grupos de organismos baseado apenas na semelhança ou diferença entre as formas básicas dos animais.

É evidente que Lineu cometeu equívocos. No entanto, sua classificação foi e é ainda bastante usada como modelo estrutural para a categorização dos organismos vivos, tanto no que concerne ao sistema cladístico básico e a divisão dos organismos em reinos, filos, classes, ordens, etc., quando ao sistema binomial de classificação biológica. Também no século XIX, os grandes naturalistas e exploradores utilizaram as informações e as comparações da anatomia/morfologia dos organismos vivos para classificar e conhecer diversas novas espécies desconhecidas dos Europeus, que passaram a serem descobertas nas Américas e por todo o globo.

Talvez o exemplo de maior repercussão tenha sido correlacionado com a diferença entre os bicos dos tentilhões (diversas espécies do gênero *Thraupidae*). Estas aves habitavam diferentes ilhas do arquipélago de Galápagos e foram modelos cruciais para Darwin erigir a teoria que viria a ser considerada a base explicativa de toda a biologia desde então (Figura 2). Segundo observações do naturalista, os tentilhões com diferentes formas de bico tinham também diferentes hábitos alimentares, o que o levou a sugerir uma adaptação local de cada pássaro à dieta mais comum em cada uma das ilhas, tendo assim o vislumbre intelectual para a descoberta do mecanismo de seleção natural.

Diante do que fora anteriormente descrito, fica evidenciado que o conhecimento da anatomia de um organismo é importante principalmente porque permite catalogar e inferir funções prováveis de órgãos. Tais funções podem ser então comparadas

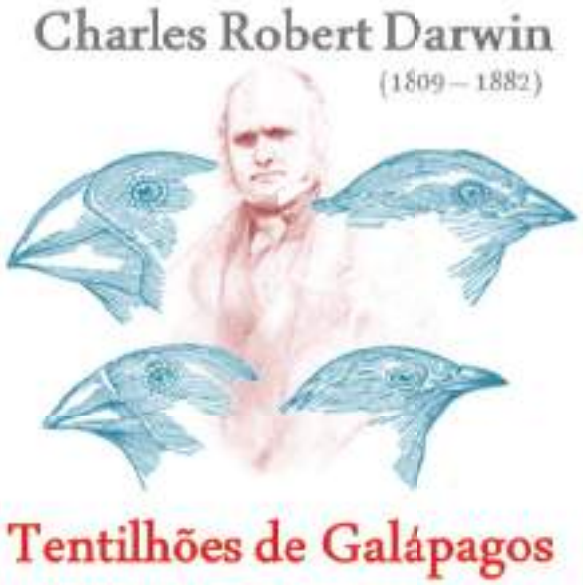


Figura 2. Evolução dos bicos dos tentilhões proposta por Charles Darwin. O formato do bico retrata uma perfeita adaptação da espécie à disponibilidade alimentar em diferentes ilhas do arquipélago de Galápagos. A partir de observações como esta surgiu uma nova linha de estudo baseada na evolução das espécies a partir da seleção natural.

entre diferentes organismos vivendo em diferentes habitats, competindo de forma diferenciada com indivíduos de sua própria espécie ou de outras espécies.

A questão da anatomia versus a genômica precisa, no entanto, ainda ser melhor explorada. Assim como os dados de anatomia, os dados de genomas podem ser considerados, rigorosamente falando, apenas informações descritivas sobre os organismos e não encerram em si mesmos nenhum tipo de informação. Entretanto, quando comparamos o número de apêndices motores ou o tamanho e a similaridade geral entre formas anatômicas de organismos, conseguimos entender as relações evolutivas entre eles, conseguimos criar hipóteses para explicar as diferenças nas anatomias e podemos também tentar entender como a seleção natural atuou ao longo da evolução daquele organismo de forma a selecionar as variantes mais adaptadas a um determinado meio ambiente (Figura 3).

Quando trabalhamos com genômica, a mesma analogia pode ser aplicada. Observamos e sequenciamos os genomas dos organismos. Com um único genoma em mãos, sabemos também todas as possíveis proteínas que podem ser codificadas pelo DNA desse organismo e podemos mapear as enzimas para as vias de metabolismo de carboidratos, lipídeos e outras moléculas, por exemplo. Ao sequenciarmos, portanto, o genoma de um organismo, sabemos de forma geral quais são os compostos que esse organismo seria capaz de metabolizar e quais metabólitos secundários seria capaz de usar na fabricação de suas próprias moléculas (Figura 3).

Os seres humanos e os animais, por exemplo, não são capazes de produzir todos os aminoácidos necessários para construir suas proteínas, tendo em vista que não apresentam genes responsáveis por codificarem as enzimas que fundamentais para

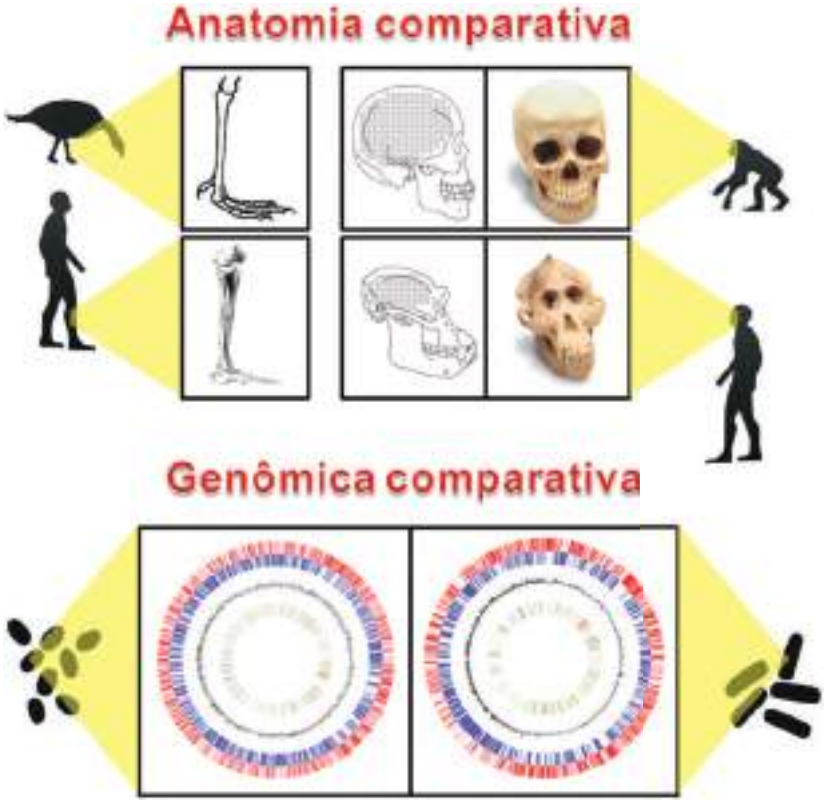


Figura 3. Relação analógica entre anatomia e genômica comparativa. Assim como podemos comparar estrutura de apêndices locomotores ou estruturas e conformações cranianas entre espécies, destacando diferenças e similaridades, podemos comparar composição de genomas (no que tange presença ou ausência de genes, organização de genes, entre outros) e relacionar isso com as características funcionais das espécies.

estas funções. Tais aminoácidos são então considerados essenciais uma vez que enzimas correlacionadas à sua biossíntese não são codificadas pelo genoma humano. Desta forma, precisam ser adquiridos através da alimentação. Neste mesmo sentido, no caso de fungos ou bactérias de difícil cultivo, a compreensão e o conhecimento de seus genomas pode ajudar os pesquisadores a criar um meio de cultura eficiente para que estes organismos possam crescer em laboratório. Entretanto, acreditamos que as verdadeiras questões que levarão à uma melhor compreensão da relação entre genótipo de um organismo (conteúdo de DNA) e seu fenótipo (características enquanto ser vivo) são mais facilmente descobertas quando realizamos comparações entre organismos evolutivamente parecidos, mas que possuem algumas diferenças em questões essenciais relacionadas ao seu metabolismo. Perspectiva esta fundamental para a ciência denominada Genômica Comparativa.

Este tipo de abordagem está ajudando cada vez mais os pesquisadores a elucidarem as adaptações moleculares em genes e proteínas necessárias para que um organismo consiga ou não viver em determinado ambiente ou a metabolizar compostos que

apresentam alguma característica direta ou indiretamente relacionadas à política financeira ou ambiental, síntese de antibióticos ou degradação de xenobióticos respectivamente. Outros exemplos são retratados por organismos que são capazes de sobreviver altas temperaturas, alta salinidade ou mesmo ambientes com escassez de água, diferença de pressão entre outras peculiaridades.

Histórico

Embora se tenha uma tendência em descrever o assunto "genômica comparativa" como se fosse uma vertente de estudos derivado dos sequenciamentos genômicos completos, a comparação entre genomas já vinha sendo feita há muito tempo. Antes mesmo que as técnicas mais modernas de sequenciamento do DNA permitissem aos pesquisadores conhecer o genoma completo dos organismos em pequena escala de tempo (dias ou mesmo horas).

Como vimos no Capítulo 2, o químico inglês **Frederick Sanger** desenvolveu por volta de 1977 a primeira técnica de sequenciamento de moléculas de DNA, o chamado método dideoxi de sequenciamento (SANGER, F *et al*, 1977). Isso permitiu que outros genes pudessem ser conhecidos e a comparação de sequencias tornou-se algo naturalmente inevitável. Desta forma, mesmo que não houvesse conhecimento acerca de toda a sequencia genômica dos organismos a serem comparados, pelo menos a sequência de genes específicos poderia ser estudada. O que, consequentemente levaria a um conhecimento acerca das modificações estruturais de tais proteínas, que poderiam ou não acarretar (decorrente da degeneração do código genético) modificações fisiológicas nos processos aos quais estavam relacionadas.

Tomemos como exemplo a sequencia das cadeias polipeptídicas que compõem a insulina (cadeias A e B), uma das proteínas mais bem estudadas e conhecidas (Figura 4). Embora nem todos os genomas das espécies retratadas na figura estejam finalizados, o estudo individualizado dos genes que codificam paras tais subunidades permite-nos fazer uma comparação ao nível de sequências de aminoácidos. Permitindo, desta forma, identificar resíduos conservados (-) ou variantes em cada posição destas cadeias, mostrando que a evolução das espécies tem direta relação com a evolução de suas biomoléculas.

Diferentes estudos envolvendo a mesma abordagem têm sido feitos desde então, permitindo assim que conheçaos melhor não só a biologia das espécies como também a possível evolução de cada uma delas. Como será discutido no decorrer deste capítulo.

Comparando o conjunto de biomoléculas produzido em uma célula

Uma vez que o sequenciamento de diversos genomas esteja finalizado e que a grande maioria dos genes codificados por eles esteja predita (Capítulo 3), os pesquisadores podem iniciar a comparação dos genomas para estudar diversos níveis de informação. A genômica comparativa pode ser feita em nível de (i) genes, (ii) transcritos de RNA (Capítulo 8), (iii) proteínas (Capítulo 9) ou mesmo (iv) vias bioquímicas e metabólitos (Capítulo 10).

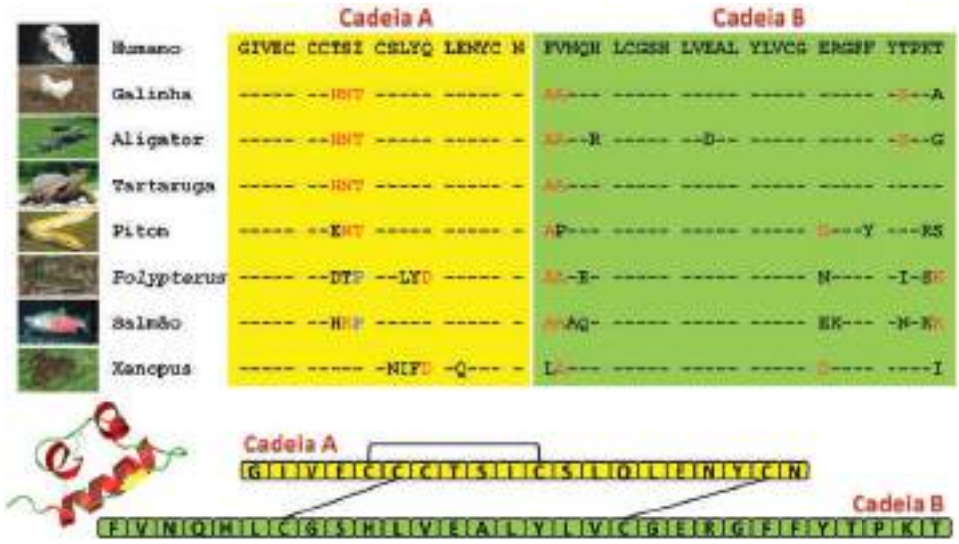


Figura 4. Análise comparativa das estruturas primárias das cadeias que compõem a insulina entre distintos organismos. Observe que mesmo pertencendo a famílias de organismos completamente distintos, grande parte das estruturas primárias é idêntica, sem levar em conta que algumas das modificações tendem a manter o mesmo perfil químico entre os aminoácidos modificados (D↔E; N↔Q; T↔S).

Em nível genômico, as comparações foram inicialmente feitas em escala quantitativa, ou seja, os pesquisadores tentaram observar se o número de genes presentes entre organismos poderia estar relacionado à complexidade dos mesmos. De fato, quando se descobriu que a mosca-das-frutas (*Drosophila* sp) possui aproximadamente 18.000 genes, enquanto o ser humano possui cerca de 20.000 genes, isso foi tido como uma surpresa pelos cientistas, que esperavam explicar a complexidade do ser humano através da presença de um número significativamente maior de genes em nossa espécie (ADAMS, MD *et al*, 2000).

Embora fosse evidente que no início do projeto genoma (meados de 1995), estimava-se que o número de genes humanos codificadores de proteínas estaria em torno de 50-100.000, uma super-estimativa, observou-se então que nós temos apenas 5 vezes mais genes do que uma bactéria de vida livre comum (que possui cerca de 5.000 genes, embora esse número possa variar de espécie para espécie).

Neste mesmo contexto, pesquisadores já haviam calculado previamente se o peso de DNA do núcleo de cada célula seria correspondente à complexidade dos organismos. Uma hipótese óbvia indicava que quanto mais DNA um organismo possui no núcleo, mais complexo ele deveria ser. Entretanto, ao longo da abrangência do conhecimento nesta área pôde-se descobrir que vários protozoários unicelulares, além de salamandras e outros organismos que não eram considerados complexos, possuíam uma grande quantidade de DNA em seus núcleos celulares, contrastando tais correlações. Ficou conhecido então como paradoxo do valor-C (*C-value*) esta inexistência de uma relação direta entre a quantidade de DNA contido no núcleo celular e a complexidade do organismo.

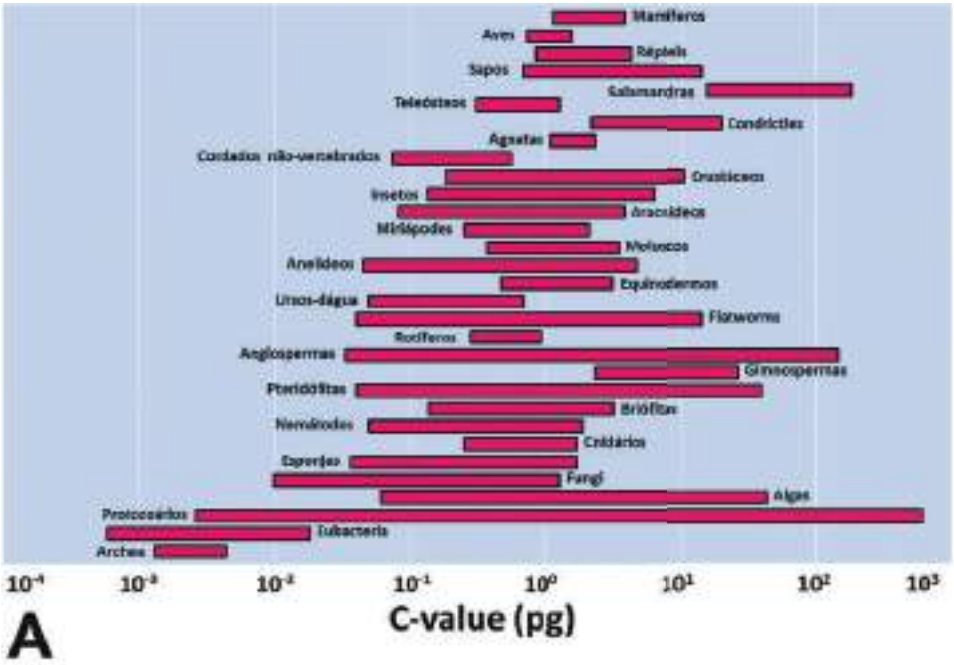
O paradoxo do valor-C foi então atualizado recentemente para a realização da contabilidade do número de genes e a complexidade do organismo, de forma que mais uma vez não se encontrou nenhuma relação direta entre esses fatores (Figura 5) e hoje acredita-se que novas camadas de informação se sobrepõem à quantidade de informação nos genomas dos seres biológicos para gerar a complexidade.

Nos eucariotos que pensávamos serem mais complexos já se observou um maior número e maior complexidade de íntrons, sequências repetitivas, elementos transponíveis e oriundos de transferência lateral (Capítulo 5) ou mesmo de micro-RNAs (MATTICK JS and MAKUNIN IV., 2005) (Capítulo 11), além de formas sofisticadas dos mecanismos epigenéticos de metilação de citosinas e de modificações pós-traducionais em histonas (Capítulo 11) e hoje acredita-se que a complexidade foi produzida principalmente pela associação entre esses fatores (Figura 6 e Tabela 1). Isto sem contar com os fenômenos de splicing-alternativo que podem fazer com que um único gene com vários éxons produza centenas a milhares de proteínas diferentes.

Ainda que o número de genes não seja um fator exatamente adequado para a descoberta de diferenças entre os genomas, muitas vezes é importante realizar a catalogação e a comparação de vários fatores em projetos de genômica comparativa. Podemos calcular, por exemplo, (i) o número de genes por cromossomo entre diversas espécies, (ii) o número médio de éxons ou íntrons de cada gene, (iii) a distribuição de nucleotídeos A, C, T ou G pelos genes, (iv) a utilização de códons preferenciais nos genes codificadores de proteínas, (v) a utilização de pares de nucleotídeos (ou dinucleotídeos) entre os genes, (vi) o compartilhamento de domínios funcionais de proteínas, (vii) a presença de promotores e o tipo de promotores que devem ser "ativados" para a expressão gênica, dentre diversos outros fatores que podem ser descritos e comparados entre diferentes genomas.

A genômica comparativa também pode ser entendida enquanto transcritômica comparativa e os genes oriundos de DNA expresso em RNA mensageiro, ou cDNA (Capítulo 8), podem também ser comparados entre diferentes espécies ou entre diferentes grupos de uma mesma espécie, entre células de uma mesma espécie sujeitas a diferentes tratamentos. A análise comparativa dos genes que estão ativos ou inativos em células normais ou tumorais, de um mesmo tecido, têm ajudado os pesquisadores a compreender melhor os mecanismos de transformação de células normais em células cancerosas. Como o conjunto de genes ativos em um determinado sistema celular muda de forma espaço-temporal, qualquer tipo de modificação em um tipo celular pode ser estudado quando comparado a outro. Assim, a genômica comparativa, baseada nos transcritos totais entre condições celulares contrastantes tem muito a nos ensinar sobre como as células respondem a diferentes estímulos ao ativar e desativar genes. Tudo isso se dá porque sabemos que embora o genoma humano possua cerca de 25-35.000 genes, nem todos esses genes estão ativos num mesmo instante (VENTER, JC *et al*, 2001 and IHGSC *et al*, 2001).

Alguns genes chamados de *house-keeping*, termo inglês que se refere à "arrumação da casa" estão normalmente expressos em todas as células e são responsáveis pelo metabolismo basal celular. Entretanto estima-se que esses genes não correspondam sequer a 10% de todos os genes presentes no genoma e dos outros 90% apenas alguns são ativos em determinadas ocasiões (EISENBERG, E and LEVANON, EY, 2013). De fato, os genes são transcritos em RNA mensageiro apenas quando algum estímulo



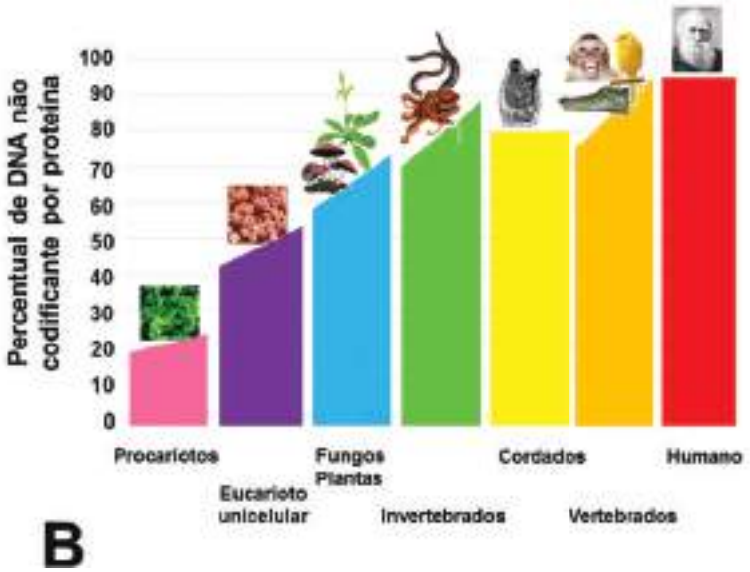


Figura 5. Relação entre o paradoxo C (A) e percentual de DNA não codificador de proteínas (B) em diferentes organismos. Observe que na figura A destacam-se protozoários com grande amplitude de valor-C (~10^{-3} a 10^{3}) em contraposição à amplitude de variação no valor-C de mamíferos (10^{0} a 10^{1}). Curiosamente, na figura B, podemos observar que a proporção de DNA não codificante de proteínas em mamíferos, inclusive humanos, é bem superior se comparado com eucariotos unicelulares ou mesmo procariotos.

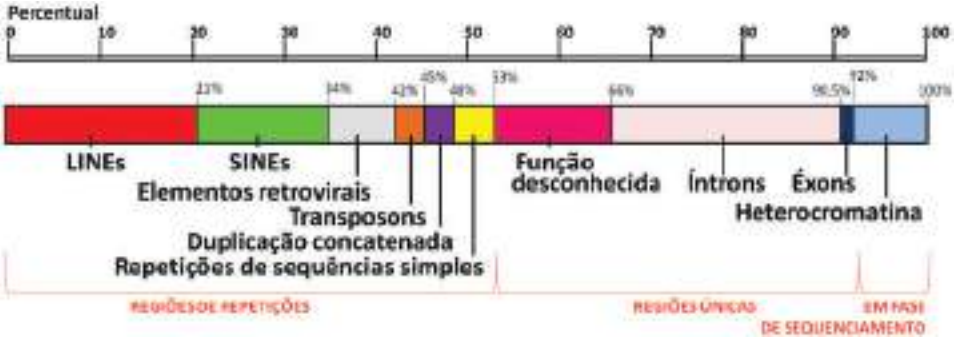
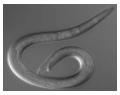
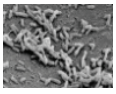
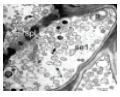
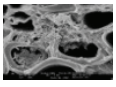


Figura 6. Estrutura composicional do genoma humano. Observe que apenas 1,5% do genoma representa regiões codificantes de proteínas, os éxons, que conjuntamente com as sequencias de íntrons formariam os supostos genes (cerca de 25.000). Ademais, mais da metade do genoma é composta por regiões repetitivas, cerca de 10% representam regiões de heterocromatina, ainda em fase de completo sequenciamento, e uma parcela de cerca de 13% representam regiões cujas funções ainda são desconhecidas.

que chega à célula promova a transcrição de um fator de transcrição ou modifique a repressão a algum gene, que então será transcrito em RNAm e então traduzido em proteína. A regulação da expressão gênica ainda é um dos maiores mistérios da biologia molecular e apenas exemplos bastante simples de regulação são ainda entendidos pelos pesquisadores, como o operon Lac (Capítulo 15) e o operon triptofano. Entretanto, ainda que não entendamos a forma segundo a qual os metabólitos modificam a expressão dos genes, sabemos que esses graus de expressão gênica são modificados e podemos medir quais genes estão ativos, e em que grau, através dos estudos de transcritômica e transcritômica comparativa. A super-expressão ou o silenciamento de determinados genes em determinadas circunstâncias celulares podem permitir o tratamento de doentes e são uma esperança para resolvermos doenças como o câncer. Sabemos hoje, por exemplo, através dos estudos de transcritômica, dentre outros, que o metabolismo energético de uma célula cancerosa é diferente do metabolismo de uma célula normal e dessa forma podemos então testar diferentes drogas que modificam a atuação de enzimas nessa via para ver o que acontece com o metabolismo da célula cancerosa. Já foram encontradas, inclusive, algumas substâncias que atuam nessas enzimas e que são realmente capazes de diminuir a proliferação celular do tumor (Figura 7).

Finalmente a comparação entre o que comumente chamamos de proteoma (Capítulo 8) e que representa o conjunto mais provável de proteínas produzido por um determinado genoma permite aos pesquisadores entender e conhecer quais as possíveis proteínas que um organismo é capaz de produzir. As proteínas, ou em alguns casos complexos proteicos, são as principais máquinas moleculares e são tidas como as biomoléculas mais importantes para a organização do metabolismo celular. São elas que possuem diferentes formas e variedades químicas e que são as enzimas *par excellence*, ou seja, que são capazes de ligar a compostos químicos de estrutura simples e transformá-los em outros que serão utilizados pela célula no ciclo de reações químicas que em que consiste o metabolismo celular (Figura 7).

Tabela 1. Comparações entre tamanho e genes codificadores de proteínas em diferentes genomas eu e procariotos.

Organismo		Número de cromossomos/ plasmídeos	Tamanho do genoma (Mpb)	Número de genes	Referência
Homo sapiens		23/0	3×10^9	~ 25.000	1 (HGPI)
Arabidopsis thaliana		5/0	125×10^6	~ 27.500	2 (TAIR)
Mus musculus		21/0	$2{,}7 \times 10^9$	~35.000	3 (MGR)
Drosophila melanogaster		4/0	139×10^6	~ 15.000	4 (FlyBase)
Caenorhabditis. elegans		6/0	100×10^6	~ 21.000	5 (GGSP)
Saccharomyces cerevisae		16/2	12×10^6	~ 6.600	6 (SGD)
Escherichia coli K12		1/0	$4{,}6 \times 10^6$	~ 4.300	7 (EcGP)
Agrobacterium tumefaciens C58		2/2	$4{,}7 \times 10^6$	~5.500	8 (Agro)
Phytoplasma asteris OY-W		1/0	$0{,}8 \times 10^6$	~750	9 (PRC)
Xylella fastiosa 9a5c		1/2	$2{,}5 \times 10^6$	~ 2.500	10 (XFGP)

1 – HGPI - Human Genome (http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml)
2 – TAIR - The *Arabidopsis* Information Resource (http://www.arabidopsis.org/index.jsp)
3 – MGI – Mouse Genome Research (http://www.ncbi.nlm.nih.gov/genome/guide/mouse/)
4 – FlyBase – A Data Base of *Drosophila* Genes & Genomes (http://flybase.org/)
5 – CGSP - *Caenorhabditis* Genome Sequencing Projects (http://www.sanger.ac.uk/Projects/C_elegans/)
6 – SGD – *Saccharomyces* Genome Database (http://www.yeastgenome.org/)
7 – EcGP – *E.coli* Genome Project (http://www.genome.wisc.edu/)
8 – Agro – *Agrobacterium* public web resource (http://depts.washington.edu/agro/)
9 – PWS – *Phytoplasma* Resource Center (http://plantpathology.ba.ars.usda.gov/phytoplasma.html)
10 – XFGP – *Xylella fastidiosa* genome Project (http://aeg.lbi.ic.unicamp.br/xf/)

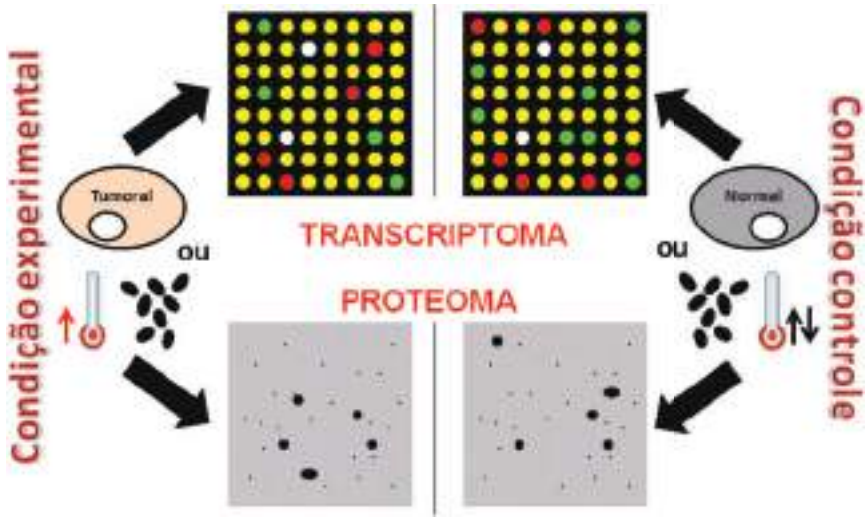


Figura 7. Exemplo de um modelo figurativo relacionando transcritômica e proteômica comparativa a partir de tecidos celulares diferenciados e em distintas condições fisiológicas.

A genômica comparativa de proteínas permite então que os pesquisadores observem quais proteínas estão ou não presentes em um determinado genoma e que, através do agrupamento das proteínas por via bioquímica, possam identificar quais vias um organismo é capaz – ou não – de produzir. Como já explicado anteriormente os humanos não são capazes de sintetizar todos os aminoácidos que precisam para montar suas proteínas através de vias de síntese *de novo*, ou seja, a partir de precursores simples. Isso significa que o ser humano não tem determinadas enzimas que seriam responsáveis por fazer com que intermediários das vias de biossíntese de aminoácidos produzissem mesmo essas moléculas dentro das células. Com relação a fungos e bactérias que podem ser cultivados em laboratório é importante saber se eles possuem proteínas para quebrar determinados tipos de carboidratos ou gorduras em suas células para que se possa produzir um meio de cultura apropriado para eles. E, além disso, pode-se hoje inclusive pensar em construir organismos sintéticos contendo vias bioquímicas inteiras que foram encontradas em outros organismos e adicionadas a um organismo hospedeiro de interesse (Capítulo 15).

Finalmente a genômica comparativa a partir de metabólitos secundários (Capítulo 10) produzidos permite compreender quais vias metabólicas ou subprodutos destas, diferentes organismos conseguem produzir para fins diversificados, quase sempre relacionados à adptação diante de condições fisiológicas as quais foram expostos.

Genes conservados, genes espécie-específicos e herança vertical

Os genes de um genoma que são compartilhados entre outros genomas são chamados de **conservados**. Para isso devemos levar em conta alguns critérios razoavelmente bem

estabelecidos, como (i) a cobertura e (ii) a identidade entre as sequências. Entende-se por cobertura o percentual de comprimento de uma sequência de um organismo que se deseja comparar (query) em relação a outra sequência conhecida (subject). Esta comparação entre tais sequencias, posição a posição (seja uma comparação de nucleotídesos ou aminoácidos) também fornecerá informação quanto ao nível de conservação, aferindo desta forma o que se chama identidade.

Portanto, para ser conservado um gene precisa estar presente ao longo de inúmeros organismos de genomas conhecidos e que sejam distantes filogeneticamente. O gene que codifica as proteínas histonas, por exemplo, que são proteínas responsáveis por ligar o DNA e permitir a compactação do DNA no núcleo celular são altamente conservados em todos os organismos eucarióticos, e estão ausentes apenas nas bactérias. A sequência de nucleotídeos dos genes que codificam a histona são bastante similares entre os organismos. Entretanto o que realmente importa na hora de definirmos se os genes são conservados é a sequência de aminoácidos da proteína codificada pelo gene em questão. Isso se dá pelo fato de que o código genético é degenerado, ou seja, sequências diferentes de DNA podem resultar em sequências parecidas de proteínas. São as sequências parecidas das proteínas codificadas pelo gene que definem se o mesmo é ou não conservado quando essas sequências são comparadas entre diversos organismos. Sabemos que proteínas que possuem a mesma estrutura primária, ou seja, a mesma ordem de aminoácidos encadeados, possuirão assim a mesma estrutura terciária ou 3D em que consiste sua estrutura nativa que dá a ela a capacidade de encaixar nos substratos e catalisar reações químicas que, sem elas, demorariam muito mais tempo para acontecer.

O fato de proteínas de genomas distintos serem consideradas conservadas sugere que elas terão a mesma estrutura nas células dos genomas em questão. E quando falamos em proteínas, a estrutura delas está altamente ligada à função molecular que irão exercer na célula. Essas proteínas conservadas, portanto, terão a mesma estrutura na célula do genoma 1 e do genoma 2 e muito provavelmente funcionarão de forma a quebrar o mesmo substrato e produzir o mesmo produto em ambas as células. De fato, testes *in vitro* demonstram que proteínas conservadas normalmente realizam funções idênticas, mesmo que algumas funcionem melhor em diferentes condições de temperatura ou pH.

Vale notar que essas proteínas conservadas que normalmente realizam a mesma função em organismos, células e genomas diferentes possuem também uma origem comum no passado. Se tivermos como base os mecanismos de evolução darwiniana, saberemos que todos os organismos vivos tiveram um ancestral no passado. Homem e cães tiveram um ancestral vivo em algum momento da história e as proteínas conservadas que existem em nossos genomas são normalmente derivadas de proteínas que já existiam nesse organismo desconhecido que foi nosso ancestral. Estudos realizados mostram que são poucos os genes que surgem "milagrosamente" a partir do nada ao longo da evolução das espécies. Genes normalmente surgem a partir da duplicação de genes antigos. Além disso, o nosso gene da insulina não é exatamente idêntico ao gene que codifica a insulina no cachorro porque nossas linhagens divergiram e nossos genes sofreram mutações diferentes ao longo do processo evolutivo. O mais interessante entretanto não é observar as diferenças e sim

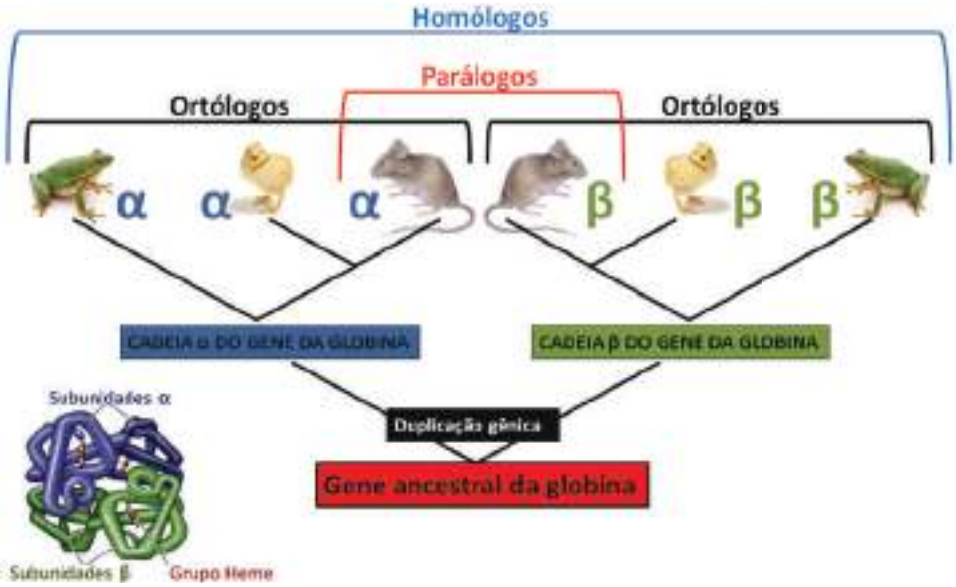


Figura 8. Relação figurativa entre os conceitos de homologia, paralogia e ortologia, baseado na presença dos genes que codificam para as subunidades de globina.

as incríveis semelhanças entre as sequências das proteínas codificadas por genes de espécies que já se distanciaram evolutivamente há milhões de anos.

Tais genes que estão nos genomas dos organismos vivos, que são conservados e que são derivados de um mesmo gene num organismo ancestral são chamados de genes ortólogos (Figura 8). Um gene é ortólogo a outro quando se prova, através de análise filogenética, que eles possuem um gene ancestral no passado que deu origem aos dois genes através de um processo contínuo de especiação. Os genes ortólogos são importantíssimos para a nossa compreensão da genômica e da biologia molecular porque uma vez que sabemos a função de um determinado gene ortólogo, ao testarmos sua atividade em bancada, normalmente os bioinformatas são capazes de inferir, ao menos de forma provisória, a função de outro gene recém-sequenciado e desconhecido através de sua alta similaridade e conservação com esse gene bem conhecido. A anotação de novos genomas é, portanto, realizada de forma a transferir informações conhecidas de genes entre espécies. Hoje em dia, quando uma nova espécie é sequenciada, os pesquisadores normalmente já são capazes de conhecer a função de uma boa parte dos genes desse genoma. Isso porque a maioria dos genes presentes em um genoma consiste em genes conservados. Como falamos anteriormente, os genes responsáveis pela manutenção do metabolismo celular basal (genes *house-keeping*) são altamente conservados e estão presentes nos mais diversos genomas com poucas alterações. Praticamente todos os genes house-keeping de um novo genoma sequenciado podem ser identificados através de comparações de sequências com genes já conhecidos em genomas previamente sequenciados.

Vale notar, entretanto, que para toda espécie sequenciada, normalmente se descobre uma determinada quantidade de genes que não apresentavam nenhuma

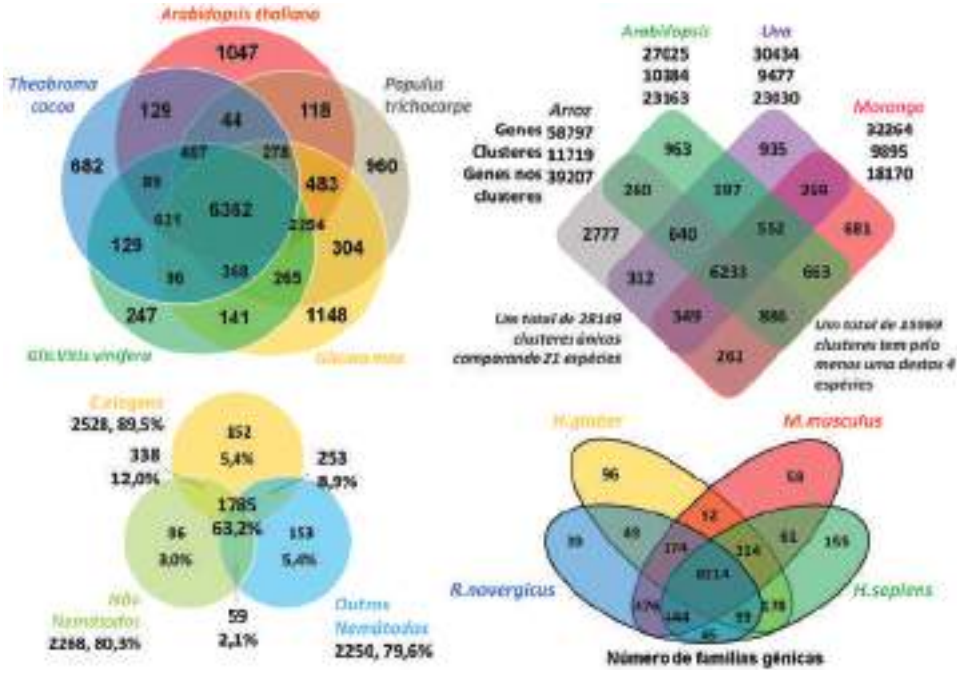


Figura 9. Diferentes exemplos de gráficos de Venn com número de genes conservados e espécies-específicos, encontrados nas espécies comparadas. Observe que o grau de complexidade do diagrama tem direta relação entre o número de genomas comparados (gráficos da esquerda), entretanto modelos distintos são elaborados com o meso número de genomas comparados (gráficos da direita, que por sua vez expressam valores de agrupamentos ou famílias gênicas).

similaridade de sequência com nenhum gene de nenhuma outra espécie conhecida. Esses genes espécie-específicos, quando são pela primeira vez encontrados, não se pode definir ao certo qual seria sua função celular (Figura 9). É interessante notar que, à medida que mais genomas vão sendo descobertos, menor é o número de genes espécie-específicos que se descobre. De forma contrária, quando sequenciamos pela primeira vez um organismo de um filo ou grupo taxonômico que não possuía nenhum organismo sequenciado, muitas vezes encontramos uma quantidade significava de genes totalmente desconhecidos. Esses genes desconhecidos e espécie-específicos podem ser inclusive utilizados como alvo para drogas caso o organismos sequenciado seja um patógeno e o desenho racional de drogas normalmente exige que (i) o gene alvo da droga não esteja presente no genoma do hospedeiro e (ii) que o gene seja essencial ao patógeno. Assim, uma droga direcionada a este gene provavelmente não atacará nenhuma proteína do hospedeiro e atingirá especificamente a proteína do parasita, inibindo-a ou modificando-a de uma maneira que produza a cura do hospedeiro. Quando sequenciamos, portanto, novos genomas, costumamos representar a conservação dos genes da espécie nova que sequenciamos e os genes de outros organismos conhecidos através de gráficos de Venn (em diferentes modelos) que mostram o número de genes na união ou interseção entre os genomas (Figura 9).

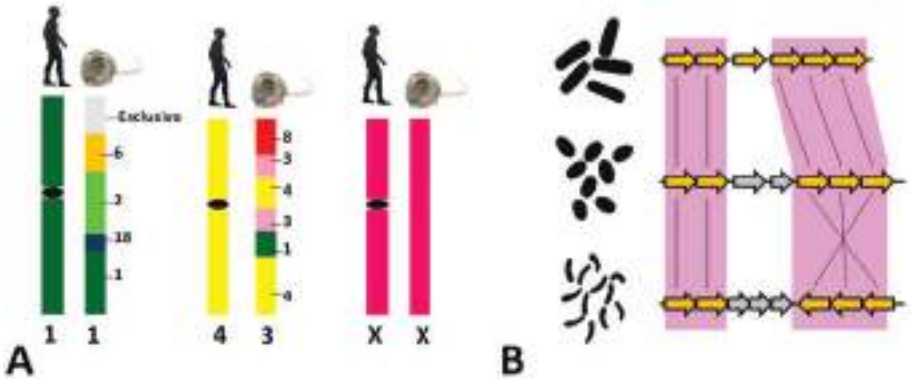


Figura 10. Sintenia e análise comparativa estrutural de cromossomos e genes. (A) Análise comparativa estrutural de cromossomos humanos e de camundongo. Abaixo de cada cromossomo há a respectiva identificação do cromossômico em cada cariótipo. Os cromossomos humanos foram usados como referência de comparação, recebendo uma cor específica para cada um deles: verde, amarelo e rosa (embora sejam apresentados apenas três). Observe que o cromossomo 1 de camundongo representa uma quimera dos cromossomos 1, 2, 6 e 18 de humano, além de possuir uma região exclusiva na ponta do braço p (em cinza). O mesmo pode ser observado entre os cromossomos 4 humano e 3 de camundongo. Curiosamente esta diferença estrutural não é observado entre os cromossomos X de ambas as espécies. (B) Sintenia in loco de um grupo de genes relacionados com uma função específica em três diferentes espécies de bactérias. No primeiro modelo 6 genes que participam de uma mesma função metabólica (representados por uma mesma cor), que por sua vez apresentam-se dispostos de formas diferenciadas em outros genomas. Observe que o terceiro genoma apresentou uma reorganização por inversão de alguns genes.

Sintenia e mudança de organização estrutural de genes

Outro tipo de estudo que é normalmente realizado quando comparamos sequências genômicas de diversos organismos são os chamados estudos de sintenia. Os estudos de sintenia se referem ao entendimento de como a ordem dos genes ao longo dos genomas foi se modificando ao longo do tempo, caso venham a ser alterados. Sabemos que os genes estão dispostos em grandes peças de DNA e proteínas nucleares chamados de cromossomos. E os genes situam-se nos chamados locos cromossômicos, que são os lugares onde os genes "residem". O endereço de um gene em humano pode ser definido como, por exemplo, 5q2.1, onde (q) sinaliza o braço longo do cromossomo 5, porção 2, segmento 1. Cada gene possui um endereço preciso no genoma dos organismos onde eles estão presentes.

Desta forma, quando falamos de estudos de sintenia estamos queremos estudar a comparação entre a ordem dos genes em diferentes organismos. Será que o gene que estão no braço longo do cromossomo 5 de humanos estará no mesmo lugar no genoma do chimpanzé? Dada a enorme plasticidade genômica que temos percebido existir, é difícil comparar a ordem de genes entre genomas eucarióticos que já se divergiram há centenas de milhares de anos no passado (Figura 10A).

Por outro lado, é razoavelmente simples e elegante comparar a diferenciação na ordem dos genes encontrados em diferentes cepas de bactérias ou mesmo espécies ou gêneros bacterianos diferentes (Figura 10B). Muitas vezes podemos observar claramente como um gene "saiu" de um determinado lugar e "entrou" em outro.

Os genomas são muitas vezes quebrados por mutações ambientais e há enzimas e elementos de transposição espalhados pelos genomas das bactérias que podem promover esses saltos genômicos. Também a conservação na ordem dos genes entre diferentes bactérias pode funcionar como marcador para o tempo de divergência entre diferentes variedades ou espécies. Quando mais próxima a ordem dos genes em determinados organismos, menos tempo os pesquisadores inferem que eles divergiram no passado.

Genoma mínimo

Com base no perfil e nas justificativas acerca da importância de se comparar genomas, do número de genes em um genoma e do número de genes compartilhados entre os mais diferentes genomas, uma pergunta interessante se faz necessária. Qual seria o tamanho mínimo possível para um genoma? Ou em outras palavras, qual seria o menor número de genes absolutamente essenciais que sejam capazes de promover a manutenção de uma espécie de vida livre? Embora muitos estudos comparativos de genomas tenham sido apresentados, em especial àqueles que envolvem microorganismos, não há um consenso exatamente preciso acerca deste conjunto de genes. De qualquer forma, sabe-se que determinados genes primordiais importantes para o metabolismo básico de açúcares, proteínas, ácidos nucléicos e lipídeos são sempre necessários.

Utilizando a lógica de que genomas menores em tamanho codificam para um menor número de genes, tendo em vista que em média há cerca de um gene por kilobase (1000 bases) de um DNA procariótico, organismos com menores estruturas cromossômicas tenderão a apresentar menor número de sequências codificadoras. A partir de observações como estas que os endossimbiontes unicelulares obrigatórios passaram a ganhar notoriedade, e organismos como Ricketisias e Mycoplasmas passaram a ser sequenciados.

O verdadeiro genoma mínimo da forma segundo a qual consideramos hoje vem a ser bem próximo do que vemos nos genomas de *Mycoplasma genitalium* ou *Ca. Phytoplasma asteris* da classe *Mollicutes*, que não padem ser caracterizados respectivamente como um organismo de vida-livre e outro um endossimbionte obrigatório, capazes de realizar todas as funções metabólicas normais de um organismo vivo. Estas espécies apresentam genomas com cerca de 600 Kbp responsáveis pela codificação de não mais que 500 genes codificadores de proteínas, tendo uma média de tamanho gênico um pouco maior do que o normal (cerca de 1 Kbp) e apresentando quase a totalidade de seu genoma constituída por genes constitutivos (Figura 11) (Oshima, K et al, 2004).

Genes hipotéticos e conservados hipotéticos

Os primeiros genomas sequenciados apresentaram dados curiosos acerca da presença de supostos genes sem função determinada, os chamados genes hipotéticos. Estes já representaram cerca de 50% de todos os genes preditos nos primeiros genomas sequenciados e hoje estão em números cada vez menores, dado o ganho de conhecimento biológico acerca destas sequências.

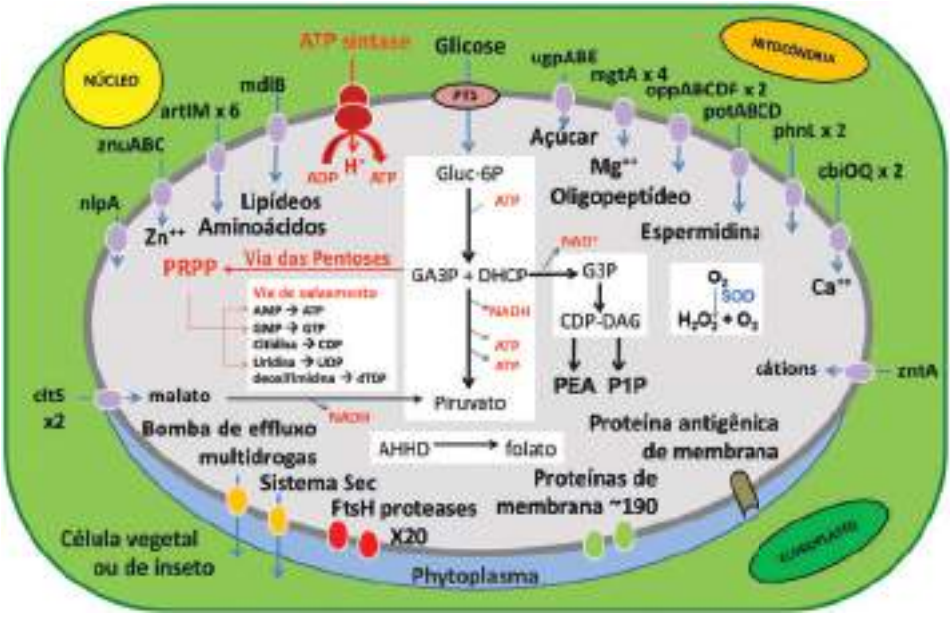


Figura 11. Resumo figurativo das principais atividades metabólicas de Phytoplasma usando como base as informações genômicas. Observe que um considerável número de genes codifica para receptores e transportadores de membrana, resposnáveis pela internalização de moléculas essenciais à sua sobrevivência e cuja síntese é impossibilitada pela ausência de genes específicos. O caso mais curioso retrata a presença de um complexo ATPsintase que utiliza o prótons do interior da célula vegetal para promover a síntese de ATP no interior da bactéria.

Além disso, quando um suposto gene hipotético passa a ser encontrado com alto nível de identidade e similaridade em outros genomas, então a classificação de gene hipotético muda para conservado hipotético. Embora continuemos a não saber a função específica destes genes, agora conservados em outras espécies, o dado reforça a possibilidade de ser uma sequencia que efetivamente codifique uma proteína, tendo em vista que outros organismos também o apresentam em seus genomas. Estes genes conservados podem retratar um ancestral comum, ou mesmo eventos derivados de herança gênica lateral.

De qualquer forma, nestas circunstâncias uma pergunta permanece e torna a discussão mais contundente. Como estes genes foram anotados com esta função se não havia qualquer outra informação sobre o mesmo em bancos de dados especializados para uma inferência comparativa? E a resposta para esta pergunta reflete a capacidade de usarmos ferramentas computacionais para predizermos a possibilidade de uma determinada sequencia de DNA ser ou não codificadora de uma proteína, portanto, um gene (Capítulo 3).

Usando ferramentas computacionais específicas (ORFFinder, Glimmer, etc), os biólogos do computador ou bioinformatas é que são os responsáveis por esta empreitada. Em procariotos, onde os genes não possuem íntrons, são estes programas que apresentam a função primordial de encontrarem supostas sequencias de códons de parada numa determinada sequencia de DNA. Encontrada uma região entre dois

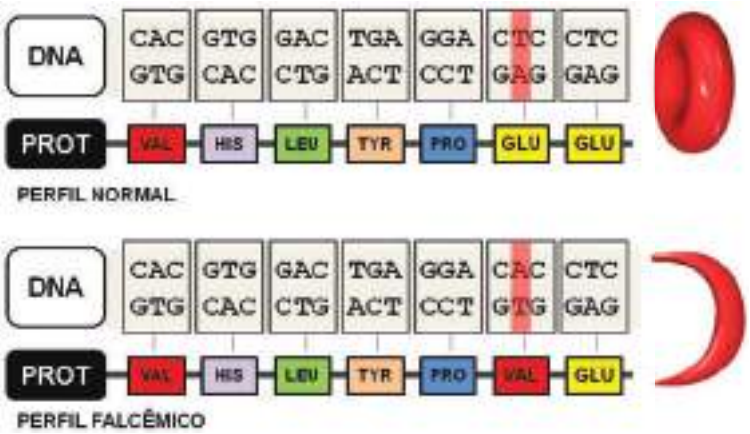


Figura 12. SNP que proporciona a modificação importante na morfologia de hemácias. A troca de um A por um T no gene α-globina acarreta a troca de aminoácido (ácido glutâmico por valina) que por consequência acarreta a alteração fisiológica da proteína levando a hemácia a apresentar morfologia de foice, daí o nome da doença, anemia falciforme.

supostos códons de parada, o próximo objetivo é determinar um suposto códon de iniciação, quase sempre um AUG que codifica para o aminoácido metionina. A partir de alguns critérios, tais como tamanho da suposta região codificadora (CDS) ou canal aberto de leitura (ORF), ou possível sobreposição com outros genes localizados em mesma região genômica, estes genes são anotados como simplesmente hipotéticos ou hipotéticos conservados, se tiverem similaridades com genes de outros organismos, e passam a ser caracterizados como um suposto gene até ser funcionalmente validado ou refutado.

SNPs

Quando comparamos organismos da mesma espécie também podemos encontrar, obviamente, enormes diferentes. Você que lê este livro e nós, que o escrevemos, somos pessoas diferentes, reagimos diferentemente a modificações no meio ambiente e temos informações diferentes armazenadas em nosso DNA. De fato, nós somos também diferentes de nós mesmos, uma vez que temos dois genomas completos em cada uma de nossas células, um deles proveniente do espermatozóide de nosso pai e outro proveniente do óvulo de nossa mãe. Podemos ser homozigotos ou heterozigotos para todos os genes.

No caso de bactérias, por exemplo, que são organismos com apenas uma cópia de seus genomas, podemos comparar sequências completas de seus genomas e identificarmos algumas substituições de nucleotídeos que podem ser catalogadas e que caracterizam a diversidade molecular de uma determinada espécie, gênero, família ou clado qualquer de organismos.

Em espécies altamente diversas como a nossa, há determinadas variações entre os genomas que são conhecidas e catalogadas. Por exemplo, em um determinado gene alguém pode ter um nucleotídeo C em uma posição e outra pessoa apresentar um

nucleotídeo T nesta mesma posição de um gene. Se essa diferença for polimórfica e isso significa que essa variação deve estar presente em pelo menos 1% dos indivíduos do grupo que se analisa, os biólogos resolveram chamar essas mudanças de SNPs. Na sigla em inglês, o SNP significa *Single Nucleotide Polymorphism* ou polimorfismo de um único nucleotídeo. Tais polimorfirmos são importantes de ser catalogados porque muitos deles podem causar ou predispor a doenças e é interessante que o indivíduo tenha um diagnóstico cada vez mais precoce das doenças que ele possa vir a ter. A doença conhecida como Anemia falciforme consiste exatamente em uma mudança de bases, de um A para T no gene da cadeia beta da hemoglobina que causa uma mutação do aminoácido ácido glutâmico para valina na sexta posição desta cadeia (Figura 12). Essa mutação faz com que a hemoglobina do anêmico tenha formas distorcidas e não carreguem eficientemente o oxigênio através do sangue.

A grande parte dos polimorfismos conhecidos em humanos, entretanto, não parece ter nenhuma função específica no fenótipo dos organismos, o que vai de acordo com a chamada teoria neutra da evolução molecular, que diz que a maioria das bases no genoma não sofre ação da seleção natural.

Vale notar ainda que as mutações de nucleotídeos únicos ou SNPs podem ser de dois tipos: aquelas que causam mudança no aminoácido codificado pela proteína e as que não causam. Uma vez que o código genético é degenerado (61 trincas de nucleotídeos - códons - para codificar 20 aminoácidos), é de se esperar que algumas mudanças no DNA não ocasionem nenhuma mudança na proteína codificada, especialmente as que ocorrem na terceira posição do códon. A comparação entre a quantidade de mutações sinonímias em um gene, essas que não mudam os aminoácidos codificados, e o número de mutações não-sinonímias, essas que mudam o aminoácido, são importantes marcadores para medir como um gene vem evoluindo entre diferentes espécies. Quando há um número excessivo de mutações não-sinonímias e quando elas excedem de maneira marcante o número de mutações sinonímias, podemos considerar que a seleção natural está de alguma forma agindo para que o gene venha a produzir uma proteína diferente, na chamada seleção molecular positiva.

Bibliografias

ADAMS, M.D., CELNIKER, S.E., HOLT, R.A., et al. The genome sequence of Drosophila melanogaster. Science. 2000, 287(5461):2185-95.

EISENBERG, E. and LEVANON, EY. Human housekeeping genes, revisited. Trends in Genetics, 2013, 29(10), 569-574.

INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM, LANDER, E.S., LAUREN M. et al. Initial sequencing and analysis of the human genome. Nature.2000, 409, 860-921.

MATTICK, J.S. and MAKUNIN, I.V. Small regulatory RNAs in mammals. Hum Mol Genet. 2005.14(1):R121-32.

OSHIMA, K., KAKIZAWA, S., NISHIGAWA, H., JUNG, H.Y., WEI, W., SUZUKI, S., ARASHIDA, R., NAKATA, D. et al.. Reductive evolution suggested from the complete genome sequence of a plant-pathogenic phytoplasma. .Nature Genetics. 2004, 36(1): 27–29.

SANGER, F., NICKLEN, S. and COULSON, A.R. DNA sequencing with chain-terminating inhibitors. PNAS. 1977, 74(12):5463-7.

VENTER, J.C., ADAMS, M.D., MYERS, E.W. et al. The Sequence of the Human Genome. Science. 291(5507): 1304-1351.

5

Plasticidade e fluxo genômico

Leandro Marcio Moreira
Alessandro de Mello Varani

Introdução

No capítulo anterior foram descritos os principais objetivos e aspectos que devem ser considerados e empregados em análises que envolvem princípios de genômica comparativa. No entanto, em nenhum momento foi abordado o papel biológico que a reorganização da estrutura genômica pode desempenhar, assim como o possível impacto que essa reorganização pode ocasionar na evolução de todas as formas de vida presentes em nosso planeta. O processo de plasticidade e fluxo genômico é um tema bastante estudado nas últimas décadas, principalmente a partir da introdução da técnica de sequenciamento de Sanger (Sanger e Coulson, 1975) e mais recentemente com as Novas Tecnologias de Sequenciamento (NGS, do inglês *Next Generation Sequencing*) (Mardis, 2008).

O sequenciamento genômico permitiu o estudo detalhado da composição dos cromossomos, genes e potêncial codificante dos seres vivos. Estes estudos levaram a resultados fascinantes e intrigantes. Atualmente sabe-se que a plasticidade e fluxo genômico, mesmo entre organismos da mesma espécie é muito maior do que se pensava em décadas atrás. Hoje sabemos que a constituição de genes pode ser até 30% diferente entre organismos da mesma espécie, sendo que a organização em que estes genes estão distribuídos ao longo dos cromossomos pode ser bastante variável (Fraser-Liggett, 2005). Os geneticistas e biólogos moleculares descobriram que grande parte desta diversidade genética pode ser ocasionada por uma variedade de mecanismos e fenômenos celulares. Dentre estes mecanismos, a transferência gênica lateral (TGL, ou LGT, do inglês *lateral gene transfer*) realizada pelos elementos genéticos móveis (EGMs, ou MGEs, do inglês *mobile genetic elements*), é atualmente considerada como um dos principais mecanismos relacionados à plasticidade e fluxo genômico observados nas mais variadas formas de vidas estudadas.

Os Elementos Genéticos Móveis (EGMs) estão presentes em grande número em todas (se não em quase todas) as formas de vidas conhecidas de nosso planeta. Também são considerados como um dos principais agentes responsáveis pela diversidade e variação genética, podendo inclusive contribuir como mediadores da adaptação a novos nichos. Já, a transferência gênica lateral é o nome que designa processos envolvendo trocas de DNA entre organismos, em contraposição à herança clonal ou vertical, onde os genes são passados entre gerações por processos de simples divisão celular (Eisen, 2000). Portanto, o estudo do fluxo da informação gênica ocasionado pelos EGMs pode permitir que possamos entender a diversidade presente em nossa bioesfera, e também permitir uma melhor compressão das chamadas heranças verticais ou horizontais de genes.

Neste capítulo serão discutidos todos esses aspectos apresentados acima, com ênfase na importância que esses EGMs desempenham na biologia e evolução dos seres vivos, e em como os eventos de recombinação gênica podem contribuir como agentes da diversidade, plasticidade e fluxo genômico.

Plasticidade genômica

O termo plasticidade tem sido usado sobre diversos contextos. Entretanto, o termo deriva das Ciências Exatas e, explicitamente no campo da Física a plasticidade pode ser definida como o ramo do conhecimento que visa estudar o comportamento de corpos que mudam sua conformação (deformam) quando submetidos à ações externas (forças), impedindo que retornem ao estado inicial. É evidente que a intensidade destas ações e a susceptibilidade do corpo a sofrer deformações interferem no efeito final. Portanto, qualquer deformação que seja observada num dado objeto tem relação direta com sua plasticidade. É importante destacar que plasticidade e elasticidade, embora sejam processos que possam ser confundidos num momento inicial, são bem distintos. A elasticidade possibilita que este corpo previamente deformação retorne a seu estado original, contrapondo ao efeito da plasticidade.

Embora esta terminologia "plasticidade" em Ciências da Vida possa parecer algo novo, a mesma vem sendo usada com frequência em algumas áreas do conhecimento, além da Genômica. Talvez um dos usos mais interessantes do termo plasticidade tenha relação com a Neurologia e ciências correlacionadas, Psiquiatria e Psicologia. Neste contexto sabe-se que o cérebro apresenta-se em constantes mudanças e as diferenciações estruturais ou comportamentais de suas unidades celulares (neurônios) podem ser decorrentes de inúmeros fatores. Um destes fatores, talvez o mais conhecido e estudado é o próprio desenvolvimento do cérebro, contextualizado sob dois aspectos: i) aumento do número de células (que ocorre em paralelo ao crescimento do indivíduo), e a própria ii) plasticidade sináptica, ou seja, as alterações da capacidade de comunicação dos neurônios entre si formando redes e ramificações dinâmicas. Esta plasticidade sináptica desencadeia papel fundamental em eventuais lesões ou durante aprendizagem/cognição. Dois exemplos interessantes que esboçam tal efeito são retratados abaixo.

Você já deve ter ouvido falar de pacientes que sofreram danos cerebrais e que vieram a recuperar, posteriormente, parte das atividades perdidas em decorrência do trauma a qual foram acometidos. Deve também ter ouvido falar que uma forma

de retardar possíveis sintomas induzidos pelo mal de Alzheimer é exigir do cérebro estímulos contínuos, ou seja, força-lo a sair da rotina colocando os cinco sentidos a se depararem com situações novas. Pois bem, em ambos os casos, sob o ponto de vista celular/molecular, estamos nos referenciando a plasticidade sináptica acima mencionada. Note o quão importante é esta adaptação do cérebro a novas condições. Em um contexto parecido, a plasticidade genômica ganha fundamental importância biológica.

Com base nesta contextualização do termo plasticidade podemos dizer precisamente que o genoma tem característica plástica, e isto é fundamental em aspectos evolutivos. Esta plasticidade será abordada daqui por diante sobre três aspectos: reorganização, inserção e deleção de sequencias genômicas, não necessariamente codificadoras de proteínas. Vale ressaltar que estes aspectos podem ser vistos complementando um ao outro em eventos biológicos, como por exemplo: a inserção de uma sequência gênica de um bacteriófago temperado durante o processo de infecção (ciclo lisogênico) poderia, por consequência, promover a reorganização de um dado "gene alvo" desta inserção presente no genoma do hospedeiro e que, por consequência, poderia induzir a perda de sua função, ou em alguns casos, quando a inserção ocorrer nas proximidades da região promotora deste "gene alvo", poderia diminuir ou aumentar a transcrição do "gene alvo" associado a este promotor. Somente com esses exemplos citados acima poderíamos imaginar uma série de consequências para a célula do organismo alvo, dentre as principais a alteração do fenótipo, seja para adaptação a uma nova condição, ou em casos mais severos para aniquilação do organismo alvo. Portanto, em uma primeira análise, a plasticidade observada nos genomas dos mais diversos organismos é um por sí só um claro indicativo da ocorrência do processo de seleção natural.

Reorganização da estrutura cromossômica

Embora deva ser considerado fundamental o estudo envolvendo a mudança no número de cromossomos de uma espécie, as chamadas alterações numéricas cromossomais (euploidias e aneuploidias), limitaremos nossas discussões aqui a modificações estruturais cromossômicas, a que se referem a organização em que os genes estão dispostos em cada cromossomo.

Existem naturalmente rearranjos cromossômicos decorrentes da própria maquinaria celular, como é o caso dos eventos de *crossing over* observado em células em divisão meiótica, durante a fase de Prófase I. Entretanto, é comum que ocorram as chamadas quebras e ligações entre frações do genoma (cromossomo), e este processo pode ocorrer espontaneamente ou se dar por ação de estresse, como se observa em células que são submetidas por longos períodos à radiação ionizante. Este evento é conhecido como *breakage and rejoining* (Figura 1A). Há também eventos de recombinação que pode ser mediados por sequencias repetitivas no genoma que vão ocasionar um evento denominando de recombinação homóloga, formando um estrutura denominada de Junção Holliday , que é composta pelas duas cadeias de DNA (a doadora e receptora) (Figura 1A),e por último pela presença e ação de EGMs, ver adiante (Figura 1b).

Durante estes eventos, pedaços da estrutura cromossômica são trocados e embaralhados, podendo provocar alterações na composição e principalmenteda ordem

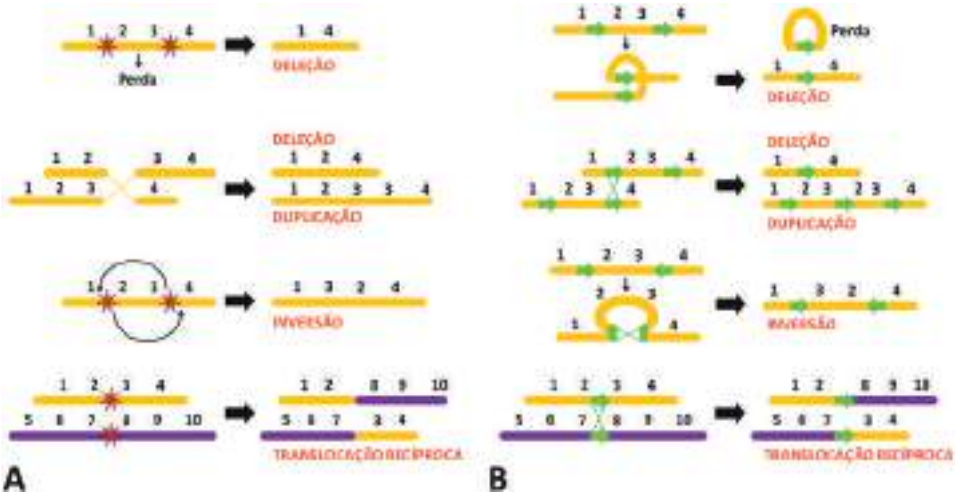


Figura 1. Eventos de plasticidade genômica. (A) Eventos ligados a *breakage and rejoining*. A estrela vermelha denota quebra do genoma, seguido de possível evento de religação destes fragmentos (translocação ou inversão) ou perda (deleção). Use os números como genes presentes nos cromossoos para facilitar a compreensão. (B) Eventos ligados a elementos repetitivos do genoma (Recombinação Homóloga). As setas verdes denotam sequencias repetitivas e, o sentido, a recíproca direção no genoma. Observe que os eventos apresentandos em A e B geram resultados similares, mas o fator indutor do processo é diferenciado.

gênica, tanto no local (*loci*) de origem do fragmento como no novo local onde o mesmo se integrou (*target loci*). Podem ocorrer também, e é muito comum, deleções, inserções ou até mesmo inversões de fragmentos cromossômicos alterando a fisiologia deste genoma e que em última instância também poderá alterar o fenótipo do organismo/célula em que está presente. Neste sentido devemos compreender melhor as chamadas alterações estruturais balanceadas e não balanceadas que podem ocorrer nos tecidos germinativos ou somáticos.

Quando balanceada estas modificações estruturais não induzem qualquer perda de função gênica e, portanto, o fenótipo do indivíduo/célula permanece inalterado. Entretanto, quando não balanceada os danos podem ser severos, podendo levar a célula/indivíduo à morte. Isto vai depender do tipo e/ou quantidade de genes perdidos durante estes eventos. Muitas vezes, a perda de um único gene leva o indivíduo a morte, tamanha a importância deste para a sobrevivência da espécie, neste caso denominamos a mutação de letal. Estas alterações estruturais balanceadas e não balanceadas são corriqueiramente observadas em populações de bactérias, onde uma linhagem bacteriana pode adquirir ou perder resistência a um determinado antibiótico através de eventos de recombinação.

É importante destacar que embora estes rearranjos possam parecer prejudiciais ao genoma, são de fato considerados parte de um processo fundamental relacionado com evolução molecular dos genes de cada indivíduo e com a geração da diversidade de todas as formas de vida de nosso planeta. Ou seja, estas modificações acontecem com freqüência, porém, caso ocorra fixação destas modificações (novos genes) numa determinada geração, as próximas gerações poderão tanto se beneficiar, como sofrer

a conseqüência deste processo. Em via de regra, o estudo desta imensa diversidade ocasionada pelos eventos de recombinação está fornecendo contribuições fascinantes e importantes para a nossa compreensão do mundo biológico.

Duplicação, inversão, deleção e translocação

Embora pareçam pouco representativos, os eventos de duplicação, inversão, deleção e translocação podem ser bastante frequentes dependendo da espécie e das condições ambientais as quais as mesmas foram expostas, a chamada pressão seletiva.

Eventos de duplicação gênica quase sempre levam uma das cópias duplicadas a se especializar com o tempo passando então a adquirir uma nova função, não necessariamente correlacionada à função anterior. Talvez o melhor exemplo seja o dos genes da globina. Embora Mioglobina e Hemoglobina possam ligar oxigênio molecular, suas composições estruturais (mioglobina como subunidade única e hemoglobina como um tetra-heterodímero), bem como localizações intracelulares e afinidade pelo composto possibilita que desenvolvam funções primordiais, porém bastante diferenciadas entre si. Hemoglobina apresenta a função de tranporte de O_2 pelo sangue, ao passo que a mioglobina apresenta a função de fixação do O_2 nos tecidos.

Muitas vezes estas duplicações sofrem alguma seleção e modificação, porém a função básica a ser desenvolvida permanece intacta. Isso é bastante observado em proteínas que atuam como receptores de membrana. A fração citoplasmática destas proteínas quase sempre apresenta os mesmos domínios de sinalização. Já a fração periplasmática (no caso das bactérias), ou externa no caso de eucariotos, apresentam composições estruturais diferenciadas para se adequarem perfeitamente a diferentes compostos presentes no meio. Um exemplo clássico está relacionado com transdução de sinal mediada por proteínas denominadas MCP (*Methyl-accepting chemotaxis protein*) em bactérias, envolvidas com resposta a eventos quimiotáticos. A duplicação pode gerar o que chamamos de genes parálogos, que são genes duplicados em um mesmo genoma e que apresentam uma mesma função. Quando este evento de duplicação gênica é observado entre organismos distintos e/ou de outras espécies, esses genes serão denominados genes ortólogos.

Inserções e deleções são mais fáceis de serem compreendidas quando associadas a elementos genéticos móveis, portanto serão explicados a seguir.

Elementos genéticos móveis

Um pouco das características dos elementos genéticos móveis já foi descrito no capítulo VII, quando foi descrito o uso destes elementos genéticos como ferramentas de indução de mutações para que fossem verificadas e estudadas as funções de genes alvos.

Neste capítulo iremos retratar apenas os eventos naturais onde os EGM estão contextualizados. Consideram-se elementos genéticos móveis os plasmídeos, vírus e bacteriófagos, transposons e elementos integrativos (Figura 2). Estes elementos integrativos são caracterizados como um DNA circular, similar aos plasmídeos, porém com sítios de reconhecimento (*att* na figura) que permitem sua integração no genoma. Quase sempre estes elementos integrativos e transposons, assim como alguns tipos de plasmídeos, carregam consigo cópias de genes que acabam por conferir alguma

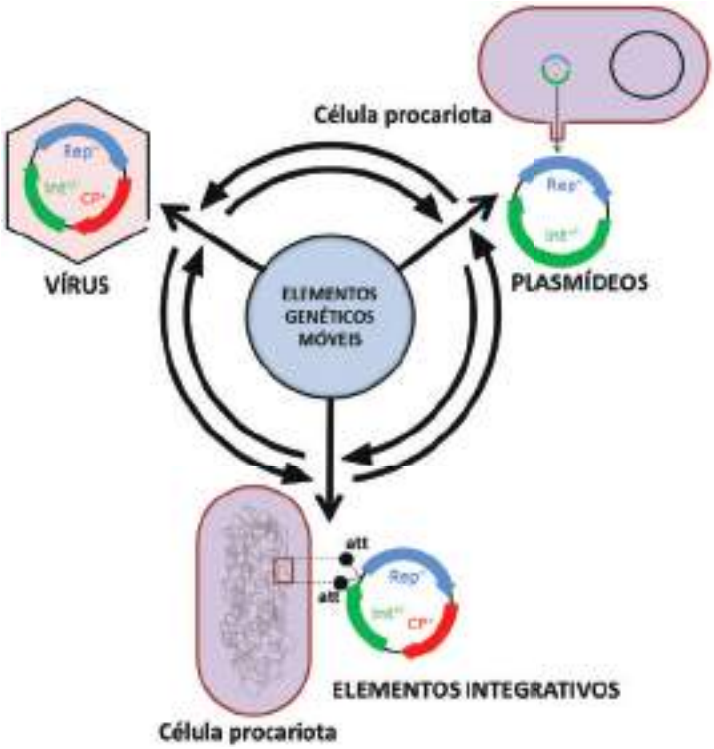


Figura 2. Classificação dos elementos genéticos móveis conhecidos. Note que a troca de informação genética pode ocorrer também entre os tipos de elementos genéticos móveis, e não apenas destes para com as células que os receberão. Adaptado de Krupovic *et al*, 2011.

característica adicional no genoma receptor, a exemplo das chamadas resistências a antibióticos, resistências a metais pesados, ou mesmo capacidade de infectar novos hospedeiros. De fato esse mecanismos já é conhecido a algumas décadas. Em 1975, Falkow apontou evidências de processos de transferência de genes, onde descreveu como a resistência à antibióticos em bactérias podia ser intermediada por genes presentes em plasmídeos capazes de realizar sua replicação e mobilização autônoma no cromossomo hospedeiro (Dröge et al., 1998).

O ganho mediado por estes elementos fica evidente cada vez que estes adentram a célula que irá hospedá-los, permitindo que estes possam ou não se integrar ao cromossoma principal. Entretanto, quando ocorre integração, este ganho de informação genética pode promover perda de alguns genes em decorrência da ruptura da unidade cromossômica principal para receber este DNA exógeno. É evidente que se a integração ocorrer num gene fundamental para o metabolismo celular, podemos dizer que ocorreu uma integração suicida, já que este evento não será perpetuado na espécie, em decorrência da morte iminente da célula receptora deste DNA exógeno. Logo, podemos concluir que todas as inserções sabidamente conhecidas hoje em genomas não foram letais, e foi graças a isso que esta dinâmica de troca de informação genética entre as espécies pode ser conhecida.

Como o próprio nome já diz, por serem móveis e uma vez integrados no genoma, há a possibilidade de saírem. Ao saírem, podem levar consigo frações do DNA da célula hospedeira, promovendo, com isso, mais deleções de genes que, geralmente, estão no entorno desta região. Um exemplo clássico deste tipo de evento é o mecanismo de transdução generalizada e especializada ocasionada pela infecção de bacteriófagos temperados em bactérias susceptíveis. Um fago lisogênico, quando induzido ao ciclo lítico, no momento da excisão de seu genoma que está integrado ao genoma da bactéria hospedeira, pode "arrancar" e carregar genes adjacentes ao sítio de integração no hospedeiro devido a uma excisão imperfeita. Neste caso, uma versão modificada do genoma do fago original será formada e encapsulada na nova partícula viral, caracterizando a transdução especializada. Na transdução generalizada, segmentos do própriocromossomo bacteriano distantes do sítio de integração do profago podem ser capturados no momento damontagem do capsídeo viral e empacotados no novo bacteriófago formado. Este fenômeno/conhecimento biológico foi usado para desenvolver uma técnica clássica em biologia molecular cujo objetivo é induzir a expressão de proteínas em capsídeo viral, denominada *Phage Display* (Smith, 1985).

Se ampliarmos esta discussão levando em conta que este cassete de genes que acabou de sair de um genoma possa adentrar outro, então esta nova célula receptora receberá fragmentos de DNA que fazem parte da própria estrutura do elemento genético móvel, mas também de ou outro genoma que fora previamente um hospedeiro desta sequencia exógena. Vamos tornar este contexto biologicamente mais interessante? (Figura 3).

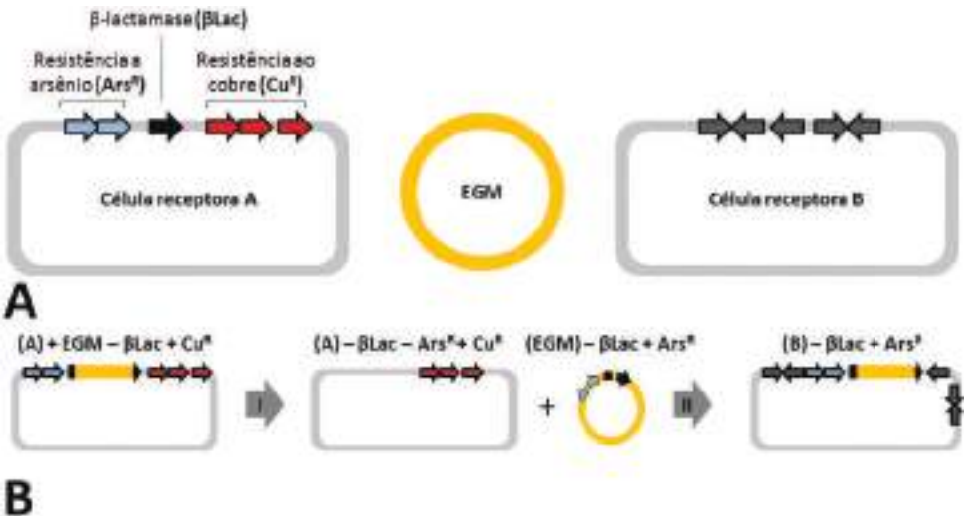


Figura 3. Modelo hipotético de ganhos e perdas gênicas mediado por EGM. (A) Composição gênica dos genomas das células hospederias A e B e do próprio EGM. (B) Perspectiva da dinâmica de integração e saída do EGM dos genomas hospedeiros. Observe que o modelo já tem início demonstrando uma prévia inserção do EGM no genoma de A, promovendo com isso perda do gene β-lac. Num evento posterior, denominado I, o EGM sai deste genoma hospedeiro, carreando consigo β-lac fragmentado bem como os genes de ArsR. Esta nova composição do EGM pode agora integrar na célula B conferindo a esta uma ArsR.

Vamos pensar que três estruturas serão à base da discussão abaixo: uma célula hospedeira nomeada A, um EGM integrativo e uma célula hospedeira nomeada B (Figura V-3A). A célula A apresenta um conjunto de genes que lhe confere resistência a arsênio (ArsR), um gene que codifica para uma beta lactamase, β-lac (resistência a antibióticos do tipo β-lactâmicos, exemplo penicilina) e um conjunto de genes que lhe confere resistência ao cobre (CuR). Vamos imaginar que esta célula sofra um dano gerado pela integração do EGM, justamente na região que codifica para β-lac. Como resultado, teremos um genoma contendo um EGM, porém com perda da função de β-lac. Como a perda deste gene não é letal, com o tempo pode ser que este EGM sai do genoma, levando consigo parte do DNA desta célula A. No nosso exemplo, ao sair o EGM carreou consigo o gene β-lac e os genes relacionados com ArsR. Caso este mesmo EGM venha adentrar e integrar uma nova célula (Hospedeira B), esta nova célula terá em seu genoma um gene de β-lac inativado, porém genes que conferem resistência ao arsênio (ArsR) íntegros, podendo inclusive serem funcionais.

Este é só um exemplo da dinâmica de integração de EGM podendo promover tanto ganhos quanto perdas de informações genômicas. Uma boa tarefa é pensar em outras possibilidades que permitam organismos a terem múltiplas resistências, bem como múltiplos ganhos e perdas de funções biológicas em detrimento de ganhos e perdas de genes em seus genomas.

Alguns estudos científicos corroboram a possibilidade do contexto apresentando acima. A variação do tamanho do genoma de diferentes linhagens da bactéria *Escherichia coli* é algo entre 4,6 – 5,5 Mpb. De fato um tamanho considerável, e que, com certeza, codificará para características fundamentais para manutenção da vida de cada linhagem. Neste mesmo estudo observaram em detalhes essas diferenças, e notaram que grande parte destavariação estava relacionada a integração de EGMs. Estes EGMs também foram responsáveis pela aquisição de fatores de virulência pela bactéria, em uma maneira similar ao processo proposto acima (Schneider et al., 2002).

Fluxo do/no genoma

Entende-se por fluxo no genoma o perfil de herança aos quais seus genes estão relacionados. Neste sentido podemos classificar os mecanismos de herança como sendo verticais (originados de um ancestral) ou laterais (quando originados de na espécie não ancestral).

Tomemos como exemplo a Figura 4. Em ambas figuras temos um perfil figurativo de fluxo de genes nos genomas. Na figura A observamos que as espécies “c,d,e e f” possuem um certo gene, e que possivelmente foi herdado de seu ancestral comum. Entretanto, observe que em “d” duas cópias deste gene é encontrada, o que é mais parcimonioso pensar que este gene foi adquirido por evento de duplicação. Ao mesmo tempo, as espécies“a, b e h” não o possuem em decorrência de um possível evento de deleção. De qualquer forma, só se observa neste caso a chamada herança vertical. Contrapondo a esta análise, na figura B temos um mesmo perfil de presença ou ausência de genes nas respectivas espécies. Entretanto, pelos eventos demonstrados é possível especular que a presença dos genes nas espécies “c,d,e e f” só podem ser frutos de um evento de transferência lateral de genes (LGT) entre as espécies correlacionadas, já que seus ancestrais não é observado estes genes.

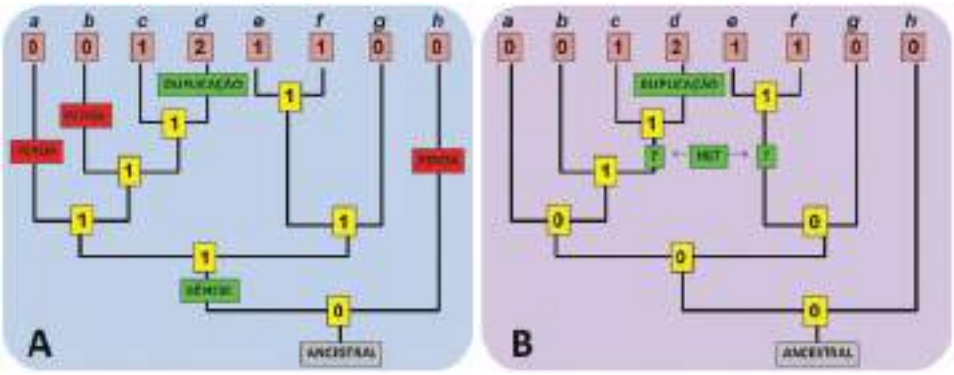


Figura 4. Modelos que demonstram possibilidades de heranças verticais e horizontais entre genomas ancestrais e derivados. (A) Modelo que enfatiza a perspectiva de ganho (por gênese e duplicação) e perda. (B) Modelo que enfatiza a perspectiva exclusiva de ganho mediado por duplicação e possível herança horizontal. Adaptado de Snel e colaboradores (2002).

Embora possa ser fácil compreender esta perspectiva a partir de uma imagem elucidativa como destacado na Figura 4, ainda fica uma pergunta que precisa ser esclarecida: Como analisar estes eventos em genomas completos?

O processo de análise destes eventos envolvem os conhecimentos adquiridos no capítulo passado sobre genômica comparativa. De certa maneira esses eventos são continuamente estudados pela comunidade científica, em um processo que depende de análises computacionais robustas, muitas vezes dispendiosas, sendo que é sempre necessário usar organismos modelos ou espécies correlacionadas que compartilham um provável ancestral em comum, para guiar para as comparações e deduções a serem realizadas (Figura 5). Tomemos como exemplo esta figura adaptada dos trabalhos de Snel e colaboradores que tentaram responder como seria o fluxo de informações em genomas procariotos.

Os dois painéis (A e B) refletem uma possível reconstrução da evolução dos genomas de Archea e Proteobactérias, respectivamente, a partir de alguns genomas completamente sequenciados. É interessante observar que tomando como base as possíveis informações de seus respectivos ancestrais (reconstruídos a partir das informações dos atuais genomas), alguns destes atuais genomas tiveram expansão, ao passo que outros apresentaram redução. Isto tem direta relação com a pressão seletiva ao qual foram expostos ao longo do tempo. O mais interessante é que mesmo que os atuais organismos tenham números de genes equivalentes a seus ancestrais, o conteúdo gênico é notoriamente diferenciado tendo em vista os diferentes eventos que atuaram nesta dinâmica alteração da composição do genoma, ou simplesmente plasticidade, a qual todos os genomas estão sujeitos.

Fica ainda mais evidente este dinamismo na composição do genoma quando a bactéria *E.coli* é analisado dentro do grupo das enterobactérias (Figura 6). Além de ser o organismo procarioto mais bem estudado em ensaios empíricos, também se trata do genoma mais bem estudado e conhecido, sendo que atualmente centenas de diferentes genomas *E. coli*, já foram sequenciados.

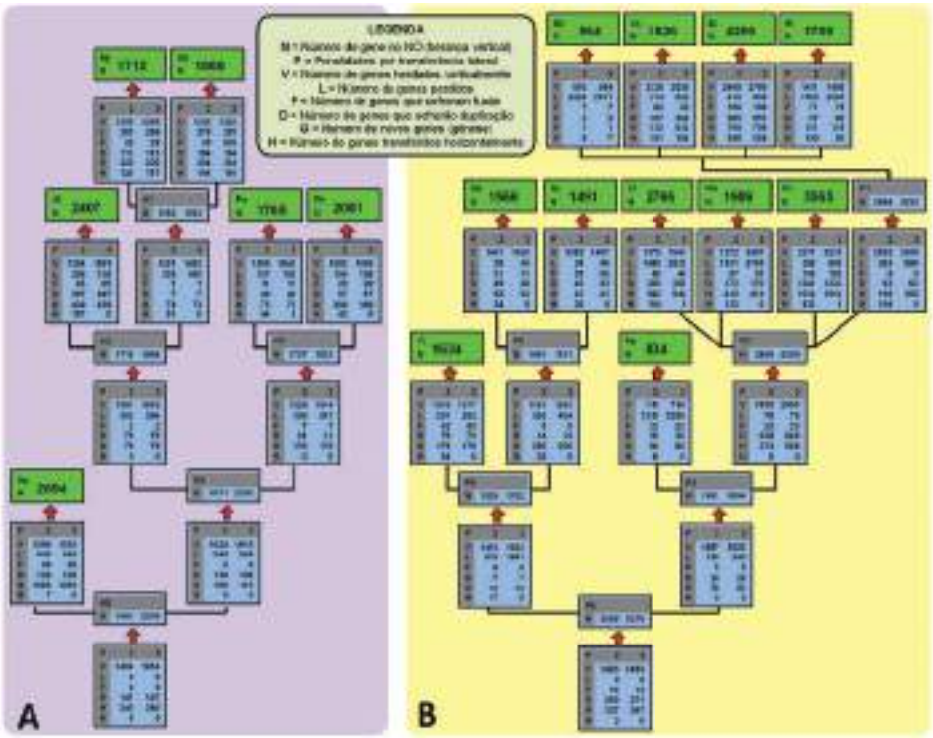


Figura 5. Reconstrução da evolução dos organismos demonstrando a plasticidade e fluxo gênico em seus genomas. As duas letras de abreviação denotam as espécies iniciais em cada nível do processo evolutivo (ver abaixo). O nó ancestral sempre denota um caractere que reflete seu respectivo táxon, e o número é usado para distinguir dois ou mais nós de um mesmo táxon. Cada nó também indica quantos genes são propostos para estarem presentes no respectivo ancestral. Cada processo que possivelmente tenha atuado na evolução destes genomas (Box cinza na figura) vem acompanhado de um número que expressa o total de genes para cada condição. (A) – Archaea (Af, *Archaeoglobus fulgidus*; Ap, *Aeropyrum pernix*; Mj, *Methanococcus jannaschii*; Mt, *Methanobacterium thermoautotrophicum*; Pa, *Pyrococcus abbysii*; e Ph, *Pyrococcus horikoshi*). (B) – Proteobacteria; (Bu, *Buchnera* sp. APS; Cj,*Campylobacter jejuni* NCTC 11168; Ec, *Escherichia coli*; Hi, *Haemophilus influenzae*; Hp, *Helicobacter pylori* 26695; Hy, *Helicobacter pylori* J99; Nm, *Neisseria meningitidis* MC58; Ps, *Pseudomonas aeruginosa* PA01; Rp, *Rickettsia prowazekii*; Vc, *Vibrio cholerae*; e Xf, *Xylella fastidiosa)*. Adaptado de Snel e colaboradores (2002).

Plasticidade do genoma e adaptação ao ambiente

Como descrito acima, a plasticidade do genoma depende intimamente das condições as quais estes foram submetidos e refletem a adaptação sob pressão seletiva imposta pelo meio, o princípio básico da seleção natural numa abordagem molecular.

Para tornar esta dinâmica mais contextualizada, será apresentado abaixo um caso interessante de plasticidade dos genomas de dois fitopatógenos capazes de infectar um mesmo hospedeiro, porém sobrevivendo em nichos completamente diferenciados no interior desta planta. Vamos retratar o caso de *Xylella fastidiosa* e *Xanthomonas citri*, respectivamamente os agentes causais da clorose variegada e cancrose dos citros.

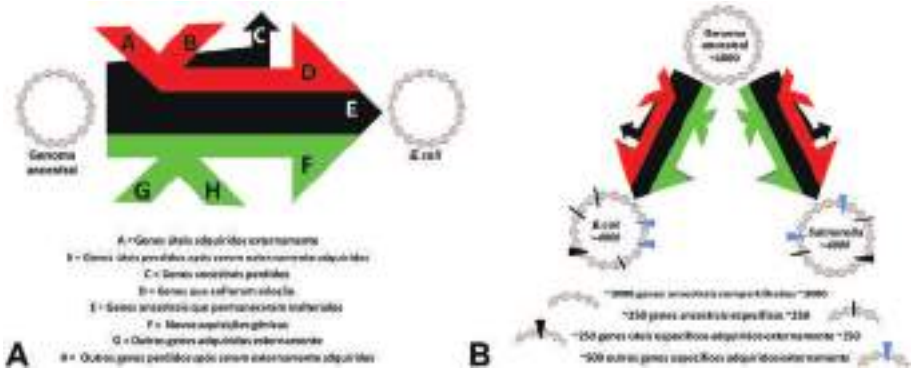


Figura 6. Fatores que determinaram a plasticidade no genoma de enteropatógenos. (A) Visão geral da influência de variados fatores (A a H) na atual composição do genoma de *E.coli*, a partir de seu genoma ancestral. Curiosamente, pela espessura da seta, é possível verificar que o o total de genes previstos no ancestral ou em *E.coli* são parecidos, porém a conteúdo destes genes é dependente da ação destes fatores (ver B). (B) Detalhamento, em número de genes, entre o conteúdo do genoma ancestral e o conteúdo dos atuais genomas de *E.coli* e *Salmonella*, destacando o número de genes compartilhados, específicos de cada espécie e adquiridos por HGT. Adaptado de Hacker, J. et al. (2006).

Xylella sobrevive no interior do xilema da planta, possui um genoma com cerca de 2,5 Mpb e um total de 2800 genes em seu genoma composto por um cromossomo e dois plasmídeos (Simpson, et al, 2000). *Xanthomonas*, por sua vez, sobrevive no espaço intersticial das células e possui um genoma com cerca de 4,4 Mpb e um total de 4284 genes distribuídos e um cromossomo e dois plasmídeos (da Silva, et al 2002).

Levando em conta que ambos apresentam um mesmo ancestral em comum, uma série de perguntas interessantes pode ser feita a partir deste quadro:

a. Qual seria o provável tamanho deste genoma ancestral?
b. Se o genoma ancestral tivesse tamanho e composição próximo ao genoma de *Xylella*, como explicar a expansão do genoma de *Xanthomonas*?
c. Se o genoma ancestral tivesse tamanho e composição mais próximos ao genoma de *Xanthomonas*, como explicar massiva redução do genoma de *Xylella*?
d. Que genes são compartilhados entre estes genomas a ponto de determinar um genoma mínimo ancestral?

Estas são só algumas perguntas que poderiam ser analisadas no contexto genômico. Porém, quando algumas informações biológicas são embutidas nesta perspectiva fica mais fácil compreender estas características genômicas.

Xylella só pode ser transferida mediante auxílio de um vetor animal, o que poderia justificar seu genoma mais reduzido e o fato de se alocar especificamente nos vasos lenhosos da planta. Em contrapartida, para *Xanthomonas*, vento, chuva, utensílios agrícolas e o homem podem ser considerados vetores de propagação.

Outro exemplo interessante que expressa bem à interferência do tempo e do ambiente na evolução dos genomas é descrito abaixo. Há muito tempo se sabe que durante o cultivo de bactérias em laboratório, é muito importante que em cada novo experimento se use alíquotas de amostras denominadas estoque. Ou seja, amostras que algum dia foram mantidas conservadas, e que são reativadas a cada novo ensaio, evitando que se faça um repique a partir das amostras já em crescimento. Por que

isso é importante? Porque cada vez que a bactéria tem seu genoma replicado, erros relacionados à própria dinâmica de replicação e pelas falhas inerentes à maquinaria de reparo, mutações são transferidas verticalmente. Além disso, possíveis elementos genéticos móveis podem alterar por completo a organização genômica que, por conseqüência, pode gerar algum distúrbio em futuras análises de resultados. Diante deste conhecimento de cunho prático A equipe de Barrick e colaboradores em 2009 resolveu sequenciar o completo genoma de uma estirpe de *E.coli* em diferentes tempos de crescimento. Sequenciaram esta mesma estirpe a partir de uma amostra conservada, após 5, 10, 15, 20 e 40 mil gerações, simulando uma possível perspectiva de evolução em curto espaço de tempo, e notaram que o genoma da mesma suposta bactéria no estágio inicial de crescimento é bastante diferente do mesmo genoma após

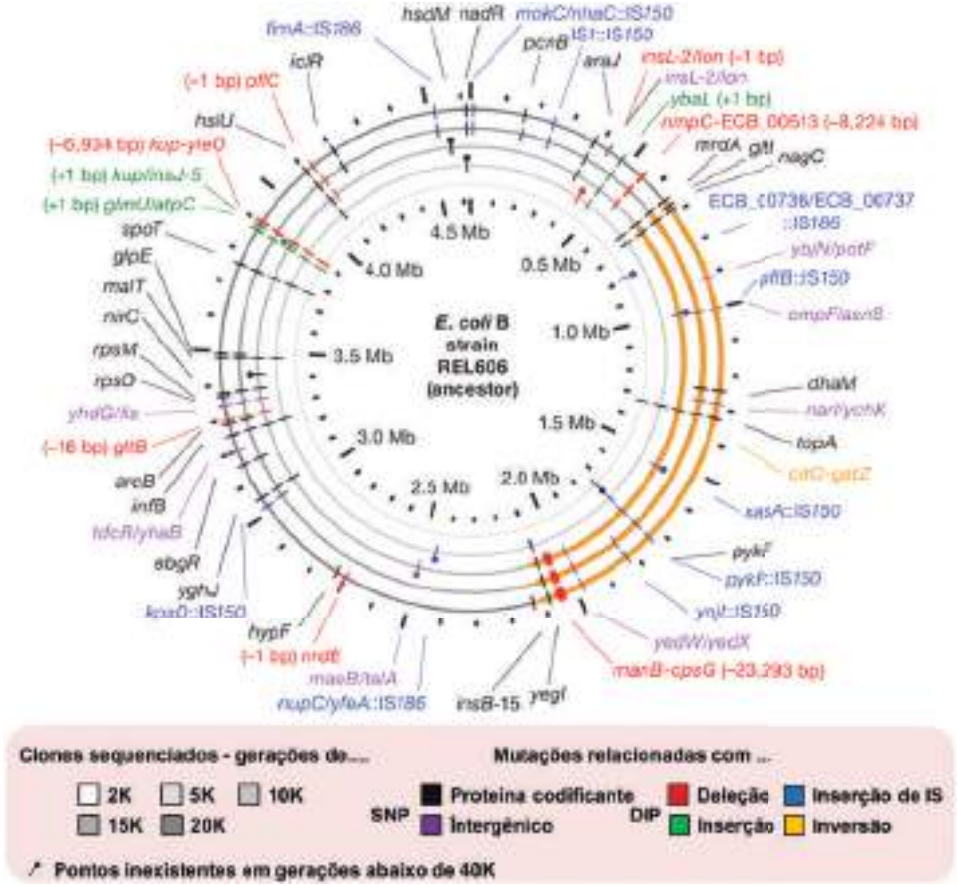


Figura 7. Evolução e modificações estruturais no genoma de *E.coli* ao longo de experimentos de longa duração. Todas as informações referentes a cores dos círculos e dos atributos presentes neles estão representadas na legenda figurativa. Na parte mais externa e com nomenclatura de quatro letras em itálico estão representados alguns genes importantes no genoma de *E.coli*. IS – denomina sequências de inserção. Note que muitos genes e regiões genômicas trocaram de posição ao longo das gerações, alguns genes também desapareceram (destacados em letras vermelhas). Adaptado de Barrick e colaboradores (2009).

40 mil gerações (Figura 7). Eles não concluíram, mas deixaram que a comunidade pense a respeito da seguinte pergunta: Será que podemos dizer que após tanto tempo replicando este organismo podemos dizer que ao final teremos a mesma bactéria?

Plasticidade genômica e aquisição de genes de virulência

Todo o contexto acima discutido passa a ser ainda mais interessante quando destacamos esta plasticidade genômica correlacionada a genes de virulência em patógenos animais e vegetais.

Veja só que perguntas interessantes:

a. Por que *Mycobaterium leprae* é causador da lepra (hanseníase) ao passo que *Mycobacterium tuberculosis* é agente causador da tuberculose se ambas pertencem ao mesmo gênero?
b. O que faz destes organismos tão próximos evolutivamente provocarem doenças completamente diferentes num mesmo hospedeiro?

Perguntas deste tipo podem ser extrapoladas a outros organismos modelos que infectam tanto hospedeiros animais quanto vegetais, e a resposta para boa parte destas perguntas está associada a diferenças nas composições gênicas, frutos de ganhos e perdas de parte de seus genomas.

Estes ganhos e perdas estão quase sempre associados a genes que conferem a estes organismos potenciais para se tornarem mais virulentos ou até mesmo patogênicos. As regiões genômicas que carregam estes genes, por sua vez, recebem o nome de ilhas de genômica e são de extremo interesse para que se possa conhecer melhor a biologia da interação entre patógeno e hospedeiro.

As ilhas genômicas (IGs) são EGMs presentes no genoma de bactérias que apresentam um conjunto de genes flexível em sua constituição, ligados geralmente com atributos de patogenicidade e virulência ou adaptação ao meio. As ilhas genômicas podem ser classificadas de acordo com as vantagens que agregam àbactéria, sendo subdivididas em: (a) ilhas adaptativas (*fitness islands*): possuem genes cujos produtos promovem um aumento na adaptação do organismo recipiente ao meio, incluindoaumento na sobrevivência, dispersão e transmissão do organismo dentro de um nicho ecológicoespecífico; (b) ilhas de patogenicidade (IPs): possuem genes cujos produtos contribuem para virulência da bactéria, como por exemplo toxinas; (c) ilhas de simbiose: possuem genes relacionados a processos de interações simbióticas com outros organismos; (d) ilhas de resistência: possuem genes responsáveis pela degradação de antibióticos ou outros compostos tóxicos para a célula (Hentschel et al., 2001).

Vamos exemplificar um caso interessante baseado nos genomas de *E. coli* (Figura 8). Existe uma série de estirpes ou isolados de *E.coli* capazes de causar diferentes enfermidades: *E.coli* enteropatogênica (EPEC), *E.coli* enterohemorrágica (EHEC), *E.coli* enterotoxigênica (ETEC), *E.coli* enteroagregativa (EAEC), *E.coli* de aderência difusa (DAEC), *E.coli* uropatogênica (UPEC), *E.coli* associada a meningite (MNEC) e *E.coli* que causa infecção extraintestinal (ExPEC). Curiosamente, diversos estudos têm mostrado que as diferenças genômicas que permitem esta diversidade patológica são pequenas e associadas a *loci* especializados, denominados, como citados acima, de ilhas de patogenicidade em *E.coli*.

Kaper e colaboradores (2004) tentaram reunir as informações existentes sobre evolução destas estirpes e ilhas de patogenicidade e chegaram um modelo que poderia explicar estas diferenças. Para isso, no entanto, foram incorporados à história evolutiva a presença de transposons, plasmídeos, bacteriófagos e ilhas de patogenicidade transferidas por outros organismos. Segundo estes autores estes EGM foram fundamentais para a geração desta diversidade de estirpes, provando a importância destes elementos na diversidade e evolução das espécies.

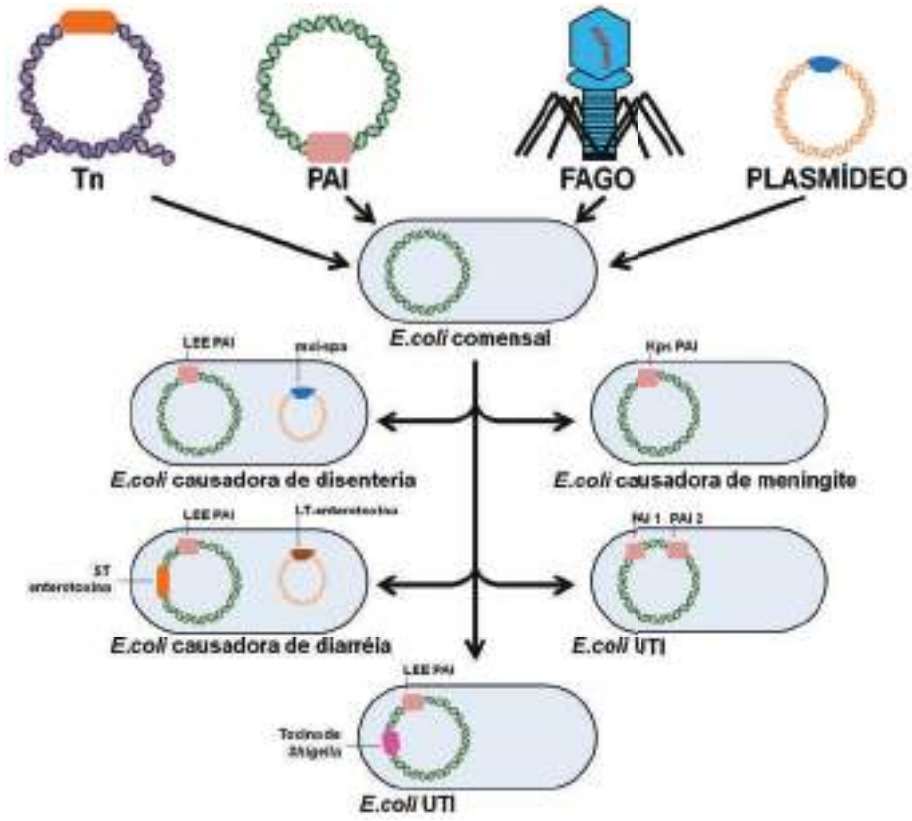


Figura 8. Fatores de virulência de *E.coli* que podem estar associados a diferentes EGM.Entre os cinco isolados de *E.coli* relacionados com as principais patologias por elas provocadas destacam-se isolados causadores de disenteria, diarréia, meningite, e dois isolados capazes de provocar infecção de trato urinário (UTI). A história evolutiva e análise da composição de seus genomas permitem inferir que um ancestral comum de característica comensal passou a apresentar características patogênicas após receber DNAs exógenos provenientes de diferentes fontes, entre elas: Tn – transposon que carrega enterotoxina estável ao calor (ST enterotoxina); Plasmideo que carrega enterotoxina sensível à temperatura (LT enterotoxina), Bacteriófago que carrega shiga toxina; e ilhas de patogenicidade (PAIs) que carregam fatores associados a invasão tecidual. Adaptado de Kaper e colaboradores (2004).

Bibliografias

BARRICK, J.E., YU, D.S., YONN, S.H., JEONG, H., OH, T.K., SCHNEIDER, D., LENSKI, R.E. and KIM, J.F. Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* 461, 1243-1247

DA SILVA, A.C., FERRO, J.A., REINACH, F.C., et al (2002) Comparison of the genomes of two *Xanthomonas* pathogens with differing host specificities. *Nature* 417, 459-463.

DRÖGE M, PÜHLER A, SELBITSCHKA W. Horizontal gene transfer as a biosafety issue: a natural phenomenon of public concern. J Biotechnol. 1998, 17;64(1):75-90.

EISEN, J.A.Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. Curr Opin Genet Dev. 2000, 10(6):606-11.

FRASER-LIGGETT, C.M. Insights on biology and evolution from microbial genome sequencing. Genome Res. 2005, 15(12):1603-10.

HACKER, J., DOBRINDT, U. and GÖBEL, W. Pathogenomics: Genome Analysis of Pathogenic Microbes. 2006, Wiley-Blackwel, 616 pages.

KAPER, J.B, NATARO, J.P. and MOBLEY, H.L.T. pathogenic Escherichia coli. Nature review Microbiology, 2004, 2, 123 – 140.

MARDIS, E.R. The impact of next-generation sequencing technology on genetics. Trends Genet. 2008, 24(3):133-41.

MART KRUPOVIC, M., PRANGISHVILI, D., HENDRIX, R.W. AND BAMFORD, D.H. Genomics of Bacterial and Archaeal Viruses: Dynamics within the Prokaryotic Virosphere. Microbiol. Mol. Biol. Rev. 75, 4610-635, 2011.

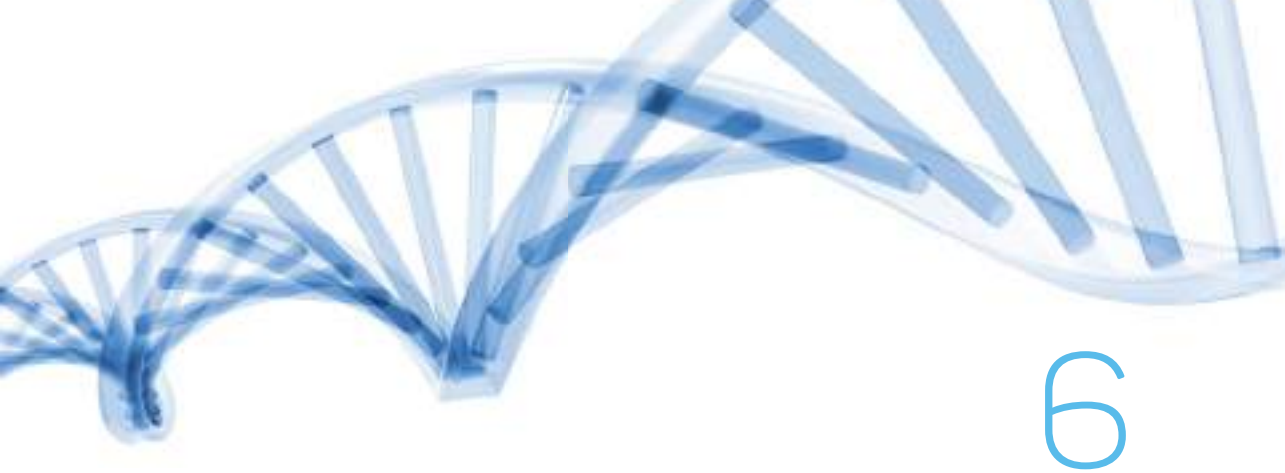
SANGER, F., NICKLEN, S. and COULSON, A.R; DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci 1977. Dec;74(12):5463-7.

SCHNEIDER, W.P., HO, S.K., CHRISTINE, J., YAO, M., MARRA, A. and HROMOCKYJ, A.E. Virulence gene identification by differential fluorescence induction analysis of Staphylococcus aureus gene expression during infection-simulating culture.Infect Immun. 2002, 70(3):1326-33.

SIMPSON, A.J., REINACH, F.C., ARRUDA, P., et al. (2000). The genome sequence of the plant pathogen *Xylellafastidiosa*. The *Xylella fastidiosa* Consortium of the Organization for Nucleotide Sequencing and Analysis. Nature 406, 151–157.

SMITH GP (June 1985). “Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface”. Science 228 (4705): 1315–7.

SNEL, B, BORK, P and HUYNEN, M.A. Genomes in Flux: The Evolution of Archaeal and Proteobacterial Gene Content. Genome Res. 2002. 12: 17-25.



6

Filogenômica

Laila Alves Nahum
Jerônimo Conceição Ruiz

Introdução

O termo filogenômica foi proposto por Jonathan Eisen, no final da década de 1990, como sendo a interseção entre filogenética e genômica com o intuito de aprimorar a predição funcional de genes (Eisen et al. 1997). Este pesquisador com formação multidisciplinar percebeu a importância de interpretar os dados genômicos gerados pelo sequenciamento de DNA usando uma plataforma evolutiva, uma abordagem que teve suas raízes nos trabalhos científicos da década de 1960 (e.g. Dayhoff 1965, Zukerdandl and Pauling 1965, Fitch and Margoliash 1967).

A filogenética (do grego, *phylon* + *genetikos*) é uma das áreas da biologia evolutiva assim como a filogeografia, evolução molecular, dentre outras (Ridley 2003, Barton et al. 2007, Futuyma 2013, Matioli and Fernandes 2011). A filogenética reconstroi as relações evolutivas entre organismos (vivos ou extintos), macromoléculas (DNA, RNA, proteínas, etc.), ecossistemas ou entre quaisquer outros elementos que compaprtilhem uma origem evolutiva comum.

A genômica, termo cunhado pelo geneticista Thomas Roderick em 1986, refere-se ao estudo do mapeamento, sequenciamento e análise do genoma (do inglês, *genome, genes* + *chromoss*ome). Por analogia, outros termos foram criados contendo o sufixo ômica, tais como transcritômica, proteômica, metabolômica, dentre muitos outros. Coletivamente, estes termos estão relacionados ao uso de diferentes tecnologias para a análise de dados biológicos não exclusivamente, porém frequentemente, em larga escala.

A motivação de Eisen baseou-se em um dos principais desafios da interpretação de dados genômicos que diz respeito à predição funcional de genomas, genes e seus produtos a partir das suas respectivas sequências moleculares (Eisen 1997). Desde então, o termo filogenômica tem sido usado em diferentes contextos e aplicações (Eisen and Hanawalt 1999, Eisen and Fraser 2003, Sjölander 2004, Delsuc et al. 2005,

Jeffroy et al. 2006, Nahum and Pereira 2008, Nahum et al. 2009, Sjölander 2010, Engelhardt et al. 2011, Burki 2014, Wang et al. 2014, Wang and Wu 2015). Em alguns casos, os termos filogenômica e filogenética são usados como sinônimos embora sejam conceitualmente distintos.

Antes mesmo do termo filogenômica ser cunhado, pesquisadores já analisavam os dados de genomas completamente sequenciados através da reconstrução de árvores evolutivas. Na literatura, os trabalhos envolvendo análise filogenômica podem ser referidos também por outras terminologias, tais como: "*whole-genome phylogeny*", "*genome-wide phylogenetic analysis*", "*whole genome-based phylogenetic analysis*", "*evolutionary genomics*", etc. (e.g. Fitz-Gibbon and House 1999, Uddin et al. 2004, Kuo et al. 2008, Bonaventura et al. 2010). Quando a filogenômica é usada na análise de genomas mitocondriais, por exemplo, esta abordagtem é frequentemente referida como mitogenômica (Pereira and Baker 2006, Pacheco et al. 2011, Wang and Wu 2015).

Cabe ressaltar ainda que a filogenômica não se limita à análise de genomas completamente sequenciados. Ela inclui também a análise de famílias gênicas e proteicas ou mesmo genes individuais em questões relacionadas à biologia evolutiva das macromoléculas e/ou dos organismos nos mais variados ambientes (Eisen 1997, Eisen and Hanawalt 1999, Nahum et al. 2009, Castoe et al. 2007, Andrade et al. 2011).

Este capítulo apresenta inicialmente conceitos fundamentais aos estudos de filogenômica que envolvem a interpretação de árvores evolutivas e as relações de homologia entre genes e seus produtos. Em seguida, o capítulo trata das principais metodologias de análise filogenômica com ênfase àquelas que utilizam dados de sequências moleculares. Apresentam-se alguns temas muito importantes como a predição de homologia, predição funcional e exemplos da análise de genomas completamente sequenciados usando a filogenômica. Em conjunto, o capítulo aborda variados temas sob uma perspectiva evolutiva e, portanto, intertidisciplinar.

Árvores evolutivas

A árvore filogenética, também chamada de árvore evolutiva ou filogenia, é a forma mais amplamente utilizada de representação dos dados evolutivos (Figura 1). Ela mostra as relações entre diferentes elementos (e.g. genes) presentes na base de dados analisada e seus possíveis ancestrais. A árvore é um tipo especial de grafo (cf. teoria de grafos) e inclui alguns componentes principais, tais como: ramos, nós (internos e terminais) e raiz no caso das árvores enraizadas.

Os ramos conectam as pontas (ou arestas) aos seus ancestrais representados pelos nós na árvore. Os ramos correspondem a um único táxon ou um único gene no caso de uma árvore de genes. O mesmo se aplica à todas as demais filogenias, sejam de famílias de proteínas, de caracteres morfológicos, dentre outras. Uma escala é fornecida quando o comprimento dos ramos (do inglês, *branch length*) é proporcional ao tempo evolutivo ou à variação genética.

Um táxon consiste em um grupo de organismos (e.g. espécies, gêneros, famílias, etc.). Frequentemente, o táxon é referido como unidade taxonômica operacional (do inglês, *operational taxonomic unit* – OTU). No caso das macromoléculas, estas unidades são representadas por sequências de nucleotídeos (DNA ou RNA) ou de resíduos de aminoácidos (proteínas).

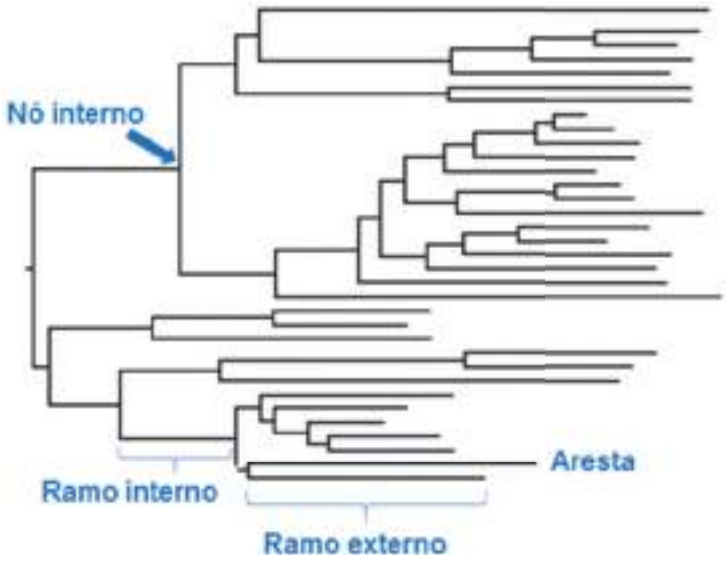


Figura 1. Componentes da árvore evolutiva. Uma árvore evolutiva hipotética mostrando ramos externos e internos, nós internos e arestas (*tips*). As arestas da árvore correspondem aos táxons (OTU) ou macromoléculas (genes, proteínas, introns, etc.).

Na mesma árvore, é possível ter diferentes níveis taxônomicos. Por exemplo, espécies e subespécies. Quando se trata de árvores de famílias gênicas, diferentes genes e pseudogenes podem ser identificados. O mesmo se observa para árvores de famílias proteicas, nas quais variantes funcionais são estudadas (Nahum and Pereira 2008, Nahum et al. 2009).

Os nós representam um ancestral (táxon, gene, proteína, etc.) dando origem a dois ou mais ramos. Geralmente, os nós vêm associados a valores de apoio estatístico (do inglês, *support values*) que indicam o grau de confiança de um determinado grupo indicado pela topologia da árvore.

O grupo monofilético, também chamado de clado, é formado por dois ou mais ramos conectados por um nó com apoio estatístico significativo, ou seja, um grupo constituído por um ancestral e todos os seus descendentes (táxons, genes, proteínas, etc.). Se mais de dois ramos emergem a partir do nó, tem-se uma politomia (do grego, *polli* + *tomi*, muitos cortes). A identificação de uma politomia indica que as relações entre os elementos analisados não foram resolvidas.

As árvores podem ser enraizadas ou não enraizadas. O enraizamento da árvore pode ser feito pela escolha de um grupo externo (do inglês, *outgroup*). A raiz estabelece a ordem na qual os eventos evolutivos ocorreram ao longo do tempo. Em alguns casos, a escolha do grupo externo representa um desafio.

Existem diferentes formas de se representar uma mesma árvore evolutiva (Figura 2). As árvores podem estar dispostas na forma retangular, radial ou circular. No caso das retangulares, elas podem ser orientadas horizontalmente com as arestas à direita e a raiz à esquerda ou vice-versa. As árvores também podem ser orientadas verticalmente com as arestas no topo e a raiz na base ou vice-versa. No cladograma,

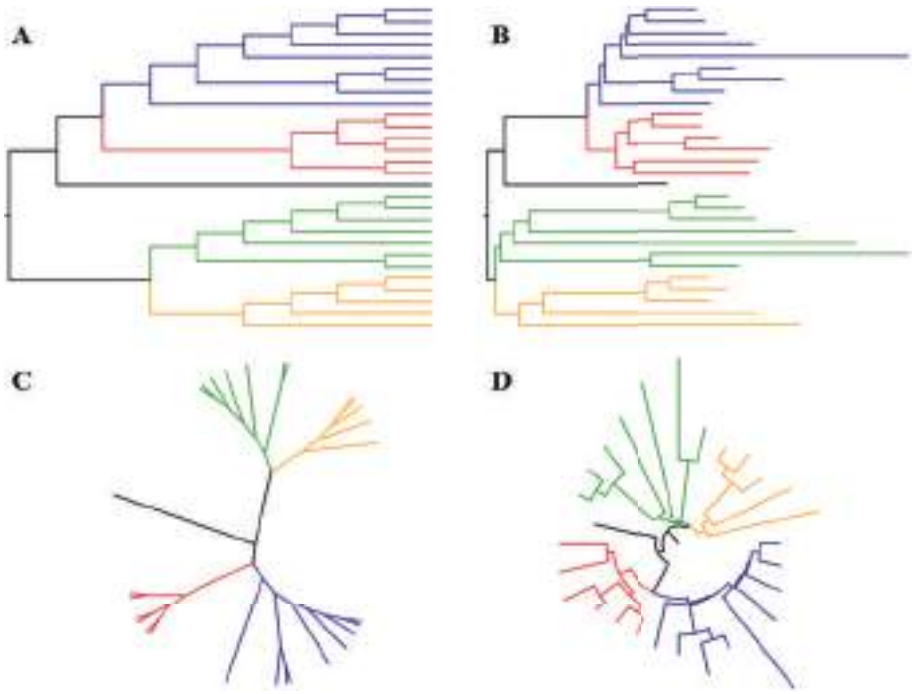


Figura 2. Representações da árvore evolutiva. Cladograma (A), filograma (B, C e D) dispostos nas formas retangular (A e B), radial (C) e polar (D). Imagens geradas a partir de uma árvore hipotética usando o programa FigTree.

todos os ramos têm igual comprimento, ao passo que, no filograma os ramos têm comprimentos distintos refletindo a diversidade dos dados analisados.

A árvore evolutiva representa uma hipótese ou um conjunto de hipóteses sobre as relações evolutivas entre os elementos de uma análise comparativa. O racional que permeia o delineamento, reconstrução e interpretação de árvores é denominado *tree-thinking* (O´Hara 1997, Bapteste et al. 2005, Baum et al. 2005, Cracraft and Bybee 2005, Omland et al. 2008, Sandvik 2008, Meisel 2010). A literatura que discute este racional é muito interessante e certamente recomendada para aqueles que pretendem construir ou expandir seu conhecimento nas diferentes áreas da biologia evolutiva.

Homologia e evolução molecular

Homologia *versus* similaridade

Homologia e similaridade **não** são termos intercambiáveis (cf. Reeck et al. 1987).

Homologia é a relação de ancestralidade entre dois ou mais elementos (e.g. genes). Dizer que genes, proteínas, sequências, estruturas ou posições de um alinhamento são homólogos significa dizer que os mesmos compartilham um ancestral comum. Sendo assim, a homologia é um termo qualitativo.

A similaridade, por sua vez, corresponde ao grau de proximidade entre duas ou mais sequências moleculares, geralmente expresso em porcentagem (%). Portanto, a similaridade é um termo quantitativo e pode ser calculada através de diferentes abordagens. Sequências ou estruturas similares podem ou não compartilhar um ancestral comum.

Por exemplo, podemos dizer que dois genes homólogos são 90% similares no nível da sequência de nucleotídeos. Porém, estes genes não podem ser referidos como 90% homólogos. O conceito de grau de homologia **não** existe!

Sobre a origem da similaridade nos níveis da sequência, estrutura e/ou função biológica, devemos considerar pelo menos dois cenários. As sequências moleculares podem se originar por evolução divergente (descendência com modificação a partir de um ancestral comum) ou evolução convergente (similaridade sem que haja um ancestral ou história evolutiva comum).

Ao longo do tempo evolutivo, homólogos podem divergir a ponto de não mais exibirem grau de similaridade detectável pelos métodos computacionais. Portanto, nem todos os homólogos são similares nos níveis da sequência, estrutura e/ou função biológica.

A convergência evolutiva (ou evolução convergente) pode levar ao surgimento de sequências altamente similares apesar de não serem relacionadas evolutivamente. Da mesma forma, a presença de duas estruturas tridimensionais altamente similares não garante que estas proteínas sejam realmente homólogas. Além da evolução convergente ser bem aceita no nível morfológico, esta é cada vez mais discutida no nível molecular através de estudos que evidenciam sua ocorrência entre distintos genes (Doolittle 1994, Galperin et al. 1998, Gherardini et al. 2007, Castoe et al. 2007).

Um dos exemplos de convergência evolutiva melhor estudados é o da tríade catalítica das cisteino e serino proteases que evoluiu independentemente em mais de 20 superfamílias de enzimas (Buller and Townsend 2013). Outro exemplo são as famílias de galactoquinase, hexoquinase e riboquinase que tem funções enzimáticas similares na fosforilação de açúcares mas evoluiram a partir de três famílias não homólogas distintas (Bork et al. 1993). Tais enzimas possuem similaridade de sequência, porém apresentam estruturas tridimensionais completamente distintas.

Genes que tem similaridade funcional ou no nível de suas sequências, mas tem uma origem evolutiva independente, ou seja, não compartilham um ancestral comum, são chamados de genes análogos. O mesmo pode-se dizer das proteínas ou quaisquer outros elementos nesse contexto. Processos evolutivos como convergência, paralelismo e reversão dando origem a sequências ou estruturas análogas são chamados, em conjunto, de homoplasia.

Relações de homologia

Conforme mencionado anteriormente, genes homólogos são aqueles que descendem de um ancestral comum. O evolucionista Walter Fitch, pioneiro na reconstrução de árvores evolutivas baseadas em sequências de DNA e proteínas, definiu diferentes tipos de homólogos baseado em dados moleculares (Fitch and Margoliash 1967, Fitch 1970).

Segundo Fitch, genes parálogos são homólogos que divergiram entre si após um evento de duplicação gênica (Fitch 1970). Por exemplo, os genes humanos que

codificam as hemoglobinas α e β são parálogos. Por sua vez, genes ortólogos são homólogos que divergiram entre si após um evento de especiação (divergência entre duas espécies). Genes que codificam a globina α de duas espécies de mamíferos (e.g. homem e camundongo) são ortólogos.

Posteriormente, outros termos foram propostos para classificar os diferentes subtipos de genes parálogos (Sonnhammer and Koonin 2002, Koonin 2005). *Inparalogs* são parálogos que se originaram a partir de duplicações linhagem-específicas após um evento de especiação. Por sua vez, os *outparalogs* são parálogos resultantes de duplicações que precederam um dado evento de especiação.

Genes xenólogos são aqueles que divergiram entre si após um evento de transferência lateral de genes (Koonin et al. 2001). Os genes de resistência a antibióticos presentes em diferentes espécies de bactéria são um bom exemplo de genes xenólogos.

Uma árvore evolutiva representando as relações entre membros de uma superfamília ou família gênica em distintos organismos deverá conter tanto parálogos quanto ortólogos. Eventualmente, os genes xenólogos também podem estar presentes.

O número de homólogos de uma família gênica pode variar entre diferentes organismos em função de ganho, perda e eventos de duplicação gênica pós-especiação (Descorps-Declère et al. 2008, Gabaldón 2007, Chothia and Gough 2009, Nahum et al. 2009, Silva et al. 2011). A inativação de genes originando pseudogenes altera o número de genes funcionais de uma família gênica. A expansão de famílias gênicas pode refletir possíveis adaptações dos organismos a diversos ambientes (Copley et al. 2003, Nahum et al. 2009, Silva et al. 2011).

Mecanismos de evolução molecular

A duplicação gênica seguida de divergência é o principal mecanismo de evolução molecular conforme postulado por Susumu Ohno e posteriormente confirmado por vários estudos independentes realizados antes mesmo do desenvolvimento das tecnologias genômicas (Ohno 1970).

Erros ocorridos durante a recombinação homóloga ou mesmo eventos de retrotransposição podem levar à duplicação parcial ou total de genes. Além da duplicação gênica, podem ocorrer também a duplicação cromossômica (polissomia parcial ou total) e a duplicação genômica (poliploidia parcial ou total), que em conjunto constituem mecanismos muito importantes na evolução de diversos grupos taxonômicos conforme amplamente descrito na literatura (Ridley 2003, Griffiths et al 2004, Clarck 2005, Barton et al. 2007, Babá et al. 2009, Futuyma 2013).

Existem diversos outros mecanismos de evolução molecular. Dentre eles, citam-se: mutação, recombinação, ganho e perda de genes, amplificação gênica, conversão gênica, embaralhamento de éxons, embaralhamento de domínios proteicos, fusão de genes (proteínas multimodulares), transferência lateral (horizontal) de genes, *trans-splicing*, *splicing* alternativo de transcritos, *lineage sorting*, etc. (Page and Holmes 1998, Ridley 2003, Griffiths et al. 2004, Barton et al. 2007, Nahum 2011, Matioli and Fernandes 2011, Futuyma 2013). Como resultado, observam-se a neofuncionalização e subfuncionalização de genes e seus produtos ou mesmo a inativação de genes (pseudogenes).

Em conjunto, estes mecanismos modelam a evolução dos genomas, transcritomas, proteomas e quaisquer outros sistemas simples ou complexos que, orquestrados pelas interações com o ambiente (células, biomas, etc.), desempenham um papel fundamental na origem e evolução da extraordinária biodiversidade observada nos organismos contemporâneos como resultado de 3.5 bilhões de anos de história da vida biológica na Terra.

Cabe ressaltar, que a maioria destes mecanismos foram evidenciados através de estudos de genética clássica e genética molecular, sendo que a identificação da maioria deles precedeu as análises de genômica comparativa (Ridley 2003, Griffith et al. 2004, Clark 2005, Barton et al. 2007, Futuyma 2013). A compreensão destes processos é de fundamental importância para a interpretação de dados biológicos no contexto evolutivo como é o caso dos estudos envolvendo a análise filogenômica.

Tipos de dados

Diferentes tipos de dados podem ser usados para testar as hipóteses evolutivas. Dentre eles, citam-se os dados morfológicos, moleculares, ecológicos, fósseis, dentre outros. Exemplos de dados moleculares incluem: dados de alozimas, sítios de enzimas de restrição no DNA, sequências moleculares (DNA, RNA e proteínas), conteúdo gênico, ordem gênica (sintenia), assinaturas genômicas, etc. A ênfase deste capítulo é a análise filogenômica usando dados de sequências moleculares.

Sequências moleculares

Com o avanço das tecnologias de sequenciamento de ácidos nucléicos e a disponibilidade de dados em bancos públicos, os dados de sequência e organização de genomas, genes e seus produtos representam a principal "matéria-prima" da análise filogenômica, genômica comparativa e outras abordagens. De fato, existem diversos bancos de dados de sequências, estruturas, função biológica, taxonomia e ontologia disponíveis na Web (cf. Bolser et al. 2012). Alguns dos principais bancos de dados e ferramentas computacionais dedicados à análise filogenômica estão listados na Tabela 1.

Apesar do desenvolvimento de técnicas de sequenciamento de proteínas, a maioria das sequências de aminoácidos depositadas nos bancos de dados ainda corresponde àquelas preditas computacionalmente a partir das sequências de nucleotídeos. No caso dos dados estruturais, um número crescente de modelos estão sendo gerados por predição computacional, além daqueles gerados por métodos experimentais, tais como a cristalografia de Raio-X e ressonância magnética.

No caso das sequências moleculares, o nucleotídeo ou aminoácido é tratado como um caráter independente. Cada tipo de caráter pode apresentar diferentes estados. Sendo assim, sequências de DNA têm quatro estados (A, C, G e T), enquanto que sequências de proteína têm 20 estados (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W e Y). Por convenção, sequências de RNA depositadas em bancos de dados contêm T ao invés de U. As inserções ou deleções (do inglês, in*sertions or* dele*tions – indels*), incluídas nos alinhamentos de sequências moleculares, podem ser consideradas

Tabela 1. Bancos de dados de análise filogenômica.

Bancos de Dados	URL
BPG	http://phylogenomics.berkeley.edu
Ensembl Compara	http://www.ensembl.org/info/docs/compara
GeneTrees	http://genetrees.vbi.vt.edu
PANTHER	http://www.pantherdb.org
PHOGs	http://bioinf.fbb.msu.ru/phogs
Phylemon	http://phylemon.bioinfo.cipf.es
PhyloExplorer	http://www.ncbi.orthomam.univ-montp2.fr/phyloexplorer
PhyloFacts	http://phylofacts.berkeley.edu
PhylomeDB	http://phylomedb.org
TreeBASE	http://www.treebase.org
TreeFam	http://www.treefam.org
eggNOG	http://eggnog.embl.de

URL: Endereço de acesso na Web.

um estado de caráter adicional. Logo, sequências de DNA/RNA e proteína teriam, respectivamente, cinco e 21 estados de caráter.

Na análise filogenômica, podem ser usados dados de genes e proteínas individuais, famílias gênicas e proteicas, genomas e transcritomas parcial ou completamente sequenciados. Note que estas análises não se limitam aos genomas nem às análises em larga escala.

A escolha de marcadores moleculares para os estudos evolutivos depende da pergunta científica e da hipótese que se pretende testar. Devem ser consideradas as taxas de substituição de nucleotídeos ou de aminoácidos, a origem e o modo de evolução das sequências moleculares, sua presença em organismos de interesse, a disponibilidade e curadoria dos dados que se pretende analisar, etc. Os critérios de seleção de alvos filogenéticos são amplamente descritos na literatura (e.g Russo 2011, Freeman and Herron 2013).

Em se tratando da filogenômica usando dados moleculares, as etapas metodológicas incluem a identificação de potenciais homólogos, o alinhamento de sequências moleculares, a reconstrução de árvores evolutivas, predição de homologia e a anotação das árvores para interpretação dos resultados.

Seleção de sequências para análise

Uma etapa crucial na análise filogenômica é a identificação de potenciais homólogos que possam ser estudados nessa plataforma evolutiva. Essa etapa é realizada primeiramente pela seleção de sequências potencialmente homólogas a partir de bancos de dados e da predição de homologia usando-se diferentes métodos.

Quando se trata de sequências moleculares, podemos aplicar métodos extrínsecos e intrínsecos usando bancos de dados e ferramentas computacionais. Os métodos extrínsecos desconsideram as características existentes nas sequências a serem analisadas enquanto que os métodos intrínsecos são baseados no reconhecimento de características específicas da sequência em associação ao conteúdo da mesma.

O método de similaridade é um exemplo de método extrínseco. Ele baseia-se na busca por sequências similares em bancos de dados a partir de uma ou mais sequências de interesse, sendo cada uma delas tratada por *query*, ou seja, uma pergunta ao banco. A estratégia mais amplamente usada neste caso é a que envolve o uso dos *softwares* do pacote *Basic Local Alignment Search Tool* (BLAST) (Altschul et al. 1997), disponível no National Center for Biotechnology Information (NCBI).

A busca por similaridade não garante a identificação de homólogos, uma vez que similaridade e homologia são conceitos distintos como discutido em maior detalhe anteriormente. Por isso, nos referimos a esta etapa como a identificação de potenciais homólogos. A confirmação ou não da homologia dependerá da análise das filogenias para as distintas bases de dados evidenciando a ancestralidade comum entre as sequências selecionadas.

Sequências análogas podem ser recuperadas na busca por similaridade e estarão, portanto, presentes na árvore filogenética. Neste caso, poderão ser evidenciados os casos de convergência evolutiva, conforme mencionado anteriormente.

Um exemplo de método intrínseco é o uso de modelos ocultos de Markov (do inglês, *Hidden Markov Models* – HMMs), que permitem modelar a probabilidade de uma sequência linear de eventos em uma dada base de dados (Durbin et al. 1999). Os HMMs são amplamente usados na análise de dados biológicos como, por exemplo, na predição de genes no genoma, alinhamento múltiplo de sequências moleculares e identificação de potenciais sequências homólogas em bancos de dados (cf. Mount 2004). Bancos de dados como o Pfam (Finn et al. 2014) e SUPERFAMILY (Wilson et al. 2009) fazem uso desta metodologia para a identificação de famílias de domínios de proteínas.

O método intrínseco costuma ser específico para determinada base de dados, uma vez que os genes e proteínas variam consideravelmente entre diferentes contextos evolutivos (e.g. presença/ausência em diferentes organismos). Além disso, o perfil de expressão de genes e proteínas, assim como variantes de *splicing* alternativo variam em diferentes estágios do desenvolvimento ou localização celular de um organismo. Portanto, deve-se construir diferentes HMMs para distintas bases de dados de modo a testar diferentes hipóteses que possam responder as perguntas de interesse.

Conteúdo gênico, ordem gênica e outros

O alinhamento de sequências de genomas completamente sequenciados pode oferecer desafios importantes à análise filogenômica em função da distribuição desigual de homólogos entre distintos organismos, ou seja, pelas diferenças quanto à presença e ausência de genes no genoma dos mesmos.

Além disso, o grau de divergência entre as sequências presentes nestes genomas pode variar significativamente comprometendo a qualidade dos alinhamentos e, consequentemente, a acurácia e robustez da reconstrução de árvores evolutivas.

Uma alternativa é se trabalhar com um perfil filogenético (Pellegrini et al. 1999). Este perfil consiste em uma matriz de dados de presença e ausência de genes ou famílias de genes em cada organismo selecionado. O mesmo se aplica aos dados de proteínas ou famílias de proteínas presentes ou não em um grupo de organismos selecionados para análise.

Esta abordagem se baseia no conteúdo gênico e não leva em consideração a organização genômica. Para tanto, usa-se a matriz de presença e ausência de genes ou proteínas, domínios protéicos, etc. Alternativamente, pode se usar a distância evolutiva baseada na proporção de ortólogos compartilhados entre dois genomas divididos pelo tamanho do menor genoma evitando assim artefatos relacionados à variação no tamanho dos genomas analisados.

A vantagem desta abordagem é que é possível analisar uma grande quantidade de dados, cobrindo praticamente todo o genoma, acessando a história evolutiva dos organismos e não apenas a história de genes ou produtos gênicos. Uma das desvantagens desta abordagem é que ela não detecta os eventos de transferência lateral de genes, embora em alguns casos ela possa fornecer indícios para a identificação de tais eventos.

Estudos baseados em ordem gênica comparam regiões ortólogas do genoma de distintos grupos taxonômicos e buscam inferir a árvore evolutiva que minimiza o número de pontos de interrupção (do inglês, *breakpoints*) que levam à mudança da organização dos genes de um genoma em outro. Esta abordagem tem sido utilizada para reconstruir a filogenia de uma grande variedade de organismos (e.g. Blanchette et al. 1999).

Outra abordagem de análise de sequências se baseia no uso de assinaturas genômicas também chamadas de *DNA strings* (Qi et al. 2004). Neste caso, o algoritmo calcula a frequência de pequenos trechos de nucleotídeos presentes nas sequências analisadas, geralmente a partir de dinucleotídeos. As frequências são representadas graficamente na forma de imagem colorida na qual as cores representam a frequência dos *strings*.

A análise de assinaturas genômicas não requer que as sequências moleculares sejam alinhadas evitando possíveis limitações relativas à identificação de homologia e grau de divergência entre as mesmas.

É possível ainda usar mudanças genômicas raras para a análise filogenômica. Estas mudanças incluem *indels* de um único ou múltiplos nucleotídeos ou aminoácidos, posição de introns, informações sobre fusão e fissão de genes, integração de elementos móveis, dentre outros (Rokas and Holland 2000).

Alinhamentos e reconstrução filogenética

Alinhamento de sequências moleculares

O alinhamento é um procedimento computacional que visa estabelecer a correspondência entre as posições (sítios ou colunas) de duas ou mais sequências moleculares (linhas) mantendo a ordem das mesmas (Figura 3). As lacunas (*gaps*) no alinhamento correspondem a um ou mais eventos de inserção ou deleção em posições específicas das sequências de nucleotídeos ou aminoácidos. Geralmente, estes *indels* são representados por hífen (–) ou ponto (.) no alinhamento de sequências.

Existem diferentes tipos de alinhamento. Com relação ao número de sequências, tem-se o alinhamento par-a-par (do inglês, *pairwise alignment*) ou seja, entre duas sequências, e o alinhamento múltiplo (*multiple sequence alignment*) entre três ou mais sequências.

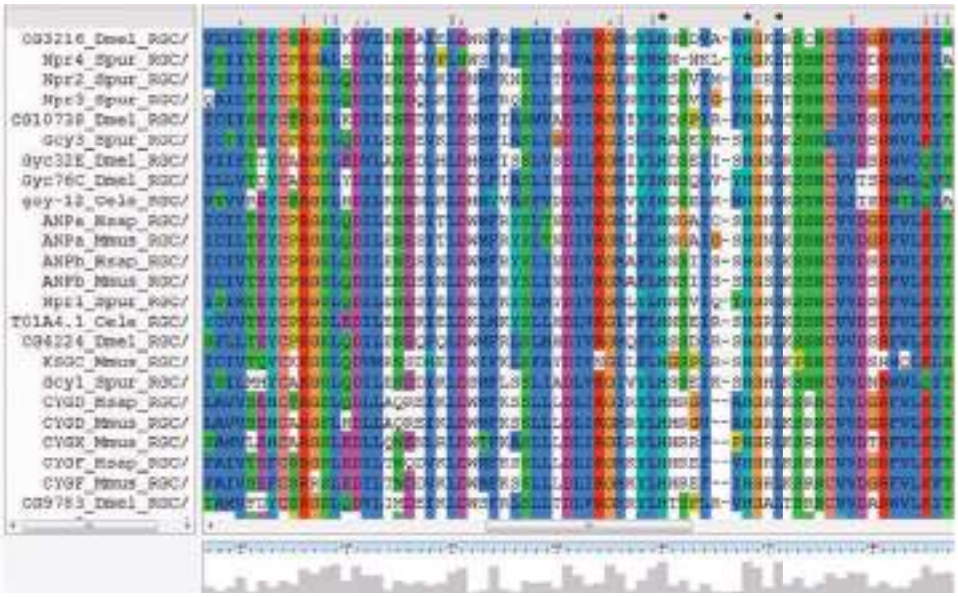


Figura 3. Alinhamento de múltiplas sequências de aminoácidos (proteína) gerado com o programa ClustalX. Os *gaps* (-) no alinhamento correspondem a *indels* nas sequências moleculares.

O alinhamento também pode ser classificado como global e local em função da estratégia usada para se alinhar duas ou mais sequências moleculares. No alinhamento global, todos os nucleotídeos ou aminoácidos de todas as sequências são alinhados uns aos outros na extensão completa da sequência de maior comprimento.

No alinhamento local, somente as regiões das sequências apresentando a mais alta densidade de identidades são alinhadas e, dessa forma, blocos de alinhamentos locais são identificados e mapeados nas sequências. Tais blocos podem cobrir grande parte da sequência original dependendo do grau de identidade e similaridade entre as sequências. Em outras palavras, quanto maior o grau de identidade/similaridade entre as sequências, maior a extensão do alinhamento local.

Os métodos de alinhamento local e global não são apropriados para a análise de dados contendo eventos de recombinação ou rearranjo de sequências. O primeiro método busca alinhar as sequências de modo a cobrir a região de sobreposição entre elas, enquanto que o segundo tenta forçar o alinhamento a fim de estender a região alinhada. Nestes casos, recomenda-se o uso de um método híbrido, denominado semiglobal ou "glocal" (global-local), que busca o melhor alinhamento possível que inclua o início e fim das sequências (Brudno et al. 2003). Este procedimento pode ser útil na análise comparativa de genomas completamente sequenciados.

Os métodos de alinhamento são baseados em programação dinâmica, método progressivo (hierárquico), método iterativo, HMMs, algoritmos genéticos e *simulated annealing* (cf. Mount 2004, Higgs and Attwood 2005). Diferentes algoritmos computacionais são usados para produzir e analisar os alinhamentos de sequências. Tais algoritmos foram implementados em diferentes *softwares*, tais como os listados

na Tabela 2. Diferentes *softwares* de alinhamento exibem distintos níveis de acurácia, robustez, dentre outras características (e.g. Edgar 2010).

Um alinhamento de sequências representa uma hipótese evolutiva. Cada posição no alinhamento contém nucleotídeos ou aminoácidos que supostamente compartilham uma mesma história evolutiva, i.e. evoluíram a partir de um ancestral comum. Trata-se de uma homologia de posição (do inglês, *positional homology*).

Note que nem todas as diferenças observadas nos sítios de um alinhamento correspondem a mutações nas sequências moleculares. Na realidade, a maioria destas diferenças refletem substituições ocorridas nas sequências homólogas ao longo do tempo evolutivo. É importante se fazer uma clara distinção entre os conceitos de mutação e substituição na análise de sequências moleculares. Embora estes termos sejam frequentemente usados como sinônimos, eles são conceitualmente distintos (Ridley 2003, Griffith et al. 2004, Barton et al. 2007, Futuyma 2013).

Uma mutação (do latim, *mutare*) corresponde a uma mudança herdável no DNA. Mutações podem alterar o fenótipo, mas isso não é uma regra. Por exemplo, uma mutação silenciosa dá origem a uma sequência diferente de DNA que especifica o mesmo aminoácido. Uma mutação neutra não altera função e a maior parte das mutações são neutras.

Por outro lado, uma substituição (e.g. nucleotídeo ou aminoácido) é uma mudança observada entre um ou mais elementos sem que haja alteração do fenótipo selvagem. Por exemplo, as substituições nas sequências de um gene de dois indivíduos ou duas populações diferentes conforme identificado na análise de alinhamentos.

A análise do alinhamento de sequências tem diversas aplicações no contexto da biologia molecular e evolução. Dentre eles, citam-se: análise de perfis e padrões,

Tabela 2. *Softwares* de alinhamento de sequências moleculares.

Softwares	URL
BioEdit	http://www.mbio.ncsu.edu/BioEdit/bioedit.html
Clustal	http://www.clustal.org
Diallign	http://dialign.gobics.de
MAFFT	http://www.ebi.ac.uk/Tools/mafft
Mugsy	http://mugsy.sf.net
Muscle	http://www.ebi.ac.uk/Tools/msa/muscle
Probalign	http://www.cs.njit.edu/usman/probalign
ProbCons	http://probcons.stanford.edu
SATCHMO-JS	http://phylogenomics.berkeley.edu/q/satchmo
T-Coffee	http://tcoffee.crg.cat
Gblocks	http://molevol.cmima.csic.es/castresana/Gblocks.html
trimAl	http://trimal.cgenomics.org
ZORRO	http://probmask.sourceforge.net

URL: Endereço de acesso na Web. *Softwares* de construção (topo) e edição/filtragem (fundo) de alinhamentos.

identificação de grupos de sequências relacionadas, montagem de genes e genomas, desenho de *primers* para a reação em cadeia da polimerase (do inglês, *polimerase chain reaction* – PCR), identificação de sequências de vetores de clonagem, estudos de polimorfismo, caracterização de inserções e deleções, análise de domínios protéicos, identificação de motivos conservados, predição de estrutura de proteínas, dentre outros (Griffith et al. 2004, Mount 2004, Babá et al. 2009).

Outras considerações sobre o alinhamento de sequências

Conforme mencionado anteriormente, a análise filogenômica pode ser realizada usando-se diferentes tipos de dados, por exemplo, sequências de genes e proteínas, famílias gênicas e proteicas, genomas e transcritomas parcial ou completamente sequenciados.

As sequências de nucleotídeos (DNA e RNA) ou de aminoácidos (proteínas) a serem alinhadas podem corresponder a sequências simples (e.g. um único gene) ou concatenadas (e.g. múltiplos genes dispostos sequencialmente na base de dados) de um ou mais organismos. No caso do alinhamento de sequências concatenadas, a ordem dos genes deve ser a mesma em todas as sequências analisadas, pois a maioria dos *softwares* de alinhamento assume essa premissa. Neste caso, o alinhamento das sequências pode ser realizado basicamente de duas formas: 1) alinhamento das sequências individuais e posterior concatenação das sequências alinhadas ou 2) alinhamento das sequências concatenadas de nucleotídeos ou aminoácidos.

Importante: Sequências de genes com diferentes padrões de organização (i.e. ocorrência e distribuição de éxons, introns e regiões não codificantes) devem ser previamente processados computacionalmente antes do uso das mesmas em *softwares* de alinhamento. O mesmo se aplica a proteínas que apresentam distintas arquiteturas (ocorrência e distribuição de domínios protéicos). Exemplos de proteínas com distintas arquiteturas estão ilustrados no Pfam (Finn et al. 2014). Este procedimento se justifica, pois a maioria dos *softwares* de alinhamento não leva em consideração estas diferenças. Portanto, o não processamento prévio das sequências com distintas organizações resulta em alinhamento de regiões não homólogas.

Obter um alinhamento de alta qualidade é etapa crucial no processo de reconstrução das árvores evolutivas (Nahum et al. 2006, Talavera and Castresana 2007, Jordan and Goldman 2011, Wu et al. 2012). A qualidade do alinhamento depende da sua acurácia, ausência ou baixa frequência de regiões de ambiguidade, dentre outros fatores. Obter um alinhamento de alta qualidade pode ser um procedimento bastante complexo considerando o número de sequências a serem analisadas e/ou o grau de divergência entre elas impondo um limite restritivo quanto ao custo computacional, ou seja, o tempo de processamento dos dados. A exclusão das posições ambíguas do alinhamento confere maior acurácia à reconstrução de árvores evolutivas. Existem diferentes *softwares* que permitem proceder à filtragem dos dados de alinhamento de sequências. Dentre eles, citam-se o Gblocks (Talavera and Castresana 2007) e o trimAl (Capella-Gutiérrez et al. 2009).

Para fins de reconstrução de árvores evolutivas, é possível combinar os alinhamentos de sequências de nucleotídeos e aminoácidos em uma mesma base de dados. Para isso,

deve-se alinhar separadamente cada tipo de sequência e posteriormente considerá-las como diferentes partições ao se usar os diferentes *softwares* de reconstrução de árvores evolutivas. Além disso, pode-se combinar outros dados como os morfológicos, estruturais, etc. à base de dados a ser analisada conforme implementado em alguns *softwares* de reconstrução de árvores evolutivas (Ronquist and Huelsenbeck 2003).

Reconstrução de árvores evolutivas

Existem basicamente duas categorias de métodos de reconstrução de árvores evolutivas: métodos de distância (geométricos) e métodos baseados em caracteres.

Os métodos de distância transformam os dados em uma medida da distância entre cada par de sequências e usam a matriz para a construção da árvore. Neste método, a análise é realizada em duas etapas principais. Primeiramente, calcula-se a matriz de distância entre cada par de sequências de um alinhamento. Posteriormente, constrói-se a filogenia usando os dados da matriz. Nesta etapa, usa-se um algoritmo de construção de árvores como o *neighbor-joining*, *stepwise addition*, *star decomposition*, etc. (Felsenstein 2003, Barton et al. 2007). Apesar de ser simples e rápido, o método de distância é pouco realista, pois se perde informação na conversão dos caracteres em medidas de distância entre as sequências. Além disso, esse método oferece limitações no caso de sequências divergentes.

Os métodos baseados em caracteres usam diretamente os caracteres alinhados, tais como sequências de nucleotídeos ou aminoácidos. Estes métodos incluem: máxima parcimônia (*maximum parcimony*), máxima verossimilhança (*maximum likelihood*) e inferência bayesiana (*bayesian inference*), sendo os dois últimos considerados métodos probabilísticos por calcularem a probabilidade dos dados serem explicados pelo modelo evolutivo (Felsenstein 2003, Barton et al. 2007, Matioli and Fernandes 2011).

O método de máxima parcimônia preconiza que a melhor hipótese evolutiva é aquela que requer o menor número de passos para explicar um dado processo. Dessa forma, a árvore que possuir um menor número de mudanças para explicar os dados do alinhamento é considerada ideal (árvore mais parcimoniosa). Neste método, as árvores são calculadas diretamente a partir dos dados do alinhamento. Não há cálculo de distância. As possíveis árvores são comparadas e cada uma delas recebe um *score* que reflete o número mínimo de mudanças no estado de caráter (e.g. substituições de nucleotídeos) necessários ao longo do tempo evolutivo para posicionar as sequências em uma dada árvore. A análise é relativamente rápida para bases de dados contendo algumas centenas de sequências e robusta quando as sequências são próximas entre si, ou seja, quando exibem altos níveis de similaridade. Entretanto, o método de máxima parcimônia tem baixo desempenho quando existe uma variação substancial entre as sequências analisadas (divergência entre as sequências).

A máxima verossimilhança é um método semelhante ao de máxima parcimônia no que diz respeito à atribuição de um *score* às diferentes topologias a serem comparadas, porém trata-se de um método probabilístico. O método de máxima verossimilhança busca a árvore que maximiza a probabilidade dos dados observados. Neste método, calculam-se as probabilidades associadas a diferentes topologias e cada uma delas com as variações nos tamanhos dos ramos, considerando o modelo evolutivo escolhido. A árvore ótima é aquela com maior valor de verossimilhança, ou seja, maior

probabilidade dos resultados terem se originado conforme o modelo de substituição de nucleotídeos ou de aminoácidos. Este é considerado um método de maior consistência e robustez, porém apresenta algumas desvantagens. Por ser um método complexo, tem um alto custo computacional, o que pode limitar as análises de bases de dados contendo um grande número de sequências. Além disso, é particularmente sensível a ambiguidades presentes no alinhamento.

A inferência bayesiana está muito relacionada ao método de máxima verossimilhança, porém pode ser realizada mais rapidamente para bases de dados contendo um grande número de sequências, além de ser menos sensível a ambiguidades. Esta análise usa o algoritmo de Monte Carlo baseado em cadeias de Markov (do inglês, *Markov chain Monte Carlo* – MCMC), uma classe de algoritmos para amostragem de distribuições de probabilidade baseadas na construção de cadeias de Markov. Na inferência bayesiana, estima-se a probabilidade posterior das hipóteses evolutivas a partir do conhecimento da topologia, comprimento dos ramos, parâmetros de substituição de nucleotídeos e probabilidades dos dados fornecidos *a priori*. As principais desvantagens deste método incluem a necessidade de se especificar a distribuição *a priori* dos parâmetros e a dificuldade em se determinar se o MCMC alcançou a convergência.

Existem distintos modelos evolutivos de sequências de nucleotídeos e de aminoácidos usados em estudos de evolução molecular e inferência filogenética (cf. Felsenstein 2003, Barton et al. 2007, Matioli and Fernandes 2011). Tratam-se de modelos matemáticos que descrevem a probabilidade de mudança de um caráter (nucleotídeo ou aminoácido) em outro. A maioria dos modelos são simplificações dos fenômenos biológicos e, portanto, são pouco realistas. Por outro lado, modelos complexos, i.e. com um maior número de parâmetros, requerem uma grande quantidade de dados a fim de testar a hipótese evolutiva. A quantidade de dados está relacionada ao número de sequências e/ou numero de sítios em um alinhamento.

O teste de múltiplos modelos para verificar qual deles melhor se adequa aos dados também pode ser feito durante a reconstrução das árvores seja por máxima verossimilhança ou inferência bayesiana. A seleção do modelo que melhor explica a base de dados a partir de um conjunto de modelos candidatos pode ser realizada usando-se ferramentas como as desenvolvidas pelo grupo do pesquisador David Posada: ModelTest (Posada 2006) e ProtTest (Darriba et al. 2011).

A topologia da árvore evolutiva por si só não é suficiente para se analisar as relações entre os táxons ou macromoléculas nas bases de dados analisados. É necessário avaliar o grau de confiança na topologia obtida após a reconstrução das árvores. Existem diferentes metodologias para se avaliar o grau de confiança na topologia de uma ou mais árvores evolutivas. Dentre elas, cita-se o *bootstrapping*. Nesta abordagem, as colunas do alinhamento original são reamostradas e novas bases de dados (réplicas) são geradas, sendo cada uma delas usada para gerar uma árvore. As árvores são comparadas e cada nó recebe um valor em porcentagem que indica quão frequente as duas sequências (arestas) ocorrem juntas nas diferentes árvores. Na inferência bayesiana, o valor de apoio estatístico atribuído a cada nó da árvore corresponde à probabilidade posterior.

Outra abordagem para se testar o grau de confiança de árvores evolutivas é o aLRT (do inglês, *approximate likelihood-ratio test*). Este teste é bastante rápido e tem demonstrado acurácia e robustez (Anisimova and Gascuel 2006). O aLRT vem sendo cada vez mais utilizado, especialmente em estudos em larga escala.

Existe um grande número de *softwares* para a reconstrução de árvores evolutivas disponíveis para instalação local ou para uso diretamente na Web. A Tabela 3 mostra uma relação de *softwares* ou pacotes desenvolvidos para essa finalidade. Uma referência importante é a lista de *softwares* reunida no *website* do pesquisador Joseph Felsenstein (University of Washington) que desenvolveu o PHYlogenetic Inference Package – PHYLIP (Felsenstein 1989).

A partir da inferência das árvores evolutivas, é importante proceder à anotação das mesmas com base nas informações disponíveis na literatura e em bancos de dados. Em se tratando de filogenias moleculares, as informações relevantes dizem respeito às sequências propriamente ditas, dados estruturais e função bioquímica dos produtos gênicos caracterizada experimentalmente por distintas metodologias.

Além disso, é importante acrescentar informações taxonômicas, ecológicas, etc. dos táxons dos quais foram obtidas as sequências. Esta é uma etapa crucial na interpretação dos resultados obtidos pela reconstrução de árvores evolutivas, especialmente para fins de predição funcional das sequências (hipotéticas ou preditas) não caracterizadas experimentalmente até o momento do estudo.

Tabela 3. *Softwares* de reconstrução de árvores evolutivas.

Softwares	URL
BAMBE	http://www.mathcs.duq.edu/larget/bambe.html
BEAGLE	http://code.google.com/p/beagle-lib
BEAST	http://beast.bio.ed.ac.uk
EDIBLE	http://www.ebi.ac.uk/goldman-srv/edible
GARLI	http://code.google.com/p/garli
GeneTree	http://taxonomy.zoology.gla.ac.uk/rod/genetree/genetree.html
MacClade	http://www.macclade.org
MEGA	http://www.megasoftware.net
Mesquite	http://www.mesquiteproject.org/mesquite/mesquite.html
MrBayes	http://mrbayes.sourceforge.net
PAML	http://abacus.gene.ucl.ac.uk/software/paml.html
PAUP*	http://paup.csit.fsu.edu
PHYLIP	http://evolution.genetics.washington.edu/phylip.html
PhyloBayes	http://www.phylobayes.org
Phylocom	http://phylodiversity.net/phylocom
Phylogeny.fr	http://www.phylogeny.fr
RAxML-VI-HPC	http://www.exelixis-lab.org
SHOT	http://coot.embl.de/~korbel/SHOT
TREE-PUZZLE	http://www.tree-puzzle.de
FigTree	http://tree.bio.ed.ac.uk/software/figtree
iTOL	http://itol.embl.de
Tree Editors	http://bioinfo.unice.fr/biodiv/Tree_editors.html
TreeDyn	http://www.treedyn.org
TreeView	http://taxonomy.zoology.gla.ac.uk/rod/treeview.html

URL: Endereço de acesso na Web. *Softwares* de construção (topo) e edição/visualização (fundo) de árvores evolutivas.

Na interpretação das árvores evolutivas, deve-se verificar se os resultados respondem a(s) pergunta(s) do estudo em questão e se os mesmos apoiam ou rejeitam a(s) hipótese(s) propostas(s) inicialmente (cf. Walsh and Sharma 2009).

Predição funcional de genes e seus produtos

Relação entre sequência e função

Conhecer a função biológica dos genes e seus produtos é de crucial importância em várias áreas da Ciência e Tecnologia. Esta tarefa se torna um grande desafio considerando o número crescente de dados de sequências moleculares depositadas em bancos de dados. Uma vez que a caracterização experimental de todas essas sequências seria inviável, torna-se necessária a utilização de metodologias que possam auxiliar na predição das possíveis funções desempenhadas pelos genes e seus produtos.

A maioria dos métodos de predição funcional baseia-se em buscas por similaridade em bancos de dados com a transferência da anotação funcional das sequências mais similares para a sequência de interesse. Esta abordagem constitui uma das principais fontes de erro na anotação dos genes individuais e/ou genomas completamente sequenciados (Bork and Koonin 1998, Galperin et al. 1998, Gilks et al. 2002, Sjölander 2004).

A similaridade de sequência pode ou não refletir a similaridade funcional dos alvos de estudo (e.g Gerlt and Babbitt 2000). Membros de famílias gênicas, por exemplo, podem compartilhar um grau de similaridade variável e divergirem quanto às funções biológicas desempenhadas em distintas condições fisiológicas. Existem alguns casos extremos nos quais a substituição de um único resíduo de aminoácido é responsável pela alteração da função bioquímica de uma dada proteína.

A identificação de homólogos pode não ser suficiente para realizar a predição funcional de um gene ou proteína ainda não caracterizados experimentalmente. Isso se deve ao fato de que nem todos os homólogos tem a mesma função. Por exemplo, a duplicação gênica seguida de divergência de sequência pode gerar genes com funções diferentes. Os mecanismos de evolução molecular mencionados anteriormente podem contribuir para a divergência funcional dos genes e proteínas em grau variável entre distintos grupos taxonômicos.

Iniciamente foi mencionada a limitação da predição baseada em similaridade. Então, citou-se que a identificação de homólogos não implica em confirmação de função. A seguir pretende-se ilustrar como a filogenômica emerge como plataforma evolutiva contribuindo efetivamente para a predição da função de genomas, genes e seus produtos bem como fornecendo *insigths* para priorização na identificação de alvos moleculares para futuras análises e delineamento experimental para a caracterização dos mesmos.

Predição funcional *via* filogenômica

Eisen e colaboradores foram os primeiros a demonstrar que a análise filogenética poderia ser usada como ferramenta para realizar a predição funcional de genes e proteínas, permitindo uma melhor identificação das relações de ortologia entre

membros da superfamília das SNF2, envolvidos em diversos processos celulares, tais como reparo de DNA, regulação da transcrição, dentre outros (Eisen et al. 1995). Posteriormente, Eisen cunhou o termo filogenômica como plataforma evolutiva para a predição funcional de genes e seus produtos, nomeando este racional a pedido do editor da revista Nature Medicine (Eisen et al. 1997).

As funções gênicas e de seus produtos podem se modificar ao longo do tempo e entre os diferentes organismos como resultado da evolução. Portanto, a reconstrução da história evolutiva dos genes e seus produtos pode auxiliar na predição funcional daqueles que ainda não foram caracterizados experimentalmente. Este é o racional em que se baseia a predição funcional a partir da filogenômica.

Conforme discutido anteriormente, o primeiro passo neste processo é a reconstrução de uma árvore evolutiva que represente uma hipótese ou um conjunto de hipóteses que representem as relações evolutivas de um alvo de interesse (genoma, gene, proteína, etc.) e seus homólogos em distintos organismos. As informações obtidas a partir da reconstrução de árvores evolutivas, tais como topologia, comprimento de ramos e apoio estatístico, podem contribuir para a predição funcional de diferentes maneiras.

Deve-se considerar que os clados (ancestrais e seus descendentes) identificados na árvore diferem dos *clusters* de similaridade de sequência, pois os primeiros resultam da inferência filogenética usando diferentes métodos computacionais que convertem padrões de similaridade em relações evolutivas tendo como premissa um modelo evolutivo definido.

Inicialmente, deve-se identificar os eventos de duplicação gênica e especiação que originaram, respectivamente, parálogos e ortólogos (e.g. Gabaldón 2007, Silva et al. 2011, Sonnhammer et al. 2014). Outros processos evolutivos podem ser identificados a partir das árvores baseados na interpretação das mesmas. Em seguida, é importante mapear na árvore todas as informações funcionais obtidas a partir da caracterização experimental descrita na literatura (e.g. Nahum et al. 2009).

A possibilidade de genes ortólogos compartilharem a mesma função é, em geral, mais alta do que quando parálogos são analisados, visto que parálogos surgem por duplicação gênica, um dos principais mecanismos de evolucão molecular. Pórem, cabe ressaltar, que isso não é uma regra.

As informações obtidas a partir da reconstrução de árvores evolutivas podem contribuir para a predição funcional de diferentes maneiras. As informações disponíveis podem ser usadas para traçar a história das modificações funcionais, identificando por exemplo, quais características são conservadas ao longo do tempo evolutivo e quais divergem entre os diferentes organismos. Eventos de neofuncionalização, subfuncionalização e inativação de genes e seus produtos podem ser revelados pela interpretação das árvores evolutivas. A hipótese de convergência evolutiva, ou seja, o compartilhamento de características (morfológicas, moleculares, etc.) similares sem que haja ancestralidade comum entre os genes ou organismos analisados também pode ser testada neste contexto.Com esta abordagem, também se pode corrigir erros de anotação funcional previamente descritos na literatura. Connforme mencionado anteriormente, é possível atribuir funções a genes ou proteínas sem caracterização experimental prévia. Além disso, pode-se identificar possíveis adaptações biológicas a partir da identificação de funções espécie-específicas e expansão ou redução de

famílias gênicas/proteicas em um organismo em relação aos demais organismos analisados (e.g. Nahum et al. 2009, Silva et al. 2011).

Existem vários bancos de dados e ferramentas commputacionais que usam a filogenômica como plataforma preditiva de funções biológicas de genes e proteínas (Tabela 1). Dentre elas, destaca-se o PhyloFacts desenvolvido pelo Berkeley Phylogenomics Group, liderado pela Dra. Kimmen Sjölander na Universidade da Califórnia, Berkeley, EUA (Krishnamurthy et al. 2006). Trata-se de uma enciclopédia filogenômica com informações estruturais e funcionais das proteínas do banco de dados UniProt (UniProt Consortium 2015), analisadas em uma plataforma evolutiva. O PhyloFacts identifica famílias de proteínas homólogas baseado na conservação da arquitetura proteica (sequência completa) ou de domínios protéicos (sequência parcial) identificados conforme o Pfam (Finn et al. 2014).

Para cada grupo de proteínas homólogas, o PhyloFacts disponibiliza o alinhamento das sequências, as árvores evolutivas, predição de ortologia, modelos ocultos de Markov, domínios protéicos de acordo com o Pfam, anotações segundo o Gene Ontology, dados experimentais, e outros tipos de dados.

Exemplos de estudos usando filogenômica

A filogenômica se aplica a um enorme número de situações abrangendo desde os estudos da biodiversidade e origem da vida até as aplicações em saúde, ambiente e sociedade (Sjölander 2004, Gabaldón 2007, Nahum and Pereira 2008, Mindell 2009, Burki 2014). Seguem-se exemplos que ilustram algumas das aplicações da filogenômica.

Dados de genes e genomas mitocondriais têm sido amplamente usados em estudos evolutivos de vários grupos taxonômicos há décadas. A análise de genes mitocondriais (e.g. citocromo c oxidase subunidade I – *cox*1), considerados como marcadores padrão na identificação de espécies, tem sido usados em estudos de código de barras de DNA (do inglês, *DNA barcoding*) que incluem reconstrução filogenética. Por outro lado, estudos de mitogenômica usando dados de genomas mitocondriais completamente sequenciados tem contribuído para revelar as relações evolutivas entre grandes ordens de aves, por exemplo.

Um estudo das relações evolutivas e tempos de divergência de representantes de ordens de Neognathae (Neoaves) ilustra o uso de dados de genomas mitocondriais completamente sequenciados (Pacheco et al. 2011). Neste estudo, foi possível resolver as politomias previamente observadas na filogenia de Neoaves analisando 80 genomas mitocondriais. Esta abordagem permitiu identificar Columbiformes (pombos, rolas, etc.) e Charadriiformes (gaivotas, maçaricos, etc.) como grupos irmãos. A partir desta amostragem taxonômica (do inglês, *taxon sampling*), foi possível resolver as relações evolutivas entre as principais ordens de aves. Além disso, as hipóteses evolutivas foram usadas para se estimar os tempos de divergência desses grupos indicando que esta diversificação ocorreu antes do limite Cretáceo/Terciário (K/T), que foi um pouco mais recente do descrito anteriormente na literatura. As árvores com as estimativas de tempo de divergência foram usadas para estimar a taxa de evolução de cada gene mitocondrial. Os autores identificaram uma grande

variação destas taxas entre os genes mitocondriais e entre as diferentes linhagens de aves analisadas.

Outro importante estudo foi realizado com os Apicomplexa. Estes incluem muitos patógenos importantes para a saúde humana e animal, tais como: *Babesia*, *Cryptosporidium*, *Plasmodium*, *Theileria* e *Toxoplasma*, cujos genomas foram completamente sequenciados. Um estudo comparativo do genoma nuclear de sete espécies de Apicomplexa identificou 268 genes de cópia única adequados à inferência filogenética (Kuo et al. 2008). Neste estudo, um ciliado de vida livre, *Tetrahymena thermophila*, foi usado como grupo externo. As filogenias obtidas foram consistentes com as concepções anteriores sobre a evolução de Apicomplexa baseadas em informações de ultraestrutura e de desenvolvimento. À primeira vista, o nível de incongruência entre as árvores de genes e árvore de espécies pareceu bastante elevado, porém a maioria dos conflitos observados não apresentou altos valores de apoio estatístico (*bootstrap*). Além disso, sequências de genes cujas análises filogenéticas geraram topologias com alto valor de apoio estatístico se mostraram robustas independentes das mudanças nos parâmetros de alinhamento ou do método filogenético utilizado. A análise de múltiplos genes não ligados exibindo forte sinal filogenético é importante para a inferência filogenética precisa, uma vez que distintos genes podem ter uma história evolutiva diferente da filogenia dos organismos. Em conjunto, este estudo forneceu uma lista de alvos filogenéticos de um grupo importante de patógenos direcionando futuras iniciativas de sequenciamento e caracterização experimental de representantes desse grupo.

Estudos de genômica comparativa têm mostrado que famílias de proteínas variam significativamente em um mesmo organismo e entre organismos distintos. Esta variação inclui o número de membros em cada família bem como as relações da sequência, estrutura e função dos mesmos.

Em outro estudo envolvendo a abordagem filogenômica, foi possível conectar a diversidade funcional de membros de famílias de enzimas à capacidade metabólica de distintos organismos contribuindo para suas características biológicas/fisiológicas particulares (Nahum et al. 2009). Para tanto, foram analisadas três famílias de proteínas em três distintas bactérias (*Escherichia coli*, *Bacillus subtilis* e *Pseudomonas aeruginosa*), cujo genoma foi completamente sequenciado e cuja biologia e ecologia são bem conhecidas e amplamente descritas na literatura. As famílias de enzimas estudadas apresentaram distintas composições e relações evolutivas entre si e entre as bactérias analisadas conforme evidenciado pela inferência bayesiana realizada neste estudo. As características funcionais conservadas entre membros de cada família incluem o mecanismo de reação, uso de cofatores e especificidade de substrato. Neste estudo, várias observações relativas à presença e ausência das funções enzimáticas correspondem ao conhecimento sobre a bioquímica e ecofisiologia destas bactérias. A análise também permitiu contribuir para a predição funcional de proteínas sem caracterização experimental prévia. Em conjunto, este tipo de abordagem pode ser bastante útil na predição da diversidade metabólica de organismos que são relativamente pouco conhecidos e/ou que ainda não são cultiváveis em laboratório como é o caso daqueles evidenciados por estudos de metagenômica.

Conclusões, desafios e perspectivas

O termo filogenômica foi cunhado para refletir a interseção entre filogenética e genômica para predição funcional de genes e seus produtos. Posteriormente, foi usado em distintos contextos e aplicações. Por se tratar de uma abordagem evolutiva, a filogenômica se baseia na reconstrução e interpretação de árvores, um racional também conhecido por *tree-thinking*, o qual assume que distintos elementos (organismos, moléculas, etc.) podem estar relacionados sob uma perspectiva histórica, temporal e espacial.

Dessa forma, a filogenômica envolve também a identificação, predição e interpretação de relações de homologia. As relações de homologia implicam em ancestralidade comum. A similaridade (quantitativa) pode ser um indicativo de homologia (qualitativa), porém os dois termos não são intercambiáveis.

Diferentes dados podem ser usados na análise filogenômica como as sequências moleculares, conteúdo e ordem gênica, dentre outros. Em se tratando de sequências moleculares, as etapas metodológicas incluem identificação de potenciais homólogos, obtenção de alinhamentos e reconstrução de árvores evolutivas para diferentes finalidades. A comparação de dados de genomas completamente sequenciados oferece alguns desafios importantes na obtenção de alinhamentos de boa qualidade. Alternativamente, usam-se dados de conteúdo e ordem gênica, além de assinaturas genômicas.

Existem diferentes métodos de reconstrução filogenética que incluem distintos modelos evolutivos e algoritmos implementados em um grande número de *softwares* amplamente descritos na literatura. A escolha do tipo de dado e metodologia de análise dependem da natureza da hipótese evolutiva que se deseja testar. Esta, por sua vez, esta intimamente relacionada às perguntas pertinentes ao objeto de estudo conforme a ótica da metodologia científica.

Outros desafios encontrados na análise filogenômica dizem respeito ao custo computacional das mesmas que incluem o tempo de processamento devido à complexidade dos dados, dos modelos, etc. A computação paralela é uma estratégia usada para contornar estes desafios. Outra possibilidade é a utilização de computação nas nuvens (do inlês, *cloud computing*) onde o processo computacional é distribuído em centenas ou milhares de computadores localizados em ampla distribuição geográfica.

Em conjunto, estas abordagens têm a capacidade de gerar um volume imenso de dados. Todavia, assim como qualquer outra análise de dados (biológicos ou não) em menor ou maior escala, o desafio reside e residirá sempre na interpretação dos mesmos e construção de conhecimento. Nesse sentido, técnicas de mineração de dados e representação de ontologias podem auxiliar tremendamente no processo.

O principal desafio nesta e em qualquer outra área da Ciência diz respeito à formação de recursos humanos com perfil inter/multidisciplinar, autônomo e criativo. O profissional deve ter sempre uma boa fundamentação teórica, ser conhecedor dos conceitos e dos seus relacionamentos e certamente ser conhecedor da história e filosofia do seu campo de atuação, seja este educacional, científico, tecnológico, de inovação ou outro. Afinal... "Quem não conhece sua própria história, arrisca-se a repeti-la" (autor desconhecido).

Agradecimentos

Dedicamos este capítulo ao saudoso Professor Dr. Henrique Lenzi, um profundo conhecedor da vida e seus sistemas, por nos falar sobre "o encanto da educação, a beleza da filogenia"... e por outros tantos ensinamentos. Agradecemos de modo especial ao Dr. Leandro Márcio Moreira pelo convite para participarmos deste livro. Aradecemos também à Dra. Larissa Lopes Silva Scholte pela revisão criteriosa deste capítulo. Agradecemos aos orientandos e discentes das disciplinas coordenadas e ministradas pelos autores deste capítulo pelas discussões construtivas que inspiraram a elaboração deste material. A preparação deste capítulo contou com o financiamento do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (CNPq-Universal 476036/2010-0) e do National Institutes of Health/Fogarty International Center (NIH/Fogarty) (D43TW007012).

Bibliografias

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389-3402.

Andrade, L.F., Nahum, L.A., Avelar, L.G., Silva, L.L., Zerlotini, A., Ruiz, J.C., and G. Oliveira. 2011. Eukaryotic protein kinases (ePKs) of the helminth parasite *Schistosoma mansoni*. *BMC Genomics*, 12:215.

Anisimova, M., and O. Gascuel. 2006. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Systematic Biology*, 55(4):539-552.

BABÁ, Elio Hideo et al. *DNA Recombinante: Genes e Genomas*. Porto Alegre: Artmed, 2009. p. 477.

Bapteste, E., Susko, E., Leigh, J., MacLeod, D., Charlebois, R.L., and W.F. Doolittle. 2005. Do orthologous gene phylogenies really support tree-thinking? *BMC Evolutionary Biology*, 5:33.

BARTON, Nicholas H. et al. *Evolution*. Cold Spring Harbor Laboratory Press, 2007. p. 833.

Baum, D.A., Smith, S.D., and S.S. Donovan. 2005. Evolution. The tree-thinking challenge. *Science*, 310(5750):979-980.

Blanchette, M., Kunisawa, T., and D. Sankoff. 1999. Gene order breakpoint evidence in animal mitochondrial phylogeny. *Journal of Molecular Evolution*, 49(2):193-203.

Bolser, D.M., Chibon, P.Y., Palopoli, N., Gong, S., Jacob, D., Del Angel, V.D., Swan, D., Bassi, S., González, V., Suravajhala, P., Hwang, S., Romano, P., Edwards, R., Bishop, B., Eargle, J., Shtatland, T., Provart, N.J., Clements, D., Renfro, D.P., Bhak, D., and J. Bhak. 2012. MetaBase--the wiki-database of biological databases. *Nucleic Acids Research*, 40(Database issue): D1250-D1254.

Bonaventura, M.P., Lee, E.K., Desalle, R., and P.J. Planet. 2010. A whole-genome phylogeny of the family Pasteurellaceae. *Molecular Phylogenetics and Evolution*, 54(3):950-956.

Bork, P., and E.V. Koonin. 1998. Predicting functions from protein sequences--where are the bottlenecks? *Nature Genetics*, 18(4):313-318.

Bork, P., Sander, C., and A. Valencia. 1993. Convergent evolution of similar enzymatic function on different protein folds: the hexokinase, ribokinase, and galactokinase families of sugar kinases. *Protein Science*, 2(1):31-40.

Brudno, M., Malde, S., Poliakov, A., Do, C.B., Couronne, O., Dubchak, I., and S. Batzoglou. 2003. Glocal alignment: finding rearrangements during alignment. *Bioinformatics*, 19 Suppl 1:i54-62.

Buller, A.R., and C.A. Townsend. 2013. Intrinsic evolutionary constraints on protease structure, enzyme acylation, and the identity of the catalytic triad. *Proceedings of the National Academy of Sciences of the United States of America*, 110(8):E653-E661.

Burki, F. 2014. The eukaryotic tree of life from a global phylogenomic perspective. *Cold Spring Harbor Perspectives in Biology*, 6(5):a016147.

Capella-Gutiérrez, S., Silla-Martínez, J.M., and T. Gabaldón. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15):1972-1973.

Castoe, T.A., Stephens, T., Noonan, B.P., and C. Calestani. 2007. A novel group of type I polyketide synthases (PKS) in animals and the complex phylogenomics of PKSs. *Gene*, 392(1-2):47-58.

Chothia, C., and J. Gough. 2009. Genomic and structural aspects of protein evolution. *Biochemical Journal*, 419(1):15-28.

CLARCK, David P. *Molecular Biology: Understanding the Genetic Revolution*. ACACL, 2005. p. 816.

Copley, R.R., Goodstadt, L., and C. Ponting. 2003. Eukaryotic domain evolution inferred from genome comparisons. *Current Opinion in Genetics & Development*, 13(6):623-628.

CRACRAFT, Joel, BYBEE, Rodger W. *Evolutionary Science and Society: Educating a New Generation*. BSCS, AIBS, 2005.

Darriba, D., Taboada, G.L., Doallo, R., and D. Posada. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, 27(8):1164-1165.

Dayhoff, M.O. 1965. Computer aids to protein sequence determination. *Journal of Theoretical Biology*, 8(1):97-112.

Delsuc, F., Brinkmann, H., and H. Philippe. 2005. Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, 6(5):361-375.

Descorps-Declère, S., Lemoine, F., Sculo, Q., Lespinet, O., and B. Labedan. 2008. The multiple facets of homology and their use in comparative genomics to study the evolution of genes, genomes, and species. *Biochimie*, 90(4):595-608.

Doolittle, R.F. 1994. Convergent evolution: the need to be explicit. *Trends in Biochemical Sciences*, 19(1):15-18.

Durbin, Richard et al. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1999. p. 356.

Edgar, R.C. 2010. Quality measures for protein alignment benchmarks. *Nucleic Acids Research*, 38(7):2145-2153.

Eisen, J.A., and C.M. Fraser. 2003. Phylogenomics: intersection of evolution and genomics. *Science*, 13,300(5626):1706-1707.

Eisen, J.A., and P.C. Hanawalt. 1999. A phylogenomic study of DNA repair genes, proteins, and processes. *Mutation Research*, 435(3):171-213.

Eisen, J.A., Kaiser, D., and R.M. Myers. 1997. Gastrogenomic delights: a movable feast. *Nature Medicine*, 3(10):1076-1078.

Eisen, J.A., Sweder, K.S., and PC Hanawalt. 1995. Evolution of the SNF2 family of proteins: subfamilies with distinct sequences and functions. *Nucleic Acids Research*, 23(14):2715-2723.

Engelhardt, B.E., Jordan, M.I., Srouji, J.R., and S.E. Brenner. 2011. Genome-scale phylogenetic function annotation of large and diverse protein families. Genome Research, 21(11):1969-1980.

Felsenstein, J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*, 5: 164-166.

FELSENSTEIN, Joseph. *Inferring Phylogenies*. Sinauer Associates, 2003. p. 664 pages.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E.L., Tate, J., and M. Punta. 2014. Pfam: the protein families database. *Nucleic Acids Research*, 42(Database issue):D222-D230.

Fitch, W.M. 1970. Distinguishing homologous from analogous proteins. Systematic Zoology. 19(2):99-113.

Fitch, W.M., and E. Margoliash. 1967. Construction of phylogenetic trees. *Science*, 155(3760):279-284.

Fitz-Gibbon, S.T., and C.H. House. 1999. Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Research*, 27(21):4218-4222.

FREEMAN, Scott, HERRON, Jon C. *Evolutionary Analysis*. Benjamin Cummings, 2013. p. 864.

FUTUYMA, Douglas. *Evolution*. Sinauer Associates Inc., 2013. p. 656.

Gabaldón, T. 2007. Evolution of proteins and proteomes: a phylogenetics approach. *Evolutionary Bioinformatics Online*, 1:51-61.

Galperin, M.Y., Walker, D.R., and E.V. Koonin. 1998. Analogous enzymes: independent inventions in enzyme evolution. *Genome Research*, 8(8):779-790.

Gerlt, J.A., and P.C. Babbitt. 2000. Can sequence determine function? *Genome Biology*, 1(5):REVIEWS0005.

Gherardini, P.F., Wass, M.N., Helmer-Citterich, M., and M.J. Sternberg. 2007. Convergent evolution of enzyme active sites is not a rare phenomenon. *Journal of Molecular Biology*, 372(3):817-845.

Gilks, W.R., Audit, B., De Angelis, D., Tsoka, S., and CA Ouzounis. 2002. Modeling the percolation of annotation errors in a database of protein sequences. Bioinformatics, 18(12):1641-1649.

GRIFFITHS, Anthony J.F. et al. *An Introduction to Genetic Analysis*. W. H. Freeman, 2004. 800 p.

HIGGS, Paul G., ATTWOOD, Teresa K. *Bioinformatics and Molecular Evolution*. Wiley-Blackwell, 2005. p. 384.

Jeffroy, O., Brinkmann, H., Delsuc, F., and H. Philippe. 2006. Phylogenomics: the beginning of incongruence? *Trends in Genetics*, 22(4):225-231.

Jordan, G., and N. Goldman. 2011. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Molecular Biology and Evolution*, 29(4):1125-1139.

Koonin, E.V. 2005. Orthologs, paralogs, and evolutionary genomics. *Annual Review of Genetics*, 39:309-38.

Koonin, E.V., Makarova, K.S., and L. Aravind. 2001. Horizontal gene transfer in prokaryotes: quantification and classification. *Annual Review of Microbiology*, 55:709-742.

Krishnamurthy, N., Brown, D.P., Kirshner, D., and K. Sjölander. 2006. PhyloFacts: an online structural phylogenomic encyclopedia for protein functional and structural classification. *Genome Biology*, 7(9):R83.

Kuo, C.H., Wares, J.P., and J.C. Kissinger. 2008. The Apicomplexan whole-genome phylogeny: an analysis of incongruence among gene trees. *Molecular Biology and Evolution*, 25(12):2689-2698.

MATIOLI, Sergio Russo, FERNANDES, Flora Maria de Campos (Org.). *Biologia molecular e evolução*. Ribeirão Preto: Holos Editora, 2011. p. 249.

Meisel, R.P. 2010. Teaching Tree-Thinking to Undergraduate Biology Students. *Evolution* (N Y), 3(4):621-628.

Mindell, D.P. 2009. Evolution in the everyday world. *Scientific American*, 300(1):82-89.

MOUNT, David W. *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press, 2004. p. 692.

Nahum, L.A., Goswami, S., and M.H. Serres. 2009. Protein families reflect the metabolic diversity of organisms and provide support for functional prediction. *Physiological Genomics*, 38(3):250-260.

Nahum, L.A., Reynolds, M.T., Wang, Z.O., Faith, J.J., Jonna, R., Jiang, Z.J., Meyer, T.J., and D.D. Pollock. 2006. EGenBio: a data management system for evolutionary genomics and biodiversity. *BMC Bioinformatics*, 7 Suppl 2:S7.

NAHUM, Laila Alves, Pereira, Sergio Luiz. *Phylogenomics, Protein Family Evolution, and the Tree of Life: An Integrated Approach between Molecular Evolution and Computational Intelligence*. In: Smolinski TG, Milanova MG, Hassanien A-E (eds), Studies in Computational Intelligence (SCI) 122. Berlin Heidelberg: Springer-Verlag, 2008. p. 259-279.

NAHUM, Laila Alves. Evolução dos Genomas. In: *Biologia Molecular e Evolução*. Ribeirão Preto: Holos Editora: 2011. p. 249.

O'Hara, R.J. 1997. Population thinking and tree thinking in systematics. *Zoological Scripta*, 26:323–329.

OHNO, S. (1970). *Evolution by gene duplication*. Springer-Verlag. ISBN 0-04-575015-7.

Omland, K.E., Cook, L.G., and M.D. Crisp. 2008. Tree thinking for all biology: the problem with reading phylogenies as ladders of progress. *Bioessays*, 30(9):854-867.

Pacheco, M.A., Battistuzzi, F.U., Lentino, M., Aguilar, R.F., Kumar, S., and A.A. Escalante. 2011. Evolution of modern birds revealed by mitogenomics: timing the radiation and origin of major orders. *Molecular Biology and Evolution*, 28(6):1927-1942.

PAGE, Roderick D.M., HOLMES, Edward C. *Molecular Evolution: A Phylogenetic Approach*. Wiley-Blackwell, 1998. p. 352.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and T.O. Yeates. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 96(8):4285-4288.

Pereira, S.L., and A.J. Baker. 2006. A mitogenomic timescale for birds detects variable phylogenetic rates of molecular evolution and refutes the standard molecular clock. *Molecular Biology and Evolution*, 23(9):1731-1740.

Posada, D. 2006. ModelTest Server: a web-based tool for the statistical selection of models of nucleotide substitution online. *Nucleic Acids Research*, 34(Web Server issue):W700-3.

Qi, J., Wang, B., and B.I. Hao. 2004. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. *Journal of Molecular Evolution*, 58(1):1-11.

Reeck, G.R., de Haën, C., Teller, D.C., Doolittle, R.F., Fitch, W.M., Dickerson, R.E., Chambon, P., McLachlan, A.D., Margoliash, E., Jukes, T.H., and E. Zuckerkandl. 1987. "Homology" in proteins and nucleic acids: a terminology muddle and a way out of it. *Cell*, 50(5):667.

RIDLEY, Mark. *Evolution*. Oxford University Press, 2003. p. 472.

Rokas, A., and P.W. Holland. 2000. Rare genomic changes as a tool for phylogenetics. *Trends in Ecology & Evolution*, 15(11):454-459.

Ronquist, F., and J.P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572-1574.

RUSSO, Claudia A. M. Como escolher genes para problemas filogenéticos específicos. In: *Biologia Molecular e Evolução*. Ribeirão Preto: Holos Editora: 2011. p. 249.

Sandvik, H. 2008. Tree thinking cannot taken for granted: challenges for teaching phylogenetics. *Theory in Biosciences*, 127(1):45-51.

Silva, L.L., Marcet-Houben, M., Zerlotini, A., Gabaldón, T., Oliveira, G., and L.A. Nahum. 2011. Evolutionary histories of expanded peptidase families in *Schistosoma mansoni*. *Memórias do Instituto Oswaldo Cruz*, 106(7):864-877.

Sjölander, K. 2004. Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics*, 20(2):170-179.

Sjölander, K. 2010. Getting started in structural phylogenomics. *PLoS Computional Biology*, 6(1):e1000621.

Sonnhammer, E.L., and E.V. Koonin. 2002. Orthology, paralogy and proposed classification for paralog subtypes. *Trends in Genetics*, 18(12):619-620.

Sonnhammer, E.L., Gabaldón, T., Sousa da Silva, A.W., Martin, M., Robinson-Rechavi, M., Boeckmann, B., Thomas, P.D., Dessimoz, C., Quest for Orthologs consortium. 2014. Big data and other challenges in the quest for orthologs. *Bioinformatics*, 30(21):2993-2998.

Talavera, G., and J. Castresana. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology*, 56(4):564-577.

Uddin, M., Wildman, D.E., Liu, G., Xu, W., Johnson, R.M., Hof, P.R., Kapatos, G., Grossman, L.I., and M. Goodman. 2004. Sister grouping of chimpanzees and humans as revealed by genome-wide phylogenetic analysis of brain gene expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9):2957-2962.

UniProt Consortium. 2015. UniProt: a hub for protein information. Nucleic Acids Research, 43(Database issue):D204-D212. doi: 10.1093/nar/gku989. Epub 2014 Oct 27. PubMed PMID: 25348405; PubMed Central PMCID: PMC4384041.

Walsh, D.A., and A.K. Sharma. 2009. Molecular phylogenetics: testing evolutionary hypotheses. *Methods in Molecular Biology*, 502:131-168.

Wang, Z., and M. WU. 2015. An integrated phylogenomic approach toward pinpointing the origin of mitochondria. *Scientific Reports*, 5:7949.

Wang, Z., Xie, Z., Cai, Y., Shu, K., and F. Huang. 2014. Advances in phylogenomics. *Yi Chuan*, 36(7):669-678.

Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., Chothia, C., and J. Gough. 2009. SUPERFAMILY--sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Research*, 37(Database issue):D380-D386.

Wu, M., Chatterji, S., and J.A. Eisen. 2012. Accounting for alignment uncertainty in phylogenomics. *PLoS One*, 7(1):e30288.

Zuckerkandl, E., and L. Pauling. 1965. Molecules as documents of evolutionary history. *Journal of Theoretical Biology*, 8(2):357-366.

7

Mutassômica

Leandro Marcio Moreira
Marcia Regina Soares

Introdução

Se pudéssemos retroceder a 1910, teríamos a oportunidade de conhecer em tempo real as propostas de pesquisa e os resultados gerados por Thomas Hunt Morgan, um dos maiores geneticistas da história. Morgan, trabalhando com drosófilas (*Drosophila* sp), pôde observar em seu laboratório que algumas destas moscas ao invés de apresentarem olhos vermelhos brilhantes, como a maioria delas, apresentavam olhos brancos (Morgan, 1909).

Baseado nestas observações envolvendo os fenótipos diferenciados e embasado em resultados similares previamente descritos para outros modelos por pesquisadores como Hugo de Vries, Carl Correns, Erich von Tschermak e até mesmo Charles Darwin, Morgan se propôs a verificar como esta característica aparecia na população e se era transmitida verticalmente (de pais para filhos) (Morgan, 1910). Graças a esta indagação e a sua curiosidade, resultados inovadores surgiram levando ao que conhecemos atualmente como a teoria cromossômica, resultados diretos de seu trabalho.

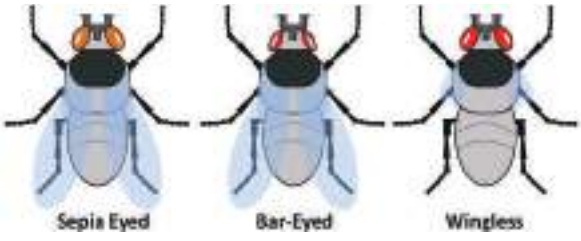


Figura 1. Perfil de coloração dos olhos observados por Morgan no início do século passado. Note que indivíduos Wingless além de ter coloração e tamanho de olhos diferenciados ainda apresenta asas vestigiais.

A elaboração desta teoria levou Morgan a ser considerado o primeiro pesquisador da história a provar que uma característica genética estava intimamente ligada ao cromossomo, e que a transferência de cromossomos específicos entre as gerações (meiose, formação de gametas e fertilização) é que definiam os genótipos e, por consequência, os fenótipos. Conhecimento hoje que representam a base do conhecimento da Genética.

Tais moscas de olhos brancos retratam a primeira evidência de que uma mutação em um gene (considerado então recessivo) acarretava mudança fenotípica específica. Estes resultados abriram as portas do conhecimento para que outros estudos que

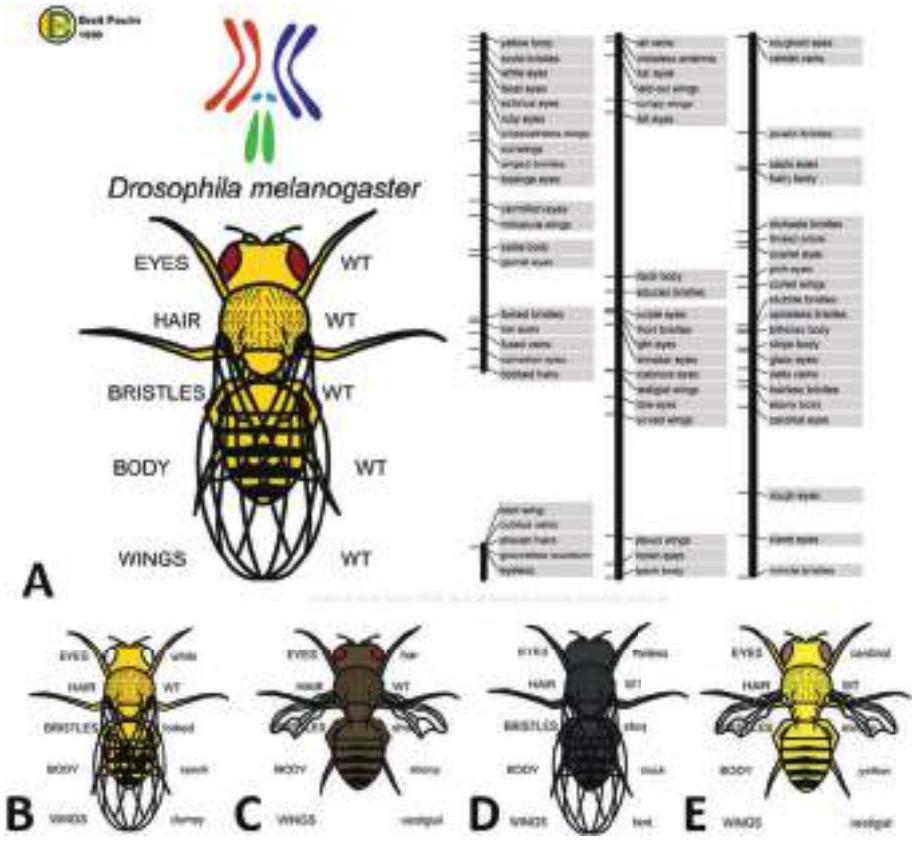


Figura 2. Teoria cromossômica e influência direta dos genes na determinação fenotípica das espécies. (A) Modelo interativo da influência dos genes na determinação fenotípica e caracterização do mapa físico e de ligação dos genes de drosófila distribuídos nos quatro cromossomos pertencentes à espécie. Este modelo está disponível em http://www.biology.ualberta.ca/facilities/multimedia/uploads/testing/fly.html, e foi elaborado por Brett Poulin da Universidade de Alberta na Califórnia e publicado pela Bio-DiTRL. Por ser interativo, o modelo permite que o usuário induza modificações no fenótipo da mosca modelo, selecionando algumas das características genéticas referenciadas em cada um dos cromossomos, alterando consequentemente o status do modelo selvagem (do Inglês *Wild type* – WT). Cinco são as características que podem ser modificadas pelo usuário, a saber: olhos, pelos, cerdas, corpo e asas. (B a E) Destacam alguns dos modelos gerados com a ferramenta. Compare em especial cor, formato dos olhos e das asas.

envolveram a interferência das mutações nas características físicas e funcionais dos organismos pudessem ser desvendados. Então, inúmeras outras mutações em genes de drosófilas foram identificadas em estudos específicos e posteriores, de tal forma que atualmente se conhece praticamente todo o mapeamento físico de localização destes genes nos cromossomos (Figura 2), que foi sucedido pelo então sequenciamento genômico completo deste modelo de estudo. De fato, pelo seu trabalho com drosófilas, Morgan recebeu o prêmio Nobel em 1933.

Embora Morgan tivesse notoriedade com sua pesquisa e colocado a Genética em destaque na comunidade científica, muito do conhecimento decorrente de seu trabalho evoluiu e, em 1940, outro importante achado científico teve sua importância. George Beadle e Edward Tatum estudando o fungo *Neurospora crassa* (Ascomycota) elaboraram a teoria que relacionava diretamente uma proteína produzida a um gene codificador (Beadle & Tatum, 1941). Embora esta teoria atualmente sofra controvérsias notórias, em detrimento das descobertas de complexos heteroprotéicos, RNAs de interferência e RNAs não codificadores, a mesma influenciou muito os avanços acerca do conhecimento da função gênica bem como função de seu produto (proteínas) no metabolismo celular.

O ponto alto deste estudo desencadeado por Beadle e Tatum foi devido à escolha do modelo biológico de estudo (Figura 3). *Neurospora* apresenta um genoma haplóide o que facilitou o estudo do efeito de genes mutados, já que não tem uma cópia que complementa fenótipo como observado em casos de heterozigose. Esta dupla de

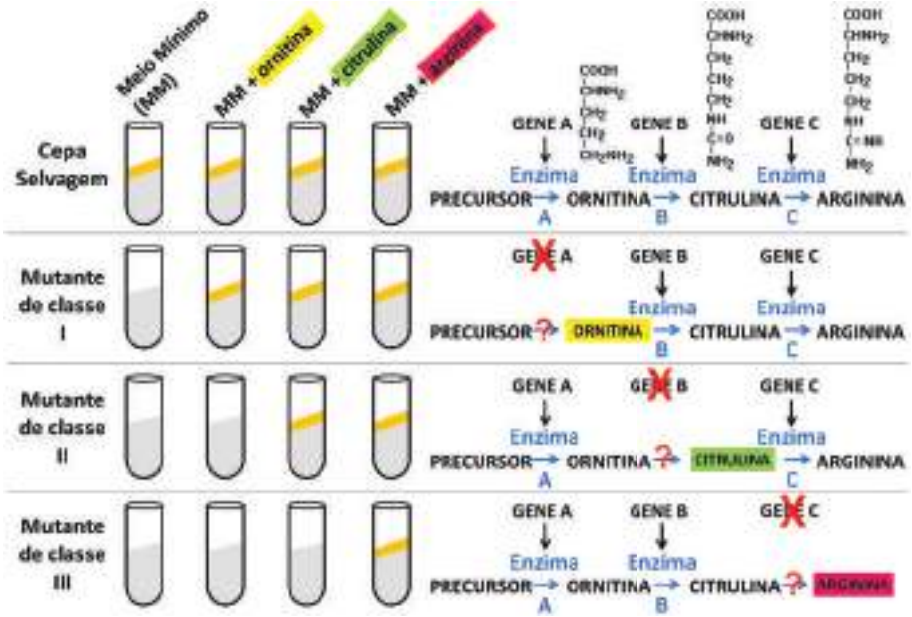


Figura 3. Caracterização dos experimentos propostos por Beadle e Tatum que permitir identificar a relação teórica "um gene uma proteína". Os tubos de ensaio retratam a presença de um meio de cultura (cinza) que pode propiciar ou não o crescimento de *Neurospora* (amarelo). Observe que na presença de um gene mutado não há crescimento de *Neurospora*, salvo quando o meio é suplementado com o produto da reação enzimática catalisada pela enzima codificada pelo suposto gene mutado.

pesquisadores então produziu três cepas mutantes, cada uma delas com a perda de função de um gene, mas que curiosamente todos os três participavam de uma mesma via metabólica, a via de biosíntese de arginina, um aminoácido essencial ao metabolismo de *Neurospora*.

Enquanto cepas selvagens cresciam bem em meios de cultura mínimos, acrescidos ou não de suplementos (aminoácidos ou precursores de aminoácidos), cepas mutadas só cresciam na presença de substratos específicos, em especial aqueles substratos que faltavam à célula. Isto ocorria já que perderam a função de sintetizá-los em detrimento da consequente perde de função do gene mutado.

Mesmo diante destes resultados sensacionais, produzir mutações não era uma tarefa fácil já que todos os agentes mutacionais usados (UV preferencialmente) para tal ocasião poderiam mutar outros genes que não só de interesse de estudo, prejudicando o andamento dos ensaios experimentais. A resolução pra este impasse veio com o tempo e a evolução do conhecimento científico como retratato abaixo.

Como induzir mutações para verificar perda de função gênica?

A grande ferramenta em biologia molecular descoberta para este fim (indução de mutações) ocorreu graças aos estudos desenvolvidos por Barbara McClintock que estudou os transposons ou sequência de inserção durante duas décadas (1940 e 1950). McClintock observou que a presença de transposons em sítios específicos do cromossomo, mais especificamente em genes que determinam coloração das sementes de milho (genes Bz), alterava o fenótipo observado, que poderia ser revertido caso esta mesma sequência viesse a se retirar, sozinha, do local prévio de inserção (Figura 4). Por isso a denominação de elementos saltadores ou transposons (McClintock, 1951). Barbara recebeu o Nobel de Fisiologia e Medicina de 1983, por esta descoberta.

Foi só a partir deste conhecimento que se tornou factível criar organismos, fazendo então o uso de elementos de transposição como indutores de mutação. Ainda assim, o grande problema que ainda persistia era como induzir mutações em genes específicos de forma rápida e em larga escala para acompanhar os avanços genômicos. A proposta deste capítulo é focar esta abordagem.

Tipos de bibliotecas de mutantes

Em genômica, embora os genes sejam anotados com base na homologia de sequências, apenas uma pequena parcela acaba tendo função validada experimentalmente. Uma das maneiras mais corriqueiramente utilizadas para verificar a função destes supostos genes anotados está condicionada à anulação de sua função por indução de mutação.

Entretanto, uma parte dos genes que compõem os genomas de todos os organismos esta possibilidade é inviabilizada. Isto porque alguns destes genes são tão fundamentais à biologia dos organismos que sua ausência seria incompatível com a sobrevivência do mesmo, independentemente da condição ao qual é exposto. Estas são consideradas mutações em genes letais. Em outros casos, genes mutados levam estes organismos a sobreviverem em condições específicas, e uma vez colocadas em condições adversas, os mesmos veem a morrer ou simplesmente não se desenvolvem. Estes são consideradas

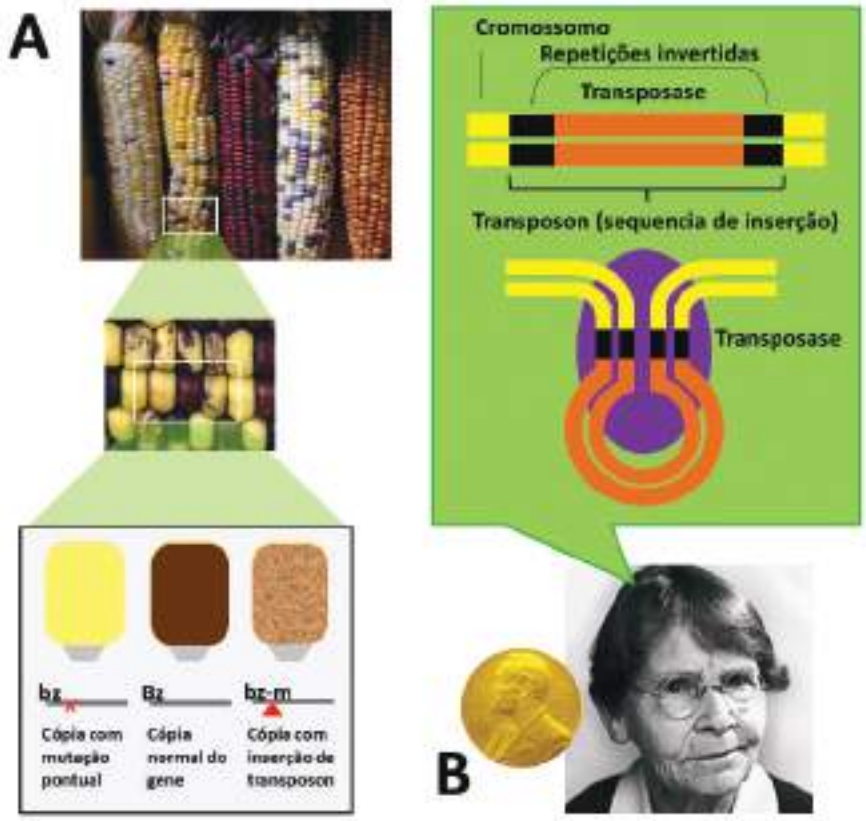


Figura 4. A) Efeito do elemento de transposição no fenótipo "cor da semente" em milho. Observe que as três cores mais frequentes na espiga de milho selecionada apresentam um genótipo diferenciado. A semente marrom carrega a cópia íntegra do gene Bz. A semente amarela carrega a cópia mutada do gene, então denominado bz. E a semente variegada carrega a cópia recessiva com a presença do elemento de inserção. B) Imagem de Barbara McClintock, autora das descobertas sobre elementos de transposição, descrevendo figurativamente a composição e efeito genético de um transposon.

mutações condicionais. Para o restante dos genes a mutação propicia a alteração de um fenótipo ou condição metabólica.

Existem basicamente duas tecnologias para indução de mutação em um gene, as chamadas mutações aleatórias, induzidas por elementos transponíveis inseridos em vetores específicos, ou por recombinação homóloga, que por convenção se costuma usar o termo *knockout*.

Mutações aleatórias usando cassetes de inserção com elementos de transposição

O primeiro passo, em qualquer experimento de indução de mutação está na construção do vetor de mutação. Na maioria dos casos se utiliza plasmídeos com estruturas especializadas. Quase sempre estes plasmídeos possuem um gene repórter de

resistência a um antibiótico, utilizado como ferramenta de seleção, uma origem de replicação, embora não seja funcional, já que o plasmídeo apresenta característica suicida, e uma região de múltiplo sítio de clonagem, local onde o cassete de transposição será inserido.

É importante esclarecer que a origem de replicação em grande parte dos plasmídeos permite que o mesmo atue como um episssomo, ou seja, se replique independentemente da replicação bacteriana. Assim, uma vez que uma célula competente é transformada, este plasmídeo pode se replicar inúmeras vezes, permitindo ao pesquisador isolar tais estruturas por técnicas especializadas, aumentando assim o número de cópias plasmidiais a serem usadas experimentalmente (Figura 5A). Como trata-se de um vetor suicida, o mesmo não pode se replicar, caso contrário poderia induzir inúmeras mutações em um único organismo o que dificultaria as análises biológicas e prejudicaria

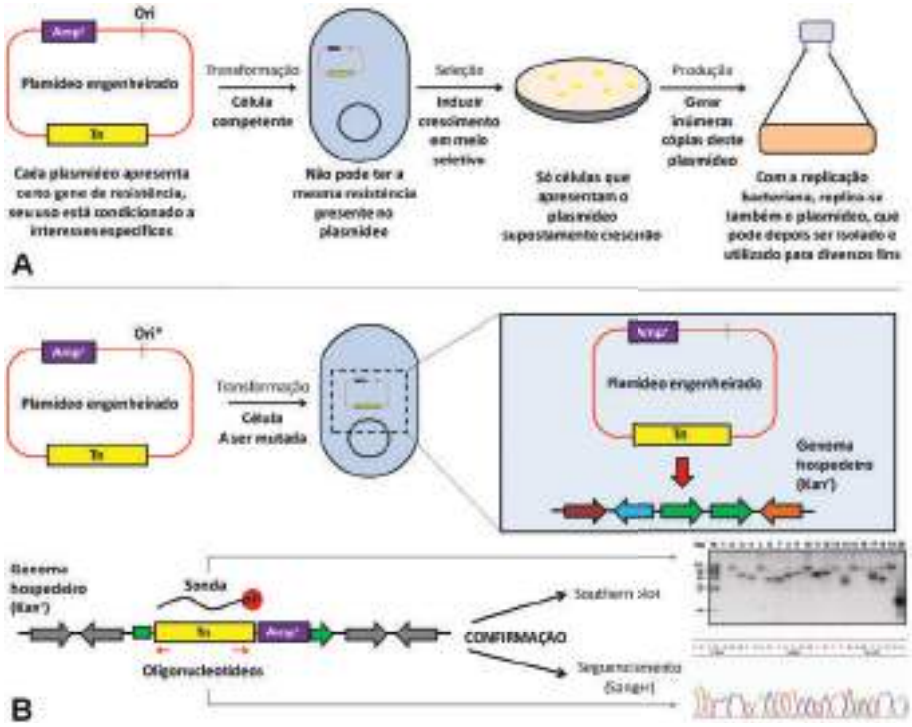


Figura 5. Etapas envolvidas na geração de mutantes por transposição. (A) Etapas preparatórias (transformação, seleção e produção) que antecedem a integração do plasmídeo ao DNA da célula receptora. (B) Uma vez averiguado que plasmídeo adentrou a célula, naturalmente este plasmídeo irá se integrar ao DNA desta de forma aleatório, provocando, obrigatoriamente, uma mutação no local de inserção. Este local de inserção poderá ser uma região codificante ou não, conferindo a este organismo perda de um possível gene funcional e ganho de uma resistência adicional (neste caso à kanamicina). A partir da inserção duas análises confirmatórias são necessárias: se faz um sequenciamento a partir das pontas do transposon direcionando os oligonucleotídeos para as bordas deste elemento, isto permitirá saber com precisão que região ou gene foi mutado por homologia de sequências; ou análise por *Southern bloting*, usando uma sonda que se liga ao vetor, permitindo identificar quantos plasmídeos adentraram e integraram o DNA celular, exigindo que o resultado demonstre apenas uma banca no gel (ver Figura 6).

mais severamente a fisiologia do organismo. Portanto, após a transformação (entrada do vetor de clonagem na célula hospedeira), espera-se que a integração plasmidial ao DNA hospedeiro ocorra rapidamente (em decorrência da presença do transposon), e a cada replicação celular a mutação induzida por inserção é passada às gerações (Figura 5B).

Por se tratar de uma inserção aleatória, não há como controlar onde ocorrerá a mutação. A mesma pode ocorrer em regiões intergênicas, que podem ou não ter função regulatória, ou em genes específicos. Logo é fundamental desenvolver ferramentas para validar os processos bem como saber que região genômica foi realmente modificada. Como a transformação vem a ocorrer por alteração da condição salina do ambiente ou mediante choque elétrico que aumenta permeabilidade de membrana, não se pode controlar quantos plasmídeos adentraram a célula hospedeira, portanto, um controle experimental nesta ocasião torna-se fundamental.

Destaca-se o fato de que quanto maior a sequência do plasmídeo menor é a taxa de transformação. Portanto, a escolha do plasmídeo e do tamanho do inserto (cassete de transposição) também deve se levar em conta no desenho experimental.

A melhor forma de se verificar a inserção do plasmídeo ao DNA hospedeiro ocorrer por intermédio da técnica de *Southern bloting*. Para isso utiliza-se uma sonda radioativa que anela à sequência do elemento de transposição. Logo, se apenas um plasmídeo adentrou a célula durante evento de transformação, apenas uma banda será notada no gel de agarose. No entanto, se o número de plasmídeos for superior serão observados números proporcionais de bandas (Figura 6). Vale ressaltar que para saber se uma célula é ou não transformante, basta crescê-las em meio de cultura seletivo, ou seja, que apresente o antibiótico cujo gene repórter codifica a enzima de resistência.

Uma vez determinado que um único plasmídeo adentrou a célula e integrou o DNA hospedeiro, o próximo passo é descobrir efetivamente que região foi mutada pelo

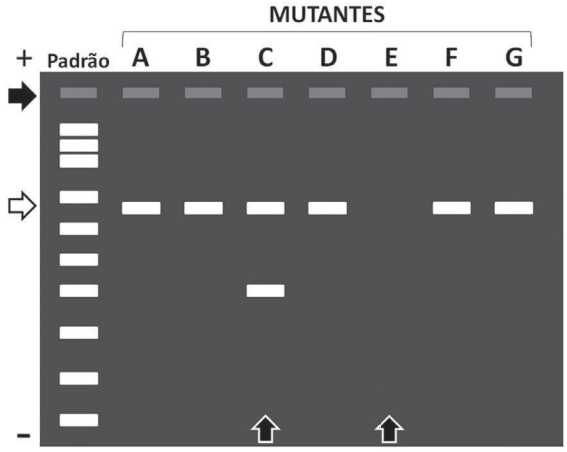


Figura 6. Modelo de validação de transformantes baseado na análise de *Southern bloting* mediante uso de sonda específica. A seta preta indica a origem de aplicação das amostras. A seta branca o resultado esperado da inserção do cassete decorrente de uma única entrada plasmidial durante vento de transformação. Setas prestas com bordas brancas determinam eventos não esperados de transformação, respectivamente uma dupla entrada plasmidial e a seleção de um organismo resistente ao antibiótico de seleção, mas sem o cassete de transposição.

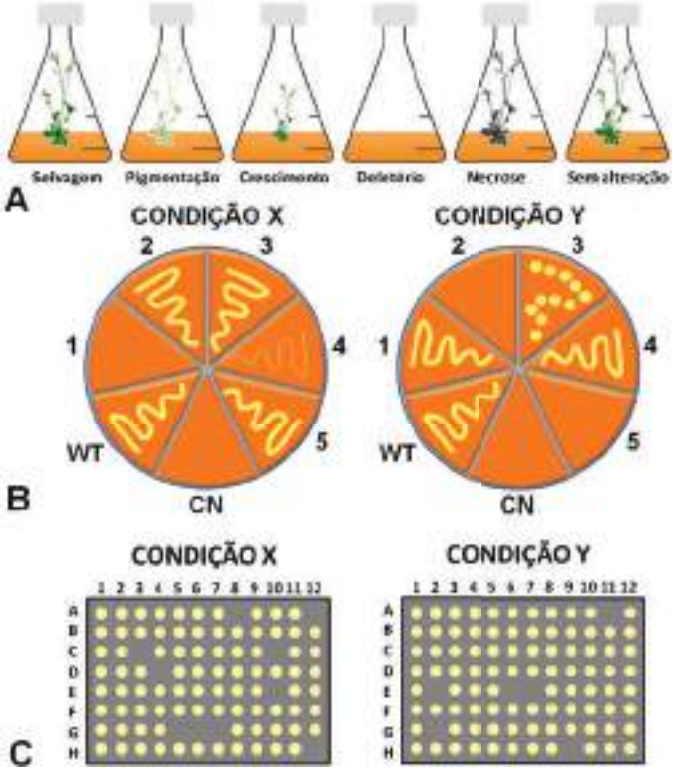


Figura 7. Ensaios para seleção de mutantes de interesse gerados aleatoriamente. (A) Modelo hipotético de seleção de plantas contendo genes modificados que alteraram o perfil de crescimento em condição padrão. Quatro distintos fenótipos são apresentados, demonstrando que os genes mutados possam ter alguma relação com produção de pigmentos, associados ao crescimento, essencial para que a planta cresça na condição avaliada e capaz de gerar necrose. (B) Seleção de mutantes bacterianos em placas de Petri ou em (C) placas de 96 poços. Em Placas de Petri a mudança nos perfis das colônias (coloração, formato ou taxa de crescimento) são indicativos de alterações. Em placas de 96 poços o não crescimento dos mutantes em uma dada condição X ou Y é indicativo da função do respectivo gene mutando em uma via metabólica essencial para a bactéria.

elemento de inserção. Para isso basta que se faça um sequenciamento desenhando-se oligonucleotídeos iniciadores para uma reação de PCR cuja ancoragem ocorra na ponta do vetor de clonagem, apontando obrigatoriamente para fora da sequência de inserção. Ou seja, quando os oligonucleotídeos anelarem-se à região do vetor os mesmos se direcionarão no sentido da região mutada. Após obtenção do cromatograma e sequência FASTA, basta por homologia de sequências ferramentas determinar se a região modificada trata-se de um gene ou região intergênica.

Acompanhando o aumento no número de genomas sequenciados, veio a necessidade de se desenvolver plataformas de análise de mutação de genes e larga escala, uma excelente metodologia para a compreensão empírica da funcionalidade de genes supostamente anotados como hipotéticos ou mesmo de regiões regulatórios.

Geralmente para focar um estudo direcionado a este conjunto de mutantes gerados aleatoriamente, ensaios exploratório são necessários. Todos estes mutantes

de um modelo são submetidos a diferentes condições de crescimento, de acordo com o interesse da pesquisa, e aqueles que demonstrarem alteração fenotípica são identificados e analisados com maior precisão. Tomemos os exemplos apresentados na Figura 7 como referência.

Nesta figura três modelos distintos de seleção dos mutantes de interesse são apresentados. Em A pode-se observar a seleção dos mutantes de plantas (*Arabidopsis*) com base na alteração do perfil de crescimento e/ou mudança em sua coloração. Para isso, sempre se usa um controle positivo, representado pela planta selvagem. Em B observa-se o perfil de seleção de mutantes de bactérias em placas de Petri. Da mesma forma que em *Arabidopsis*, tem-se interesse nos mutantes que apresentaram alguma modificação fenotípica em relação a cepa selvagem (WT). Neste modelo de experimento é interessante incluir um controle negativo (CN), representado por um espaço na placa sem bactéria. Certamente nada poderá crescer neste espaço. Em C destaca-se um método para seleção de mutantes também de bactérias, porém usando placas de 96 poços e diferentes condições analisadas.

Mutações regiões específicas usando cassetes de recombinação homóloga

Mesmo sendo possível gerar uma infinidade de mutantes por cassetes de mutação aleatória, muitas vezes um gene de interesse pode não aparecer na biblioteca. Por esta razão é importante ter um sistema de indução de mutação que altere a característica de um gene em específico. A principal ferramenta hoje existente para esta finalidade é a recombinação homóloga.

Para se fazer uma recombinação homóloga também se utiliza um plasmídeo, porém o cassete de recombinação deverá apresentar diferentes configurações: a) uma sequência de nucleotídeos que se pareça muito com o gene de intesse, mas que possua algumas alterações que não permita que a respectiva proteína possa ser sintetizada sem sua forma nativa; b) contenha no inserto um gene repórter com seu respectivo promotor funcional. Em qualquer uma das situações, para que ocorra uma recombinação, ou troca desta sequência não codificadora ou codificadora de um gene repórter, pela sequência codificadora do gene de interesse no genoma, é necessário incluir sítios de repetições invertidas flanqueando a sequência de interesse. Desta forma, uma vez que o plasmídeo já está no interior da célula e o que ocorra um pareamento destas regiões invertidas do plasmídeo para com as sequências do genoma (junção de Holliday), a própria maquinaria de recombinação e reparo da célula ficará responsável pela troca.

Uma vez que a troca possa ter ocorrido, conferir se tudo funcionou perfeitamente é fundamental antes de promover os ensaios de indução de variação fenotípica. Para isso diferentes abordagens podem ser utilizadas. Para o caso de troca entre uma sequência codificadora por uma não codificadora se aconselha fazer o seqüenciamento da região de interesse tentando enocntrar no genoma a cópia não codificante do gene. *Southern bloting* também é uma boa forma de confirmar o número de plasmídeos que adentraram, de forma similar ao já descrito acima para o caso de mutações aleatórias. Sequenciamento também poderia ser feito para o cassete contendo o

gene repórter, entretanto, para este, é mais fácil se verificar testando a função deste gene. Caso seja uma proteína verde fluorescente, ou GFP, na sigla em inglês, iremos observar o transformante recombinado na cor verde, caso seja uma proteína repórter do tipo luciferase (enzima), veremos a amostra transformada e recombinada brilhar na presença do substrato luciferina, e se for uma proteína repórter que confere resistência a um antibiótico, então o transformante recombinado crescerá em meio seletivo (Figura 8).

Observando os modelos apresentados nas Figuras 6 (mutação aleatória) e 8 (recombinação homóloga), bem como as explicações sumarizadas associadas a cada um deste exemplos, tudo parece muito simples. Quando o modelo de estudo trata-se de um procarioto podemos dizer que o processo é mais simples, porém se o modelo de estudo é eucarioto com genoma diplóide a situação se complica bastante, uma vez que para se ter um fenótipo preciso da perda do gene os pesquisadores terão de ter o mutante

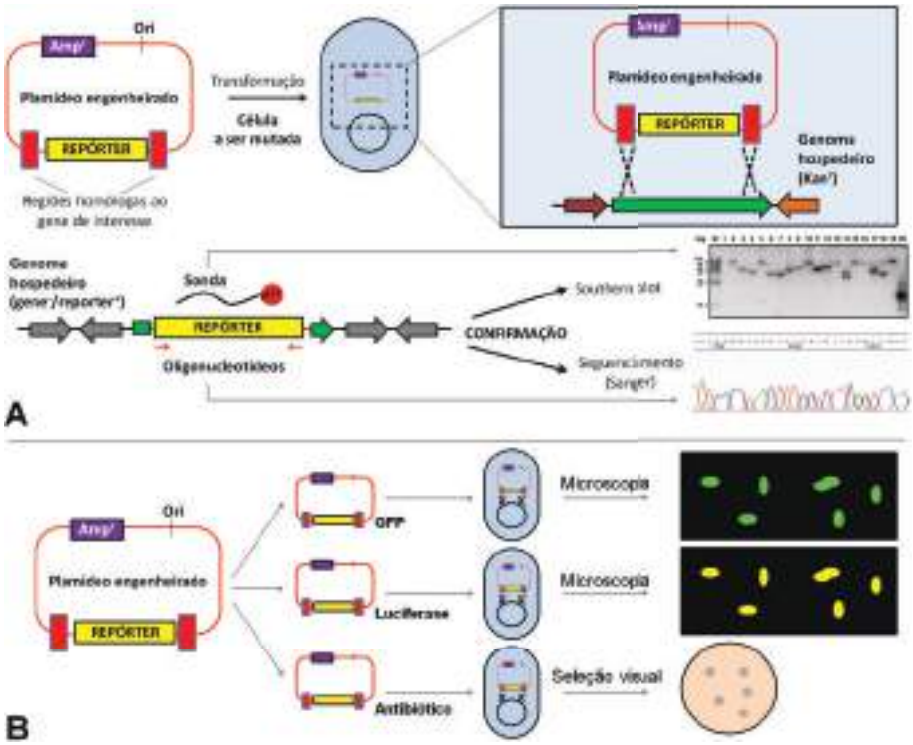


Figura 8. Etapas envolvidas na geração de mutantes por recombinação homóloga. Para que se entenda as perspectivas apresentadas em A e B considere a explicação associada à imagem A da Figura 6. (A) Uma vez averiguado que plasmídeo adentrou a célula, naturalmente este plasmídeo irá se integrar ao DNA desta porém em um gene específico usando para isso o pareamento das regiões repetitivas que flanqueiam o cassete de inserção, formando uma estrutura de junção Holliday. Por ação da maquinaria de recombinação e reparo, as cópias são trocadas e tem-se início ao processo de checagem da recombinação, similar ao item B da Figura 6, por sequenciamento. Os oligonucleotídeos estão apontados para dentro do cassete de recombinação. (B) Etapas de checagem da recombinação a partir do uso de genes repórteres. Três modelos são apresentados, conforme descrito no texto: recombinação com uso de GFP, Luciferase ou resistência a um antibiótico.

nulo. Ou seja, um organismo que tenha as duas cópias em cromossomos homólogos modificadas, caso contrário, haverá um sistema de compensação da cópia única mutada ou uma redução no perfil fenotípico que não permitirá a seleção com precisão.

A Figura 9 mostra um conjunto de imagens que destacam o perfil de alteração fenotípica em vários modelos que permitem compreender um pouco melhor o que acima foi descrito.

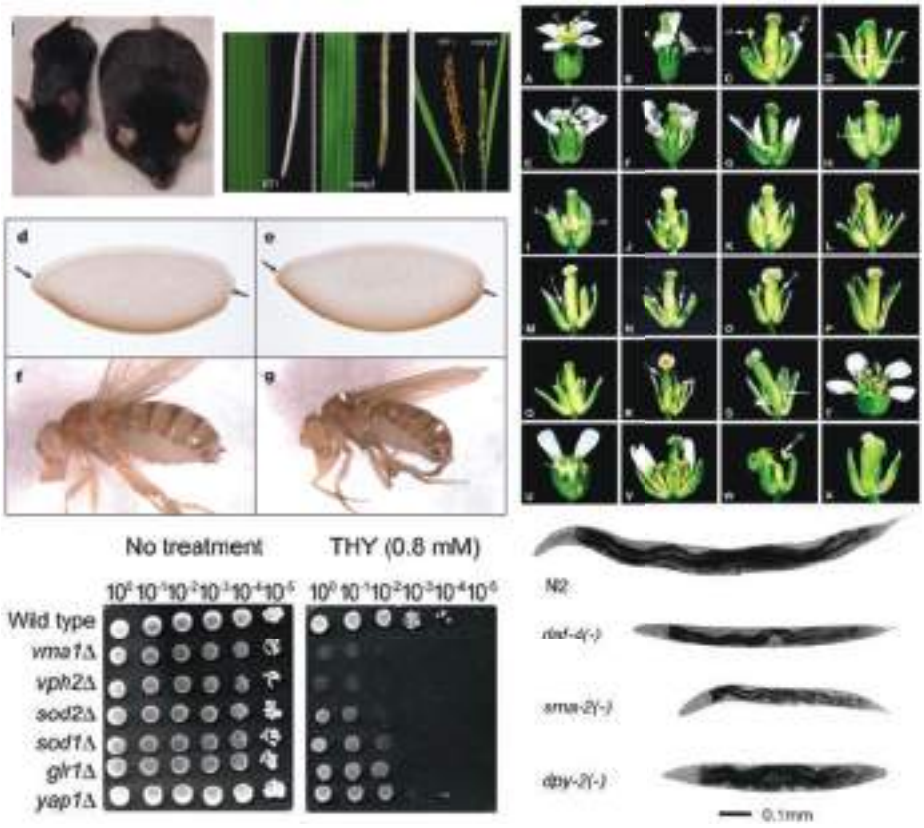


Figura 9. Exemplos de mutantes que apresentaram notória alteração fenotípica em diferentes modelos experimentais. Detalhes envolvendo a função biológica de cada gene mutado nos exemplos apresentados podem ser melhor compreendidos a partir dos artigos originais dispostos nas URLs abaixo.

- Mutantes de flores de *Arabidopsis thaliana*.(http://dev.biologists.org/content/128/14/2735.figures-only)
- Mutante para o gene *Mc4R* relacionado com obesidade em camundongo. (http://www.informatics.jax.org/allele/MGI:3723065)
- Mutantes de *C.elegans* envolvidos com alterações morfológicas. (http://www.pnas.org/content/97/10/5285.figures-only)
- Mutante para o gene *Nramp* relacionados com aquisição de manganês em plantas de arroz. (http://www.nature.com/ncomms/2013/130912/ncomms3442/full/ncomms3442.html)
- Mutante para o gene *wntD* em Drosophila associado com ativação dorsal ectópica. (http://www.nature.com/nature/journal/v437/n7059/fig_tab/nature04073_F3.html)
- Mutantes associados com sensibilização a estresse redox em *Aspergillus fumigatus* (http://journal.frontiersin.org/Journal/10.3389/fmicb.2012.00088/full)

Geração de um duplo mutante em organismos diploides

Em organismos procariotos, por apresentarem um genoma haploide, basta uma única mutação em um gene de interesse para se possa verificar a implicação deste na biologia do organismo. Neste caso desconsiderando a possibilidade de haverem genes parálogos funcionais. Em eucariotos, pelo fato de possuírem genomas diploides uma simples mutação faria com que o organismo saísse de uma condição de homozigose dominante para heterozigose, e a alteração fenotípica desta mutação poderá não ser observada em decorrência de um possível mecanismo de compensação gerado pelo alelo funcional. Diante desta condição, para se ter certeza absoluta da importância funcional deste par de genes o chamado duplo mutante deverá ser construído. Entretanto, esta condição não é tão simples de se obter, embora metodologias diferenciadas tenham sido desenvolvidas para alguns modelos biológicos.

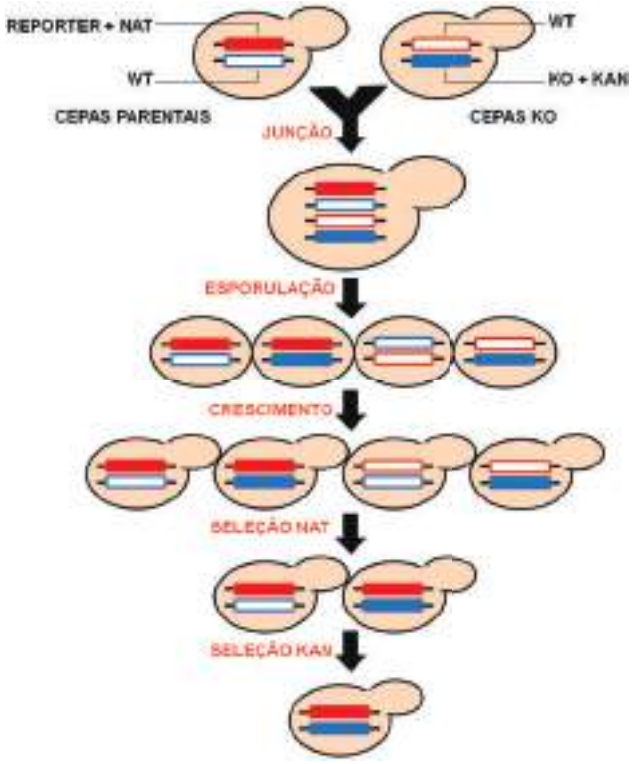


Figura 10. Modelo de formação de mutantes nulos em *Saccharomyces cerevisae*. Dois mutantes independentes são gerados para um mesmo gene de interesse, um contendo o gene de resistência a KAN (kanamicina) e o outro a NAT (nourseothricin). Ambos os mutantes são colocados juntos em um meio de cultura para induzir junção das estruturas celulares e, consequentemente, formação de um tetraploide. Em seguida a célula tetraploide é submetida em uma condição que induz esporulação que, por meiose, permite a formação de quatro possíveis composições entre o gene mutado e a condição selvagem. Uma primeira seleção destas linhagens de esporos é selecionada em meio contendo NAT, e posteriomente em um meio contendo KAN. Apenas as células contendo os dois cassetes de resistência permanecerão ativos após todo o processo, gerando desta forma um mutante nulo.

Para demonstrar a dificuldade de se gerar um duplo mutante em organismos que apresentam um genoma diploide, será exemplificado um processo simplificado deste a partir de uma metodologia experimental interessante e prática usada em *Saccharomyces cerevisae* (Figura 10). Esta metodologia se baseia na criação de um duplo mutante fazendo uso da própria maquinaria genética do organismo. A perspectiva biológica e o resultado final podem ser extrapolados a outros eucariotos que apresentam reprodução sexuada.

Dois mutantes independentes para um mesmo gene são gerados, fazendo uso da metodologia de recombinação homóloga, por exemplo. Cada um destes mutantes carrega consigo um gene repórter diferenciado, por exemplo NAT (resistência a nourseothricin) e KAN (resistência a Kanamicina). Para selecionar cada um destes mutantes de interesse, basta induzi-los a crescerem em meios seletivos, contendo estes antibióticos. Os isolados que crescerem nestas condições apresentarão o respectivo cassete de resistência. Dando continuidade, caso ambos sejam colocados juntos em uma mesma condição de crescimento, ocorrerá a fusão de seus materiais celulares que, por consequência, induzirá uma mistura de ácidos nucléicos, acarretando a formação de um tetraploide híbrido, momentaneamente. Este organismo tetraploide sofrerá esporulação, podendo gerar quatro combinações dos *loci* para um mesmo gene. Três destas combinações não apresentarão interesse, já que refletem o genótipo selvagem, ou os heterozigotos para cada uma das mutações (NAT+ ou KAN+). Entretanto, uma das combinações carregará os dois genes mutados, e isso poderá ser observado submetendo os esporos a duas condições seletivas seriadas. Como mostra o desenho da Figura 9, primeiro uma seleção baseada na resistência a NAT seguida de uma seleção baseadas na resistência a KAN.

A seguir, é só submeter o organismo na condição fisiológica de interesse e verificar o que acontecerá. Vale lembrar que em alguns mutantes o efeito, ou os efeitos observados, podem ser frutos de um fenômeno denominado pleiotropia. Em outras palavras, o gene está relacionado com tantas funções biológicas que é possível verificar alteração fenotípica sob diversas condições as quais o mutante é submetido. Além disso, mutações denominadas polares também podem interferir no entendimento da função biológica do gene mutado, ver adiante.

Mutações polares

Denominam-se mutações polares todas e quaisquer mutações induzidas que não afetem o gene pretendido, mas genes regulados pela região genômica mutada, praticamente genes que estejam a jusante do sítio de mutação, em qualquer uma das fitas do DNA.

Para tentar esclarecer um pouco melhor este efeito, usaremos como exemplo a Figura 11. Dez são os possíveis genes nos quais o cassete de transposição gerado poderá se inserir (nomeados de *a* a *j*). Não podemos descartar a possibilidade de que o cassete seja inserido fora dos genes, mas em putativas regiões regulatórias, como é o caso demosntrado em X (vermelho). Mais complexo ainda é a situação representada pelo possível sítio de mutação destaca por Y (em azul). Cada um destes casos será descrito com detalhes abaixo:

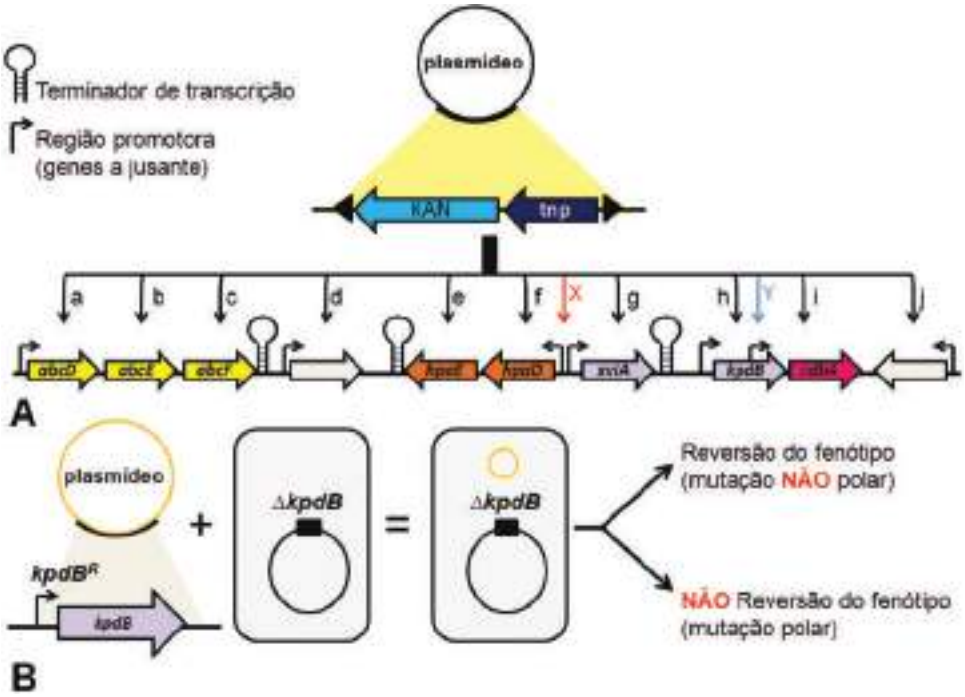


Figura 11. Diagnosticando se uma mutação é polar. (A) As setas no genoma representam genes, cujas cores estão associadas a uma mesma função biológica. Há genes únicos regulados por uma única região promotora, e há também dois ou três genes controlados por uma única região promotora, os denominados operons. De *a* a *j* são apresentados possíveis sítios de inserção do cassete de transposição, afetando cada um dos genes apresnetados nesta fração do genoma modelo. X denota uma mutação em região regulatória e Y uma mutação em região codificante e ao mesmo tempo regulatória. (B) Ensaio para constatar se uma mutação é ou não polar. Um plasmídeo deverá ser elaborado contendo o gene modificado em sua versão íntegra e funcional bem como sua região regulatória também funcional. Este plasmídeo deverá ser inserido no organismo mutado (transformação). Este plasmídeo não tem perfil integrativo, portanto permanecerá como uma unidade de DNA independente do genoma do organismo. Caso a maquinaria celular reconheça o inserto e permita a correta transcrição do gene íntegro, a proteína codificada pelo gene desencadeará sua função e o suposto fenótipo alterado será revertido na condição avaliada, demonstrando que a mutação não é polar. Caso o fenótipo alterado se mantenha, isso significa dizer que este é decorrente da perda de função do gene cdbA (Figura A), uma vez que não foi possível ocorrer a transcrição em trans deste gene.

a. Mutação em um gene não contido em operon (*d,g,h,i* e *j*).
   Muito provavelmente estas mutações não se caracterizarão como polar, já que se inseriram em uma região genômica que sabidamente está associada a um único gene e que também não apresneta qualquer perfil de região regulatória.
b. Mutação em genes que compõem um operon (*a-c* e *e-f*).
   Embora todas estas mutações possam afetar um sistema ou via metabólica aos quais os genes do operon estão relacionados, mutações nos genes que ocupam as primeiras posições do operon tenderão a serem mais prejudicias do que as que modificam a estrutura dos últimos genes. Isto porque a mutação no primeiro gene acarretará prejuízo nos genes subsequentes. Logo o efeito da mudança fenotípica

poderá ser fruto indireto da perda de função de todos os genes do operon e não necessariamente só do gene mutado, portanto um efeito polar da mutação. Em contraposição, alteração fenotípica induzida por uma mutação no último gene de um operon certamente e resultado direto da falta deste gene.

c. A região modificada representa uma região regulatória (X).
   Para deixar a discussão mais integrada, X representa uma mutação em uma região genômica que regula dois genes com funções distintas e em sentidos opostos de transcrição. O efeito fenotípico desta mutação é decorrente da perda dos genes *hpaED* ou *xviA*? Portanto, esta também é considerada uma mutação polar.
d. Mutação em região regulatória codificante, mas que ao mesmo tempo atua como uma região regulatória de outro gene (Y).
   Este é sem dúvida um dos casos mais intrigantes para definir se uma mutação é polar ou não polar. A alteração fenotípica induzida por esta mutação poderá ser fruto do próprio gene modificado (*kdpB*), ou da perda de função do gene que está a jusante (*cdbA*). Isto porque o gene *cdbA* está sob regulação de um promotor que na prática ocupa a mesma posição genômica codificada pelo gene *kdpB*.

Diante de tudo o que foi apresentado sobre estes tipos de mutação, uma pergunta ainda fica pendente: Na prática, como descobrir se uma mutação é ou não polar? Para responder esta pergunta, uma prática experimental que não é simples deverá ser considerada. Devemos para isso fazer o que se chama complementação, que consiste em devolver ao genoma mutado uma sequência íntegra e funcional da região modificada, tentando verificar se o fenótipo previamente modificado volta ao normal (Figura 11B).

Bancos de dados integrativos

Com o intuito de promover integração entre o conhecimento gerado pelo sequenciamento genômico e pelas plataformas funcionais, bancos de dados integrativos têm sido criados. O objetivo é reunir de forma organizada e estruturada toda a informação relativa ao estudo da função de genes em organismos modelos, facilitando a análise e levantamento de informações biológicas pela comunidade científica internacional. Em muitas destas plataformas o depósito de informação pode ser feita por qualquer pessoa, desde que devidamente cadastrada, sob supervisão de uma curadoria de manutenção destes bancos, responsável pelo aval final sobre a informação ali apresentada. De qualquer forma, a responsabilidade pela autoria da informação é sempre do depositante.

Bancos de dados contendo informações de mutantes em organismos-específicos

Alguns bancos de dados relacionados a organismos específicos estão em uma fase avançada de organização destas informações genômicas. A Tabela 1 lista alguns destes bancos de importantes e bem estudados organismos modelos.

O objetivo secundário destes bancos é promover intercâmbio de informações e troca de experiências entre as equipes que se envolvem no estudo dos respectivos modelos biológicos.

Tabela 1. Lista de bancos de dados contendo informações de mutantes em alguns organismos modelos.

Banco	Modelo	Sitio na WWW
FlyBASE	*Drosophila melanogaster*	http://flybase.org/static_pages/allied-data/external_resources5.html
TAIR	*Arabidopsis thaliana*	http://www.arabidopsis.org/portals/mutants/stockcenters.jsp
CGSC	*Escherichia coli*	http://cgsc.biology.yale.edu/
SGDP	*Saccharomyces cerevisae*	http://www-sequence.stanford.edu/group/yeast_deletion_project/deletions3.html
MGI	*Mus musculos*	http://www.informatics.jax.org/
WormBase	*Caenorhabditis elegans*	http://www.wormbase.org/#01-23-6

Bancos de dados envolvendo mutações relacionadas com patologias-específicas

Embora menos frequente que os bancos de organismos específicos, alguns são especializados na compreensão de doenças específicas. Tomemos como exemplo o banco de dados que armazena informações sobre a síndrome da progeria (Síndrome de Werner, também chamada de velhice precoce, é uma doença que faz com que o portador envelheça 5 a 10 vezes mais rapidamente que o normal) retratado com detalhes na Figura 12. Neste banco de dados é possível obter informações sobre o gene envolvido com a caracterização desta doença, sua organização modular em éxons e íntrons, a posição de SNPs (polimorfismos de nucleotídeo único) já descritos envolvendo este gene, além das mutações mais frequentes relacionadas com substituição, deleção e inserção de nucleotídeos.

Exemplos envolvendo outras patologias podem ser encontrados no OMIM (*Online Mendelian Inheritance in Man*) mantido pelo NCBI (*National Center for Biotecnology Information*).

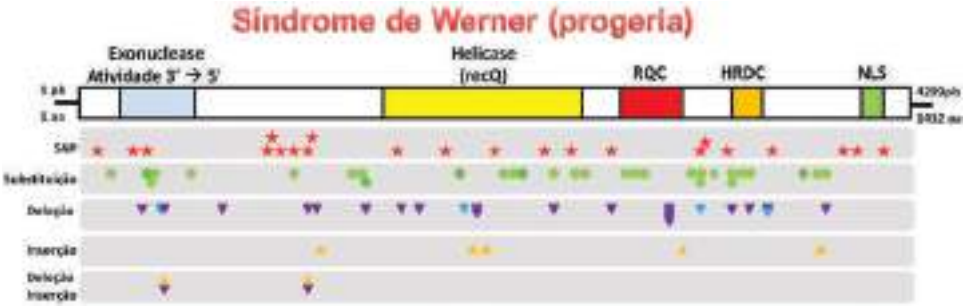


Figura 12. Esquema figurativo contendo todas as informações mutacionais no gene RECQL2, associado à síndrome de Werner ou síndrome da progeria. O gene contem 4299 pares de bases, codificando 5 éxons (cores). As linhas abaixo do esquema do gene destacam os tipos de mutação com suas respectivas representações ao longo do eixo deste gene. Maiores informações http://omim.org/entry/277700.

Considerações finais

O estudo da função de um gene baseado na indução de uma mutação e verificação de sua alteração fenotípica tem sido e será por muito tempo uma das metodologias mais empregadas nos mais diversos modelos experimentais. Daí a importância de se conhecer pelo menos a base fundamental das técnicas envolvidas com este propósito. Assim, fica a dica que muitos outros assuntos correlacionados e técnicas mais elaboradas estão disponíveis em literatura específica a quem queira se aprofundar.

Bibliografias

MORGAN, T. H. What are "factors" in Mendelian explanations? *American Breeders Association Reports*. 1909, 5, 365-368.

MORGAN, T. H. Sex-limited inheritance in Drosophila. *Science* 1910. 32, 120-122.

BEADLE, G. & E. TATUM.. Genetic control of biochemical reactions in Neurospora. *Proc. Nat. Acad. Sci*. 1941, 27: 499-506

MCCLINTOCK, B. Mutable loci in maize. Carnegie Institution of Washington Yearbook. 1951. 50, 174-181.

EPSTEIN CJ, MARTIN GM, SCHULTZ AL, MOTULSKY AG: Werner's syndrome a review of its symptomatology, natural history, pathologic features, genetics and relationship to the natural aging process. Medicine (Baltimore). 1966. 45: 177-221.

SNUSTAD, D. P., SIMMONS, M.J., and JENKINS, J.B., 1997, Principles of Genetics, John Wiley & Sons, Inc., New York.

8

Transcrissômica

Leandro Marcio Moreira
Renata Guerra de Sá Cota
Camila Carrião M. Garcia

Introdução

Durante grande parte do século XX, a biologia tentou explicar os fenômenos associados à manutenção da estrutura e função celular analisando o comportamento de moléculas de forma individualizada. Um bom exemplo disso veio através dos estudos de Gregor Mendel que começou a investigar variações de herança genética em plantas e que, posteriormente, foi referendado e validado por Thomas Hunt Morgan analisando a cor dos olhos de *Drosophila melanogaster* (Capítulo 7). Nestes estudos, os cientistas inferiram a existência de genes dispostos estrututuralmente ao longo do cromossomo, de forma organizada. Mais tarde, análises aprofundadas sobre a estrutura do DNA e a funcionalidade dos genes serviram de base para a proposta de criação do dogma fundamental da biologia molecular, descrito de forma sumarizada em 1970 por Francis Crick (Crick, F. 1970) (Figura 1). Esse dogma afirmava que "O *DNA sofre replicação dando origem a novas moléculas de DNA, depois é transcrito em RNA, e este por sua vez, traduz o código genético em proteínas*". Assim, a informação genética contida na molécula de DNA ao ser expressa na forma de proteínas produzirá um fenótipo que pode variar de acordo com o estado fisiológico, estímulos físicos, químicos e biológicos aos quais uma célula ou indivíduo é submetido/exposto. Entretanto, para que estas proteínas possam realmente desenvolver seus papéis biológicos, antes, moléculas de RNA mensageiros (mRNA) precisam ser sintetizadas a partir de um DNA molde, para que então a tradução protéica ocorra, e é sobre esta base conceitual que todo este capítulo focará seu estudo. Embora todos estes processos tenham sido sumarizados acima, hoje sabemos que tratam-se de eventos biológicos extremamente complexos, e muito bem orquestrado durante a biologia celular.

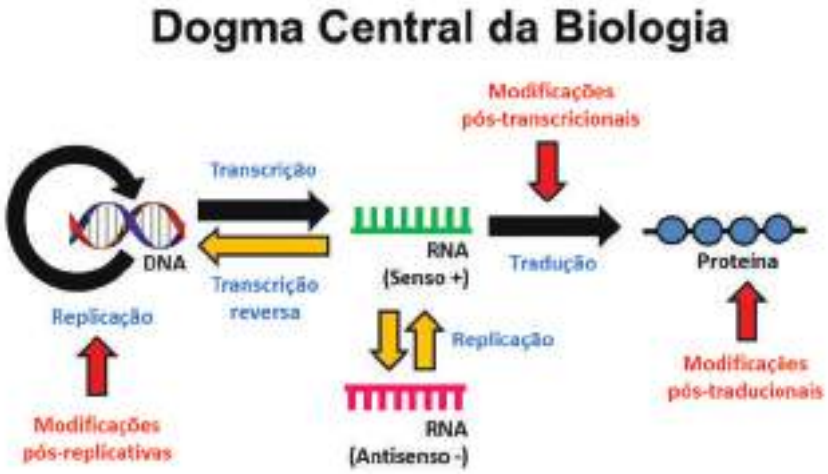


Figura 1. O Dogma central da Biologia. Este modelo, ainda de forma simplificada, foi proposto por Francis Crick em 1970. Porém algumas etapas deste dogma foram surgindo em decorrência dos avanços científicos, é o caso da transcrição reversa e da replicação de RNA. Atualmente não podem ser desconsideradas deste dogma as etapas envolvidas com modificações pós-replicacionais, transcricionais e traducionais, e envolvimento de moléculas de RNAs na regulação da biologia celular (não presentes neste modelo).

É importante ressaltar que algumas descobertas posteriores geraram questionamentos a este dogma, obrigando-o a sofrer adequações. Como exemplo, podemos citar o fato de que moléculas de RNA também são capazes de auto replicação ou mesmo servirem de molde para a síntese de moléculas de DNA mediante ação da enzima transcriptase reversa, encontrada nos retrovírus (Figura 1). Atualmente, a extensão do dogma central da biologia molecular exige que sejam representadas a importância de moléculas de RNA em praticamente todos os processos biológicos, justificada pela teoria do mundo de RNA (RNA world) (Gilbert, W, 1986). Além disso, existem mecanismos importantes para a manutenção da molécula de DNA (modificações pós-replicacionais, exemplo metilação e empacotamento), formação de um RNAm maduro em eucariotos (modificações pós-transcricionais, exemplo *splicing*, *splicing* alternativo e adição de CAP e poliA) e adequação e manutenção do estado nativo e funcional de algumas proteínas (modificações pós-traducionais, exemplo glicosilação, sulfatação, metilação, acetilação, fosforilação, dentre outras) que não podem ser deixados de lado quando uma discussão mais integrada deste dogma é colocado em pauta.

Assim, embora o dogma central desse uma perspectiva geral de como os sistemas biológicos funcionam ao nível molecular, foi só a partir da década de 90 do século passado, com o surgimento das ciências genômicas, que pesquisadores questionaram o fato de que as abordagens metodológicas da época (até então utilizadas para este ganho de conhecimento biológico) haviam atingido o limite na contribuição para o avanço da ciência. Isto porque os sistemas biológicos são extremamente complexos e têm propriedades emergentes que não podem ser explicados estudando suas partes individuais (DNA, RNA ou proteínas). Esta abordagem, denominada reducionista, embora bem sucedida nos primeiros períodos da biologia molecular, subestimava

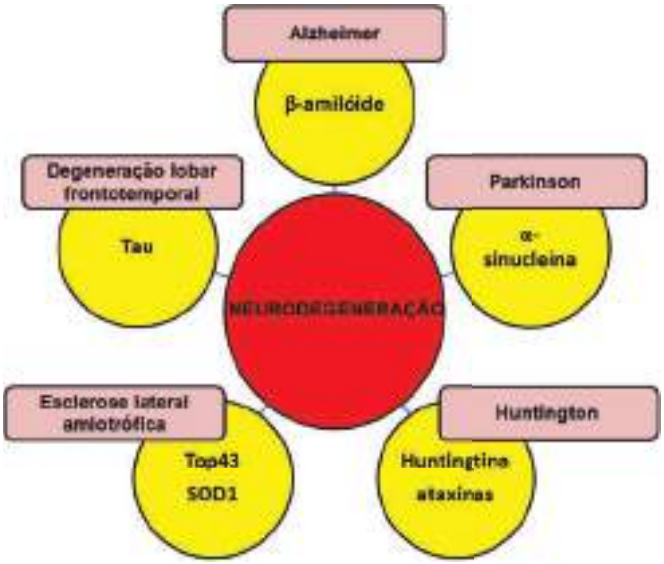


Figura 2. Função das proteínas envolvidas no desencadeamento e progressão de doenças neurodegenerativas. Para cada uma das cinco doenças apresentadas, uma ou mais proteínas foram identificadas como estando diretamente associadas á patologia. Fatores intrínsecos (genética) e fatores extrínsecos (condições ambientais e estilo de vida) não foram incorporados a este modelo, mas são de fundamental importância na caracterização da doença. O mau enovelamento ou a agregação destas proteínas leva o paciente a ter morte progressiva de neurônios caracterizando tais quadros patológicos com seus respectivos sintomas associados.

esta complexidade, chegando a ter certa influência negativa em muitas áreas da pesquisa. Por exemplo, no campo da biomedicina esta subestimação limitou durante algum tempo o entendimento sobre doenças neurodegenerativas e a descoberta de drogas para os seus tratamentos. Hoje sabe-se que tais doenças são multigênicas e multifatoriais, o que reflete as dificuldades para a descoberta de novos fármacos eficientes para o tratamento de tais patologias. Embora todas as principais doenças neurodegenerativas sejam decorrentes de alteração no perfil de proteínas com funções especializadas, os fatores que levam estas proteínas a perderem suas funções naturais, levando ao quadro patológico, são inúmeros e de relação bastante complexa (Figura 2).

Juntamente com a evolução do conhecimento científico e da expansão das pesquisas Genômicas no final do século passado, vieram também as necessidades de se compreender toda esta informação gerada em larga escala de maneira funcional. As primeiras abordagens envolvendo esta possível funcionalidade de forma integrada em uma célula ou tecido vieram com a criação da transcrissômica. E é sobre este conhecimento que este capítulo irá trabalhar os conhecimentos e evolução do pensamento científico.

Conceituando e destacando a importância dos transcrissomas

Denomina-se transcrissoma o conjunto completo de transcritos (mRNA, sRNA, tRNA e ncRNA) de um organismo, órgão, tecido ou linhagem celular. Como os mRNAs são

traduzidos em proteínas, foi estabelecido que o transcrissoma abrange o conjunto desta espécie de RNA e dos microRNAs. Esta última classe controla a expressão gênica ao nível pós-transcricional, bloqueando a tradução dos mRNA em proteínas, daí sua importância na fisiologia celular.

Esta visão em larga escala, dos transcritos gerados, permite interrogar uma informação genômica em pelo menos três perspectivas principais:

- Quais são os transcritos possíveis, incluindo todas as formas de processamento alternativo e RNA não codificador, de um organismo sob uma determinada condição fisiológica?
- Qual o modelo de expressão espaço-tempo dos transcritos e como eles variam entre os diferentes tecidos e sob diferentes condições sob as quais uma célula ou organismo possa ser exposto?
- Como e que mecanismos de expressão gênica são regulados numa dada condição fisiológica?

Tais abordagens são as mais amplamente investigadas pela comunidade científica que, de maneira geral, busca a resposta para uma pergunta ainda mais complexa: Como uma célula ou organismo se comporta sob o aspecto molecular frente às diferentes condições fisiológicas às quais podem ser expostas? É evidente que pelo grau de complexidade desta pergunta, nos limitaremos a tentar esclarecer a importância das três primeiras, uma vez que a resposta pra esta quarta está longe de ser compreendida pela comunidade científica. Até porque quando esta última pergunta for respondida, a ciência estará desprovida de quaisquer outras perguntas investigativas para um determinado modelo experimental. Isso é praticamente impossível de acontecer, e é isso que torna a Ciência tão instigante.

Com o objetivo de responder estas perguntas, uma série de técnicas foi desenvolvida ao longo dos últimos 60 anos na tentativa de se compreender melhor esta dinâmica de expressão gênica. Estas técnicas envolvem análise de um transcrito individual, codificado por gene de interesse, ou mais recentemente, a chamada expressão gênica em larga escala. Limitaremos-nos aqui a descrever estas técnicas de investigação global da expressão gênica, com algumas ponderações importantes envolvendo técnicas para análise de um conjunto limitado de genes.

É importante esclarecer que embora trabalhos em larga escala permitam entender o transcrissoma completo de um organismo ou célula numa determinada condição, muitas vezes a confirmação/validação dos resultados é obtida com algumas análises individuais de genes que foram detectados neste experimentos. Vários métodos específicos foram desenvolvidos e utilizados para dectectar e quantificar mRNA individuais. Destaque especial para as técnicas de Northern Blot e qRT-PCR (Saiki, RK *et al*., 1985), abaixo descritas.

Um breve histórico envolvendo a análise de expressão gênica

Os primeiros experimentos de transcrissoma foram realizados utilizando o método ou curva de Rot. Neste método, inicialmente é sintetizado uma cópia do RNA na forma de DNA complementar (cDNA), utilizando um precursor radioativo de modo

que as moléculas de cDNA resultantes são uniformemente radioativas. A seguir, as moléculas de cDNA marcadas são desnaturadas e hibridadas com o RNA total. As fitas complementares formarão híbridos. Neste método, as moléculas de RNA mais abundantes serão ligadas preferencialmente em relação às menos abundantes, de forma que a taxa de hibridação de uma molécula de cDNA será equivalente à abundância do seu RNA correspondente na amostra em estudo. Quanto mais abundante for uma espécie de RNA em particular, mais rapidamente irá se anelar ao seu complementar.

Esta metodologia também é conhecida como hibridação cDNA-mRNA ou hibridação cruzada e foi utilizada por Hastie & Bishop em 1976 para demonstrar que havia um conjunto de RNAs expressos em quantidades equivalentes em rins, fígado e cerebro de camundongos (Hastie, ND & Bishop, JO, 1976). Também foi demonstrado que algumas especies mais abundantes de RNAm encontrada no rim ou eram ausentes ou presente em níveis muito baixos no fígado (Figura 3). Este trabalho, além da originalidade, serviu de base para praticamente todos as outras propostas de estudo de transcrissômica hoje investigados pela comunidade científica, daí sua importância.

Já no final da década de 80 do século passado, a tecnologia de *Northern blot* foi então desenvolvida (Hayes, P.C. *et al*., 1989). Após a completa extração do RNA total

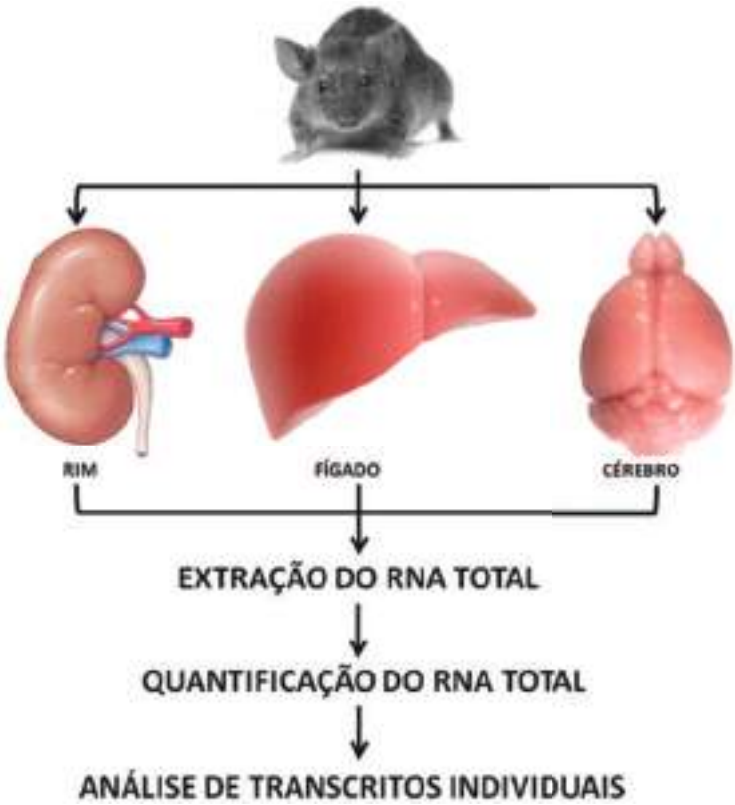


Figura 3. Sumarização da proposta dos trabalhos de Hastie e Bishop (1976) envolvendo análise diferencial de transcritos em três tecidos.

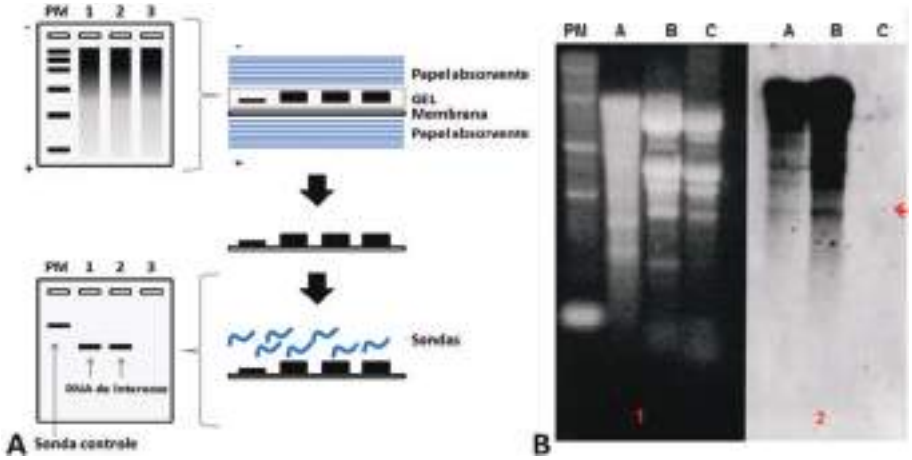


Figura 4. **A tecnologia de Northern Blot.** (A) Modelo simplificado explicitando como funciona a técnica de Northern Blot. (B) Modelo experimental de identificação de trasncritos alvos em RNAs totais: (1) Análise do RNA viral "A", RNA total extraído de planta infetada pelo *Southern bean mosaic* virus (SBMV-BSP) "B" e de planta sadia "C" em gel desnaturante de agarose 1%. PM - marcador RNA Ladder, Gibco-BRL (2) Northern blot de um gel similar ao (1) hibridado com sonda específica para o genoma do SBMV-BSP. A seta indica banda na qual a intensidade é proporcional à quantidade de RNAm. Imagem do resultado de Northern foi adaptada a partir dos trabalhos de Moreira, AE e Gaspar, JO, (2002).

de uma célula, este é separado utilizando a técnica de eletroforese (o que caracteriza a formação de um esfregaço ou *smear* como é tecnicamente conhecido). Posteriormente estas amostras são transferidas para membrana de nitrocelulose ou nylon, e hibridizado com uma sonda radioativa marcada de forma isotópica ou não-isotópica (Figura 4A). Estas sondas são derivadas das sequências de DNA a partir da qual o mRNA de interesse foi transcrito. A presença do RNA na amostra resultará na ligação da sonda e na visualização de uma banda por autoradiografia da membrana. A intensidade da banda detectada é dependente da quantidade de RNA presente na amostra (Figura 4B). Embora extremamente sensível, é uma técnica que requer muita habilidade do operador e limita o uso de poucas sondas a serem avaliadas simultaneamente.

Nas últimas duas décadas, o uso da tecnologia de qRT-PCR ou qPCR (PCR quantitativo ou PCR em tempo real) tem ganhado notoriedade frente a tecnologia de *Northern*. Embora menos sensível, permite a investigação de muitiplos transcritos simultaneamente, e equipamentos robustos possibilitam uma automatização na realização dos ensaios envolvendo esta técnica. Da mesma forma, o RNA total de um tecido ou célula deve ser extraído. Caso seja um organismo eucarioto adiciona-se uma fase de seleção dos RNAm usando uma coluna de oligo dT. Em seguida o RNAm de interesse é convertido em cDNA com o uso de uma transcriptase reversa e oligonucleotídeos específicos dando sequencia a uma reaçao de PCR (Figura 5A). Denomina-se em tempo real porque cada vez que fluoróforos são incorporados à novas fitas sintetizadas pela técnica de PCR um sensor detecta este nível de intensidade de fluorescência emitido pelo cromóforo e cria uma curva de rot automaticamente (Figura 5B).

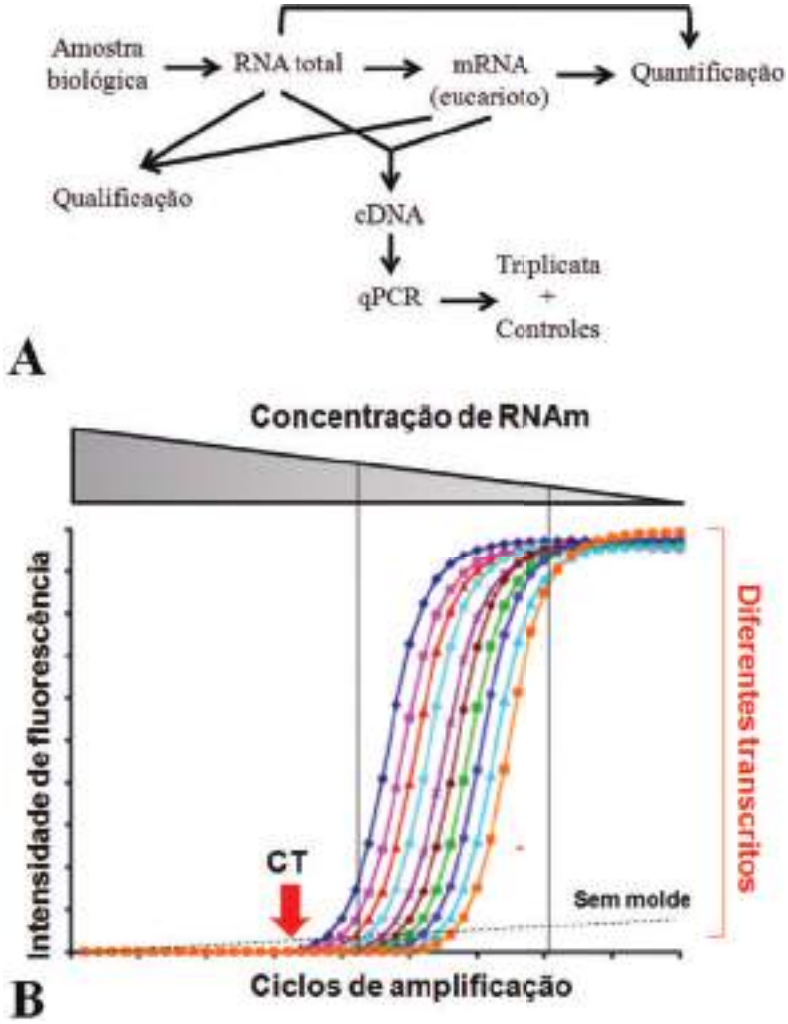


Figura 5. qPCP e curva de Rot. (A) Curva de rot envolvendo 10 diferentes transcritos por qRT-PCR. Cada linha em cor diferenciada representa um mRNA qualquer, diferente dos outros. Quanto maior a quantidade deste mRNA presente na análise, mais rápido o híbrido é formado, aumentando, consequentemente, o nível de fluorescência detectado por sensores especializados da máquina, até que se estabeleça um platô. Pode-se concluir portanto, que o mRNA representado pela linha azul à esquerda está mais abundante na amostra do que o mRNA representado pela cor laranja, estando todos os outros em concentrações intermediárias a estes exemplos.

É importante destacar que o desenho dos oligonucleotídeos nesta técnica determinará a qualidade dos ensaios, uma vez que devem se ligar a regiões específicas do gene de interesse, em sentidos opostos para que se tenha a geração de um amplicon, e que apresente Tm muito próximos uns dos outros. Este amplicon por sua vez não deve ultrapassar 100 pb (pares de bases) de tamanho e deve ter tamanho médio equivalente a todos os outros amplicons gerados simultaneamente numa análise de multiplos alvos.

Análise global da expressão gênica

Existem vários métodos para a analise em larga escala da expressão de um genoma, classificados em três categorias:

- baseado em sequenciamento, como por exemplo o sequenciamento *shotgun* de bibliotecas de cDNA (EST/ORESTES), sequenciamento do RNA (RNA seq) e bibliotecas SAGE (Capítulo 2);
- métodos baseados em amplificação, com o por exemplo PCR array;
- métodos baseados em hibridação, tais como os microarranjos, o foco deste capítulo.

Análise em larga escala utilizando microarranjos

Mais recentemente, foram desenvolvidos métodos que combinaram a capacidade de olhar para a variação na população de RNA total de diferentes tecidos com a capacidade de analisar especificamente a variação em um mRNA específico. Estes métodos são denominados de DNA chip ou microarranjos de DNA. Basicamente estes consistem num arranjo pré-definido de moléculas de DNA (fragmentos de DNA genômico, cDNAs ou oligonucleotídeos) quimicamente ligadas à uma superfície sólida (Schena et al., 1995), geralmente lâminas de vidro ou de silicone revestidas com compostos como epoxissilanos, aminossilanos, polilisina, poliacrilamida, entre outros (Grainger et al., 2007) (Figura 6).

Os microarranjos também podem ser preparados em membranas de *nylon* positivamente carregadas. Quando em membranas de nylon, por apresentar uma resolução mais grosseira frente ao microaranjo em vidro, são denominados macroarranjos. Em ambos os casos, o princípio da técnica está na detecção e quantificação de ácidos nucléicos (DNA genômico ou RNAm na forma de cDNA) provenientes de amostras biológicas, as quais são postas para hibridar com o DNA/oligonucleotídeo fixado no *chip/membrana* (hibridação por complementariedade de bases). A detecção é possível, pois são geradas sondas, normalmente marcadas com fluorocromos, como por exemplo, cianina 3 (Cy3) ou cianina 5 (Cy5) quando o ensaio for preparado em lâmina de vidro, ou com o isótopo 33-P quando o ensaio for preparado em membranas de nylon (Figura 7). Para ambos os casos se faz necessário

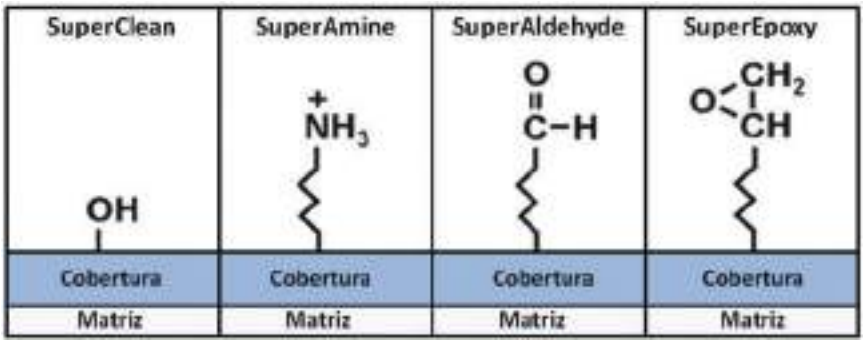
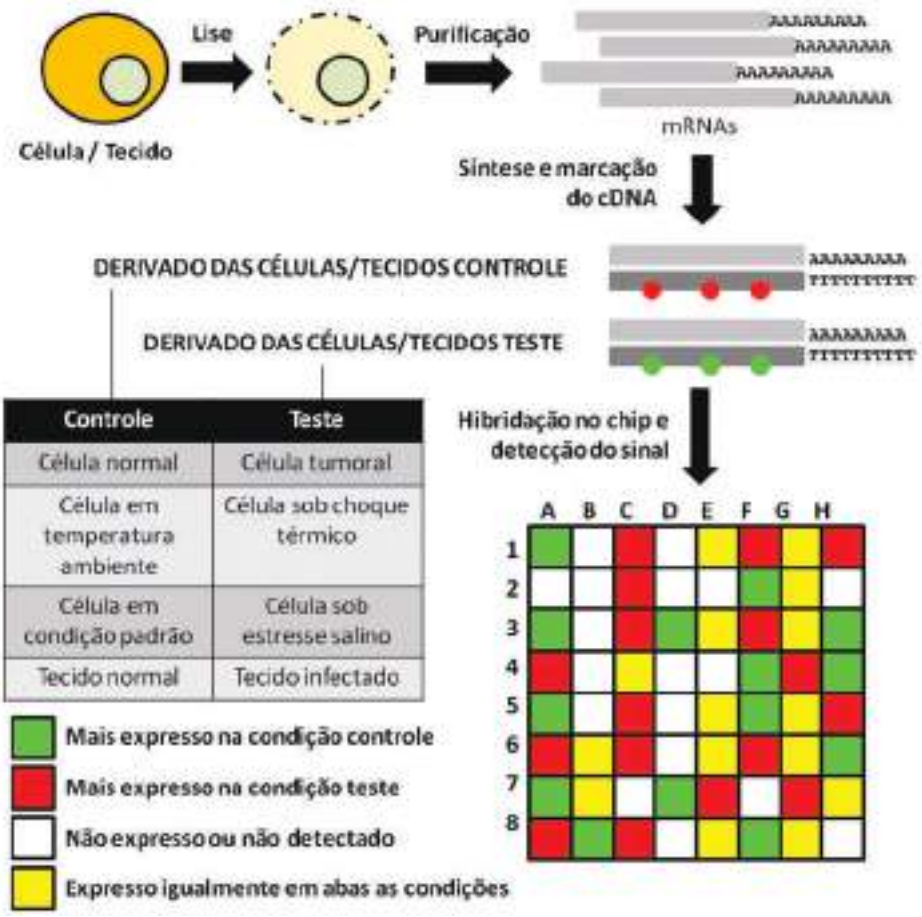


Figura 6. Exemplos de compostos ligados à superfície sólida (matriz).



Controle	Teste
Célula normal	Célula tumoral
Célula em temperatura ambiente	Célula sob choque térmico
Célula em condição padrão	Célula sob estresse salino
Tecido normal	Tecido infectado

Figura 7. Chip carregando cDNA ou oligonucleotídeos e perspectiva de hibridação sonda-alvo. Os RNAm são extraídos de tecidos ou células e utilizados como molde para síntese de cDNA, utilizando nucleotídeos marcados. O RNA não retrotranscrito é eliminado e o cDNA é utilizado como sonda nos arrays. A detecção do sinal de hibridação vai depender da abundância de cada RNA mensageiro. Por exemplo, em B2 e E2 os respectivos genes não foram transcritos em nenhuma das condições analisadas, enquanto que em E1 e B7, por exemplo, foram transcritos igualmente em ambas as condições. Nesta mesma perspectiva, em A5 e H4, por exemplo, observa-se maior intensidade de fluorescência das sondas derivadas da condição teste (verde), indicando abundancia deste mRNA na condição estudada, ao passo que em C5 e E7, por exemplo os respectivos mRNA são escassos, favorecendo a ligação das sondas associadas à condição controle.

à geração de uma imagem de hibridação, que é obtida por meio de leitores (*scanners*) a laser (para os fluorocromos) ou leitores de fósforo (para o isótopo 33-P).

Para compreendermos melhor como a técnica funciona, tomemos como exemplo um dos primeiros trabalhos envolvendo análise de expressão gênica diferencial em *Saccharomyces cerevisae* (DeRisi JL, 1997). Neste trabalho os autores queriam não apenas validar a funcionalidade da primeira plataforma de microrarnjos para estudo deste organismo, como também investigar o comportamento metabólico desta

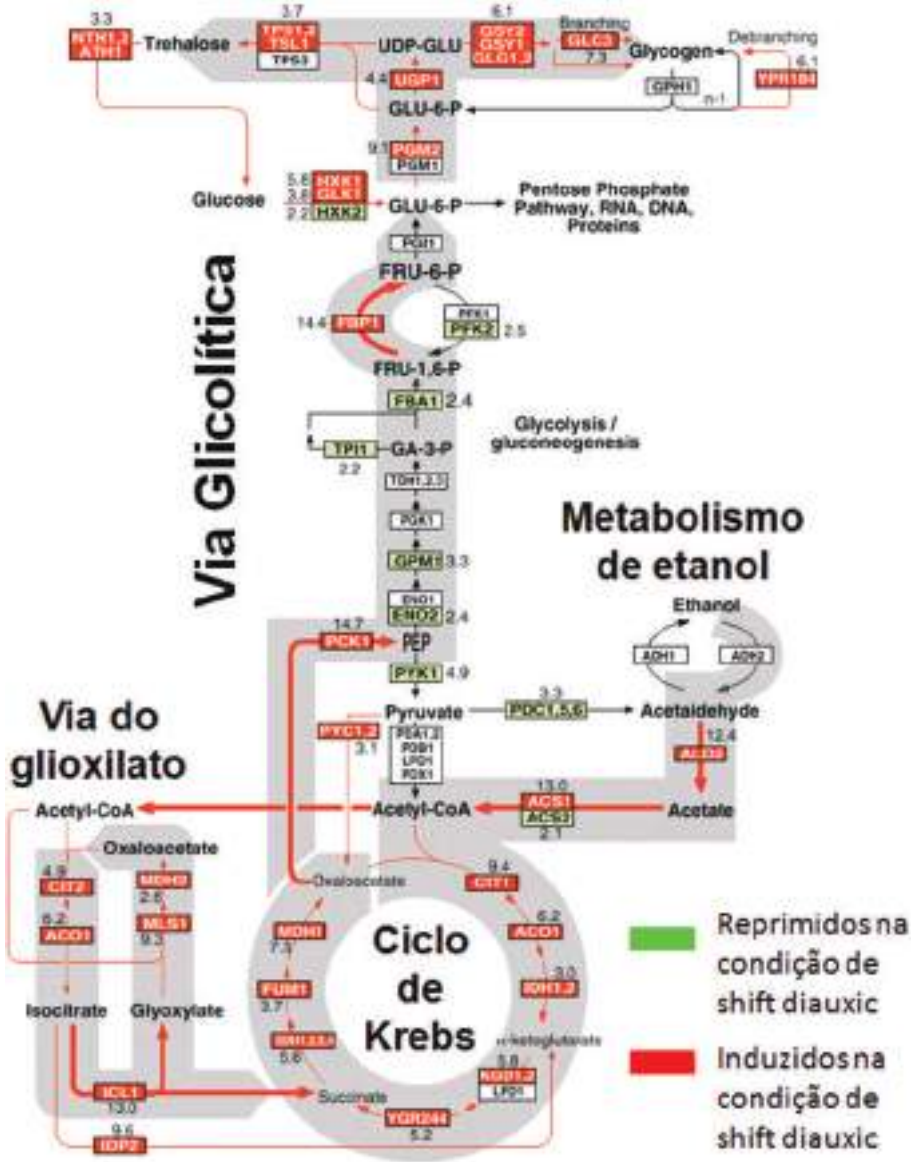


Figura 8. Perfil de expressão diferencial dos genes envolvidos com o fenômeno de *shift diauxic* em *Saccharomyces cerevisae*. Em verde destacam-se os genes ativados na condição de *shift diauxic* (presença de etanol como substrato energético) e em vermelho reprimidos nesta mesma condição (portanto, abundantes na condição controle, com oferta de glicose). Adaptado de Derisi, JL, et al. (1997).

levedura frente a um fenômeno biológico conhecido como *shift diauxic* (Figura 8). Este fenômeno em *Saccharomyces* funciona da seguinte maneira: ao fornecermos carboidratos simples a este organismo, ele rapidamente os degrada formando como metabólito final do processo fermentativo, o etanol; ao esgotar todo o carboidrato

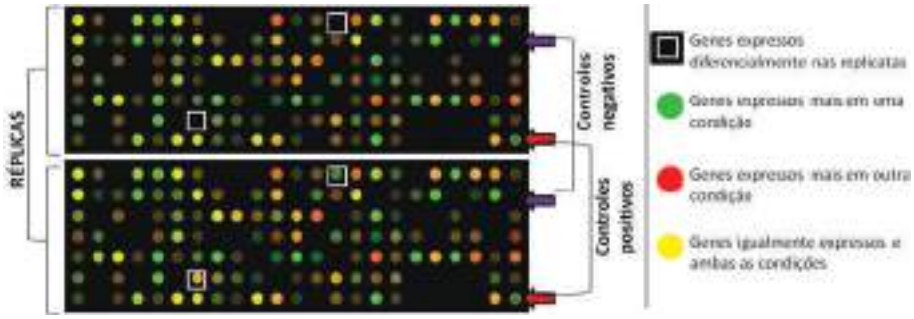


Figura 9. Perfil de hibridação de um chip de cDNA. Os pontos estão separados por uma distância média de 380 μm. Os controles positivos são indicados por setas vermelhas e os controles negativos por setas roxas. A figura mostra dois painéis (pontos em replicata) com diferentes sinais de hibridação (ver quadrados com contorno na cor branca). Esta replicata é justificada como controle interno de reprodutibilidade da reação de hibridação.

existente no meio, *Saccharomyces* passa então a metabolizar etanol para gerar energia. Isto é uma forma bastante dinâmica de competir para com outros organismos pelo carboidrato, fonte primária de energia. Como a maioria dos organismos que poderia competir com *Saccharomyces* em um determinado nicho não conseguem metabolizar o etanol, todo este substrato passa a ser metabolizado apenas por ele.

Neste trabalho os autores extraíram mRNA de Saccharomyces submetido ao crescimento na presença de glicose ou de etanol como fontes de carbono, e estes foram convertidos a cDNA com adição de fluoróforos de cores diferentes (verde e vermelho respectivamente). Estas sondas foram colocadas para hibridar num chip contendo alvos que representavam genes anotados no genoma de Saccharomyces e os resultados mostraram que na presença de glicose como substrato, genes associados a via glicolítica são mais expressos, ao passo que na presença de etanol genes associados com ciclo de Krebs (ciclo do TCA), via do glioxilato, degradação de glicogênio e degradação de etanol foram mais expressos.

Na últimas duas décadas, chips de cDNA foram desenvolvidos por várias companhias, como por exemplo Syntheni™ e Hyseq™. Os cDNAs selecionados são amplificados por PCR utilizando oligonucleotídeos específicos e então imobilizados numa lâmina com uma densidade de 10.000 amostras/cm2. Posteriormente, a lâmina será hibridada com cDNAs marcados com fluorescência e para cada cDNA haverá uma mediada e uma quantificação da intensidade de hibridação (Figura 9). Assim, quando um RNAm específico estiver presente na amostra, um sinal será obtida no ensaio, com a força do sinal proporcional à quantidade do RNA que está presente na amostra. Da mesma forma, se o RNAm estiver ausente, então também haverá ausência na sonda e nenhum sinal será obtido.

Mais recentemente, no entanto, foi desenvolvido oligo-microarranjo pela tecnologia de síntese *in-situ* por um mecanismo denominado fotolitografia (mais apropriadamente referidos como DNA-chips) (Figura 8). Neste caso, os oligonucleotídeos são sintetizados na própria lâmina numa densidade de 30-50.000 pontos/cm^2. Cada ponto representará um segmento gênico em particular. Quanto mais pontos houver no microarranjo, mais abrangente será na análise do trancrissoma. Mais especificamente,

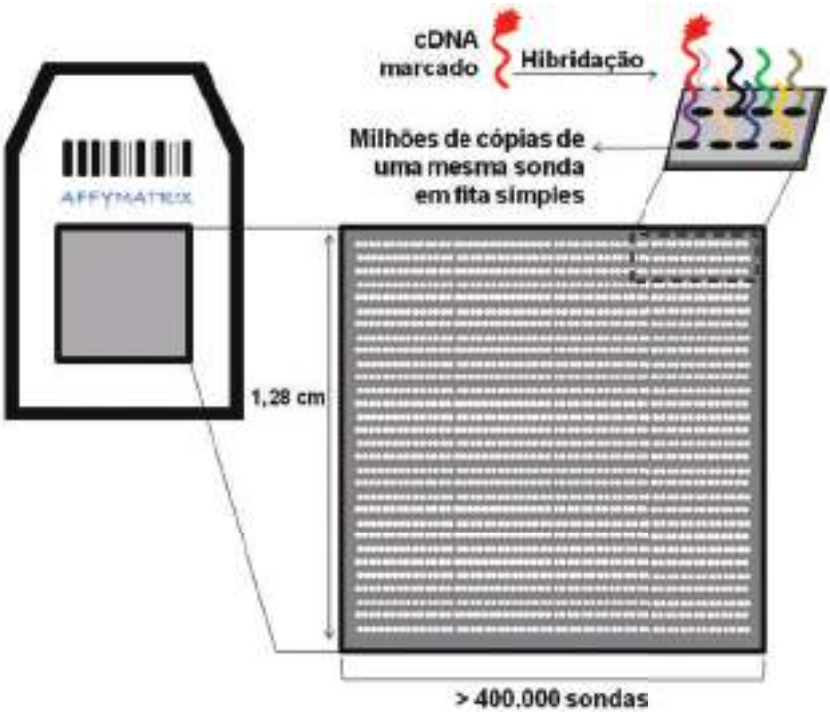


Figura10. Modelo de um GeneChip da empresa Affimatrix™.

o procedimento utilizando esta plataforma, o denominado GeneChip® (Affymetrix Inc., Santa Clara, CA, USA), é a plataforma de microarrays mais utilizada, com a qual foram publicados mais de vinte mil artigos nos últimos anos 10 anos (*http://www.affymetrix.com*) (Figura 10).

A tecnologia já foi aplicada na demonstração de padrões específicos de expressão gênica que caracterizam diferentes tipos de câncer, para predizer prognósticos em uma série de doenças, para compreender condições fisiológicas em uma série de contextos comparativos, e até respostas a terapias específicas (farmacotranscrissômica). Essa tecnologia pode ser aplicada para outras abordagens além da determinação do padrão de expressão gênica, como por exemplo, genômica comparativa, Chip, detecção de SNPs, splicing alternativo e *tallinn array* (Figura 11).

Mais uma vez usando Saccharomyces como organsimos modelo, Winzeler EA e colaboradores (1998), usaram uma plataforma de rearranjos para identificar distinções genotípicas entre duas cepas/estirpes desta levedura. Em uma das cepas (cepa 1), amostras de DNA foram extraídas e marcadas com fluoróforos vermelhos. O mesmo se procedeu com a cepa 2, porém as sondas de DNA foram marcadas com fluoróforos verdes. O grau de interação sonda alvo permite esboçar sinais de fluorescência mais intensos quando a ligação é perfeita, ou menos intensos decorrentes de mal pareamento (mismatch). Na auência de um dos sinais (verde ou vermelho) implica em dizer automaticamnete que em uma das cepas a porção de DNA correspondente está ausente (Figura 11A). Como era de se esperar os resultados demonstraram que a

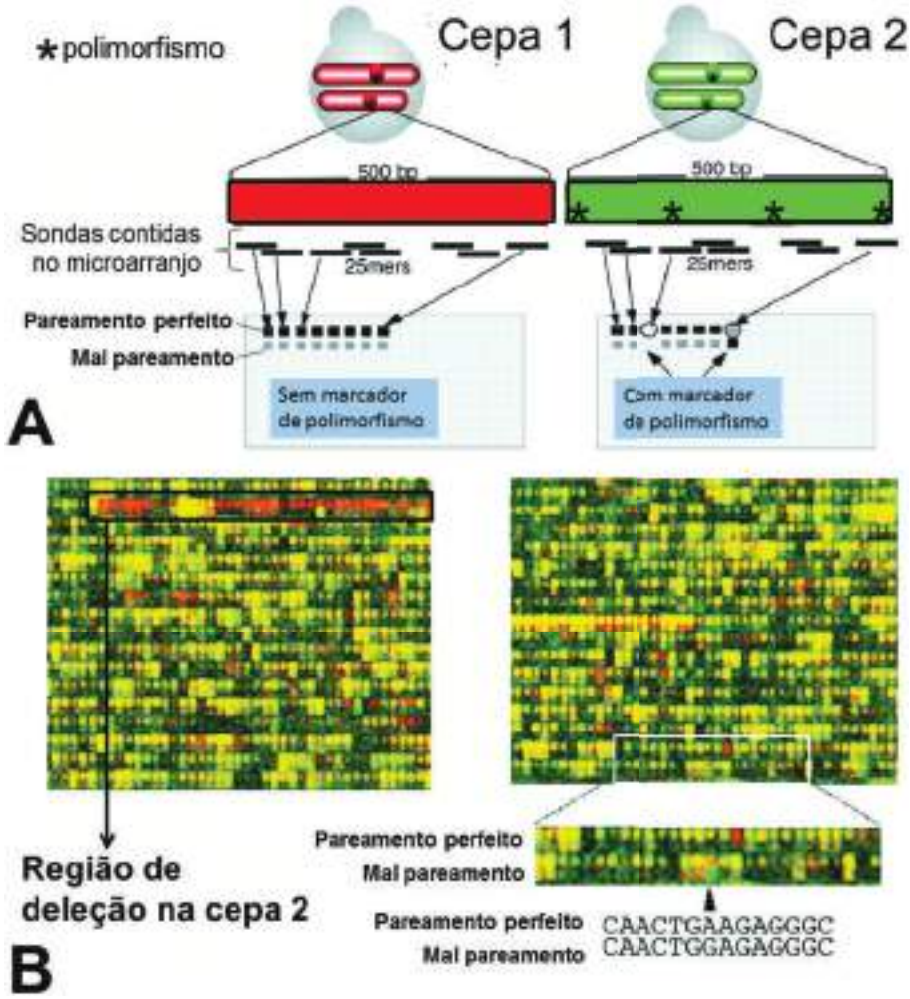


Figura 11. Uso da plataforma de microarranjos para detecçãod e SNPs, polimorfismos e deleções entre os genomas de duas cepas de *Saccharomyces cerevisae*. (A) Amostras de DNA da cepa 1 (controle) foram marcadas com fluoróforo vermelho ao passo que amostras de DNA da cepa 2 foram marcadas com fluoróforos verdes. Caso o pareamento sonda-alvo seja perfeito, então o sinal de fluorescência será intenso, se for mal pareado o sinal será mais fraco e se existir polimorfismo não haverá sinal em posições específicas do genoma da cepa a ser investigada. (B) Resultado da hibridação sonda alvo envolvendo material genético das duas cepas. A maioria dos pontos estão destacados na cor amarala, resultado de uma hibridação proporcional entre sondas da cepa controle e sondas da cepa experimento. Entretanto, dois resultados interessantes são destacados. À esquerda dá para se verificar uma região toda deletada no genoma da cepa 2, uma vez que todos os pontos em série estão na cor vermelha, logo hibridaram apenas sondas da cepa 1. À direita destaca-se uma região de polimorfismo, destacado pelo ponto verde, demonstrando que só hibridou sonda verde. Adaptado de Winzeler EA et al. (1998).

maior parte do DNA de amabas as cepas são idênticos, já que os pontos apresnetaram sinais de hibridaçao amarelo (hibridação proporcional de sondas verdes e vermelhas). Entretanto, alguns pontos chamaram atenção, como por exemplo uma grande região ausente no genoma da cepa 2, uma vez que só houve hibridação de sondas vermelhas, e pontos de polimorfismos, identificados por hibridação exclusiva de sondas verdes.

Fatores que interferem e prejudicam análises usando microarranos de DNA

Um ponto importante na técnica de microarranjo é que o perfil de expressão gênica pode ser afetado por um conjunto de diferentes fatores técnicos, que refletem mudanças nas condições experimentais. A identificação da fonte de variação e a avaliação da sua influência são fundamentais para assegurar a reprodutibilidade dos arrays. Assim, a variabilidade técnica pode ser minimizada, controlando uma série de fatores, abaixo sumarizados. São fatores que interferem a qualidade dos resultados derivados de microarranjos

- Todo RNA extraído deve antes passar por um crivo de qualidade. Quantificar este RNA não resolve. Uma análise visual da qualidade pode ser feita por eletroforese. RNA de baixa qualidade implica em sondas ruins que refletem baixa qualidade de hibridação e geração de sinal.
- Durante a produção das sondas marcadas, sejam de cDNA ou DNA, deve-se ter certa padronização e rigor metodológico, pois isso pode interferir abundantemente na qualidade do sinal de hibridação.
- Deve-se adicionar a mesma quantidade de sondas entre as condições analisadas. Diferenças nestas concentrações gerarão interferência na competição das sondas pela hibridação a seus respectivos alvos.
- Durante a hibridação de sondas deve-se evitar que se formem bolhas na interface de contato desta solução com a matriz sólida. Estas bolhas são geralmente decorrentes da prsença de contaminantes sólidos (poeira) pré-aderidos na matriz.
- Após período de hibridação sonda-alvo, deve-se fazer uma lavagem cuidadosa das amostras para eliminar qualquer sonda não ligada, mas que possa ficar aderida na matriz. Isso promoverá interferência no chamado plano de fundo (background) que por conseqüência resultará em uma captura de dados ruins pelo sistema de escâner.
- Qualidade e uso adequado do escâner é outro fator que interfere nos resultados. Evite escanear a lâmina sucessivamente, pois a cada submissão ao laser, o poder de intensidade dos fluoróforos diminui, logo o sinal também será proporcionalmente menor.
- A análise da intensidade dos pontos deve ser feito mediante uma gradeamento de pontos (spots) preciso. Caso este gradeamento tenha falhas, sinais de background serão incorporados às análises.
- Deve haver uma precisa correlação entre a matriz de grademaneto e o banco de dados que carrega as informações biológicas que representam os microarrajos. Qualquer erro nesta correção pode interferir em toda a discussão biológica dos resultados.
- Além disso, a escolha do algoritmo de processamento de imagem pode impactar diretamente no resultado do microarranjo (Shedden K ET al 2005). Existem vários

métodos, como por exemplo: dCHIP (Li C and Wong WH., 2001); MAS (Hubbell E, et al., 2002) e RMA (Irizarry RA, 2003).

Meso que haja qualidade e precisão com relação a todos estes fatores acima descritos, alguns cuidados adicionais devem surgir durante a construção de uma matriz (do microarranjo propriamente dito). Listamos abaixo algumas destas precauções:

- Toda plataforma deve conter os chamados controles negativos de hibridação. Este controle poderá representar genes que sabidamente porções de DNA que sabidamente não existem no modelo a ser analisado, ou simplesmente representar pontos onde são depositados água ou tampão ao invés de um alvo de DNA.
- Controles positivos. Muitos kits de confecção de plataformas de microarranjos acompanham frações de DNA de um organismo alvo, mais as respectivas sondas a serem marcadas em conjunto com o RNA de interesse. Logo, ao se promover interação sonda-alvo, estas sondas organismo específico se ligarão obrigatoriamente aos alvos de DNA do mesmo organismo. Durante as análises estatísticas este spoderão servir como normalizadores do processo de hibridação.
- É interessante que toda plataforma de microarranjos apresente as chamadas réplicas totais dos genes numa mesma matriz. Ou seja um conjunto de alvos de DNA fixados numa determinada posição da matriz (chamado de set) deverá estar em outra região desta mesma matriz (outro set). Isso reduzirá erros pontuais de hibridação (ver Figura 9)
- Além das réplicas em sets independentes, é interessante que hajam réplicas dentro de um mesmo set. A justificativa para ambos os casos está na qualidade de análise dos dados evitando possíveis hibridações não desejáveis ou qualquer fator que gere resultados falso positivos ou falso negativos.

Obtenção das imagens de hibridação e análise preliminar dos resultados

Como vimos até agora, a metodologia de microarranjos explora o conceito da hibridização, ou seja, o fato de que sequências complementares se ligam de modo estável. Ao se fazer as leituras das reações (escaneamento), as moléculas são excitadas com um laser e a intensidade de fluorescência capturada pelo detector é proporcional a quantidade de mRNA original referente ao respectivo gene do material analisado. Apesar da proporcionalidade é necessário obter resultados de abundância relativa a um controle. Nos microarranjos que estudamos até agora, utilizamos uma hibridização competitiva utilizando duas populações de cDNAs sendo que cada uma delas é marcada com um fluoróforo diferente, para que posteriormente, a origem da intensidade de fluorescência detectada seja reconhecida, e os resultados finais, representados como serão expressos como razão de hibridização de um cDNA em relação a amostra controle, também denominada de referência. Desta forma, da captura da imagem até a análise efetiva dos dados, são necessários vários passos para que o resultado final possa de fato refletir o transcrissoma.

Durante o processo de captura das diferentes intensidades de fluorescência obtidas são utilizados softwares específicos que localizam automaticamente ou semi-automaticamente cada um dos elementos da imagem (Figura 12). Isso acontece através

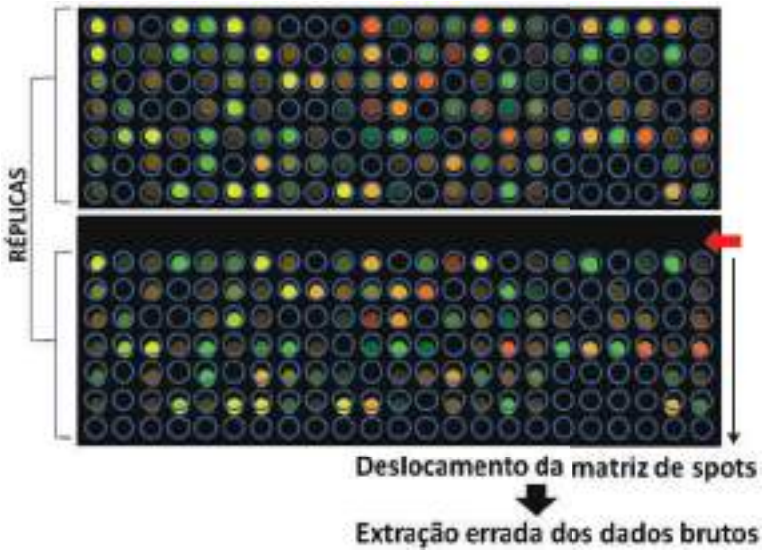


Figura 12. Imagem de resultados de microarranjos mostrando os prejuízos decorrentes de um gradeamento mal inserido pelo programa automatizado. Observe que a matriz de spots é colocada de forma errada pelo programa automatizado, o que gera erros na obtenção dos dados brutos. A perda do sinal da primeira linha na segunda réplica de hibridação deslocou a matriz para a próxima linha em evidência, adicionando, conseqüentemente, um erro nos resultados decorrentes da última linha do grademento.

da determinação de um gradeamento que deve ser precisamente definido para que o sistema só capte a fluorescência presente nos respectivos pontos representativos dos microarranjos. Um gradeamento mal inserido acarretará uma captação de imagem distorcida, podendo inclusive captar borrões do chamado background (fundo), propiciando resultados falsos positivos.

A segunda fase é a definição e retirada do background, que vai evidenciar a abundância de um transcrito e por fim, os dados são normalizados, que na análise de microarranjos torna a razão entre as intensidades de fluorescências detectadas uma medida mais precisa.

A seguir são abordadas algumas fontes de variações de imagem comumente observadas na área dos *spots* (*foreground*), no plano de fundo dos *spots* (*background*) e na informação de intensidade. Dependendo do tipo de marcação do cDNA durante a preparação do microarranjo (hibridação), pode-se obter: únicas, duplas ou múltiplas fluorescências numa mesma imagem. É mais comum encontrar fontes de dados que representam imagens duplamente fluorescentes produzidas por dispositivos que operam em dois comprimentos de onda.

Em geral, os dados da imagem de microarranjo podem consistir em um número qualquer de canais. Outra variação consiste na forma como o arquivo é armazenado, se houve compressão dos dados e qual foi a precisão utilizada (número de bytes por pixel). Por exemplo, um arquivo armazenado num formato com perda de dados introduz um "borramento" espacial dos *spots* e a análise da imagem torna-se menos precisa. Similarmente, o número de bytes por pixel precisa acomodar a faixa do sinal analógico

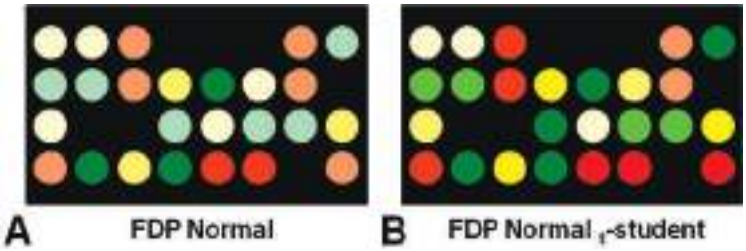


Figura 13. Ruídos de *background* modelados de formas diferenciadas. Observe a diferença nos sinais de intensidade de um mesmo ponto (gene) entre os modelos FDP Normal (A) e FDP Normal t-student (B). A Qualidade da resolução de FDP modelado por t-student é bem superior.

produzido pela excitação dos corantes fluorescentes. Essa faixa corresponde a máxima menos a mínima medida de amplitude, e qualquer valor fora do intervalo [*min* <-> *max*] é mapeado para um dos extremos. Para um número fixo de bytes, o aumento da faixa conseqüentemente diminui a precisão de cada medida de intensidade.

A localização dos *spots* também é afetada pelo material da lâmina (vidro, nylon) e os tipos de sondas utilizadas (marcação com elementos radioativos ou químicos fluorescentes) (Steinfath, M. et al., 2001). Estas variações têm diferentes causas:

- Esforços mecânicos (nylon);
- Baixa potência de discriminação (vidro);
- Forte sinal de *background* (marcação fluorescente);
- Interferência do sinal pelos sinais dos *spots* vizinhos (marcação radioativa).

A baixa potência de discriminação merece atenção especial porque devido a ela, muitos *spots* deixam de ser detectados (Buhler, J. et al., 2000). A ausência de *spots* introduz um desafio para o alinhamento automático do *grid*. Por exemplo, um método de alinhamento totalmente automático de *grid*s deverá falhar em detectar corretamente um *grid* se uma linha de *spots* longe da borda estiver faltando completamente (nenhuma evidência de existência da linha) (ver Figura 12).

As diferenças de *background* ocorrem devido a:

- Preparação da lâmina do microarranjo;
- Procedimentos inapropriados de aquisição (presença de poeira ou sujeira);
- Instrumentos de aquisição.

Os tipos (a) e (b) de variações de *background* devem ser detectadas pela avaliação da qualidade do microarranjo. Por isso as réplicas experimentais, e duplicatas de genes no interior da lâmina são fundamentais para a melhor qualidade dos resultados.

Muitos algoritmos de processamento de imagens compensam as variações de *background* através da modelagem via funções de distribuição de probabilidade (FDP) (Balagurunathan, Y. et al., 2002). O modelo mais utilizado é a FDP Gaussiana (também chamada de Normal). Outros modelos estatísticos considerados são a distribuição Uniforme e a distribuição Funcional, dependendo das propriedades observadas nas imagens adquiridas. A Figura 13 mostra exemplos de *background* modelados por distribuições de probabilidade. Embora todos os canais das imagens dos microarranjos possam seguir a mesma FDP, cada canal precisa de seus próprios parâmetros para o modelo de distribuição escolhido.

Diferentes métodos para se analisar resultados de transcrissoma

Existe uma série de métodos estatísticos para analisar dados derivados de experimentações em larga escala, incluindo os dados de transcrissoma. Destacamos a seguir aqueles que são mais corriqueiramente apresentados em trabalhos científicos.

Análise por diagrama de Venn

Embora mais simplificado e menos informativo, o diagrama de Venn é corriqueiramente usado em propósitos de investigação em larga escala (Figura 14). Embora simples e de fácil confecção/compreensão, caracteriza-se como um modelo pouco informativo já que destaca apenas o total de genes induzidos ou reprimidos em cada uma das condições analisadas.

Análise baseada em *heat map* (mapas de intensidade)

Os mapas de intensidade permitem agrupar genes de acordo com o seu perfil de expressão (induzido ou reprimido), ao longo de um período cronológico, em diferentes condições de experimentação, ou entre diferentes células/organismos submetidos a uma mesma condição fisiológica. Este agrupamento considera todas estas condições analisadas e gera um dendrograma associado a este perfil de expressão que facilita a compreensão sobre o quão parecido são estes perfis de expressão gênica (Figura 15).

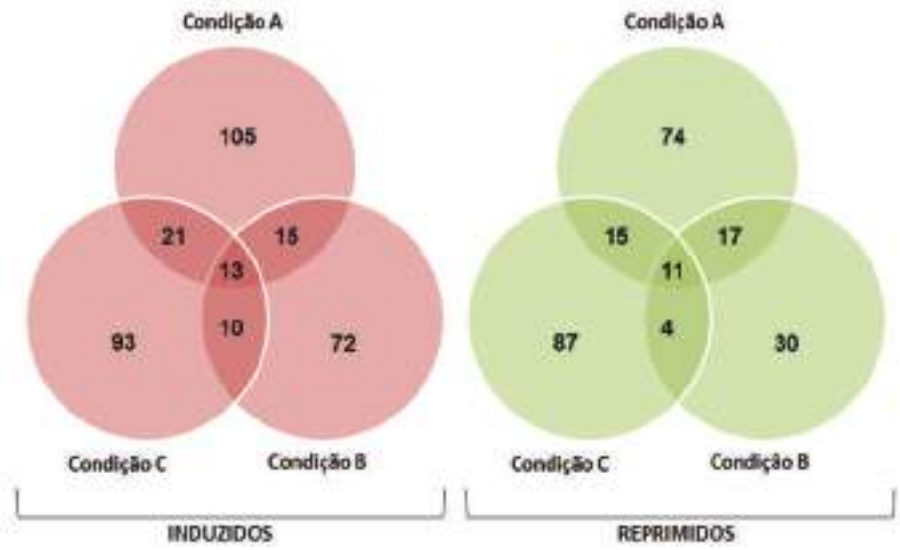


Figura 14. Diagramas de Venn para representação de expressão gênica diferencial em larga escala. Os números internos representam os genes reprimidos (vermelho) ou induzidos (verde) em cada uma das condições avaliadas, ou entre as intersecções destas condições.

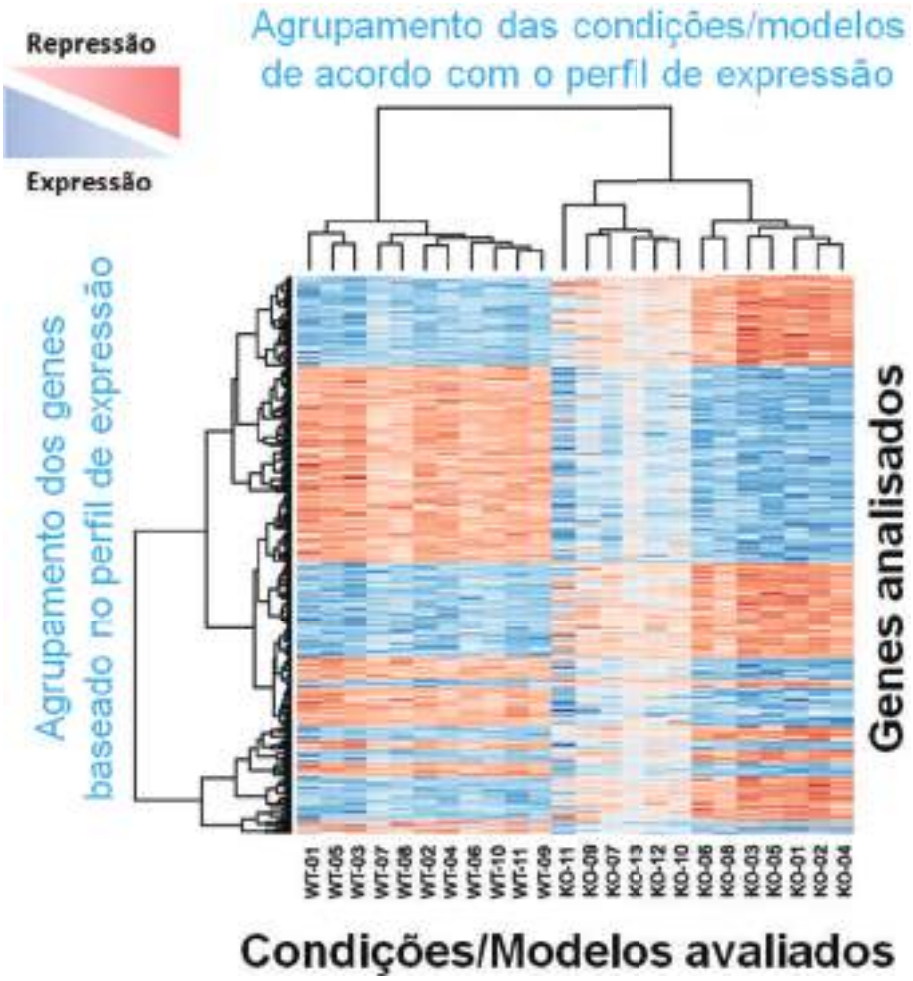


Figura15. Modelo de *heat map* gerados a partir de experimentações usando plataformas de microarranjos. Cada linha representa um gene analisado e cada coluna uma condição/modelo. As cores azuis representam genes ativados, e as cores vermelhas reprimidos. A intensidade destas cores tem relação direta com o grau de indução ou repressão. Note que dois dendrogramas foram estabelecidos para estes dados, um deles agrupa somente genes quanto aos seus perfis de expressão (esquerda), ao passo que o outro agrupa as condições/modelos avaliados (acima). Adaptado de Harding, P. et al. (2010).

Análise de agrupamentos de genes baseada em *K-means*

De forma análoga ao *heat map*, *K-means* permite agrupar genes com mesmo perfil de expressão e de acordo com os ensaios propostos. Porém, diferentemente, *K-means* não demonstra se um gene foi mais ou menos induzido do que outro se apresenta o mesmo perfil de expressão. Isto porque o objetivo central deste modelo de análise é justamente o perfil de expressão apresentado por um conjunto de genes (Figura 16).

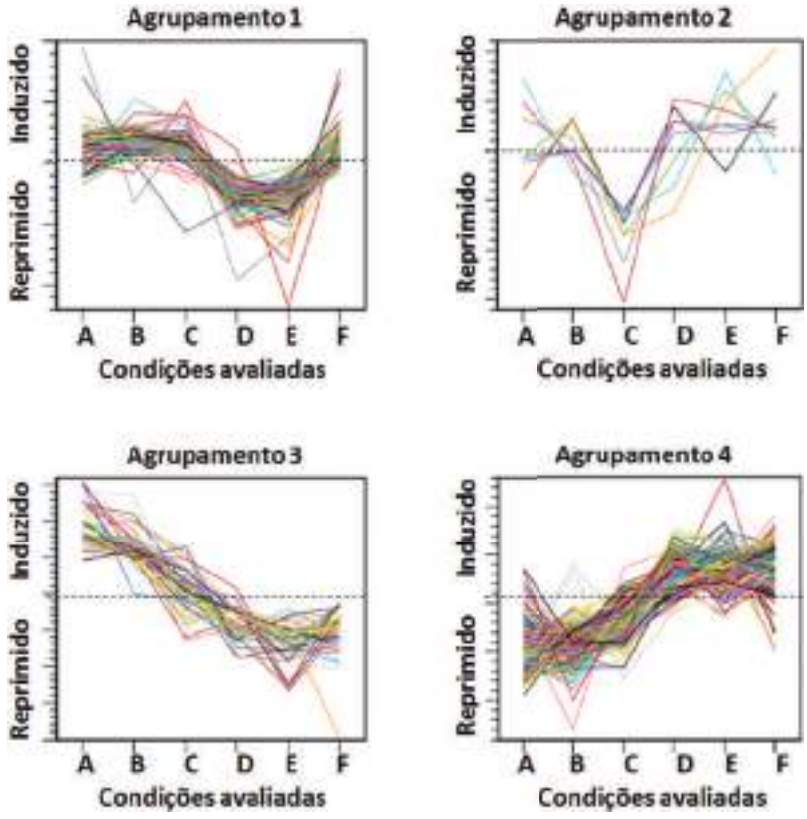


Figura 16. Análise de expressão gênica por K-means. São demonstrados apenas 4 agrupamentos de genes que apresentam expressão diferencial nas seis condições avaliadas (A a F). As linhas em cores diferenciadas representam cada um dos genes que foram agrupados de acordo com um perfil estabelecido. Os gráficos dos agrupamentos 3 e 4 permitem comparar o propósito deste modelo de análise de maneira mais clara: enquanto o gráfico da esquerda denota uma contínua repressão ao longo das condições avaliadas, o da direita denota uma progressiva indução ao longo destas mesmas condições. Adaptado de ClCBio, Qiagen company™ a partir do sítio (URL) http://www.clcbio.com/desktop-applications/features/.

Esta análise apresenta um potencial muito grande para agrupar genes tidos como hipotéticos (sem função biológica conhecida) a outros de funções já estabelecidas, permitindo especular que o perfil de expressão destes genes hipotéticos acompanha o de outros com função biológica já descrita.

Gráficos de indução × repressão gênica

Estes gráficos, com diferentes variáveis a serem inclusas nas ordenadas e abscissas, permitem compreender como está se procedendo a expressão de todos os genes de um genoma. Duas formas são freqüentemente apresentadas em trabalhos que reportam ensaios de expressão gênica global (Figura 17). Em uma delas os genes de um genoma são colocados em ambos os eixos (de forma sequencial ao seu respectivo

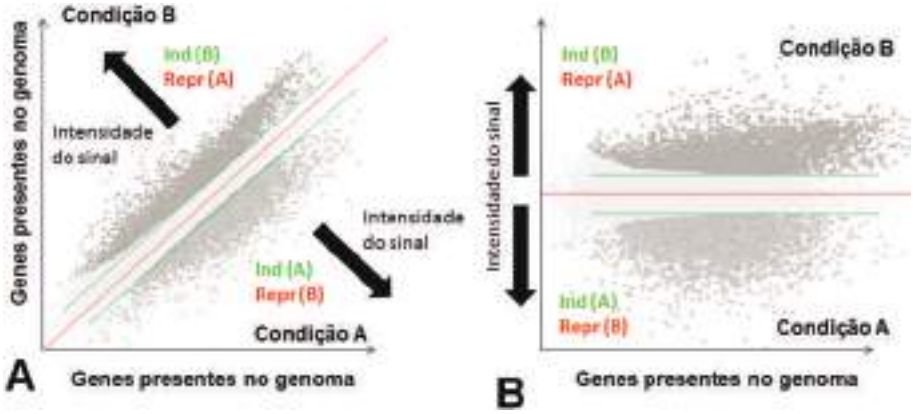


Figura 17. Gráficos de relação indução × repressão de genes em experimentos de microarranjos de DNA. Cada um dos pontos em ambos os gráficos representa um determinado gene do genoma. (A) A expressão diferencial é tomada com base no distanciamento dos pontos ao longo de um eixo de 45º que denota a expressão gênica similar nas duas condições avaliadas (eixos X e Y). (B) As mesmas condições experimentais apresentadas em A, são dispostas num plano vertical. Quanto maior o sinal de intensidade a partir da linha central, mais um determinado gene foi expresso. Em ambos os casos parâmetros normalizadores foram adicionados (pontos mais claros).

posicionamento no cromossomo) e a expressão diferencial é tomada com base no distanciamento dos pontos ao longo de um eixo de 45º, que denotaria expressão similar nas duas condições avaliadas. Em outra forma de apresentar os dados, um dos eixos representa os genes de um genoma e o outro eixo as condições avaliadas sob a perspectiva de indução ou repressão. Em ambos os casos alguns parâmetros normalizadores devem ser embutidos, fazendo com que o interesse mais evidente seja restrito aos genes (pontos) mais distantes das regiões que denotam expressão similar nas duas condições avaliadas.

Perspectivas

Embora com os avanços da proteômica (Capítulo 9) e Metabolômica (Capítulo 10) a transcrissômica tenha ficado num segundo plano de atenção, já que se propõe a investigar o comportamento de RNAs, ainda assim grupos e laboratórios especializados vêm demonstrando resultados bem relevantes em suas respectivas área do conhecimento biológico.

Diferentemente do que muitos cientistas especulam, estas novas ciências ômicas dificilmente promoverão a extinção da transcrissômica, até porque a análise de microRNAs tem ganhado notoriedade e esta tem sido a grande temática do uso desta metodologia.

O que se espera num futuro próximo é que todas estas ciências ômicas se associem para que possamos compreender um pouco mais a biologia dos organismos de forma integrada, algo que já vem sendo proposto pela chamada biologia de sistemas (Capítulo 13).

Bibliografia

BALAGURUNATHAN, Y., DOUGHERTY, E.R., CHEN, Y., BITTNER, M.L., TRENT, J.M. *Simulation of cDNA Microarrays via a Parameterized Random Signal Model*, Journal of Biomedical Optics. 2002;7(3):507-23.

BUHLER, J.; IDEKER, T.; HAYNOR, D. *Dapple: Improved Techniques for Finding Spots on DNA Microarrays.* UV CSE Technical Report UWTR. 2000, 1-12

CRICK, F. Central dogma of molecular biology. Nature.1970, 227(5258): 561–3.

DERISI JL, IYER VR, BROWN PO. Exploring the metabolic and genetic contol of gene expression on a genomic scale. Science. 1997 278:680-686.

GILBERT, W. Origin of life: The RNA world. *Nature* 1986, 319, 618

GRAINGER DW, GREEF CH, GONG P, LOCHHEAD MJ. Current microarray surface chemistries. Methods Mol Biol. 2007;381:37-57.

HARDING , P., YANG, X-P., YANG, J., SHESELY, E., HE, Q. , LAPOINTE, M.C. Gene expression profiling of dilated cardiomyopathy in older male EP4 knockout mice. American Journal of Physiology - Heart and Circulatory PhysiologyPublished. 2010, 298, H623-H632.

HASTIE ND, BISHOP JO.The expression of three abundance classes of messenger RNA in mouse tissues. Cell. 1976 Dec;9(4 PT 2):761-74.

HAYES PC, WOLF CR, HAYES JD. Blotting techniques for the study of DNA, RNA, and proteins. BMJ. 1989 Oct 14;299(6705):965-8.

HUBBELL E, LIU WM, MEI R. Robust estimators for expression analysis. Bioinformatics 2002;28:2585–92.

IRIZARRY RA, HOBBS B, COLLIN F, BEAZER-BARCLASY YD, ANTONELLIS KJ, SCHERF U, SPEED TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics, 2003 4(2):249–264.

LI C and WONG WH. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci 2001;98:31–36.

MOREIRA, AE e GASPAR, JO. Propriedades moleculares de um isolado brasileiro do Southern bean mosaic vírus. Fitopatol. bras.2002, 27 (3)

SAIKI, R.K., S. SCHARF, F. FALOONA, K.B. MULLIS, G.T. HORN, H.A. ERLICH, and N. ARNHEIM. 1985. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science 230:1350-1354

SCHENA M, SHALON D, DAVIS RW, BROWN PO.Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995 Oct 20;270(5235):467-70.

SHEDDEN K, CHEN W, KUICK R, GHOSH D, MACDONALD J, CHO KR, GIORDANO TJ, GRUBER SB, FEARON ER, TAYLOR JM, HANASH S. Comparison of seven methods for producing Affymetrix expression scores based on False Discovery Rates in disease profiling data. BMC Bioinformatics 2005;6:26.

STEINFATH, M., WRUCK, W., SEIDEL, H., LEHRACH, H., RADELOF, U. and O'BRIEN, J. *Automated image analysis for array hybridization experiments*, Bioinformatics, v. 17, p. 634-641. 2001.

WINZELER EA, RICHARDS DR, CONWAY AR, GOLDSTEIN AL, KALMAN S, MCCULLOUGH MJ, MCCUSKER JH, STEVENS DA, WODICKA L, LOCKHART DJ, DAVIS RW (1998) Direct allelic variation scanning of the yeast genome. Science 281:1194-1197.

9

Análise proteômica: princípios e aplicações

William de Castro Borges
Leandro Xavier Neves

Introdução

De maneira simplista, Proteômica refere-se a uma tentativa de investigação simultânea do repertório total de proteínas expressas por uma célula. Neste sentido, credita-se a Marc R. Wilkins a autoria do termo 'PROTEOME', durante a '*First Siena conference, 2D electrophoresis: from protein maps to genomes*', realizada no período de 5 a 7 de setembro, 1994. Nesta ocasião, Wilkins definiu o termo como: "*the* PROTE*in complement expressed by a gen*OME".

A informação genética armazenada no núcleo de uma célula, ou seja, a sequência, número e sintenia dos genes, constituem elementos estáticos permanecendo essencialmente inalteráveis durante o ciclo celular. Por outro lado, a comunicação direta entre célula e meio ambiente promove um equilíbrio dinâmico entre transcrição gênica e tradução proteica, definindo diferentes proteomas para uma mesma célula ao longo do seu processo de diferenciação (de Hoog and Mann 2004). Fatores externos como a influência de drogas, alterações de pH, temperatura e disponibilidade de nutrientes também constituem elementos fundamentais na definição do proteoma (Figura 1). Enquanto o número total de genes codificadores de um determinado organismo pode ser relativamente baixo (cerca de 30.000 para o genoma humano), o número de proteínas constituintes quase sempre atinge a ordem de centenas de milhares (Venter, Adams et al. 2001). Este número torna-se consideravelmente maior quando modificações pós-traducionais contribuem para o aumento da diversidade molecular e funcional proteica de uma célula.

Este capítulo tem por objetivo apresentar os conceitos fundamentais e os alicerces de instrumentação relacionados a uma investigação proteômica. Enquanto as seções

Figura 1. Diagrama esquemático para a definição de PROTEOMA (S) - Situação hipotética em que se verifica o equilíbrio dinâmico do conteúdo de uma célula quando submetida a diferentes condições. Por outro lado, sobre influência de quaisquer fatores, a informação contida no núcleo celular permanece essencialmente invariável.

iniciais abordam as estratégias de proteômica clássica por eletroforese bidimensional, as últimas páginas são dedicadas às análises de identificação de proteínas em larga escala. Em ambos os casos, a proteômica aqui descrita refere-se àquela atualmente denominada *Bottom-up*. Os métodos apresentados visam à obtenção e detecção de peptídeos por espectrometria de massas para a geração de dados espectrais os quais, por sua vez, possibilitam inferir a identidade das moléculas proteicas de origem.

A técnica de eletroforese bidimensional (2-DE)

O primeiro grande desafio para uma análise proteômica consistiu-se na necessidade de visualização de um número significativo de proteínas em um único gel. A eletroforese de proteínas em uma dimensão, embora de grande utilidade, não se presta à comparação de proteomas pelo fato de proteínas distintas poderem apresentar diferenças mínimas de massas moleculares e pontos isoelétricos. Neste contexto, a aplicação de tecnologia para o ressurgimento da técnica de eletroforese em gel bidimensional, originalmente descrita por (O'Farrell 1975), contribuiu sobremaneira para que proteomas pudessem ser comparados com altíssima reprodutibilidade. Através desta metodologia, as proteínas são inicialmente separadas de acordo com a carga líquida em gel de primeira dimensão e, em seguida, submetidas à separação de acordo com a massa molecular na segunda dimensão. A recente utilização de géis de primeira dimensão, com variados gradientes de pH imobilizados para isoeletrofocalização, permite reprodutibilidade

nas preparações e grande facilidade operacional (Gorg, Drews et al. 2009). Estima-se que o poder de resolução de um gel 2D, 16 x 18 cm, esteja em torno de 3000 proteínas (Figura 2).

A técnica de eletroforese em gel de poliacrilamida em duas dimensões apresenta grande utilidade na comparação de proteomas de um mesmo organismo, quando submetido a condições diversas, ou mesmo para comparação de isolados diferentes de uma mesma cepa. Neste sentido, podem ser detectadas alterações como aumento ou diminuição nos níveis de expressão, modificações pós-traducionais que modificam a massa molecular como glicosilação, adição de lipídeos e ainda modificações que

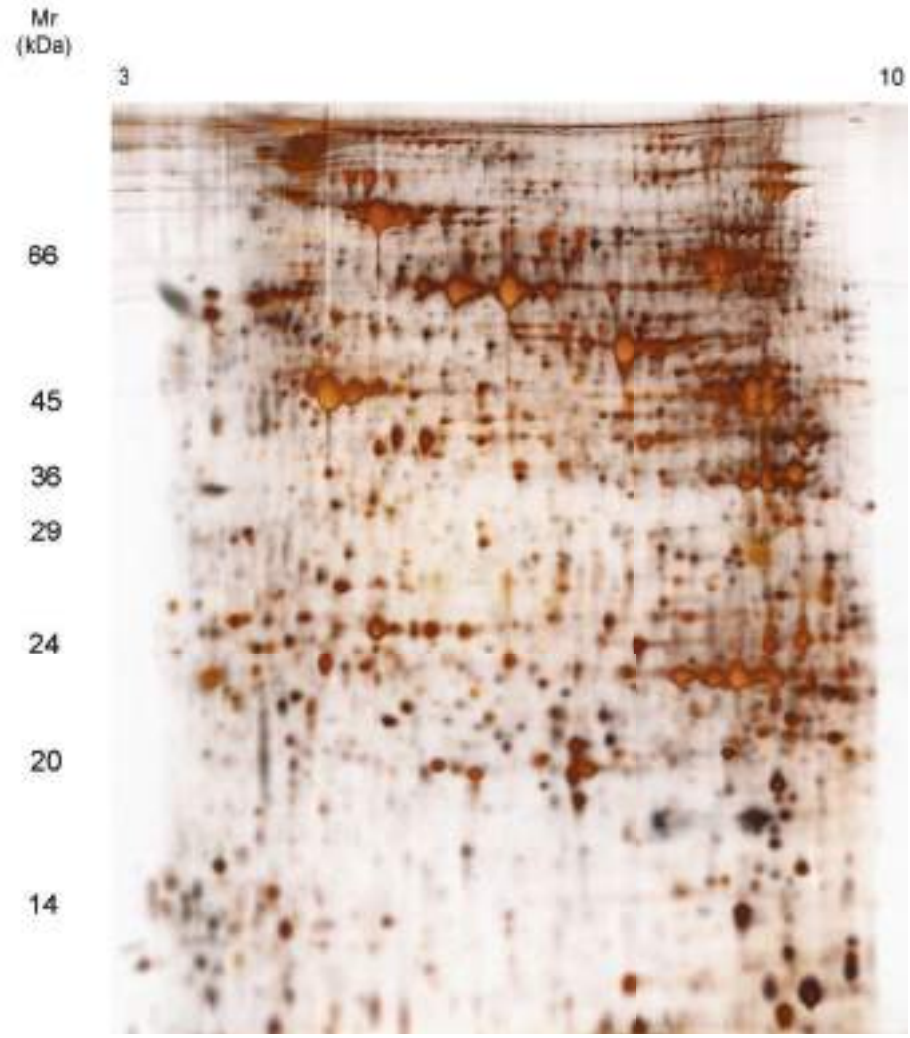


Figura 2. Eletroforese em gel bidimensional (2D-PAGE). Gel representativo de extrato solúvel de pares de vermes adultos do parasito *Schistosoma mansoni*. Notar a presença de várias proteínas apresentando pontos isoelétricos semelhantes, distribuídas em ampla faixa de massa molecular. Coloração por prata. Cortesia de Leandro X. Neves - Laboratório de Enzimologia e Proteômica, UFOP.

interferem com a carga líquida da molécula como fosforilação e defosforilação. A recente proposição de protocolos para marcação fluorescente de amostras a serem comparadas, seguido de separação simultânea nos géis de primeira e segunda dimensão, permitiu a criação de metodologia alternativa para a separação por 2-DE denominada **DIGE** - 2D Fluorescence Difference Gel Analysis Technology (Alban, David et al. 2003; Marouga, David et al. 2005). Esta tecnologia constitui atualmente a ferramenta de maior reprodutibilidade para comparação de proteomas em gel, através da qual, alterações mínimas de expressão gênica podem ser inferidas com alta confiabilidade.

A primeira dimensão: isoeletrofocalização

Proteínas são moléculas anfotéricas, ou seja, por serem constituídas de aminoácidos, muitos dos quais contendo cadeias laterais com capacidade de ionização, o somatório de cargas positivas e negativas pode conferir diferentes cargas líquidas às moléculas, dependendo do pH no qual se encontram. Em linhas gerais, abaixo do pH no qual a carga líquida é igual a zero (pI) as proteínas apresentam carga líquida positiva enquanto que em pH superior ao pI elas estão carregadas negativamente. Quando um campo elétrico é aplicado sobre um gradiente de pH imobilizado (gel de primeira dimensão), as proteínas a serem isoeletrofocalizadas migram para a posição na qual apresentam carga líquida igual à zero. Neste ponto elas deixam de sofrer a influência de ambos os eletrodos. Por exemplo, uma proteína que apresenta carga líquida negativa irá migrar em direção ao anodo (+), ocorrendo então um decréscimo gradual dessa carga até que a proteína atinja e estacione no pH correspondente ao seu ponto isoelétrico. Obviamente, a situação inversa também é válida, o que resulta num processo capaz de separar proteínas presentes numa mistura tomando por base diferenças mínimas de cargas elétricas (Castro-Borges, Cartwright et al. 2007).

A segunda dimensão: SDS-PAGE

Após o processo de isoeletrofocalização, a segunda etapa de construção de um mapa proteico bidimensional, refere-se ao processo de separação das proteínas tomando por base o parâmetro massa molecular, essencialmente conforme descrito por (Laemmli 1970). Para este fim, procedimentos prévios de redução e alquilação das proteínas no gel de primeira dimensão garantem a completa desnaturação das moléculas a serem resolvidas na segunda dimensão. Este processo é ainda comumente facilitado pelo tratamento com o detergente dodecil sulfato de sódio, o qual confere carga líquida negativa às moléculas. Desta maneira, após a montagem do gel de primeira dimensão no aparato de eletroforese contendo o gel de separação, seguido de aplicação de uma diferença de potencial, as proteínas migrarão pela matriz do gel em direção ao polo positivo. Proteínas de maior massa molecular encontrarão maior resistência de migração enquanto as de menor massa apresentarão maior mobilidade no polímero. Através do emprego de géis de tamanhos variados, aliado à possibilidade do procedimento ser realizado em gradientes distintos do polímero, é possível a

detecção de moléculas apresentando diferenças mínimas de massas moleculares (Gorg, Obermaier et al. 1999; Gorg, Weiss et al. 2004).

A identificação de proteínas por espectrometria de massas e bioinformática

Geração de peptídeos a partir de proteínas intactas

Paralelamente ao emprego da tecnologia associada ao grande poder de reprodutibilidade dos géis 2D foram desenvolvidos protocolos com o objetivo de extrair do gel a sequência de aminoácidos contida nas proteínas de interesse. Com este propósito, o desenvolvimento da técnica de digestão *in gel* revolucionou o processo de identificação das proteínas uma vez que moléculas presentes numa mistura complexa puderam ser individualmente sequenciadas, sem necessidade de purificação ou fracionamento prévio. O princípio da técnica baseia-se na excisão da proteína ou *spot* de interesse seguido de hidrólise enzimática ou química para a geração de peptídeos e posterior análise por espectrometria de massas (Shevchenko, Wilm et al. 1996; Shevchenko, Tomas et al. 2006). Em linhas gerais, anteriormente ao processo de hidrólise da proteína, ciclos sucessivos de redução e alquilação dos resíduos de cisteína garantem que fragmentos lineares de aminoácidos serão gerados quando do emprego do método de hidrólise escolhido. Sendo de massa molecular menor que a proteína original, os peptídeos produzidos difundem do gel para o sobrenadante, por onde podem ser coletados para análise por espectrometria de massas (Figura 3). Várias enzimas proteolíticas podem ser empregadas neste processo, no entanto a tripsina é mais comumente utilizada. Esta protease pode apresentar características interessantes como alta estabilidade, conseguida a partir de metilação de seus resíduos de lisina,

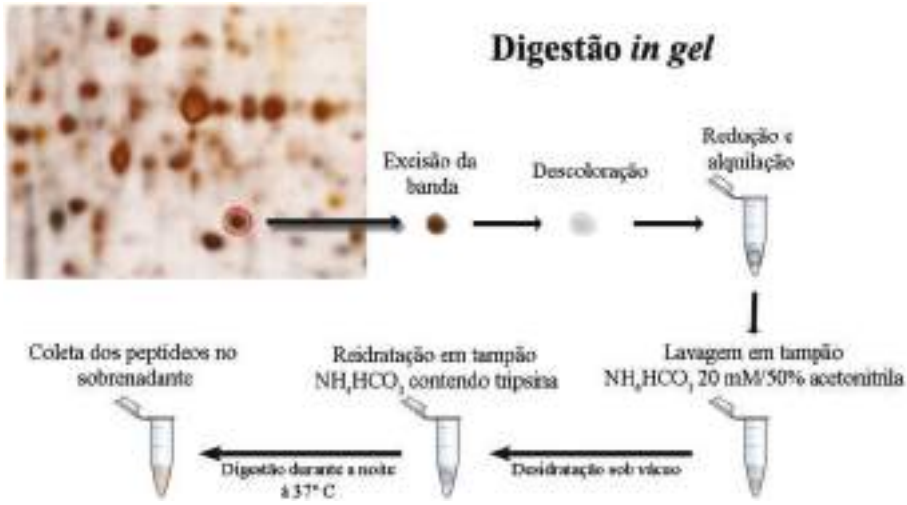


Figura 3. Sequência de etapas do protocolo de digestão *in gel*.

e especificidade para a hidrólise de ligações peptídicas C-terminais aos resíduos de lisina e arginina de seus substratos.

De um modo geral, a digestão de proteínas por tripsina gera peptídeos na faixa de massa molecular ideal para o sequenciamento por espectrometria de massas. A hidrólise de proteínas *in gel* também pode ser realizada com o emprego de reagentes químicos tais como brometo de cianogênio e ácido fórmico, os quais promovem clivagens de ligações peptídicas frente aos resíduos de metionina e ácido aspártico, respectivamente (Hua, Low et al. 2006; Samyn, Sergeant et al. 2006). Contudo, independente do método de hidrólise empregado, nem todos os peptídeos gerados pelo processo de digestão *in gel* serão devidamente ionizados e detectados pelo espectrômetro de massas. Somente aqueles que possuírem tamanhos ideais e carga líquida positiva poderão alcançar o detector para o registro da relação massa/carga (*m/z*), como será discutido posteriormente.

A espectrometria de massas aplicada à análise de peptídeos

Para melhor compreendermos o funcionamento de um espectrômetro de massas podemos subdividi-lo em três constituintes fundamentais: a fonte de ionização, o analisador de massas e o detector (Figura 4).

Primeiramente, a amostra a ser analisada é introduzida na fonte de ionização do instrumento, onde as moléculas (em particular os peptídeos) serão ionizadas e transferidas para a fase gasosa. Este processo de ionização é essencial, uma vez que íons podem ser facilmente manipulados através da aplicação de campo elétrico, ao contrário de moléculas neutras. Após essa etapa os íons são conduzidos por este campo até o analisador de massas do equipamento, onde ocorre a separação de acordo com suas respectivas razões *m/z*. Na terceira, e última etapa do processo, dados sobre abundância relativa e os valores *m/z* são individualmente registrados pelo detector e, posteriormente, apresentados através do espectro de massas.

Outro grande desafio no campo da espectrometria de massas aplicada à Proteômica consistiu na elucidação de métodos que permitissem com que peptídeos pudessem ser ionizados e detectados no espectrômetro de massas. Criados no início do século XX, os espectrômetros foram primariamente utilizados para a determinação da massa molecular de pequenas moléculas, devido à facilidade de ionização e incorporação das amostras na fase gasosa. Neste sentido, um problema fundamental no emprego da espectrometria para análise de constituintes celulares tais como proteínas, carboidratos complexos e ácidos nucleicos, consistiu na deficiência de técnicas que permitissem a transferência de moléculas polares e de alta massa molecular para a fase gasosa, sem destruí-las estruturalmente.

Figura 4. Constituintes fundamentais de um espectrômetro de massas.

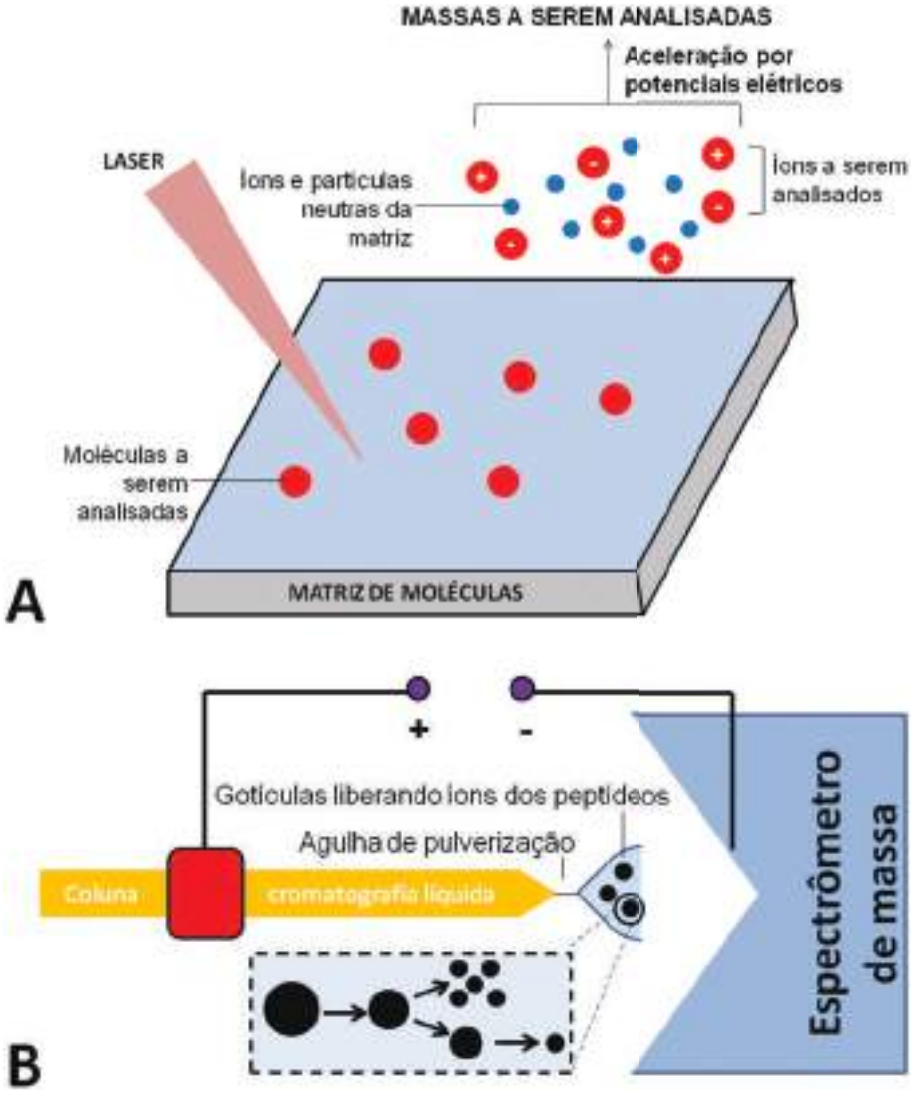


Figura 5. Métodos de ionização da mistura peptídica - Em (a), a técnica de MALDI, ilustrando a transferência dos íons peptídicos + matriz para a fase gasosa. Em (b), a técnica de eletropulverização, ilustrando o processo de "droplet fission". Adaptado de: **THE ABC'S (AND XYZ'S) OF PEPTIDE SEQUENCING** - STEEN, H.; MANN, M., 2004.

Este problema foi resolvido como o advento das técnicas denominadas *soft ionization* ou técnicas brandas de ionização tais como MALDI (Matrix-Assisted Laser Desorption Ionization) e ESI (ElectroSpray Ionization), Figura 5.

No processo de ionização por MALDI (Hillenkamp, Karas et al. 1991), a amostra peptídica a ser analisada é inicialmente combinada a um excesso de matriz orgânica, cuja característica estrutural permite absorção de radiação ultravioleta. A análise da mistura produzida ocorre após evaporação total do solvente, sendo dependente do

estado sólido dos peptídeos a serem analisados. No interior do equipamento, a partir da incidência de um feixe de laser, em comprimento de onda adequado, a energia absorvida pelas moléculas de matriz é transferida para as moléculas dos peptídeos, possibilitando com que estes últimos, em particular, passem do estado sólido para a fase gasosa. Durante este processo, os peptídeos em geral adquirem carga positiva +1. Neste momento, sob vácuo e diferença de potencial em torno de 20 kV os peptídeos carregados, os quais podem apresentar diferenças mínimas de massas moleculares, alcançam o detector de maneira diferenciada. A combinação mais comum da técnica de ionização por MALDI consiste no registro do tempo de vôo (T*ime of* F*light*) dos íons produzidos, os quais alcançam o detector em tempos variados de acordo com suas massas. A configuração final do equipamento é conhecida como MALDI-ToF. A versatilidade desse método de ionização permite inclusive com que constituintes teciduais, presentes num corte histológico, possam ser analisados de maneira global pela técnica de *MALDI-Imaging* (Schwamborn and Caprioli 2010). Por meio dessa técnica, e utilizando-se de algoritmos apropriados, é possível a obtenção de uma imagem tecidual que reflete composição molecular. Desse modo, informações sobre distribuição espacial e quantificação de constituintes presentes *in situ* podem ser reveladas.

Através da técnica de ionização por eletropulverização (ESI), os peptídeos resultantes da digestão proteolítica alcançam o espectrômetro de massas a partir de uma interface prévia com um sistema de cromatografia líquida (Yamashita and Fenn 1984). Desta maneira, uma diferença fundamental da ionização por MALDI refere-se à natureza líquida da amostra a ser ionizada por ESI. Por este processo, a amostra é dissolvida num solvente polar e volátil sendo bombeado na direção de um capilar de pequeno diâmetro (75 a 150 μm), o qual alcança o interior da fonte de ionização do espectrômetro de massas. A aplicação de uma alta voltagem na extremidade deste capilar garante a criação de forte campo elétrico nesta região, proporcionando a formação de um aerossol composto por gotículas altamente carregadas. Este processo é auxiliado pela nebulização de um gás, o qual direciona o *spray* emergindo do capilar para o interior do espectrômetro de massas. As gotículas carregadas diminuem em tamanho pela evaporação do solvente, sendo este processo auxiliado por um fluxo de nitrogênio, conhecido como gás de secagem. O processo de perda de solvente por evaporação gera gotas de diâmetro cada vez menor, um evento denominado *droplet fission* (fissão da gota) o qual produz, ao mesmo tempo, a dispersão dos íons constituintes. O processo continua até que, em média, todos os íons presentes na amostra original estejam espacialmente dispersos um do outro. Os peptídeos ionizados são então acelerados para o interior do espectrômetro onde terão suas razões *m/z* resolvidas de acordo com o tipo de analisador de massas do instrumento. A etapa de separação através do parâmetro *m/z* acontece de maneira bastante semelhante ao discutido acima. Entretanto, é válido mencionar que ao contrário da técnica de MALDI, a qual produz principalmente íons com carga +1, por ESI, os íons em geral apresentam múltiplas cargas (Wilm 2011).

Após a determinação dos valores *m/z* e o registro da intensidade de todos os picos no espectro, os modernos espectrômetros de massas prosseguem no processo de obtenção de informação sobre a sequência primária de aminoácidos contida nos íons. A sequência de eventos é denominada MS Sequencial ou MS^n, devido ao fato

dos equipamentos poderem acoplar mais de um estágio de análise por espectrometria de massas. Para a grande maioria dos casos, uma análise por MS^2 é suficiente para extrair informação sobre a composição de aminoácidos de um determinado peptídeo.

Neste contexto, um peptídeo precursor é inicialmente isolado e, através de inúmeras colisões para transferência de energia com um gás inerte, o qual pode ser nitrogênio, hélio ou argônio, o peptídeo sofre fragmentações de ligações internas gerando íons de menor massa molecular. O emprego de condições controladas de energia de colisão permite com que o processo de fragmentação do íon precursor não ocorra de maneira simplesmente aleatória, devido às diferenças nas forças de ligação que compõem a cadeia peptídica. A princípio, vários íons podem ser observados, entretanto, na prática prioriza-se a obtenção de íons correspondentes a fragmentos opostos e complementares entre si, tais como aqueles das séries -y e -b, úteis na determinação da composição de aminoácidos contida na sequencia do íon precursor. Em particular, íons de série -y e -b são gerados a partir da fragmentação da ligação peptídica e são exemplos de íons nos quais a carga retida situa-se nas regiões C- e N-terminais, respectivamente (Figura 6) (Roepstorff and Fohlman 1984). Quando a clivagem ocorre simultaneamente nas posições C- e N-terminais de um resíduo de aminoácido, o fragmento liberado é denominado íon imônio (Tabela 1). Estes íons embora não tenham utilidade na determinação da sequência dos aminoácidos nos peptídeos, apresentam valor diagnóstico por fornecerem informações valiosas a respeito de composição molecular dos íons precursores. Em linhas gerais, considera-se um espectro MS^2 elucidativo quando o mesmo é composto por um número adequado de íons das séries -y e -b, associado a um número significativo de íons imônios (Figura 7) (Steen and Mann 2004; Cantú, Carrilho et al. 2008).

Levando-se em consideração que, de um modo geral, a fragmentação da ligação peptídica produzirá íons que diferem na razão *m/z* correspondente à massa molecular de um resíduo de aminoácido, aparentemente torna-se simples o processo de determinação dos mesmos na sequência do íon precursor. Entretanto, durante a

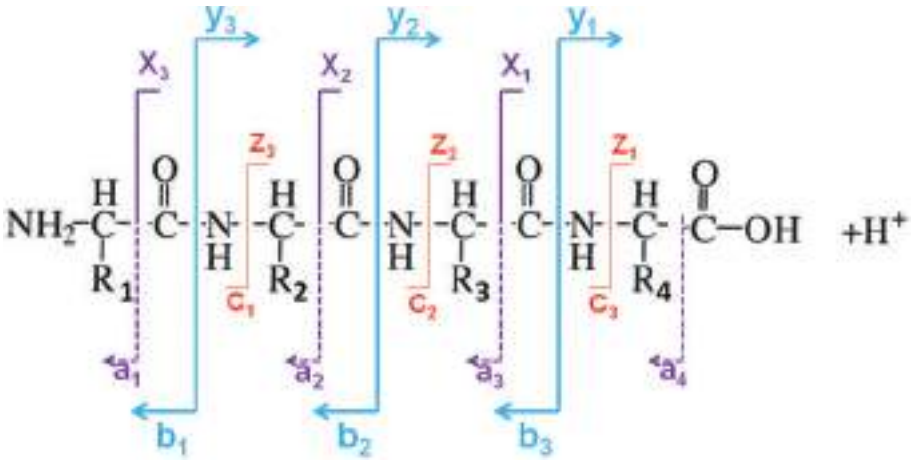


Figura 6. Principais séries de íons obtidas a partir da fragmentação de íons precursores. Em geral, íons de série -y e -b são suficientes para determinação da sequência peptídica.

Tabela 1. Massa dos resíduos de aminoácidos de ocorrência natural e seus respectivos íons imônio.

Aminoácidos	Abreviação	Massa (Dalton)	Íon imônio (*m/z*)
Glicina	G	57.02146	30.03439
Alanina	A	71.03711	44.05004
Serina	S	87.03203	60.04496
Prolina	P	97.05276	70.06569
Valina	V	99.06841	72.08134
Treonina	T	101.04768	74.06061
Cisteína	C	103.00919	76.02212
Leucina/Isoleucina	L/I	113.08406	86.09699
Asparagina	N	114.04293	87.05586
Ácido aspártico	D	115.02694	88.03987
Glutamina	Q	128.05858	101.07151
Lisina	K	128.09496	101.10789
Ácido glutâmico	E	129.04259	102.05552
Metionina	M	131.04049	104.05342
Histidina	H	137.05891	110.07184
Fenilalanina	F	147.06841	120.08134
Arginina	R	156.10111	129.11404
Tirosina	Y	163.06333	136.07626
Triptofano	W	186.07931	159.09224

fragmentação vários eventos podem ocorrer, dentre eles destacam-se às perdas de massas devido às modificações pós-traducionais, frequentemente presentes nos íons precursores, tais como fosforilação / glicosilação e a ocorrência de rearranjos intramoleculares. Esses eventos podem resultar na presença de picos no espectro cuja interpretação nem sempre poderá ser facilmente conseguida.

Contudo, dependendo da qualidade do espectro obtido e da experiência do pesquisador, é possível a determinação da sequência do íon precursor manualmente, ou através de auxílio via algoritmos apropriados, num processo denominado *De novo sequencing* (Samgina, Kovalev et al. 2010; Seidler, Zinn et al. 2010). Este método consiste na identificação e anotação de valores relacionados à fragmentação de aminoácidos ou peptídeos presentes no espectro original de modo a determinar, através de cálculos variados, a sequência de aminoácidos nele contida.

Bioinformática aplicada à proteômica

Rotineiramente, os estudos de proteômica baseiam-se na identificação e busca de homologia automática da informação espectral fornecida pelo espectrômetro de massas (Eng, Mccormack et al. 1994; Mann and Wilm 1994). Neste sentido, por

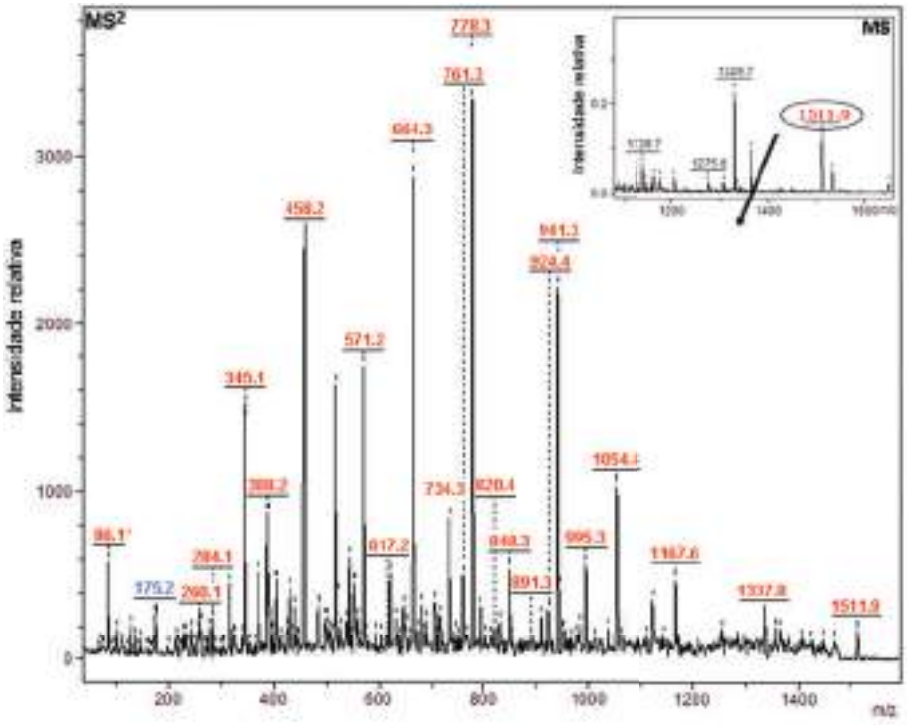


Figura 7. Espectros de massas MS e MS². O espectro MS ilustra a relação *m/z* dos íons gerados pela hidrólise proteica. O espectro MS² mostra os íons obtidos após fragmentação do íon precursor *m/z* 1511.9. * Pico referente ao íon amônio dos resíduos de leucina/isoleucina. Cortesia de Adam Dowle – Centre of Excellence in Mass Spectrometry

tratar-se de comparação entre sequências de aminoácidos, sucesso tem sido obtido na identificação e caracterização de proteínas provenientes de organismos para os quais dados de sequenciamento genômico ou transcrissômico estão disponíveis. Neste contexto, algoritmos delineados exclusivamente com esta finalidade são capazes de interpretar dados espectrais e, desta forma, revelar as sequências peptídicas preditas, contidas em bancos de sequências de nucleotídeos ou aminoácidos. Para este fim, esses algoritmos inicialmente simulam uma proteólise ou hidrólise *in silico* de todas as sequencias codificadoras presentes no banco de dados escolhido. Obviamente, esta simulação leva em consideração a enzima ou reagente químico utilizado durante a geração experimental dos íons precursores. Além disso, a informação sobre a natureza do espectrômetro de massas, com particular importância para o método de ionização e o analisador de massas utilizado, orienta a busca para determinados grupos de íons precursores (+1, ou múltiplas cargas para aqueles gerados por MALDI e ESI, respectivamente). Outros parâmetros a serem considerados na predição das massas dos peptídeos referem-se a modificações químicas induzidas como a alquilação dos resíduos de cisteína, ou aquelas geradas espontaneamente tais como a oxidação de metioninas. Deste modo, o algoritmo terá temporariamente obtido as massas de todos os possíveis íons precursores contidos nas sequencias de proteínas preditas

no banco de dados. A partir da comparação dos dados de fragmentação obtidos experimentalmente com aqueles gerados pela simulação *in silico* obtém-se, por meio de tratamentos estatísticos refinados, as prováveis sequencias dos peptídeos sob investigação. Com base no altíssimo poder de resolução dos espectrômetros de massas atuais é possível determinar a sequencia de um peptídeo com uma diferença de massa < 0.001 Da, entre os valores preditos e os observados experimentalmente. Na prática, a identificação de vários peptídeos com *scores* significativos, provenientes de uma mesma proteína, eleva o grau de confiabilidade de uma identificação positiva (Cottrell 2011).

Identificação de proteínas em larga escala - *Shotgun Proteomics*

A proteômica por análise do tipo *shotgun* foi desenvolvida em paralelo ao aprimoramento das técnicas de cromatografia líquida e espectrometria de massas. Por essa estratégia um número significativamente maior de proteínas presentes numa mistura complexa pode ser obtido num tempo reduzido de análise. Esta abordagem tem sido largamente empregada na determinação de composição de proteomas por permitir alta processividade, quando comparada ao método clássico de análise por 2-DE (Rabilloud, Chevallet et al. 2010). Mesmo considerando o alto poder de resolução dos géis 2D, esta técnica não é recomendada para a análise de proteínas que compartilham determinadas características físico-químicas tais como hidrofobicidade acentuada, alta massa molecular e ponto isoelétrico extremo (Wilkins, Gasteiger et al. 1998; Corthals, Wasinger et al. 2000). Desta maneira, moléculas potencialmente relevantes para um estudo proteômico podem ser desconsideradas e, portanto, excluídas da análise, baseando-se apenas em separações por géis (Rabilloud 2009; Makarov and Scigelova 2010).

Através de uma abordagem de proteômica via *shotgun* inicialmente promove-se a hidrólise proteolítica de todos os componentes da preparação. Esta etapa, denominada digestão em solução permite com que moléculas incompatíveis com o método de separação por gel bidimensional, possam ter seus peptídeos constituintes analisados por espectrometria de massas (Figura 8). Convém ressaltar que o resultado final desta etapa consiste numa mistura complexa de peptídeos, os quais deverão ser separados e analisados individualmente. Neste contexto, a proteômica *shotgun* depende essencialmente do acoplamento dos métodos de cromatografia líquida e espectrometria de massas. Por sua vez, os processos de separação cromatográfica exploram as diferentes propriedades físico-químicas dos peptídeos gerados. A separação depende de interações peptídicas individuais favorecidas ou desfavorecidas com a fase estacionária, podendo ser iônicas (aniônicas e catiônicas) ou não-iônicas (hidrofílicas ou hidrofóbicas).

Rotineiramente os protocolos de análises proteômicas via *shotgun* incluem etapas de purificação dos peptídeos totais obtidos, anteriores à aplicação dos métodos de separação cromatográfica. Através de extração em fase sólida, os peptídeos resultantes da digestão em solução podem ser facilmente purificados, em conjunto, de modo a excluir substâncias que poderiam interferir em análises posteriores por espectrometria de massas. Substâncias como agentes redutores, alquilantes, sais e

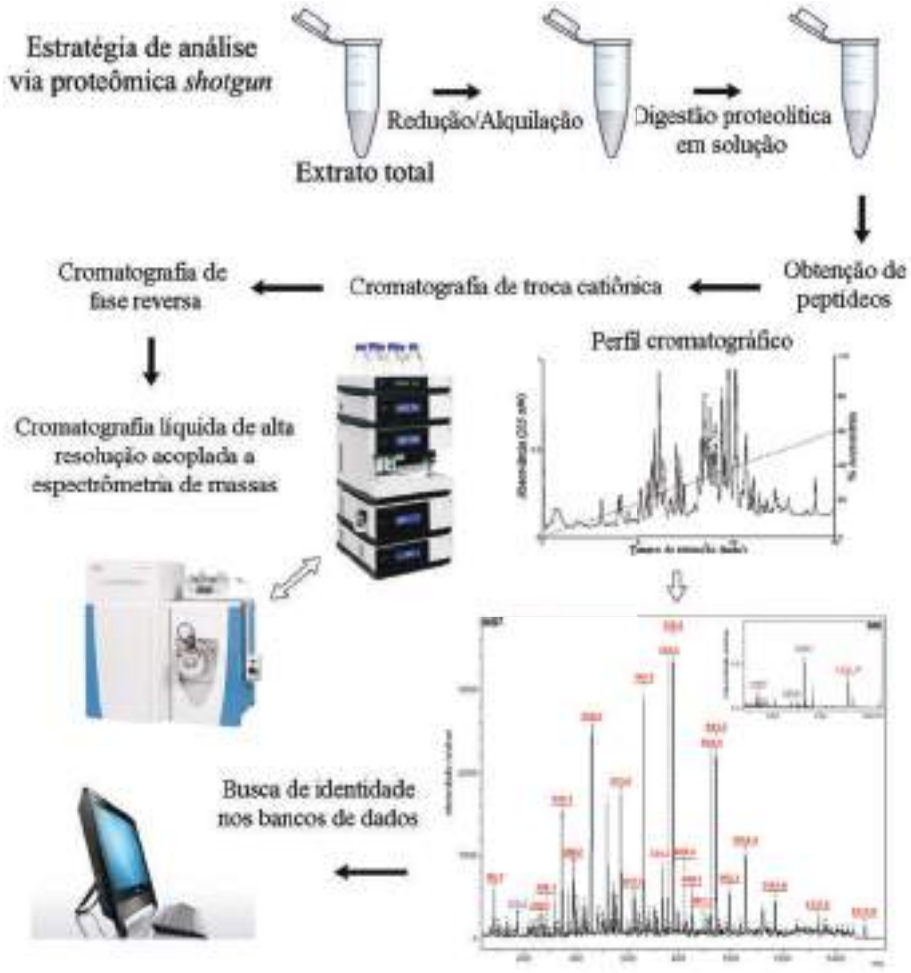


Figura 8. Etapas envolvidas no processo de elaboração e análise de dados via proteômica shotgun.

até mesmo detergentes, frequentemente presentes no tampão de digestão, podem ser removidos por completo, utilizando-se de até duas extrações sequenciais em fase sólida. Na primeira, geralmente pelo emprego de uma troca catiônica, consegue-se a remoção de detergentes, agentes redutores e alquilantes. No entanto, a eluição dos peptídeos ligados à fase sólida requer a utilização de um tampão contendo quantidades proibitivas de sais para análise por espectrometria de massas. Deste modo, numa segunda etapa de extração em fase reversa, utilizando-se de um solvente orgânico e meio ácido para eluição, obtêm-se peptídeos em condições ideais para serem separados por cromatografia de alta resolução e analisados via espectrometria (Castro-Borges, Dowle et al. 2011; Ferret-Bernard, Castro-Borges et al. 2012). Obviamente, esses dois processos podem estar acoplados no sistema cromatográfico, quando se pretende fracionar uma mistura complexa peptídica em duas dimensões (Fournier, Gilmore et al. 2007). Neste caso, a troca catiônica pode gerar sub-frações de peptídeos

as quais podem ser individualmente separadas por cromatografia de fase reversa. Esta estratégia permite uma alta resolução na separação dos peptídeos por cromatografia líquida e, desta forma, propicia com que moléculas presentes em baixos níveis na preparação possam ser detectadas durante a análise. O processo de identificação dos peptídeos e, por conseguinte, das proteínas constituintes da preparação, segue-se de maneira bastante similar ao descrito na seção anterior.

A análise de proteínas por *shotgun* tem sido uma valiosa ferramenta para a caracterização em larga escala de proteomas complexos. A abordagem tem sido valiosa no aprofundamento dos estudos proteômicos, uma vez que contorna alguns dos desafios impostos pela complexa natureza das proteínas. Em paralelo à alta capacidade de identificação, a análise por *shotgun* permite ainda a possibilidade de quantificação relativa e absoluta dos constituintes individuais de uma preparação (Ishihama, Oda et al. 2005; Ong and Mann 2005; Shinoda, Tomita et al. 2010). Experimentos de proteômica quantitativa podem, por exemplo, orientar a escolha racional de novos alvos vacinais (Castro-Borges, Simpson et al. 2011) e apontar marcadores de relevância para o prognóstico de doenças (McDonald and Yates 2002; Spivey 2009). Avanços significativos de instrumentação relacionados tanto aos processos cromatográficos quanto à resolução e sensibilidade dos espectrômetros de massas, tem ocorrido de maneira surpreendente (Cox and Mann 2011). Análises recentes mostraram que cerca de 4μg de peptídeos totais foram suficientes para a identificação da quase totalidade do proteoma de *Saccharomyces cereviseae*, sem o emprego de fracionamento prévio da complexa mistura peptídica (Nagaraj, Kulak et al. 2012). Em consonância, tem sido demonstrado que a partir de alguns nanolitros recuperados (via microdissecção a *laser)* de um corte histológico, alterações moleculares presentes especificamente em determinadas células ou tecidos podem ser detectadas através de análise por proteômica *shotgun* (Wisniewski, Ostasiewicz et al. 2011).

Considerações finais

O cenário atual aplicado à tecnologia de análise proteômica tem permitido análises globais e altamente refinadas dos sistemas biológicos. A cada dia, novos biomarcadores relacionados a doenças e processos diversos vêm sendo revelados com o auxílio de métodos proteômicos. Aos pesquisadores cabe a busca de inspiração para utilizá-los da maneira mais apropriada, de modo a decifrar a complexidade das inter-relações que coletivamente contribuem para a definição do proteoma celular.

Bibliografias

ALBAN, A., S. O. DAVID, et al. (2003). "A novel experimental design for comparative two-dimensional gel analysis: two-dimensional difference gel electrophoresis incorporating a pooled internal standard." Proteomics 3(1): 36-44.

CANTÚ, M. D., E. CARRILHO, et al. (2008). "Sequenciamento de peptídeos usando espectrometria de massas: um guia prático." Química Nova 31: 669-675.

CASTRO-BORGES, W., J. Cartwright, et al. (2007). "The 20S proteasome of Schistosoma mansoni: a proteomic analysis." Proteomics 7(7): 1065-1075.

CASTRO-BORGES, W., A. Dowle, et al. (2011). "Enzymatic shaving of the tegument surface of live schistosomes for proteomic analysis: a rational approach to select vaccine candidates." PLoS Negl Trop Dis 5(3): e993.

CASTRO-BORGES, W., D. M. SIMPSON, et al. (2011). "Abundance of tegument surface proteins in the human blood fluke Schistosoma mansoni determined by QconCAT proteomics." Journal of Proteomics 74(9): 1519-1533.

CORTHALS, G. L., V. C. WASINGER, et al. (2000). "The dynamic range of protein expression: a challenge for proteomic research." Electrophoresis 21(6): 1104-1115.

COTTRELL, J. S. (2011). "Protein identification using MS/MS data." Journal of Proteomics 74(10): 1842-1851.

COX, J. and M. MANN (2011). "Quantitative, high-resolution proteomics for data-driven systems biology." Annu Rev Biochem 80: 273-299.

de HOOG, C. L. and M. MANN (2004). "Proteomics." Annu Rev Genomics Hum Genet 5: 267-293.

ENG, J. K., A. L. MCCORMACK, et al. (1994). "An Approach to Correlate Tandem Mass-Spectral Data of Peptides with Amino-Acid-Sequences in a Protein Database." J Am Soc Mass Spectrom 5(11): 976-989.

FERRET-BERNARD, S., W. CASTRO-BORGES, et al. (2012). "Plasma membrane proteomes of differentially matured dendritic cells identified by LC-MS/MS combined with iTRAQ labelling." Journal of Proteomics 75(3): 938-948.

FOURNIER, M. L., J. M. GILMORE, et al. (2007). "Multidimensional separations-based shotgun proteomics." Chem Rev 107(8): 3654-3686.

GORG, A., O. DREWS, et al. (2009). "2-DE with IPGs." Electrophoresis 30 Suppl 1: S122-132.

GORG, A., C. OBERMAIER, et al. (1999). "Recent developments in two-dimensional gel electrophoresis with immobilized pH gradients: wide pH gradients up to pH 12, longer separation distances and simplified procedures." Electrophoresis 20(4-5): 712-717.

GORG, A., W. WEISS, et al. (2004). "Current two-dimensional electrophoresis technology for proteomics." Proteomics 4(12): 3665-3685.

HILLENKAMP, F., M. KARAS, et al. (1991). "Matrix-Assisted Laser Desorption Ionization Mass-Spectrometry of Biopolymers." Anal Chem 63(24): A1193-A1202.

HUA, L., T. Y. LOW, et al. (2006). "Microwave-assisted specific chemical digestion for rapid protein identification." Proteomics 6(2): 586-591.

ISHIHAMA, Y., Y. ODA, et al. (2005). "Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein." Mol Cell Proteomics 4(9): 1265-1272.

LAEMMLI, U. K. (1970). "Cleavage of structural proteins during the assembly of the head of bacteriophage T4." Nature 227(5259): 680-685.

MAKAROV, A. and M. SCIGELOVA (2010). "Coupling liquid chromatography to Orbitrap mass spectrometry." J Chromatogr A 1217(25): 3938-3945.

MANN, M. and M. WILM (1994). "Error-tolerant identification of peptides in sequence databases by peptide sequence tags." Anal Chem 66(24): 4390-4399.

MAROUGA, R., S. DAVID, et al. (2005). "The development of the DIGE system: 2D fluorescence difference gel analysis technology." Anal Bioanal Chem 382(3): 669-678.

MCDONALD, W. H. and J. R. YATES, 3rd (2002). "Shotgun proteomics and biomarker discovery." Dis Markers 18(2): 99-105.

NAGARAJ, N., N. A. KULAK, et al. (2012). "System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap." Mol Cell Proteomics 11(3): M111 013722.

O'FARRELL, P. H. (1975). "High resolution two-dimensional electrophoresis of proteins." J Biol Chem 250(10): 4007-4021.

ONG, S. E. and M. MANN (2005). "Mass spectrometry-based proteomics turns quantitative." Nat Chem Biol 1(5): 252-262.

RABILLOUD, T. (2009). "Membrane proteins and proteomics: love is possible, but so difficult." Electrophoresis 30 Suppl 1: S174-180.

RABILLOUD, T., M. CHEVALLET, et al. (2010). "Two-dimensional gel electrophoresis in proteomics: Past, present and future." J Proteomics 73(11): 2064-2077.

ROEPSTORFF, P. and J. FOHLMAN (1984). "Proposal for a common nomenclature for sequence ions in mass spectra of peptides." Biomed Mass Spectrom 11(11): 601.

SAMGINA, T. Y., S. V. KOVALEV, et al. (2010). "N-terminal tagging strategy for de novo sequencing of short peptides by ESI-MS/MS and MALDI-MS/MS." J Am Soc Mass Spectrom 21(1): 104-111.

SAMYN, B., K. SERGEANT, et al. (2006). "A method for C-terminal sequence analysis in the proteomic era (proteins cleaved with cyanogen bromide)." Nat Protoc 1(1): 318-323.

SCHWAMBORN, K. and R. M. CAPRIOLI (2010). "MALDI imaging mass spectrometry--painting molecular pictures." Mol Oncol 4(6): 529-538.

SEIDLER, J., N. ZINN, et al. (2010). "De novo sequencing of peptides by MS/MS." Proteomics 10(4): 634-649.

SHEVCHENKO, A., H. TOMAS, et al. (2006). "In-gel digestion for mass spectrometric characterization of proteins and proteomes." Nat Protoc 1(6): 2856-2860.

SHEVCHENKO, A., M. WILM, et al. (1996). "Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels." Anal Chem 68(5): 850-858.

SHINODA, K., M. TOMITA, et al. (2010). "emPAI Calc--for the estimation of protein abundance from large-scale identification data by liquid chromatography-tandem mass spectrometry." Bioinformatics 26(4): 576-577.

SPIVEY, A. (2009). "Amplify, amplify: shotgun proteomics boosts the signal for biomarker discovery." Environ Health Perspect 117(5): A206-209.

STEEN, H. and M. MANN (2004). "The ABC's (and XYZ's) of peptide sequencing." Nat Rev Mol Cell Biol 5(9): 699-711.

VENTER, J. C., M. D. ADAMS, et al. (2001). "The sequence of the human genome." Science 291(5507): 1304-1351.

WILKINS, M. R., E. GASTEIGER, et al. (1998). "Two-dimensional gel electrophoresis for proteome projects: the effects of protein hydrophobicity and copy number." Electrophoresis 19(8-9): 1501-1505.

WILM, M. (2011). "Principles of electrospray ionization." Mol Cell Proteomics 10(7): M111 009407.

WISNIEWSKI, J. R., P. OSTASIEWICZ, et al. (2011). "High recovery FASP applied to the proteomic analysis of microdissected formalin fixed paraffin embedded cancer tissues retrieves known colon cancer markers." Journal of Proteome Research 10(7): 3040-3049.

YAMASHITA, M. and J. B. FENN (1984). "Electrospray Ion-Source - Another Variation on the Free-Jet Theme." Journal of Physical Chemistry 88(20): 4451-4459.

10

Metabolômica

Jan Schripsema
Denise Saraiva Dagnino

Introdução

O termo metaboloma foi introduzido em 1998 em um artigo de Oliver et al. (1998) para indicar o conjunto dos metabólitos de um organismo. Um organismo contém uma gama de compostos dentre eles íons, macromoléculas e micromoléculas. São entendidas como metabólitos as micromoléculas orgânicas do sistema biológico. Estes metabólitos derivam da interação do genoma, do transcriptoma e do proteoma do organismo com o meio ambiente dando origem ao seu fenótipo (Figura 1).

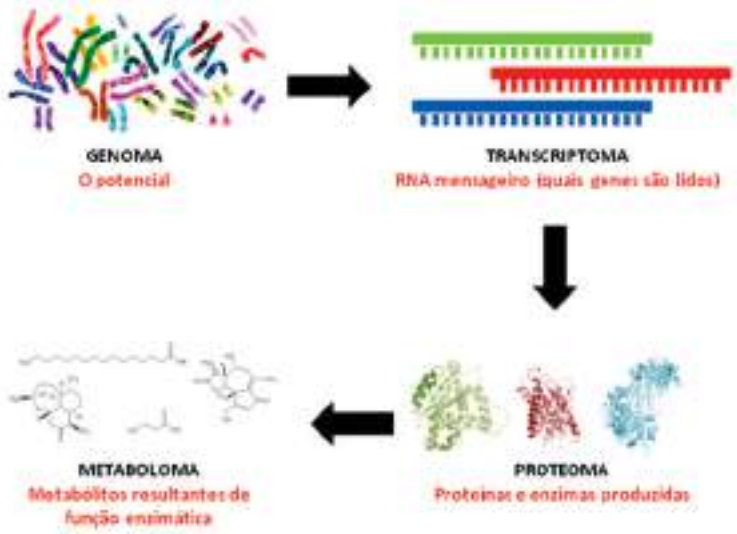


Figura 1. A relação entre as ciências "ômicas".

A metabolômica tem como objetivo estudar o conjunto desses metabólitos, assim como a genômica envolve o estudo dos seus genes, o genoma, a transcriptômica o estudo do RNA mensageiro, o transcriptoma e a proteômica o estudo das proteínas, o proteoma. Trata-se de mais uma área de estudo que tem como as outras, o objetivo de aumentar a nossa compreensão e possivelmente de controlar o metabolismo. Sendo assim, o estudo extensivo do metaboloma de um organismo leva à identificação e quantificação do conjunto de seus metabólitos. No entanto, diferente das outras ômicas, ainda não há uma técnica única capaz de analisar a diversidade molecular dos metabólitos do metaboloma. Na prática, muitos estudos de metabolômica analisam apenas uma parte do metaboloma sendo que para algumas delas já foram propostos nomes específicos como glicômica, lipidômica e peptidômica.

A seguir a definição de alguns termos específicos utilizados em trabalhos da área de metabolômica:

- **Perfil metabólico (**Metabolic profiling**):** Análise quantitativa de um conjunto de metabólitos de uma rota biossintética selecionada ou de uma classe de compostos específicos. Inclui também "target analysis": a análise focada em compostos específicos.
- **Impressão digital metabólica (**Metabolic fingerprinting**):** Análise geral de metabólitos, não focada, para classificar amostras, baseada nos padrões dos metabólitos ou impressões digitais, que mudam em resposta a doenças, perturbações ambientais ou genéticas, com a finalidade de identificar marcadores.
- **Metabonômica (**Metabonomics**):** Não há diferença real entre Metabonômica e Metabolômica: o termo metabonômica é tradicionalmente mais usado nas pesquisas biomédicas, que têm como objetivo obter impressões digitais das perturbações bioquímicas causadas por doenças, drogas ou toxinas.

Em diversos estudos o uso de técnicas analíticas para a identificação e quantificação de metabólitos é rotina. A diferença destes estudos para a metabolômica é que esta tem como objetivo a análise de uma ampla gama de compostos, sem que haja necessidade de composto(s) alvo específico(s). Isto foi possível graças ao avanço das técnicas analíticas capazes, atualmente, de determinar um amplo espectro de compostos em uma única análise. Muitos estudos em metabolômica visam identificar, dentro deste amplo espectro, aqueles compostos que podem caracterizar determinado estado metabólico. Assim pode ser detectada rapidamente qualquer alteração no metabolismo. Usando esta estratégia já foram identificadas alterações no perfil metabólico características para diversas patologias humanas como cânceres, doenças cardíacas e diabetes (Rappaport, 2012). Estudos como esses apontam para compostos que possam servir para diagnosticar e monitorar estas doenças bem como para as vias metabólicas alteradas pela patologia.

A metabolômica também é utilizada para o monitoramento do metabolismo de organismos geneticamente modificados. Isto é particularmente importante para avaliar os efeitos da modificação sobre o metabolismo e por consequência a segurança do consumo destes organismos. Por exemplo, a introdução de genes para melhorar a qualidade nutritiva de alimentos tem sido o objetivo de muitas modificações e esta tecnologia já foi aplicada com sucesso em culturas como arroz e tomate. No entanto, devido à complexidade da rede metabólica, incluindo a sua regulação, nem sempre

são obtidas as modificações almejadas sendo acumulados pela planta compostos diferentes daqueles desejados. Assim procedimentos de segurança são necessários, e já são adotados, para avaliar se estes organismos são próprios para o consumo.

Ultimamente tem sido verificada a importância de alimentos, os chamados alimentos funcionais, que contenham novos ou quantidades maiores de certos compostos, como por exemplo, vitaminas, que não fazem parte do grupo considerado nutrientes como proteínas, carboidratos e gorduras. O consumo de alimentos de origem vegetal contendo compostos como fitoestrogênios que atuam no balanço hormonal, ou flavonoides que podem atuar como antioxidantes levam a perceptíveis decréscimos de certas patologias em populações que tradicionalmente incluem maiores quantidades destes alimentos em suas dietas. O valor do consumo destes compostos se deve à sua ação na manutenção da homeostase metabólica adequada, levando à promoção da saúde. Devido ao seu efeito benéfico, estes compostos receberam a classificação de "nutraceuticals". Estes compostos são encontrados em plantas, muitas vezes em concentrações relativamente baixas, e fazem parte do grupo de compostos classificados como metabólitos secundários. Muitos estudos de metabolômica visam verificar quantitativamente e qualitativamente estes compostos em, por exemplo, diferentes cultivares ou condições de cultivo, aumentando assim a qualidade dos produtos.

As aplicações citadas acima são apenas algumas de um número crescente de aplicações que podem ser verificadas na literatura. Deve ser salientado que a metabolômica é assunto de estudo recente e que, apesar de grandes avanços, as técnicas ainda estão em desenvolvimento. Ainda não há uma única técnica que possa ser aplicada para o estudo extensivo do metaboloma de um organismo e, portanto, várias técnicas complementares devem ser utilizadas para uma visão mais completa. Outro aspecto em desenvolvimento é a análise dos dados obtidos, já que a metabolômica, como as outras "ômicas", gera uma quantidade enorme de dados a serem trabalhados (uma única análise pode levar à separação de centenas de compostos). Sendo assim, há a necessidade de automatizar ao máximo a análise dos dados para detectar as diferenças entre conjuntos de cromatogramas e/ou espectros. Isso exige a comparação dos dados utilizando a análise multivariada. A automação da identificação de compostos com bancos de dados públicos ainda esbarra em dificuldades, como diferenças de condições de análises. É certo, no entanto que estas dificuldades serão superadas em um futuro próximo.

Os metabólitos

Metabólitos são micromoléculas, compostos de baixo peso molecular, quando comparado ao peso molecular de macromoléculas como DNA, proteínas ou amido. São formados por vias metabólicas diversas e, a julgar pela diversidade das vias, pode-se deduzir a enorme variedade estrutural encontrada nos metabólitos contidos na célula. Atualmente existem mais de 200 000 metabólitos conhecidos. Estes apresentam, além de grande variedade de propriedades químicas e físicas, diferenças em concentração enormes no organismo.

Metabólitos podem ser divididos em metabólitos primários ou secundários (Figura 2). Os metabólitos primários fazem parte das vias anabólicas e catabólicas comuns a muitos organismos. Exemplos são as vias biossintéticas dos ácidos graxos e o catabolismo da glicose (glicólise). Apesar de comuns a muitos organismos os metabólitos primários são muito diversos quanto à sua estrutura, pois não são simples polímeros lineares de unidades repetitivas como os ácidos nucléicos ou as proteínas. A análise destes metabólitos é útil para detectar distúrbios no metabolismo e assim entender melhor, por exemplo, patologias. Atualmente isto é feito rotineiramente em exames médicos como os de sangue. Neste exame, a comparação da concentração de metabólitos como glicose, colesterol, triglicerídeos e ácido úrico, no sangue do paciente com os valores normalmente encontrados na população, auxilia o médico a estabelecer o diagnóstico. Fica claro então que o nível de certos metabólitos reflete o estado metabólico do organismo.

Metabólitos secundários são compostos formados a partir de vias biossintéticas de distribuição relativamente restrita e os precursores destas vias derivam do metabolismo primário. O termo metabolismo secundário é mais frequentemente usado para vias metabólicas específicas de diversos organismos como bactérias, fungos e plantas sendo raramente utilizado para compostos do metabolismo animal. Exemplos

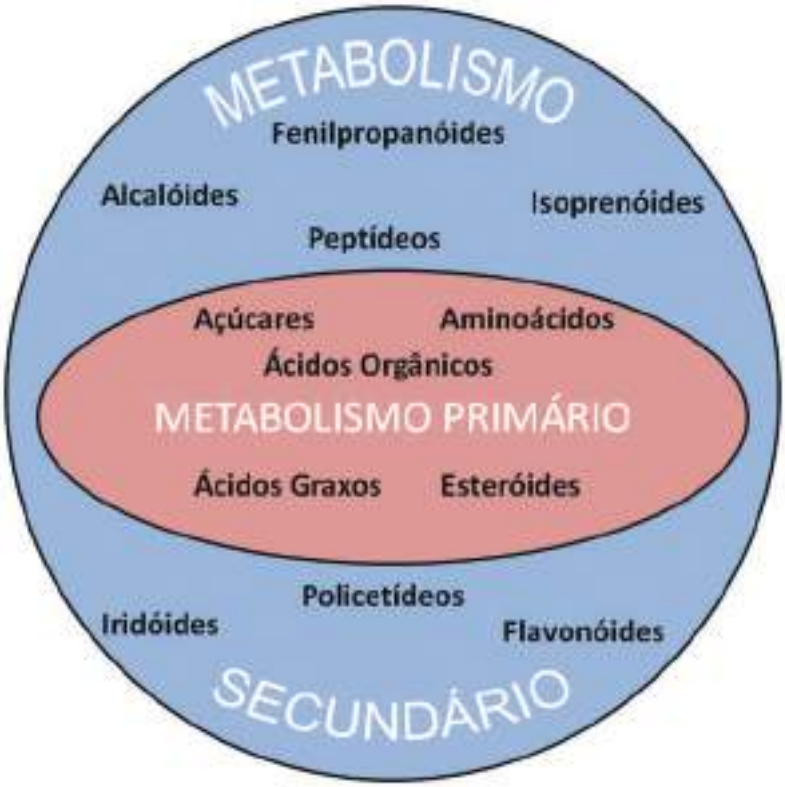


Figura 2. Metabolismo secundário como extensão do metabolismo primário.

de metabólitos secundários, produzidos por plantas e micro-organismos, são os alcaloides, uma classe de compostos, que contém diversas substâncias com ação no sistema nervoso de vertebrados. Os alcaloides indólicos, um subgrupo de alcaloides, têm como precursor o amino ácido triptofano. A biossíntese destes compostos está restrita a algumas famílias de Angiospermas. No passado acreditava-se que estes compostos não teriam função específica para o organismo. Atualmente sabe-se que muitos metabólitos secundários atuam, na interação do organismo com o ambiente. Por exemplo, como compostos de defesa química (ex: antioxidantes), física (ex: proteção contra a radiação) ou biológica (ex: proteção contra predação ou patógenos). Muitos medicamentos são metabólitos secundários extraídos de plantas ou microorganismos que pela complexidade estrutural não podem ser obtidos via síntese em laboratório de maneira economicamente viável. Sendo assim a análise desses compostos e dos intermediários de sua via biossintética permite entender melhor a regulação de sua biossíntese e como consequência auxilia o processo de obtenção destes produtos.

A seguir serão discutidas diferentes classes de compostos, suas propriedades e sua relevância na metabolômica.

Açúcares

Açúcares, também conhecidos como carboidratos, são encontrados em todos os seres vivos. Podem ser divididos em mono-, di-, oligo- e polissacarídeos (Figura 3). Os açúcares mais simples são os monossacarídeos que podem ser aldoses ou

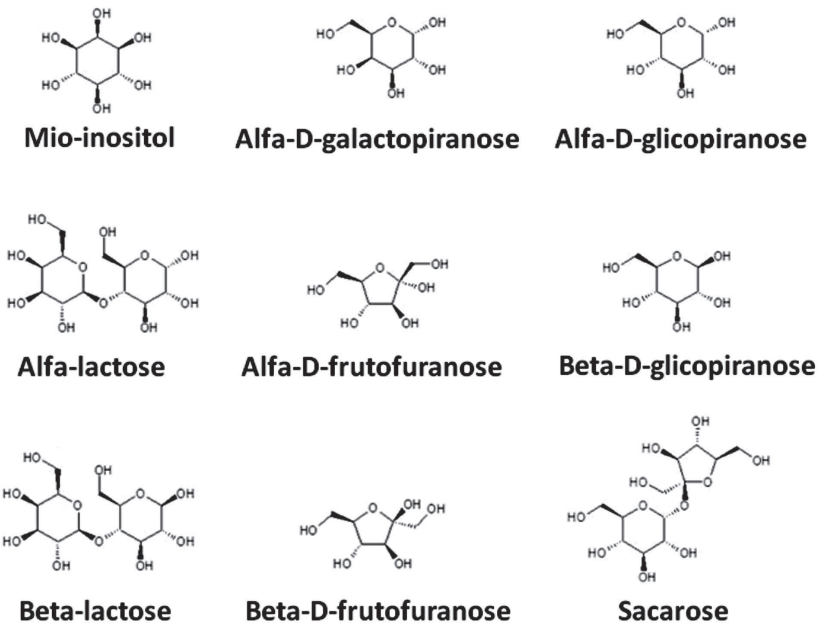


Figura 3. Estruturas de açúcares.

cetoses. Os monossacarídeos mais comuns contêm de três a sete carbonos e os mais frequentemente encontrados são a frutose, a glicose e a galactose (todas as hexoses). Os dissacarídeos mais conhecidos são sacarose (dímero de frutose e glicose) e lactose (dímero de galactose e glicose). Os carboidratos são importantes para estocagem de energia, na forma de mono-, di- ou polissacarídeos. A polimerização de glicose dá origem a diversos polissacarídeos que diferem entre si por serem lineares ou não e pelo tipo de ligação glicosídica formada. Plantas acumulam o polissacarídeo amido enquanto animais acumulam glicogênio para estocagem de energia. Polissacarídeos também são componentes estruturais como é o caso da celulose, um dos componentes da parede celular de vegetais. Além dos açúcares comuns do metabolismo primário há vários exemplos de açúcares incomuns e de distribuição restrita (Thibodeaux et al., 2007). Os mono-, di- e oligossacarídeos são solúveis em água, já os polissacarídeos frequentemente não são dissolvidos por solventes. Muitas vezes carboidratos fazem parte de moléculas mais complexas, de origem biossintética mista, como é o caso das pentoses ribose e desoxirribose, que fazem parte da estrutura dos nucleotídeos em RNA e DNA, respectivamente. No reino vegetal há muitos metabólitos secundários que ocorrem na forma de glicosídeos, por exemplo, flavonoides ou iridoides. Uma das subdivisões da metabolômica, a glicômica, se concentra na análise de todos os carboidratos livres e ligados covalentemente a proteínas e lipídeos.

Amino ácidos

Há em todos os seres vivos polímeros de L-amino ácidos, as chamadas proteínas, formadas pela combinação linear dessas moléculas nos ribossomos. São 20 os amino ácidos normalmente encontrados em proteínas (Figura 4). A proteômica tem como objetivo estudar vários aspectos do perfil protéico de organismos. Também são encontrados nos seres vivos polímeros mais curtos desses amino ácidos, os peptídeos (até 50 unidades) e oligopeptídeos (até 20 unidades), não sendo estes objetos de estudo da proteômica, devido ao seu baixo peso molecular. A peptidômica estuda estes peptídeos, que envolvem compostos com atividades importantes, como hormônios (por exemplo: insulina, glucagona e calcitonina) ou compostos antimicrobiais (como microcinas).

Alguns amino ácidos são também precursores para a biossíntese de outras classes de compostos tanto do metabolismo primário quanto do secundário. Podem ser mencionados os alcalóides indólicos e quinolínicos, derivados do triptofano, os flavonoides e taninos (compostos fenólicos), derivados de fenilalanina.

Peptídeos e oligopeptídeos podem conter os amino ácidos típicos de proteínas e também outros amino ácidos, os chamados amino ácidos atípicos. Muitos polímeros de amino ácidos não são formados via ribossomo. Principalmente em micro-organismos, vem crescendo o número de peptídeos identificados que são formados por peptídeo sintetases não ribossomais. Estes costumam conter amino ácidos atípicos ligados entre si, muitas vezes por ligações peptídicas também atípicas, sendo necessária a identificação rigorosa, inclusive da estereoquímica de cada amino ácido.

Todos os amino ácidos, como o nome já intui, contêm um grupamento amino e outro carboxílico. Nos amino ácidos típicos de proteínas estes grupamentos estão

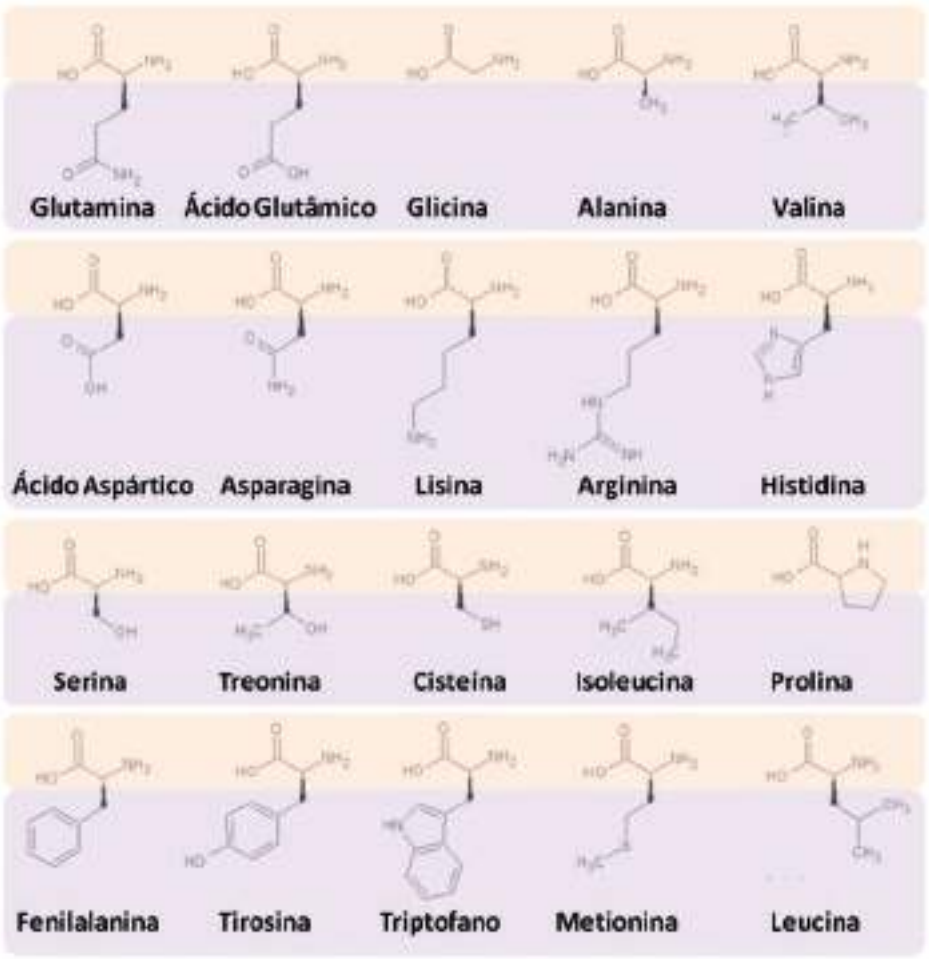


Figura 4. Estruturas dos 20 *L*-amino ácidos comuns.

ligados ao mesmo átomo de carbono. A configuração do carbono quiral pode ser L ou D sendo que, nos amino ácidos típicos de proteínas, a configuração é sempre L. De acordo com a estrutura das cadeias laterais, os amino ácidos são classificados como polar ou apolar. No entanto, devido às suas propriedades anfótericas, os amino ácidos podem ser bem dissolvidos na água, sendo que a solubilidade pode variar com mudanças de pH.

Amino ácidos livres estão presentes em muitos tecidos. Em plantas, frequentemente altos níveis de certos amino ácidos podem ser encontrados. Por exemplo, quando há alta disponibilidade de nitrogênio, a planta pode estocar o excesso de N na forma de arginina, glutamina ou asparagina, todos com a razão entre os números de N e C alta.

Ácidos graxos

A biossíntese de ácidos graxos ocorre em todos os seres vivos pela via do acetato-malonato, pois estes são componentes essenciais das membranas celulares. Os ácidos graxos diferem entre si quanto ao comprimento da cadeia de carbonos e o número, a posição e a configuração das suas insaturações (Figura 5). Os ácidos graxos mais comuns contêm número par de C e, quando presentes, insaturações de configuração cis. Ácidos graxos são importantes fontes de energia metabólica e são normalmente estocados na forma de triglicerídeos, ésteres formados a partir da reação de glicerol com três moléculas de ácido graxo. As gorduras de animais e muitos óleos vegetais, como azeite de oliva e óleo de soja, contêm quase que exclusivamente triglicerídeos. Além dos compostos que podem estar presentes em grandes quantidades nas células ou tecidos, como os citados acima, são encontrados derivados de ácidos graxos entre os quais muitos hormônios, tanto animais quanto vegetais. Exemplos são as prostaglandinas (derivadas do ácido araquidônico) e o ácido jasmônico (derivado de ácido-linolênico).

Ácidos graxos de cadeia longa e triglicerídeos são apolares e, portanto, praticamente insolúveis em água. Extratos de tecidos feitos com solventes apolares como clorofórmio e hexano contêm, quase que exclusivamente, ácidos graxos e/ou triglicerídeos.

Ácidos graxos em conjunto com os terpenóides são classificados como lipídeos. Há uma subdivisão da metabolômica interessada em caracterizar especificamente estas moléculas: a chamada lipidômica. Estes dois grupos de metabólitos apresentam em comum uma baixa polaridade, e, portanto, são extraídos com solventes orgânicos apolares. Apesar de estarem reunidos sob o termo lipídio sua origem biossintética é distinta: os ácidos graxos são formados pela via do acetato-malonato enquanto que os terpenos são formados a partir de unidades isoprênicas.

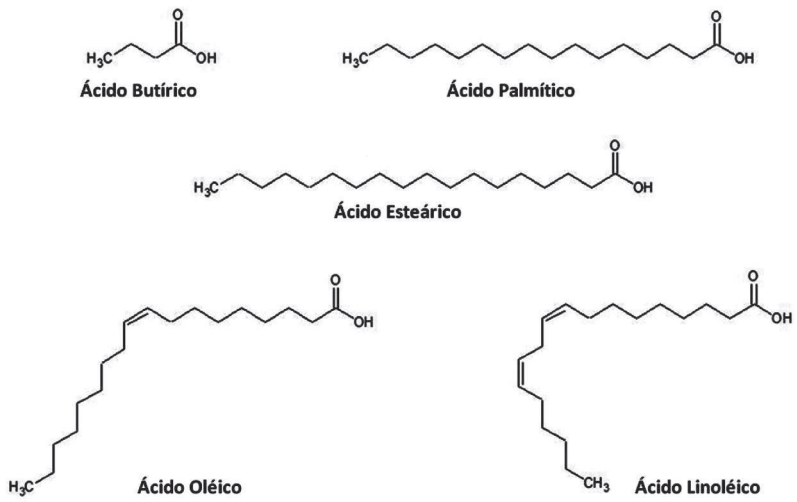


Figura 5. Estruturas de ácidos graxos.

Terpenóides

Muitos compostos do metabolismo primário são terpenos ou terpenóides. Por exemplo, os esteróides colesterol e ergosterol são derivados de triterpenos tetracíclicos (Figura 6). Fazem parte do grupo de triterpenos modificados, os esteróides, conhecidos hormônios, tanto animais (testosterona) como vegetais (brassinoesteróides). Outro exemplo é o grupo de compostos chamados de vitamina D formados também a partir de triterpenos tetracíclicos. Apesar de serem estruturalmente muito diversos, todos estes compostos são formados por uma sequência de reações que levam a junção linear de compostos de cinco átomos de C (derivados de isopentenil pirofosfato).

Por muito tempo considerou-se que o isopentenil difosfato era biossintetisado somente a partir do ácido mevalônico. No final dos anos 80 foi comprovado que as unidades isoprênicas também podem ser sintetizadas a partir de 1-deoxy-D-xilulose-5-fosfato (Rohmer, 1999). Os terpenos são classificados de acordo com o número de unidades de cinco carbonos que os compõe em hemiterpenos (5 C), monoterpenos (10 C), sesquiterpenos (15 C), diterpenos (20 C), triterpenos (30 C). A variedade estrutural do grupo é grande, pois apesar de se tratar de unidades isoprênicas ligadas entre si de forma linear, as possibilidades de ciclização e modificações das moléculas são inúmeras.

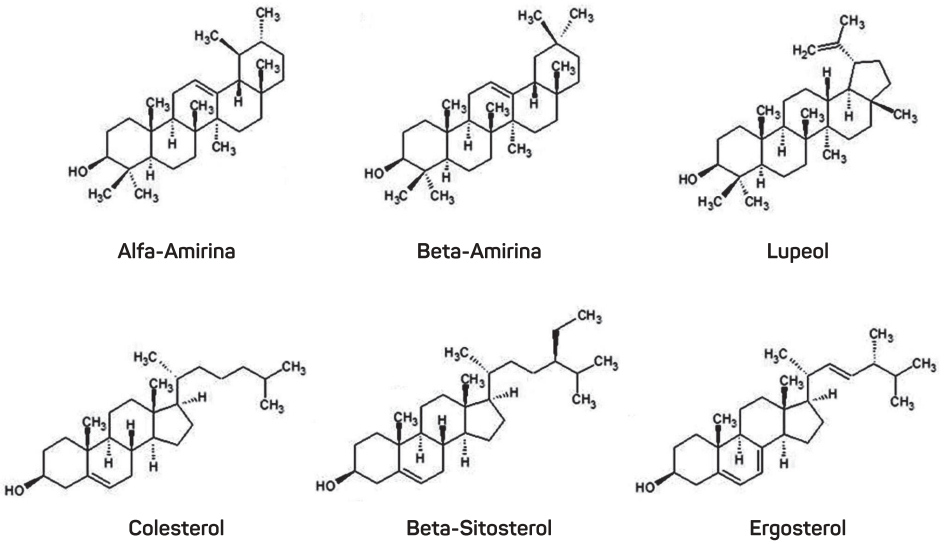


Figura 6. Estruturas de esteroides.

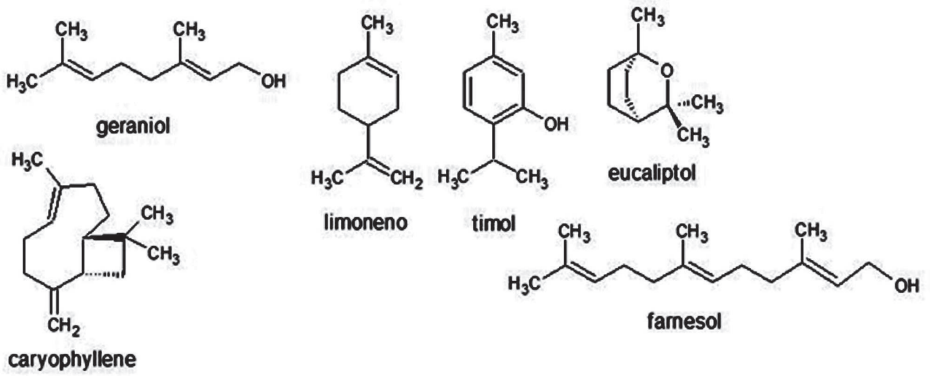


Figura 7. Estruturas de terpenos e terpenóides.

Os monoterpenos formam um grupo grande que contém muitos componentes voláteis dos óleos essenciais, como limoneno, carvacrol e mentol (Figura 7).

Sesquiterpenos são também constituintes importantes de óleos essenciais. Derivados oxigenados que tem atividades importantes, como lactonas sesquiterpênicas são encontrados em muitas plantas. Exemplos são artemisinina da *Artemisia annua*, e lactucina da alface (*Lactuca sativa*) (Figura 8). Diterpenos e diterpenóides são comuns em diversas famílias de plantas, e.g. paclitaxel (taxol) é um derivado de diterpenos. Os triterpenos lupeol e alfa- e beta-amirina têm uma ampla distribuição no reino vegetal (Figura 6). Carotenóides são tetraterpenos e são encontrados em altas concentrações em, por exemplo, a variedade de cenoura mais consumida. Carotenóides são também alguns dos pigmentos encontrados em flores e penas de pássaros. Carotenóides obtidos na ingestão de alimentos podem ser modificados sendo precursores de metabólitos como é o caso do β-caroteno, que é modificado e transformado em vitamina A, pelo nosso metabolismo.

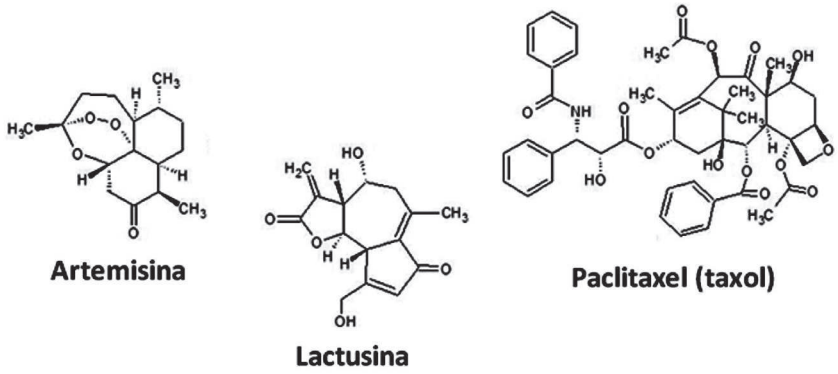


Figura 8. Derivados oxigenados de terpenos.

Os terpenóides não oxigenados ou pouco oxigenados são compostos apolares podendo ser extraídos com solventes como diclorometano. Precauções especiais devem ser tomadas para a análise dos metabólitos de baixo peso molecular, pois estes são voláteis e podem ser perdidos ao longo do processo de preparo de amostras.

Em certas famílias de plantas iridóides são comuns. São monoterpenos e geralmente ocorrem na forma de glicosídeos. São considerados importantes para defesa contra herbívoros ou microorganismos. O iridoide secologanina é um precursor dos alcaloides indólicos monoterpênicos. Devido à polaridade dos glicosídeos eles têm boa solubilidade em solventes polares.

Ácidos orgânicos

Muito comum são os ácidos carboxílicos que são intermediarios em muitas reações bioquímicas, por exemplo no ciclo de Krebs. Ácidos orgânicos comuns são o ácido acético, o cítrico, o fórmico, o láctico, o oxálico e outros (veja Figura 9). Estes compostos são geralmente compostos majoritários em extratos polares de qualquer tipo de organismo.

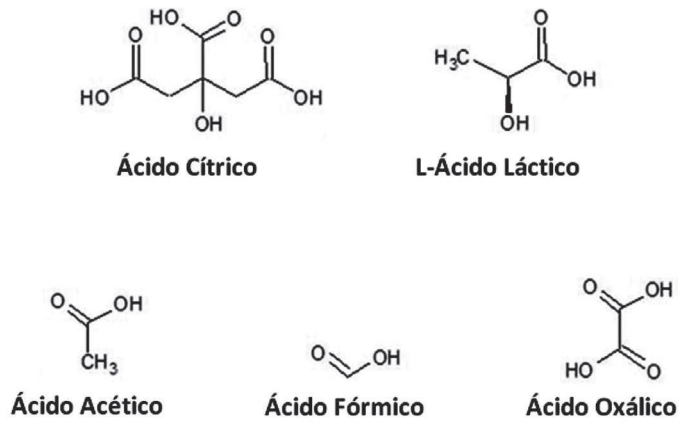


Figura 9. Estruturas de ácidos orgânicos.

Flavonoides

Flavonoides são compostos com ampla distribuição no reino vegetal (Figura 10). São comuns como pigmentos em flores. Eles são apontados como os princípios ativos de muitas plantas medicinais, e assim despertam grande interesse. A estrutura geral contém dois anéis aromáticos conectados a uma unidade de três carbonos, e existem muitas variações em substituintes. Biossinteticamente eles são de origem mista, formados pela condensação de um fenilpropanoide com um policetídeo de três unidades. Eles podem ocorrer na forma livre, ou como glicosídeos. A solubilidade de compostos individuais depende muito dos substituintes e pode variar bastante.

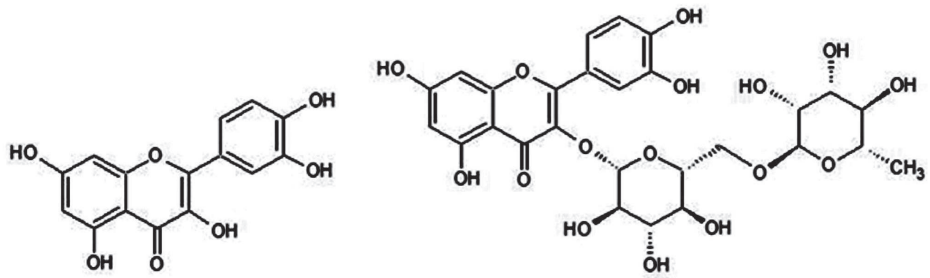


Figura 10. Estruturas de flavonoides.

Outros fenólicos

Além dos flavonoides há também outros fenólicos, por exemplo os fenólicos simples como o ácido salicílico, a vanilina, o ácido sinápico, cumárico entre outros (Figura 11). Estes compostos, considerados como metabólitos secundários, são geralmente derivados diretos da descarboxylação de fenilalanina. Para estes compostos foram relatadas entre outras atividades antibióticas. Estes Compostos têm normalmente uma boa solubilidade em solventes polares.

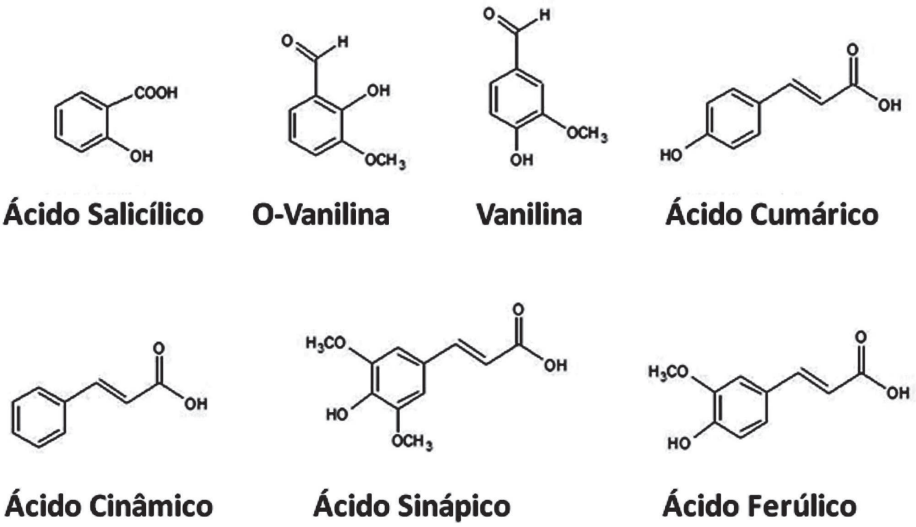


Figura 11. Estruturas de fenólicos.

Alcaloides

Alcaloides são compostos naturais que normalmente contém átomos de nitrogênio na forma básica, dentro uma estrutura cíclica (Figura 12). Não há nenhuma definição exata dos compostos considerados alcalóides e a distinção entre alcalóides e outros compostos nitrogenados muitas vezes não é clara. Um grande grupo de metabólitos secundários e muitos dos compostos biologicamente ativos de plantas medicinais são alcalóides, por exemplo cafeína, morfina, cocaína, atropina, quinina, estricnina e nicotina. Devido ao caráter básico dos alcalóides, extratos específicos destes compostos podem ser obtidos com extração ácido-base.

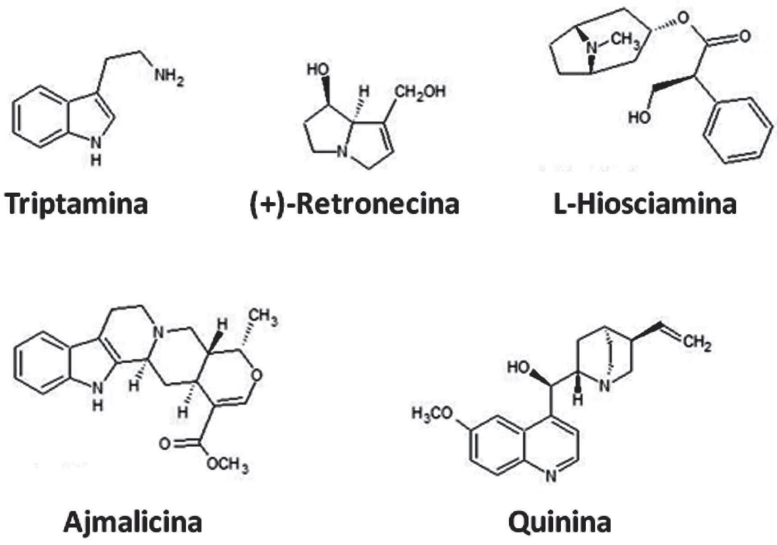


Figura 12. Estruturas de alcaloides.

Das quantidades

Assim como a diversidade de metabólitos em um organismo é enorme o mesmo pode-se dizer da concentração destes compostos no organismo. Mesmo intermediários das principais vias do metabolismo primário como, por exemplo, a glicólise ou o ciclo de Krebs, podem não acumular, pois, apesar de presentes, são metabolizados rapidamente. Além das diferenças devido ao metabolismo enzimático, há diferenças inerentes à função do metabólito. Sendo assim açúcares ou ácidos graxos que participam do metabolismo energético são, usualmente, encontrados em concentrações muito superiores a, por exemplo, da maioria dos hormônios ou vitaminas. Mesmo os compostos encontrados em concentrações relativamente baixas podem ser importantes marcadores para distúrbios no metabolismo. Portanto, detectar diferenças que possam existir em componentes minoritários de um extrato é mais um dos desafios nos estudos de metabolômica.

Previsão dos metabolitos presentes na amostra

Nos organismos vivos há uma série de vias metabólicas. Nestas vias há a transformação sucessiva de substratos em produtos catalisada por enzimas. Sendo assim, teoricamente, seria possível encontrar no organismo todos os substratos e produtos de suas reações enzimáticas. Como muitas das vias metabólicas são comuns aos organismos, pode-se prever parte do seu metaboloma. Esta previsão pode ser feita mesmo que este organismo não tenha sido estudado anteriormente. Por exemplo, em toda planta (com raríssimas exceções), espera-se encontrar todas as vias do metabolismo primário típicas de um organismo eucariótico fotoautotrófico do grupo das plantas e, portanto pelo menos parte dos intermediários destas vias. Pode-se também derivar seu metabolismo secundário ao nível de classe de compostos encontrados. De uma planta da família das *Apocynaceae*, por exemplo, há grande chance de encontrar alcalóides indólicos e iridóides. Pode ser feita também uma busca bibliográfica dos compostos já identificados em determinado organismo ou grupo de organismos. Neste caso é bom sempre lembrar que para muitos organismos pode haver diferentes nomes, especialmente plantas, o que exigirá a identificação de sinônimos.

Apesar de ser possível prever uma grande parte dos metabólitos presentes em um organismo pode haver grandes variações, entre grupos ou até dentro a mesma espécie. Isso devido a mudanças no ambiente, o parte do organismo estudado ou outros fatores. Essas diferenças normalmente são mais pronunciadas no metabolismo secundário. Assim não há como prever, por exemplo, se uma determinada cepa de *Microcystis* (cianobactéria) produz microcistinas ou outros polipeptídeos não ribossomais, sendo necessária sempre uma análise química para verificar a sua presença. Em *Microcystis* já foi verificado que a produção ou não de microcistinas se deve ao genótipo da cepa.

A gama de metabólitos encontrados em um organismo pode variar conforme o meio ambiente ao qual o organismo está exposto (variação fenotípica). Em fungos, por exemplo, foi verificado recentemente que existem vias metabólicas no genoma que não são geralmente expressas pelo organismo, as chamadas vias metabólicas crípticas ou silenciosas (Zazopoulos et al. 2003). A existência dessas vias foi proposta pelo fato, de apesar de contidas no genoma, os seus metabólitos não serem encontrados no organismo. Por exemplo, o produto de uma policetídeo sintase pode ser previsto conhecendo-se a sequência de genes do lócus, mas, frequentemente, em condições de cultivo usuais, o policetídeo não é detectado. Os mecanismos que induzem a expressão dessas vias não são bem conhecidos, no entanto, o reconhecimento de sua existência abriu uma nova estratégia de busca por metabólitos secundários de interesse econômico. Fica claro que a gama de metabólitos encontrados pode também variar conforme o meio ambiente ao qual o organismo está exposto. Estas também podem ser induzidas ou reprimidas caso os mecanismos de controle sejam conhecidos, alterando assim a gama de compostos acumulados. Em plantas são conhecidos os chamados fitoalexinas, que são compostos produzidos em resposta a um ataque de um patógeno (Hammerschmidt, 1999).

Deve ser enfatizado que os compostos encontrados em um organismo nem sempre derivam de suas próprias vias metabólicas. Compostos podem ser adquiridos pela alimentação ou pelo convívio com outros organismos. Alguns dos compostos

exógenos podem ser de vital importância para a sobrevivência do organismo. Exemplos evidentes de compostos indispensáveis ao metabolismo de humanos são os amino ácidos essenciais, assim chamados porque há a necessidade de sua ingestão, já que sua biossíntese não ocorre em nosso organismo. Outro exemplo são as vitaminas que também devem ser obtidas na alimentação. Portanto fica evidente que nem todos os metabólitos encontrados no corpo humano derivam de nossas próprias vias metabólicas. Não há porque supor que o mesmo não ocorra em outros os organismos heterotróficos. Existem também muitos exemplos de acúmulo de compostos sintetizados por outros organismos que não são vitais à existência dos seres que os acumulam, mesmo em organismos autotróficos, aqueles capazes de sintetizar todos os compostos orgânicos necessários para a sua sobrevivência. A grande maioria destes compostos é acumulada devido à ingestão ou convivência estreita do organismo produtor com o que os acumulam como as toxinas de dinoflagelados, acumuladas por moluscos bivalvos. Também há diversos besouros que acumulam metabólitos das plantas que eles usam como alimento (veja, por exemplo, Kuhn et al. 2004).

Coleta e extração da amostra

Em estudos de metabolômica há a necessidade de comparação da identidade e quantidade dos metabólitos em diversas amostras. Portanto é essencial utilizar métodos bem padronizados para minimizar o efeito de variações que se devam à metodologia. Estes cuidados na padronização vão desde a coleta da amostra até a sua análise.

Logo após a coleta é necessário impedir que reações metabólicas ocorram, por exemplo, extraindo as amostras imediatamente com solvente orgânico, desnaturando as enzimas. No caso das amostras serem armazenadas, deve-se congela-las e eventualmente liofiliza-las até que sejam processadas.

A metodologia de extração deve ser cuidadosamente escolhida e avaliada para que os objetivos do trabalho possam ser alcançados. A etapa de extração tem como objetivo não só a inativação de enzimas como a separação das micromoléculas (metabólitos) do resto dos componentes macromoleculares da amostra, a chamada matriz. Durante o processo as células são rompidas e os metabólitos solubilizados. O tipo de matriz varia. Por exemplo, em tecidos vegetais os metabólitos devem ser separados da principal macromolécula, a parede celular cujos componentes são celulose além de outros polissacarídeos e fenólicos. Em tecidos animais deve ser eliminada nesta etapa a maior parte da proteína da amostra. Os metabólitos extraídos são separados da matriz por centrifugação ou filtração permitindo então a sua análise.

Quando há necessidade de extração das amostras deve se ter em mente que nenhum solvente é capaz de dissolver todos os metabólitos. Isto porque a diversidade estrutural encontrada nos metabólitos traz consigo diversidade em propriedades físicas incluindo a sua solubilidade (Figura 13). Por exemplo, existe uma grande diferença de polaridade nos metabólitos encontrados nas células: estes vão desde os hidrossolúveis, como os açucares, até os lipossolúveis, como os esteróides.

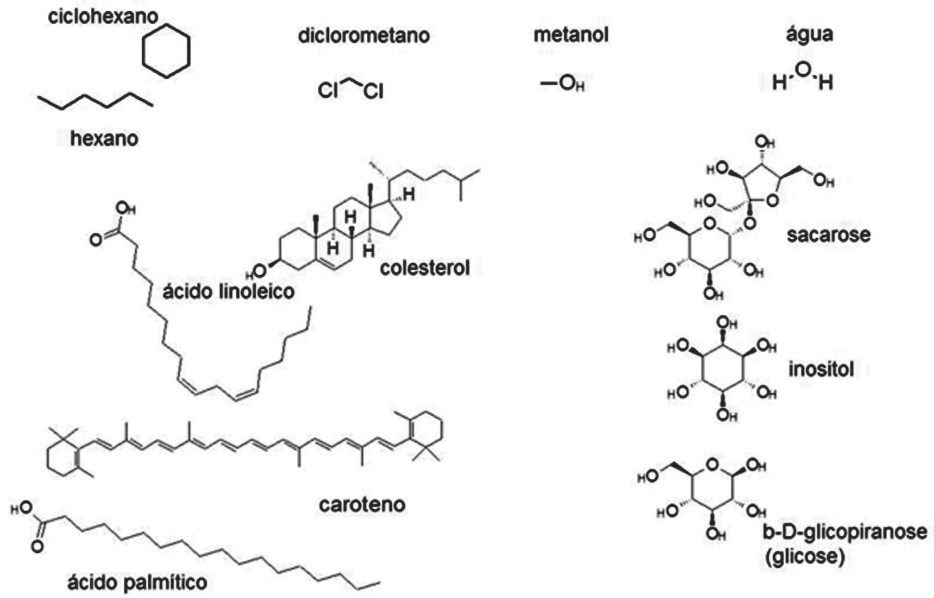


Figura 13. Solventes dissolvem especialmente os compostos com polaridades parecidos.

Polaridade e solubilidade

A polaridade é uma das propriedades de uma ligação química. Ela é dependente da eletronegatividade dos átomos que participam da ligação. A eletronegatividade é definida como a capacidade do átomo, estando ligado a outro, de atrair elétrons para si. Quando dois átomos de eletronegatividades distintas estão ligados, a distribuição dos elétrons no orbital molecular não é uniforme e, portanto a ligação é polar. De maneira geral, pode-se concluir que quanto maior a diferença de eletronegatividade dos átomos ligados maior será a polaridade da ligação. Na Tabela 1, a listagem das eletronegatividades de elementos comumente presentes nos metabólitos e solventes usados nas extrações.

Tabela 1. Eletronegatividades de elementos comumente presentes nos metabólitos e solventes usados nas extrações.

elemento	eletronegatividade
H	2,20
C	2,55
N	3,04
O	3,44
P	2,19
S	2,55
Cl	3,16

A polaridade de uma molécula depende da diferença de eletronegatividade dos elementos que a compõe e da sua forma. Por exemplo, a molécula de CO_2 não é polar porque, apesar da diferença de eletronegatividade do C e do O, a molécula é linear e, portanto, simétrica. Já na água a diferença de eletronegatividade dos elementos que compõe a molécula e a sua geometria faz com que esta seja polar.

O grau de solubilidade de um soluto em um solvente depende do balanço de forças inter-moleculares entre solvente e solvente, soluto e soluto e solvente e soluto. Utiliza-se a generalização "igual dissolve igual" significando que um solvente polar dissolverá melhor solutos polares e vice versa (veja Figura 13). Sendo assim não há solvente que dissolva todos os metabólitos, pois estes podem ser tanto polares quanto apolares.

Seja qual for o solvente escolhido para a extração, ele revelará um perfil específico de metabólitos, pois extrairá com melhor eficiência, os metabólitos que nele sejam mais solúveis. Sendo assim é importante frisar que este fato deve ser levado em consideração em análises quantitativas.

Para aumentar a diversidade dos metabólitos extraída é possível usar uma sequência de solventes de polaridades crescentes. Outro fator a ser considerado é a quantidade de solvente utilizado na extração. Por mais solúvel que compostos sejam em determinado solvente, se o volume usado na extração não for o suficiente ocorrerá saturação no solvente que levará a erros na etapa de quantificação dos metabólitos. No caso de análises por ^{1}H RMN deve-se considerar a possibilidade de extrair a amostra direta com solventes deuterados. Apesar de mais caros, a extração nesses solventes permite a análise direta com RMN de ^{1}H.

Para a escolha de um solvente para a extração, diferentes fatores devem ser levados em consideração além da solubilidade dos compostos: o seu grau de pureza, a possibilidade de gerar artefatos, a facilidade de manuseio e o método de análise a ser utilizado posteriormente. A pureza do solvente deve ser alta para evitar a contaminação dos extratos por compostos que não pertençam à amostra. Além disso, impurezas no solvente podem levar à formação de artefatos. Fatores como ponto de ebulição e toxicidade influenciam a facilidade com que as extrações são realizadas. Alguns solventes reagem com certos compostos e isto deve ser levado em consideração durante a escolha. Por fim deve ser levada em consideração a toxicidade do solvente dando preferência àqueles menos tóxicos e/ou menos poluentes.

Os fatores mais críticos, que devem ser padronizados para minimizar variações na etapa de extração, são: o tipo de solvente (eventualmente mistura de mais de um solvente) a proporção entre material extraído e solvente (peso/volume), o tempo de extração, a etapa de concentração e o armazenamento dos extratos até a análise. Como já mencionado anteriormente, o volume de solvente utilizado deve ser suficiente para evitar a sua saturação pelos solutos. Também o volume deve ser o menor possível para diminuir o tempo necessário para concentração/secagem dos extratos e minimizar a concentração de impurezas contidas solvente. O método de concentração e secagem dos extratos depende do solvente utilizado. Para solventes orgânicos utiliza-se uma centrífuga a vácuo que permite a secagem simultânea de muitos extratos. É necessário padronizar as condições de secagem evitando a exposição a temperaturas altas que podem levar à degradação de compostos. No caso de extratos aquosos um método adequado é a liofilização que, como ocorre a baixas temperaturas, minimiza a possibilidade de degradação. Quando secos os extratos devem ser armazenados a

baixas temperaturas e luminosidade sendo necessário, dependendo da estabilidade dos extratos, minimizar os efeitos da atmosfera oxidante substituindo-a por N_2.

Os grupos trabalhando com metabolômica geralmente desenvolvem um procedimento padrão para processar as amostras. Cada grupo tende a adaptar estes protocolos básicos de acordo com o material investigado. Sendo assim são inúmeros os protocolos descritos na literatura para a extração de amostras. O importante é que, ao final, seja possível a comparação entre o grupo tratado e o controle.

Às vezes é possível evitar a etapa de extração. Grupos trabalhando com a análise de amostras de humanos, como urina e sangue, normalmente analisam as amostras diretamente. Estes grupos têm interesse em investigar marcadores de estados metabólicos (patologias) em fluidos facilmente amostrados. Nestes casos é importante a desativação de enzimas e, eventualmente, a remoção de proteínas ou outros componentes que possam prejudicar a análise.

As técnicas analíticas

Como descrito acima a metabolômica lida com uma ampla gama de compostos com uma grande variedade de propriedades químicas e físicas. Não ocorre o mesmo com as outras "ômicas", a genômica, a transcriptômica e a proteômica, que analisam moléculas relativamente uniformes, tanto quanto na sua estrutura quanto nas suas propriedades físicas e químicas. Sendo assim fica evidente que não há um procedimento padrão para o estudo do metaboloma.

A intenção de um experimento de metabolômica é verificar e identificar as diferenças qualitativas e/ou quantitativas de metabólitos entre grupos de amostras. O objetivo pode ser, por exemplo, investigar a influência de fatores ambientais, ou a influência da introdução de novos genes no metabolismo de um organismo. Dada a complexidade da rede metabólica fica claro então que não se trata de tarefa fácil e depende de desenvolvimento de metodologia prática e precisa para agilizar a obtenção de resultados confiáveis. Na metabolômica para detectar diferenças entre grupos de amostras são obtidas as impressões digitais metabólicas. O objetivo final é a identificação das diferenças moleculares e de identificar o maior número de compostos responsáveis por estas diferenças, e isso com um mínimo de investimento de mão de obra, tempo e custo. Em princípio tanto as técnicas cromatográficas como técnicas espectroscópicas são capazes de fornecer impressões digitais do metabolismo, e assim podem ser usadas em metabolômica. É possível realizar estudos de metabolômica somente com cromatografia líquida ou cromatografia gasosa. Também pode ser feito com somente espectrometria de massas ou ressonância magnética nuclear. Atualmente as técnicas mais usadas são Ressonância Magnética Nuclear (RMN), Cromatografia Líquida acoplada à Espectrometria de Massas (LC-MS) ou Cromatografia Gasosa acoplada à Espectrometria de Massas (GC-MS). Os seguintes fatores são importantes para a escolha da técnica mais adequada:

- A quantidade de compostos a ser analisado
- A reprodutibilidade da técnica.
- A linearidade da resposta quantitativa.
- A facilidade da preparação das amostras.

Em cromatografia, compostos de uma mistura são separados de acordo com a afinidade com as fases, a fase estacionária e a fase móvel. Cada pico na cromatograma corresponde com um composto, e a quantidade do composto é refletida na intensidade do pico. O método de detecção determina se um composto é detectado e com qual sensibilidade. Grandes diferenças na resposta de compostos individuais são comuns e para a quantificação de compostos individuais o fator de resposta do composto deve ser conhecido ou determinado.

Todo método cromatográfico tem como objetivo separar compostos contidos em uma amostra. Existem várias técnicas cromatográficas desenvolvidas com o objetivo de separar compostos com características distintas. Assim a cromatografia gasosa é a mais adequada para a separação de compostos apolares e de baixo peso molecular e, portanto, mais voláteis. Já a cromatografia líquida é geralmente utilizada para compostos, ou mais polares e/ou de maior peso molecular ou que não possam ser analisados por cromatografia gasosa devido às altas temperaturas utilizadas durante a corrida.

Espectrometria de massas normalmente não é usada isoladamente para analisar extratos brutos em experimentos de metabolômica. Isso se deve a algumas limitações da técnica. Primeiramente é necessária a ionização dos compostos para possibilitar a sua detecção. Na ionização em extratos brutos há uma grande possibilidade de certos compostos interferirem na ionização de outros, um processo chamado *supressão iônica* (ion suppression). Assim normalmente a espectrometria de massas é usada em combinação com alguma técnica cromatográfica.

Em GC-MS a ionização costuma ser por impacto de elétrons, fornecendo portanto espectros bem característicos para os compostos. Em certas classes de compostos, como os terpenos, é comum ter muitos isômeros sendo difícil distinguir seus espectros de massa. No entanto, quando há um padrão do composto disponível existe a possibilidade de verificar a co-eluicão com o padrão para verificar o tempo de retenção. Muito raramente isômeros têm espectros de massa e tempos de retenção iguais.

Nem todas as moléculas são voláteis o suficiente para serem analisadas por cromatografia gasosa. Isto foi parcialmente solucionado derivatizando as amostras: estas são submetidas a reações químicas (normalmente silanização) diminuindo a polaridade dos compostos tornando-os mais voláteis. Permanece, no entanto a limitação quanto ao peso molecular máximo que pode ser analisado por cromatografia gasosa. Além disto, a técnica é inadequada para compostos termicamente instáveis devido às altas temperaturas necessárias para a cromatográfica gasosa.

O segundo método cromatográfico frequentemente acoplado ao espectrômetro de massas é a cromatografia líquida. A primeira dificuldade técnica que surgiu foi o acoplamento do espectrômetro de massas a um método cromatográfico cuja fase móvel é líquida e cujos fluxos de eluente eram na ordem de 1 ml/min. Atualmente estas dificuldades foram superadas e há vários tipos de interfaces entre o cromatógrafo e espectrômetro de massas que são adequadas para a análise de uma variedade ampla de compostos. Um desenvolvimento mais recente disponibilizou cromatógrafos líquidos e colunas cromatográficas no mercado que permitem trabalhar a fluxos baixíssimos (UHPLC), diminuindo enormemente as dificuldades inerentes ao acoplamento de equipamentos usuais de LC-MS.

A interface entre a cromatografia liquida e o espectrômetro de massas mais utilizada é a eletrospray. Nesta interface o líquido ao entrar no espectrômetro recebe uma carga elétrica e é evaporado, levando a transferência de carga para as moléculas a serem analisadas. Este método de ionizaçao é bem mais suave que a ionização por impacto de elétrons e, consequentemente, há pouca fragmentação da molécula e portanto pouca informação estrutural além do peso molecular. Com MS-MS a fragmentação pode ser induzida, mas os bancos de dados deste tipo de espectro de massas ainda são bem mais limitados quando comparados aos bancos de dados de EI-MS. Como no caso do GC-MS, a coeluição de um padrão pode ser utilizada para conferir o tempo de retenção e o espectro de massas sendo suficiente para a identificação.

Diferente que em LC-MS e GC-MS, RMN de ^{1}H fornece espectros que refletem as quantidades exatas dos compostos visíveis. Nestes espectros são observados todos os átomos de hidrogênio de todos os compostos presentes em quantidades suficientes. Sendo assim, um espectro de RMN de um extrato contém o somatório de todos os espectros dos componentes individuais sendo que boa parte dos espectros estão sobrepostos. Portanto, geralmente não e possível identificar mais de algumas dezenas de componentes dentro dos espectros do extrato sendo que estes, na maioria das vezes, são os componentes principais da amostra. A identificação dos componentes é geralmente mais confiável com RMN do que com espectrometria de massas, porque o espectro completo esta presente. A preparação das amostras é mais simples e há uma maior reprodutibilidade. A grande desvantagem de RMN de ^{1}H é o fato de que somente algumas dezenas de compostos podem ser observadas, enquanto com as outras técnicas centenas de compostos podem ser analisados (Veja Figuras 14 e 15 para exemplos).

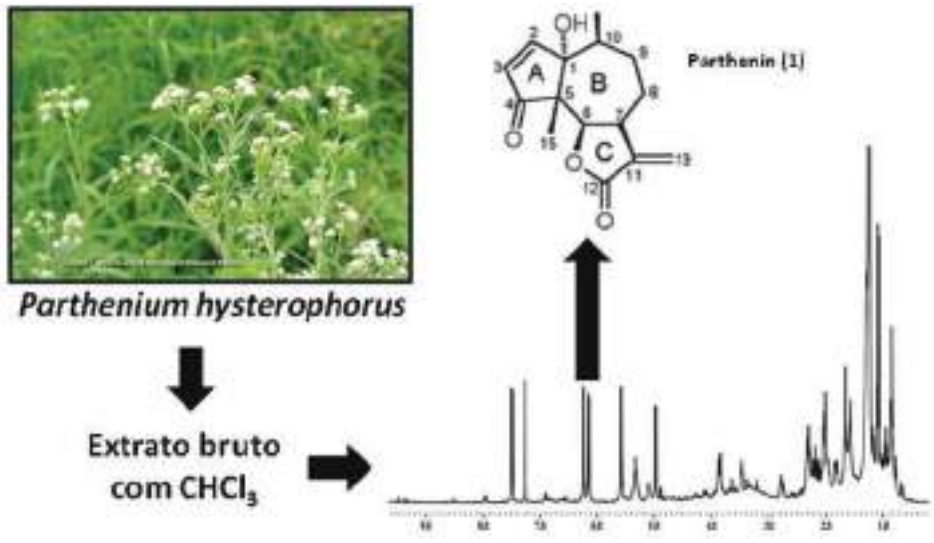


Figura 14. Espectro de RMN de ^{1}H de um extrato bruto feito com $CHCl_3$. O espectro mostra os sinais de lipídeos (sinais altos na região de 1 a 2 ppm), mas também sinais altos de um metabolito secundário, a partenina.

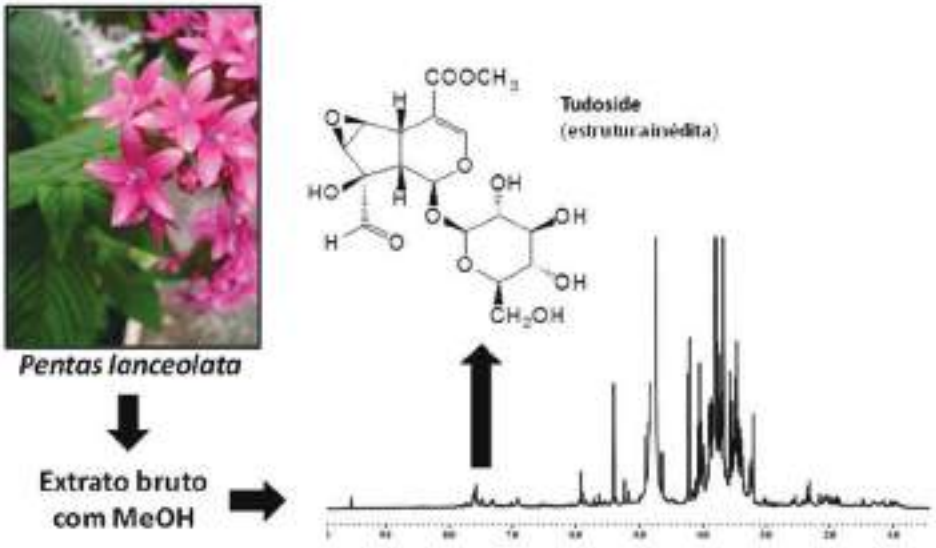


Figura 15. Espectro de RMN de ^{1}H de um extrato bruto feito com metanol. O espectro mostra os sinais altos de carboidratos (especialmente sacarose), mas também típicos de um iridoide (sinais perto de 7,6 e 5,9 ppm). O iridoide em questão foi identificado depois isolamento com cromatografia liquida preparativa como um composto inédito, que recebeu o nome tudosídeo.

Como em qualquer experimento, todos os fatores que possam introduzir variações indesejáveis devem ser minimizados ou controlados. Isso inclui uniformizar todos os parâmetros como o tipo e a coleta de material, e a preparação e o armazenamento das amostras mantendo a variação dentro de cada um dos grupos a menor possível. Só assim será possível detectar diferenças entre os grupos tratados e controle. Como os experimentos em metabolômica visam verificar diferenças quantitativas e qualitativas de metabólitos, são dois os fatores críticos do experimento: a extração das amostras e a sua análise.

Na obtenção da impressão digital metabólica é importante que os picos correspondam com compostos específicos e que não haja confusão em relação a identidade dos picos.

Dados adicionais obtidos junto com a cromatografia podem ajudar bastante para chegar a uma identificação dos picos: os máximos de absorção em UV, o espectro de massas, os tempos de retenção. Especialmente compostos conhecidos podem ser identificados logo após a obtenção da impressão digital metabólica com RMN, LC-MS ou GC-MS, usando bancos de dados.

Processamento de dados

Após a obtenção dos espectros e/ou cromatogramas, os dados das diferentes amostras devem ser comparados. Para analisar os conjuntos de dados e verificar se existem diferenças significantes entre os conjuntos, a análise multivariada é necessária.

Idealmente, cada amostra deve fornecer uma listagem contendo a identificação de todos os compostos presentes, com suas respectivas quantidades, mas isso ainda é uma utopia: com GC-MS ou LC-MS uma grande parcela dos metabólitos pode ser detectada, mas tanto a identificação quanto a quantificação são complexas.

Com RMN somente os compostos majoritários podem ser detectados, mas com este método a quantificação é facilitada pelas características da própria técnica. Na comparação também deve ser levada em consideração a possível distorção de sinais ou picos.

Num experimento de metabolômica diferentes fases podem ser distinguidas (Figura 16, Van den Berg et al. 2006). Após o experimento dados brutos são obtidos. Em LC-MS ou GC-MS os dados são tempos de retenção, intensidades dos picos, e espectros de massas. Em análises de RMN de ^{1}H são obtidos os deslocamentos químicos dos picos e as intensidades.

Figura 16. Diferentes fases num experimento de metabolômica com ênfase no processamento dos dados.

Os dados brutos das análises precisam passar por diferentes processos para obter os dados limpos. Este processo é chamado de pré-processamento dos dados e pode envolver os seguintes passos:

- Adaptação da escala dos picos: geralmente padrões internos são incluídos na análise para obter o adaptar a escala. Todos os espectros ou cromatogramas individuais devem passar por este processo aonde a intensidade é ajustada com a ajuda da intensidade dos picos do padrão interno.
- Alinhamento dos picos: Em cromatografia é comum pequenas variações no tempo de retenção de picos individuais devido à variação de fatores externos ou internos. Isso pode ser devido a temperatura da análise, a qualidade da coluna ou outros compostos presentes no extrato, que influenciam por exemplo o pH. (veja Tistaert et al. 2011). Em RMN o alinhamento é feito com uso do padrão interno para o deslocamento químico, mas nesta técnica também, devido à pequenas variações de pH, picos podem se deslocar em relação a outros, tornando o alinhamento complicado.

- "Binning or bucketing". Este processo é muito usado em espectros de RMN. Em "binning" o espectro é dividido em regiões e todas as regiões são integradas. Em RMN de ^{1}H são geralmente utilizadas regiões com uma largura de aproximadamente 0,04 ppm. Isso reduz os 16 K pontos do espectro para somente 250 pontos. O resultado é uma enorme perda de resolução, mas a comparação dos espectros é facilitada. A grande vantagem deste processo é que pequenas variações no deslocamento por causa de fatores como pH, concentração ou temperatura são eliminadas, mas ao mesmo tempo informação de pequenos sinais são perdidas.
- "Peak picking". Neste processo todos os picos são registrados com as suas intensidades. Antes deste processo é essencial fazer um alinhamento dos picos.
- Deconvolução: O melhor procedimento para extrair os dados de espectros é por deconvolução, assim que os componentes individuais são reconhecidos pelas relações com outros sinais. Em RMN o procedimento foi chamado de targeted profiling (Weljie et al. 2006). Nele o espectro é modelado matematicamente tendo como base os espectros dos componentes individuais.

Depois destes procedimentos são obtidos os dados limpos, que ainda devem passar por um processo chamado pré-tratamento dos dados (Figura 16). Este pré-tratamento tem influência direta nos resultados da análise multivariada. Os passos envolvidos são os seguintes:

- "Centering". Neste processo intensidades ou concentrações são convertidas para flutuações em volta de zero. Desse modo somente a variação é considerada para a análise.
- Ajustamento da escala de cada componente pela divisão do valor por um fator específico para cada um. Isso é necessário para possibilitar a comparação entre compostos presentes em altas concentrações com outros em baixas concentrações. Diferentes métodos existem, cada um com vantagens e desvantagens (auto-, range-, pareto-, vast-, ou level scaling). (Van den Berg et al. 2006).
- Transformação dos dados: Pode ser usado uma transformação logarítmico ou "power". Ao mesmo tempo esta transformação da também uma mudança na escala.
- Os processos de pré-tratamento tem influencia grande nos resultados da análise multivariada e podem ser usados para aumentar a importância dos metabolitos menos abundantes nestas análises.

Os dados limpos são submetidos à análise multivariada. As técnicas mais comuns são PCA (Principal Component Analysis) e PLS-DA (Partial Least Squares - Differential Analysis).

PCA é um procedimento matemático que converte o conjunto das observações num conjunto de variáveis linearmente não correlacionadas, os componentes principais. São verificados os componentes principais que mostram a maior variabilidade, dentro do conjunto total de dados.

Em PLS-DA a correlação dos dados com variáveis independentes é verificada. Neste caso existe um grande risco de obter correlações não-existentes, especialmente quando a quantidade de variáveis é maior que o numero de amostras (Broadhurst e Kell, 2006). Isso é chamado "over-fitting": Quando há muitas variáveis e relativamente poucas amostras, algumas variáveis podem mostrar separação de classes por acaso.

Assim sempre o processo de PLS-DA deve ser acompanhado por um processo de validação dos resultados.

Como em todos os métodos estatísticos, muito cuidado deve ter tomado com a interpretação dos resultados. Já foi escrito que na maioria das vezes conclusões de experimentos não são justificadas levando a falsas suposições.

A identificação dos metabolitos

A análise multivariada leva idealmente à indicação dos sinais (compostos) que diferenciam os conjuntos de amostras. Em seguida é essencial que os compostos, responsáveis por estes sinais sejam identificados para gerar as hipóteses da origem dos fatores que causam as diferenças entre os conjuntos de amostras.

Isso também é essencial para excluir a possibilidade de "over-fitting" ou a possibilidade que as diferenças foram causadas pelo desenho experimental.

Na identificação de metabólitos dois casos podem ser distinguidos:

- O composto é conhecido e a estrutura deve ser confirmada.
- O composto é desconhecido e a estrutura deve ser elucidada e confirmada.

Para chegar a identificação primeiro devem ser verificadas quais informações estão disponíveis das analises anteriores: Das análises com LC-MS ou GC-MS, há o tempo de retenção no sistema cromatográfico, e os dados de espectrometria de massas, a massa molecular e eventualmente um espectro de massas mostrando fragmentação.

Em análises de RMN o deslocamento químico do sinal indica o tipo de hidrogênio, e eventualmente já pode indicar o composto, especialmente quando há informações adicionais sobre outros hidrogênios na molécula.

Com estes dados muitas vezes já é possível sugerir a identidade, certamente quando se trata de um composto comum. Para confirmar a identidade de um composto em RMN, primeiramente a presença de todos os seus sinais deve ser verificada (Veja Figura 17).

Quando a estrutura sugerida é de um composto comum ou conhecido, e há padrão disponível, a identidade poderia ser confirmada pela adição do padrão à amostra e verificar a coincidência exata de todos os sinais do composto com os do padrão. Em RMN não poderá haver duplicação ou alargamento dos sinais. Além isto o aumento da intensidade de cada um dos sinais do composto deve ser idêntica.

Quando não há padrão disponível e se trata de um composto com sinais claros no espectro de RMN, espectros adicionais de RMN podem ser obtidos da mesma amostra para fornecer informações adicionais, como o espectro de RMN de ^{13}C, ou espectros bidimensionais como COSY, HSQC ou HMBC. Quando isso não é possível por se tratar de compostos minoritários ou quando há muito sobreposição de sinais, o composto deveria ser isolado. O isolamento do composto permite a obtenção de vários espectros, não só de RMN, mas de outros métodos físicos, como espectros de ultravioleta, infravermelho, rotação ótica e eventualmente outros dados. Na Figura 18 estão listadas as informações que periódicos exigem para publicação de estruturas inéditas.

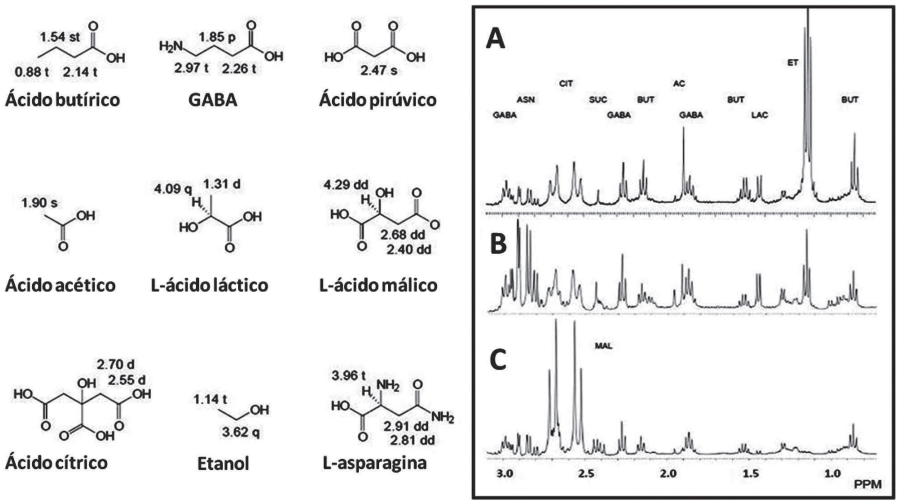


Figura 17. Fragmentos dos espectros de RMN de ^{1}H de amostras de mamão. (A) Fruto supermaduro. (B) Fruto com o distúrbio fisiológico de gelificação. (C) Fruto maduro normal. A identidade dos diferentes sinais está indicada. GABA: acido gama-aminobutirico, ASN: Asparagina, CIT: citrato, SUC: succinato, BUT: butirato, AC: acetato, LAC: lactato, ET: etanol, MAL: malato.

Compound characterization: Physical and spectroscopic data for **new** compounds must be comprehensive, and follow the order shown below: compound name (and assigned number in text); physical state of compound (e.g. oil, crystal, liquid, etc.), melting and/or boiling point; optical rotation and/or circular dichroism measurements, if optically active; UV; IR, ^{1}H NMR; ^{13}C NMR; MS. For all new compounds, either high-resolution mass spectral or elemental analysis data are required. See later section for method of data presentation.

Figura 18. As exigências da revista "Phytochemistry" para a publicação de compostos inéditos.

Para a elucidação estrutural de compostos desconhecidos, especialmente espectros bidimensionais de RMN são muito importantes. Os seguintes tipos de espectros são os mais usados:

- 2D-COSY: mostra as correlações entre átomos de hidrogênio devido aos acoplamentos.
- 2D-HSQC: mostra as correlações entre os átomos de hidrogênio e os átomos de carbono com os quais eles têm ligação.
- 2D-HMBC: mostra as correlações entre os átomos de hidrogênio e os átomos de carbono que tem entre si duas ou três ligações.
- 2D-NOESY: mostra as correlações entre átomos de hidrogênio que estão espacialmente próximos um do outro. A correlação não depende de ligações entre os átomos em questão.

Espectrometria de massas também e importante para a elucidação estrutural porque fornece informações acerca da fórmula molecular e do arranjo dos átomos na molécula. Informações a cerca do arranjo dos átomos são obtidas após a fragmentação da molécula o chamado espectro de massas (Figura 19). O íon mais frequente gera

o sinal mais alto e é chamado de pico base. Apesar de ser uma técnica destrutiva, a quantidade de amostra necessária para realização das análises é na ordem de nano- ou picogramas e, portanto, tão pequena que na maioria das vezes esta desvantagem é apenas secundária.

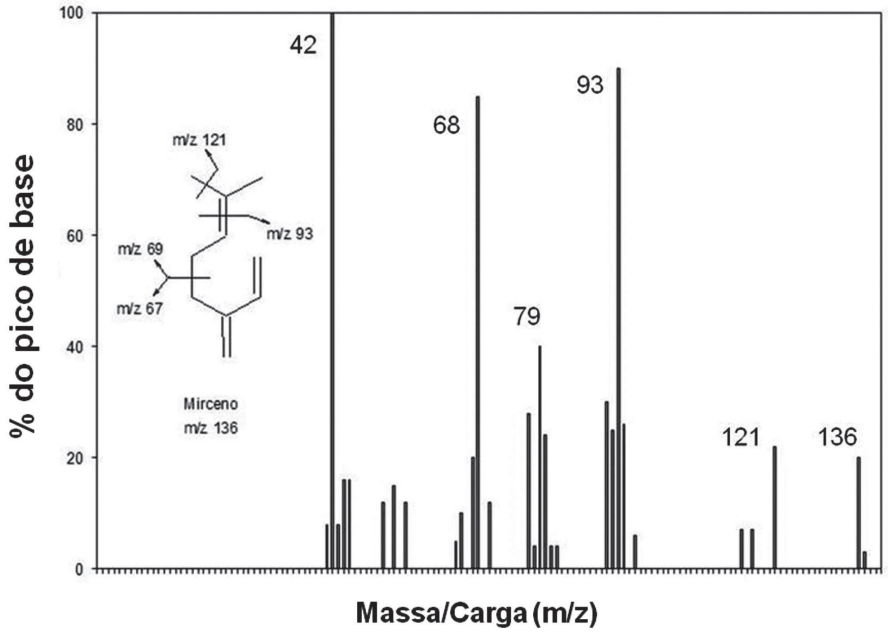


Figura 19. O espectro de massas de mirceno, obtido por Impacto de Elétrons (EI-MS).

Quando o espectro apresenta íon molecular a fórmula molecular pode ser derivada quando são utilizados espectrômetros de alta resolução, pois as massas dos átomos que compõe a molécula analisada não são integrais. Portanto a massa do íon molecular é o somatório das massas dos isótopos mais abundantes de cada elemento e a abundância de cada elemento pode ser calculada. Para obtenção da fórmula molecular é importante a grau de resolução do espectrômetro de massas. Quanto maior a exatidão na determinação da massa, menores serão as possibilidades de fórmula molecular para cada massa nominal. Também a medição exata da intensidade dos picos de isótopos facilita a determinação da formula molecular, porque os elementos têm isótopos com abundâncias naturais bem definidas.

O espectro de massas também fornece informação sobre o arranjo dos átomos em uma molécula, pois o padrão de fragmentação pode ser racionalizado. São várias as técnicas utilizadas para induzir a fragmentação de uma molécula que, dependendo da energia utilizada, levarão a graus diferentes de fragmentação.

A probabilidade de uma ligação ser quebrada depende, entre outros fatores, da força desta ligação e da estabilidade dos fragmentos gerados. Existe bibliografia específica em que este processo é explicado em profundidade. Também existem bancos

de espectros de massa que facilitam a identificação de compostos por comparação sendo que os bancos mais úteis foram gerados com ionização por impacto de elétrons.

Pelo descrito acima fica claro que a elucidação estrutural dos compostos no extrato geralmente é a parte mais complexa do experimento de metabolômica, mas estudos sem a identificação formal dos compostos não tem sentido. (Scalbert et al. 2009)

Bibliografias

BROADHURST D, KELL D. 2006. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics* 2: 171-196.

HAMMERSCHMIDT, R. 1999. Phytoalexins: What have we learned after 60 years? Annu. Rev. Phytopathol. 37: 285-306.

KUHN, J., PETTERSSON, E.M., FELD, B.K., BURSE, A., TERMONIA, A., PASTEELS, J.M., and BOLAND, W. 2004. Selective transport systems mediate sequestration of plant glucosides in leaf beetles: A molecular basis for adaptation and evolution. Proc. Natl. Acad. Sciences. 13808-13813.

OLIVER SG, WINSON MK, KELL DB, BAGANZ F. 1998. Systematic functional analysis of the yeast genome. *Trends Biotechnol* 16: 373-378.

RAPPAPORT, SM. 2012. Discovering environmental causes of disease. Journal of Epidemiology and Community Health 66: 99-102.

ROHMER, MICHEL. 1999. The discovery of a mevalonate-independent pathway for isoprenoid biosynthesis in bacteria, algae and higher plants. *Nat Prod Rep* 16: 565–574.

SCALBERT A., BRENNAN, L., FIEHN, O., HANKEMEIER, T., KRISTAL, B.S., VAN OMMEN, B., PUJOS-GUILLOT, E., VERHEIJ, E., WISHART, D., WOPEREIS, S.. 2009. Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research. *Metabolomics*, 5: 435-458.

THIBODEAUX, CJ, MELANCON, CE, LIU, HW. 2007. Unusual sugar biosynthesis and natural product glycodiversification. Nature 446: 1008-1016.

TISTAERT, C., DEJAEGHER, B., and VANDER HEYDEN, Y. 2011. Chromatographic separation techniques and data handling methods for herbal fingerprints: A review. Anal. Chim. Acta 690: 148-161.

VAN DEN BERG, R.A., HOEFSLOOT, H.C.J., WESTERHUIS, J.A., SMILDE, A.K., and VAN DER WERF, M.J. 2006. BMC Genomics 7: 142-156.

WELJIE AM, NEWTON J, MERCIER P, CARLSON E, SLUPSKY CM. 2006. Targeted profiling: Quantitative analysis of 1H NMR metabolomics data. *Anal. Chem*. 78: 4430-4442.

ZAZOPOULOS, E., HUANG, K., STAFFA, A., LIU, W., BACHMANN, B.O., NONAKA, K., AHLERT, J., THORSON, J.S., SHEN, B. and FARNET, C.M. 2003. A genomics-guided approach for discovering and expressing cryptic metabolic pathways. Nature Biotechnology 21, 187-190.

11

Epigenômica

Paulo de Paiva Amaral
Gonçalo Castelo-Branco

"O método pelo qual tecidos e órgãos diferenciam-se no curso do desenvolvimento é atualmente o maior espaço em branco no mapa da Biologia... Sabemos pouco sobre os precisos passos dados pelos processos epigenéticos, os fatores bioquímicos envolvidos e, acima de tudo, o que determina a especificidade reprodutível dos tecidos diferenciados." (Huxley, 1957)

Introdução

Princípios de regulação epigenética

O sequenciamento de genomas tem como objetivo central a identificação completa dos elementos funcionais presentes no DNA de um organismo ou de um determinado tipo celular, tal como células tumorais (Capítulos 2 e 4). No estado atual do conhecimento, acredita-se que todas as células em um indivíduo de uma determinada espécie tenham a mesma constituição genética. No entanto, tal como os diferentes cantores em um coro lendo a mesma partitura de acordo com sua voz e papel na canção, cada tipo de célula lê o seu genoma, interpretando as regiões regulatórias de seus genes e decidindo como e se os expressam (transcrição). A expressão específica de determinados genes em cada célula leva a um fenótipo adequado à função da célula no contexto do organismo, resultando assim em uma composição harmoniosa. A coordenação requerida para atingir esta composição é notável dado que, enquanto um coral possui em geral algumas

dúzias de músicos, um organismo complexo como um mamífero possui trilhões de células, cada uma com sua idiossincrasia.

Assim sendo, estudos genômicos são normalmente acompanhados de análise "Transcriptômica", na qual o conjunto de genes transcritos ("transcriptoma") nos tipos celulares de interesse é definido pelo sequenciamento de mRNAs (Capítulo 8). Os genes e as regiões genômicas que controlam sua expressão são essenciais para a transmissão hereditária de fenótipos celulares presentes no organismo. Estes elementos também são frequentemente alterados em condições patológicas, tais como cânceres e diversas doenças genéticas.

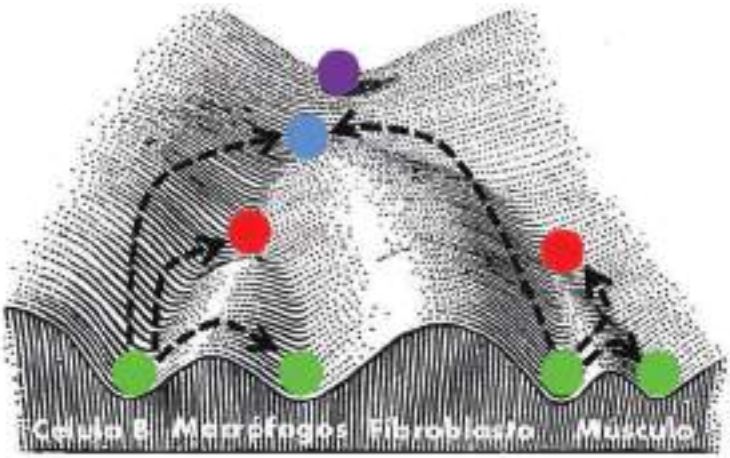
O que é epigenética?

A diferente interpretação do genoma pelas células é definida pelo seu estado epigenético, que está em constante alteração durante o desenvolvimento do organismo e mesmo em um estado final de diferenciação, onde modula as interações da célula com o ambiente circundante. O termo Epigenética (prefixo grego "epi", acima ou sobre) foi proposto originalmente por Conrad Waddington em 1942, descrevendo-a como sendo "o ramo da Biologia que estuda as interações causais entre os genes e seus produtos, e que levam à manifestação do fenótipo" (Waddington, 1942). Mais tarde ilustrou suas ideias com o conceito de "Paisagem" ou "Relevo" Epigenético (Figura 1), no qual células eram equiparadas a bolas rolando de uma montanha com superfície irregular e percorrendo caminhos alternativos ladeira abaixo, determinados pelo relevo da montanha. Em analogia, no processo de expansão e diferenciação das células originalmente presentes no embrião, as células adotam diferentes rotas que levam aos diferentes fenótipos, apesar de todas elas possuírem o mesmo conteúdo genético.

Todos os diferentes tipos celulares presentes no organismo adulto são derivados de uma única célula, o zigoto. Esta célula pode igualmente originar um novo organismo por divisão celular e tem, deste modo, a característica única de ser totipotente (capaz de originar todos os tipos celulares e de gerar um organismo). Ao longo do desenvolvimento do organismo, há uma perda de potência ou capacidade de originar diferentes tipos celulares, passando-se por estágios de pluripotência (células-tronco embrionárias, também conhecidas como células estaminais, que podem dar origem a todos os tipos celulares do organismo adulto, mas não a um novo organismo); multipotência (células-tronco fetais, neonatais ou adultas, que só têm capacidade de diferenciar-se na sua linhagem embrionária); unipotência; até ao estágio final de diferenciação, em que a célula se especializa na função a exercer no organismo adulto. No topo da montanha do revelo de Waddington, encontramos o zigoto totipotente, com um estado epigenético definido. À medida que há uma perda de potência durante o desenvolvimento, ocorrem alterações no estado epigenético, de modo a permitir a expressão de certos genes e a repressão de genes não associados ao fenótipo da célula.

Apesar de cunhada em 1957, a frase inicial deste capítulo, do evolucionista Julian Huxley, ainda é bastante atual. No entanto, nas últimas duas décadas, o papel dos processos epigenéticos e os fatores bioquímicos que transmitem e modulam essa camada de informação têm sido elucidados rapidamente. O estado epigenético é determinado fundamentalmente pela camada de informação imposta sobre a

sequência de nucleotídeos do genoma, na forma de modificações covalentes das proteínas que se ligam a DNA, as histonas, ou diretamente às bases do DNA, sobre os quais este capitulo irá focar. É um assunto em aberto se esta camada de informação pode ser considerada por si só epigenética ou se requer um caráter hereditário. Para alguns cientistas, um estado epigenético consiste em informação não codificada pela sequência de DNA que é transmitida à próxima geração por divisão celular, por mitose (no contexto de um indivíduo) ou mesmo por meiose (pela linha germinativa, para gerações seguintes).



Potencial de desenvolvimento	*Status* Epigenético
Totipotente Zigoto.	(i) Metilação global do DNA
Pluripotente Células ICM/ES; Células EC; Células GS; Células iPS.	(i) Apenas cromossomos X ativos; (ii) Repressão global da diferenciação de genes por proteínas Polycomb; (iii) Metilação de promotores.
Multipotente Células tronco adultas Células parcial reprogramadas (células pré-iPs)	(i) Inativação do cromossomo X; (ii) Repressão de genes de linhagem específica por proteínas Polycomb; (iii) Metilação de promotores.
Unipotente Tipos celulares diferenciados.	(i) Inativação do cromossomo X; (ii) Desrepressão de linhagem de genes silenciados por proteínas Polycomb; (iii) Metilação de promotores.

Figura 1. "Paisagem" ou "Relevo" Epigenético de Waddington (Waddington, C. H. *The Strategy of the Genes. A Discussion of Some Aspects of Theoretical Biology* (Alen & Unwin, 1957)). Epigenetic reprogramming and induced pluripotency. Konrad Hochedlinger and Kathrin Plath Development 136, 509-523 (2009) doi:10.1242/dev.020867 . ICM - Células da massa celular interna do blastocito; ES - Células tronco embrionarias; GS - Células pluripotentes da linhagem germinativa; iPS- Células pluripotentes induzidas.

O estado epigenético depende de interações com o ambiente. A sua manutenção depende de sinais do ambiente circundante, através dos sistemas de processamento e "retro-alimentação" dessas informações com o controle de expressão gênica, conhecidos como *feed back loops* (Berger et al., 2009). A transição para outro estado epigenético pode ser determinado por uma alteração do ambiente, o que, considerando transmissão epigenética através da linha germinativa, leva a possibilidade de que a teoria de Jean-Baptiste Lamarck de transmissão hereditárias de características adquiridas durante a vida de um organismo possa ocorrer em casos específicos. Durante o ciclo de vida de um mamífero, a maior parte da informação epigenética é transientemente apagada em dois estágios: nas células primordiais germinativas, que vão dar origem aos gametas, e durante a fertilização, no zigoto. No entanto, é possível que alguma informação epigenética seja mantida apesar do processo de reprogramação. Em plantas, um exemplo deste fenômeno é a paramutação, um processo no qual o estado epigenético de um alelo de um determinado lócus afeta o estado epigenético do outro alelo, sendo esta interação hereditária (Chandler e Stam, 2004).

Nos últimos anos, fenômenos deste gênero foram descritos em espécies animais, como a transmissão hereditária de hipertrofia cardíaca em camundongos (Wagner et al., 2008). Outros exemplos são a transmissão hereditária da cor do pelo através do lócus agouti (Morgan et al., 1999), de respostas metabólicas a dieta (Carone et al., 2010; Ng et al., 2010), longevidade (Greer et al., 2011) ou resposta antiviral (Rechavi et al., 2011) em camundongos ou no verme *Caenorhabditis elegans*. Em humanos, a transmissão de informação epigenética parece estar na base de alguns casos individuais de doenças como câncer colorretal hereditário não-poliposo e as síndromes de Prader-Willi e Angelman (Migicovsky e Kovalchuk, 2011). Um aspecto emergente em muitos destes casos é a intervenção de RNA não-codificadores de proteínas (RNAs regulatórios) ou de metilação de DNA como veículos epigenéticos (ver "Regulação da estrutura da cromatina").

Cromatina

Como é definida a informação epigenética a nível molecular? As fibras de DNA encontram-se no núcleo da célula organizadas numa estrutura denominada cromatina. Se estendido, o DNA armazenado em uma célula humana teria o comprimento de cerca de 2 metros, o que equivale a aproximadamente 200 mil vezes o diâmetro médio de um núcleo celular. A associação do DNA com proteínas formando a cromatina permite a contorção do DNA em diversos graus de compactação (Figura 2A). Para além de resolver o problema do armazenamento do DNA em um espaço tão pequeno como o núcleo da células, a organização em cromatina permite igualmente o controle da acessibilidade de proteínas a diferentes partes do genoma. Deste modo, a cromatina é a base estrutural central de informações epigenéticas.

Para adquirir a organização em cromatina, o DNA em organismos eucariotos associa-se a proteínas altamente conservadas evolutivamente denominadas histonas. As histonas têm um caráter básico (de carga positiva) devido à presença de aminoácidos como lisina e arginina, o que facilita a interação com as fitas negativas de DNA. Cinco tipos principais de proteínas histonas ocorrem em células eucarióticas, e se organizam

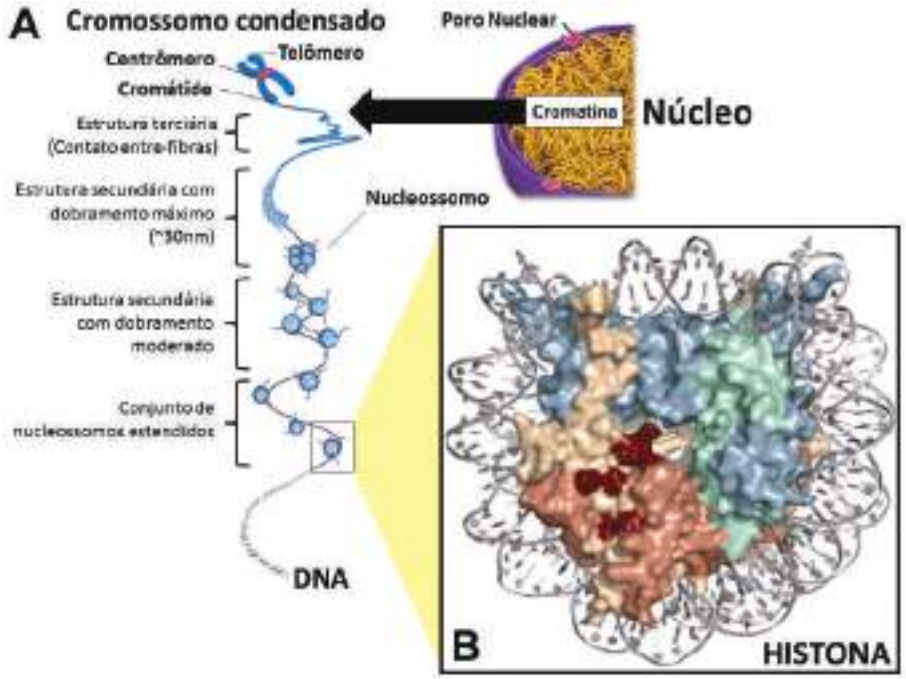


Figura 2. (A) Estrutura da cromatina. (B) Estrutura do nucleossomo. Adaptado de http://micro.magnet.fsu.edu/cells/nucleus/chromatin.html e Chromatin structure depends on what's in the nucleosome's pocket. Tamara L Caterino and Jeffrey J Hayes, Nature Structural and Molecular Biology 14, 1056 - 1058 (2007).

em unidades básicas chamadas de "nucleossomos". Os nucleossomos correspondem a octâmeros de histonas contendo duas moléculas de cada uma das histona "centrais" (H2A, H2B, H3 e H4), que são organizadas em dois heterodímeros de H2A-H2B e um heterotetrâmero de H3-H4, em volta dos quais cerca de 147 bases de DNA são enroladas (Figura 2B) (Xhemalce B, Dawson M, Bannister, AJ, 2011). Além disso, histonas H1, conhecidas como "histonas de ligação", associam-se às regiões entre os nucleossomos, as quais atuam como "espaçadores" e possibilitam a compactação da cromatina em níveis superiores.

O estado de compactação é relacionado à função da cromatina, impondo restrições quanto à acessibilidade ao DNA e desempenhando um papel importante em todos os processos DNA-dependentes nas células, como a transcrição, replicação e reparo de danos no DNA. Em termos gerais, estabelecidos originalmente com base em observações de microscopia de luz, a cromatina pode encontrar-se em dois estados:

- **Eucromatina** (observada de forma translúcida em microscopia), em que a densidade de nucleossomos é menos elevada, permitindo que fatores de transcrição recrutem RNA polimerase e permitindo produção de RNA.
- **Heterocromatina** (zonas do núcleo mais elétron-densas em microscopia), um estado de alta condensação do DNA e histonas, no qual há uma elevada densidade em redor do gene, não tendo a maquinaria de transcrição acesso ao gene e não

havendo produção massiva de RNA. Existem dois tipos de heterocromatina, a facultativa, que ocorre dentro de regiões eucromáticas do genoma, de modo a reprimir a expressão de determinados genes; e a constitutiva, presente em parte do genoma com elevado grau de sequências repetitivas do DNA, como telômeros e centrômeros (ou centrossomos). A heterocromatina constitutiva é essencial para a estabilidade e integridade do genoma (Beisel e Paro, 2011).

A "estrutura da cromatina" é determinada também pela associação do DNA com proteínas "não histonas", as quais modulam os estados de compactação locais e globais da fibra de cromatina. Algumas dessas proteínas são responsáveis pela regulação enzimática das histonas e dos nucleossomos.

Regulação da estrutura da cromatina

Sabe-se que diferentes processos atuam em sinergia para controlar a estrutura da cromatina, de modo a determinar o estado epigenético da célula, entre os quais os principais são: (a) modificação do DNA por metilação; (b) modificações covalentes de histonas; (c) substituição de histonas por variantes; (d) remodelamento de nucleossomos dependente de ATP. Mais recentemente, RNAs não-codificadores têm emergido com importantes reguladores da cromatina.

Modificação do DNA por metilação

Resíduos de citosina no DNA podem ser alvos da metilação, a adição de um grupo -CH_3 na posição 5 da (pirimidina) citosina (5mC, ou 5-metil-citosina). Em geral, a metilação de DNA ocorre em regiões de alto conteúdo de dinucleotídeos CG, denominados de "ilhas CpG" (do Inglês "Cytosine—phosphate—Guanine"), as quais ocorrem com pouca frequência e de forma não aleatória no genoma, estando frequentemente associadas a sequências promotoras. A metilação de DNA em regiões promotoras é classicamente associada ao silenciamento da expressão gênica. Descobertas mais recentes indicam que metilação pode estar igualmente presente no corpo dos genes (isto é, entre a região promotora e o sitio de terminação da transcrição) e pode também estar associada a genes ativos, sem a presença de ilhas CpG (Lister et al., 2009). A metilação de DNA é igualmente importante no processo de *imprinting*, no qual a expressão de uma determinada zona do genoma ocorre somente do alelo paterno mas não do materno, ou vice-versa (ou seja, expressão monoalélica). Neste casos, regiões especificas nessa zona do genoma encontram-se hipo-metiladas num dos alelos (permitindo expressão) e hiper-metiladas no outro alelo (reprimindo a expressão) (Koerner et al., 2009).

A metilação do DNA é controlada por uma família de proteínas denominadas Dnmts (*de novo DNA metilases*). Dnmt1 é responsável pela manutenção da metilação durante a replicação do DNA durante o ciclo celular. Até recentemente, pensava-se que metilação de DNA era irreversível, mas vários estudos indicam que este não é o caso, especialmente durante as fases iniciais do desenvolvimento, quando há reprogramação epigenética (Wu e Zhang, 2011), e também notavelmente no cérebro adulto (Guo et al., 2011). Proteínas da família Tet (*Ten-eleven translocation*) têm um

papel importante no processo de "de-metilação" do DNA, através da conversão de 5mC em 5 hidroximetil-citosina (5hmC), 5-formil-citosina (5fC) e 5-carboxil-citosina (5caC). Apesar de suas funções ainda não estarem bem elucidadas, estas novas formas de citosina podem estar igualmente envolvidas em estados epigenéticos específicos (Wu e Zhang, 2011). Alterações no perfil de distribuição da metilação de DNA estão implicadas no desenvolvimento de diversas doenças, especialmente em câncer, por exemplo em casos que envolvem a metilação anormal de genes supressores tumorais (Esteller, 2007).

Modificações covalentes de histonas

As histonas centrais contêm nos seus terminais amino (N) e carboxi (C) peptídeos protuberantes que podem ser alvo de modificações pós-traducionais. Há uma enorme variedade de modificações covalentes que podem ocorrer nas "caudas" das histonas, as quais contribuem decisivamente para a estrutura e regulação da cromatina. Ocorrem predominantemente na porção N-terminal das histonas centrais H3 e H4, em vários resíduos de aminoácidos, e incluem acetilação, metilação, fosforilação, ADP-ribosilação, sumoilação e ubiquitinação, entre outras (Xhemalce B, Dawson M, Bannister, AJ, 2011) (exemplos na Tabela 1). Novas modificações têm sido identificadas na parte globular das histonas e igualmente nas variantes de histonas (ver abaixo).

Estas modificações exercem a sua função por dois mecanismos principais. Por um lado, podem levar a alteração da carga do aminoácido onde ocorrem. A acetilação de lisinas leva a perda da carga positiva do aminoácido (Figura 3), o que leva a uma menor interação entre as histonas e o DNA e a uma configuração mais eucromática. Já a metilação de lisinas ou argininas não altera a carga do aminoácido. Por outro lado, certas proteínas funcionam como "leitoras" de específicas modificações de histonas (Xhemalce B, Dawson M, Bannister, AJ, 2011). Por exemplo, proteínas contendo domínios "Bromo" (bromodomínio) interagem com histonas acetiladas, enquanto proteínas com domínios "Cromo" (cromodomínio) interagem com histonas metiladas. Estas "proteínas leitoras" estão associadas em complexos com outras proteínas, com diversas atividades enzimáticas que influenciam a função da região genômica. Assim sendo, uma modificação de um aminoácido de uma histona num determinado nucleossomo pode levar ao recrutamento de certas proteínas que podem afetar a estrutura e a função da cromatina circundante.

A acetilação de histonas é promovida por enzimas denominadas histona acetil-transferases (HATs), e é uma modificação reversível através da ação de enzimas denominadas histona deacetilases (HDACs). Já a metilação de histonas, catalisada por histona metil-transferases, foi, durante muitos anos, considerada como uma modificação altamente estável. Em 2004, a primeira demetilase de histonas foi identificada, a demetilase especifica de lisina 1 (LSD1). Nos últimos anos, muitas outras demetilases têm sido identificadas (Xhemalce B, Dawson M, Bannister, AJ, 2011).

Uma questão que é ainda amplamente investigada é a função e a interação entre essas numerosas modificações covalentes de histonas. Em 2000, Strahl e Allis propuseram a teoria do "código histônico" , pelo qual a combinação de diferentes marcas epigenéticas em uma região do genoma (tal como um promotor gênico) é o fator determinante na propriedade funcional dessa região (por exemplo, regulando

Tabela 1. Tabela de modificações de histonas. (Adaptado de Allis CD, Jenuwein T, Reinberg D. Eds. 2006. Epigenetics. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.)

		S. cerevisiae	*S. pombe*	*N. crassa*	*C. elegans*	*Drosophila*	**Mammals**	*A. thaliana*
	Caracteristicas genômicas							
	Tamanho do genoma	12 Mb	14 Mb	40 Mb	100 Mb	180 Mb	3,400 Mb	150 Mb
	Número de genes	6,000	5,000	10,000	20,000	14,000	~25,000	25,000
	Tamanho médio dos genes	1.45 kb	1.45 kb	1.7 kb	2 kb	5 kb	35-46 Kb	2 kb
	Média do número de íntrons por gene	≤1	2	2	5	3	6-8	4-5
	% codificadora do genoma	70	60	44	25	13	1-1.5(Hs)	26
	Caracteristicas Epigenéticas							
ON	Acetilação de Histonas	+	+	+	+	+	+	+
ON	Metilação de H3K4	+	+	+	+	+	+	+
ON	Metilação de H3K36	+	+	+	+	+	+	+
ON	Metilação de H3K79	+	+	+	+	+	+	+
ON	Histona variante H3.3 should be ON/OFF	+	+	+	+	+	+	+
OFF	Metilação de H3K9[a]	-	+	+	+	+	+	+
OFF	Proteínas "HP1-like"	-	+	+	+	+	+	+
OFF	RNA de interferência	-	+	+	+	+	+	+
OFF	Metilação de H4K20[b]	-	+	+	+	+	+	+
OFF	Metilação de H3K27	-	-	+	+	+	+	+
OFF	Complexos Polycomb	-	-	-	+	+	+	+
OFF	Metilação de DNA	-	-	+	-	(+)[c]	+	+
OFF	Proteínas de ligacão a DNA metilado	-	+[d]	+	+[e]	+[f]	+	+
OFF	"Imprinting"	-	-	-	-	+[g]	+	+

Abreviação: (Hs) Homo sapiens.
a Há evidência que H3K9 também seja encontrada em regiões de eucromatina, mas a significância funcional é desconhecida.
b H4K20 tri-metilada não ocorre em S. cerevisiae, enquanto todos os três estados de metilação de H4K20 estão presentes em organismos multicelulares.
c Drosophila possui níveis muito baixos de metilação de DNA.
d Dnmt2 mutada.
e Dnmt2 (Pp) e proteínas com domínio MBD (Ce, Cb, Pp).
f Dnmt2 e proteínas com domínio MBD (Dm).
g Ao nível global de cromossomo ou genoma, ao invés de gene especifico.

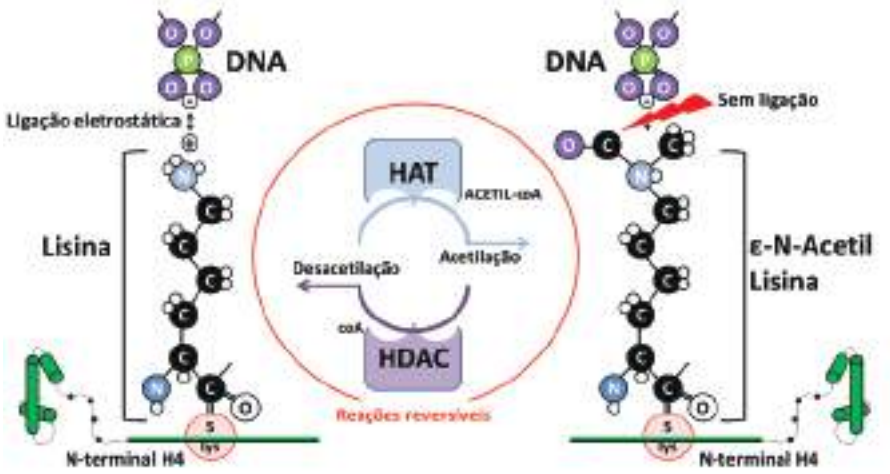


Figura 3. Exemplo de mecanismo de regulação por modificação de histonas: perda de carga positiva da lisina após acetilação. Adaptado de http://www.nobelprize.org/educational/medicine/dna/a/transcription/acetylation.html

a atividade transcricional do promotor) (Strahl e Allis, 2000). Esta hipótese prevê que determinadas marcas "atraem" a adição (ou remoção) de outras marcas, e envolvem a ação coordenada de complexos de proteínas que depositam as marcas enzimaticamente ("escritores") e de proteínas que reconhecem quimicamente as marcas ("leitores"), as quais mediam a propriedade regulatória dessa região da cromatina (tal como o estado de compactação local).

As modificações têm consequências diretas na expressão gênica (Figura 4). Por exemplo, a presença de modificações como acetilações em diversos aminoácidos das diferentes histonas ou metilação da lisina 4 na histona 3 (H3K4me3) na região promotora de um determinado gene é associada com ativação da transcrição, assim como H3K36me3 na região codificadora do gene. Em contraste, H3K27me3, H3K9me3 e H4K20me1 nas regiões promotoras marcam genes reprimidos. Em células-tronco embrionárias, genes associados com o desenvolvimento embrionário estão reprimidos, mas têm a RNA polimerase associada de modo a permitir uma ativação rápida durante diferenciação. Estes genes tem nos seus promotores nucleossomos com marcas ativadoras (H3K4me3) e repressoras (H3K27me3), num estado denominado "bivalência" (Mikkelsen et al., 2007). Durante a diferenciação das células-tronco, esta bivalência é resolvida, com os promotores mantendo exclusivamente H3K4me3 em tipos celulares em que o gene é expresso ou H3K27me3 em tipos celulares em que o gene é reprimido (Mikkelsen et al., 2007).

Outras regiões regulatórias do genoma, tais como os "super-enhancers" e "regiões de controle de lócus" (LCRs), que podem controlar diversos genes, mesmo posicionadas a distâncias consideráveis dos mesmos, também são "marcadas" por modificações específicas, tal como H3K4me1. Frequentemente, também são caracterizadas pela ligação de proteínas regulatórias específicas e associadas a essas marcas, como os fatores p300 em *enhancers* e Heterochomatin protein 1 (HP1) em regiões heterocromáticas.

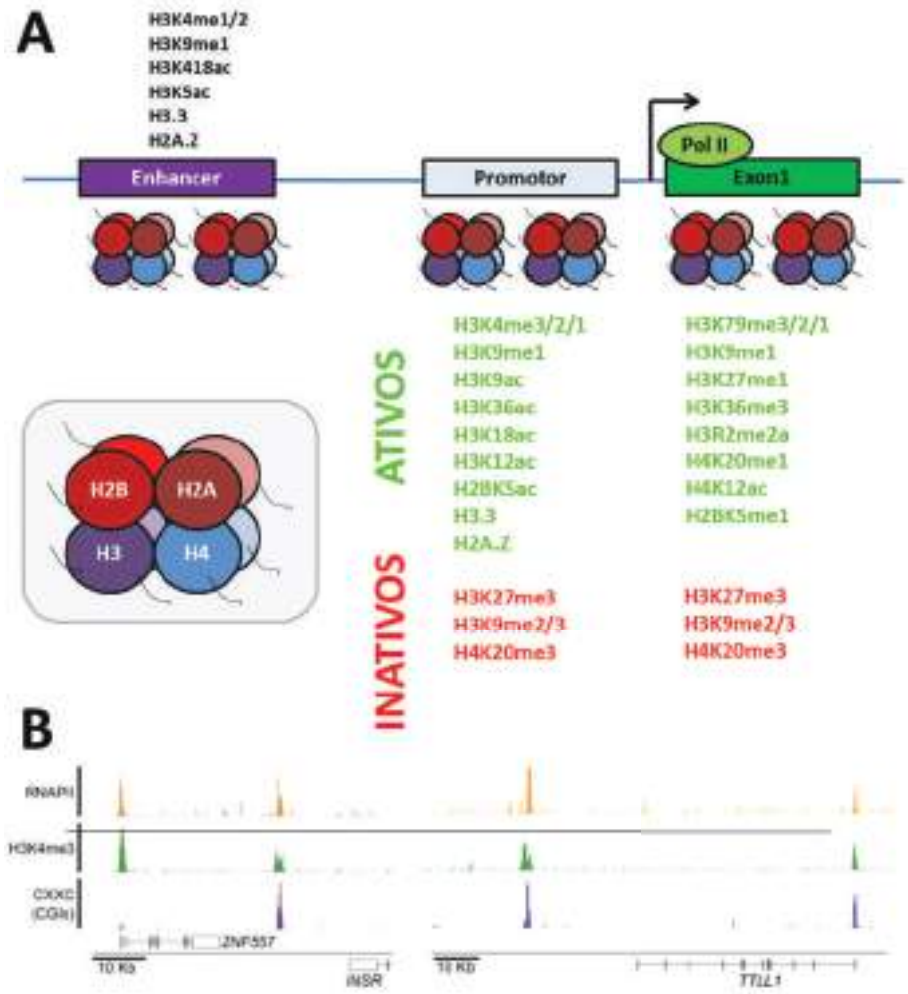


Figura 4. (A) Localização genômica das modificações de histonas. (B) Reflexo da variação na expressão gênica modulada por histonas - visão linear do genoma mostrando o mapeamento de marcas associadas à transcrição e promotores de genes e intergênicas, com sobreposição da RNA Polimerase 2 (RNAPII), H3K4me3 e proteínas CXXC que se ligam a ilhas CpG (CGIs) (adaptado de Deaton AM e Bird A. CpG islands and the regulation of transcription. Genes Dev. 2011 25: 1010-1022.).

Substituição de histonas por variantes

As histonas centrais são depositadas na cromatina durante a replicação do DNA, durante o ciclo celular. Além destas, existem variantes de histonas que são depositadas independentemente de replicação (Banaszynski et al., 2010). Estas variantes (Tabela 2) são proteínas que, apesar de possuírem alta similaridade de sequência com as histonas centrais, possuem diferenças que conferem propriedades distintas às regiões nas quais ocorrem.

Tabela 2. Tabela de tipos de variantes de histonas.

Nome específico da histona			Comentários
Histona H3	Mamíferos	Leveduras	
	H3.1		Deposicão dependente da replicação
	H3.2		Deposicão dependente da replicação
	H3.3	H3.3	Deposicão dependente da replicação
	Cenp-A		Histona centromérica H3
Histona H2A	H2A		Histona canônica H2A
	H2A.Z		Múltiplas funções
	H2A.X	H2A	Múltiplas funções incluindo reparo de DNA
	Macro H2A		Repressão transcricional
	H2A.Bdb		Participa da formação dos nucleossomos

Adaptado de Meyers: Encyclopedia of Molecular Cell Biology and Molecular Medicine: Epigenetic Regulation and Epigenomics, Second Edition. Edited by Robert A. Meyers. 2011 Wiley-VCH Verlag GmbH & Co. KGaA. Published 2011 by Wiley-VCH Verlag GmbH & Co. KGaA.

Todas as histonas centrais, com exceção de H4, têm variantes conhecidas. Exemplos incluem a histona H2A.Z, uma variante da histona H2A, que ocorre em promotores e estão associadas à ativação da expressão gênica, mas também à formação de heterocromatina em fases iniciais do desenvolvimento (Banaszynski et al., 2010). H2A.X é outra variante da histona H2A, que está envolvida na reparação do DNA após dano e também em remodelagem da cromatina (Talbert e Henikoff, 2010). Como exemplo para a histona H3, uma variante chamada de H3.3 ocorre igualmente ao longo de genes que são ativamente transcritos e heterocromatina, e parece ter um papel importante na manutenção da memória epigenética (Ng e Gurdon, 2008). Outra variante de histona H3 com papel importante para a memória epigenética são as histonas CenP-A, que se associam ao centrômeros e tem um papel essencial na sua localização e na segregação de cromossomos durante a mitose e meiose.

Além dessas variantes, cabe mencionar-se as protaminas, proteínas igualmente pequenas e de caráter básico mas não relacionadas com as histonas, que associam-se à maior parte do DNA nos gametas masculinos. Uma pequena porcentagem do DNA paterno continua associado a histonas, o que pode estar relacionado com memória epigenética em gametas. Após fertilização, o genoma paterno volta a associar-se a histonas (Migicovsky e Kovalchuk, 2011).

Remodelamento de nucleossomos dependente de ATP

O remodelamento é a manipulação física do estado de compactação da cromatina pela modificação da posição dos nucleossomos (Ho e Crabtree, 2010). É executada por enzimas que dependem de ATP para a sua atividade. Existem 4 famílias de complexos de remodelamento de cromatina: SWI/SNF, ISWI, CHD e INO80; bem como cerca de 30 genes codificadores de proteínas que fazem parte dos complexos

de remodelamento. Apesar destas subunidades de ATPases terem semelhantes atividades *in vitro*, elas não são redundantes, dado que a mutação individual de muitas das ATPases leva a fenótipos severos durante o desenvolvimento em modelos de camundongo. Em vertebrados, diferentes subunidades formam complexos multiméricos de remodelamento e podem estar envolvidos quer na ativação como repressão da transcrição. Estes complexos contêm diferentes unidades em diferentes tipos de células e podem associar-se com fatores de transcrição, podendo também ter funções adicionais à regulação de transcrição (Ho e Crabtree, 2010).

RNAs não-codificadores de proteínas e cromatina

Tradicionalmente, os estudos de regulação gênica e epigenética têm sido centrados na atuação de proteínas que agem como fatores de transcrição e das enzimas que participam da modificação covalente de histonas e DNA. Mais recentemente, no entanto, a participação de RNAs regulatórios na determinação da estrutura da cromatina tem despertado grande interesse.

Além de DNA, histonas e outras proteínas, a cromatina também tem em sua composição RNA que, além de moléculas resultantes da transcrição momentânea de genes ativos (os "transcritos nascentes"), incluem transcritos associados mais estavelmente à cromatina ou às proteínas que regulam a cromatina (Rodriguez-Campos e Azorin, 2007; Guttman et al., 2011). Contudo, apesar de saber-se há décadas que RNAs fazem parte da composição da eucromatina e heterocromatina (Paul e Duerksen, 1975), apenas recentemente estes têm sido identificados e suas funções estudadas (Mattick et al., 2009).

Uma das razões para esse novo interesse é que estudos de transcriptômica nos últimos anos (Capítulo 8) mostraram que milhares de ncRNAs (os RNAs não-codificadores de proteínas), cujas funções estão apenas começando a ser estudadas, são produzidos em células eucarióticas. Em paralelo, tem-se revelado que moléculas de RNA podem desempenhar diversos papéis nas células, incluindo funções regulatórias e estruturais, independentes das atividades já bem caracterizadas exercidas por RNAs ribossomais (rRNAs), RNAs transportadores (tRNAs) e RNAs mensageiros (mRNAs), na síntese de proteínas (Amaral et al., 2008) .

De relevância para o entendimento dos processos epigenéticos, um conjunto crescente de evidência indica que o controle das modificações estruturais dinâmicas da cromatina pode ser regulado por RNAs (Figura 5). Apesar de os mecanismos não serem ainda bem compreendidos, os casos estudados até agora indicam que RNAs podem atuar de diversas formas nestes processos. Um aspecto importante é que tais mecanismos requerem a interação de RNAs com proteínas modificadoras e remodeladoras da cromatina, as quais frequentemente possuem típicos domínios de ligação a RNA. Por exemplo, o domínio conservado cromodomínio está presente em proteínas das famílias de histona metil-transferases, histona acetilases, HP1 e proteínas do complexo Polycomb (Mattick et al., 2009). Proteínas que interagem com RNA também são componentes de sistemas de metilação de DNA em mamíferos e.g., (Jeffery e Nakielny, 2004).

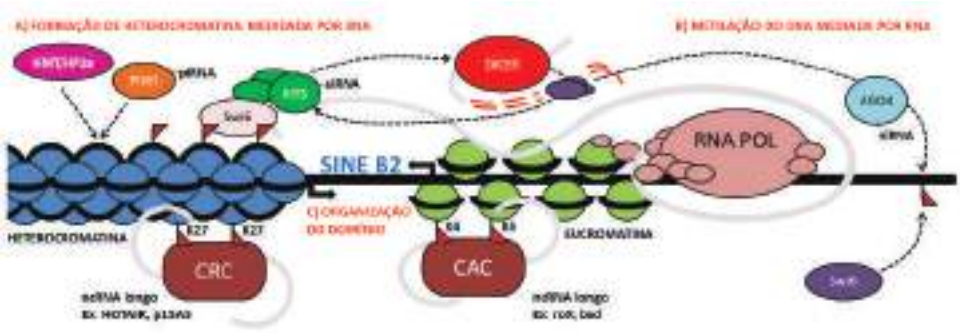


Figura 5. Exemplos da participação de RNAs na regulação da estrutura da cromatina, atuando tanto na ativação quanto na repressão gênica, por exemplo recrutando Complexos Repressores da Cromatina (CRC) ou Complexos Ativadores da Cromatina (CAC). (Adaptado de Mattick JS, Amaral PP et al. 2009. RNA regulation of epigenetic processes. BioEssays 31:51–59)

Os transcritos com propriedades regulatórias incluem RNAs de baixo ("RNAs pequenos") ou elevado ("RNAs longos") número de nucleotídeos. Os RNAs pequenos mais bem estudados em eucariotos pertencem às classes dos microRNAs (miRNAs) e dos "pequenos RNAs de interferência" (siRNAs, do Inglês *small interfering RNAs*). Estes, em geral, possuem entre 19 e 22 nucleotídeos e são produzidos por uma maquinaria celular dedicada que constitui a via de RNAi (ou "RNA de interferência"). Apesar de as funções de pequenos RNAs e da via de RNAi serem melhor conhecidas na regulação pós-transcricional de "genes-alvo" (por exemplo, inibição da tradução e/ou degradação de mRNAs direcionada por miRNAs), espécies de RNAs pequenos também estão envolvidas na regulação transcricional e na regulação da estrutura da cromatina.

RNAs pequenos têm sido associados principalmente à formação de heterocromatina (Buhler e Moazed, 2007). siRNAs produzidos a partir de regiões heterocromáticas são essenciais em processos de silenciamento da cromatina e dinâmica estrutural dos cromossomos em animais, plantas, fungos e protozoários (Mattick et al., 2009; Amaral e Mattick, 2008). Por exemplo, em leveduras *Schizosaccharomyces pombe*, a via de RNAi está envolvida na formação de heterocromatina na região dos centrômeros e é importante nos processos de meiose e mitose. Estes processos envolvem a metilação na lisina 9 de histonas H3 e a atuação do complexo RITS (*RNA-Induced Initiation of Trascriptional gene Silencing*), o qual contém proteínas que interagem com pequenos RNAs complementares a trechos do DNA-alvo na cromatina (Buhler e Moazed, 2007). Apesar de diferenças nos mecanismos, há evidências que processos similares envolvendo proteínas da via de RNAi ocorram não só em animais tão diversos como a mosca-das-frutas e mamíferos (Buhler e Moazed, 2007; Pal-Bhadra et al., 2004; Kanellopoulou et al., 2005), mas também em plantas (Chen, 2009). Cabe notar que é possível que pequenos RNAs também estejam envolvidos na ativação gênica via modulação da cromatina, apesar de a evidência ainda ser preliminar.

Por outro lado, os RNAs longos também estão envolvidos com a regulação da cromatina, tanto para a ativação quanto para a repressão da expressão gênica (Mattick et al., 2009). RNAs longos são, por definição, maiores que 200 nucleotídeos, incluindo

uma grande diversidade de transcritos, tais como RNAs que são processados por *splicing* ou não, que podem ser poli-adenilados ou não, e que podem ser transportados para o citoplasma ou permanecerem no núcleo celular. Dentre os longos, os RNAs melhor estudados em mamíferos são os RNAs *Xist* e *Air*, que possuem cerca de 17 kilobases (kb) e 100kb, respectivamente. Ambos são RNAs nucleares e estão envolvidos na formação de heterocromatina e no silenciamento gênico. O RNA *Xist* (acrônimo do Inglês *X-inactive specific transcript*) é essencial no processo conhecido como "compensação de dose" em mamíferos (Clerc e Avner, 2011). Em fêmeas, que possuem dois cromossomos X por célula (XX), um dos cromossomos X é inativado via heterocromatinização, gerando um cromossomo compactado e fazendo com que a expressão da maioria dos genes presentes no cromossomo X se iguale à expressão desses genes em células de machos (XY), os quais contem apenas um cromossomo sexual X. A forma pela qual *XIST* causa esse silenciamento é complexa e tem sido objeto de intenso estudo (Clerc e Avner, 2011). Ela envolve a transcrição do RNA *Xist* nos estágios iniciais de desenvolvimento embrionário a partir apenas do cromossomo a ser inativado. Este RNA inicia a heterocromatinização do cromossomo a partir do seu sítio de transcrição e a espalha ao longo de todo o cromossomo, envolvendo o recrutamento de proteínas que promovem a formação de heterocromatina via modificações de histonas, tais como proteínas do complexo Polycomb, que mediam metilação de histonas, e metilação de DNA.

Já o RNA *Air* ("*Antisense Igf2r RNA*") atua no processo de *imprinting* genômico (ver acima), pelo qual um grupo de genes no lócus *Igf2r* é transcrito apenas a partir do cromossomo herdado da mãe (e são, portanto, "silenciados" no cromossomo paterno), enquanto o RNA *Air* é transcrito apenas do cromossomo paterno e silenciado no materno (Koerner et al., 2009). *Air* atua de uma maneira análoga ao RNA *XIST*, recrutando proteínas modificadoras da cromatina para silenciar especificamente certos genes presentes no mesmo cromossomo, mas sua atuação se limita a genes em sua vizinhança na região do lócus *Igf2r* . Por sua vez, a expressão monoalélica do RNA *Air* é controlada via metilação da ilha CpG na sua região promotora. Notavelmente, diversos outros *loci* regulados por *imprinting* também são regulados por ncRNAs longos, de forma semelhante à regulação do lócus *Igf2r* (Koerner et al., 2009).

Além disso, importantes para o contexto da regulação genômica, diversos outros RNAs longos também tem sido encontrados regulando outros genes, não apenas genes que se encontram na vizinhança e no mesmo cromossomo do ncRNA (atuação em *cis*), mas também genes presentes à distância e às vezes localizados em outros cromossomos (atuação em *trans*). Por exemplo, em células humanas, o RNA chamado *HOTAIR* ("*Hox antisense intergenic RNA*") é transcrito a partir do lócus *HOXC* no cromossomo 12, mas regula genes do lócus *HOXD* no cromossomo 2, também envolvendo a interação com Polycomb e outras proteínas modificadoras da cromatina (Rinn et al., 2007; Tsai et al., 2010). Estudos recentes indicam que milhares de ncRNAs longos podem atuar de maneira similar via interação com Polycomb e outras proteínas que modificam a cromatina (Guttman e Rinn, 2012). Dado que o RNA possui informação na sua sequência primária, é possível que muitos dos RNAs não-codificadores possam servir igualmente como "guias" dos complexos em que estão associados para regiões especificas do genoma. Esta possibilidade poderá envolver a formação de estruturas tríplex entre o RNA e a dupla fita de DNA (Schmitz et al., 2010; Martianov et al., 2007). Embora a função de RNA não-codificadores tenha sido descrita fundamentalmente

como repressora de transcrição, RNAs como *HOTTIP* (“*HOXA transcript at the distal tip*”), em mamíferos, e roX (“*RNA on the X*”) em *Drosophila*, podem igualmente estar associados a complexos que promovem a ativação gênica via controle da estrutura da cromatina (Wang et al., 2011; Ilik e Akhtar, 2009).

De fato, estes representam apenas alguns exemplos de RNAs que controlam processos epigenéticos e muitos mais têm emergido recentemente. Isto sugere que esta é somente a ponta do iceberg e que estamos perante um admirável mundo novo de regulação epigenética por RNA, cuja compreensão será muito importante para entender perturbações epigenéticas envolvidas em câncer e ouras doenças.

A organização tridimensional da cromatina

Apesar de muito do foco em estudos genômicos estar na determinação de sequências de DNA ou RNA, relativamente pouco se sabe sobre como esta informação está organizada nas células.

O uso da microscopia de fluorescência em tempo real para observação de células eucarióticas revelou uma organização altamente dinâmica do genoma no núcleo interfásico. Essas observações, em conjunto com as observações com técnicas de biologia molecular descritas abaixo, têm promovido uma mudança na percepção da estrutura do genoma unidimensional, para uma visão multi-dimensional (Figura 6) (Van Berkum e Dekker, 2009; Misteli, 2007).

Certos elementos funcionais do genoma, tais como sequências regulatórias, agem a longas distâncias e entram em contato com os promotores de genes específicos via

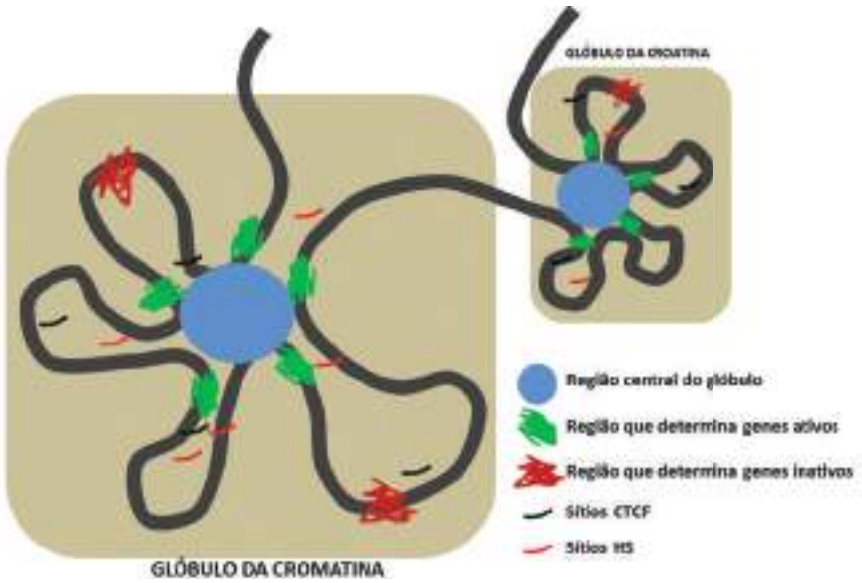


Figura 6. Estrutura tridimensional da cromatina. Adaptado de “The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules”. Davide Baù, Amartya Sanyal, Bryan R Lajoie, Emidio Capriotti, Meg Byron, Jeanne B Lawrence, Job Dekker & Marc A Marti-Renom. Nature Structural & Molecular Biology 18, 107–114 (2011) doi:10.1038/nsmb.1936

formação de *loops*, mediados por proteínas que interagem com DNA, histonas e outras proteínas (Van Berkum e Dekker, 2009) (Figura 6). Estes elementos de DNA incluem os *enhancers*, que modulam a atividade de genes em tipos celulares específicos, e "insuladores", que limitam as regiões do genoma controladas por determinados elementos regulatórios (análogos a pontuações em sentenças). Além disso, não apenas um cromossomo pode ter este tipo de regulação tridimensional, mas também diferentes cromossomos podem interagir entre si e são organizados no núcleo de forma não aleatória. Por exemplo, frequentemente (mas não exclusivamente), regiões transcricionalmente inativas (com alta concentração de heterocromatina) tendem a localizarem-se na periferia do núcleo, associadas ao envelope nuclear, enquanto regiões ativas, ricas em genes expressos, são posicionadas mais no interior do núcleo. Além disso, tem-se identificado "compartimentos" nucleares, configurados por diversos "territórios cromossômicos", que frequentemente correspondem a regiões ricas ou pobres em genes que podem ser co-regulados (Cremer e Cremer, 2010).

Estas observações têm sido expandidas de forma dramática com o uso de tecnologias genômicas, possibilitando o surgimento de um novo entendimento da organização funcional dos cromossomos e do núcleo celular com um todo. Estas serão exploradas brevemente nas sessões seguintes.

Epigenômica

O conjunto de "marcas" epigenéticas em um determinado tipo celular constitui o seu "Epigenoma". A Ciência da Epigenômica refere-se ao estudo da distribuição de marcas epigenéticas ao longo de todo o genoma, bem como o mapeamento global dos compartimentos funcionais da cromatina.

As revoluções na Genômica têm tido um grande impacto em diversas áreas de ciência biomédicas, incluindo de forma bastante significativa a Epigenômica (Box 1). Mais especificamente, os métodos de sequenciamento por tecnologias de segunda geração, também conhecido como sequenciamento de nova geração ("*Next Generation Sequencing*" ou NGS), têm beneficiado diretamente o estudo de estados epigenéticos para amplas regiões do genoma ou para todo o genoma.

De fato, como passo lógico após o Projeto Genoma Humano e sequenciamento do genoma de diversos outros organismos, esforços recentes tem-se concentrados em mapear os elementos funcionais presentes na sequência genômica estabelecida. O projeto ENCODE (*ENCyclopedia Of DNA Elements*), por exemplo, é desenvolvido por um consórcio internacional de laboratórios (lançado pelo *National Human Genome Research Institute*, NHGRI, nos Estados Unidos) e visa a catalogação de todos os elementos funcionais do genoma humano em diversos tecidos e tipos celulares. Tal complexa tarefa baseia-se em anotações computacionais (tais como a identificação de elementos evolutivamente conservados) e bioquímicas, que vão desde análise de sequência transcritas até o mapeamento de sítios de ligação de fatores de transcrição, metilação de DNA e de modificações de histonas. Além disso, diversos projetos paralelos têm objetivos similares aplicados a organismos modelo, como mosca-das-frutas e o verme *C. elegans* (nos consórcios chamados de "ModENCODE"). Já um exemplo de projeto internacional voltado especificamente para o mapeamento de estados epigenéticos é o "Projeto Epigenoma Humano", que também envolve diversos

laboratórios coordenados para compilação sistemática dos padrões de metilação de DNA em um grande número de tecidos humanos.

Além de permitirem um entendimento básico do funcionamento do genoma, tais esforços visam facilitar a compreensão dos estados epigenéticos que caracterizam células em estados normais e patológicos. Isto porque modificações de histona e DNA em diferentes tipos celulares não são apenas fundamentais para determinarem o estado celular durante o desenvolvimento, mas também estão frequentemente alterados em doenças, principalmente nas patologias complexas ou multifatoriais, que não são explicadas por mutações em um ou poucos genes, como a maiorias dos cânceres e doenças comuns como a diabetes. Deste modo, espera-se que o entendimento de como alterações epigenéticas ("epimutações", em analogia às mutações genéticas) que contribuem para o desenvolvimento de doenças específicas poderá proporcionar não apenas novas ferramentas de diagnóstico, mas também alvos para tratamento. Um exemplo recente é a identificação de drogas específicas que afetam a ligação de proteínas bromodomínio a histonas acetiladas e que se mostraram eficazes em testes de controle de leucemia em células em cultura e em modelos animais (Dawson et al., 2011) Tais descobertas indicam que compostos químicos com atuação específicas podem ser empregados no tratamento de doenças com base epigenética.

BOX 1 – Algumas distinções da detecção epigenômica

Ao contrário dos estudos Genômicos, que envolvem apenas a purificação do DNA celular e determinação da sequência, os estudos Epigenômicos envolvem o isolamento das sequências de DNA associadas às "marcas" epigenéticas em estudo, sejam elas regiões de DNA metiladas ou associadas a histonas modificadas. Como pré-requisito dos métodos atuais, é essencial que se tenha a sequência do genoma já determinada para o desenvolvimento de *tiling arrays* ou para o mapeamento de dados de sequenciamento obtidos com os protocolos de Epigenômica (ver abaixo). Com os *tiling arrays*, nos quais sondas cobrem inteiras regiões genômicas (ao contrário dos microarranjos convencionais que contém uma ou poucas sondas por gene), pode-se escolher as regiões a serem analisadas, por exemplo restringindo-se a sequências promotoras ao longo de todo o genoma; ou até mesmo explorar toda a porção não repetitiva do genoma e representada no microarranjo. Já com sequenciamento, não há necessidade de se definir a priori quais regiões serão analisadas, mas o conhecimento do genoma é necessário para o mapeamento das sequências obtidas e para a análise dos dados.

Tecnologias epigenômicas: mapeamento global de domínios funcionais da cromatina

Diversos métodos para o estudo de regiões regulatórias do genoma têm sido desenvolvidos nas últimas duas décadas, dentre eles experimentos com o objetivo de avaliar a estrutura da cromatina e seu estado de modificação epigenética, os quais tem sido amplamente explorados em estudos de Epigenômica. Estes procedimentos frequentemente empregam métodos imunológicos (tal como imunoprecipitação), mas também englobam uma ampla gama de métodos bioquímicos e biofísicos, os quais

devem ser empregados de acordo com o problema Biológico e necessidades técnicas específicas. Algumas destas metodologias serão exploradas a seguir.

Imunoprecipitação da cromatina (ChIP)

Desenvolvida inicialmente em 1988 (Hebbes et al., 1988), a técnica de ChIP é o método mais amplamente utilizado para a identificação de sítios de ligação de proteínas associadas ao DNA e as enzimas que as regulam. A técnica é baseada no uso de anticorpos específicos para o isolamento (pelo procedimento de imunoprecipitação) de proteínas ou complexos protéicos de interesse que se associam com a cromatina estavelmente (tais como Histonas) ou mais transientemente (tais como fatores de transcrição). Devido a grande importância e a generalidade do seu uso, essa técnica é descrita em maior detalhe no Box 2. Após o procedimento de imunoprecipitação (Das et al., 2004), a detecção dos fragmentos de DNA associados a estas proteínas e imunoprecipitados pode ser feita de diversas maneiras, incluindo PCR quantitativo para análise de regiões selecionadas ou, para estudos epigenômicos, utilizando-se técnicas para detecção em larga-escala de regiões genômicas co-imunoprecipitadas, tais como hibridação em microarranjos de DNA (por exemplo, os "*tiling arrays*" que representam grande porções do genoma simultaneamente) ou sequenciamento dos fragmentos de DNA isolados (Ku et al., 2011) (Figura 7).

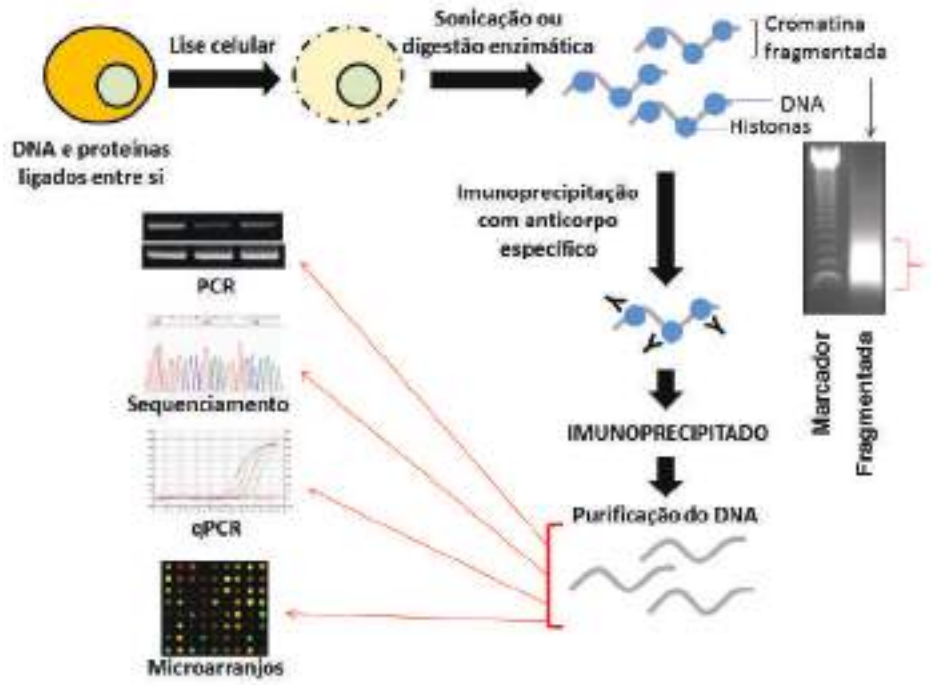


Figura 7. Imunoprecipitação da cromatina (ChIP) e métodos de detecção. Adaptado de Philippe Collas, John Arne Dahl. 2008. Chop it, ChIP it, check it: the current status of chromatin immunoprecipitation. Frontiers in Bioscience 13, 929-943.

Inicialmente, a combinação das técnicas de ChIP e hibridação em microarrays de DNA, conhecida como "ChIP-chip", foi aplicada com sucesso para mapear a posição de modificações epigenéticas nos genomas eucarióticos de humanos a leveduras (Bernstein et al., 2002; Schubeler et al., 2004; Bernstein et al., 2005). Mais recentemente, a combinação de ChIP com técnicas de sequenciamento em alta densidade ("deep sequencing"), denominada de "ChIP-seq", tem tornado o método de escolha e revolucionado o campo da Epigenômica (Ku et al., 2011).

A análise dos dados para o mapeamento das sequências de DNA enriquecidas pela imunoprecipitação em ChIP-Seq e ChIP-chip é definida por critérios estatísticos e é fortemente dependente de recursos computacionais e técnicas de **bioinformática.** Para os experimentos de ChIP-Seq, o enriquecimento das regiões do DNA associados a proteína de interesse é avaliado relativamente aos fragmentos sequenciados oriundos de todo o genoma. Assim, o "sinal" representa apenas uma fração pequena das sequências isoladas ao fim do procedimento. Após o mapeamento das sequências isoladas no genoma, a análise bioinformática permite a determinação de "picos" de sinal significativamente acima do ruído, definindo-se deste modo os sítios de ligação mais robustos das proteínas em estudo.

Desde a implementação da combinação das técnicas de ChIP e "*tiling arrays*" ou sequenciamento tem sido possível avançar rapidamente o entendimento da localização de fatores de transcrição ao longo dos cromossomos, bem como da complexidade da distribuição das modificações na cromatina, suas funções específicas e a inter-relação entre elas (o chamado "código histônico"). Dentre as numerosas descobertas, uma observação importante foi a grande complexidade de ligações de diferentes fatores de transcrição à cromatina, com grupos de fatores atuando em combinação para regulação de genes específicos (Bernstein et al., 2007). Além disso, observou-se a presença de diversos novos elementos regulatórios, com uma grande fração dos sítios de ligação a fatores de transcrição ocorrendo em regiões "intergênicas" ou "intrônicas", longe de regiões promotoras canônicas esperadas (frequentemente dezenas de milhares de bases de distância). O mapeamento dessas regiões tem possibilitado inclusive a identificação de novos genes, tais como milhares de novos genes que codificam "RNAs não-codificadores de proteínas" denominados "lincRNAs" (do inglês *Long Intergenic ncRNAs*), cujas funções tem sido alvos de intenso estudo (Guttman et al., 2009).

BOX 2 – O procedimento de chip

Um protocolo de ChIP convencional (Figura 7) em geral envolve passos iniciais de indução de ligações covalentes entre proteínas associadas a cromatina e DNA, a fim de estabilizar tais associações durante a purificação dos complexos proteína-DNA. Este procedimento de ChIP com ligações cruzadas é denominado "X-ChIP"; em contraste ao procedimento que exclui essa etapa, conhecido com ChIP Nativo ou "N-ChIP", e que pode ser usado para isolar proteínas que se ligam ao DNA mais estavelmente e com afinidade robusta (por exemplo, Histonas). No X-ChIP, a fixação das proteínas associadas à cromatina é normalmente promovida pela adição de *agentes químicos* ou por radiação com luz ultravioleta. O uso de agentes químicos, como formaldeído ou glutaraldeído, é o mais usual.

De forma resumida, para um protocolo de ChIP convencional inclui as seguintes etapas (Figura 7): (A) ligação cruzada, (B) lise celular (ou isolamento dos núcleos seguida de lise nuclear); (C) fragmentação da cromatina, (D) incubação com o anticorpo de interesse, (E) precipitação do complexo anticorpo-cromatina, (F) lavagens do imunoprecipitado com soluções salinas de estringência crescente para remoção de interações inespecíficas e (G) eluição da cromatina. Segue-se então a reversão das ligações cruzadas, digestão das proteínas da cromatina com proteinases, purificação e detecção dos fragmentos de DNA co-imunoprecipitados (Das et al., 2004).

Para a avaliação direcionada para regiões específicas e pré-determinadas do genoma (por exemplo, a região promotora de um gene), estratégias como PCR ou PCR quantitativo (PCR em tempo real) são as opções mais tradicionais para detecção dos fragmentos de DNA associados à proteína de interesse. Em geral, o sucesso do experimento é avaliado pelo "enriquecimento" dos fragmentos de DNA da região (ou regiões) de interesse, os quais são co-imunoprecipitados com o anticorpo específico (i.e. o "sinal"), em relação à ("normalizado" pela) quantidade inicial de DNA presente na amostra (denominado "input"). Este enriquecimento é avaliado de duas formas: primeiro, comparado com a quantidade de DNA da mesma região genômica obtida após imunoprecipitação com um anticorpo não específico, como IgG normal (i.e. o que representa o "ruído" da técnica); e, segundo, comparado com o enriquecimento de fragmentos das regiões que não devem associar com o antígeno imunoprecipitado, que são usadas como controles negativos de amplificação.

Estratégias de Captura da Conformação da Cromatina: 3C e suas variantes (4C, 5C e "HiC")

Apesar de a técnica ChIP ser útil para mapear-se interações entre proteínas e trechos do DNA, ela não proporciona informação acerca da estrutura tridimensional da cromatina ou como as diferentes regiões do genoma interagem entre si. Como mencionado anteriormente, é cada vez mais evidente que essas informações são fundamentais para entender a regulação gênica e as propriedades fisiológicas da cromatina.

Uma das técnicas originais desenvolvidas para o estudo de interações a longas distâncias é denominada "3C", ou "Captura da Conformação da Cromatina" (do Inglês, "*Chromatin Conformation Capture*") (De Wit e Laat, 2012). Os métodos mais comuns de análise da topologia do genoma, cujo objetivo equivale a uma "Genômica 3D", são baseados em Captura de Conformação da Cromatina (Figura 8) e serão apresentados nesta seção.

3C (Captura de Conformação da Cromatina)

A técnica de 3C foi desenvolvida em 2002 (Dekker et al., 2002) para analisar contatos entre distantes regiões genômicas pré-determinadas, usando o tipo celular ou tecido de escolha. Estas interações tridimensionais estudadas incluem conformações com papéis regulatórios, por exemplo a formação dos *loops* que trazem juntos elementos

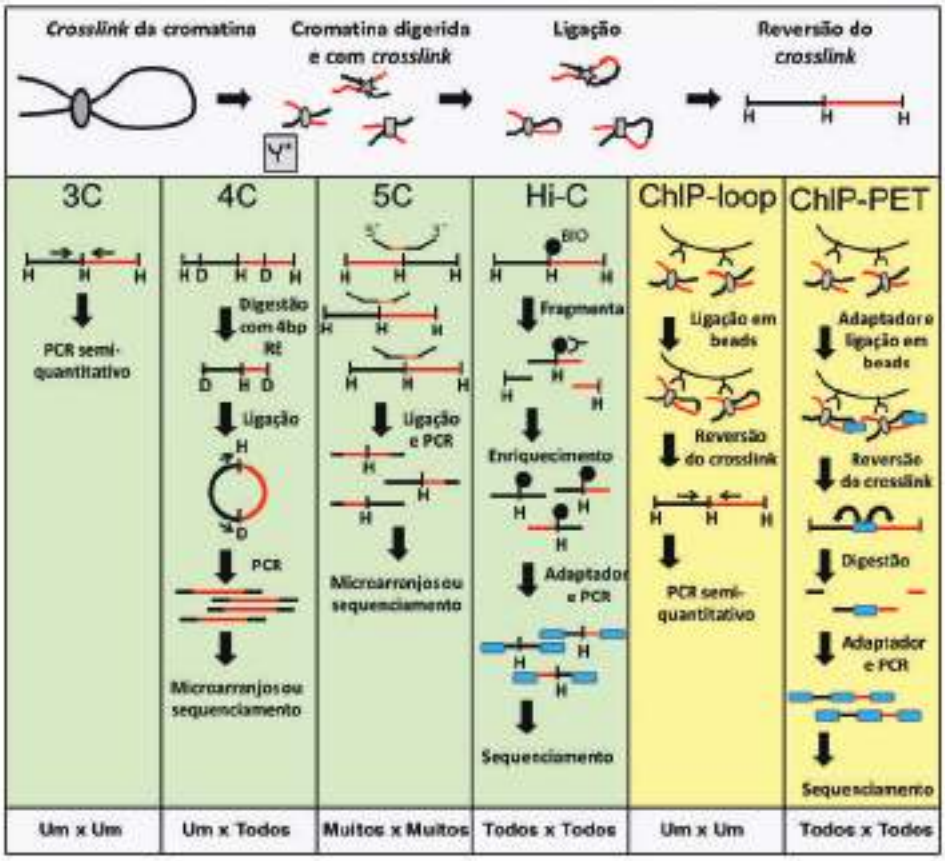


Figura 8. Metodologias comuns para os estudos da organização tridimensional da cromatina: 3C e técnicas derivadas. Adaptado de Wit E, de Laat W. A decade of 3C technologies: insights into nuclear organization. Genes Dev. 2012;26(1):11-24 (RE: Restriction Enzyme).

regulatórios em *enhancers* e promotores (Figura 6 e Figura 8) (Palstra et al., 2003; Fraser e Bickmore, 2007).

O 3C é baseado no princípio de "ligação por proximidade", no qual regiões distintas do genoma que se encontram próximas devido à conformação 3D do cromossomo são imobilizadas e ligadas umas às outras, seguida pela quebra da cromatina em fragmentos pequenos e detecção por PCR das interações entre as regiões estudadas (Dekker et al., 2002) (Figura 8). O método tradicional envolve os seguintes passos: (1) ligações cruzadas da cromatina (em geral com formaldeído), (2) fragmentação ou "digestão" da cromatina com o uso de enzimas de restrição com sítios de clivagem comuns (por exemplo, *Hind*III), (3) ligação dos fragmentos de DNA imobilizados em proximidade e contendo extremidades coesivas geradas pelas enzimas de restrição, e, por último, (4) detecção das associações por PCR quantitativo ou semi-quantitativo usando oligonucleotídeos específicos para as regiões de interesse (Figura 8). Devido

ao fato de o segundo passo (de ligação dos fragmentos) ser realizado em condição de alta diluição (reduzindo assim a chance de ligações aleatórias de fragmentos de DNA em solução), os produtos de amplificação no PCR refletem a proximidade original das sequências analisadas no genoma, indicando co-localização espacial mesmo que estas sequências encontrem-se distantes na sequência linear do genoma.

Uma limitação do 3C é que as interações investigadas devem ser já conhecidas ou pelo menos preditas, possibilitando a definição das regiões a serem analisadas por PCR. Contudo, baseado nesta técnica, várias adaptações têm sido desenvolvidas, com variações que proporcionam melhor poder quantitativo e que permitem uma expansão para a análise da associação de várias regiões genômicas simultaneamente (De Wit e Laat, 2012; Fullwood e Ruan, 2009). Mais recentemente, estas análises tem sido estendidas para todo o genoma, gerando dados de enorme complexidade que revelam a organização dos cromossomos no núcleo celular (De Wit e Laat, 2012). A seguir, serão apresentadas algumas das principais adaptações do 3C (Figura 8).

4C ("Captura de Conformação da Cromatina Circular" ou "Chromosome Conformation Capture on Chip")

O método de 4C refere-se a um grupo de adaptações do 3C nas quais as interações entre uma região escolhida do genoma são analisadas simultaneamente com múltiplas outras regiões (Simonis et al., 2007) (Figura 8).

Tal sequência escolhida é denominada "ponto de vista" e a técnica é conhecida como "um contra todos". De fato a análise pode se estender para a interação da sequência escolhida com todo o genoma. Para atingir esse objetivo, ao invés do uso de PCR para a análise das interações, são utilizados microarranjos de DNA (daí o nome "*Chromosome Conformation Capture on Chip*") representando diversas regiões do genoma (Splinter e De Laat, 2011) ou até mesmo sequenciamento dos fragmentos contendo as interações ("4C-seq").

Além do método de detecção das interações, as modificações básicas para a execução do 4C são a inclusão de uma segunda etapa de digestão com enzimas de restrição e a ligação dos fragmentos de cromatina de modo que formem pequenos círculos de DNA (daí o termo "*Circular Chromosome Conformation Capture*"). As interações entre a sequência escolhida e as outras regiões do genoma estarão representadas dentre os círculos formados. Antes da detecção das interações, é incluído um passo de amplificação das regiões interagindo com a sequência escolhida, com o uso da técnica de "PCR invertido" e oligonucleotídeos específicos para a sequência escolhida ("ponto de vista"). Os fragmentos analisados podem então ser detectados por microarranjo ou sequenciamento de DNA.

A técnica de 4C é o método de escolha para o estudo dos contatos de uma região específica com sítios distais no genoma. Ela tem sido utilizada, por exemplo, para determinar-se as interações entre elementos regulatórios de interesse com genes-alvo e outras regiões genômicas em diferentes tecidos, bem como a dinâmica e regulação dessas interações. Por exemplo, genes regulados por Polycomb são encontrados preferencialmente em grupos (*clusters*) (Bantignies et al., 2011).

No estado atual de desenvolvimento da técnica, o 4C apresenta certa limitação da resolução, com pouca sensibilidade para a detecção de interações de curtas distâncias, já que geralmente emprega enzimas de restrição que digerem o DNA com uma frequência média de aproximadamente um corte a cada mil bases de DNA, gerando fragmentos longos, maiores do que elementos regulatórios como *enhancers* e promotores. Avanços recentes têm aumentado essa resolução, por exemplo, com o emprego de enzimas de restrição que cortam o DNA com maior frequência .

5C (3C cópia de carbono ou "3C-carbon copy")

Outra adaptação do 3C é o método de 5C, na qual são simultaneamente analisadas as interações de múltiplas regiões de interesse no genoma com diversas outras regiões (Dostie et al., 2006) (Figura 8). Esta estratégia é conhecida como "Muitos contra Muitos".

Para a execução do 5C, os produtos de ligação do 3C (que representam as diferentes regiões genômicas interagindo) são hibridizados com uma mistura de diferentes pares de oligonucleotídeos. Os diferentes oligonucleotídeos contêm em sua porção final 3' sequências que são complementares às diferentes regiões genômicas de interesse e que contém o sítio de restrição para enzimas utilizadas na digestão. Oligonucleotídeos que se hibridizam às regiões que estiverem em proximidade no genoma serão justapostos e, após uma reação de ligação, são unidos covalentemente em pares. Além disso, no 5C, cada par de oligonucleotídeos contém na região 5' uma sequência "universal" que é compartilhada por todos os oligos "Forward" ou outra sequência universal compartilhada por todos os oligos "Reverse". Desta forma, um PCR com oligos que correspondem a essas duas sequências universais ("Forward" + "Reverse") é capaz de amplificar simultaneamente todas as regiões que interagem ("PCR multiplex"). Assim como no 4C, a análise e identificação destas interações podem ser feitas com microarranjos de DNA (Capítulo 8) ou sequenciamento (Capítulo 2).

Os resultados são visualizados como uma matriz de frequência de interações entre os pares. Deste modo, esta técnica possibilita uma reconstrução das conformações 3D de inteiras regiões genômicas de interesse (Figura 8). Cabe notar, no entanto, uma das limitações desta técnica é que nem todas as sequências que contém os sítios de restrição na região de interesse são adequados para o desenho de oligonucleotídeos. Logo, a matriz de interações obtidas não é exaustiva.

HiC

Dentre os métodos baseados em Captura de Conformação da Cromatina, um dos mais recentes e abrangentes é o denominado Hi-C (Figura 8), que tem como objetivo o mapeamento de todas as interações genômicas em um grupo de células, utilizando sequenciamento em larga-escala (Lieberman-Aiden et al., 2009). Este tipo de método pode ser denominado "todos contra todos".

A primeira diferença em relação ao protocolo de 3C ocorre após a digestão enzimática da cromatina. As clivagens com enzimas de restrição usadas rotineiramente em Biologia Molecular geram as chamadas extremidades "coesivas", as quais terminam

com DNA de fita simples. Após este passo, e antes da etapa de ligação, nucleotídeos "marcados" com a molécula de Biotina

> *A Biotina é uma molécula amplamente utilizada para "marcar" – via ligação covalente – moléculas em protocolos de purificação, devido à sua alta afinidade à molécula Estreptoavidina, a qual é usada para "pescar" as moléculas contendo Biotina em colunas de purificação ou esferas magnéticas, incluindo proteínas, ácidos nucléicos e lipídeos*.

são incorporados aos fragmentos de DNA, "preenchendo" as extremidades coesivas dos fragmentos digeridos para recompor a dupla fita em suas extremidades (agora contendo Biotina). A ligação entre os fragmentos de DNA dupla-fita é então executada em condição de alta diluição e, assim como no 3C, deverá incluir as interações entre diferentes regiões do genoma que estavam próximas no momento da ligação cruzada com formaldeído.

Em seguida, um passo adicional de fragmentação física ("*shearing*") do DNA ligado é realizado, gerando fragmentos menores. Os fragmentos que incluem as regiões de DNA que contém o sítio de restrição para a enzima utilizada estarão 'marcados' com biotina; os fragmentos que correspondem a regiões de DNA mais distantes dos sítios de restrição não conterão biotina. Em um passo que também caracteriza o processo de HiC, os fragmentos de DNA contendo junções entre regiões do genoma são então enriquecidos via purificação das moléculas contendo biotina.

Nas últimas etapas, oligonucleotídeos adaptadores são adicionados nas extremidades dos fragmentos selecionados e estes são amplificados por PCR e submetidos a sequenciamento em larga escala. As sequências obtidas são então mapeadas no genoma e as interações são determinadas quando sequências de DNA de regiões diferentes no genoma são encontrados em um mesmo fragmento sequenciado.

Dada a complexidade das interações estudadas, a análise de HiC requer uma grande quantidade de dados de sequenciamento, mas proporciona um catálogo das interações ao nível de todo o genoma. Uma das observações importantes possibilitadas pela técnica de HiC foi a demonstração da organização tridimensional da cromatina em domínios de eucromatina e heterocromatina (Lieberman-Aiden et al., 2009) , bem como domínios tridimencionais definidos de associação topológica (chamados TADs ou "Topologically Associating Domains").

ChIP-loop e ChIA-PET

De modo a obter-se uma identificação mais específica de interações entre duas regiões da cromatina, é possível adicionar uma etapa de ChIP com anticorpos específicos antes da realização do 3C. Tal estratégia é denominada "ChIP-loop" ou "ChIP-3C", para qual também há variantes (Figura 8). O método de ChIA-PET tem como objetivo a análise da conformação da cromatina especificamente em regiões ligadas às proteínas de interesse, e envolve a fusão das técnicas de ChIP e 3C (Fullwood et al., 2009).

Técnicas comuns para o estudo de metilação de DNA

Dada a importância da metilação de DNA na regulação gênica e o envolvimento em diversos cânceres, há um grande interesse no mapeamento dessa modificação nas regiões promotoras de genes e outras regiões regulatórias do genoma, o chamado "Metiloma". De fato, uma ampla gama de métodos foi desenvolvida pra o estudo de metilação de DNA e vários deles foram adaptadas para estudos de Epigenômica (Harris et al., 2010), incluindo o "Projeto Epigenoma Humano" e outros projetos similares (Jones et al., 2008).

Conversão com bissulfito

Diferentemente dos métodos utilizados para se mapear modificações de histonas, o estado de metilação de DNA pode ser analisado por um método químico, conhecido como "conversão com bissulfito", ou sequenciamento de DNA por bissulfito (Figura 9). Empregado originalmente por Frommer e Clark (Clark et al., 1995), este método tem sido utilizado para a maioria dos estudos sobre a metilação do DNA.

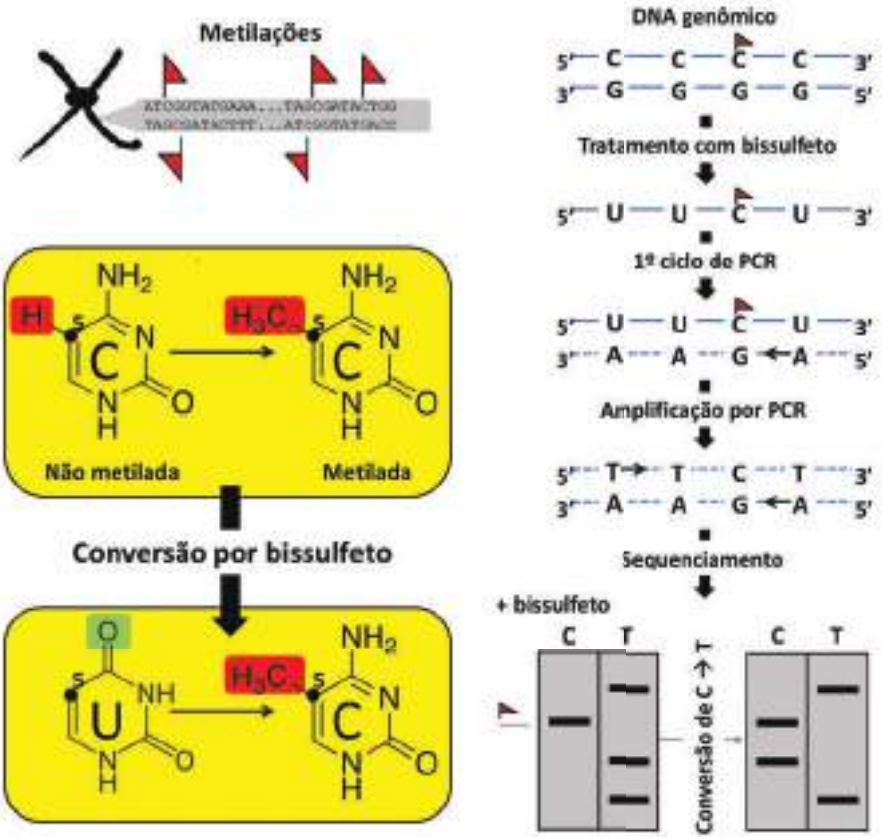


Figura 9. Análise de metilação de DNA pela técnica de bissulfito.

Tal técnica permite a identificação (e quantificação) da metilação do DNA com a resolução de um nucleotídeo. O tratamento com bissulfito causa uma modificação na base de citosina que permite distingui-la da forma metilada (5-mC). Mais especificamente, as bases de citosina no DNA são deaminadas e convertidas em uracila, enquanto as 5-metil-citosinas são resistentes a tal modificação. Uma etapa de desnaturação do DNA antes da adição do bissulfito é necessária, já que a conversão só ocorre em citosinas presentes em DNA de fita simples. Durante o PCR de preparação para o sequenciamento do DNA após a conversão, a uracila resultante é amplificada como timina, enquanto a 5-mC permanece como citosina, permitindo que CpGs metilados sejam distinguidos após o sequenciamento.

Nos métodos tradicionais, as regiões específicas do genoma cujo estado de metilação analisado pela técnica de bissulfito, tais como promotores gênicos, são amplificadas por PCR utilizando *primers* específicos e em seguida clonadas e sequenciadas individualmente. Contudo, com o advento das tecnologias de sequenciamento de nova geração, tornou-se possível o sequenciamento em larga escala do genoma após tratamento com bissulfito, a fim de determinar-se o estado de metilação das citosinas ao longo de todo o DNA genômico (Lister et al., 2008). Tal metodologia tem recebido diferentes nomes, conhecida como Metil-C-seq, BS-seq ou WGBS (do Inglês "whole-genome bisulfite sequencing") e, assim como na análise convencional de regiões específicas, as regiões metiladas são definidas pela comparação com a sequência genômica não tratada com bissulfito.

Apesar de ser ainda um método caro, principalmente em estudos comparativos incluindo muitas amostras, os custos tem sido bastante reduzidos com o desenvolvimento e popularização dos métodos de sequenciamento.

Digestão diferencial de DNA metilado

Outra estratégia utilizada para o estudo de metilação é o uso de enzimas de restrição sensíveis ao estado de metilação do DNA. Certas enzimas, tais como *Hpa*II (cujo sítio de restrição é CCGG) são capazes de clivar apenas sequências não metiladas, e portanto permitem identificar o estado de metilação das regiões de CpG do DNA contendo o sítio de restrição (Horsthemke et al., 1992). Esta técnica também pode empregar outras enzimas sensíveis ao estado de metilação de di-nucleotídeos CG, tais como *Aci*I e *Hin*6I.

De forma similar, tal análise pode ser estendida para todo o genoma, em uma estratégia conhecida como MRE-seq (do Inglês, *Methylation-sensitive Restriction enzyme Sequencing*) (Harris et al., 2010). Contudo, esta também apresenta etapas experimentais distintas e limitações inerentes à técnica. Por exemplo, a digestão com enzimas de restrição gera extremidades coesivas nos fragmentos de DNA que devem ser levados em consideração para a ligação dos adaptadores para sequenciamento. Naturalmente, uma limitação é que a análise é restrita a regiões contendo o sítio de clivagem. Uma das formas de estender a região analisada é o uso de múltiplas enzimas de restrição.

Técnicas de enriquecimento de DNA metilado

Utilizando o mesmo princípio da técnica de ChIP explorada acima, a metilação de DNA também pode ser estudada por imunoprecipitação e cromatografia de afinidade (Harris et al., 2010). Tal enriquecimento pode ser realizado de duas formas: (A) empregando-se anticorpos que reconhecem diretamente o DNA metilado, ou (B) pelo isolamento por associação com proteínas recombinantes que se ligam ao DNA metilado com alta afinidade e especificidade.

Desta forma, pode-se enriquecer para DNA metilado a partir de DNA celular purificado, o qual pode ser detectado da mesma maneira utilizada para detecção de DNA isolado por ChIP (direcionado a uma região específica do genoma, ou em larga escala com o uso de *tiling arrays* ou sequenciamento). O uso destas técnicas seguida de sequenciamento de nova geração tem sido denominadas de "MeDIP-seq" ("Methylated DNA Immunoprecipitation sequencing") e "MBD-seq ("Methylated DNA Binding sequencing") (Serre et al., 2010; Jacinto et al., 2008).

Na primeira estratégia, os anticorpos gerados contra DNA metilado são específicos para o reconhecimento de DNA em fita simples. Logo, assim como na análise por conversão por bissulfito, para o procedimento de imunoprecipitação deve-se partir de DNA desnaturado. Já para o isolamento de DNA metilado pela associação com proteínas ligantes, utiliza-se as proteínas MECBP2 ou MBD2 produzidas de forma recombinante e imobilizadas em *beads*, utilizando-se o princípio de cromatografia de afinidade (também é possível a realização de ChIPs utilizando anticorpos contra tais MBDs, possibilitando a identificação de regiões de DNA metiladas e ligadas endogenamente as essas proteínas). Estas são utilizadas para o enriquecimento de fragmentos de DNA contendo CpGs metilados, os quais podem ser eluídos com diferentes concentrações de sal, o que permite a diferenciação dos fragmentos contendo um baixo ou alto conteúdo de CpGs metilados, eluídos em baixas ou altas concentrações de sal, respectivamente. Portanto, uma vantagem desta técnica é a aquisição inerente de informação da densidade de metilação das regiões detectadas. Vale ressaltar que, ao contrário dos métodos anteriores, que permitem a detecção de metilação do DNA pela atividade preferencial sobre regiões não metiladas, tais técnicas permitem o enriquecimento direto de sequências contendo citosinas metiladas. Devido a essas distinções e às limitações das diferentes técnicas, as quais refletem em diferentes graus de sensibilidade e acurácia dos métodos, estudos mais amplos têm utilizado uma combinação de diferentes métodos para determinar-se os estados de metilação do genoma, por exemplo com o emprego de MeDIP-seq em paralelo a MRE-seq.

De forma geral, estudos comparativos de metilação de DNA em diferentes tecidos e estados fisiológicos têm sido informativos, por exemplo, para a definição de novas regiões diferencialmente metiladas (DMRs, do Inglês "differentially methylated regions") bem como na descoberta de novas regiões que são alvos de metilação, inclusive fora de regiões promotoras, tanto em regiões intergênicas como no corpo dos genes (Deaton e Bird, 2011).

Técnicas para determinação do estado de compactação da cromatina

Estas técnicas determinam quais as regiões da cromatina estão "abertas" e ligadas a fatores de transcrição (eucromatina), e quais estão compactadas, inacessíveis a fatores de transcrição (heterocromatina).

Há várias décadas que a digestão da cromatina com a enzima *DNAse I* tem sido usada para se identificar quais regiões estão "expostas" a fatores que se ligam ao DNA, tais como fatores de transcrição, as quais estão associadas com regiões ativas em transcrição (Wu et al., 1979). Essas regiões ligadas a fatores de transcrição não estão enoveladas em nucleossomos e por isso são acessíveis à DNase. A técnica é tradicionalmente conhecida como "footprinting" ou ensaio de "Hipersensibilidade a DNase I" (*DNAse I HS,* ou *DHS*) e historicamente tem sido instrumental para a identificação de elementos regulatórios, tais como promotores, *enhancers* e insuladores (Bell et al., 2011).

Inicialmente, a forma de análise da hipersensibilidade a DNase I envolvia o uso da técnica de *Southern blot* com sondas radioativas e o perfil de separação eletroforética (Bell et al., 2011). Mais recentemente, sequenciamento de nova geração tem sido útil para a determinação de regiões hipersensíveis a DNase I e para a determinação de sítios de ligação a proteínas ao longo de todo o genoma. Para isto, após o tratamento da cromatina com DNase, o DNA purificado é submetido a métodos de detecção e as regiões não digeridas são identificadas. Um dos princípios explora o isolamento de fragmentos gerados pela DNaseI e detecção por microarranjos ou sequenciamento das extremidades destes fragmentos, as quais correspondem aos sítios de clivagem da DNase in vivo (Sabo et al., 2006). Técnica similares têm recebido diferentes denominações, incluindo "*digital genomic footprinting*", "Dnase-seq", "DHS-seq" ou "*DHS mapping*" (Boyle et al., 2008; Hesselberth et al., 2009).

Outras enzimas utilizadas para o estudo do estado de compactação da cromatina são a micrococcal nuclease (Mnase)(Noll et al., 1975), que digere o DNA de ligação entre os nucleossomos, dando informação da posição destes no genoma, e metilases de bactéria como M.CviPI, que deixam uma marca de metilação nas regiões expostas com GpCs, que depois é detectada por bissulfito, como referido anteriormente. Este método permite a detecção simultânea da metilação do DNA e da ocupação de nucleossomos (Taberlay et al., 2011). Um método que não envolve enzimas é *Formaldehyde-Assisted Isolation of Regulatory Elements* (FAIRE) (Giresi et al., 2007), no qual os nucleossomos são alvo de ligação cruzada com o DNA. Após fragmentação e extração por fenol-clorofórmio, é possível extrair na parte solúvel o DNA que não interage com os nucleossomos.

Todos este métodos podem ser combinados com sequenciamento NGS ou microarranjo, possibilitando a obtenção de um panorama geral da acessibilidade do genoma, o monitoramento do estado regional de compactação da cromatina e da ligação de fatores regulatórios, e identificando, dentre outras coisas, diversas novas regiões regulatórias.

EWAS

Como apenas uma fração dos traços hereditários e predisposição a doenças podem ser explicados por variações e mutações em um ou poucos genes, a identificação dos elementos funcionais dos genomas têm proporcionado uma nova perspectiva para o seu entendimento. Em particular, nos últimos anos os estudos de variantes de alelos ao longo de todo o Genoma, conhecidos como GWAS ("*Genome-wide association studies*"), os quais buscam a associação de polimorfismos de nucleotídeo em sequências genômicas com traços genéticos complexos analisando um grande número de indivíduos "afetados" e controles (tais como portadores e não-portadores de câncer). Tais esforços têm sido importantes para o entendimento da diversidade genética e da sua relação com traços complexos; contudo tais variações conseguem explicar apenas uma pequena proporção dos traços hereditários, indicando que muitos fatores adicionais podem estar envolvidos, incluindo mecanismos epigenéticos, não representados na sequência de DNA.

Neste sentido, estudos de associação em escala epigenômica, conhecidos como "EWAS" (do Inglês "Epigenome-wide association studies") têm ganhado crescente importância para identificação de variações epigenéticas na população e associação destas com traços complexos. Doenças comuns atualmente estudadas por EWAS incluem diabetes, hipertensão, arteriosclerose, asma, Alzheimer, autismo, esquizofrenia, desordem bipolar, dentre outras (Rakyan et al., 2011).

Conclusão

A definição do epigenoma apresenta a complicação de se tratar de um estado transiente e dinâmico, que varia entre diferentes tipos celulares e mesmo em uma célula individual em diferentes estados fisiológicos. Consequentemente, uma das dificuldades destas técnicas, compartilhadas com muitas abordagens moleculares, é que elas medem as interações utilizando populações de células, e não células individuais. Isto dificulta a análise e interpretação dos resultados, já que as células individuais podem ter (e certamente têm) estados diferentes. A "média" obtida ao agrupar os dados provavelmente corresponde às interações mais comuns e estáveis, e métodos recentes têm sido desenvolvidos para ampliar a gama de interações detectadas em diferentes estados celulares. Além disso, tais técnicas caminham para formas de detecção mais sensíveis, principalmente devido ao maior poder de sequenciamento das tecnologias emergentes. Com estas, é possível examinar o estado epigenético de um número cada vez menor de células, possibilitando inclusive o estudo de células individuais, por exemplo, com o uso de tecnologias de "single-molecule, real-time (SMRT) DNA sequencing".

Apesar de as técnicas apresentadas nas seções anteriores serem úteis individualmente, estudos epigenômicos mais abrangentes frequentemente beneficiam-se da combinação dos diversos métodos, incluindo técnicas de espectrometria de massa e imunológicas, associadas a outras abordagens bioquímicas, genéticas e análises funcionais. À medida que estas análises são estendidas para um número cada vez maior de tipos celulares, tecidos e estados fisiológicos, a comparação dos mapas epigenéticos obtidos tende a revelar o papel dos elementos regulatórios

e marcas epigenéticas identificados. Epigenética e Epigenômica são campos muito dinâmicos nos quais processos moleculares são frequentemente descobertos, e novas técnicas implementadas. Com um foco cada vez maior neste tipo de estudo espera-se que muitas novas metodologias inovadoras surgirão.

Ao contrário das décadas de 1990 e 2000, que foram marcadas pelos projetos Genoma, esta nova década já tem sido considerada como a década do Epigenoma (Martens et al., 2011). Ao propor a hipótese do "código histônico" em 2000, Strahl e Allis predisseram: "o entendimento das regras e consequências do código histônico provavelmente terá um impacto em muitos, senão em todos, os processos dependentes de DNA, com amplíssimas implicações para biologia humana e doença" (Strahl and Allis, 2000). Passado pouco mais de uma década, essas ideias são cada vez mais palpáveis, devido à melhor compreensão dos mecanismos de regulação da cromatina, bem como à rápida evolução das tecnologias Epigenômicas, com a extensão das aplicações e a integração com o entendimento dos outros aspectos do controle epigenético, como a remodelagem de cromatina, metilação de DNA, interações tridimensionais e participação de RNAs regulatórios (Figura 10). Pela primeira vez na história, estamos conseguindo casar o entendimento de como a Genética e Epigenética interagem para determinação dos fenótipos celulares e de organismos completos, na saúde e na doença.

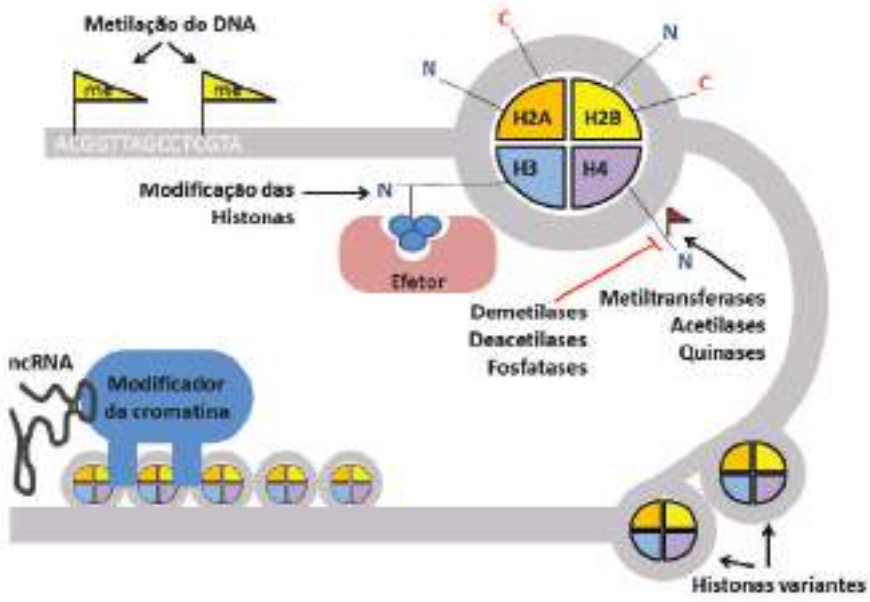


Figura 10. Integração de mecanismos regulatórios da cromatina em estudos epigenômicos.

Bibliografia

AMARAL PP, DINGER ME, MERCER TR, MATTICK JS: The eukaryotic genome as an RNA machine. *Science* 2008, 319(5871):1787-1789.

AMARAL PP, MATTICK JS: Noncoding RNA in development. *Mamm Genome* 2008, 19(7-8):454-492.

BANASZYNSKI LA, ALLIS CD, LEWIS PW: Histone variants in metazoan development. *Dev Cell* 2010, 19(5):662-674.

BANTIGNIES F, ROURE V, COMET I, LEBLANC B, Schuettengruber B, Bonnet J, Tixier V, Mas A, Cavalli G: Polycomb-dependent regulatory contacts between distant Hox loci in Drosophila. *Cell* 2011, 144(2):214-226.

BEISEL C, PARO R: Silencing chromatin: comparing modes and mechanisms. *Nat Rev Genet* 2011, 12(2):123-135.

BELL O, TIWARI VK, THOMA NH, SCHUBELER D: Determinants and dynamics of genome accessibility. *Nat Rev Genet* 2011, 12(8):554-564.

BERGER SL, KOUZARIDES T, SHIEKHATTAR R, SHILATIFARD A: An operational definition of epigenetics. *Genes Dev* 2009, 23(7):781-783.

BERNSTEIN BE, MEISSNER A, LANDER ES: The mammalian epigenome. *Cell* 2007, 128(4):669-681.

BOYLE AP, DAVIS S, SHULHA HP, MELTZER P, MARGULIES EH, WENG Z, FUREY TS, CRAWFORD GE: High-resolution mapping and characterization of open chromatin across the genome. *Cell* 2008, 132(2):311-322.

BUHLER M, MOAZED D: Transcription and RNAi in heterochromatic gene silencing. *Nat Struct Mol Biol* 2007, 14(11):1041-1048.

CARONE BR, FAUQUIER L, HABIB N, SHEA JM, HART CE, Li R, BOCK C, LI C, GU H, ZAMORE PD *et al*. Paternally induced transgenerational environmental reprogramming of metabolic gene expression in mammals. *Cell* 2010, 143(7):1084-1096.

CAUDRON-HERGER M, MULLER-OTT K, MALLM JP, MARTH C, SCHMIDT U, FEJES-TOTH K, RIPPE K: Coding RNAs with a non-coding function: maintenance of open chromatin structure. *Nucleus* 2011, 2(5):410-424.

CHANDLER VL, STAM M: Chromatin conversations: mechanisms and implications of paramutation. *Nat Rev Genet* 2004, 5(7):532-544.

CHEN X: Small RNAs and their roles in plant development. *Annu Rev Cell Dev Biol* 2009, 25:21-44.

Clark SJ, Harrison J, Frommer M: CpNpG methylation in mammalian cells. *Nat Genet* 1995, 10(1):20-27.

CLERC P, AVNER P: New lessons from random X-chromosome inactivation in the mouse. *J Mol Biol* 2011, 409(1):62-69.

CREMER T, CREMER M: Chromosome territories. *Cold Spring Harb Perspect Biol* 2010, 2(3):a003889.

DAS PM, RAMACHANDRAN K, VANWERT J, SINGAL R: Chromatin immunoprecipitation assay. *Biotechniques* 2004, 37(6):961-969.

DAWSON MA, PRINJHA RK, DITTMANN A, Giotopoulos G, Bantscheff M, Chan WI, Robson SC, Chung CW, HOPF C, SAVITSKI MM *et al*. Inhibition of BET recruitment to chromatin as an effective treatment for MLL-fusion leukaemia. *Nature* 2011, 478(7370):529-533.

DE WIT E, de LAAT W: A decade of 3C technologies: insights into nuclear organization. *Genes Dev* 2012, 26(1):11-24.

DEATON AM, BIRD A: CpG islands and the regulation of transcription. *Genes Dev* 2011, 25(10):1010-1022.

DEKKER J, RIPPE K, DEKKER M, KLECKNER N: Capturing chromosome conformation. *Science* 2002, 295(5558):1306-1311.

DOSTIE J, RICHMOND TA, ARNAOUT RA, SELZER RR, LEE WL, HONAN TA, RUBIO ED, KRUMM A, LAMB J, NUSBAUM C *et al.* Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res* 2006, 16(10):1299-1309.

ESTELLER M: Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat Rev Genet* 2007, 8(4):286-298.

FRASER P, BICKMORE W: Nuclear organization of the genome and the potential for gene regulation. *Nature* 2007, 447(7143):413-417.

FULLWOOD MJ, LIU MH, PAN YF, LIU J, XU H, MOHAMED YB, ORLOV YL, VELKOV S, HO A, MEI PH *et al.* An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* 2009, 462(7269):58-64.

FULLWOOD MJ, RUAN Y: ChIP-based methods for the identification of long-range chromatin interactions. *J Cell Biochem* 2009, 107(1):30-39.

GIRESI PG, KIM J, MCDANIELL RM, IYER VR, LIEB JD: FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res* 2007, 17(6):877-885.

GREER EL, MAURES TJ, UCAR D, HAUSWIRTH AG, MANCINI E, LIM JP, BENAYOUN BA, SHI Y, BRUNET A: Transgenerational epigenetic inheritance of longevity in Caenorhabditis elegans. *Nature* 2011, 479(7373):365-371.

GUO JU, SU Y, ZHONG C, MING GL, SONG H: Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. *Cell* 2011, 145(3):423-434.

GUTTMAN M, AMIT I, GARBER M, FRENCH C, LIN MF, FELDSER D, HUARTE M, ZUK O, CAREY BW, CASSADY JP *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 2009, 458(7235):223-227.

GUTTMAN M, DONAGHEY J, CAREY BW, GARBER M, GRENIER JK, MUNSON G, YOUNG G, LUCAS AB, ACH R, BRUHN L *et al.* lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* 2011, 477(7364):295-300.

GUTTMAN M, RINN JL: Modular regulatory principles of large non-coding RNAs. *Nature* 2012, 482(7385):339-346.

HARRIS RA, WANG T, COARFA C, NAGARAJAN RP, HONG C, DOWNEY SL, JOHNSON BE, FOUSE SD, DELANEY A, ZHAO Y *et al.* Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat Biotechnol* 2010, 28(10):1097-1105.

HEBBES TR, THORNE AW, CRANE-ROBINSON C: A direct link between core histone acetylation and transcriptionally active chromatin. *Embo J* 1988, 7(5):1395-1402.

HESSELBERTH JR, CHEN X, ZHANG Z, SABO PJ, SANDSTROM R, REYNOLDS AP, THURMAN RE, NEPH S, KUEHN MS, NOBLE WS *et al.* Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat Methods* 2009, 6(4):283-289.

HO L, CRABTREE GR: Chromatin remodelling during development. *Nature* 2010, 463(7280):474-484.

HORSTHEMKE B, CLAUSSEN U, HESSE S, LUDECKE HJ: PCR-mediated cloning of HpaII tiny fragments from microdissected human chromosomes. *PCR Methods Appl* 1992, 1(4):229-233.

ILIK I, AKHTAR A: roX RNAs: non-coding regulators of the male X chromosome in flies. *RNA Biol* 2009, 6(2):113-121.

JACINTO FV, BALLESTAR E, ESTELLER M: Methyl-DNA immunoprecipitation (MeDIP): hunting down the DNA methylome. *Biotechniques* 2008, 44(1):35, 37, 39 passim.

JEFFERY L, NAKIELNY S: Components of the DNA methylation system of chromatin control are RNA-binding proteins. *J Biol Chem* 2004, 279(47):49479-49487.

KANELLOPOULOU C, MULJO SA, KUNG AL, GANESAN S, DRAPKIN R, JENUWEIN T, LIVINGSTON DM, RAJEWSKY K: Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. *Genes Dev* 2005, 19(4):489-501.

KOERNER MV, PAULER FM, HUANG R, BARLOW DP: The function of non-coding RNAs in genomic imprinting. *Development* 2009, 136(11):1771-1783.

KU CS, NAIDOO N, WU M, SOONG R: Studying the epigenome using next generation sequencing. *J Med Genet* 2011, 48(11):721-730.

LIEBERMAN-AIDEN E, VAN BERKUM NL, WILLIAMS L, IMAKAEV M, RAGOCZY T, TELLING A, AMIT I, LAJOIE BR, SABO PJ, DORSCHNER MO *et al*. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 2009, 326(5950):289-293.

LISTER R, O'MALLEY RC, TONTI-FILIPPINI J, GREGORY BD, BERRY CC, MILLAR AH, ECKER JR: Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* 2008, 133(3):523-536.

LISTER R, PELIZZOLA M, DOWEN RH, HAWKINS RD, HON G, TONTI-FILIPPINI J, NERY JR, LEE L, YE Z, NGO QM *et al*. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 2009, 462(7271):315-322.

MARTENS JH, STUNNENBERG HG, LOGIE C: The decade of the epigenomes? *Genes Cancer* 2011, 2(6):680-687.

MARTIANOV I, RAMADASS A, SERRA BARROS A, CHOW N, AKOULITCHEV A: Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature* 2007, 445(7128):666-670.

MATTICK JS, AMARAL PP, DINGER ME, MERCER TR, MEHLER MF: RNA regulation of epigenetic processes. *Bioessays* 2009, 31(1):51-59.

MIGICOVSKY Z, KOVALCHUK I: Epigenetic memory in mammals. *Front Genet* 2011, 2:28.

MIKKELSEN TS, KU M, JAFFE DB, ISSAC B, LIEBERMAN E, GIANNOUKOS G, ALVAREZ P, BROCKMAN W, KIM TK, Koche RP *et al*. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 2007, 448(7153):553-560.

MISTELI T: Beyond the sequence: cellular organization of genome function. *Cell* 2007, 128(4):787-800.

MONDAL T, RASMUSSEN M, PANDEY GK, ISAKSSON A, KANDURI C: Characterization of the RNA content of chromatin. *Genome Res* 2010, 20(7):899-907.

MORGAN HD, SUTHERLAND HG, MARTIN DI, WHITELAW E: Epigenetic inheritance at the agouti locus in the mouse. *Nat Genet* 1999, 23(3):314-318.

Moving AHEAD with an international human epigenome project. *Nature* 2008, 454(7205):711-715.

NG RK, GURDON JB: Epigenetic memory of an active gene state depends on histone H3.3 incorporation into chromatin in the absence of transcription. *Nat Cell Biol* 2008, 10(1):102-109.

NG SF, LIN RC, LAYBUTT DR, BARRES R, OWENS JA, MORRIS MJ: Chronic high-fat diet in fathers programs beta-cell dysfunction in female rat offspring. *Nature* 2010, 467(7318):963-966.

NOLL M, THOMAS JO, Kornberg RD: Preparation of native chromatin and damage caused by shearing. *Science* 1975, 187(4182):1203-1206.

PAL-BHADRA M, LEIBOVITCH BA, GANDHI SG, RAO M, BHADRA U, BIRCHLER JA, ELGIN SC: Heterochromatic silencing and HP1 localization in Drosophila are dependent on the RNAi machinery. *Science* 2004, 303(5658):669-672.

PALSTRA RJ, TOLHUIS B, SPLINTER E, NIJMEIJER R, GROSVELD F, DE LAAT W: The beta-globin nuclear compartment in development and erythroid differentiation. *Nat Genet* 2003, 35(2):190-194.

PAUL J, DUERKSEN JD: Chromatin-associated RNA content of heterochromatin and euchromatin. *Mol Cell Biochem* 1975, 9(1):9-16.

RAKYAN VK, DOWN TA, BALDING DJ, BECK S: Epigenome-wide association studies for common human diseases. *Nat Rev Genet* 2011, 12(8):529-541.

RECHAVI O, MINEVICH G, HOBERT O: Transgenerational inheritance of an acquired small RNA-based antiviral response in C. elegans. *Cell* 2011, 147(6):1248-1256.

RINN JL, KERTESZ M, WANG JK, SQUAZZO SL, XU X, BRUGMANN SA, GOODNOUGH LH, HELMS JA, Farnham PJ, Segal E *et al*. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* 2007, 129(7):1311-1323.

RODRIGUEZ-CAMPOS A, AZORIN F: RNA is an integral component of chromatin that contributes to its structural organization. *PLoS One* 2007, 2(11):e1182.

SABO PJ, KUEHN MS, THURMAN R, JOHNSON BE, JOHNSON EM, CAO H, YU M, ROSENZWEIG E, GOLDY J, HAYDOCK A *et al*. Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nat Methods* 2006, 3(7):511-518.

SCHMITZ KM, MAYER C, POSTEPSKA A, GRUMMT I: Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. *Genes Dev* 2010, 24(20):2264-2269.

SERRE D, LEE BH, TING AH: MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. *Nucleic Acids Res* 2010, 38(2):391-399.

SIMONIS M, KOOREN J, DE LAAT W: An evaluation of 3C-based methods to capture DNA interactions. *Nat Methods* 2007, 4(11):895-901.

SONG L, CRAWFORD GE: DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc* 2010, 2010(2):pdb prot5384.

SPLINTER E, DE LAAT W: The complex transcription regulatory landscape of our genome: control in three dimensions. *Embo J* 2011, 30(21):4345-4355.

STRAHL BD, ALLIS CD: The language of covalent histone modifications. *Nature* 2000, 403(6765):41-45.

TABERLAY PC, KELLY TK, LIU CC, YOU JS, DE CARVALHO DD, MIRANDA TB, ZHOU XJ, LIANG G, JONES PA: Polycomb-repressed genes have permissive enhancers that initiate reprogramming. *Cell* 2011, 147(6):1283-1294.

TALBERT PB, HENIKOFF S: Histone variants--ancient wrap artists of the epigenome. *Nat Rev Mol Cell Biol* 2010, 11(4):264-275.

TSAI MC, MANOR O, WAN Y, MOSAMMAPARAST N, WANG JK, LAN F, SHI Y, SEGAL E, CHANG HY: Long noncoding RNA as modular scaffold of histone modification complexes. *Science* 2010, 329(5992):689-693.

VAN BERKUM NL, DEKKER J: Determining spatial chromatin organization of large genomic regions using 5C technology. *Methods Mol Biol* 2009, 567:189-213.

WAGNER KD, WAGNER N, GHANBARIAN H, GRANDJEAN V, GOUNON P, CUZIN F, RASSOULZADEGAN M: RNA induction and inheritance of epigenetic cardiac hypertrophy in the mouse. *Dev Cell* 2008, 14(6):962-969.

WANG KC, YANG YW, LIU B, SANYAL A, CORCES-ZIMMERMAN R, CHEN Y, LAJOIE BR, PROTACIO A, FLYNN RA, GUPTA RA *et al*. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* 2011, 472(7341):120-124.

WU C, BINGHAM PM, LIVAK KJ, HOLMGREN R, ELGIN SC: The chromatin structure of specific genes: I. Evidence for higher order domains of defined DNA sequence. *Cell* 1979, 16(4):797-806.

WU H, ZHANG Y: Mechanisms and functions of Tet protein-mediated 5-methylcytosine oxidation. *Genes Dev* 2011, 25(23):2436-2452.

XHEMALCE B, DAWSON M, BANNISTER, AJ: Histone Modifications. In: *Meyers: Encyclopedia of Molecular Cell Biology and Molecular Medicine: Epigenetic Regulation and Epigenomics*. Edited by Meyers RA; 2011.

12

Metagenômica

Luciana Principal Antunes
Julio Cézar Franco de Oliveira

Introdução

A elaboração de estimativas de diversidade filogenética da microbiota presente em nosso planeta é um desafio que persiste desde a descoberta da vida microbiana no final do século XIX e que trouxe implicações em diferentes áreas como na saúde humana e no desenvolvimento de processos e produtos tecnológicos (biotecnológicos). A Metagenômica é uma abordagem de acesso a essa diversidade, a qual possibilita a análise genômica de comunidades microbianas sem a necessidade da etapa prévia de cultivo. Trabalhos nesta área vêm sendo desenvolvidos a partir do final da década de 1980, porém de forma mais expressiva após o advento das tecnologias de sequenciamento de alto-desempenho (Pace, RN, 1997). Ela é particularmente poderosa porque possibilita o estudo de genomas de microrganismos cultiváveis e não-cultiváveis, uma vez que as estimativas mais otimistas apontam que apenas cerca de 1% da microbiota existente na biosfera é passível de cultivo em condições de laboratório, fato que pode ser atribuído ao desconhecimento das necessidades nutricionais destes organismos e, muitas vezes, o crescimento e multiplicação condicionado a sistemas de consórcio complexos em que o produto do metabolismo de um microrganismo é essencial para o estabelecimento de colônia microbiana de outro (Head, I.M et al, 1998). Como os meios de culturas não reproduzem as condições presentes nestes consórcios, os estudos de microbiologia historicamente estiveram limitados antes do advento da metagenômica.

Essa estratégia permite não apenas identificar quais microrganismos estão presentes em uma amostra/ambiente, com também a realização de análises mais aprofundadas de diversidade, dinâmica e ecologia de comunidades, promovendo uma melhor compreensão da microbiota existente a partir da enorme quantidade de dados disponibilizados com o sequenciamento de metagenomas; outra possibilidade

é a triagem de enzimas e de biomarcadores de interesse para o emprego em diferentes segmentos da indústria e da medicina (Lorenz, P and Eck, J, 2005). Na tarefa de caracterizar uma amostra microbiológica, o emprego da metatranscriptômica e da metaproteômica, estratégias derivadas da metagenômica e complementares a ela, permitem a integração entre as análises de ácidos nucléicos e de proteínas e a realização de inferências a dinâmica desta comunidade, o que está "fazendo" e "como", a partir da análise de expressão gênica e da determinação de quais proteínas estão presentes na mesma condição (Figura 1).

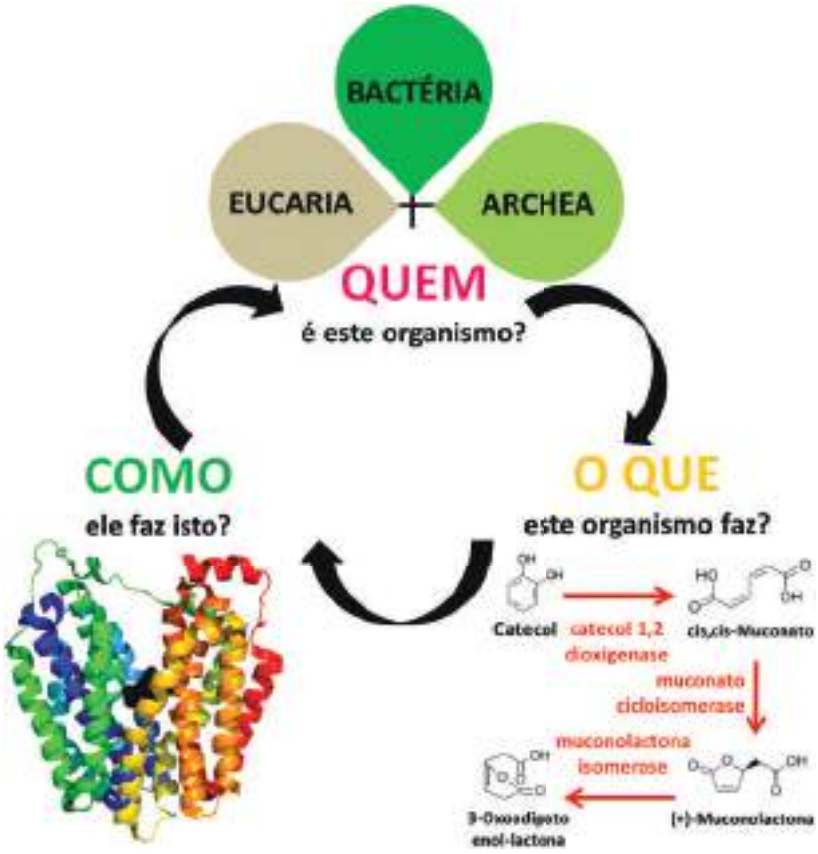


Figura 1. Análise integrada de comunidades microbianas com a caracterização: filogenética, metabólica e protéica, por Metagenômica, Metatranscriptômica e Metaproteômica.

A possibilidade de se isolar DNA a partir de amostras coletadas nos mais diversos *hábitats* e proceder à análise deste DNA representando o genoma coletivo das espécies contidas ali, constitui-se no fundamento básico da metagenômica (Fierer, N and Jackson, RB, 2006). Isto permite inserção num contexto no qual a sociedade, a indústria e a agropecuária se comprometem cada vez mais com o emprego de práticas menos nocivas ao meio ambiente e a saúde humana. Além disso, relaciona-se diretamente com a crescente necessidade de se descobrir novos arsenais terapêuticos

mais eficazes no tratamento de diversas patologias. Diante disso, novos paradigmas se colocam na pauta atual das pesquisas científicas e tecnológicas, para os quais a metagenômica tem muito a contribuir em termos de novas possibilidades para os desafios atuais e futuros (Banik, JJ and Brady, SF, 2010).

Neste capítulo abordaremos os desafios que se colocam para a obtenção, processamento e utilização racional da imensa quantidade de dados que vem sendo disponibilizados à medida que o censo da diversidade microbiana é refinado. Desafios estes, que envolvem o isolamento, sequenciamento e caracterização de ácidos nucléicos (DNA, RNA) e proteínas obtidas a partir da coleta de amostras provenientes dos mais diversos ambientes, busca de novas moléculas com funções desejadas, bem como novos processos metabólicos capazes de atender às demandas crescentes das sociedades contemporâneas.

Estudo da diversidade microbiana a partir de amostras ambientais através de análise do rRNA por PCR

A primeira etapa de qualquer trabalho de metagenômica é a obtenção do material genético de modo íntegro e com elevado grau de pureza a partir dos microrganismos presentes no ambiente que se pretende investigar, e que pode requerer a padronização de protocolos e otimizações de kits comerciais. Dependendo da amostra, isso terá um grau maior ou menor de complexidade e pode limitar a abordagem por metagenômica, sobretudo em ambientes onde predominam processos de decomposição da matéria orgânica e, portanto, apresentam um alto teor de substâncias húmicas originadas no mesmo, que por compartilharem propriedades físico-químicas similares aos dos ácidos nucléicos tornam ineficientes muitos métodos de purificação tradicionalmente empregados. Um agravante deste tipo de contaminação é que essas substâncias inibem reações enzimáticas empregadas em etapas subsequentes envolvendo técnicas de biologia molecular.

Superada a primeira etapa é possível dar sequência à caracterização da microbiota, tarefa que consiste em um dos maiores desafios da metagenômica e para a qual é amplamente utilizado o gene codificante do 16S rRNA como marcador filogenético para a identificação das espécies presentes no metagenoma e elaboração de estimativa da sua abundância relativa (Tringe, SG and Hugenholtz, P, 2008). Para isso, pode ser sequenciado exclusivamente o gene do 16S rRNA (em estudo de diversidade) ou extrair suas sequências, por bioinformática, a partir dos *reads* (sequências) gerados com o sequenciamento total do metagenoma. As sequências obtidas podem então ser comparadas com bancos de dados como: RDP (*Ribosomal Database Project*), SILVA e *Greengenes*. Uma alternativa ao sequenciamento é a marcação direta deste gene por hibridização de sondas fluorescentes (FISH) e posterior análise por microscopia (Amann et al, 1995).

Apesar da larga utilização do 16S rRNA como marcador filogenético, muitas vezes ele é alvo de críticas devido a alguns trabalhos relatarem o seu envolvimento em eventos de transferência horizontal. Esse gene também pode estar presente em diferentes números de cópias dependendo do organismo, o que restringe os estudos de abundância por promover a sobre-estimação de organismos que apresentam um

alto número de cópias e a sub-estimação daqueles com baixo número. Existem outros genes indicados para essa função, como o rpoB (subunidade beta da RNA polimerase) e o GyrB (subunidade beta da Girase), os quais estão presentes em uma única cópia por genoma. Contudo, o 16S rRNA ainda é o gene marcador filogenético mais utilizado e para o qual vários bancos de dados são disponíveis, o que facilita a análise e permite a comparação dos dados obtidos com de outros trabalhos publicados.

Nos estudos de diversidade a partir do 16S rRNA, novamente, é desejável que o sequenciamento do metagenoma tenha uma boa cobertura para que os membros da comunidade presente em baixo número possam ser detectados e não apenas os mais abundantes, o que é especialmente desafiador em ambientes altamente diversos. A cobertura do metagenoma é avaliada utilizando uma curva de rarefação com base no 16S rRNA (curva que correlaciona o número de espécies encontradas versus números de sequência analisadas), sendo que a cobertura ideal é obtida quando a curva atinge o platô (Figura 2). Neste ponto, o aumento do número de sequências analisadas não promove um aumento no número de espécies devido à diversidade total da comunidade já estar representada. O pacote de ferramentas QIIME (*Quantitative Insights Into Microbial Ecology*) de análise e comparação de comunidades microbianas permite a obtenção da curva de rarefação, da estimativa das diversidades alfa (diversidade dentro do hábitat) e beta (diversidade entre *hábitats*) com base em diferentes índices - filogenéticos e não-filogenéticos (Caporase, JG et al, 2010). A partir da estimativa da diversidade beta é possível obter a PCoA (*Principal Coordinate Analysis*), que é comumente utilizada para comparar grupos de amostras com a plotagem das mesmas em um gráfico 2D ou 3D e, desta forma, visualizar e identificar amostras com maior ou menor grau de correlação de grupos.

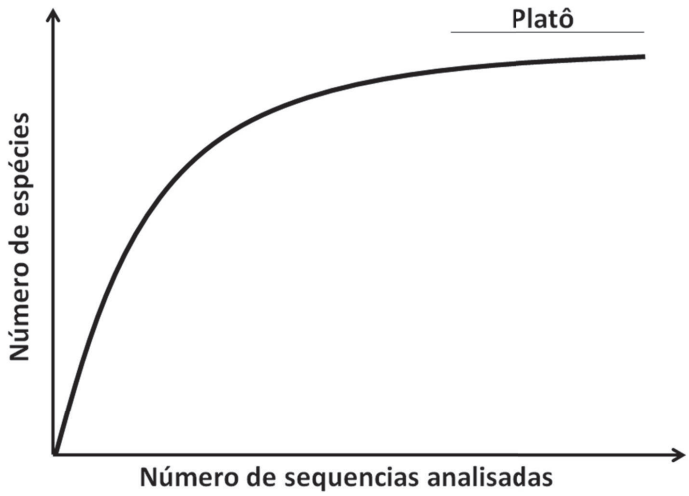


Figura 2. Curva de rarefação.

Um modo menos robusto de se caracterizar uma amostra microbiana ambiental, por ter uma resolução limitada, é a eletroforese em gel com gradiente desnaturante (DGGE) utilizando o DNA metagenômico. Esta técnica fornece um perfil ("impressão

digital") que pode ser utilizado como parâmetro de comparação entre metagenomas com o auxílio de softwares específicos. A "impressão digital" obtida pode ser um primeiro indício de diferenças populacionais da microbiota e justificar a utilização de outras estratégias mais sofisticadas a fim de elucidar essa diferença.

Metagenomas da microbiota ambiental e associada a hospedeiros

A estratégia de estudar a microbiota ambiental a partir do seu material genético (Metagenômica) desde que foi proposta por Pace e colaboradores (1997) possibilita a caracterização microbiana de diversos ambientes naturais e engenheirados pela ação humana como: marinho, fluvial, lacustre, fontes termais, gelo, salinas, solo, lodo e de processos envolvendo transformações microbianas em biorreatores e unidades de compostagem. Entre os projetos desenvolvidos, o do Mar do Sargasso de 2004 (o primeiro projeto metagenômico de grande escala) identificou mais de 1,2 milhões de proteínas e com isso promoveu a duplicação do número de proteínas conhecidas até então (Venter, JC et al, 2004) (Figura 3). Mais de 95% das proteínas identificadas neste trabalho eram novas e a maioria proveniente de organismos ainda não estudados. Isto exemplifica o quão limitado era o conhecimento disponível até recentemente sobre o mundo microbiano e o grande impulso que a metagenômica trouxe para a expansão do mesmo e, por conseguinte do número de proteínas conhecidas. Em outro grande projeto, o Global Ocean Survey (GOS), foram coletadas amostras de águas dos oceanos ao redor do mundo e preditas 6,1 milhões de proteínas (Nelson, KH and Venter, JC, 2007) (Figura 3). O volume de dados gerados em projetos destas proporções é assustador e o seu processamento ainda constitui em um grande desafio para a bioinformática, apesar do desenvolvimento de interfaces de anotação automatizada.

Figura 3. Crescimento no número de proteínas, depositadas em bases de dados, com o advento dos projetos de metagenômica.

Outra vertente da metagenômica é o estudo dos microrganismos associados a um hospedeiro, a qual é alvo de investigação de diversos projetos que visam estabelecer correlações entre microbiota e o estado de saúde, idade, sexo e nutrição do hospedeiro. Essa microbiota também pode ser explorada para a triagem de enzimas, uma vez que, já no primeiro estudo dos microrganismos do trato intestinal humano, com apenas dois indivíduos amostrados, foram identificadas 50.000 novas proteínas (Ley, RE et al, 2008).

Neste contexto, o corpo do hospedeiro pode ser considerado como uma "ilha" constituída por distintos *habitats*, que se tornam disponíveis após o seu nascimento para colonização microbiana. Esta ilha passa a ser, portanto, crucial para o desenvolvimento de um determinado hábito, como a dieta baseada na degradação de celulose de Térmita (cupim) e de ruminantes, que é possibilitada pela associação dos mesmos com microrganismos produtores das enzimas necessárias para a degradação da celulose. Esses microrganismos despertam um especial interesse para a triagem de enzimas celulolíticas e culminou no trabalho do metagenoma do rúmen bovino com a identificação de 27.755 genes associados com a degradação de carboidratos, sendo que 43% apresentaram similaridade menor do que 50% com os depositados nos bancos de dados (Ross, EM, et al, 2012). Ou seja, isso indica a presença de enzimas novas, entre as quais podem estar enzimas mais robustas do que as disponíveis atualmente para aplicação em processos industriais como na produção de biocombustíveis a partir de matéria orgânica vegetal.

A metagenômica de microrganismos associados a um hospedeiro também tem o potencial de fornecer biomarcadores para o diagnóstico de doenças e pré-disposição a obesidade, permitir o estabelecimento de uma nutrição mais adequada visando suprir as necessidades também da microbiota e a otimização da absorção de medicamentos, entre inúmeras outras possibilidades que este conhecimento pode proporcionar para a saúde humana, em particular. Com relação à obesidade, estudos utilizando o modelo murino (modelo animal para o estudo de diversas doenças humanas) revelaram possíveis alvos de modulação da microbiota intestinal como controle da obesidade, somado a isto o monitoramento da microbiota intestinal de murino ao longo da administração de antibióticos revelou a ocorrência de uma mudança na sua composição, com a diminuição de representantes do filo Firmicutes. Este filo, por consequência, é encontrado em maior frequência em indivíduos obesos e é associado com uma maior eficiência da obtenção de energia a partir de carboidratos da dieta (Turnbaugh et al., 2006; Million et al., 2013).

Semelhante ao corpo de animais, superfícies de plantas e algas também podem abrigar comunidades microbianas e o seu estudo pode permitir a inferência de modelos de ecologia microbiana. Como é caso do metagenoma associado à superfície da macroalga *Ulva australis,* que indicou que o modelo de "Loteria" competitivo era o responsável pela estruturação dessa comunidade (Burke, C et al, 2011). Este modelo, em analogia a uma loteria, cujo conjunto de bolinhas apresenta o mesmo peso e volume e, portanto, são igualmente prováveis de serem sorteadas independentemente da sua numeração, assume que todas as espécies são funcionalmente equivalentes na comunidade, logo, possuem a mesma chance de colonizar o ambiente uma vez que, o mesmo esteja vago e, posteriormente, com as espécies competindo pelo espaço após a colonização e a coexistência entre elas mediada por eventos aleatórios. No estudo,

foi verificada uma baixa correspondência das espécies associadas a diferentes *U. australis* (17%) enquanto que a similaridade de função foi alta (75%), sugerindo que este sistema pode ser melhor caracterizado pela função de genes, os quais estão distribuídos amplamente entre os grupos presentes nesta comunidade. O mesmo pode ser aplicado a outras microbiotas, como a do intestino humano devido à acentuada variação em termos de espécies encontrada entre diferentes indivíduos, porém com o compartilhamento de um conjunto de genes, que a caracteriza em termos funcionais e não taxonômico.

Sequenciamento de alto-desempenho de metagenomas

O poder da metagenômica em propiciar o estudo de microrganismos cultiváveis e não-cultiváveis foi significativamente expandido pelas novas tecnologias de sequenciamento de alto-desempenho, com obtenção de metagenomas de modo rápido e a custos abordáveis. Além disso, permitiu o sequenciamento de microrganismos sem a necessidade da etapa prévia de clonagem e multiplicação em células hospedeiras, que consistia em um dos gargalos aos estudos de metagenoma. O primeiro sequenciador de alto desempenho lançado foi o pirossequenciador 454 da Roche em 2005. Em seguida outras plataformas foram desenvolvidas: Illumina da Solexa, SOLID da Applied Biosystems, Heliscope da Helicos e o PacBio da Pacific Biosciences, as quais não necessitam da infra-estrutura de grandes centros de genômica e podem ser empregados em laboratórios convencionais. Estas tecnologias possibilitam a leitura de diferentes tamanhos de sequências (*reads*) a uma taxa de obtenção das sequências e de custos variados, podendo apresentar diferentes indicações dependendo da amostra e do objetivo. A seguir há descrição sucinta das tecnologias de sequenciamento de alto-desempenho (nova geração) disponíveis:

Os métodos de sequenciamento do 454 e SOLID (Capítulo 2) compartilham algumas etapas: iniciam com a fragmentação do DNA e a ligação dos fragmentos a oligonucleotídeos adaptadores, que são então imobilizados em esferas (*beads*), as quais são adicionadas a uma emulsão de água-óleo e englobadas em micelas, que funcionam como mini-reatores na reação de PCR (PCR em emulsão) para amplificação dos fragmentos de DNA. No sequenciamento utilizando a plataforma 454, as esferas contendo DNA são isoladas e depositadas nos poços de uma placa de fibra óptica. São adicionados sequencialmente à reação os reagentes dNTP, tampão, e a DNA polimerase na presença da enzima quimioluminescente luciferase, a qual utiliza o ATP produzido a partir do pirofosfato liberado com a incorporação de um dNTP complementar à fita molde do DNA para produzir luz, cujo feixe gerado é registrado e analisado computacionalmente. Ao final são sequenciados fragmentos de DNA de cerca de 400 nucleotídeos, podendo chegar até 700 nucleotídeos a depender do kit de sequenciamento utilizado. Já com a plataforma SOLID, o DNA amplificado nas esferas são depositados em uma placa e, então, hibridizados sequencialmente com sondas aleatórias de oligonucleotídeos contendo um dinucleotídeo conhecido na extremidade 3′ e um fluoróforo. Após cinco ciclos de replicação a fita sintetizada é des-hibridizada da fita molde de DNA e o ciclo de replicação se repete, porém começando a partir de uma base imediatamente *upstream* com relação ao ciclo

anterior. A decodificação de cada base é permitida pela repetição de vários ciclos, os quais reduzem erros na leitura das sequências e é indicada para a identificação de polimorfismos de um único nucleotídeo (SNPs). Com a Plataforma SOLID são produzidos fragmentos de sequências (*reads*) de 35-50 nucleotídeos para cada leitura ao final do processo.

No sequenciamento com a plataforma Illumina–Solexa (Capítulo 2), o DNA é fragmentado, ligado a adaptados e então ancorado em uma plataforma sólida contendo adaptadores complementares aos ligados ao DNA a ser sequenciado. Os fragmentos são amplificados por PCR para produzir aglomerados clonais, os quais posteriormente são sequenciados utilizando nucleotídeos-terminadores reversíveis com uma marcação fluorescente específica para cada uma das quatro bases. Após cada ciclo de síntese, é feita uma imagem identificando qual base foi incorporada e a marcação fluorescente é removida antes que um novo ciclo de polimerização reinicie. Essa tecnologia produz sequências de até 600 nucleotídeos de comprimento.

No método de Heliscope são gerados fragmentos de DNA de 100-200 nucleotídeos, os quais são ligados a adaptadores e imobilizados em uma plataforma de microfluídica. São adicionados, sequencialmente, nucleotídeos com uma marcação luminosa que é registrada após cada ciclo de amplificação e utilizada na análise das imagens geradas para identificação de qual base foi incorporada em cada fita molde.

No método desenvolvido pela Pacific Biosciences, a DNA polimerase é imobilizada em poços com volume de zeptolitros. Os fragmentos de DNA molde, de fita simples (não é necessária amplificação prévia), são adicionados aos poços juntamente com dNTPs com fósforo marcado com uma sonda fluorescente, o qual é removido e detectado após a incorporação de um novo nucleotídeo a fita. Nos poços os nucleotídeos marcados são mantidos em concentrações elevadas, permitindo uma alta eficiência de reação. Neste método, pode ser gerada sequências com até milhares de nucleotídeos (Quail et. al., 2012).

As tecnologias de sequenciamento de alto-desempenho geram quantidades volumosas de sequências que precisam ser posteriormente montadas em *contigs* e analisadas por bioinformática. A montagem pode ser feita pelo método de sobreposição, com a obtenção de uma sequência consenso (*contig*) a partir de pequenas sequências total ou parcialmente complementares, ou através da montagem gráfica como com o método de Bruijn, que divide as sequências em *k-mers* e calcula a distância entre elas. Com a montagem, é obtido um rascunho do genoma sequenciado, o qual pode ser completamente resolvido com sequenciamento tradicional (Método de Sanger – ver Capítulo 2), complementar para a obtenção das sequências de lacunas que persistiram após a montagem ou pela utilização de outra tecnologia de alto desempenho (Figura 4). Por exemplo, uma amostra pode ser sequenciada utilizando as plataformas Roche 454 e Illumina-Solexa, de forma que limitações inerentes a cada uma das abordagens possam ser amenizadas com a combinação das duas metodologias. Entretanto, em muitos casos o rascunho do genoma é suficiente e, deste modo, não é necessário a realização de sequenciamentos adicionais para o "acabamento", que pode consistir em uma etapa economicamente custosa e demorada.

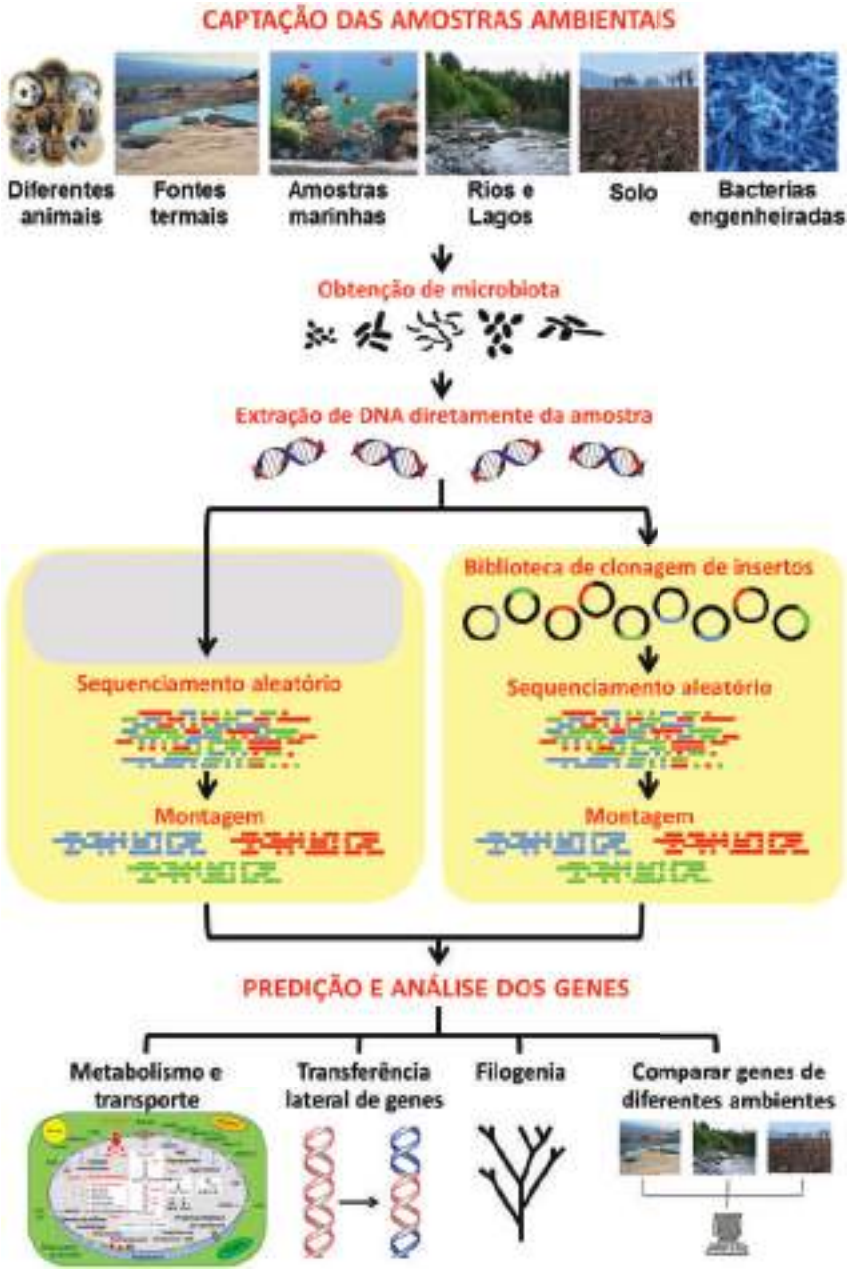


Figura 4. Etapas do estudo de metagenomas.

Análise de metadados obtidos com sequenciamento de alto-desempenho

O grande volume de metadados gerados pelas tecnologias de sequenciamento de alto-desempenho, seja na forma de *reads* ou de *contigs* (após montagem), podem ser depositados e analisados em servidores como MG-RAST e JGI-IMG. Estas interfaces

realizam a anotação automática do metagenoma em um intervalo curto de tempo, permitindo atribuir funções e determinar quais organismos estão presentes (Figura 4) através da busca por similaridade com sequências depositadas em bases de dados, como: de 16S rRNA (RDP), COG (*Clusters of orthologous groups of proteins*), NOG (*Non-supervised orthologous group*), KO (*KEGG orthology*) e SEED, que estão reunidos nestes servidores, permitindo a comparação de um metagenoma com outros previamente disponíveis. O MG-RAST também fornece uma etapa preliminar de de-replicação de sequências, um artefato inerente as plataformas 454 e SOLID, que pode resultar em um viés nas análises caso essa etapa não seja realizada. Adicionalmente pode ser necessária a realização de análise local dos dados.

Na tarefa de se analisar os dados de metagenômica, a abordagem denominada *Binning* ou classificação é particularmente poderosa por não se restringir a utilização de genes marcadores filogenéticos, como o gene do 16S rRNA, o qual apesar de permitir um alto grau de resolução taxonômico possui a limitação de estar presente em baixas proporções nos metagenomas (cerca de 1% das sequências). Com o binning é possível, a partir de uma mistura de sequências provenientes de organismos distintos identificar a qual organismo cada uma delas pertence. Isso pode ser conseguido utilizando como base a similaridade entre sequências do metagenoma com as depositadas em bancos de referência através da ferramenta BLAST, o MEGAN (*Metagenome Analyzer*) e o CARMA, que são exemplos de interfaces que utilizam a estratégia de binning por similaridade. *Binning* também pode utilizar a composição das sequências como: o conteúdo CG (Figura 5), preferência de códon e a frequências dos nucleotídeos, as ferramentas PhyloPythia e TETRA permitem a atribuição de uma dada sequência a um organismo por confrontar essas características com a de genomas de referência.

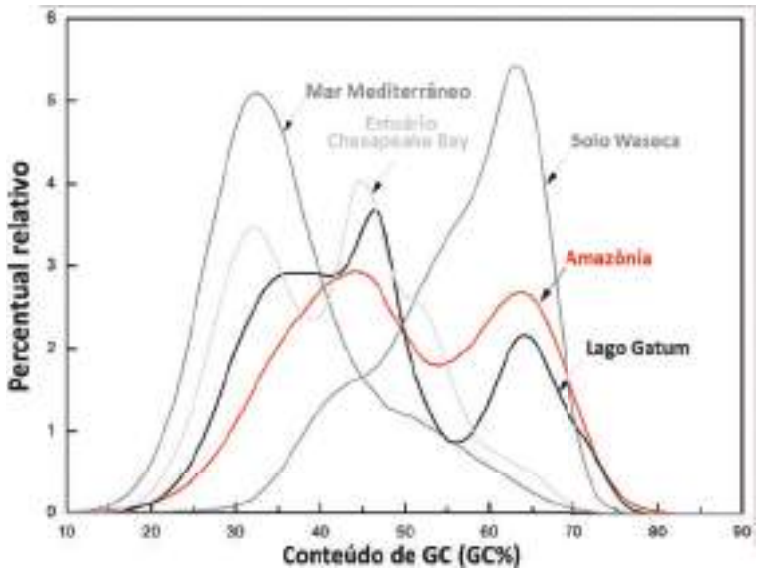


Figura 5. Perfil de distribuição do conteúdo CG obtido no MG-RAST dos metagenomas: do Mar Mediterrâneo, Estuário Chesapeake, Solo Waseca, Amazônia e Lago Gatum. Fonte: (*Ghai et. al.*, 2011).

O sequenciamento pode ser um fator limitante na análise do metagenoma caso ele forneça uma baixa cobertura do mesmo, isso é especialmente crítico em ambientes com alta diversidade de espécies e com a presença de muitos membros raros na comunidade. Em alguns casos, a caracterização funcional do metagenoma, ou seja, a identificação de quais funções gênicas estão representadas, pode ser mais apropriado do que tentar caracterizá-lo taxonomicamente, já que o sequenciamento pode ser capaz de apenas revelar quais espécies estão mais abundantes e não toda a diversidade presente na amostra. Com a caracterização funcional é obtido um conjunto de funções gênicas que é utilizada para caracterizar um metagenoma, e para confrontá-lo com outros, além de permitir a realização de inferências entre a utilização/disponibilidade de um determinado substrato e condições físico-químicas com as funções encontradas. Esta abordagem de caracterização funcional de metagenomas foi proposta por Tringe, SG et al. (2005), que também sugeriu o termo Tag-gene ambiental (EGT) em analogia a etiquetas de sequências expressas (EST – Capítulo 2), devido à natureza fragmentada dos dados, proporcionando uma "impressão digital" de cada ambiente.

O link funcional que estava faltando nas análises da metagenômica

Existe uma necessidade crescente de se fazer um "link" entre o potencial genômico revelado com a publicação de metagenomas de diversos ambientes, obtido com o sequênciamento "total" dos genes e não apenas do 16S rRNA (gene marcador filogenético), e as reais transformações que estão ocorrendo nos mesmos em função da presença de enzimas específicas e em resposta a mudanças ambientais, uma vez que a simple presença de um gene não implica que ele esteja sendo expresso e, menos ainda, que a sua proteína relacionada foi traduzida e esteja ativa.

Para fazer essa ponte entram em cena a Metatranscriptômica e a Metaproteômica, derivadas de abordagens que antes se restringiam a análise de culturas de microrganismos puras e agora permitem a detecção de perfis de expressão derivados diretamente da microbiota ambiental. A metatranscriptômica, análise do trascriptoma de comunidades microbianas, permite analisar a variação da expressão gênica global como resultado de alterações ambientais ou da disponibilidade de um recurso com a identificação de quais genes estão sendo expressos em cada condição, o que pode direcionar e facilitar a triagem de enzimas de interesse. A etapa crítica dessa abordagem, e que consiste em um desafio, é a purificação do mRNA a partir de amostras ambientais complexas devido a sua conhecida instabilidade e a presença de enzimas RNases no meio. Uma vez superada essa limitação, podem ser empregadas tecnologias baseadas na hibridização de ácidos nucléicos em arranjos de sequências gênicas (DNAs genômicos ou cDNAs) imobilizados em suportes sólidos ("microarrays" – Capítulo 8) e na construção e sequenciamento de bibliotecas de cDNA obtidas a partir de mRNAs isolados de amostras ambientais (Capítulo 2). Alternativamente o sequenciamento pode ser feito diretamente dos cDNAs sem a construção de bibliotecas em hospedeiro, utilizando tecnologias sequenciamento de alto desempenho relatadas anteriormente neste capítulo.

A expressão gênica de amostras ambientais já podem ser avaliadas utilizando microarranjos comerciais que abrigam milhares de sondas, o que é especialmente

interessante quando o objetivo é o diagnóstico da contaminação ou não de um ambiente com um metal, por exemplo, analisando o perfil de expressão de enzimas da microbiota envolvida em vias de biodegradação ou de resistência ao mesmo. No entanto, apesar da significativa contribuição na metatranscriptômica na geração de conhecimento de diversas comunidades microbianas em diferentes ecossistemas, as tecnologias de arranjos de DNA e de sequenciamento de bibliotecas de cDNA implicam na geração de resultados tendenciosos capazes de diminuir a amplitude das análises. Ensaios de microarranjo só podem detectar a expressão de genes previamente conhecidos e selecionados para a montagem do arranjo, fator que limita a análise e priva a descoberta de novos genes. A abordagem de sequenciamento de cDNA não tem essa limitação e, assim, pode ser utilizada para a descoberta de novos genes e de vias metabólicas devido ao caráter aleatório do processo de construção de bibliotecas e o seu posterior sequenciamento, o que contorna a limitação da escolha de sequências a serem analisadas como no caso anterior. Contudo, podem ocorrer outros desvios na geração dos dados como, por exemplo, de abundância relativa de sequências de origem eucariótica (amplificadas a partir de mRNA) em relação a sequências de origem bacteriana. Outro fator limitante a ambas abordagens é que a simples presença de um mRNA de uma amostra não significa, necessariamente, que o mesmo esteja sendo traduzido com a produção da sua proteína correspondente, já que nem todo gene expresso é efetivamente traduzido em proteína.

Adicionalmente pode ser realizado o estudo do conjunto de proteínas expressas coletivamente por uma microbiota ambiental, abordagem convencionalmente denominada metaproteoma, assim uma mesma amostra coletada para o isolamento de ácidos nucléicos e destinada a análises de metagenoma e metatranscriptômica pode ser ainda analisada por metaproteômica. A metaproteômica foi proposta por Wilmes e Bond, 2006 e talvez seja a estratégia mais direta para relatar quais reações estavam ocorrendo em um dado ambiente no momento da amostragem. Tem o potencial não apenas de identificar quais proteínas estavam presentes, como também as proporções entre elas, interações poteína-proteína, taxa de renovação e se há modificações pós-traducionais que regulam a sua atividade. Para isso, são empregadas técnicas como eletroforese bidimensional em gradiente desnaturante (SDS-PAGE) seguida por espectrometria de massas (Figura 6) (Capítulo 9).

Vale salientar que qualquer que seja a metodologia utilizada na geração de Meta dados em larga escala é imprescindível o uso de ferramentas robustas de análise computacional e a integração dos mesmos é denominada de metainteractômica.

Metagenômica e biotecnologia

Processos produtivos sustentáveis de base biotecnologia e pesquisas visando terapias médicas inovadoras são geradores de demandas constantes em termos de novidades que atendam às expectativas da indústria e da medicina com o suprimento de enzimas mais robustas como: lípases, amidases, celulases e proteases, além de novos antibióticos, imunossupressores, antineoplásicos, entre outros. A metagenômica tem potencial de atender a estas demandas, pois permite acesso inédito a uma fonte sem precedentes de conhecimentos advindos de uma diversidade biológica e molecular

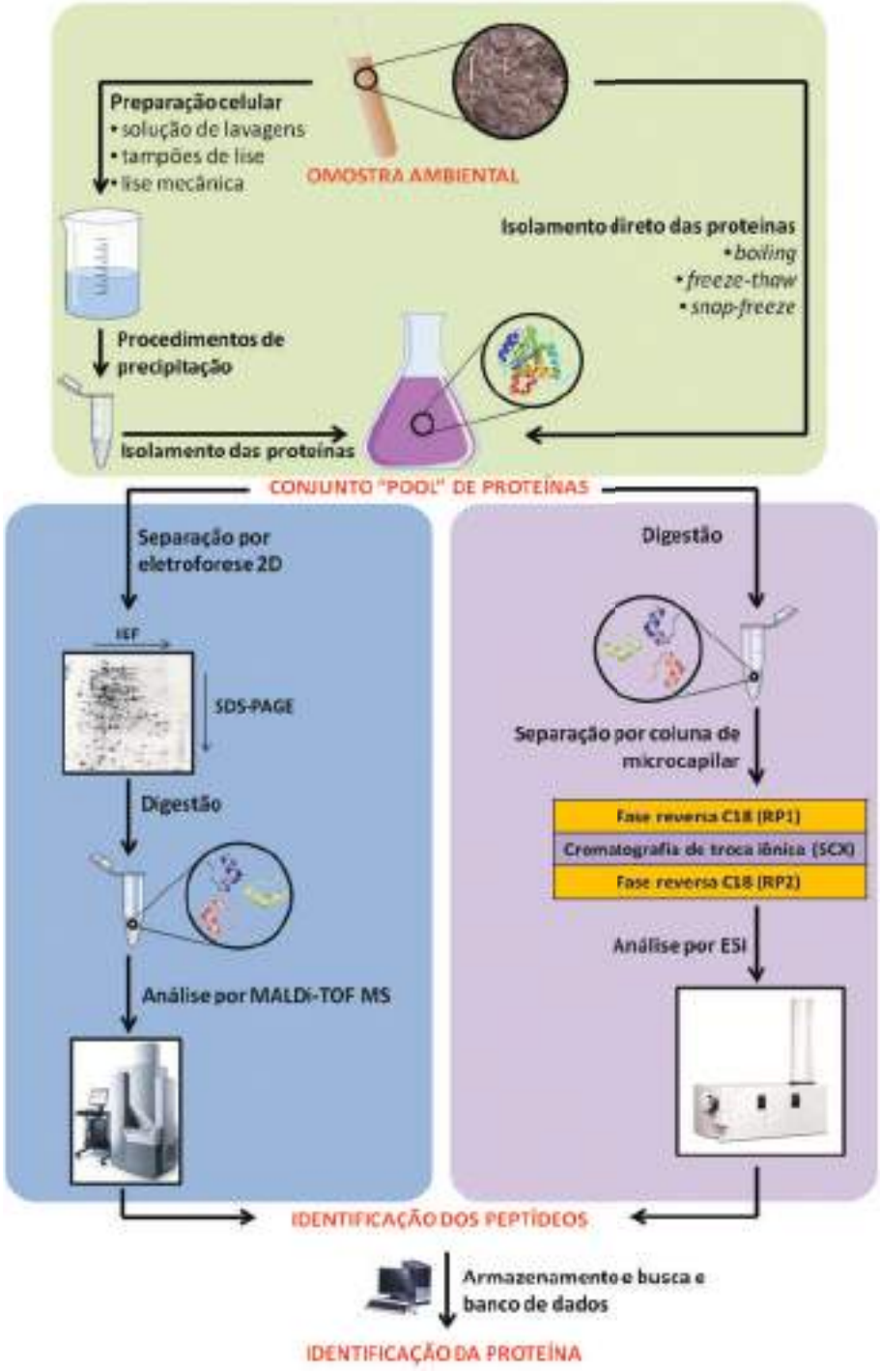


Figura 6. Etapas envolvidas no processo de metaproteômica.

ainda inexplorada em nosso planeta. Para isso, podem ser construídas bibliotecas de expressão a partir do DNA metagenômico total com a fragmentação e ligação destes em vetores adaptados para a expressão em hospedeiros heterólogos (E. coli, leveduras, etc) e posterior triagem de enzimas específicas nos clones ou o sequenciamento de alto desempenho do DNA, análise por bioinformática das sequências, síntese do gene de interesse triado neste metagenoma e a sua clonagem em um vetor apropriado para expressão em larga escala em células hospedeiras compatíveis. A escolha do vetor determina o tamanho do inserto de DNA que pode ser clonado, o número de clones necessários para a cobertura da metagenoma e quais produtos gênicos poderão ser triados, uma vez que, genes como os envolvidos na biossíntese de antibióticos normalmente fazem parte de grandes *Operons* e, portanto, neste caso a construção de uma biblioteca de insertos pequenos em vetores plasmidiais não é indicada.

Na triagem dos clones são empregadas diferentes estratégias de acordo com a enzima pretendida, todavia a mais utilizada é o crecimento dos clones em meios seletivos ou ricos em uma dada fonte de nutriente para a observação da formação de halo quando essa fonte é degradada por ação enzimática. Neste caso, a formação de halo indica que o clone contém um segmento de DNA com um gene codificante de uma proteína envolvida na degradação do substrato específico fornecido e que a mesma foi expressa, traduzida e secretada na forma ativa pelo maquinário da célula hospedeira. A automatização da triagem dos clones com o emprego de robôs diminui extremamente o tempo dispendido nesta etapa, sobretudo em bibliotecas de fragmentos pequenos (2-5kb) com um número elevado de clones, o que facilita e deixa de ser um fator limitante desta abordagem.

Alternativamente, podem ser desenhados oligonucleotideos iniciadores para amplificação por PCR de sequências específicas a partir do DNA metagenômico e o produto desta amplificação é que é utilizado na clonagem e construção de biblioteca. Seja no caso de bibliotecas de DNA de metagenoma ou produtos de PCR amplificados a partir do mesmo, a triagem funcional subsequente estará sempre limitada à necessidade destes genes serem expressos em organismos representando sistemas heterólogos de expressão. Ou seja, não correspondendo ao organismo original de onde o gene de interesse foi isolado. Os sistemas heterólogos mais utilizados para triagem funcional de clones a partir de bibliotecas de metagenoma tem sido *Escherichia coli*, *Streptomyces lividans*, *Pseudomonas putida*, *Bacillus subtilis*, bactérias do gênero *Rhyzobium* e sistemas eucarióticos de expressão (leveduras, células de insetos, células de mamíferos), compondo um arsenal genético, bioquímico e fisiológico adequados para a produção de biomoléculas ativas codificadas a partir de metagenomas.

Bibliografias

AMANN, R.I.; LUDWIG, W.; SCHLEIFER, K. Phylogenetic Identification and In Situ Detection of Individual Microbial Cells without Cultivation. Microbiological Reviews, Mar. 1995; p. 143–169 Vol. 59, No. 1 0146-0749/95/$04.0010.

BANIK, J.J. and BRADY, S.F. Recent application of metagenomic approaches toward the discovery of antimicrobials and other bioactive small molecules. Current Opinion in Microbiology, 2010.13(5): p. 603-609.

BURKE, C., STEINBERG, P., RUSCH, D. KJELLEBERG, S. and THOMAS, T. Bacterial community assembly based on functionalgenes rather than species.PNAS, July 14, 2011, 1-

CAPORASO, J.G., et al., QIIME allows analysis of high-throughput community sequencing data. Nat Methods, 2010.7(5): p. 335-6.

FIERER, N. and JACKSON, R.B. The diversity and biogeography of soil bacterial communities. Proc Natl Acad Sci U S A, 2006. 103(3): p. 626-31.

HEAD, I.M., SAUNDERS, J.R. and PICKUP,R.W. Microbial evolution, diversity, and ecology: A decade of ribosomal RNA analysis of uncultivated microorganisms. Microb. Ecol., 1998. 35(1): p. 1-21.

LEY, R.E. et al., Evolution of mammals and their gut microbes.Science, 2008.320(5883): p. 1647-51.

LORENZ, P. and ECK,J. Metagenomics and industrial applications. Nat Rev Microbiol, 2005. 3(6): p. 510-6.

MILLION, M; LAGIER, G.C.; YAHAV, D.; PAUL, M. Gut bacterial microbiota and obesity. Clin Microbiol Infect march 2013; 19: 305–313.

NEALSON, K.H. and VENTER, J.C. Metagenomics and the global ocean survey: what's in it for us, and why should we care? The ISME Journal (2007) 1, 185–187.

PACE, N.R., A molecular view of microbial diversity and the biosphere. Science, 1997. 276(5313): p. 734-40.

QUAIL, M.; SMITH, M.; COUPLAND, P.; OTTO, T. D.; HARRIS, S.R.; CONNOR, T.R; BERTONI, A.; SWERDLOW, H.P.; GU, Y. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 2012, 13:341.

ROSS, E.M., MOATE, P.J., BATH, C.R., DAVIDSON, S.E., SAWBRIDGE, T.I., GUTHRIDGE, K.M., COCKS, B.G. and HAYES, B.J. High throughput whole rumen metagenome profiling using untargeted massively parallel sequencing. BMC Genetics 2012, 13:53

TRINGE, S.G. and HUGENHOLTZ,P.A renaissance for the pioneering 16S rRNA gene. Current Opinion in Microbiology, 2008.11(5): p. 442-6.

TRINGE, S.G., VON MERING, C., KOBAYASHI, A., SALAMOV, A.A., CHEN, K., CHANG, H.W., PODAR, M., SHORT, J.M., MATHUR, E.J., DETTER, J.C., BORK, P., HUGENHOLTZ, P. and RUBIN, E.M. Comparative metagenomics of microbial communities.Science. 2005 Apr 22; 308(5721): 554-7.

TURNBAUGH, P.J.; LEY, R.E.; MAHOWALD, M.A.; MAGRINI, V.; MARDIS, E.R.; GORDON, J.I. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature, 1027-1031 (21 December 2006) | doi:10.1038.

VENTER, J.C., et al., Environmental genome shotgun sequencing of the Sargasso Sea.Science, 2004.304(5667): p. 66-74.17.

WILMES, P.; BOND, P.L. Metaproteomics: studying functional gene expression in microbial ecosystems. Trends Microbiol. 2006 Feb;14(2):92-7.

13

Genômica e biologia de sistemas

Diego Bonatto
Helder Takashi Imoto Nakaya

Introdução

O contínuo avanço da Biologia Molecular revolucionou o modo como estudamos o funcionamento de um organismo, nas suas mais diferentes formas e condições fisiológicas, de uma maneira inimaginada até algumas décadas atrás. A compreensão dos fenômenos biológicos em uma escala molecular e a aplicação direta destes conceitos como ferramentas práticas para a dissecação dos diferentes componentes bioquímicos, elevou as Ciências Biológicas a um patamar próximo ao das Ciências Exatas em termos de formalismo e de experimentação. Os organismos vivos passaram a ser vistos como sistemas altamente complexos e organizados, compostos por milhares de unidades de informações, que incluem genes, proteínas e metabólitos.

Em termos práticos, a compreensão de como um organismo funciona e como este responde ao seu ambiente ocorreu de forma gradual ao longo da história das Ciências Biológicas. Esta busca gradual se valeu de métodos denominados reducionistas, onde pequenas partes de um todo maior (como algumas proteínas ou algumas sequencias de ácidos nucléicos, por exemplo) são isoladas e estudadas de forma sistemática ao longo de muito tempo. De certa forma, o reducionismo ainda é fundamental para que possamos compreender como os componentes menores ou individuais são formados, para que possamos determinar a sua composição físico-química e como estes estão inseridos dentro de um contexto maior, seja uma organela, uma célula ou mesmo um organismo como um todo.

Entretanto, está cada vez mais claro que um ser vivo é muito mais do que a simples soma ou junção das suas partes. Este tipo de definição, quando o todo observado é maior do que apenas a junção dos seus componentes, é conhecido como "propriedades emergentes". Estas propriedades emergentes só podem ser explicadas quando o organismo é analisado por completo, com cada um dos seus elementos precisamente identificados em um contexto mais amplo. Por exemplo, só podemos definir o cérebro e o seu funcionamento quando olhamos para todas as suas partes integradas, algo que não é possível pelo simples reducionismo de seus componentes.

É interessante notar que, a partir do final do século XX e início do século XXI, as ferramentas de análises moleculares tornaram-se cada vez mais amplas em suas análises, possibilitando o estudo dos organismos de uma forma abrangente. Por exemplo, a introdução dos microarranjos de DNA, da espectrometria de massa e dos equipamentos de sequenciamento de DNA automatizados no final do século XX geraram as chamadas "Ciências Ômicas". Denominadas assim por abranger o estudo do genoma (a sequência completa de DNA de um organismo), transcriptoma (o conjunto de transcritos de RNA), proteoma (o conjunto de proteínas), dentre outros (Figura 1).

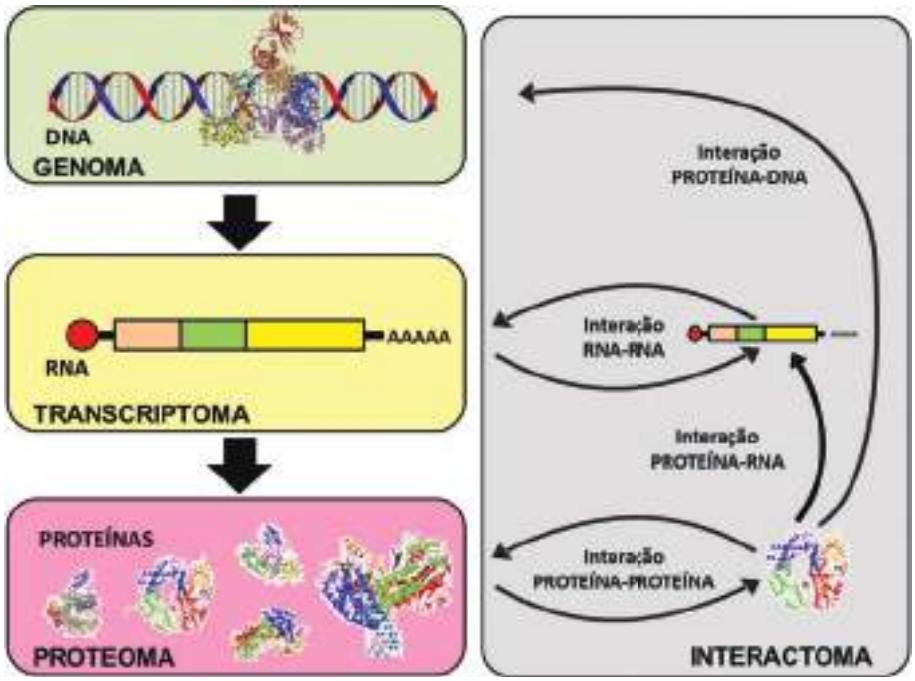


Figura 1. Principais "Ciências Ômicas". Graças aos avanços significativos de áreas como nanotecnologia, robótica, processamento de dados e óptica, podemos estudar os milhares de componentes de um sistema biológico em diferentes níveis. O conjunto de todos os genes (ou a sequência completa do DNA) faz parte do genoma de um organismo. Os conjuntos de todas as moléculas de RNA e de todas as proteínas são chamados, respectivamente de transcritoma e proteoma. O estudo da interação entre esses diferentes componentes é chamado de interactoma.

Cada ciência "ômica" visa analisar ou comparar a estrutura e o comportamento de seus milhares de componentes em diferentes condições fisiológicas ou em diferentes espécies. Um aspecto interessante dos projetos de pesquisas "ômicos" foi o fato de que estes têm mostrado que uma função biológica é a conseqüência de uma série de interações complexas que ocorrem entre as inúmeras moléculas encontradas em uma célula. Por exemplo, várias doenças degenerativas ou de caráter plurimetabólico, como a Doença de Parkinson e o diabetes melito, são o resultado final da interação global entre os componentes moleculares e fatores ambientais, caracterizando a sua propriedade emergente.

Desta maneira, considerando os sistemas biológicos como sendo complexos e emergentes, e somado ao fato de que os dados obtidos a partir de experimentos de larga escala ou ômicos necessitam ser interpretados de uma forma simples e mais completa, torna-se necessário a busca de ferramentas que possibilitem a visualização dos princípios biológicos básicos que regem um determinado organismo. Estas ferramentas devem ter uma precisão adequada para que os componentes individuais sejam estudados e, ao mesmo tempo, possibilitem gerar uma visão ampla da interação entre estes componentes. Deve-se fortemente salientar que o desenvolvimento das técnicas de análise em larga escala de processos biológicos permitiu a obtenção de uma vasta quantidade de informações a respeito do comportamento celular em determinadas condições fisiológicas. Neste sentido, estas técnicas de análise em larga escala também permitiram determinar como e quando as moléculas biológicas interagem entre si para gerar uma resposta fisiológica do organismo.

Definindo a biologia de sistemas

O preenchimento desta lacuna analítica foi resolvido com o desenvolvimento de um campo de pesquisas que integra diferentes áreas do conhecimento científico conhecida como Biologia de Sistemas. Desta maneira, a Biologia de Sistemas busca observar, de uma forma bastante ampla, as relações entre os componentes de um sistema biológico e dos seus respectivos processos, enfocando na geração de hipóteses que possam ser testadas experimentalmente. Devido à complexidade e a enorme quantidade de informações geradas pelos projetos "ômicos", um dos principais desafios da Biologia de Sistemas é integrar todas essas informações a fim de mapear, entender e modelar, em termos quantitativos, o comportamento celular ou fisiológico de um organismo.

Da mesma forma, a Biologia de Sistemas pode ser vista como um processo interativo e sempre continuo. Interativo e continuo porque uma vez obtido o modelo biológico ou a descrição de um mecanismo biológico, torna-se fundamental aperfeiçoá-lo por meio da comparação constante do modelo biológico com os dados experimentais e também pela integração de características biológicas geradas pelas técnicas experimentais mais pontuais ou reducionistas.

Assim, movido pela contínua busca em aperfeiçoar os modelos biológicos gerados e estudados pela Biologia de Sistemas, há uma necessidade inerente em obter ferramentas computacionais mais refinadas que possam realizar este aperfeiçoamento. Dentre as ferramentas computacionais disponíveis, os bancos de dados constituem a fonte primordial de dados biológicos computáveis que capturam a informação a

respeito das interações funcionais relacionadas as principais macromoléculas, como proteínas, DNA e RNA.

Apesar do crescimento dramático dos bancos de dados de informações biológicas nos últimos anos, a sua exploração no sentido de gerar informações úteis enfrenta um importante desafio computacional. A fim de suplantar este desafio, uma das principais formas computacionais utilizadas para integrar todas estas informações está presente na Teoria dos Grafos, um ramo da matemática que lida com os fundamentos das chamadas redes de integração. As redes de integração são estruturas complexas, formadas por elementos únicos (nós ou vértices) ligados por conectores, e que permitem descrever uma ampla variedade de sistemas, sejam esses biológicos, tecnológicos ou sociais. Por exemplo, a rede mundial de computadores é uma complexa rede de roteadores e computadores (nós; Figura 2) unidos por ligações físicas (conectores; Figura 2), onde a internet é uma enorme rede virtual de páginas (nós) conectadas por *hyperlinks* (conectores). Do mesmo modo, uma via metabólica pode ser descrita como uma complexa rede de substâncias químicas (nós; Figura 2) conectadas por reações químicas (conectores; Figura 2). Esses sistemas representam apenas alguns dos inúmeros exemplos de redes de integração que têm atraído a atenção de pesquisadores para o seu estudo.

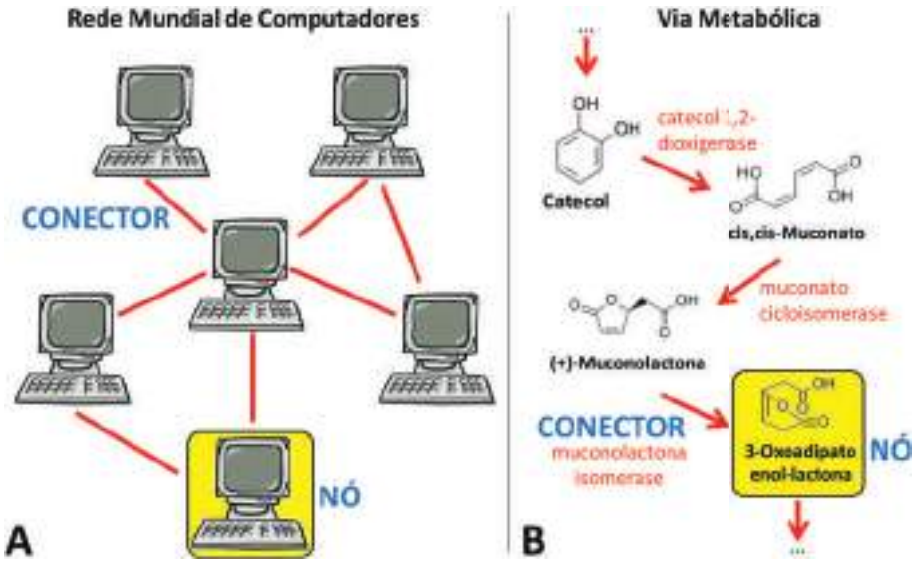


Figura 2. Exemplos de Redes de Integração. (A) Na rede mundial de computadores, cada conector representa uma ligação física entre dois computadores (nós). (B) Em uma via metabólica, os conectores podem ser representados pelas reações químicas promovidas pelas enzimas e os nós por seus substratos e produtos.

Além dos conceitos básicos associados a redes de integração entre elementos biológicos, veremos como estas ferramentas estão ampliando o nosso conhecimento na forma como os componentes biológicos interagem entre si. Por fim, observaremos que, durante a evolução dos organismos, estes componentes biológicos organizaram-se

de forma a compor sistemas altamente integrados e com funções bastante específicas. Cada um destes sistemas altamente integrados pode ser assim definido como um sistema modular, em uma analogia bastante precisa com o termo "módulo" usado pelas engenharias. Cada módulo tem a capacidade de integrar-se a outro módulo, onde o conjunto de módulos define as inúmeras vias bioquímicas e interações gênicas que compõem uma célula.

Histórico da biologia de sistemas

Como colocado no início deste capítulo, a Biologia de Sistemas emergiu como uma conseqüência dos projetos de larga escala (ou "ômicos"), sendo considerada como um conjunto de análises matemáticas que visam complementar ou mesmo substituir a visão tradicional de pesquisa biológica, em sua maioria restrita a um nível de complexidade. Mas o que significa o termo "nível de complexidade" para Sistemas Biológicos? De uma forma geral, os "sistemas" ou "níveis de complexidade" da Biologia de Sistemas podem ser categorizados como a interação entre duas moléculas necessárias para realizar uma determinada função na célula, seguida das interações observadas entre todos os componentes celulares que compõem um tecido até as interações vistas entre tecidos e órgãos que compõe um organismo (Figura 3). A Biologia de Sistemas, desta maneira, é integrativa e busca entender e predizer o comportamento das propriedades "emergentes" de processos biológicos complexos e formados por múltiplos elementos.

Um nível sistêmico ou de complexidade de um processo biológico tenta responder três questões principais. Primeiramente, quais são as partes componentes do sistema (para as células, quais são os genes, proteínas e metabólitos) que estão envolvidas? Segundo, como as partes funcionam e em que condições? E terceiro, como as partes trabalham e interagem para atingirem um determinado objetivo em resposta a um estímulo interno do organismo ou do meio ambiente? Neste sentido, a visão tradicional da biologia, como colocado anteriormente, está preocupada em enumerar e caracterizar todos os blocos construtores dos sistemas vivos, o que é fundamental para responder a primeira questão, mas também essencial para a segunda e para a terceira questão relacionadas aos níveis de complexidade de um sistema. Por exemplo, certas proteínas possuem a capacidade de ligarem-se a regiões particulares do DNA, ativando ou inibindo a transcrição do RNA mensageiro que, posteriormente, será traduzido para proteínas. Estas mesmas proteínas que controlam a transcrição do DNA, em sua grande maioria, também possuem a capacidade de ligarem-se a outras proteínas, formando complexos que nenhuma das proteínas constituintes individuais poderia fazer. Seguindo esta linha de raciocínio e considerando as informações "ômicas" disponíveis, existem milhares (se não milhões) de diferentes tipos e estados funcionais de proteínas em um organismo vivo, de forma que o número de interações entre estas é gigantesco.

É importante ressaltar que, até pouco tempo atrás, as análises de sistemas complexos não eram parte da corrente principal das biociências, e a maioria das tecnologias requeridas para a análise de organismos em nível sistêmico não estava disponível. As equações matemáticas não estavam presentes nas principais publicações

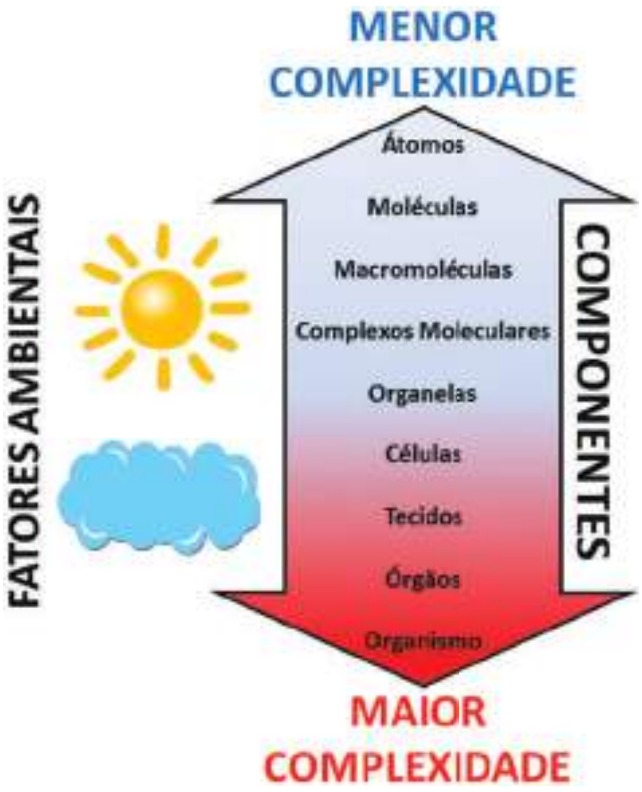


Figura 3. Níveis de Complexidade de um Sistema Biológico. Um sistema biológico pode ser estudado em vários "níveis de complexidade", desde interações entre componentes simples como átomos até componentes complexos como órgãos. Componentes de um nível de complexidade afetam diretamente componentes de outros níveis e todos podem ser afetados por fatores ambientais.

da área da biologia molecular e a biologia experimental desconhecia em larga escala a literatura matemática. Também no passado recente, outras disciplinas científicas, de natureza não biológica, desenvolveram os conhecimentos na área de sistemas que foram fundamentais para o estabelecimento destas dentro da Ciência. Alguns exemplos destas disciplinas podem ser citados pelo seu impacto no mundo atual e pela sua aplicação cada vez mais crescente: (i) a teoria dos sistemas dinâmicos e não lineares (caos e bifurcação) na matemática, (ii) o entendimento de fenômenos estocásticos e da auto-organização usando os conceitos de desequilíbrio termodinâmico e de física estatística; (iii) o uso da biofísica e da bioquímica na compreensão da cinética enzimática cooperativa e (iv) o uso da engenharia para o desenvolvimento da teoria do controle, sistemas de identificação e engenharia metabólica. Assim, as perspectivas de sistemas usadas pela Biologia estavam limitadas, em sua grande maioria, à ecologia.

Atualmente, e como veremos a seguir, a Biologia de Sistemas começou a permear vários campos das Ciências Biológicas. Infelizmente, esta área ainda não está tão conectada aos campos experimentais da biologia que estão passando por uma grande expansão, como a Biologia Molecular, devido, em grande parte, à falta de

dados moleculares quantitativos e o pouco conhecimento matemático por parte dos pesquisadores. Notáveis exceções são feitas a esta "regra", como a análise do controle metabólico, a teoria de sistemas bioquímicos, a modelagem cinética (de circuitos metabólicos e genéticos) e a mecânica de desequilíbrio termodinâmico onde, junto com a genômica, a bioinformática e a engenharia metabólica, são consideradas como as disciplinas fundadoras da Biologia de Sistemas.

Fluxo de informações nos sistemas biológicos

Um completo entendimento de qualquer um dos sistemas biológicos conhecidos requer uma quantidade de dados muito maior do que as atuais tecnologias podem oferecer. Intuitivamente, podemos imaginar que é possível construir um modelo de um sistema conhecendo todas as moléculas envolvidas, as suas concentrações, como interagem entre si, os efeitos de cada parte individual em seus vizinhos e parâmetros dinâmicos como a concentração, interações e mudanças mecanísticas com o tempo. Entretanto, esta visão não é real considerando o fato de que as tecnologias necessárias para medir muitos destes parâmetros ainda não saíram do papel, se é que existem. Assim, é necessário conhecer todos os detalhes de um processo para sermos capazes de desenvolver modelos de sistemas biológicos que sejam úteis ou preditivos? As análises de mapas de interações protéicas ou entre proteínas e DNA sugerem que, mesmo sendo os dados esparsos ou pouco conhecidos, é possível construir modelos de redes biológicas iniciais ou mesmo rudimentares. Um exemplo interessante pode ser dado considerando as chamadas "vias metabólicas", muito usadas para estudos em bioquímica e biologia molecular. Uma via metabólica é geralmente representada por substratos e produtos metabólicos conectados entre si por reações químicas mediadas por enzimas, que por sua vez podem ser reguladas por outras proteínas da mesma ou de outra via. Os exemplos de várias vias metabólicas incluem a glicólise, o ciclo do ácido cítrico, a via da pentose fosfato, entre outros. Em termos simplificados, a construção de uma via metabólica envolve a coleção de todas as informações metabólicas relevantes de um organismo e a sua compilação em uma rede que faça sentido para os vários tipos de análise que serão feitas com base na informação obtida. Neste caso, é importante que haja uma correlação entre as infomações ômicas e os processos metabólicos conhecidos, o que pode ser feito por meio do uso de banco de dados especializados, como o KEGG (http://www.genome.jp/kegg/). Assim, apesar de rudimentares, as redes metabólicas são ferramentas poderosas para o estudo e a modelagem do metabolismo. As redes de regulação genética ou gênica são outros bons exemplos a serem considerado. Uma rede de regulação genética descreve a regulação da expressão gênica, ou seja, a produção de proteínas a partir do genoma pela transcrição e pela tradução. A expressão de um gene pode ser controlada pela presença de outras proteínas ativadoras ou repressoras que, por sua vez, formam uma rede onde os nós representam as proteínas e os conectores indicam uma dependência da síntese de uma proteína particular por meio de outras proteínas. A expressão dos mRNAs também pode ser regulada por moléculas de RNAs não codificadores (principalmente os micro-RNAs), aumentando ainda mais a complexidade do sistema. Em outras palavras, as redes de regulação genética são como interruptores elétricos (ligado-

e-desligado) e reostatos de uma célula operando em nível gênico. Eles orquestram dinamicamente os níveis de expressão para cada gene presente no genoma. Cada transcrito de RNA funciona como molde para a síntese de uma proteína específica pelo processo da tradução. Da mesma forma, as redes transcricionais (regulatórias) bacterianas mostram as relações entre os fatores de transcrição e os operons, grupos de genes contíguos que são transcritos em uma única molécula de mRNA, que eles regulam. Nestas redes, cada nó representa um operon e os conectores representam as interações transcricionais diretas.

Finalmente, as redes de interações entre proteínas (ou também conhecido como interatoma) são compostas por nós que representam as proteínas e onde cada conector indica uma interação. Assim, uma rede incluindo todas as proteínas de um organismo e todas as suas possíveis interações podem ser chamadas de uma rede proteômica. As interações nestas redes são importantes para a maioria dos processos biológicos, considerando que muitas destas proteínas necessitam interagir com outras para exercerem as suas funções de forma adequada.

Tipos de biologia de sistemas

Levando em conta os diferentes tipos de redes biológicas, a Biologia de Sistemas faz uso de duas linhas de raciocínio que, conforme veremos a seguir, podem ser aplicadas para a geração de modelos biológicos úteis: a chamada Biologia de Sistemas top-down (uma expressão em inglês que significa do "mais complexo ao menos complexo" ou "de cima-para-baixo") e a Biologia de Sistemas bottom-up (do "menos complexo ao mais complexo" ou "de baixo-para-cima"). Cada um dos ramos possui pontos fortes e fracos, como veremos a seguir, e caberá ao pesquisador escolher aquele que melhor caracteriza o seu sistema. É importante ressaltar que as duas visões da Biologia de Sistemas não são excludentes.

Inicialmente, e com a introdução das ferramentas "ômicas", o método top-down emergiu como a ferramenta dominante, onde o sistema biológico é visto na sua totalidade (obtido pelos dados de larga escala) a fim de descobrir e caracterizar mecanismos biológicos mais simples, ou seja, as suas partes e as suas interações (Figura 4). Neste tipo de método, o objetivo principal é desdobrar novos mecanismos moleculares usando um ciclo interativo, como vimos no início deste capítulo: (i) obtenção de dados experimentais, (ii) análise e integração destes dados para determinar as correlações entre as partes do sistema e (iii) a formulação de novas hipóteses experimentais, as quais predizem novas correlações e permitem um novo conjunto de experimentos. A principal característica da Biologia de Sistemas top-down está na sua visão geral do sistema. E, por isso, seus maiores desafios estão na compreensão de bancos de dados experimentais obtidos a partir de organismos modelos submetidos a poucas "perturbações".

E o que são "perturbações" para um biocientista experimental? De uma forma geral, estas "perturbações" são definidas como todas as alterações de natureza genética (como as obtidas a partir de uma ou mais mutações ou mesmo superexpressão de proteínas), ambientais (mudanças na concentração de nutrientes, em fatores de crescimento ou níveis de estresse), induzidas por RNA de interferência, da natureza

Figura 4. Método Top-down. Este método inicia-se com a obtenção de dados experimentais de larga escala ("ciências ômicas"), seguido pela análise e integração destes dados que por sua vez irão gerar novas hipóteses. Estas hipóteses podem ser testadas experimentalmente através de perturbações no modelo e nova coleta e análise de dados.

intrínseca da dinâmica celular (como o ciclo celular) ou pela administração de drogas que modificam o comportamento fisiológico de um sistema. Desta maneira, estes estudos visam obter uma visão geral do comportamento de um sistema submetido a estas perturbações tendo como objetivo a descoberta de padrões comportamentais genéricos o suficiente para realizar predições biológicas. Os modelos que usam a Biologia de Sistemas top-down são fenomenológicos, ou seja, eles não são baseados em mecanismos e, em sua grande maioria, não necessitam empregar o conhecimento a respeito da relação entre os componentes moleculares. Este tipo de Biologia de Sistemas é usado, principalmente, para a caracterização de subsistemas celulares que ainda não foram elucidados e com detalhes que ainda precisam ser descobertos (como a elucidação das interações entre proteínas ou redes genéticas). Assim, a integração dos estudos "ômicos" que, simultaneamente, analisa os dados transcriptômicos, proteômicos e fluxômicos (área que estuda o fluxo de metabólitos em um sistema), pode ajudar a Biologia de Sistemas top-down por meio da chamada "genômica vertical". A genômica vertical busca traçar, quantitativamente, as mudanças na taxa de um processo biológico, por meio da mudança nos níveis de substrato, produtos,

efetores, enzimas e de mRNAs. Tal desenvolvimento oferece uma nova forma de estudo para a biologia celular e permite visualizar a extensão de várias vias regulatórias presentes internamente à rede e que podem contribuir para um determinado comportamento em particular.

Diferente da Biologia de Sistemas top-down, a Biologia de Sistemas bottom-up busca deduzir as propriedades funcionais que podem emergir de um subsistema que foi profundamente caracterizado em termos mecanísticos por meio de técnicas moleculares, muitas vezes reducionistas (Figura 5). A Biologia de Sistemas bottom-up inicia-se das partes constitutivas de um sistema por meio da simulação dos comportamentos interativos (equações de taxas) de cada processo (processos enzimáticos). Então, integra estas formulações para predizer o comportamento do sistema (Figura 5). Apesar de parecer uma definição bastante complicada para os iniciantes na Biologia de Sistemas, na verdade, a Biologia de Sistemas bottom-up tem como objetivo principal combinar estes diferentes modelos de vias enzimáticas (juntamente com taxas de fluxo de metabólitos, constantes enzimáticas e outros parâmetros bioquímicos) em um único modelo para o sistema como um todo. Tais modelos possibilitam a incorporação de mais processos em um nível de detalhe mecanístico elevado.

Exemplos deste método, obtidos neste presente momento, incluem uma série de trabalhos associados de modelagem e de obtenção de dados experimentais como a caracterização da rede de sinalização ativada pelo receptor de fator de crescimento dérmico e a modelagem do metabolismo central de carbono de *Escherichia coli* e *Trypanosome brucei*. Assim, a Biologia de Sistemas bottom-up está assentada: (i) nos estudos experimentais que determinam as propriedades cinéticas e fisico-químicas dos componentes (cinética enzimática, propriedades de difusão, dentre outras), sejam estes pelo estudo de enzimas isoladas ou pelo uso de estratégias de estimação de parâmetros; (ii) na obtenção de dados referentes as respostas dos subsistemas as perturbações, enquanto contexto celular; (iii) a construção de modelos detalhados para o cálculo de dados obtidos a partir do item (ii) e para melhorar o desenho experimental e (iv) o desenvolvimento de ferramentas para a análise e a representação de modelos (Figura 5). Infelizmente, a determinação dos parâmetros cinéticos usando as estratégias de estimação de parâmetros possui um risco inerente devido à simplificação da análise. Assim, considerando que os modelos cinéticos são simplificados, as estratégias aplicadas para a estimação destes parâmetros cinéticos talvez não mostrem os "verdadeiros" parâmetros cinéticos. Na medida em que os modelos são acrescidos de dados e estes são usados para a estimação dos parâmetros, cada um destes poderá assumir valores diferentes. Adicionalmente, as falhas em nosso conhecimento a respeito destes parâmetros cinéticos podem ser suprimidas de forma errônea assim que nós as colocamos no modelo. Por outro lado, se podemos incorrer a erros ao medir os parâmetros cinéticos de um sistema, em alguns casos é extremamente difícil, se não impossível, medir os parâmetros cinéticos *in vitro*, como é o caso da sinalização celular e da expressão gênica eucariótica. Atualmente, existe a possibilidade de contornar este problema por meio do desenvolvimento de tecnologias que medem a cinética enzimática *in vivo*. Tal ferramenta está se tornando cada vez mais disponível através do aperfeiçoamento de técnicas de imagens de alta

resolução usando fluorescência ou outros métodos que permitam a observação do fluxo de metabólitos dentro de um sistema ou subsistema celular.

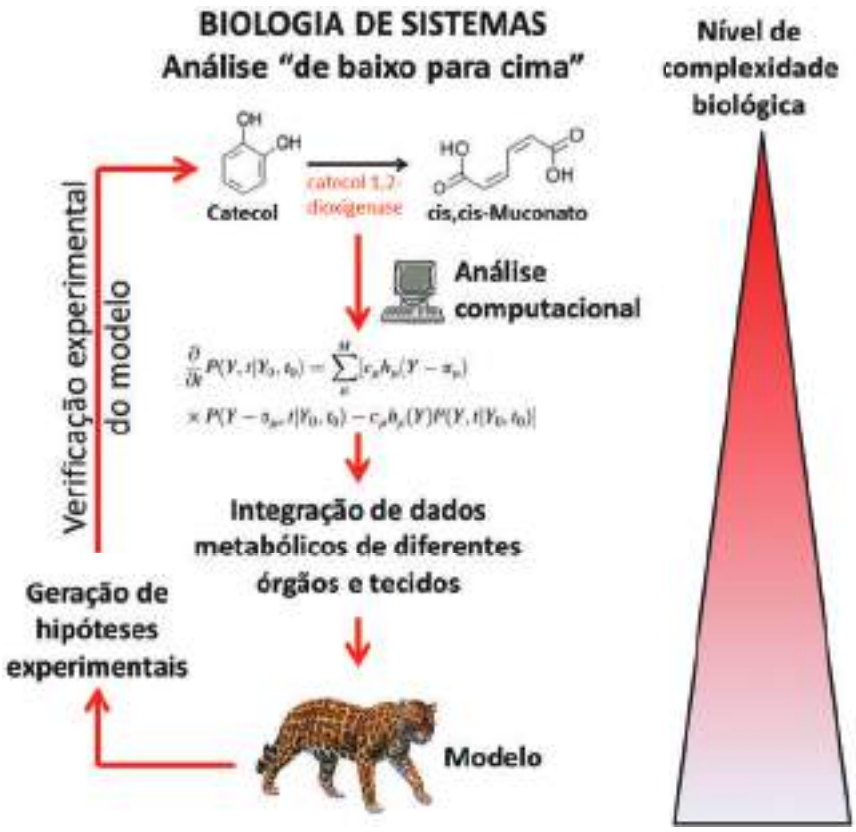


Figura 5. Análise Bottom-up. Este método parte de processos já caracterizados experimentalmente (reações enzimáticas, regulação gênicas, interações proteicas, etc), que são depois integrados e modelados computacionalmente em diferentes contextos biológicos. As hipóteses geradas nessas análises, quando verificadas experimentalmente, fornecem uma visão mais ampla dos mecanismos que regem os sistemas biológicos.

Fundamentos teóricos da biologia de sistemas

Como vimos até o momento, a compreensão da Biologia de Sistemas está embasada na linguagem ou, usando um termo mais técnico, na "abstração" matemática. Neste caso, a chamada Teoria dos Grafos oferece a abstração matemática necessária para a descrição dos diferentes níveis de complexidade de um organismo. A beleza e a utilidade desta linguagem permitem o desenvolvimento de conceitos e ferramentas para uma aplicação real, um conceito que é pó uco familiar para nós, biocientistas, mas que já provou a sua essencialidade em outras áreas do conhecimento humano. Por exemplo, muitos cientistas e engenheiros estão familiarizados com os benefícios da abstração que reside na álgebra linear e na teoria do cálculo ou da probabilidade.

Afinal, a construção de uma casa ou de um prédio requer um conjunto de ferramentas matemáticas que prediga como estas habitações irão se comportar e, principalmente, que possibilite afirmar como as mesmas resistirão as diferentes condições ambientais a que estarão sujeitas por muito tempo, como variações de temperatura, de umidade, poluição, dentre outros fatores.

No que consiste a Teoria dos Grafos? Como o próprio nome diz e como foi descrito no início deste capítulo, um grafo nada mais é do que um conjunto de nós e linhas que conectam os nós, compondo a base da Biologia de Sistemas que é uma rede de interações. Assim, os nós são as entidades de interesse e as linhas (ou conectores) representam as relações entre estas entidades. Por exemplo, as entidades podem ser um conjunto de proteínas de uma célula e as relações são modeladas pela existência de uma interação física entre duas proteínas quaisquer. Matematicamente, um grafo é especificado por um conjunto de nós "V" e um conjunto de conectores "E". Cada elemento de E contém um par de elementos de V. Cada conector pode ter pesos, direções ou ser formado por diferentes tipos de dados experimentais (Figura 6). Algumas vezes, formas especializadas de grafos, como multigrafos, grafos bipartidos e hipergrafos são úteis.

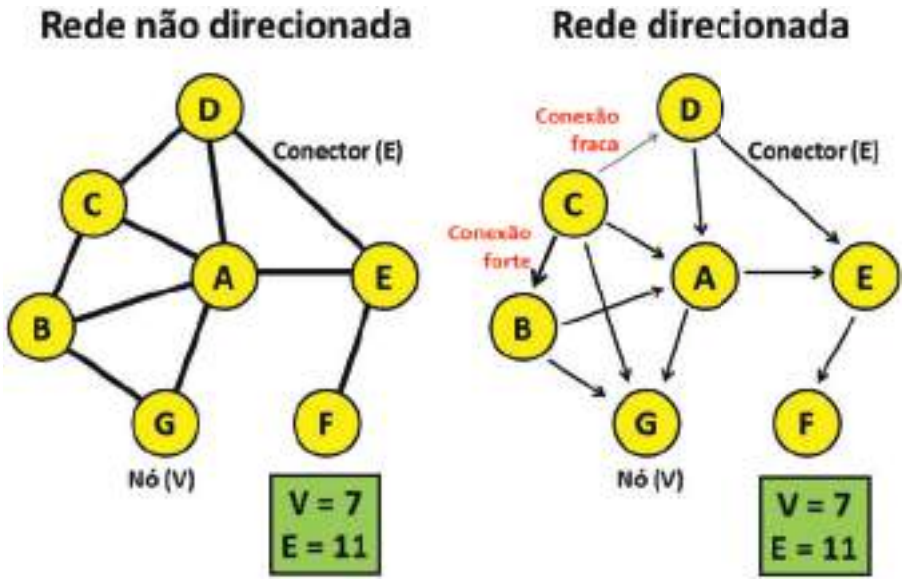


Figura 6. Principais Tipos de Grafos. O grafo da esquerda representa uma "rede não direcionada". A informação contida em E apenas revela a existência de uma conexão entre 2 nós (V). Em uma "rede direcionada" (grafo da direita), os conectores E informam não apenas a existência da conexão entre 2 nós, mas também a direção (sentido da flecha) e a força (espessura da linha) de cada conexão.

Os termos "grafo" e "rede" serão usados como sinônimos neste capítulo. Entretanto, o termo grafo enfatiza o conceito matemático, enquanto que o termo rede está associado à funcionalidade de um sistema. Os grafos possuem papéis em três grandes áreas complementares. Primeiro, os grafos geram uma estrutura de dados para a representação do conhecimento obtido a partir de dados "ômicos". Como

exemplos, podemos incluir os mecanismos regulatórios, a transdução de sinais ou as redes metabólicas. Uma segunda aplicação dos grafos na biologia molecular é para a geração de modelos onde é possível mensurar os dados biológicos. Por exemplo, em um ensaio dois-híbridos de leveduras, os dados são gerados a partir da observação de um possível par de proteínas trabalhando juntas para criar um complexo de início de transcrição. Em um experimento de microarranjos de imunoprecipitação de cromatina (chromatin immuno-precipitation microarray ou ChIP-chip), os dados referem-se à força de ligação de uma proteína a uma determinada região ou sequência de DNA o qual, por si própria, pode estar ligada a um ou vários genes que também são regulados. E, por fim, os grafos possuem um papel importante na modelagem estatística. Por exemplo, é possível construir um modelo que descreva quais conjuntos de proteínas podem ser observadas formando um complexo protéico, considerando os dados de interações entre proteínas a partir de experimentos de co-precipitação. Neste caso, diferentes modelos estatísticos podem ser aplicados para a geração de hipóteses testáveis. Outro exemplo se refere ao fato de que as proteínas interatoras também podem ser co-reguladas transcricionalmente. Neste caso, a resposta pode ser obtida olhando-se para as respectivas redes, para os grafos de interações entre proteínas e para os grafos de co-expressão, e verificando se a sobreposição destes grafos é muito maior do que aquela que poderia ser esperada pelo acaso.

Assim, as redes biológicas oferecem uma descrição quantificável das redes que caracterizam os vários sistemas biológicos. Neste sentido, é importante definir as medidas básicas das redes que permitem a comparação e a caracterização de diferentes redes biológicas, que são o grau ou conectividade, a distribuição dos graus, os expoentes do grau, o caminho mais curto e o caminho médio e, por fim, o coeficiente de agrupamento. Existem outras medidas que são utilizadas para a análise de grafos, mas as citadas acima são as mais utilizadas e geram informações importantes a respeito do sistema em que o pesquisador está trabalhando. Iniciaremos pela definição do grau e, posteriormente, das outras medidas básicas.

Propriedades matemáticas das redes de interações

Matematicamente, o grau de um nó é definido pela letra k, o qual diz quantas ligações ou conectores este nó possui com os outros nós. Neste caso, é importante salientar que os grafos podem ser considerados de dois tipos: não direcionados ou direcionados. No primeiro tipo, os conectores não apresentam uma ordem específica de entrada ou saída (Figura 6) enquanto que, no segundo caso, os nós podem ter conectores que entram e que saem deste (Figura 6). Quando existe uma direção específica o número de conectores que entram no nó é definido como kentrada, enquanto que o número de conectores que saem do nó é definido como ksaída (Figura 6). Assim, um exemplo prático de redes direcionadas e não direcionadas pode ser considerado na Figura 6 descrita abaixo. Na primeira parte da figura, que descreve uma rede direcionada, o nó "A" possui kentrada=3 e um ksaída=2 (representados pelas flechas escuras; Figura 6). Por sua vez, para a rede não direcionada, descrita na segunda parte da figura, o nó "A" possui um grau k=5 (Figura 6). Um outro aspecto importante das redes não direcionadas é a possibilidade de calcular o seu grau "médio" ou <k> (onde <>, na

linguagem matemática, denota a média), considerando que a mesma apresente N nós e L conectores. Para tanto, a fórmula <k> = 2L/N pode ser aplicada de uma forma simples e que dá uma idéia do número médio de conectores por nó (mais adiante veremos que este conceito será importante para a definição de módulos).

Outra propriedade importante das redes é a distribuição de graus, ou P(k). Neste caso, P(k) pode ser obtido por meio da contagem do número de nós, ou N(k), que possuam um k=1, 2, 3,... seguido pela divisão do número total de nós N. Esta probabilidade de distribuição é importante pois permite distinguir as diferentes classes de redes. Por exemplo, a Figura 7 mostra uma típica rede de interação biológica (neste caso, uma rede de interação entre proteínas humanas baseada na proteína fosfofrutocinase hepática ou PFKL) que apresenta o que é definido por uma distribuição de nós seguindo uma lei de potenciação, onde alguns poucos nós conectam inúmeros nós pequenos, com poucos conectores. Todos os nós que apresentam esta característica, de se apresentarem altamente conectados com outros nós pouco conectados, são denominados de *hubs*, um termo emprestado da computação que define os centros de controle ou de distribuição de dados em uma rede qualquer. Por outro lado, este exemplo também é útil para definirmos a terceira propriedade matemática das redes, conhecida pelo expoente do grau. A maioria das redes biológicas são livres de escala, o que significa que a distribuição dos graus segue uma lei de potenciação, ou $P(k) \sim k^{-\gamma}$, onde γ é o expoente do grau e "~" indica "proporcional a". Os valores de γ determinam as inúmeras propriedades de um sistema, onde quanto menor for o valor de γ, mais importante é o papel dos *hubs* na rede. Por exemplo, para valores de $\gamma > 3$, os *hubs* não são relevantes, onde $2 > \gamma > 3$ indica uma hierarquia de *hubs*, no qual o *hub* mais conectado está em contato com uma pequena fração de todos os nós. Esta propriedade matemática possibilita definir algumas das características mais importantes das redes livres de escala, como o grau de robustez de uma rede, que está associado com a resistência da rede a uma falha de um ou vários nós.

A quarta e a quinta propriedade matemática de uma rede é definida como a distância mais curta entre dois nós e o caminho médio da rede. Neste sentido, a distância em uma rede é medida por meio do tamanho de uma via, o qual diz respeito a quantos passos são necessários para viajar entre dois nós quaisquer. Como existem muitos passos ou caminhos alternativos entre dois nós, o caminho mais curto (ou o caminho com o menor número de passos entre dois nós) possui um significado especial. Em redes direcionadas, a distância entre um nó A e um nó D (ou lAD) pode ser diferente da distância entre o nó B e o A (lDA). Muitas vezes não existe um caminho direto entre dois nós. Por outro lado, o caminho médio (<l>) representa a média de todos os caminhos mais curtos entre todos os pares de nós e oferece uma medida da "navegabilidade" de uma rede.

A sexta propriedade matemática de uma rede está relacionada ao coeficiente de agrupamento. Neste caso, em muitas redes, se um nó A está conectado a B e este, por sua vez, está conectado a C, então existe a probabilidade de que A também esteja conectado com C. Este fenômeno pode ser quantificado aplicando o coeficiente de agrupamento CI=2nI/k(k-1), onde nI é igual ao número de conectores que ligam um número kI de nós vizinhos ao nó I. Em outras palavras, CI representa o número de triângulos que passam pelo nó I. Por sua vez, kI(kI-1)/2 é o número total de triângulos que podem passar pelo nó I considerando que todos os nós vizinhos de I estejam

conectados. Assim, o coeficiente de agrupamento médio ou <C> é caracterizado pela tendência geral dos nós em formar agrupamentos. Assim, uma medida importante da estrutura de uma rede é a função C(k), o qual é definido como a média do coeficiente de agrupamento de todos os nós com um número k de conectores. A Tabela 1 sumariza as propriedades matemáticas associadas a redes, conforme visto neste item.

Tabela 1. Principais propriedades matemáticas associadas a redes.

Propriedade	Representação	Função	Definição
Número de nós	N		Número de nós de uma rede
Número de conectores	L		Número de conectores de uma rede
Grau ou conectividade	K (para redes não direcionadas); kentrada e ksaída (para redes direcionadas)		Número de ligações ou conexões de um nó
Grau ou conectividade médio	<k> (para redes não direcionadas)	<k>=2L/N	Grau ou conectividade média de uma rede
Distribuição de graus ou conectividade	P(k)	P(k)=N(k)/ N	Probabilidade de que um nó N com k graus esteja presente na rede
Distância ou caminho mais curto entre dois nós	ℓAB		Número mínimo de passos para chegar ao nó B partindo do nó A
Caminho ou distância média	<ℓ>		Média do número de caminhos mais curtos entre todos os pares de nós de uma rede
Coeficiente de agrupamento de nós	C	CA=2nA/k(k-1)	Número de nós conectados entre si em relação a um nó central (neste caso, definido como A)
Média do coeficiente de agrupamento	C(k)		Média do coeficiente de agrupamento de todos os nós com um número k de conectores

O grau médio <k>, o caminho ou distância média <l> e a média do coeficiente de agrupamento <C> dependem do número de nós e conectores presentes em uma determinada rede, enquanto que as funções P (k) e C(k) são independentes do tamanho da rede. Isto é importante porque estas duas funções podem mostrar as características genéricas de uma rede, permitindo a classificação de vários tipos diferentes de redes (Tabela 1).

Um ponto essencial a respeito da abordagem matemática das redes está relacionado aos modelos gerados. Estes modelos incluem métodos de agrupamento e testes estatísticos desenvolvidos com sucesso para a análise de informações "ômicas". Apesar de os modelos variarem em complexidade e propósito, em todos os casos, seus

valores são julgados com base nos resultados experimentais obtidos. Entretanto, deve sempre ser salientado que todos os modelos atuais (inclusive as redes vistas até este momento) são estáticos em sua natureza e não consideram as informações bioquímicas e moleculares obtidas nestes últimos 50 anos de pesquisas em biociências.

Os modelos baseados em redes neurais e cadeias de Markov, por exemplo, são propostas avançadas para a análise de dados, e o uso de tais modelos possibilita o desenvolvimento de algoritmos com um certo poder preditivo. Apesar do valor destes modelos para a reconstrução de vias bioquímicas, todos são dependentes de dados biológicos obtidos por vias experimentais e não podem ser usados para simulações, como o efeito de diferentes condições ambientais no crescimento celular. Na maioria das vezes, modelos cinéticos são empregados para simular a resposta dinâmica de sistemas biológicos a diferentes estímulos ou diferentes perturbações do sistema. Estes modelos ainda são bastante raros nas ciências biológicas visto que dependem de uma enorme quantidade de informação proveniente de experimentos baseados em cinética enzimática ou protéica. A maioria destes modelos é construída utilizando o balanço de massas para os componentes individuais por meio de equações diferenciais. Além disto, a cinética das diferentes reações e processos são descritos por equações matemáticas bastante complexas. Alguns exemplos podem ser encontrados atualmente na literatura que caracterizam a Biologia de Sistemas bottom-up como visto anteriormente, mas os modelos apresentados até o presente momento são sistemas relativamente pequenos, pelo menos da perspectiva de uma célula como todo. Assim, estes modelos não são úteis para a integração de dados "ômicos" e são incapazes de descrever as funções gerais de uma célula. Contudo, espera-se que a longo prazo seja possível coletar todos os parâmetros cinéticos e definir as equações matemáticas para todas as reações e processos em uma célula, permitindo a montagem de um modelo cinético completo para a célula.

A questão da modularidade biológica

Uma conseqüência inesperada da Biologia de Sistemas foi a observação de que as redes de interações entre moléculas são formadas por inúmeras sub-redes distintas, também chamadas de módulos.

Um módulo é definido, na área da engenharia, como uma estrutura funcional e padronizada que pode ser unida a outros módulos iguais ou que possuam características diferentes para a construção de um objeto mais complexo. Aparentemente, grupos de genes, proteínas ou metabólitos são capazes de formarem módulos específicos, onde a união destes módulos constitui processos biológicos fundamentais para que uma célula consiga exercer uma determinada função. Estes processos biológicos, também denominados de ontologias gênicas, constituem um importante campo de pesquisas dentro da Biologia de Sistemas e são de extrema importância para a visualização do funcionamento de uma rede no contexto celular.

Assim, os módulos, como uma conseqüência matemática da Teoria dos Grafos, são vistos em inúmeros sistemas, como os círculos de amizades em redes sociais ou páginas da Internet que tratam de tópicos similares. E, como salientado anteriormente, uma estrutura altamente modular é um componente fundamental de sistemas de

engenharia complexos, como aviões modernos ou processadores de computadores. Seguindo esta linha de raciocínio, a biologia está repleta de exemplos de modularidade. Os complexos de proteína-proteína ou proteína-RNA que são estáveis e que variam muito pouco em termos de interação constituem o núcleo central de muitas funções básicas biológicas, desde a síntese de ácidos nucléicos até a degradação protéica. De um modo similar, os grupos de moléculas que são co-regulados temporalmente são conhecidos por atuarem em vários passos do ciclo celular, na transdução de sinais externos em bactérias para a quimiotaxia ou nas vias de resposta a feromônios em leveduras. De fato, a maioria das moléculas em uma célula faz parte de um complexo intracelular com atividade modular, como os ribossomos.

Desta maneira, para que seja possível estudar a natureza modular de uma rede, bem como se os módulos estão relacionados a um determinado mecanismo fisiológico, são necessários ferramentas e métodos de medida que são fornecidos pela Teoria dos Grafos. De uma forma bastante simplificada, um módulo ou agrupamento aparece em uma rede como um grupo de nós altamente interconectados. Um exemplo prático da presença de módulos específicos em redes de interações protéicas pode ser observado na Figura 7. Neste caso, usando novamente os dados de interações entre a enzima fosfofrutocinase hepática humana com outras proteínas, é possível observar dois módulos bastante distintos, situados acima e abaixo do nó correspondente a fosfofrutocinase hepática humana (Figura 7).

Matematicamente, cada módulo pode ser reduzido a um conjunto de triângulos, onde uma alta densidade de triângulos se reflete no coeficiente de agrupamento (Tabela 1). Por sua vez, um alto coeficiente de agrupamentos influencia no coeficiente de agrupamento médio (<C>; Tabela 1), gerando valores elevados quando comparado com redes que são geradas de forma aleatória. A maioria das redes celulares estudadas até o momento, incluindo as de interação protéica, possui um alto valor de <C>, o qual indica que o agrupamento elevado é uma característica de redes biológicas.

Por outro lado, os módulos observados em uma rede biológica são formados por inúmeros elementos unitários denominados de subgrafos. Cada subgrafo possui um conjunto de nós que estão conectados entre si por um padrão muito específico e que constituem, na maioria das vezes, um mecanismo biológico bastante pontual e discreto. Estes subgrafos, assim denominados de motivos, estão na maioria das vezes super-representados em uma rede biológica e também são muito conservados evolutivamente. Por exemplo, o elevado número de motivos conservados em uma rede de interação protéica de leveduras, e a convergência evolutiva de redes de regulação transcricional, presentes em diferentes espécies de organismos, indicam que os motivos e os módulos por eles constituídos são fenômenos biológicos universais e relevantes para a fisiologia do organismo.

Uma questão importante relacionada com a modularidade e a formação de motivos em redes biológicas diz respeito como os componentes moleculares de um motivo específico interagem com os nós que estão fora do motivo. As observações empíricas, especialmente aquelas realizadas com as redes de regulação transcricional de *E. coli*, indicam que tipos específicos de motivos podem se agregar com estes nós de forma espontânea para formar grandes módulos, sendo esta uma propriedade geral das redes.

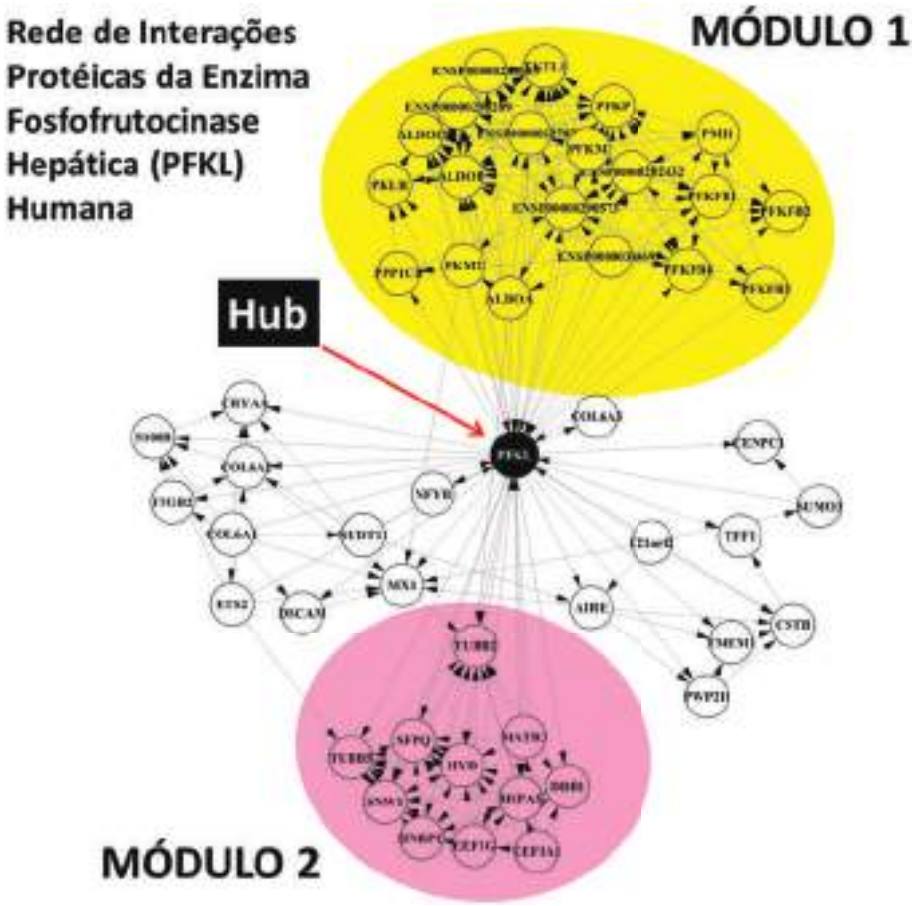


Figura 7. Rede de Interações Protéicas da Enzima Fosfofrutocinase Hepática (PFKL) Humana. A enzima Fosfofrutocinase Hepática (PFKL) representa um "hub", pois está conectada a muitas outras proteínas (nós). Estas proteínas por sua vez podem ser altamente conectadas entre si formando os 2 módulos vistos na figura.

É interessante que, à medida que o número de subgrafos distintos cresce exponencialmente com o número de nós que estão nestes subgrafos, o estudo de grandes motivos torna-se bastante complicado matematicamente, e a definição de mecanismos biológicos particulares torna-se muito difusa. Um método alternativo para o estudo de grandes motivos envolve a identificação de grupos de nós altamente interconectados (ou os módulos), diretamente da topologia da rede (ou seja, da estrutura da rede), seguida da correlação destas entidades com suas funções potenciais. A identificação de módulos é complicada pelo fato de que as propriedades matemáticas de redes livres de escala e dos módulos parecem ser contraditórias em um primeiro momento. Os módulos, por definição, implicam que eles são um grupo de nós relativamente isolados do restante do sistema, mas uma rede livre de escalas possui *hubs* que estão em contato com um grande número de nós, o que torna a existência de módulos relativamente isolados bastante improvável. Mas tanto os módulos quanto os *hubs* coexistem naturalmente em sistemas biológicos, o que também mostra que

os módulos devem se combinar para formar uma rede de natureza hierárquica. Neste caso, as assinaturas de módulos hierárquicos estão presentes em todas as redes celulares investigadas até o momento, em todos os níveis de complexidade conhecidos.

Matematicamente, o estudo dos módulos requer a quebra de uma rede celular em um conjunto de subgrafos biologicamente relevantes, o qual pode ser atingido usando-se métodos de agrupamento (e que levam em conta a topologia da rede) ou combinando o estudo da topologia com dados genômicos. Contudo, em todos os casos, deve-se ter em mente que o início e o fim de um módulo (ou as fronteiras de um módulo), não são claramente definidas devido à natureza hierárquica destas estruturas e de sua interpolação com a rede como um todo.

Quantificando e modelando o sistema

Um dos maiores desafios da Biologia de Sistemas consiste não apenas em como lidar com a enorme quantidade de informação gerada pelas tecnologias "ômicas" mas também na complexidade do próprio sistema biológico. Estima-se que nas 3 bilhões de pares de bases do genoma humano existam cerca de 30.000 genes, dos quais, devido a modificações pós-transcricionais e pós-traducionais poderiam gerar, respectivamente, 180.000 moléculas de RNA e 1.800.000 proteínas. Adicione a essa diversidade, os milhares de RNAs não-codificadores (micro-RNAs, piRNAs, snoRNAs, etc) presentes no genoma humano e as quase infinitas interações possíveis entre esses diferentes tipos de moléculas. Estudar a dinâmica de um sistema tão complexo, em qualquer nível (genes, proteínas, metabólitos, interações) parece ser uma tarefa quase impossível. O objetivo da Biologia de Sistemas, especialmente a do tipo top-down, está em justamente extrair as informações relevantes de dentro desse verdadeiro mar de dados e usar modelos matemáticos para representar e entender as complexas interações dos componentes do sistema.

Microarranjos de DNA, capazes de analisar os padrões de expressão gênica global de uma população de células é a tecnologia "ômica" mais utilizada na Biologia de Sistema top-down. Esta técnica permite a identificação de genes e vias de sinalização, possivelmente envolvidos em centenas de processos biológicos. As mudanças nos níveis de expressão de cada gene podem ser testadas individualmente usando testes estatísticos univariados simples como teste t e ANOVA. Porém, como milhares de genes são medidos em paralelo nos experimentos de microarranjos, uma correção para testes múltiplos é necessária (por exemplo, *False Discovery Rate*). Métodos não-paramétricos univariados como o teste Mann-Whitney e programas de análise populares (por exemplo, SAM) também são utilizados para identificar listas de genes diferencialmente expressos. Um outro modo de análise consiste em agrupar genes que possuem padrões similares de expressão visando revelar grupos gênicos funcionais ou genes co-regulados (mesmos fatores de transcrição ou interações entre eles). O próximo passo nessas análises tenta geralmente interpretar as listas de genes ou grupos de genes de acordo com suas funções biológicas. Utilizando-se bancos de dados como KEGG, Reactome, MsigDB, Gene Ontology e Nature/NCI que organizam os genes em módulos funcionais ou de co-expressão, é possível testar se os genes diferencialmente expressos entre duas ou mais condições estão enriquecidos em

determinadas funções biológicas. Testes estatísticos como o teste exato de Fisher e programas como GSEA, DAVID e GS-SAM são capazes de avaliar se o enriquecimento de genes em um determinado módulo é estatisticamente significativo.

Embora as análises descritas acima sejam certamente informativas, o desafio maior ainda é a integração e a modelagem dos dados gerados pelas tecnologias "ômicas" visando entender e prever o sistema. Inúmeros modelos computacionais e algoritmos foram desenvolvidos para identificar grupos de genes cuja atividade poderia prever, por exemplo, o avanço de uma doença ou a resposta a um tratamento. Métodos como KNN (*k-nearest neighbors*), Support Vector Machine, Random Forest e programas como ClaNC, DAMIP e PAM são bons exemplos dessas ferramentas. Para entender a estrutura de vias metabólicas e de sinalização e analisar seu comportamento em diferentes condições, vários modelos computacionais distintos foram propostos. Estes modelos (não descritos nesse capítulo) vão desde equações diferenciais ordinárias e parciais (geralmente aplicadas à análise do metaboloma), até métodos estocásticos e "petri nets".

Um modelo interessante consiste em aplicar lógica booleana para definir uma via biológica. Sinais podem ser propagados numa via através de pequenas ou grandes variações na atividade de seus componentes. Em sistemas onde o comportamento da sinalização é dominado por grandes variações de atividade gênica, os dados de nível de expressão podem ser usados para descrever um gene em 2 estados ("Ligado" e "Desligado") de acordo com um limite de detecção. Esta discretização faz com que as interações representadas em uma via fiquem conceitualmente e computacionalmente mais simples. Esse método foi aplicado com sucesso para explicar o tipo de resposta de células T após a ligação a um antígeno. No entanto, reduzir a atividade biológica em apenas 2 estados pode afetar diretamente a análise de expressão global.

Aplicações da biologia de sistemas

Estudos de interação proteína-proteína

As redes de interação bioquímica representam a química que permeia todo um sistema biológico e indicam as relações estequiométricas entre biomoléculas que podem ser interconvertidas conforme as necessidades fisiológicas da célula. Assim, uma das questões mais interessantes relacionadas à Biologia de Sistemas está centrada nas interações entre proteínas em um sistema biológico durante uma condição fisiológica específica. Há muito se sabe que as interações entre as proteínas são fundamentais para todos os processos biológicos, variando desde a formação de estruturas celulares e complexos enzimáticos, até a regulação de vias de sinalização ou de sistemas transcricionais. As proteínas freqüentemente funcionam como complexos estáveis ou transientes com outras proteínas, e as interações entre as proteínas podem servir a diferentes funções, como conferir especificidade para as interações entre enzimas e substratos em eventos de transdução de sinais, para a proteção de proteínas em seu ambiente, para facilitar a canalização de um substrato por uma via bioquímica ou para a construção de máquinas moleculares, como o citoesqueleto. Assim, torna-se bastante óbvio que um mapa de interações entre proteínas de baixa e alta afinidade (solúveis ou

associadas a membranas) seja essencial para o entendimento dos processos biológicos e dos mecanismos moleculares em termos de Biologia de Sistemas. Tal mapa deverá incluir pares de interações entre proteínas (também chamado de interação binária) bem como grandes complexos protéicos. Este conhecimento é um pré-requisito fundamental para o entendimento da maioria das funções celulares, especialmente as redes regulatórias e de sinalização. Desta maneira, as análises e as construções de redes de interações entre proteínas requerem métodos que sejam passíveis de escalonamento em larga escala, como o ensaio dois-híbridos de leveduras e a purificação por afinidade e espectrometria de massas para análises sistemáticas.

Apesar deste e de outros métodos estarem atualmente disponíveis para a análise em larga escala de complexos protéicos (Tabela 2), existem importantes limitações destes sistemas que precisam ser levados em conta durante a construção de uma rede de interações entre proteínas. Não entraremos em detalhes de cada método atualmente existente, mas o leitor poderá encontrar excelentes revisões a respeito da aplicabilidade e restrições de cada método.

Tabela 2. Métodos utilizados para a análise e a detecção de complexos protéicos *in vivo* e *in vitro*.

Métodos	**Análise** *In vivo*	**Análise** *In vitro*
Purificação de proteínas por afinidade	Espectrometria de massas	Espectrometria de massas
Genéticos	Ensaio dois-híbridos; mbSUSa;	
Fluorescência Split-GFP	FRETa; BRETa	
Ressonância plasmônica		Ressonância plasmônica de superfície quantitativa
Estrutura cristalina		Estrutura de complexos
Calorimetria		Análise quantitativa de interações protéicas
Microscopia de força atômica		Detecção e quantificação de interações protéicas
Ressonância magnética nuclear (RMN)	Análise quantitativa de grandes complexos protéicos	Mapeamento de interações estruturais usando RMN na célula
Arranjos de proteínas		Identificação e análise de interações protéicas seletivas

Abreviações: bioluminescence resonance energy transfer (BRET); fluorescence resonance energy transfer (FRET); mating-based split ubiquitin system (mbSUS).
bhttp://www.stratagene.com

Em todos os casos, por exemplo, as análises em larga escala são limitadas pelo número de réplicas experimentais, muitas vezes em número bastante limitado e que pode introduzir falsos resultados em uma análise de Biologia de Sistemas. Outro problema está relacionado ao fato de que as interações entre proteínas são avaliadas em um esquema de tudo-ou-nada, como no caso do ensaio de dois-híbridos de leveduras, onde um resultado de auxotrofia (ausência de crescimento em meios de cultura onde um aminoácido ou base nitrogenada esteja faltando) e um resultado de prototrofia (crescimento em meios de cultura onde um aminoácido ou base

nitrogenada esteja faltando) são avaliados conforme um código binário, sem levar em conta as interações transientes que são encontradas nas células. Além disso, a superexpressão de proteínas, comumente observada no ensaio de dois-híbridos de leveduras, pode alterar as concentrações relativas dos potenciais parceiros de interação em uma condição *in vivo*, também gerando falsos positivos ou falsos negativos. Por outro lado, as análises de interações entre proteínas usando extratos protéicos, como as tipicamente feitas para os experimentos de espectrometria de massas, podem resultar na mistura de proteínas de diferentes compartimentos que não refletem a condição *in vivo*. As interações detectadas em tais procedimentos são denominadas "interações potenciais" e são necessárias para obter uma visão geral do interatoma, mas uma posterior comprovação experimental, por métodos estabelecidos de Biologia Molecular e de Bioquímica, se faz necessária para determinar se as interações são relevantes para as funções celulares.

Estudos da interação DNA-proteína

As proteínas ligantes de DNA realizam uma série de funções importantes para a célula, incluindo a regulação da transcrição, a manutenção cromossômica, a replicação e a reparação de DNA. Por outro lado, as interações entre os fatores de transcrição e seus sítios de ligação ao DNA são de grande interesse, visto que estas interações controlam os passos cruciais no desenvolvimento de um organismo e na resposta a estresses de origem ambiental. Além disso, as disfunções relacionadas com os fatores de transcrição podem contribuir para a progressão de várias doenças em humanos.

Outro aspecto que possui forte interesse para todos os pesquisadores que trabalham com controle da expressão gênica e, conseqüentemente, com a análise de dados de interação entre proteínas e sequências de DNA, está relacionado com a modificação pós-traducional de histonas e a sua localização na cromatina, a qual pode afetar, de forma significativa, a regulação gênica de um organismo. Assim, o estudo das interações entre proteínas e DNA tem sido facilitado pelas técnicas relacionadas aos microarranjos de DNA. Apesar de não tão desenvolvidos quanto as técnicas associadas à análise de interações entre proteínas, os métodos de análise de interação DNA-proteínas são similares aos de interação entre proteínas. Por exemplo, a imunoprecipitação de cromatina com uma proteína de interesse seguida da detecção de fragmentos de DNA por meio de microarranjos (ou também chamado de "ChIP-chip") tem sido o método de larga escala corrente para a identificação de sítios de ligação de fatores de transcrição in vivo. Outros métodos alternativos baseados em microarranjos também começaram a ser aplicados recentemente. Por exemplo, a fusão de potenciais proteínas ligantes de DNA com a enzima DNA adenina metilltransferase (Dam) possibilita a identificação de sítios de ligação em sequências de DNA por meio de metilação. O microarranjo ligante de proteínas é outro exemplo, e permite uma rápida caracterização (em larga escala) de proteínas ligantes de DNA. Em todos os casos, deve ser salientado que os métodos em larga escala usados para a identificação de interações entre DNA e proteínas continuam em sua infância, e os exemplos de sua aplicabilidade na Biologia de Sistemas não são tão comuns quantos os de interação protéica.

Organização e evolução dos sistemas biológicos

É interessante observar que a presença quase universal de redes livres de escalas e de *hubs* em sistemas tecnológicos, biológicos e sociais nos remete ao fato de que se trata de um mecanismo natural de organização. Por ser natural, a organização e a evolução destas redes está sujeita as mesmas pressões (seleção natural, deriva genética, entre outros) que controlam a maioria dos sistemas complexos neste planeta. Apesar de pouco estudadas em termos evolutivos, sabe-se que dois processos fundamentais possuem um papel chave no desenvolvimento das redes. Em primeiro lugar, as redes são o resultado de um processo de crescimento, na qual nós ligam-se a ela durante um período de tempo (como, por exemplo, a internet, que cresceu de 1 para mais de 3 bilhões de páginas em um período de dez anos). Em segundo lugar, os nós preferem se conectar a outros nós que já possuam muitas conexões, um processo conhecido por "ligação preferencial". Por exemplo, na internet nós estamos mais familiarizados com páginas altamente conectadas e, conseqüentemente, tendemos a nos associar com elas.

Tanto o crescimento quanto a ligação preferencial são responsáveis pela emergência das propriedades matemáticas das redes livre de escalas e, onde houver um nó com muitas conexões, novos nós tenderão a se conectar com este a uma alta taxa de probabilidade. Este nó, por sua vez, irá ganhar novas ligações em velocidades cada vez maiores em comparação com aqueles que possuem poucos nós e, conseqüentemente, irá se tornar um *hub*. Em termos biológicos, tanto o crescimento quanto a ligação preferencial possuem uma origem comum em redes de proteínas, o que deve estar ligado à duplicação gênica, um fenômeno evolutivo comum entre os organismos e que confere novas características genotípicas e fenotípicas a estes.

Os genes duplicados produzem proteínas idênticas em um primeiro momento e que interagem com os mesmos parceiros (Figura 8). Assim, cada proteína que está em

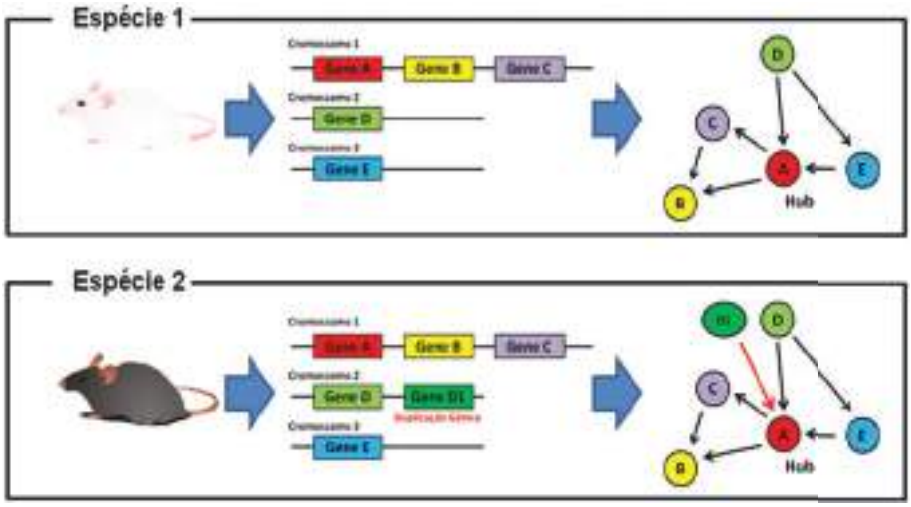


Figura 8. O Papel da Duplicação Gênica nas Redes de Interações de Proteínas. A duplicação gênica em um dado organismo irá criar um gene cuja proteína possui funções e interações idênticas ao gene que foi duplicado. Ao longo da evolução, essa cópia gênica pode ficar livre para sofrer alterações que irão gerar proteínas com novas funções ou diferentes interações.

contato com uma proteína duplicada ganha um conector extra. As proteínas altamente conectadas possuem uma vantagem natural que é a capacidade de montar uma conexão com outras proteínas quando comparado com as proteínas fracamente conectadas e podem, desta forma, adquirir novas ligações. Embora o papel da duplicação gênica tenha sido mostrado apenas para redes de interação de proteínas, este também explica a emergência de propriedades livres de escalas em redes metabólicas e regulatórias. Entretanto, e como salientado anteriormente, o estudo dos mecanismos de evolução de redes livres de escalas continua em sua infância mas, com o conhecimento atual, é possível prever outras características importantes das redes livres de escala, como a robustez.

A robustez de uma rede livre de escalas diz respeito à capacidade desta em tolerar a falha de muitos nós sem perder a sua atividade. De fato, muitos sistemas complexos, como é o caso da internet ou de sistemas de comunicação (como os telefones celulares), são bastante resilientes (em outras palavras, toleram as mudanças sem perder as funções aos quais foram designados) contra as falhas dos componentes. Para as redes livres de escalas, tem sido observado que a perda de 80% dos nós escolhidos de forma aleatória não compromete a rede (esta perda é chamada de "falha acidental"), onde os 20% remanescentes continuam a formar agrupamentos compactos que possibilitam a ligação de dois nós quaisquer. Isto se deve ao fato de que as falhas aleatórias afetam, principalmente, os nós com baixo número de conectores, o que não ocasiona a quebra da integridade da rede. Por outro lado, a falha dos *hubs* promove o que é chamado de "vulnerabilidade de ataque", onde a rede se parte em pequenos agrupamentos de nós. Neste sentido, tem sido descrito para *S. cerevisiae* e para *E. coli* que a deleção de proteínas que possuam mais de 15 conectores gera um fenótipo letal, onde a célula não consegue sobreviver sem a presença do *hub*. Estes mesmos *hubs*, pela sua importância para a célula, também são os mais conservados e antigos em termos evolutivos, e normalmente são encontrados na maioria dos organismos estudados.

Outra característica importante de redes livres de escalas e que está associada a sua organização é chamado de "falha em cascata". A falha em cascata ocorre quando um *hub* é removido de uma rede e, conseqüentemente, a rede sofre uma redistribuição de nós que fica mais propensa a falhas. À medida que as falhas vão acontecendo, estas vão se somando até que toda a rede colapsa e o sistema cai por completo. Isto pode ser observado em sistemas tecnológicos, como as redes de distribuição de energia, onde a falha em uma grande estação de transmissão promove a propagação desta por todas as estações menores, resultando em um apagão energético. Em sistemas biológicos, o efeito de uma falha em cascata pode promover a morte do organismo.

Assim, a compreensão da organização e da evolução de redes livres de escala é fundamental para o entendimento dos processos que governam estas entidades. Como vimos até este momento, é possível quantificar as interações e os processos entre os diferentes componentes de um sistema biológico por meio de ferramentas disponibilizadas pela Teoria dos Grafos. Em alguns casos, as informações estão presentes em quantidade suficiente, e a tecnologia usada para a obtenção de dados empíricos é madura, para que seja possível realizar inferências experimentais, como é o caso da Biologia de Sistemas de top-down. Em outros casos, como é o caso da Biologia de Sistemas de bottom-up, as informações são muito mais restritas e pontuais, e as tecnologias empregadas para a obtenção dos dados empíricos estão em sua infância.

Entretanto, deve ser ressaltado que a Biologia de Sistemas está revolucionando a nossa forma de compreender os organismos do ponto de vista "ômico" e evolutivo, além de estar gerando ferramentas experimentais extremamente avançadas para aplicações biológicas. À medida que o nosso conhecimento "ômico" avança, a Biologia de Sistemas tem acompanhado este ritmo, tornando-se uma área do conhecimento científico cada vez mais presente e fundamental para as pesquisas com organismos em todos os seus níveis.

Bibliografias

ALBERT, Réka, BARABÁSI, Albert-László. Statistical physics of complex networks. *Rev. Mod. Phys.* 74: 47, 2002.

ALBERT, Réka. Scale-free networks in cell biology. *J. Cell. Sci.* 118: 4947-4957, 2005.

BADER, S., KÜHNER, S., GAVIN, A. C. Interaction networks for systems biology. *FEBS Lett.* 582: 1220-1224, 2008.

BARABÁSI, Albert-László, OLTVAI, Zoltán. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5: 101-113, 2004.

BAXEVANIS, Andreas. The molecular biology database collection: 2003 update. *Nucleic Acids Res.* 31: 1-12, 2003.

BORODINA, I., NIELSEN, J. From genomes to in silico cells via metabolic networks. *Curr Opin Biotechnol.* 16: 350-355, 2005.

BRUGGEMAN, F. J., WESTERHOFF, H. V. The nature of systems biology. *Trends Microbiol.* 15: 45-50, 2007.

CARY, Michael, BADER, Gary, SANDER, Chris. Pathway information for systems biology. *FEBS Lett.* 579: 1815-1820, 2005.

ERDÖS, Paul, RÉNYI, Alfréd. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* 5, 17–61 (1960).

HANLON, S.E., LIEB, J. D. Progress and challenges in profiling the dynamics of chromatin and transcription factor binding with DNA microarrays. *Curr. Opin. Genet. Dev.* 14: 697-705, 2004.

HERNANDEZ, Thomas, KAMBHAMPATI, Subbarao. Integration of biological sources: current systems and challenges ahead. *SIGMOD Record* 33, 51-60, 2004.

HUBER, W., CAREY, V. J., LONG, L., FALCON, S., GENTLEMAN, R. Graphs in molecular biology. *BMC Bioinformatics.* 8: S8, 2007.

KWOH, Chee Keong, NG, P. Y. Network analysis approach for biology. *Cell. Mol. Life Sci.* 64: 1739-1751, 2007.

LALONDE, S., EHRHARDT, D. W., LOQUÉ, D., CHEN, J., RHEE, S. Y., FROMMER, W. B. Molecular and cellular approaches for the detection of protein-protein interactions: latest techniques and current limitations. *Plant J.* 53: 610-635, 2008.

LEE, Der-Tsai.; PREPARATA, F. P. Computational geometry-A survey. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 33, 1072-1101, 1984.

LEI, Li, ROOP, G. Singh, ZHENG, Guangzhi, VANDENBERG, Art, VAISHNAVI, Vijay K., NAVATHE, Shamkant B. A methodology for semantic integration of metadata in bioinformatics data sources. *ACM Southeast Regional Conference* 1: 131-136, 2005.

LEVY, E. D., PEREIRA-LEAL, J. B. Evolution and dynamics of protein interactions and networks. *Curr Opin Struct Biol.* 18, 349-357, 2008.

MASHANOV, G. I., NENASHEVA, T. A., PECKHAM, M., MOLLOY, J. E. Cell biochemistry studied by single-molecule imaging. *Biochem Soc Trans.* 34: 983-988, 2006.

SAKO, Y. Imaging single molecules in living cells for systems biology. *Mol Syst Biol*. 2: 56, 2006.

SRIVASTAVA, R., VARNER, J. Emerging technologies: systems biology. *Biotechnol Prog*. 23: 24-27, 2007.

STEPHANOPOULOS, G., ALPER, H., MOXLEY, J. Exploiting biological complexity for strain improvement through systems biology. *Nat Biotechnol*. 22: 1261-1267, 2004.

STRANGE, K. The end of "naive reductionism": rise of systems biology or renaissance of physiology? *Am. J. Physiol. Cell Physiol*. 288: C968-C974, 2005.

UETZ, P., FINLEY, R. L. Jr. From protein networks to biological systems. *FEBS Lett*. 579: 1821-1827, 2005.

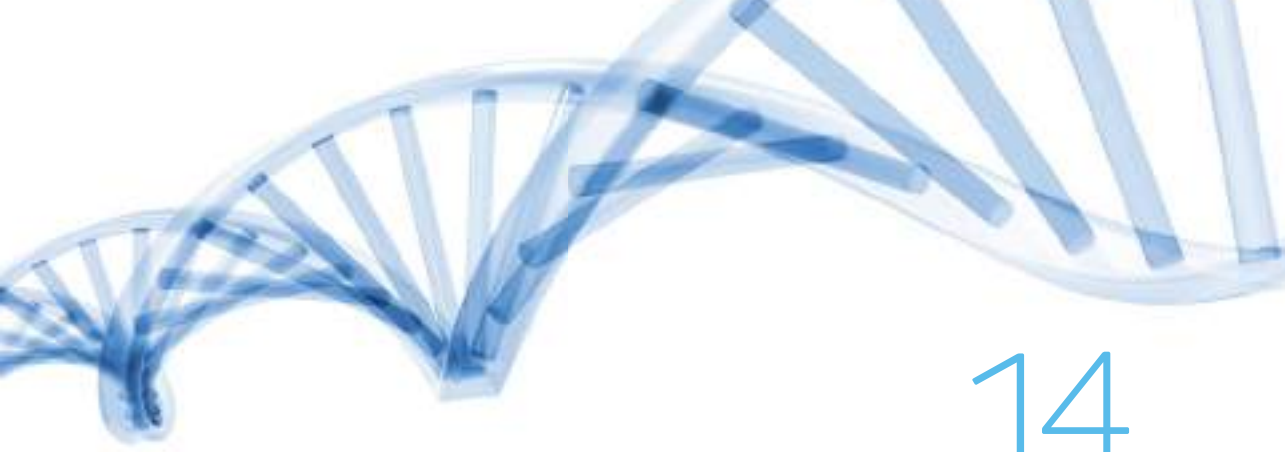
VISWANATHAN, G. A., SETO, J., PATIL, S., NUDELMAN, G., SEALFON, S. C. Getting started in biological pathway construction and analysis. *PLoS Comput Biol*. 4: 16, 2008.

YOU, L. Toward computational systems biology. *Cell Biochem Biophys*. 40: 167-184, 2004.

Páginas na Internet

String (http://string.embl.de/)

Systems Biology (http://www.systems-biology.org/)



14

Genômica e o código de barras da vida

Luiz Henrique Garcia Pereira
Claudio Oliveira

Introdução

O desenvolvimento das técnicas de sequenciamento de DNA, principalmente aquelas baseadas em sequenciadores de segunda geração, tem levado a estudos de composição de genomas de diversas espécies, nos quais são identificados milhares ou milhões de nucleotídeos que compõe cada genoma. Como visto em diversos capítulos anteriores desse livro, atualmente a Genômica é uma ciência muito bem consolidada e é crescente o número de genomas que se tem sequenciado. Porém, a aplicação dessa metodologia é ainda direcionada para umas poucas espécies, quando comparamos toda a diversidade da vida na Terra. Assim, sob essa ótica, hoje sabemos muito sobre o genoma de muito poucas espécies. No sentido contrário, um grupo de pesquisadores se propôs a realizar um estudo diferente: sequenciar um (ou alguns poucos) genes de todas as espécies do planeta! Nasceu assim a técnica conhecida como *DNA barcoding* que vamos discutir em detalhes abaixo.

O conhecimento sobre a diversidade biológica é o ponto de partida para todos os estudos básicos ou aplicados relacionados às ciências da vida e o reconhecimento de espécies, bem como a atividade de nomeá-las é fundamental para o estudo da ecologia, comportamento, evolução e todas as outras disciplinas relacionadas aos organismos (Savage, 1995).

Apesar de muito difundido nos dias atuais, o termo diversidade biológica foi utilizado pela primeira vez em 1968 por Raymond F. Dasmann em seu livro *"A Different Kind of Country"* no qual advogava para a conservação (Dasmann, 1968). No entanto, o termo só foi amplamente adotado a partir da década de 1980 com as publicações do livro *"Conservation Biology"* (Soulé e Wilcox, 1980) e do trabalho de Norse e

McManus (1980), os quais apresentaram o termo à comunidade científica. Já o termo concatenado, biodiversidade, tem seu uso mais recente, sendo utilizado pela primeira vez em 1985, por W. G. Rosen para uma reunião do Foro Nacional Americano sobre diversidade biológica e tornou-se conhecido a partir, principalmente, da publicação do livro organizado por Wilson e Peter em 1988, intitulado *"Biodiversity"* (Wilson e Peter, 1988). Os dois termos são utilizados, atualmente, como sinônimos e sua definição é ampla. A convenção sobre Diversidade Biológica (CDB) realizada pela Organização das Nações Unidas (ONU) em 1992 (Eco-92) definiu diversidade biológica como sendo *"a variabilidade de organismos vivos de todas as origens, compreendendo, dentre outros, os ecossistemas terrestres, marinhos e outros ecossistemas aquáticos e os complexos ecológicos de que fazem parte; compreendendo ainda a diversidade dentro das espécies, entre espécies e de ecossistemas"* (Dias, 2000). A diversidade dentro das espécies aqui apresentada deve ser entendida como toda a variação existente entre indivíduos de uma população e entre populações de uma mesma espécie a qual pode ser designada como diversidade genética. Assim, os termos diversidade biológica ou biodiversidade englobam todos os níveis hierárquicos de variabilidade, desde os genes aos ecossistemas.

A espécie é uma unidade de comparação fundamental em todos os campos da biologia, da anatomia ao comportamento, desenvolvimento, ecologia, evolução, genética, biologia molecular, paleontologia, fisiologia, sistemática, etc (de Queiroz, 2005). Ao longo da história, muitos conceitos de espécie foram propostos, incluindo o tipológico, morfológico, biológico, por isolamento reprodutivo, etc. Ainda que extensos debates sejam constantemente travados em relação a esses conceitos de espécie (de Queiroz, 2005, 2007; Waugh, 2007), do ponto de vista prático, os taxonomistas são os profissionais responsáveis pela caracterização dessas entidades biológicas e sua classificação, tornando-as palpáveis e reconhecíveis pela atribuição de um nome, erigido de acordo com os códigos internacionais de nomenclatura (Köhler, 2007).

A atribuição de um nome não constitui uma simples aplicação de regras de nomenclatura, mas sim a elaboração de uma hipótese, segundo a qual um determinado conjunto de caracteres (usualmente morfológicos) é capaz de identificar uma entidade (espécie) com características biológicas próprias e história evolutiva independente de outras entidades biológicas similares. Essas hipóteses podem ser testadas de diversas maneiras e, como todas as hipóteses, podem ser refutadas ou não. Adicionalmente, quando as descrições de espécies são baseadas em uma ampla base de dados, elas se tornam hipóteses científicas interessantes permitindo a elaboração de predições explícitas sobre os atributos dos organismos (Lipscomb et al., 2003).

Como identificamos espécies

O modelo tradicional de descrição e identificação de espécies é baseado em características morfológicas. Os dados morfológicos foram, historicamente, os primeiros a serem utilizados na identificação de espécies, simplesmente, pelo fato de que foram os primeiros disponíveis aos pesquisadores que iniciaram a sistematização do conhecimento sobre os seres vivos. É atribuído ao filósofo grego Aristóteles (384 a.C. – 322 a.C.) o primeiro sistema de classificação biológica, o qual consistia na separação

dos organismos de acordo com suas semelhanças, diferenças e atributos (posição na natureza), consistindo dessa forma num sistema de classificação artificial. O sistema de classificação atual, sistema binomial, somente foi proposto em 1735 por Carl Linnaeus (1707-1778) com a publicação do seu livro *"Systema naturae"* (Figura 1). Contudo, embora houvessem regras mais criteriosas para a classificação dos organismos, os mesmos eram agrupados de acordo com suas semelhanças e diferenças, consistindo assim, também, num sistema de classificação artificial. Somente em 1950, com o estabelecimento da sistemática filogenética por Willi Hennig e a consolidação da escola cladista, a qual se baseia, não somente nas semelhanças e diferenças entre os organismos para a classificação dos mesmos, mas também na atribuição de ancestrais comuns às espécies. Assim, o sistema de classificação passou a procurar grupos naturais de organismos.

REINO - Animalia
FILO - Chordata
SUBFILO - Vertebrata
CLASSE - Mammalia
ORDEM - Primates
FAMÍLIA - Hominidae
GÊNERO - *Homo*
ESPÉCIE - *Homo sapiens*

Figura 1. Classificação taxonômica da espécie humana (*Homo sapiens*) seguindo a classificação binomial de Linneu.

Com o avanço da ciência e o aumento do número de espécies estudadas, novas fontes de dados foram sendo adicionadas aos processos de descrição e classificação das espécies como caracteres embriológicos (ontogenéticos), fisiológicos, comportamentais, moleculares, dentre outros, complementando os dados morfológicos clássicos.

Atualmente, a descrição de espécies é baseada na comparação da espécie a ser descrita com todas as espécies relacionadas a ela, dentro do nível taxonômico em questão (espécie, gênero, família, etc). Para isto utiliza-se, normalmente, um conjunto de caracteres morfológicos previamente estabelecidos e característicos para cada grupo, que permitem a comparação dos mesmos com as espécies mais próximas (Figura 2) sem, contudo, limitar a inserção de novos caracteres sempre que os mesmos forem necessários. Assim, são estabelecidos caracteres diagnósticos (caracteres ou combinações de caracteres exclusivos) que definem espécies, gêneros, famílias, etc.

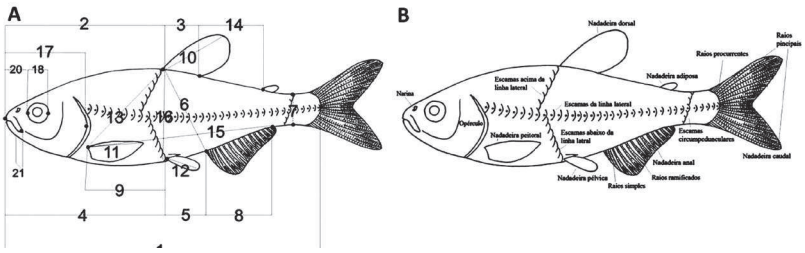


Figura 2. Identificação de espécies pelo método morfológico. Diagrama ilustrativo de características morfológicas utilizadas na identificação de espécies de peixes. (A) dados morfométricos (medidas; ex. distância pedúnculo caudal a nadadeira dorsal); (B) dados merísticos (contagem, ex. número de escamas na linha lateral).

Uma vez descritas as espécies, são construídas, pelos taxonomistas, chaves de identificação dicotômicas (Figura 3) de acordo com o nível de especificidade desejada (Ex.: espécies de um gênero, gêneros de uma região específica, famílias de uma ordem, etc) que permitem a atribuição de um espécime à sua espécie. Outra alternativa, correntemente em uso, é o envio dos espécimes que se deseja identificar aos respectivos especialistas (taxonomistas) dos grupos para que estes façam as análises comparativas necessárias para identificar a espécie a que pertence esse espécime.

Chave dicotômica de identificação para anfíbios anuros do noroeste paulista

1a. Cristas craniais presentes (Figura 7a); pele áspera; membranas interdigitais entre os dedos ausentes e bastante reduzidas ou ausentes entre os artelhos; glândulas parotóides presentes (Figura 7a) ..... BUFONIDAE: *Rhinella* 2

1b. Cristas craniais ausentes; pele lisa ou pouco granulosa, sem protuberâncias; glândulas parotóides ausentes 3

2a. (1a) Glândula parotóide grande (Figura 7a), com comprimento superior a 25% do CRC; glândula paracnêmica presente na região da tíbia (Figura 7b); CRC variando entre 120 e 160 mm. Maioria dos indivíduos sem faixa creme na maxila superior ..... *R. schneideri*

2b. Glândula paratóide pequena, com comprimento inferior a 20% do CRC; glândula paracnêmica ausente; CRC variando entre 52 e 70 mm; a maioria dos indivíduos com uma faixa creme na maxila superior, que se estende do lábio superior até a margem do olho; presença de uma faixa branca vertebral no dorso ..... *R. ornata*

3a. (1b) Corpo ovoide ou globular, com exceção de *Chiasmocleis*, que apresenta corpo esguio e alongado; focinho acuminado; cabeça pequena e triangular; com um único tubérculo metatarsal pouco desenvolvido (Figura 7h); tímpano indistinguível; artelho III maior que artelho V (Figura G); ventre sem grânulos; membros reduzidos, com dedos pequenos sem discos adesivos e com tubérculos subarticulares .... MICROHYLIDAE 4

3b. Corpo alongado na maioria das espécies; com dois ou mais tubérculos metatarsais (ou um único tubérculo metatarsal bem desenvolvido em *Pseudis platensis*); tímpano distinguível (Figura F) (indistinguível em *Pseudopaludicola*); membro posterior bem desenvolvido 7

Figura 3. Chave de identificação dicotômica. Parte da chave de identificação de anfíbios do noroeste paulista (Provete et al., 2011).

No entanto, a forma tradicional de identificação de espécies enfrenta algumas limitações como a plasticidade (variação) fenotípica dos caracteres utilizados para o reconhecimento das espécies, que podem levar a erros de identificação, devido à existência de espécies chamadas de crípticas (espécies quase idênticas do ponto de vista morfológico) em muitos grupos, as quais não podem ser identificadas facilmente e; ao fato de as chaves de identificação para diversos grupos serem formuladas para um único estágio de vida e/ou sexo (ex.: insetos – alguns grupos são identificados pela genitália do macho adulto; angiospermas – muitas espécies são identificadas pela fórmula floral) impossibilitando, ou tornando muito difícil, a identificação de muitos espécimes (Hebert et al., 2003).

Além disso, o número cada vez maior de espécies descritas (atualmente são aprox. 1,2 milhões) tem exigido um progressivo aumento no número de caracteres utilizados para a descrição e correta identificação das espécies, tarefa nem sempre fácil ou até mesmo possível, devido a limitação nos caracteres utilizados.

Números da biodiversidade

Em 250 anos de classificação biológica, desde Carl Linnaeus (1707-1778), foram reconhecidas e descritas formalmente aproximadamente 1,2 milhões de espécies

(Mora et al., 2011). No entanto, o número de espécies estimadas para o planeta é da ordem de 8 a 10 milhões (Figura 4 – Tabela 1) (Hammond, 1992; Howksworth e Kallin-Arroyo, 1995; Cox e Moore, 2000; May e Harvey, 2009; Mora et al., 2011). Estima-se que ainda são desconhecidas 86% das espécies do planeta e que 91% das espécies presentes nos oceanos ainda esperam por ser descritas (Mora et al., 2011).

No Brasil são conhecidas e descritas formalmente aproximadamente 200 mil espécies e estima-se que a biodiversidade brasileira esteja compreendida entre 1,4 e 2,4 milhões de espécies (10% a 24% de toda biodiversidade mundial) (Lewinsohn, 2005) (Figura 4 – Tabela 1), ou seja, entre 86% e 91% das espécies brasileiras permanecem desconhecidas para a ciência. Estima-se que 13% de todas as espécies de anfíbios (Silvano e Segalla, 2005), 10% das espécies de mamíferos (Costa et al., 2005), 17,8% das borboletas (Brown e Freitas, 1999) e 21% dos peixes de águas continentais (Agostinho et al., 2005) encontram-se no Brasil.

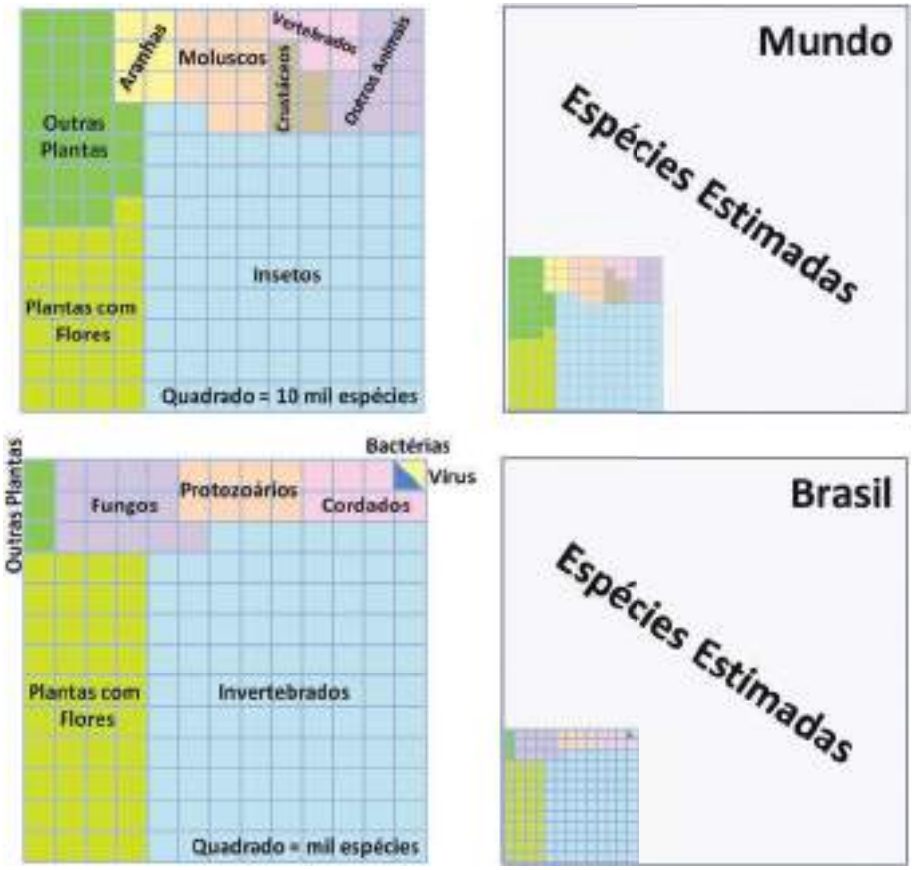


Figura.4. Número de espécies conhecidas e estimadas para o Brasil e para o mundo.

Tabela 1. Números de espécies conhecidas e estimadas para o Brasil e para o Mundo. Dados extraídos do trabalho de Lewinsohn e Prado (2005) com modificações.

	Brasil		Mundo	
Grupos	**Conhecidas**	**Estimadas**	**Conhecidas**	**Estimadas**
Vírus	310-410	40.100-70.400	3.600	400.000
Bactérias e Arquéas	800-900	100.200-175.900	4.300	1.000.000
Fungos	13.090-14.510	150.300-263.900	70.600-72.000	1.500.000
Protistas	7.650-10.320	60.100-105.600	76.100-81.300	600.000
Plantas	43.020-49.520	48.500-54.500	263.800-279.400	320.000
Nematódeos	1.280-2.880	40.100-70.400	15.000-25.000	400.000
Crustáceos	2.040	15.000-26.400	36.200-39.300	150.000
Aracnídeos	5.600-6.500	75.200-132.000	80.000-93.000	750.000
Insetos	80.750-109.250	801.800-1.407.600	950.000	8.000.000
Moluscos	2.400-3.000	20.000-35.000	70.000-100.000	200.000
Cordados	7.120-7.150	7.200-8.800	60.800	50.000
Outros	4.580-8.870	25.100-44.000	67.200-89.800	250.000
Total	168.640-212.650	1.383.600-2.394.700	1.697.600-1.798.500	13.620.000

Extinção

Um fator preocupante frente ao grande número de espécies ainda por se conhecer é a extinção. No Brasil, a lista vermelha de espécies ameaçadas de extinção inclui 1173 espécies da fauna (ICMBio – MMA) e 2479 espécies da flora (CNCFlora) (Tabela 2). O maior número de espécies ameaçadas nos dois grupos encontra-se nos dois biomas mais degradados do Brasil, Mata Atlântica e Cerrado (ICMBio – MMA; CNCFlora). No mundo, segundo a *International Union for Conservation of Nature* (IUCN), o mais importante órgão internacional de levantamento de dados sobre espécies ameaçadas de extinção e responsável pela elaboração da lista vermelha mundial, são apontadas 22.413 espécies ameaçadas de extinção entre animais, vegetais e fungos (Tabela 3) (IUCN, 2014). Os valores aqui apresentados, tanto para o Brasil como em nível mundial, são reconhecidamente subestimados, uma vez que se referem apenas às espécies já descritas e reconhecidas, e para as quais já existem dados disponíveis. Assim, extrapolando esses valores para toda a biodiversidade estimada para o planeta, centenas a milhares de espécies devem se extinguir por ano sem antes mesmo serem conhecidas. Estima-se que o planeta perca aproximadamente 0,25% de suas espécies por ano, o que representa uma perda de 25 a 37,5 mil espécies, baseados nas estimativas de 10 a 15 milhões de espécies existentes (Wilson, 1993), e já se acredita que o planeta esteja passando pela sexta extinção em massa (Ananthaswany, 2004; Wake e Vendreburg, 2008). Ao longo da história, o planeta Terra já passou por outros cinco períodos de extinção em massa, dos quais o mais recente e famoso foi o que extinguiu os dinossauros cerca de 65 milhões de anos atrás (Figura 5).

Tabela 2. Número de espécies animais e de plantas sob algum nível de ameaça de exti nção no Brasil. Dados extraídos das listas vermelhas de espécies em extinção do Brasil, portarias MMA nº 444 e 445, de 17 de Dezembro de 2014 (fauna) e Livro Vermelho da Flora do Brasil (flora) disponíveis nos sites do ICMBio – MMA e CNCFlora, respectivamente, com modificações.

Grupos	Espécies ameaçadas	Total de espécies	% de espécies ameaçadas
Aves	234	1.800	13
Mamíferos	110	658	16,7
Répteis	80	641	12,5
Anfíbios	41	776	5,3
Peixes	409	2.868	14,3
Total Vertebrados	*874*	*6.743*	*13,0*
Hemicordados	1	7	14,3
Equinodermos	10	329	3,0
Insetos	137	89.000	0,1
Aracnídeos	53	5.600	0,9
Colêmbolos	15	-	-
Miriápodes	15	320	4,7
Moluscos	24	2.400	1,0
Crustáceos	28	2.040	1,4
Anelídeos	1	1.000	0,1
Cnidários	6	470	1,3
Poríferos	5	300	1,7
Onicóforos	4	4	100,0
Total Invertebrados	*299*	*101.470*	*0,3*
Total animais	1173	108.200	1,1
Plantas	2479	48.500	5,1
Total Geral	3652	156.700	2,3

Tabela 3. Lista resumida das espécies sob algum nível de ameaça de extinção no Mundo. Dados extraídos da lista vermelha de espécies ameaçadas de extinção da IUCN (2014) com modificações.

Grupos	Espécies ameaçadas	Total de espécies	% de espécies ameaçadas
Vertebrados	7,678	66.178	11,6%
Invertebrados	4.140	1.305.250	0,3
Plantas	10.584	307.674	3,4
Fungos e Protistas	11	51.623	0,02
Total	22413	1.730.725	1,3

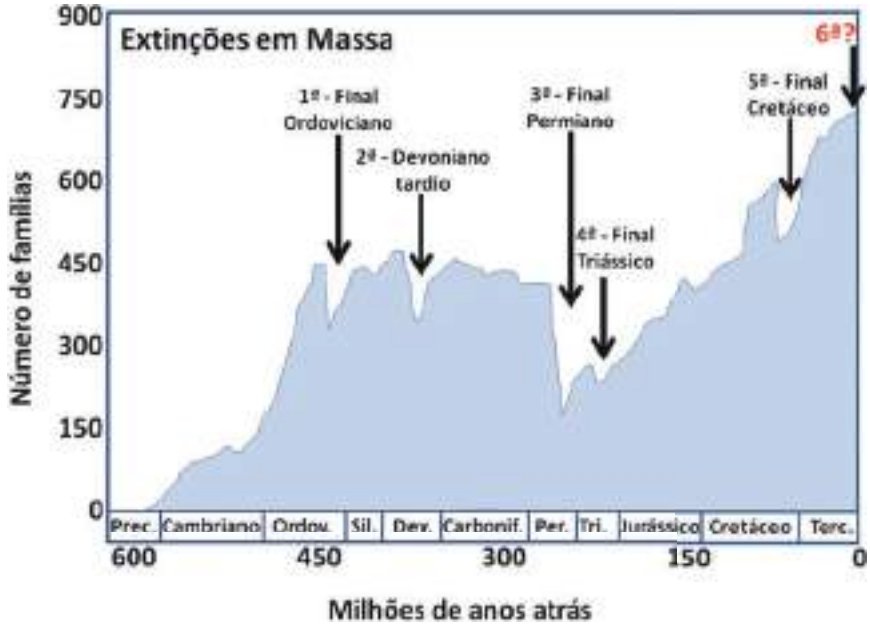


Figura 5. Gráfico mostrando as cinco extinções em massa ocorridas no planeta.

Impedimento taxonômico

No período de 1978 a 1995 foram descritas no Brasil 7.320 espécies dos mais diversos grupos, ou seja, uma taxa média de descrição de 430 espécies/ano (Marques e Lamas, 2006). Atualmente, estima-se que a taxa de descrição média no Brasil seja de 1.500 espécies/ano (Lewinshon e Prado, 2005). Desta forma, se essa taxa se manter, serão necessários aproximadamente 1.600 anos para descrever toda biodiversidade brasileira estimada (1,4 a 2,4 milhões de espécies), estimativa que pode ser extrapolada para toda biodiversidade do planeta. Além disso, existe uma escassez mundial de taxonomistas que podem identificar as espécies, descrever espécies que são novas para a ciência, determinar suas relações taxonômicas e fazer previsões sobre suas propriedades. Para alguns grupos, sequer existem taxonomistas treinados. Esta escassez tende a piorar, porque a força de trabalho taxonômico atual está envelhecendo, juntamente com um declínio do número de alunos formados em taxonomia. Para completar o quadro, há um declínio no número de cargos pagos que permitem que uma pessoa possa se dedicar exclusivamente à taxonomia. O que não falta é interesse em taxonomia e taxonomistas em potencial. Até mesmo os taxonomistas treinados estão sendo subutilizados devido ao investimento insuficiente de recursos para a pesquisa em taxonomia. Todo grande museu sofre atualmente com o acúmulo de amostras não estudadas e novas espécies já reconhecidas, porém aguardando para serem descritas. Sem citar o fato dos curadores e taxonomistas verem o número de seus alunos diminuindo por falta de financiamento e/ou bolsas de estudos ou mesmo por não conseguirem um trabalho remunerado na área.

Atualmente, se discute bastante, em nível mundial, questões sobre a biodiversidade como sua utilização e preservação. Assim, parece contraditório o declínio observado de taxonomistas, profissionais essências para inventariar e caracterizar a biodiversidade. A descoberta de novos materiais de uso potencial para a humanidade passa pelo conhecimento e descrição das espécies, além do fato de que sem conhecer não é possível preservar.

O Brasil, através do Ministério da Ciência e Tecnologia, lançou em 2005 um programa de Capacitação em Taxonomia que previa a formação de 60 novos doutores num período de sete anos. No entanto, sabe-se que a formação de especialistas nessa área é lenta e gradual e esse número, apesar de aumentar em 46% o número de taxonomistas no Brasil, é ainda insuficiente. Iniciativas similares têm surgido ao redor do mundo, porém insuficientes para sanar o déficit de profissionais na área.

Outro fator a ser considerado é o difícil acesso à materiais comparativos, os quais são necessários, principalmente, na fase de descrição de novas espécies. Grande parte destes materiais está espalhada por diversos museus ao redor do mundo que, associados à falta de financiamento para a realização de empréstimos entre instituições e, a burocracia, muitas vezes existente, dificultam ou até mesmo impedem o acesso à eles.

Soma-se ao que foi apresentado neste tópico, os problemas apresentados nos anteriores como as limitações do método tradicional (morfológico) de descrição e identificação de espécies, o enorme número de espécies estimadas e ainda desconhecidas e o fator extinção. Este conjunto de fatores e limitações na área taxonômica é conhecido pelo termo "impedimento taxonômico", o qual tem gerado uma enorme lacuna no conhecimento da biodiversidade.

Diante desse cenário, novas metodologias se fazem necessárias para auxiliar as metodologias tradicionais na identificação, estudo e conhecimento da biodiversidade. Nesse contexto, a biologia molecular tem se mostrado uma ferramenta valiosa, oferecendo uma nova e rica fonte de dados (caracteres) que permite o estudo dos organismos nos seus mais variados níveis de organização, de indivíduos a níveis taxonômicos elevados (ex.: ordens), estudos comportamentais, ecológicos, populacionais e evolutivos, além do uso intensivo em biotecnologia.

Métodos alternativos de identificação de espécies

Com o desenvolvimento de novos métodos de estudos, novas metodologias foram se tornando disponíveis para o conhecimento da biodiversidade. Dessa maneira, há quase 50 anos, a eletroforese de proteínas em géis de amido foi, pela primeira vez, utilizada para identificar espécies (Manwell e Baker, 1963). Há aproximadamente 35 anos, a análise das sequências de nucleotídeos do DNA ribossômico foi utilizada para investigar as relações evolutivas em níveis taxonômicos superiores (Woese e Fox, 1977) e as pesquisas em DNA mitocondrial dominaram a sistemática molecular no final da década de 70 e início da década de 80 (Avise, 1994) e hoje constituem um dos principais sustentadores desse tipo de investigação, com várias revistas dedicadas exclusivamente à esse campo como: Molecular Phylogenetics and Evolution, Molecular Biology and Evolution e Journal of Molecular Evolution.

Diversas técnicas moleculares têm sido utilizadas com o intuito de identificar espécies, tais como: isozimas (Aron e Solé-Cava, 1991; Gusmão et al., 2000), fragmentos obtidos por enzimas de restrição (RFLP) (Moysés e Almeida-Toledo, 2002), DNA *arrays* (Hajibabaei et al., 2007), SNPs (single-nucleotide polymorphism) (Shaffer e Thonsom, 2007), PCR-Multiplex (Mendonça et al., 2009), sequências de DNA dos mais variados genes (Pook e McEwing, 2005; Lemer et al., 2007), dentre outros. Nos estudos taxonômicos essas 'novas' categorias de dados têm sido sempre adicionadas aos dados morfológicos, nunca pretendendo substituí-los. Exemplos desse tipo de integração são cada vez mais comuns na literatura, como na descrição de uma nova espécie de tainhas, do gênero *Mugil* (Harrison et al., 2007), a qual só pode ser reconhecida como nova e descrita formalmente após o acumulo de evidências citogenéticas (estudo dos cromossomos) e moleculares que demonstravam a singularidade das amostras em estudo com relação a seus respectivos congêneres (espécies pertencentes ao mesmo gênero) (Nirchio et al., 2007).

Vantagens dos métodos moleculares

A identificação de espécies por métodos moleculares apresenta uma série de vantagens em relação ao método tradicional, embora não vise substituí-lo e sim complementá-lo. São elas:

- *método rápido* – o avanço da ciência molecular tem tornado a obtenção de dados mais rápida, precisa e barata. Um método rápido de identificação de espécies se faz necessário para acelerar a compilação da biodiversidade.
- *Identificação a partir de fragmentos* - a fonte de dados primária nos métodos moleculares é a molécula de DNA (ácido desoxirribonucléico), esteja ela presente no núcleo celular (organismos eucariotos) ou nas organelas citoplasmáticas tais como mitocôndrias e cloroplastos. Através do processo de divisão celular (mitose), que se inicia no zigoto, cada célula do organismo carrega uma cópia completa do seu genoma (conjunto total de material genético de um organismo). Desta forma, independente da fonte de material biológico, é possível (dependendo do estado de conservação do material) obter amostras de DNA e proceder a identificação. A identificação a partir de fragmentos torna possível o estudo de uma grande gama de material biológico antes inacessível do ponto de vista de identificação. Muitas espécies são raras e/ou difíceis de obter, estando disponível, muitas vezes, apenas vestígios (ex. pêlos, penas), os quais podem ser coletados e submetidos aos métodos de identificação. O uso de fragmentos torna possível análises forenses. Outras espécies, por seu tamanho ou dificuldade de transporte, exigem que a identificação seja realizada em campo, o que nem sempre é possível, desta forma a aquisição de um pequeno fragmento do espécime permite sua posterior identificação em laboratório. Além disso, é possível identificar espécies a partir de materiais industrializados, provenientes do tráfico ilegal de animais e plantas dentre outros.
- *identificação a partir de qualquer estágio do ciclo de vida* – a identificação de diversos grupos de organismos é baseada apenas em um estágio de vida e/ou sexo específicos (ex.: insetos (machos adultos), plantas fanerógamas (flor)), inviabilizando, ou

tornando muito difícil, a identificação de inúmeros espécimes. O DNA esta presente em todas as células do organismo, podendo ser obtido a partir de qualquer estágio de vida da espécie, o que permite a identificação a partir de ovos, larvas, ninfas, pupas e adultos.

- *identificação de espécies semelhantes* – muitos grupos de organismos são difíceis de identificar devido à grande semelhança morfológica existente entre suas espécies, reflexo da grande diversidade existente. Desta forma, o uso de caracteres moleculares permite uma identificação mais precisa nesses casos. O uso de ferramentas moleculares permitiu, nas últimas duas décadas, a identificação e o reconhecimento de inúmeras espécies crípticas (espécies muito similares morfologicamente) (Figura 6) (Bickford et al., 2007).

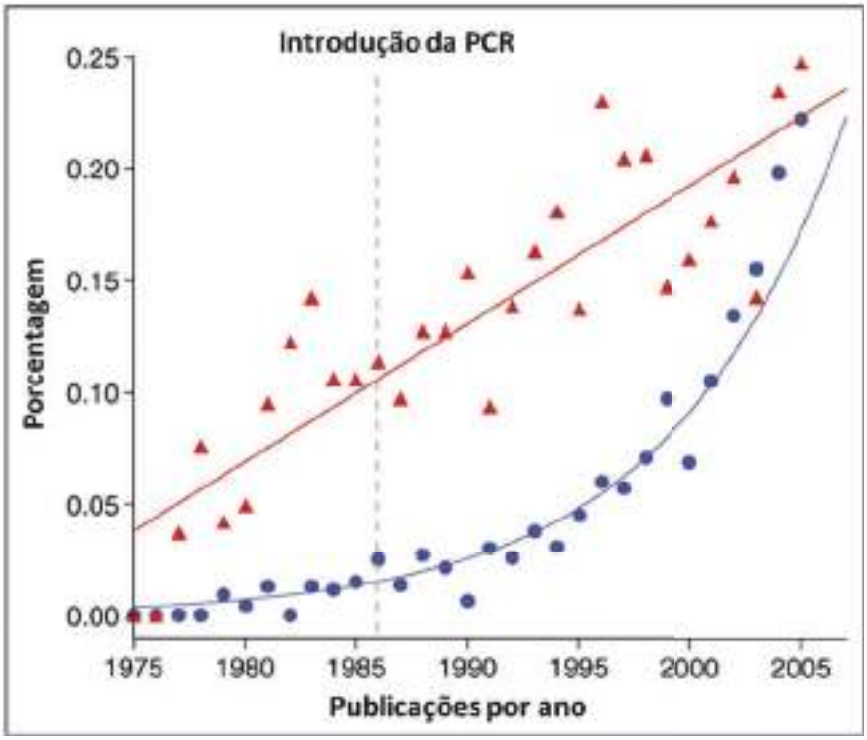


Figura 6. Número de espécies crípticas reconhecidas por ano. Número de espécies crípticas reconhecidas após o advento das técnicas moleculares (Bickford, 2007 – modificado).

- *redução de ambigüidades* – a grande plasticidade (variabilidade) presente nos caracteres utilizados na identificação de espécies, muitas vezes, confunde e até mesmo inviabiliza a atribuição de um espécime à sua espécie correspondente. A molécula de DNA corresponde à uma sequência linear de quatro tipos de nucleotídeos (Adenina, Guanina, Citosina e Timina), tornando o método de identificação mais objetivo e preciso, pois, em determinada posição da molécula de DNA estará presente apenas um dos quatro nucleotídeos possíveis.

- *democratização do acesso a identificação* – o atual número de taxonomistas é muitas vezes menor que o necessário. Para muitos grupos, não há um taxonomista treinado ou este se encontra no exterior. Fatores integrantes do chamado "impedimento taxonômico" apresentado em tópicos anteriores. A maior parte das identificações é realizada por parataxonomistas, indivíduos que possuem algum conhecimento do grupo com o qual trabalham e que com a ajuda de chaves-de-identificação, tentam identificar os espécimes que possuem. Contudo, a identificação por parataxonomistas, frequentemente gera erros, uma vez que os mesmos não são taxonomistas especializados. Assim, o uso de ferramentas moleculares, torna possível a identificação de espécies por qualquer pessoa interessada, uma vez que basta obter uma amostra de DNA, a qual será processada e confrontada com um banco de dados com as sequências de DNA das espécies.
- *acelera a obtenção de dados* – muitos trabalhos científicos dependem grandemente da correta identificação dos organismos de trabalho, um processo normalmente demorado e insatisfatório para muitos grupos devido à falta de estudos, chaves-de-identificação ou até mesmo taxonomistas. Mesmo quando esses existem, o processo de identificação pode demorar. Além disso, muitas chaves-de-identificação usam termos específicos de difícil entendimento para os não-taxonomistas.
- *acelera a compilação da biodiversidade* – com um método de identificação mais rápido e democrático será possível acelerar o processo de obtenção de dados, identificação de espécies conhecidas e novas espécies, contribuindo para a compilação da biodiversidade do planeta.
- *cria a perspectiva de um guia de campo eletrônico* – com o avanço da ciência, poderá ser possível, num futuro próximo, a construção de um dispositivo eletrônico portátil, que permitiria a identificação das espécies em campo. Este aparelho seria capaz de processar as amostras de DNA e comparar as sequências obtidas com o banco de dados, em questão de minutos.

O código de barras da vida - DNA barcoding

Embora ferramentas moleculares tenham fornecido uma ampla gama de novas oportunidades para estudar questões em Biologia Evolutiva (como nos processos de especiação) e em Sistemática Filogenética, o uso de diferentes técnicas e/ou genes por diferentes grupos de cientistas, para diferentes grupos taxonômicos, inviabiliza a criação de um sistema de identificação molecular universal de espécies pelo fato dos dados não serem comparativos. Com isso, em 2003, foi proposto o uso de um único e curto segmento de 648 nucleotídeos da extremidade 5' do gene mitocondrial *Citocromo Oxidase C subunidade I* (*COI* ou *Cox1*), o qual seria suficiente, em muitos metazoários, para identificá-los ao nível de espécie (Hebert et al., 2003a; 2003b), padronizando a técnica. Esta metodologia foi denominada *DNA Barcoding*, em analogia ao sistema de identificação de produtos por códigos de barras. Da mesma forma que um produto é identificado por sua combinação única e exclusiva de código de barras, uma espécie poderia ser identificada por uma combinação única e exclusiva de nucleotídeos desse fragmento de aproximadamente 650 pb do gene *COI* (Hebert et al., 2003) (Figura 7).

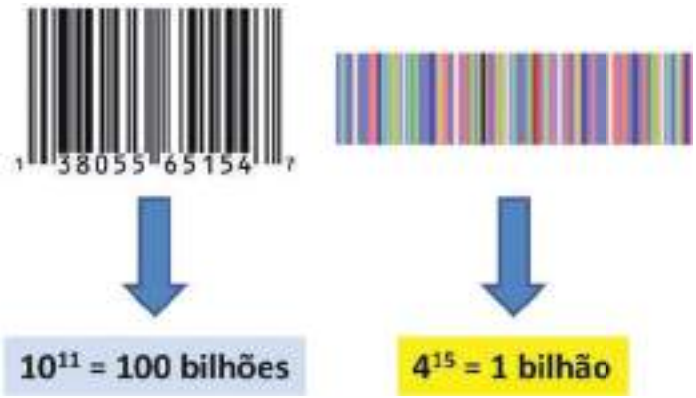


Figura 7. Códigos de barras. Códigos de barras de produtos e genético mostrando sua analogia e número de combinações possíveis.

O principal objetivo da metodologia *DNA barcoding* é a criação de um Sistema Global de Bioidentificação (GBS) que visa automatizar, simplificar, acelerar e democratizar a identificação de espécies através da criação de um banco de dados público, contendo as sequências *barcode* de todos os seres vivos. Secundariamente, o GBS permitiria a sinalização de inúmeras espécies novas, baseadas nas variações genéticas, acelerando assim a compilação de todos os seres vivos do planeta e facilitaria a identificação nos casos de espécies crípticas, microscópicas e de organismos com morfologia complexa ou inacessível.

Fundamentos da técnica

O fundamento básico no qual se baseia a metodologia de *DNA barcoding* está na exploração da variabilidade existente entre as sequências de DNA de um mesmo gene nas diferentes espécies.

O genoma é composto por regiões codificantes (genes = proteínas) e não-codificantes, podendo estas últimas, serem funcionais (DNA ribossômicos, regiões reguladoras de genes, micro-RNAs) ou não-funcionais (introns, DNA espaçador). O mecanismo básico de geração de variabilidade genética é a mutação (mudança de um nucleotídeo na cadeia de DNA). A mutação pode ocorrer por mecanismos intrínsecos moleculares, como erros da enzima DNA polimerase na construção de uma nova fita de DNA durante o processo de replicação do material genético, ou por mecanismos externos tais como radiações ultravioleta e ionizante, agentes químicos ou vírus. As mutações podem ser benéficas (geram uma característica vantajosa), maléficas ou deletérias (geram uma variante desfavorável) ou neutras (não influenciam na aptidão do indivíduo). O mecanismo evolutivo de seleção natural trata de eliminar ou reduzir a frequências mínimas as mutações maléficas ou deletérias, enquanto tende a favorecer os indivíduos que possuem mutações benéficas, as quais podem se acumular, resultando em mudanças evolutivas adaptativas. As mutações neutras, por não influenciarem a aptidão de seus portadores, não sofrem ação da seleção natural, podendo se acumular devido à ação de um segundo mecanismo evolutivo

denominado deriva genética (variação nas frequências de alelos (variantes de um gene) devido a processos aleatórios). Acredita-se que a maior parte das mutações existentes ao longo do genoma das espécies correspondem à mutações neutras. Além disso, o código genético é degenerado, ou seja, existem mais de um códon (sequências de três nucleotídeos que determinam um aminoácido) para um mesmo aminoácido (unidade de construção das proteínas) (Figura 8). Nota-se que as variações nos códons para um mesmo aminoácido ocorrem principalmente na terceira posição do códon (terceiro nucleotídeo da trinca). Logo, se espera que a maior parte das mutações seja encontrada nessa posição, uma vez que as mesmas tendem a ser neutras. De fato, é o observado na maioria dos organismos estudados.

As mutações podem ocorrer tanto em células somáticas (constituintes dos organismos) quanto em células germinativas (gametas), sendo somente essas últimas as responsáveis por passar as mutações para seus descendentes levando a sua fixação na espécie.

Independente do mecanismo gerador das mutações, essas tendem a ocorrer à uma taxa aproximadamente constante, a qual pode variar entre os diferentes segmentos de DNA de acordo com suas características. Regiões genômicas não-codificantes e não-funcionais tendem a acumular mais mutações devido ao fato de não sofrerem ação da seleção natural, enquanto que regiões codificantes e/ou funcionais possuem uma baixa taxa de mutação, devido à eliminação de muitas delas por seleção natural, acumulando-se apenas as mutações benéficas e neutras.

Figura 8. Código genético.

A metodologia *DNA barcoding* utiliza das características descritas acima na identificação das espécies. O acúmulo de mutações ao longo da história evolutiva de cada espécie gera diferenças (variações) nas sequências de DNA. Como as histórias evolutivas são independentes e o processo de mutação é aleatório, cada espécie tende a ter uma sequência de nucleotídeos única e exclusiva à ela, capaz de identificá-la (Figura 9).

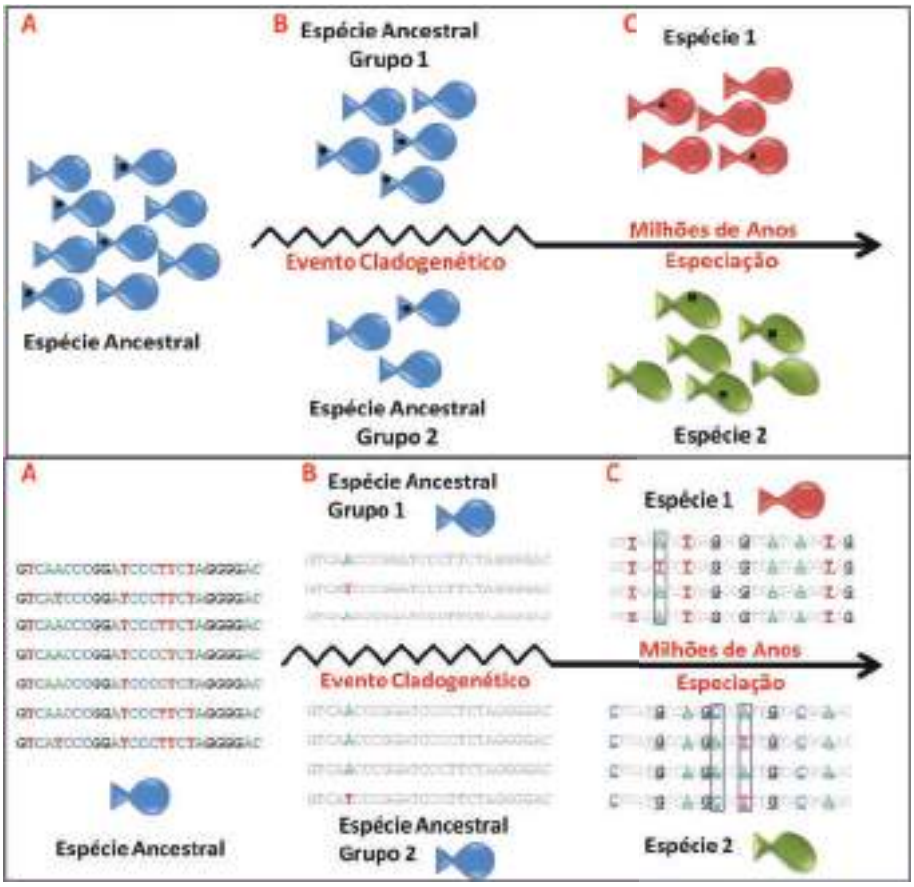


Figura 9. Fundamentos da técnica de DNA *barcoding*. Esquema do processo de acúmulo de diferenças entre espécies. A = do ponto de vista morfológico; B = do ponto de vista genético.

Um segundo fundamento da metodologia *DNA barcoding,* consequente do primeiro, reside na observação de que a variação entre as sequências de DNA entre indivíduos de uma mesma espécie (variação intra-específica) é, na vasta maioria das vezes, significativamente menor que a variação entre as sequências de DNA de espécies diferentes (variação interespecífica), mesmo entre espécies muito próximas do ponto de vista evolutivo (Figura 9). Indivíduos e/ou populações de uma mesma espécie possuem uma mesma história evolutiva, ou seja, existe fluxo gênico entre seus indivíduos (troca de material genético), o que os leva a compartilharem uma maior

homogeneidade genética. Embora variações populacionais existam, elas são muito menores que quando comparadas com as variações existentes entre espécies distintas. No outro extremo, a comparação das sequências de DNA de espécies diferentes revela uma porcentagem de variação significativamente maior. No momento em que se inicia o processo de especiação (processo que leva ao surgimento de uma nova espécie – isolamento reprodutivo) a partir de um ancestral comum, cada nova entidade tende a seguir caminhos evolutivos distintos, resultando no acúmulo de diferenças observadas nos segmentos de DNA (Figura 9). As diferenças observadas entre as variações intra e interespecíficas, na metodologia de *DNA barcoding*, é denominada de *"barcoding gap"*, o qual representa a descontinuidade nos valores destes parâmetros. Contudo, deve-se ter cautela na determinação desses intervalos, os quais podem variar de acordo com o grupo em estudo. Espécies com histórias evolutivas recentes (pouco tempo de especiação) podem ter esse intervalo reduzido ou mesmo com valores sobrepostos. Nestes casos, vale usar mão do conhecimento prévio que se tem do grupo estudado.

Com o descrito acima, a metodologia de *DNA barcoding*, se utiliza de uma sequência de DNA ou conjunto de sequências de DNA muito similares (contempla a variação populacional) para a identificação de espécies. Tais sequências funcionam como um código de barras genético.

Mas qual o tamanho do segmento de DNA a ser utilizado?

O código de barras dos produtos em geral é baseado na combinação de 10 números (0 a 9) em 11 posições diferentes, o que permite a obtenção de 100 bilhões de códigos. No caso do DNA, existem apenas quatro nucleotídeos possíveis (A, T, C e G), porém uma sequência de apenas 15 nucleotídeos produziria um bilhão de combinações (Figura 7), número bem maior que o total de 7 a 10 milhões de espécies estimadas para o planeta (Mora et al., 2011). No entanto, nem todas as posições de uma molécula de DNA podem carregar uma mutação, devido às características deletérias de muitas delas. Como apresentado anteriormente, o código genético é degenerado (Figura 8), permitindo que a terceira posição do códon possa carregar mutações. Assumindo que as mutações ocorram apenas na terceira posição do códon, um segmento de 45 nucleotídeos apresentaria 15 terceiras posições passíveis de carregar mutações, resultando em um bilhão de combinações possíveis. Esses valores representam apenas estimativas, sabe-se que nem todas as terceiras posições sofrem mutações, além do fato de que a taxa de mutação nos segmentos de DNA é variável. Assim, um fragmento de aproximadamente 650 nucleotídeos foi escolhido para proporcionar um excesso de posições nucleotídicas passíveis de carregarem mutações.

A escolha do segmento de DNA

O genoma é composto pelos DNA nuclear e os presentes em organelas como mitocôndrias e cloroplastos. Um segmento adequado para a identificação de espécies deve contemplar pelo menos dois quesitos: possuir uma taxa de evolução adequada e ser de fácil obtenção pelas técnicas existentes. Dentro deste contexto o uso do genoma mitocondrial mostra-se mais adequado. Nos animais, o genoma mitocondrial possui herança exclusivamente materna (provenientes apenas da mãe – presentes no óvulo) e consiste em uma molécula de DNA circular com 16 a 20 mil nucleotídeos, altamente

empacotado com 13 genes codificantes de proteínas, dois RNA ribossômicos, 22 RNA transportadores e uma região conhecida como "região controle" responsável pelo processo de transcrição e replicação da molécula (Figura 10).

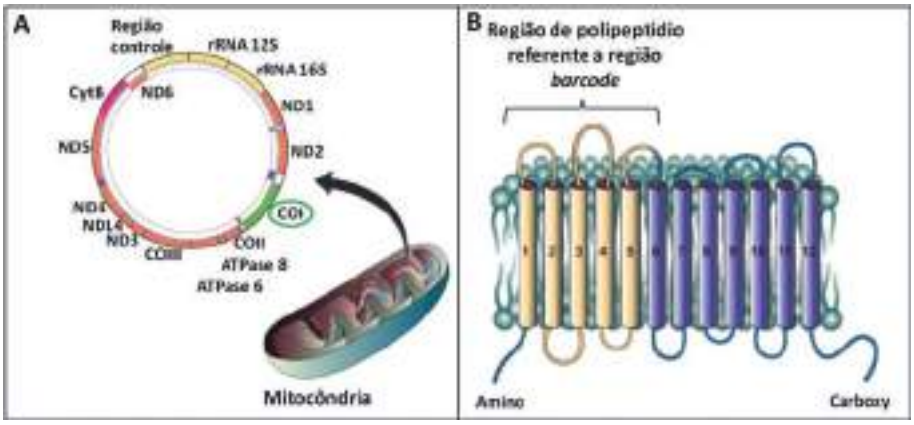


Figura 10. Gene COI. (A) mitocôndria e genoma mitocondrial – em verde, região do gene COI; (B) representação esquemática do polipeptídio codificado pelo gene COI – em destaque a porção correspondente a sequência *barcode*.

O genoma mitocondrial possui uma taxa de mutação relativamente maior que o genoma nuclear. Essa característica é atribuída ao ambiente mitocondrial mais oxidativo. A mitocôndria é responsável pelo processo de respiração celular, o qual produz radicais livres, moléculas altamente reativas e que causam danos na molécula de DNA. Outra característica a favor do DNA mitocondrial é a presença de milhares de cópias por célula. Cada célula possui de centenas a milhares de mitocôndrias, as quais fornecem uma fonte rica de material genético. Além disso, os genes mitocondriais não possuem introns (espaçadores gênicos), os quais, usualmente atrapalham as análises dos segmentos de DNA, pela grande variação existente nesses segmentos. Outra característica altamente favorável é o fato de DNA mitocondrial ser haplóide, ou seja, possui apenas uma cópia do material genético em comparação ao genoma nuclear (diplóide) onde cada molécula de DNA possui duas cópias (uma materna e uma paterna). Assim, eliminam-se possíveis problemas com a presença de diferentes alelos nas análises do DNA.

O segmento de DNA escolhido precisou preencher algumas características, como taxa de evolução e tamanhos adequados e a existência de *primers* universais (iniciadores necessários a síntese *in vitro* de moléculas de DNA). O gene *Citocromo Oxidase C subunidade I* (*COI ou Cox1*) preenchia esses requisitos além de possuir um número considerável de sequências de DNA já depositadas no GenBank (banco de dados de sequências de DNA do *National Center for Biotechnology Information* (NCBI) representantes de diversas espécies. Após análises estatísticas para verificar a capacidade de essas sequências identificarem corretamente as diferentes espécies, o gene *COI* foi escolhido. O gene *COI* codifica uma proteína transmembrana (Figura 10) que participa da cadeia fosforilativa no processo de respiração celular. Esse gene possui

aproximadamente 1500 nucleotídeos. Foi selecionado o primeiro terço de gene como região *barcode* perfazendo um segmento de aproximadamente 650 nucleotídeos da extremidade 5' do gene.

A proposta inicial do projeto *DNA barcoding* era a padronização do processo de identificação através de um único segmento de DNA. No entanto, era esperado que um único segmento de DNA não seria capaz de identificar todas as espécies do planeta. A busca por segmentos alternativos ou associação entre segmentos de DNA eram necessários para os grupos nos quais o *COI* não fosse resolutivo. Assim, para fungos está sendo utilizado além do *COI* um fragmento do espaçador do DNA ribossômico (*ITS*), para alguns grupos de insetos o gene *28S* tem demonstrado um melhor desempenho. No caso de plantas, os genes cloroplastais *matk*, *rbcla* e *rbclb* tem sido utilizados. Para alguns grupos de anfíbios utiliza-se o gene ribossomal *16S*. Já no caso de vírus e bactérias, a identificação tem sido possível através do sequenciamento de todo genoma e da associação de seis a nove genes, respectivamente.

Protocolos

A Figura 11 mostra um esquema geral do processo de obtenção das sequências *barcade*, os quais serão discutidos a seguir:

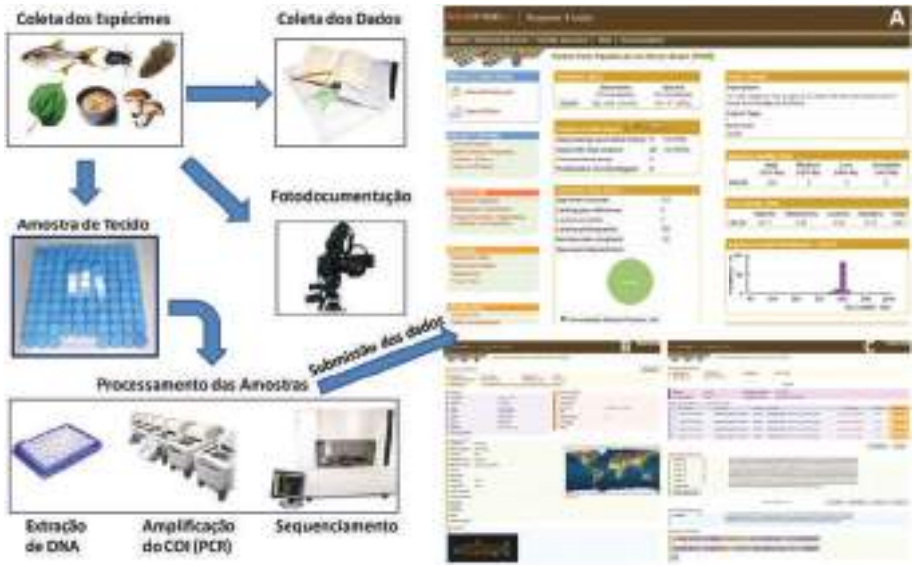


Figura 11. Fluxograma simplificado do processo de obtenção e processamento das sequências *barcode*.

- *obtenção de amostras* – a princípio qualquer tipo de material biológico é passível de análises moleculares. A obtenção de amostras pode se dar junto a natureza, utilizando-se os métodos e apetrechos de coleta adequados para cada grupo de estudo e com autorização dos órgãos responsáveis (ICMBio). Outras fontes de material biológico são as feiras e mercados, dependendo do grupo e

objetivo de estudo. Os museus e coleções biológicas também se caracterizam numa rica fonte de material biológico. No caso de museus e coleções, deve-se atentar ao método de preservação utilizado. Os métodos de preservação que utilizam formol inviabilizam as análises moleculares, pois essa substância degrada o DNA. Os exemplares coletados para a obtenção das sequências *barcode* devem necessariamente ser depositados em alguma coleção biológica como testemunho, exceto em casos específicos. Quando o espécime é de grande porte (ex. baleia) ou representam espécies raras ou ameaçadas de extinção é possível realizar o depósito de um *e-voucher*, que é a representação do indivíduo por fotos e uma coleção de dados coletados que permitam sua identificação. O número recomendado de espécimes que deve ser coletado para a obtenção das sequências *barcode* é de cinco indivíduos, obtidos, preferencialmente, de diferentes regiões/populações para contemplar a possível variação intra-específica existente. No entanto, esse número deve ser encarado como um projeto piloto e se adequar à realidade de cada espécie. Espécies com grande distribuição geográfica e com reconhecida alta taxa de variação populacional devem ser representadas por um número maior de indivíduos a fim de cobrir o máximo da variação intra-específica possível. Contudo, na falta de exemplares, de difícil coleta ou raros, as sequências de DNA *barcode* devem ser obtidas dos indivíduos disponíveis.

- *preservação das amostras de tecido* – após a coleta dos espécimes, uma amostra de tecido deve ser imediatamente retirada, se possível em campo ainda, a acondicionada em tubos devidamente identificados contendo etanol 95-100%. A retirada de amostras tardiamente pode comprometer as análises moleculares devido ao processo de degradação do DNA. No entanto, é possível obter DNA de boa qualidade mesmo a partir dessas amostras, porém não é um procedimento confiável. Alternativamente, se possível, a preservação dos organismos pode ser realizada em campo (Ex.: fixação em etanol), com posterior retirada de material biológico em laboratório. Podem ser utilizados fragmentos de músculo, sangue, pêlo, penas, folhas, flores, o indivíduo inteiro (caso seja pequeno), dentre outros. Deve-se ressaltar que nem sempre é necessário o sacrifício do organismo. Principalmente para espécies ameaçadas de extinção, pode-se retirar um pequeno fragmento de tecido (sem comprometer o organismo) e fazer um registro fotográfico do espécime (em todos os ângulos necessários a sua identificação taxonômica). O material coletado não deve ultrapassar 1/3 do volume final de etanol, para garantir a preservação do mesmo. O material coletado deve, sempre que possível, ser armazenado em freezers -20ºC para garantir a preservação dos mesmos por longo tempo. Amostras preservadas em etanol e mantidas à temperatura ambiente e em boas condições, mantém sua qualidade pelo período de 1 a 2 anos, enquanto que amostras preservadas em etanol e mantidas em freezers podem durar muitas vezes mais.
- *coleta de dados* – deve ser sempre preenchida a ficha de campo e demais fichas pertinentes ao material obtido. Os dados que devem ser obtidos, sempre que possível, são: nome dos coletores; data da coleta; localidade com coordenadas geográficas; altitude; apetrechos de coleta; notas sobre o habitat, microhabitat e associações; sexo e estágio de vida. Posteriormente devem ser incluídas

informações a respeito da coleção biológica de depósito (instituição e número de depósito (*voucher*)) e o nome do taxonomista que procedeu a identificação.

- *fotodocumentação* – sempre que possível deve-se proceder a fotodocumentação do indivíduo ressaltando as características úteis para sua identificação. As fotos devem ser obtidas em alta resolução e seguindo os critérios específicos para cada grupo.
- *obtenção das sequências barcode* – a obtenção das sequências *barcode* seguem os protocolos de rotina de um laboratório de biologia molecular.
  - *Extração de DNA* – a extração de DNA visa a obtenção e purificação das moléculas de DNA a partir de amostras de material biológico. Diversas técnicas de extração estão disponíveis desde o uso de metodologias ditas "caseiras" como a de fenol e salinas até o uso de kits comerciais próprios para extração de DNA. Sempre verifique o método mais adequado para o tipo de amostra a ser utilizada. Independente da técnica adotada preza-se pela obtenção de DNA de boa qualidade.
  - *Amplificação do segmento do gene COI* – estão disponíveis na literatura diversos conjuntos de *primers* para os mais variados grupos de seres vivos para a amplificação do fragmento do gene *COI*. A amplificação (aumento do número de cópias do fragmento de interesse) é realizada pela técnica da "reação em cadeia da polimerase" (PCR). Essa técnica permite a obtenção de milhões de cópias do segmento de interesse a partir de pequenas amostras de DNA inicial. Após o processo de amplificação, a produto da reação passa por procedimentos simples de laboratório para purificação da amostra (eliminação de excessos de reagentes).
  - *Sequenciamento do fragmento do gene COI* – o produto da amplificação, depois de purificado é submetido à técnica de sequenciamento. Existem no mercado, inúmeras plataformas e kits de sequenciamento disponíveis. O processo de sequenciamento, que consiste na determinação da sequência de nucleotídeos de uma molécula de DNA, passa por três etapas básicas. Na primeira é realizada a reação de sequenciamento, procedimento semelhante à PCR, onde são obtidas cópias do fragmento de interesse, só que neste caso, utilizam-se, além dos nucleotídeos normais (desoxinucleotídeos), nucleotídeos marcados com fluorescência (dideoxinucleotídeos - uma cor para cada tipo de nucleotídeo A, T, C e G). A reação é realizada em dois tubos separados, onde em cada um, ocorre a amplificação de apenas uma das fitas de DNA. A segunda etapa consiste na purificação dessas amostras para eliminação de excessos de reagentes. A terceira e última etapa consiste na aplicação das amostras em um sequenciador automático que fará a leitura da sequência de nucleotídeos de cada molécula de DNA. O método de sequenciamento mais utilizado é método de Sanger que se utiliza dos dideoxinucleotídeos marcados com fluorescência descritos acima. Outros métodos de sequenciamento surgiram recentemente, com o objetivo de sequenciar grandes porções do genoma. Tais equipamentos, no caso do *DNA barcoding*, são utilizados no sequenciamento de amostras difíceis de separar, como fungos, algas, bactérias, plâncton, dentre outros, onde não é possível individualizar as amostras para sequenciamento. A evolução dos sequenciadores e dos métodos de sequenciamento tem tornado o processo de obtenção de sequências de DNA cada vez mais confiáveis, rápidos e baratos.

- *Análise dos dados de DNA barcoding* - A análise de sequências de DNA exige o uso de ferramentas específicas pertencentes à área conhecida como Bioinformática. Bioinformática refere-se à aplicação das técnicas da informática na obtenção, processamento e análises de dados biológicos, combinando conhecimentos de química, física, biologia, ciência da computação, informática e matemática/estatística. O processo de análise dos dados *barcoding* seguem, pelo menos, três etapas: 1- triagem e obtenção das sequências consenso; 2 – alinhamento das sequências; 3 – identificação das espécies.
  - *1 – Triagem e obtenção das sequências consenso* – nesta etapa cada sequência de DNA obtida é checada para verificar se há erros de sequenciamento, que podem gerar inserção e/ou deleção de bases ou mesmo códons de parada precoces, uma vez que se trata de uma região gênica. Cada amostra é sequenciada, pelo menos, duas vezes para a obtenção da sequência consenso, que é a construção da sequência *barcode* final baseada na comparação das duas sequências obtidas. Este método auxilia na identificação e eliminação de erros de sequenciamento e aumenta a confiabilidade dos dados. As sequências consenso são então checadas quanto a presença de possíveis contaminantes por comparação com as sequências já depositadas no *GenBank* (NCBI) por meio da ferramenta "*Blast*" (www.blast.ncbi.nlm.nih.gov/Blast.cgi).
  - *2 – Alinhamento das sequências* – o alinhamento consiste em organizar as sequências *barcode* obtidas de modo que as sequências se alinhem de acordo com as posições de cada nucleotídeo, ou seja, a base que ocupa a posição 15, por exemplo, no gene COI está alinhada com as bases de mesma posição nas demais sequências. Isto permite que sejam comparadas regiões homólogas em todas as sequências. Existem diversos algoritmos que fazem este alinhamento, dentre eles o ClustalW (Larkin et al., 2007) e o Muscle (Edgar, 2004). O BoldSystems disponibiliza ferramentas de alinhamento *on line*.
  - *3 – identificação das espécies* – nesta etapa são aplicados ferramentas e métodos analíticos para verificar se as sequências e/ou conjunto de sequências *barcode* são capazes de discriminar corretamente cada espécie. Em outras palavras, verificar se as sequências *barcode* de determinada espécie possuem um conjunto de nucleotídeos que são lhe são exclusivos (gerados ao longo da história evolutiva da espécie por meio de mutações), de forma a permitir sua identificação. Diversos métodos e abordagens são utilizados para este fim, alguns mais simples que permitem uma análise mais básica, outros mais robustos e desenvolvidos especificamente para delimitar espécies a partir de sequências *barcode* que permitem uma análise mais refinada dos dados. A seguir são apresentados apenas alguns dos métodos/abordagens de análises, sendo os três primeiros, os métodos mais clássicos e primeiros a serem utilizados e os três últimos, métodos com maior refinamento de análise.
- *Análise por distância genética* – a análise por distância genética se dá pela mensuração dos valores de divergência genética observados entre as espécies. A análise é realizada por comparações par-a-par, onde a sequência de cada espécime é compara com todos os demais (Figura 12). O resultado final da análise é uma matriz de distância onde é possível visualizar os valores de diferenças entre as espécies. Esse método permite também verificar os valores de divergência

genética intra-específica para se observar a presença e dimensão do *barcode gap*. Além disso, é possível verificar a distribuição dos valores de divergência nos mais diferentes níveis taxonômicos (Figura 13). No entanto, com grandes números de espécimes analisados, a matriz de distância de se torna demasiado grande e de difícil visualização. Assim, esta abordagem é utilizada, principalmente, na análise de pequenos grupos e para a obtenção dos valores médios de divergência genética e sua distribuição. No cálculo das distâncias genéticas convencionou-se a utilização do modelo de evolução de Kimura-2-Parâmetros (K2P) (Kimura, 1980). A escolha desse método se justifica no fato de levar em consideração outros parâmetros pertinentes as características da molécula de DNA, como substituições múltiplas (mutação ocorre mais de uma vez numa mesma posição), transições (mudança entre nucleotídeos de mesma natureza – purina com purina, pirimidina com pirimidina) e transversões (mudança entre nucleotídeos de natureza distintas – purina com pirimidina).

	Hoplosternum littorale	*Callichthys callichthys*	*Trichomycterus maracaya*	*Trichomycterus diabolus*
Hoplosternum littorale	0			
Callichthys callichthys	19,9%	0,3%		
Trichomycterus maracaya	23,8%	23,8%	0,5%	
Trichomycterus diabolus	24,5%	22,2%	8,5%	0

Figura 12. Método de análise por distância genética. Matriz de comparação par-a-par; em preto = valores médios observados entre as espécies (variação interespecífica); em vermelho = valores médios observados entre indivíduos da mesma espécie (variação intra-específica).

	N	Táxons	Comparações	Mín.	Méd.	Máx	Desvio-padrão
Dentro das espécies	1330	229	7815	0	1,3	8,5	0,02
Entre espécies de um mesmo gênero	1354	121	19705	0	7,1	24,9	0,04
Entre gêneros de uma mesma família	1360	33	122906	1,4	20,1	31,5	0,01
Entre famílias de uma mesma ordem	-	-	184686	15,2	23,4	33,4	0,01

Figura 13. Método de análise por distância genética. Valores de divergência genética observados por categoria taxonômica para um conjunto de dados.

- *Análise por dendrogramas – grupamentos recíprocos* – a análise por construção de dendrogramas permite uma representação gráfica dos dados de *DNA barcoding* (Figura 14). A construção de dendrogramas é baseada na matriz de distância genética gerada por K2P através do algoritmo de *Neighbour-Joining*, o qual é utilizado por apresentar ótimo desempenho na análise de grandes conjuntos de dados. A eficácia do método de identificação por *DNA barcoding* pode ser verificada com este método pela observação de grupamentos recíprocos, que é a obtenção de grupos de indivíduos de uma mesma espécie (Figura 14). Esses grupamentos

evidenciam os fundamentos da metodologia de *DNA barcoding*, mostrando os baixos valores de divergência genética intra-específicas e os altos valores de divergência genética interespecíficas.

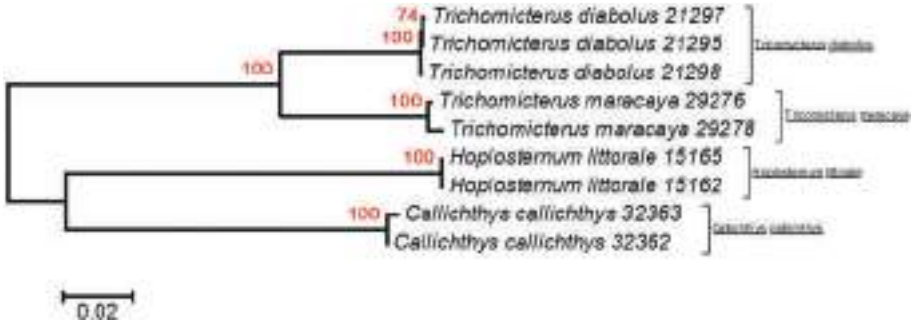


Figura 14. Método de análise por dendrogramas. Dendrograma de Neighbour-Joining mostrando a correta separação das espécies, as quais formam grupos recíprocos. Escala = valor de divergência genética, a qual corresponde aos comprimentos dos ramos. Em vermelho = suporte estatístico aplicado à análise (na metodologia *barcoding* é opcional).

- *Análise por caracteres diagnósticos* –Esta metodologia é a que mais se aproxima do método tradicional (morfológico) de identificação de espécies que usa caracteres exclusivos daquela espécie como diagnósticos na sua identificação. Da mesma forma, essa metodologia identifica a combinação de nucleotídeos que é exclusiva de uma dada espécie, tornando-se assim um caráter/nucleotídeo diagnóstico. Ao final monta-se um mapa mostrando as posições e os nucleotídeos que são diagnósticos para cada espécie (Figura 15).

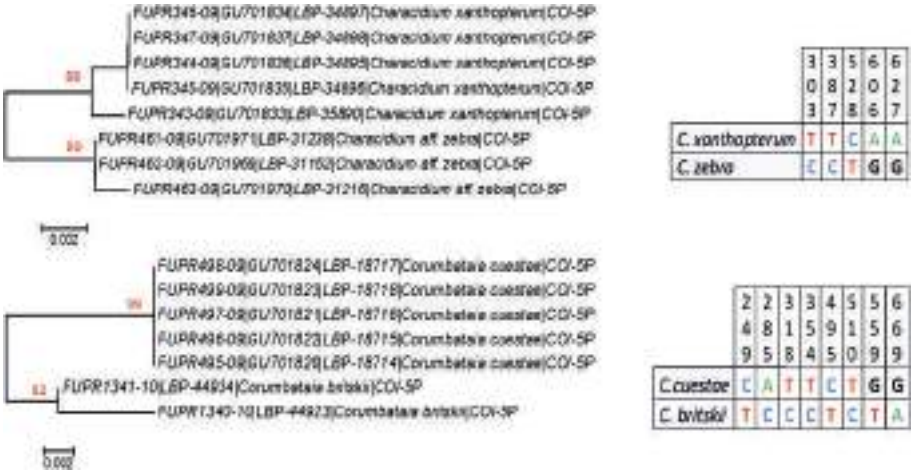


Figura 15. Método de análise por caracteres diagnósticos. Dois exemplos mostrando a utilização de caracteres (bases nucleotídicas) diagnósticos na identificação de espécies. Valores acima das letras = posição na sequência de DNA do gene COI.

- *Análise ABGD (“Automatic Barcode Gap Discovery)* – o método ABDG foi desenvolvido para identificar automaticamente o valor de *barcode gap* dentro de um conjunto de dados, para delimitar espécies, a partir da comparação das divergências genética intra e interespecíficas. A partir do *barcode gap* calculado o programa agrupa as diferentes sequências. Sequências com divergência genética abaixo do valor calculado são agrupadas e sequências com divergência genética acima do valor calculado formam um novo grupo. Ao final cada grupo de sequências representa uma espécie ou possível espécie. De forma resumida, o método calcula o valor de *barcode gap* a partir de uma matriz de distância par-a-par obtida para todas as sequências do conjunto de dados utilizando um intervalo de valores de distâncias genéticas intraespecíficas fornecidos pelo usuário (*default* do programa utiliza valores entre 0,001 e 0,1) a partir de um conhecimento prévio do grupo em estudo. Após plotar em gráfico cada valor da comparação par-a-par, o método detecta o *gap*, reconhecido como um intervalo onde os valores de distâncias genéticas par-a-par apresentam um salto na escala de valores como representado na Figura 16. Este intervalo (“*gap*”) é utilizado como *barcode gap* para todo o conjunto de dados. A partir deste valor o programa faz um primeiro agrupamento das sequências. Porém, como sabemos que o valor de *barcode gap* pode variar de grupo para grupo, o método refaz a matriz de distância par-a-par para cada agrupamento obtido em busca de novos “saltos” nos valores de divergência genética em cada grupo de sequências. Uma vez encontrado, o agrupamento inicial é então subdividido. Este passo é repetido até não se encontrarem mais “saltos” que permitiriam novas subdivisões. Ao final se obtem o número de agrupamentos (espécies) possíveis/prováveis no conjunto de dados. Os desenvolvedores do método mostram que há uma alta correspondência entre os grupamentos obtidos e as espécies previamente estabelecidas, sendo especialmente útil na descoberta/delimitação de espécies crípticas. Para saber mais veja Puillandre et al. (2012).

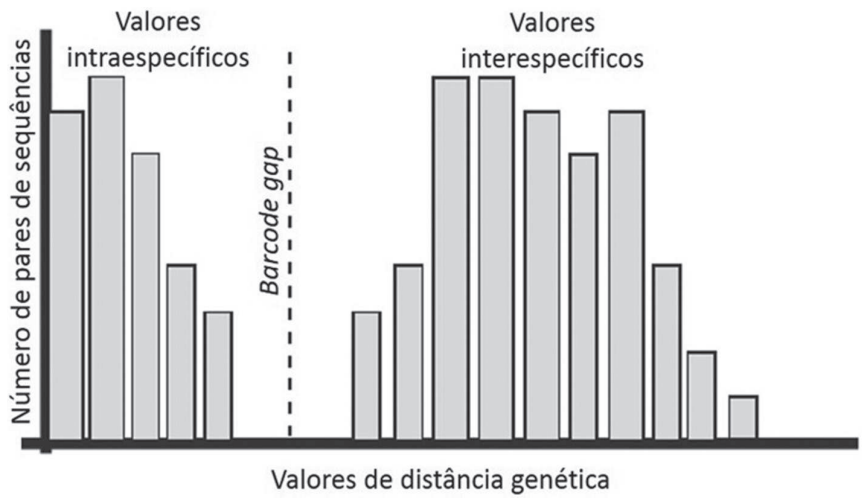


Figura 16. Gráfico ilustrativo do método ABGD na identificação do *barcode gap* (“salto” nos valores de divergência genética par-a-par).

- *Análise GYMC (Generalized Mixed Yule Coalescent)* – este método de análise foi desenvolvido para delimitar espécies que evoluem independentemente por meio da análise de um único lócus (gene ou fragmento de gene). Entende-se como espécies evoluindo independentemente aquelas que acumularam mutações exclusivas em suas sequências de modo a formarem grupamentos genéticos distintos. Este método é bem mais robusto que os anteriores, pois tenta identificar a transição entre eventos intraespecíficos (populacionais) e interespecíficos (especiação) combinando um modelo de especiação (modelo Yule – ver Nee et al., 1994) com um modelo de populações (modelo neutro de coalescência – ver Hudson, 1990) utilizando-se do método estatístico de Máxima verossimilhança para testar a probabilidade do ponto de transição encontrado. Resumidamente, o método utiliza uma árvore gerada a partir da matriz de dados com algum método filogenético que forneça os comprimentos de ramos (métodos bayesianos, por exemplo), sendo que ramos mais curtos representam linhagens mais recentes e ramos mais longos, linhagens mais antigas. De acordo com os modelos empregados, espera-se que os eventos de cladogenese ocorram à uma taxa relativamente constante dentro do grupo, ou seja, a cada período de tempo, o número de linhagens dobra. Se em dado momento, se verifica um aumento brusco no número de linhagens (cladogeneses) em um intervalo de tempo mais curto (ramos curtos), interpreta-se como eventos coalescentes ocorrendo dentro das espécies (Figura 17). Neste ponto, o método identifica a transição entre os processos de especiação (interespecíficos) e populacionais (intraespecíficos). Assim, os nós localizados atrás deste ponto de corte (mais antigos) são reconhecidos como espécies e os nós localizados à frente deste ponto, como populações/linhagens dentro da espécie. Ao final, delimita-se as espécies dentro do conjunto de dados. De modo geral o método é bastante eficaz e confiável, contudo devido ao uso de árvores e testes estatísticos demanda um tempo computacional elevado, podendo ser inviável para grandes conjuntos de dados. Para saber mais veja Pons et al., 2006 e Fujisawa e Barraclouch, 2013.
- *Análise por BIN (Barcode Index Number System)* – Este sistema acoplado ao banco de dados do projeto *barcoding* (BoldSystems) gera uma identificação numérica única (BIN) para cada grupamento de sequências *barcode* depositadas no banco de dados que estão abaixo de um limite de corte, independentemente de sua identificação a priori. De modo geral, o sistema BIN trabalha com o que chamamos de Unidades Taxonômicas Operacionais (OTUs), em última instância espécies, baseados no padrão de similaridades e divergências genéticas. Para delimitar as OTUs, o sistema BIN baseia-se no algoritmo de delimitação de espécies RESL (*Refined Single Linkage*) (Ratnasingham e Hebert, 2013), o qual, de modo resumido, segue os seguintes passos: 1 – as sequências são agrupadas de acordo com um limite de corte estabelecido em 2,2% de divergência genética (este é valor médio de divergência intraespecífica máxima encontrada nos conjuntos de dados depositados no BoldSystems), de modo que todas as sequências abaixo desse valor formam um único grupamento; 2 – os grupamentos iniciais (OTUs) que apresentarem divergência genética menor que 4,4% (o dobro do valor de corte inicial) entre si serão agrupados para uma a busca de um novo valor de corte mais refinado. São feitos agrupamentos usando novos valores de corte entre 1,0-2,4% (em intervalos de 0,2%); 3 – um teste estatístico é aplicado para

verificar o melhor padrão de grupamento obtido, onde se verifica um novo valor de corte para delimitação de OTUs (dentro de cada grupo), obtidos por meio da análise do padrão de divergência genética par-a-par entre as sequências (para saber mais ver Ratnasingham e Hebert, 2013). De modo geral o método é bastante robusto e eficaz e tem um tempo computacional relativamente baixo. Uma vez definidas as OTUs, é atribuído um único BIN para cada, independentemente da classificação a priori. Com isso, é possível atribuir espécimes à uma espécie (OTUs); sinalizar/comparar sequências geradas por diferentes pesquisadores (as vezes com classificação taxonômica distintas) agrupadas sob um único BIN; sinalizar/comparar sequências *barcode* agrupadas sob um único BIN oriundas de regiões geográficas distintas, entre outras. Com isso, por um lado, é possível identificar possíveis novas espécies, especialmente em grupos especiosos, e por outro, sinonimizar outras. Assim, os autores do método acreditam facilitar a resolução de problemas taxonômicos e contribuir com a descoberta e descrição de uma biodiversidade ainda desconhecida. Até a confecção do presente capítulo, já estavam identificadas no BoldSystems 384.080 BINs de espécies animais, mas apenas 157 mil espécies formalmente descritas e identificadas. O sistema BIN também pode ser utilizado para atribuir a sequência de um espécime desconhecido à uma espécie cuja sequência *barcode* (BIN) já está disponível no banco de dados. Para saber mais veja Ratnasingham e Hebert (2013) e www.boldsystems.org. A Figura 18 ilustra os resultados obtidos a partir do sistema BIN.

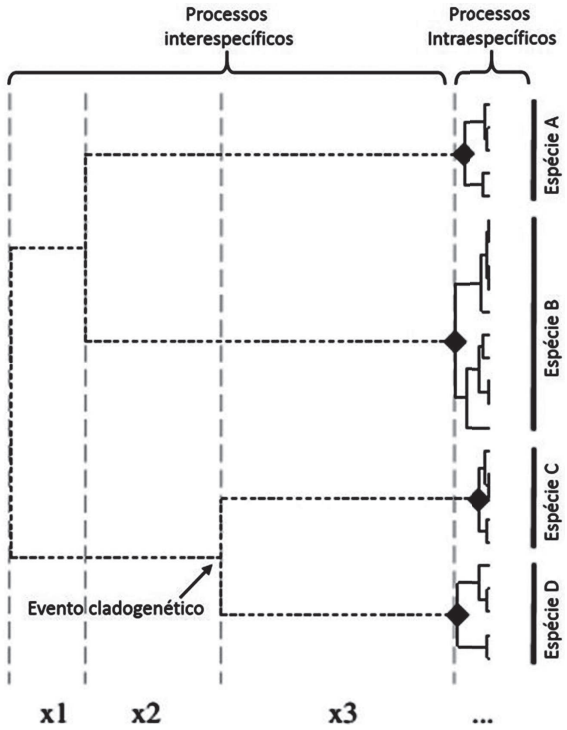


Figura 17. Esquema ilustrativo do método GYMC. Losangos = espécies.

Figura 18. Sistema BIN. 1 = número BIN e seus detalhes; 2 = classificação taxonômica e taxos ligados ao BIN; 3 = comentários; 4 = gráfico com a distribuição das distâncias genéticas dentro do BIN; 5 = publicações; 6 = dendrograma gerado com as sequências dentro do BIN; 7 = rede de haplótipos; 8 = localização das coleções (*vouchers*); 9 = link para dados complementares; 10 = foto da espécie, se disponível; 11 = mapa de localização de cada sequência (organismo) dentro do BIN; 12 = depositório dos espécimes.

Construção de bancos de dados

O uso da metodologia *DNA barcoding* ganhou muita relevância com a criação em 2004 do *Consortium for the BarCode of Life* (CBOL), uma organização internacional devotada ao desenvolvimento da metodologia do *DNA barcoding* como padrão para identificação de espécies. O CBOL conta com membros de 200 organizações de 50 países (ver sítio do CBOL em www.barcodeoflife.org). Em 2007 foi criado um banco de dados para a organização, análise, armazenamento e manutenção das sequências de DNA *barcode* denominado *The Barcode of Life Data System* (BOLD) (Ratnasingham e Hebert, 2007). O BOLD constitui-se numa plataforma de bioinformática integrada que dá suporte a todas as fases de obtenção, análise e validação das sequências *barcode*. Em 2010, foi lançado o *International Barcode of Life project* (iBOL), uma colaboração internacional com 26 países integrantes, dentre eles o Brasil, que visa a obtenção das sequências *barcode* de cinco milhões de espécimes representantes de 500 mil espécies no período de 2010 a 2015 (ver sítio do iBOL em www.ibol.org). Outras iniciativas, integradas ao CBOL e iBOL, visando à obtenção e organização das sequências *barcode* para grupos específicos foram criadas tais como *The Formicidade Barcode of Life* (www.formicidaebol.org), *Bee Barcode of Life* (Bee-BOL - www.bee-bol.org), *All Birds Barcoding Initiative* (ABBI - www.barcodingbirds.org), *Trichoptera Barcode of Life* (www.trichopterabol.org), *Coral Reef Barcode of Life* (www.reefbarcoding.org), *Fish Barcode of Life Initiative* (FISH-BOL - www.fishbol.org), *All Fungi Barcode* (www.allfungi.org), *HealthBOL* (www.healthbol.org), *Lepdoptera Barcode of Life* (www.lepbarcoding.org), *Mammalia Barcode of Life* (www.mammaliabol.org), *Mosquito Barcode Initiative* (MBI), *Marine Barcode of Life* (MarBOL - www.marinebarcoding.org), *Polar Barcode of Life Campaign* (www.ibolproject.org/polar), *Shark Barcode of Life* (www.sharkbol.org), *Sponge Barcode Project* (www.spongebarcoding.org), além de inúmeras outras iniciativas menores (ver em www.ibol.org). No final do ano de 2010, o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) em parceria com Ministério de Ciência e Tecnologia (MCT) e o Fundo Nacional de Desenvolvimento Científico e Tecnológico (FNDCT) iniciaram o financiamento de uma rede brasileira de identificação molecular da biodiversidade denominada BrBoL (www.brbol.org), cujo principal objetivo é gerar os códigos de barra de todas espécies presentes no Brasil.

Operação do banco de dados

O banco de dados BOLD possui uma série de ferramentas e facilidades para auxiliar no depósito e análises das sequências *barcode*. Para o depósito das sequências *barcode,* inicialmente cria-se um projeto junto ao banco de dados. Uma vez criado o projeto, na sua página inicial haverá uma série de ferramentas para a submissão e posterior análise dos dados.

A submissão dos dados consiste de quatro etapas. Na primeira são submetidos os dados referentes ao espécime tais como: número de *voucher;* instituição de depósito e respectivos números de controle; a classificação taxonômica; nome e instituição do taxonomista responsável pela identificação; dados referentes a sexo e estágio de vida; dados de coleta como data, local, coordenadas geográficas, dentre outros. Para a submissão desses dados o BOLD disponibiliza uma planilha para preenchimento e posterior submissão (Figura 19).

Figura 19. Planilha do Excel para submissão dos dados dos espécimes no banco de dados (BOLD). Planilha disponível no site do banco de dados.

A segunda etapa consiste na submissão das sequências *barcode* obtidas. Após serem editadas e analisadas, as sequências devem ser enviadas ao banco de dados, no campo próprio existente na página inicial do projeto, no formato FASTA (sinal de > antes do nome) e identificadas como o número de *voucher* ou do *process ID* gerado (número de entrada do espécime junto ao banco de dados) (Figura 20).

Figura 20. Arquivo FASTA. Arquivo de entrada de sequências de DNA no formato FASTA (>). Arquivo de texto sem formatação (.txt).

Na terceira etapa são enviados ao banco de dados os eletroferogramas (*trace files* – dados brutos gerados pelo sequenciador). O BOLD disponibiliza uma planilha onde são preenchidas informações a respeito dos conjuntos de *primers* utilizados na PCR e na reação de sequenciamento (Figura 21).

Figura 21. Planilha em Excel para submissão dos eletroferogramas (dados brutos) no BOLD. A= planilha para submissão disponível no site do banco de dados; B = exemplo de um eletroferograma.

A quarta etapa, não obrigatória, mas estimulada, consiste no envio das imagens obtidas dos espécimes analisados ao banco de dados através de um formulário próprio (Figura 22).

Figura 22. Planilha do Excel para submissão de fotos/imagens no BOLD. Planilha disponível no site do banco de dados.

Submetidos os dados, o BOLD possui uma série de ferramentas de análise para construção de dendrogramas, cálculo das distâncias genéticas, determinação a composição nucleotídica das sequências (conteúdo de A, T, C e G), alinhamento de sequências, construção de mapas, dentre outras.

Ao final, cada espécime depositado possui duas páginas próprias. A primeira reúne todas as informações pertinentes ao espécime, bem como o mapa de localização, gerado a partir da coordenada geográfica informada e as fotos do espécime (Figura 23). A segunda página reúne todas as informações referentes à sequência de DNA *barcode* depositada, a sequência de DNA propriamente dita, os eletroferogramas e a representação esquemática em código de barras da sequência *barcode* (Figura 24).

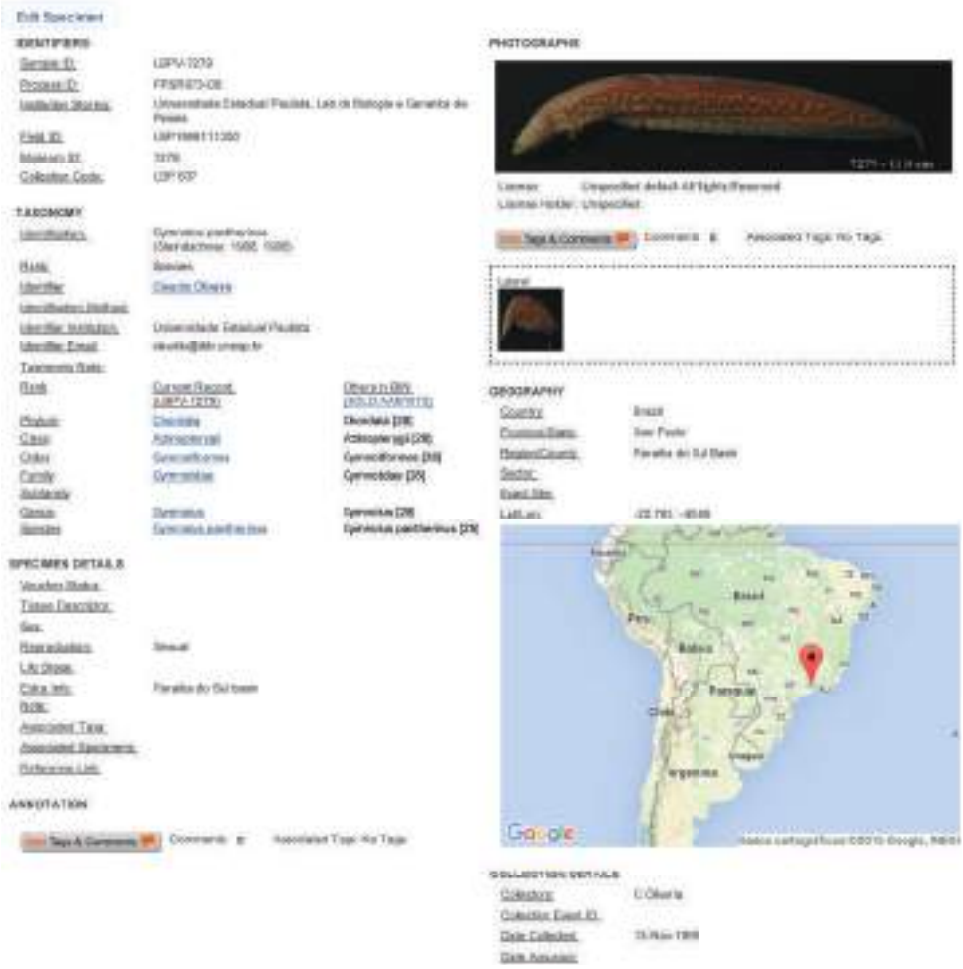


Figura 23. Página do espécime. Página gerada pelo banco de dados reunindo todas as informações cadastradas do espécime.

Figura 24. Página das sequências *barcode*. Página gerada pelo banco de dados reunindo todas as informações referentes às sequências *barcode* geradas e representação gráfica do código de barras genético, para cada espécime.

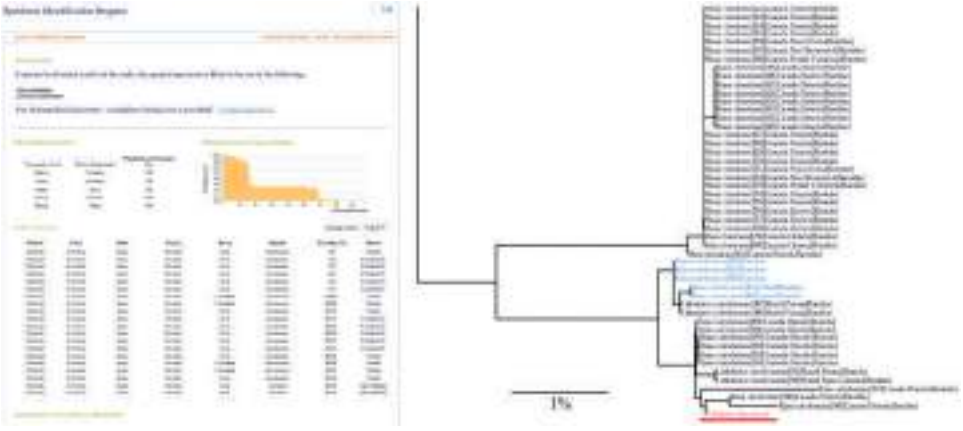


Figura 25. Identificação de espécimes pelo BOLD. Exemplo de identificação de um espécime desconhecido (circulo vermelho no dendrograma) através do banco de dados BOLD. São mostradas até 100 sequências por ordem de similaridade e um dendrograma mostrando a relação entre elas.

Como objetivo final e principal do banco de dados existe uma ferramenta para identificação de espécies. Obtida a sequência do fragmento do gene COI referente a região *barcode* de um espécime de interesse, essa sequência pode ser analisada pela ferramenta de identificação do BOLD (*"identify specimen"*), a qual irá comparar essa sequência como todas as depositadas no banco de dados, e por similaridade irá apontar a espécie a que esse espécime pertence. Caso não haja no banco de dados a sequência *barcode* da respectiva espécie serão apontadas as espécies mais próximas a espécie em análise, por ordem de similaridade (Figura 25).

Validação das sequências barcode

Para garantir a padronização na obtenção das sequências *barcode,* o BOLD criou um protocolo básico a ser seguido. Para o depósito das sequências no banco de dados são necessários apenas um número de identificação e a atribuição de um nome, que pode ser em um nível taxonômico elevado. No entanto, para o reconhecimento formal da sequência como *barcode* de uma espécie é necessário satisfazer sete itens: 1 – nome da espécie (que pode ser provisório); 2 – a existência de um *voucher* com seus dados de catalogação e instituição de depósito; 3 – dados de coleta como nome dos coletores, data e localização com coordenadas geográficas (GPS); 4 – nome de quem identificou o espécime; 5 – a sequência da região *barcode* com pelo menos 500 nucleotídeos; 6 – informações sobre os conjuntos de *primers* utilizados na amplificação e sequenciamento do fragmento de DNA e; 7 – os respectivos eletroferogramas gerados de ambas as fitas de DNA. Após o depósito, as sequências de DNA passam por análises para checar sua confiabilidade. As sequências são verificadas quanto à correspondência ao gene *COI*; à existência de códons de parada (*stop codons*) que poderiam flagrar a presença de pseudogenes; são comparadas com as sequências de possíveis contaminantes (ex. DNA humano, camundongo, bactérias) e; os eletroferogramas passam por uma análise detalhada para estimar a qualidade da sequência de DNA. A exigência desses pré-requisitos fornece um alto grau de confiabilidade às sequências depositadas no BOLD, uma vez que as informações e/ou até mesmo o *voucher* de determinado espécime está disponível e pode ser consultado a qualquer momento.

Aplicações do DNA barcoding

- *Identificação de espécies* - a principal aplicação da metodologia de identificação molecular por *DNA barcoding* é a atribuição de um espécime a uma espécie conhecida. O fundamento do processo de identificação foi apresentado acima. Com isso, espera-se criar um sistema global de identificação molecular que tornará o método mais rápido, preciso, barato e democrático. Diversos grupos de organismos já tiveram suas sequências *barcode* definidas, as quais tiveram sua eficácia demonstrada. Grupos de gastrópodes (Remigio e Hebert, 2003), anelídeos (Chang et al., 2009), colêmbolos (Hogg and Hebert, 2004), lepidópteras (Hebert et al., 2004b; Hajibabaei et al., 2006b), aves (Hebert et al., 2004a; Kerr et al., 2007; Cai et al., 2010), efemerópteros (Ball et al., 2005), aracnídeos

(Greenstone et al., 2005), crustáceos (Costa et al., 2007), peixes (Ward et al., 2005; Hubert et al., 2008; Valdez-Moreno et al., 2009; Ward, 2009; Lara et al., 2010; Ardura et al., 2010; Pereira et al., 2010, Carvalho et al., 2011), dentre outros, já foram analisados, com eficácia acima de 95%. Como exemplo, vamos citar o trabalho de Ward (2009) que analisou os dados disponíveis de 657 espécies de aves e 1088 espécies de peixes no BOLD, os quais apresentaram eficácia de identificação de 93,6% e 97,9%, respectivamente.

- *Sinalização de novas espécies* - outra aplicação importante da metodologia de identificação por *DNA barcoding* diz respeito à capacidade das sequências em sinalizar possíveis novas espécies. A primeira demonstração dessa aplicação veio do estudo de 460 amostras de borboletas da região nordeste da Costa Rica da espécie *Astraptes fulgerator*, descrita em 1775, que demonstrou a existência de pelo menos dez espécies crípticas nessa região (Hebert et al., 2004b), por meio das diferenças encontradas nas sequências de DNA, as quais foram correspondentes à observação de que havia diferentes padrões de coloração nas lagartas dessa espécie. Embora, a identificação de novas espécies não seja o principal objetivo da metodologia de *DNA barcoding*, inevitavelmente ela servirá à este propósito. A identificação de possíveis novas espécies só é possível com base nos mesmos fundamentos utilizados na identificação de espécies já descritas. Como visto, é esperado que o número de mutações (divergência genética) encontradas nas moléculas de DNA do fragmento do gene *COI* sejam muito menores entre indivíduos representantes de um mesma espécie (unidade taxonômica) do que entre indivíduos representantes de espécies distintas, mesmo que proximamente aparentadas. Logo, se um indivíduo analisado apresentar um número de mutações (divergência genética) elevado e não condizente com todas as demais espécies conhecidas do grupo, muito provavelmente esse indivíduo representa uma unidade taxonômica distinta. A observação de indivíduos que apresentam esta condição é altamente esperada devido ao grande número de espécies ainda por se descrever, principalmente dentro de grupos pouco estudados e/ou com reconhecido número elevado de espécies crípticas. No entanto, esses casos devem ser analisados com critério, levando-se em conta a extensão da diversidade genética observada e a biologia e história evolutiva do grupo, para se evitar erros.

Diante desses achados foram propostos possíveis valores de corte para a delimitação de espécies. A primeira proposta sugeriu que este valor deve ser de pelo menos 10 vezes o valor da média de divergência intra-específica observada no grupo em estudo (Hebert et al., 2004b). Outra proposta, mais utilizada, baseada em análises estatísticas, sugere que se um espécime desconhecido apresentar um valor de divergência genética maior que 2% em relação a um espécime conhecido, esse espécime terá 95% de probabilidade de representar uma espécie diferente (Ward et al., 2009). Porém, a atribuição de um valor de corte fixo para delimitação entre espécies deve ser considerada com cautela. Apesar da vasta maioria das espécies já analisadas mostrarem a presença do *barcoding gap*, outras não o possuem bem definido. Assim, o uso do valor de 2% como corte na delimitação de espécies deve ser utilizado como ponto de partida (projeto piloto) nas análises para a identificação de possíveis candidatos a novas espécies. Cada caso deve ser avaliado com cuidado, levando-se em conta os conhecimentos já obtidos da espécie em questão, ajustando-se esse

valor de acordo a historia da espécie. Um dado bastante interessante de se observar é a amplitude de divergência genética observada dentro do grupo em questão, tanto valores intra-específicos quanto interespecíficos, os quais podem servir de mecanismo de calibração para uma aproximação mais real do valor de corte para o grupo em estudo.

Para ilustrar o potencial de aplicação dos dados de *DNA barcoding* na sinalização de novas espécies vamos citar alguns exemplos. Um trabalho que analisou as sequências *barcode* de duas espécies de crustáceos anfípodes de difícil identificação do gênero *Hyalella,* encontrou uma elevada diversidade genética. *Hyalella sandra* e *Hyalella azteca* apresentaram, respectivamente, duas e 33 diferentes linhagens com divergência genética variando de 4,4% a 29,9% (Witt et al., 2006). Até mesmo em grupos amplamente estudados como o das aves, foi possível, por meio da metodologia *barcoding*, identificar novas espécies. Hebert et al. (2004a), analisando 260 espécies de aves norte-americanas, identificaram quatro possíveis novas espécies.

- *Descrição de espécies* – a descrição de espécies pelos métodos tradicionais apresentam algumas limitações como as citadas no texto. Assim o uso associado de outras fontes de caracteres se torna interessante e mesmo necessárias para atestar com confiança o *status* de espécie a um grupo de indivíduos. Dentro desse contexto, o uso das sequências de DNA *barcode* no processo de descrição contribuiria sobremaneira com o conjunto final de dados. Nesses casos, os dados visam complementar o processo de descrição, reforçando a hipótese de uma nova espécie. Os dados de *DNA barcoding* já tem sido utilizados como fonte de caracteres na descrição formal de espécies, os quais representam os primeiros passos no desenvolvimento de uma taxonomia integrativa que reúne diversas fontes de caracteres para a descrição de espécies. Quanto mais caracteres são utilizados, mais robustas se tornam as hipóteses de espécie.
- *Identificação de fraudes* – outra área de aplicação do *DNA barcoding* é na identificação de fraudes. O DNA pode ser obtido, a princípio, a partir de qualquer material biológico, mesmo que processado. Já foram obtidas amostras de DNA de qualidade de produtos enlatados, defumados, assados, secos, salgados, dentre outros. É possível, desta forma, verificar se o produto anunciado corresponde ao produto comercializado. Como exemplo, uma análise realizada com 69 amostras de pescado na Itália, revelou que 32% delas pertenciam a espécies diferentes das anunciadas (Filonzi et al., 2010). O *Food and Drug Administration* (FDA), órgão americano responsável pela fiscalização dos produtos comercializados, passará a utilizar, a partir de 2012, a metodologia de *DNA barcoding* na fiscalização de produtos de origem de frutos do mar (Handy et al., 2011).
- *Análise de dieta alimentar* – como são necessários apenas fragmentos para se obter amostras de DNA suficientes para proceder com o método de identificação molecular, a análise do conteúdo estomacal se torna bastante interessante. É possível determinar com mais precisão a dieta das espécies e construir redes e teias alimentares mais complexas e completas. Aplicação com grande potencial de valorização dos trabalhos de ecologia.
- *Estudo comparativo de organismos* – o acúmulo das sequências *barcode* permitirá o estudo comparado de organismos pertencentes à diferentes regiões geográficas, composições faunísticas/florísticas, populações, dentre outros. Com isso é possível

identificar fauna/flora compartilhada, espécies crípticas, invasão de espécies, distribuição de espécies, dentre outras inúmeras possibilidades.

- *Análises forenses* – identificação de espécies a partir de vestígios como pêlos, fezes, penas, fragmentos de folhas e frutos, sangue, etc.
- *Identificação de pragas invasoras* – a identificação mais rápida e precisa de determinadas espécies de pragas permitirá o seu manejo adequado e a sua erradicação precoce.
- *Descoberta de espécies patogênicas de interesse médico, ecológico e agronômico* – assim como casos das pragas, a identificação rápida e precisa de espécies patogênicas permitirá seu controle precocemente.

Além das aplicações citadas acima, a metodologia de *DNA barcoding* pode contribuir, como vem sendo demonstrado, com a Taxonomia, Sistemática e Genética de Populações. Na taxonomia, o DNA *barcode* pode ser utilizado para identificar espécimes atípicos e contribuir para revisão da nomenclatura de vários grupos, assim como pode ser utilizado como método de rotina para auxiliar na identificação de espécies. Na sistemática o *DNA barcoding* pode servir como ponto de partida para a seleção de táxons e as sequências de DNA obtidas nos projetos de *DNA barcoding* podem ser adicionadas ao conjunto de sequências utilizadas para elaboração de filogenias. Na genética de populações o *DNA barcoding* pode fornecer um primeiro sinal sobre a extensão e natureza das divergências populacionais o que facilitará os estudos comparativos da diversidade de várias espécies.

Limitações da metodologia de DNA barcoding

Como em todas as metodologias existentes, o *DNA barcoding* também apresenta diversas limitações. Entre elas podemos citar:

- *Hibridação de espécies* - casos de hibridação e introgressão representam limitações à metodologia de identificação por *DNA barcoding*. A limitação se encontra no modo de herança do genoma mitocondrial que, normalmente, nos animais é exclusivamente materna. Assim, em casos de hibridação, o espécime seria atribuído erroneamente à sua espécie materna. Embora casos de hibridação sejam reportados para diversos grupos, aparentemente são raros. Como exemplo, em peixes acredita-se que exista menos 1% de casos de hibridação em uma diversidade de aproximadamente 30.000 espécies (Ward et al., 2009). Nos casos de hibridação, a associação com genes nucleares que tem origem de ambos os parentais, é necessária para a correta identificação da espécie. A identificação de híbridos é difícil em muitos grupos, levando a erros de identificação. Nos casos em que há indícios de hibridação, análises mais cautelosas são necessárias.
- *Baixa taxa de evolução do gene COI* - alguns grupos de espécies apresentam baixa taxa de evolução no gene *COI*, muitas vezes com valores de divergência genética tão baixos que impedem sua discriminação. De fato, se sabe que a taxa de evolução do gene *COI* varia entre grupos e até mesmo entre espécies (Krieger e Fuerst, 2002; Frézal e Leblois, 2008; Ward et al., 2009). Em plantas, por exemplo, o uso do *COI* na identificação de espécies é impraticável devido à baixa evolução do gene dentro deste grupo. Este fato também é reportado para alguns grupos

de gastrópodes (Meyer e Paulay, 2005), moscas (Meier et al., 2006), borboletas (Brower, 2006; Wiemers e Fiedler, 2007), cnidários (Hebert et al., 2003; Meyer e Paulay, 2005) e anfíbios (Vences et al., 2005). Nestes casos, o uso de segmentos gênicos que possuem uma taxa de evolução maior são necessários para permitir a correta identificação dessas espécies, os quais podem ser utilizados sozinhos ou em conjunto com o gene *COI* ou mesmo outros genes. O banco de dados BOLD permite a entrada de sequências de outros segmentos de DNA para complementar a identificação de determinados grupos.

- *Grupos com história evolutiva recente* - existem grupos de espécies que evoluíram recentemente e que, embora possam ter taxas de evolução para o gene *COI* semelhantes às encontradas para as espécies com boa resolução em sua separação, não podem ser prontamente identificadas pela metodologia do *DNA barcoding*, por não terem tido tempo suficiente para acumular o número de mutações no gene *COI* necessário para se proceder a identificação. Para a resolução desses casos, o uso de regiões do DNA que evoluem mais rapidamente e/ou o uso combinado de genes que apresentem boa resolução para a identificação desses táxons é necessário.
- *Pseudogenes* - outra limitação apontada no uso do *DNA barcoding* é a existência de pseudogenes no genoma (NUMTs – nuclear mitochondrial DNAs) que são cópias de genes mitocondriais translocadas para o genoma nuclear. A amplificação desses pseudogenes poderia levar a erros de identificação, uma vez que não estando sobre pressão de seleção, acumulam mais rapidamente mutações criando uma aparente divergência genética. No entanto, dificilmente esses pseudogenes são amplificados com os conjuntos de *primers* utilizados devido à possível existência de mutações na região de anelamento do *primer* e, mesmo que amplificados, normalmente são menores que o gene de origem e possuem códons de parada ao longo de sua extensão, sendo facilmente reconhecidos. Porém, análises cautelosas e o uso de ferramentas que permitem identificar esses pseudogenes são necessários para se evitar erros de identificação.

Críticas e controvérsias

Algumas críticas têm sido levantadas a respeito da metodologia de *DNA barcoding*. A princípio, alguns críticos sugeriram que o *DNA barcoding* não seria uma atividade científica porque não visaria testar hipóteses e gerar conhecimento, mas sim simplesmente produzir informações (Lipscomb et al., 2003; Ebach e Holdrege 2005). Entretanto, qualquer experimento gera informações que necessitam ser interpretadas sob a luz de hipóteses e essa é uma atividade científica. Segundo as palavras de Lipscomb et al. (2003) reduzir a taxonomia somente à identificação de espécies a torna uma simples tarefa técnica ao invés de uma ciência baseada em hipóteses. Esse mesmo raciocínio se encaixa perfeitamente nos estudos de *DNA barcoding*, uma vez que esses nunca se limitam a relacionar as sequências encontradas para cada indivíduo, mas sim procuram interpretar as semelhanças e diferenças entre essas sequências e suas relações com as espécies reconhecidas por outros métodos. Assim, é forçoso concluir que taxonomia e *DNA barcoding* são igualmente atividades científicas. Waugh (2007) argumenta também que a aplicação da técnica de *DNA barcoding* serve ainda para testar a hipótese de que as espécies podem ser identificadas utilizando essa técnica e,

no futuro, pode ser uma fonte de dados que gerará outras hipóteses, o que é também uma atividade essencialmente científica.

Uma crítica mais recente, apresentada por Wiemers e Fiedler (2007), diz respeito ao chamado problema de *barcoding gap*. Os proponentes do uso do *DNA barcoding* sugeriram que a diferença genética interespecífica excede a diferença intra-específica de tal maneira que um claro *gap* permitiria atribuir um espécime desconhecido à sua espécie com uma taxa de erro insignificante (Hebert et al., 2004b). Os desvios a essa regra seriam atribuídos a um pequeno número de pares de espécies incipientes, com separação incompleta de linhagens (Hebert et al., 2004b). Como conseqüência, o estabelecimento da quantidade de divergência entre duas amostras acima de um determinado limite iria indicar uma distinção no nível de espécie, enquanto uma diferença abaixo desse limite indicaria uma identidade taxonômica entre as amostras. Além disso, a existência de um *barcoding gap* tornaria possível a identificação de espécies não descritas (Hebert et al., 2004b; Smith et al., 2006). Possíveis erros com essa abordagem incluem falsos positivos e falsos negativos (Wiemers e Fiedler, 2007). Falsos positivos ocorreriam quando populações dentro de uma espécie são muito distintas geneticamente (populações distantes com fluxo gênico limitado ou populações alopátricas com fluxo gênico interrompido) sendo identificadas como entidades distintas. Falsos negativos, por outro lado, ocorreriam quando pouca ou nenhuma variação nas sequências do fragmento de DNA utilizado é encontrada entre diferentes espécies as quais seriam consideradas erroneamente como uma única espécie (Wiemers e Fiedler, 2007). Meyer e Paulay (2005) sugerem que a amostragem insuficiente a nível interespecífico e intra-específico poderia criar, artificialmente, um *barcoding gap*. Os proponentes do *DNA barcoding* argumentam, entretanto, que a principal razão para essa sobreposição seria o pouco conhecimento taxonômico disponível para alguns grupos e a necessidade de revisão taxonômica dos mesmos.

Uma proposição alternativa e extremamente importante em relação ao estudo das sequências geradas nos projetos de *DNA barcoding* foi apresentada por DeSalle et al. (2005). Segundo esses autores, um dos principais problemas com relação à análise dos dados gerados nos projetos de *DNA barcoding* diz respeito ao uso extensivo da construção de árvores por métodos fenéticos (como *Neighbour-Joining*). Eles ressaltam que os equívocos do uso dessa metodologia têm levado a conclusões também equivocadas quanto ao uso do *DNA barcoding*. Segundo os autores, a metodologia taxonômica corrente usa a descoberta de caracteres diagnósticos, independentemente de árvores, para estabelecer sistemas taxonômicos e, principalmente para identificar espécies. Assim, concluem que o uso dos caracteres de DNA em um contexto de diagnose seria muito mais compatível com os processos correntemente empregados em taxonomia, superando muito a abordagem por árvores. Além disso, DeSalle et al. (2005), propõe explicitamente que deve haver uma ponte entre as pesquisas moleculares e morfológicas e que isso deve aprimorar o processo de identificação de espécies.

Outra crítica levantada por oponentes do uso da metodologia de *DNA barcording* diz respeito ao reduzido número de indivíduos amostrados por espécie. As recomendações em curso sugerem que cinco exemplares deveriam ser amostrados de cada espécie procedentes, sempre que possível, de diferentes pontos dentro da área estudada. Rosenberg (2007), em um estudo estatístico sobre capacidade de determinação de monofilia em comparações inter-pares, demonstrou que uma pequena amostra,

de apenas dez indivíduos para cada grupo testado, pode ser suficiente para uma discriminação altamente significativa do ponto de vista estatístico. Considerando que existem grandes diferenças biológicas entre grupos de organismos quanto a esse número mínimo, o emprego inicial de cinco indivíduos pode ser uma escolha metodologicamente viável, principalmente se encararmos essa escolha inicial como um 'experimento piloto'. Nos estudos biológicos há um consenso de que havendo disponibilidade de um grande número de amostras essas devem ser analisadas, mas havendo impedimentos, as análises devem ser feitas com o número possível de amostras.

Estado atual da arte

Independente das dificuldades, limitações e críticas existentes com a metodologia de identificação por *DNA barcoding,* o número de projetos e os depósitos de sequências no banco de dados tem crescido rapidamente (Figura 26). Durante a confecção deste capítulo já se encontravam depositadas no BOLD as sequências *barcode* de aproximadamente 3,9 milhões de espécimes representantes de cerca de 236 mil espécies dos mais diferentes grupos de organismos e regiões do planeta. Destes, aproximadamente 157 mil espécies são de animais, 62 mil de plantas e 17 mil de fungos e outros grupos.

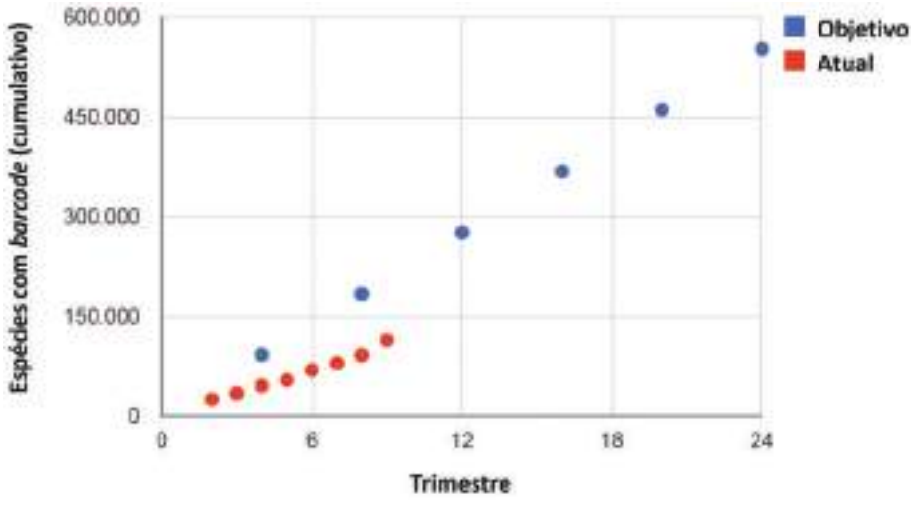


Figura 26. Gráfico do crescimento de sequências *barcode* depositadas no BOLD.

O iBOL (*International Barcode of Life*), maior projeto em andamento, congrega 27 diferentes países representantes de todos os continentes, dos quais o Brasil faz parte. No entanto, mais de 80 nações tem trabalhado no projeto *DNA barcoding*. O projeto prevê investimentos da ordem de 176 milhões de dólares para o desenvolvimento das sequências *barcode,* 120 milhões de dólares investidos no suporte taxonômico e outros 60 milhões no desenvolvimento de novos métodos laboratoriais. O projeto teve início

em 2010 e prevê num período de cinco anos (até 2015) obter as sequências *barcode* de 5 milhões de espécimes representantes de 500 mil espécies ao redor do mundo.

A primeira conferência do *DNA barcoding* ocorreu em 2005 em Londres e reuniu 220 participantes de 44 países enquanto que a mais recente, realizada no final de 2013 na China (na sua quinta edição) reuniu mais de 450 participantes de 43 países.

No Brasil teve início em dezembro de 2010, o projeto intitulado Rede Brasileira de Identificação Molecular da Biodiversidade (BrBoL) financiado pelo Ministério da Ciência e Tecnologia (MCT) e o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Os investimentos são da ordem de 5 milhões de reais, os quais serão utilizados na obtenção das sequências *barcode* e no desenvolvimento de um banco de dados nacional. O projeto prevê o sequenciamento de aproximadamente 120 mil espécimes representantes de 12 a 24 mil espécies. Ao todo são 10 subprojetos dos mais diferentes grupos taxonômicos. O projeto envolve diversas instituições

Figura 27. Mapa do Brasil mostrando a distribuição das instituições envolvidas no projeto BrBoL.

de todas as regiões brasileiras e congrega aproximadamente 200 participantes (Figura 27).

Até o período de elaboração deste capítulo, já se encontravam depositadas e disponíveis no BOLD as sequências *barcode* de aproximadamente 28 mil espécimes, representantes de 4.300 espécies presentes no Brasil, seja dentro do projeto BrBoL ou de projetos independentes.

Perspectivas

Além dos estudos que vem sendo desenvolvidos com o uso da técnica de DNA *barcoding*, alguns pesquisadores têm trabalhado no desenvolvimento dos chamados 'mini-barcodes' (Meusnier et al., 2008). Nesse caso fragmentos entre 120 e 150 pares de bases do gene COI são amplificados. Os experimentos iniciais mostraram que é possível designar *primers* universais para mini-barcodes e que é possível utilizar os fragmentos amplificados na identificação de espécies com bastante segurança (Meusnier et al., 2008). O uso desses mini-barcodes abre a possibilidade de se trabalhar com DNA altamente degradado, como aquele possível de ser retirado de espécimes depositados em museus por muitos anos, ou ainda de material processado, como na industria alimentícia.

Outro ponto extremamente interessante tem sido a recente utilização destes mini-barcodes na análise de amostras ambientais. Nesse caso os pesquisadores têm recolhido amostras de água ou solo e, depois de uma extração total de DNA, feito um sequenciamento em sequenciadores de segunda geração o que permite a obtenção de sequências de um grande número genes presentes nessas amostras. Essa técnica tem também sido referida como 'barcoding ambiental'. Vários trabalhos, com elevado nível de sucesso, têm sido realizados no sentido de desenvolver plataformas capazes de identificar as sequências de mini-barcodes presentes nessas amostras ambientais e buscar nos bancos de dados, como o BOLD, as possíveis espécies presentes nesses ambientes (Bellemain et al., 2010; Hajibabaei et al., 2011; Stern et al., 2011).

Bibliografias

AGOSTINHO, A.A.; THOMAZ, S.M.; Gomes, L.C. Conservacion of the biodiversity of Brazil's inland waters. Conservation Biology, v. 19, n. 3, p. 646-652, 2005.

ANANTHASWAMY, A. Scene set for next mass extinction. New Scientist Mar, v. 27, 2004.

ARDURA, A et al. DNA barcoding for conservation and management of Amazonian commercial fish. Biological Conservation, v. 143, n. 6, p. 1438-1443, 2010.

ARON, S. L.; SOLÉ-CAVA, A. M. Genetic evaluation of the taxonomic status of two varieties of the cosmopolitan ascidian *Botryllus niger* (Ascidiacea: Botryllidae). Biochem. Syst. Ecology, v. 19, p. 271-276, 1991.

AVISE, J. C. Molecular markers, natural history and evolution. New York: Chapman & Hall, 1994.

BALL, S. L. et al. Biological identification of mayflies (Ephemeroptera) using DNA barcodes. J. North Am. Benthol Soc, v. 24, p. 508-524, 2005.

BELLEMAIN E. et al. ITS as an environmental DNA barcode for fungi: an *in silico* approach reveals potential PCR biases. BMC Microbiology, v. 10, p. 189, 2010.

BICKFORD, D. et al. Cryptic species as a window on diversity and conservation. Trends in Ecology and Evolution, v. 22, n. 3, p. 148-155, 2007.

BROWER, A. V. Z. Problems with DNA barcodes for species delimitation: 'ten species' of Astraptes fulgerator reassessed (Lepidoptera: Hesperidae). Systematics and Biodiversity, v. 4, p. 127-132, 2006.

BROWN Jr., K. S; FREITAS, V. L. Lepidoptera, p. 227-243. *In:* BRANDÃO, C.R.F.; CANCELLO, E.M. (Eds.). Biodiversidade do Estado de São Paulo, Brasil. Invertebrados terrestres. São Paulo, FAPESP, XVI + 279 p., 1999.

CAI, Y. et al. DNA barcoding on subsets of three families in Aves. Mitochondrial DNA, v. 21, n. 3-4, p. 132-137, 2010.

CARVALHO, D. C. et al. Deep barcode divergence in Brazilian freshwater fishes: the case of the São Francisco River Basin. Mitochondrial DNA, v. 22(S1): p. 1-7, 2011.

CHANG, C. H.; ROUGERIE, R.; CHEN, J.H. Identifying earthworms through DNA barcodes: Pitfalls and promise. Pedobiologia, v. 52, n. 3, p. 171-180, 2009.

COSTA, F. O. et al. Biological identifications through DNA barcodes: the case of the Crustacea. Canadian Journal of Fisheries and Aquatic Sciences, v. 64, p. 272-295, 2007.

COSTA, L. P. et al. Mammal Conservation in Brazil. Cons. Biol., v. 19(3), p. 672-679, 2005.

COX, C. B; MOORE, P.D. Biogeography, an ecological and evolutionary approach. Blackwell Science, London, 2000.

DASMANN, R. F. A Different Kind of Country. MacMillan Company, New York. ISBN 0-02-072810-7, 1968.

DE QUEIROZ, K. Ernest Mayr and the modern concept of species. PNAS, v. 102, p. 6600-6607, 2005.

DE QUEIROZ, K. Species concepts and species delimitation. Systematic Biology, v. 56, n. 6, p. 879-886, 2007.

DESALLE, R.; EGAN, M. G.; SIDDALL, M. The unholy trinity: taxonomy, species delimitation and DNA barcoding. Philos. Trans. R. Soc. B, v. 360, p. 1905-1916, 2005.

DIAS, B.F.S. Convenção sobre diversidade biológica – CDB. Série Biodiversidade 2. MMA, Brasíla, 2000, 30 p.

EBACH, M. C.; HOLDREGE, C. DNA barcoding is no substitute for taxonomy. Nature, v. 434, p. 697, 2005.

EDGAR, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, v. 32(5), p. 1792-1797, 2004.

WILSON, E., editor, Frances M.Peter, associate editor, Biodiversity, National Academy Press, March 1988. ISBN 0-309-03783-2; ISBN 0-309-03739-5 (pbk.), online edition

FILONZI, L. et al. Molecular barcoding reveals mislabelling of commercial fish products in Italy. Food Research International, v. 43, p. 1383-1388, 2010.

FRÉZAL, L. E.; LEBLOIS, R. Four years of DNA barcoding: current advances and prospects. Infection, Genetics and Evolution, v. 8, n. 5, p. 727-736, 2008.

FUJUSAWA, T.; BARRACLOUGH, T. G. Delimiting Species Using Sing le-Locus Data and the Generalized Mixed Yule Coalescent Approach: A Revised Method and Evaluation on Simulated Data Sets. Syst. Biol., v. 62(5), p. 707-724, 2013.

GREENSTONE, M. H. et al. Barcoding generalist predators by polymerase chain reaction: carabids and spiders. Molecular Ecology, v. 14, p. 3247-3266, 2005.

GUSMÃO, J.; LAZOSKI, C.; SOLÉ-CAVA, A. M. A new species of *Penaeus* (Crustacea: Penaeidae) revealed by allozyme and cytochrome oxidase I analyses. Mar. Biol., v. 137, p. 435-446, 2000.

HAJIBABAEI M. et al. Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. PLoS ONE, v. 6(4), e 17497, 2011.

HAJIBABAEI, M. et al. Design and applicability of DNA arrays and DNA barcodes in biodiversity monitoring. BMC Biology, v. 5 n. 24, p. 1-15, 2007.

HAJIBABAEI, M. et al. DNA barcoding distinguishes species of tropical Lepidoptera. Proceedings of the National Academy of Sciences of the USA, v. 103, p. 968-971, 2006.

HAMMOND, P. Species inventory. In: Groombridge, B. (Ed). Global biodiversity: status of the earth´s living resources. London: Chapman & Hall, 1992, p. 17-39.

HANDY, S. M. et al. A Single-Laboratory Validated Method for the Generation of DNA Barcodes for the Identification of Fish for Regulatory Compliance. Journal of AOAC International, v. 94(1), p. 1-10, 2011.

HARRISON, I. J. et al. A new species of mullet (Teleostei: Mugilidae) from Venezuela, with a discussion on the taxonomy of *Mugil gaimardianus*. J. Fish Biol., v. 71, p. 76-97, 2007.

HEBERT, P.D.N. et al. Biological identifications through DNA barcodes. Proc R Soc Lond B, v. 270, p. 313-321, 2003.

HEBERT, P. D. N. et al. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astrapes fulgerator*. Proc. Natl. Acad. Sci. USA, v. 101, p. 14812-14817, 2004b.

HEBERT, P. D. N. et al. Identification of birds through DNA barcodes. PLoS Biol., v. 2, p. 1657-1663, 2004a.

HOGG, I. D.; HEBERT, P. D. N. Biological identification of springtails (Hexapoda: Collembola) from the Canadian Arctic, using mitochondrial DNA barcodes. Can. J. Zool., v. 82, p. 749-754, 2004.

HOWKSWORTH, D. L.; KALLIN-ARROYO, M. T. Magnitude and distribution of biodiversity. In: HEYWOOD, V. H. (Ed). Global biodiversity assessment. Cambridge University Press, 1995, p. 107-191.

HUBERT, N. et al. Identifying Canadian freshwater fishes through DNA *barcode*s. PLoS ONE, v. 3, n. 6, p. e2490, 2008.

HUDSON, R.R. 1990. Gene genealogies and coalescent process.In: Futuyma D.J., Antonivics J., editors. Oxford Surveys in Evolutionary Biology. Oxford: Oxford University Press. p. 1-44, 1990.

IUCN 2011. IUCN Red List of Threatened Species. Version 2011.2. Disponível em: <http://www.iucnredlist.org>. Acesso em: 15 dez. 2011.

KERR, K. C. R. et al. Comprehensive DNA barcode coverage of North American birds. Molecular Ecology Notes, v. 7, p. 535-543, 2007.

KIMURA, M. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J Mol Evol, v. 16. p. 111-120, 1980.

KÖHLER, F. From DNA taxonomy to barcoding — how a vague idea evolved into a biosystematic tool. Mitteilungen Aus Dem Museum für Naturkunde in Berlin, Zoologische Reihe, v. 83, p. 44-51, 2007.

KRIEGER, J.; FUERST, P. A. Evidence for a slowed rate of molecular evolution in the order Acipenseriformes. Molecular Biology and Evolution, v. 19, p. 891-897, 2002.

LARA, A. et al. DNA barcoding of Cuban freshwater fishes: evidence for cryptic species and taxonomic conflicts. Molecular Ecology Resources, v. 10, p. 421-430, 2010.

LARKIN, M.A. et al. Clustal W and Clustal X version 2.0. Bioinformatics, v. 23(21), p. 2947-2948, 2007.

LEMER, S. et al. Cytochrome *b* barcoding, molecular systematics and geographic differentiation in rabbitfishes (siganidae). C.R. Bioligies, v. 330, p. 86-94, 2007.

LEWINSOHN, T. M. Avaliação do estado de conhecimento da biodiversidade brasileira – volumes I e II. Brasília: MMA, 2005, 520p. (Série Biodiversidade 15).

LEWINSOHN, T. M.; PRADO, P. I. 2005. Quantas species há no Brasil? Megadiversidade, v. 1(1), p. 36-42.

LIPSCOMB, D.; PLATNICK, N.; WHEELER, Q. The intellectual content of taxonomy: a comment on DNA taxonomy. Trends Ecol. Evol., v. 18, p. 65-66, 2003.

MANWELL, C.; BAKER, C.M.A. A sibling species of seacucumber discovered by starch-gel electrophoresis. Comp. Biochem. Physiol., v. 10, p. 39-53, 1963.

MARQUES, A.C.; LAMAS, C.J.E. Taxonomia zoológica no Brasil: estado da arte, expectativas e sugestões de ações futuras. Papéis Avulsos de Zoologia, v. 46, n. 13, p. 139-174, 2006.

MAY, R. R.; HARVEY, P. H. Species uncertainties. Science, p. 323:687, 2009.

MEIER, R. et al. DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. Systematic Biology, v. 55, p. 715-728, 2006.

MENDONÇA, F.F. et al. Identification of the shark species *Rizoprionodon lalandii* and *R. porosus* (Elasmobranchii: Carcharinidae) by multiplex PCR and PCR-RFLP techniques. Molecular Ecology Resources, v. 32, p. 45-47, 2009.

MEUSNIER, I et al. A universal DNA mini-barcode for biodiversity analysis. BMC Genomics, v. 9, p. 214, 2008.

MEYER, C.P., PAULAY, G. DNA Barcoding: error rates based on comprehensive sampling. PLoS Biol., v. 3, n. 12, p. 1-10, 2005.

MORA C. et al. How many species are there on Earth and in the ocean? PLoS Biology, v. 9(8), p. e1001127, 2011.

MOYSÉS, C. B.; ALMEIDA-TOLEDO, L. F. Restriction fragment length polymorphisms of mitochondrial DNA among five freshwater fish species of the genus *Astyanax* (Pisces, Characidae). Genetics and Molecular Biology, v. 25, n. 4, p. 402-407, 2002.

NEE, S.; MAY, R.M.; HARVEY, P.H. 1994. The reconstructed evolutionary process. Phil. Trans. R. Soc. Lond. B, v. 344, p. 305-311, 1994.

NIRCHIO, M. et al. Comparative cytogenetic and allozyme analysis of *Mugil rubrioculus* and *M. curema* (Teleostei: Mugilidae) from Venezuela. Interciencia, v. 32(11), p. 757-762, 2007.

NORSE, E.A.; McMANUS, R.E. Ecology and living resources biological diversity. In Council on Environmental Quality (Ed). Environmental Quality: the eleventh annual report of the Council on Environmental Quality. Council Environmental Quality, Washington, DC, 1980.

PEREIRA, L. H. G. et al. DNA barcodes discriminate freshwater fishes from the Paraíba do Sul River basin, Brazil. Mitochondrial DNA, v. 21(S2), p. 1-9, 2010.

PONS J. et al. Sequence based species delimitation for the DNA taxonomy of undescribed insects. Syst. Biol., v. 55, p. 595-609, 2006.

POOK, C. E.; MCEWING, R. Mitochondrial DNA sequences from dried snake venom: a DNA barcoding approach to the identification of venom samples. Toxicon., v. 46, p. 711-715, 2005.

PROVETE, D. B. et al. Anurofauna do noroeste paulista: lista de espécies e chave de identificação para adultos. Biota Neotropica, v. 11(2), http://www.biotaneotropica.org.br/v11n2/en/abstract?identification-key+bn01111022011, 2011.

PUILLANDRE, N.; LAMBERT, A.; BROUILLET, S.; ACHAZ, G. ABGD, Automatic Barcode Gap Discovery for primary species delimitation. Molecular Ecology, v. 21, p. 1864-1877, 2012.

RATNASINGHAM, S.; HEBERT, P. D. N. A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System. PLOS ONE, v. 8(8), e66213, 2013.

RATNASINGHAM, S.; HEBERT, P. D. N. BOLD: the barcode of life data system (www.barcodinglife.org). Molecular Ecology Notes, v. 7, p. 355-364, 2007.

REMIGIO, E. A.; HEBERT, P. D. N. Testing the utility of partial COI sequences for phylogenetic estimates of gastropod relationships. Mol. Phylogenet. Evol., v. 29, p. 641-647, 2003.

ROSENBERG, N. A. Statistical tests for taxonomic distinctiveness from observations of monophyly. Evolution, v. 61, p. 317-323, 2007.

SAVAGE, J.M. Systematics and the biodiversity crisis. BioScience, v. 45, p. 673-679, 1995.

SHAFFER, H.B.; THONSOM, R.C. Delimiting species in recent radiations. Systematic. Biology, v. 56, n. 6, p. 896-906, 2007.

SILVANO, D.L.; SEGALLA, M.V. 2005. Conservação de anfíbios no Brasil. Megadiversidade, v. 1(1), p. 79-86.

SMITH, M. A. et al. DNA barcode reveal cryptic host-specificity within the presumed polyphagous members of a genus of parasitoid flies (Diptera: Tachinidae). Proceedings of the National Academy of Sciences, USA, v. 103, p. 3657-3662, 2006.

SOULÉ, M. E.; Wilcox, B. A. Conservation Biology: An Evolutionary-Ecological Perspective. Sinauer Associates. Sunderland, Massachusetts, 1980.

STERN, R. F. et al. Environmental barcoding reveals massive dinoflagellate diversity in marine environments. PLoS ONE, v. 5(11), p. e13991, 2011.

VALDEZ-MORENO, M. et al. Probing diversity in freshwater fishes from Mexico and Guatemala with DNA barcodes. Journal of Fish Biology, v. 74, p. 377-402, 2009.

VENCES, M. et al. Deciphering amphibian diversity through DNA barcoding: chances and challenges. Phil. Trans. R. Soc. B, v. 360, p. 1859-1868, 2005.

WAKE, D.B.; VENDREBURG, V.T. Are we in the midst of the sixth mass extinction? A view from the world of amphibians. PNAS, v. 105, p. 11466-11473, 2008.

WARD, R. D. DNA *barcode* divergence among species and genera of birds and fishes. Molecular Ecology Resources, v. 9, p. 1077-1085, 2009.

WARD, R. D. et al. DNA barcoding Australia´s fish species. Phil. Trans. R. Soc. B, v. 360, p. 1847-1857, 2005.

WAUGH, J. DNA barcoding in animal species: progress, potential and pitfalls. BioEssays, v. 29, p. 188-197, 2007.

WIEMERS, M.; FIEDLER, K. Does the DNA barcoding gap exist? - a case study in blue butterflies (Lepidoptera: Lycaenidae). Frontiers in Zoology, v. 4, p. 1-16, 2007.

WILSON, E. O. The diversity of the life. Harvard Univ Press, Cambridge, MA, 1993.

WITT, J. D. S.; THRELOFF, D. L.; HEBERT, P. D. N. DNA barcoding reveals extraordinary cryptic diversity in an amphipode genus: implications for desert spring conservation. Molecular Ecology, v. 15, p. 3073-3082, 2006.

WOESE, C. R.; FOX, G. E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl Acad. Sci. USA, v. 97, p. 8392-8396, 1977.

15

Biologia sintética

Paulo Adriano Zaini
Mateus Schreiner Garcez Lopes

Introdução e bases conceituais

Biologia Sintética (BS) é uma nova área de pesquisa que une Engenharia com as Ciências Biológicas buscando o desenvolvimento de sistemas biológicos que realizem novas funções de maneira robusta. Para isso, visa também a criação de métodos e tecnologias para tornar a bioengenharia mais fácil, segura e padronizada. Através dessa nova abordagem o campo também está desvendando mecanismos das funções biológicas naturais e a organização dos sistemas biológicos, com o objetivo de entender, aprimorar e desenvolver novos sistemas biológicos. Estamos vivendo o início de uma era com a produção de fármacos mais baratos, de novos materiais e combustíveis renováveis, de novas terapias para tratamento de doenças e de construção de novas formas de vida (Figura 1). A habilidade de engenheirar sistemas biológicos se tornou possível devido a descobertas cientificas que culminaram no desenvolvimento de disciplinas como engenharia metabólica e a biologia sintética. Biologia sintética é o "desenho e construção de novos sistemas biológicos que não existem na natureza através de montagem de componentes bem caracterizados, padronizados e reutilizáveis". Engenharia metabólica é amplamente definido como "o desenvolvimento de métodos e conceito para a análise de redes metabólicas, tipicamente com o objetivo de engenheirar fábricas celulares" (Nielsen et al., 2014).

Sistemas biológicos são caracterizados como sistemas genéticos complexos (celulares ou não) que interagem de maneira dinâmica baseados em reações de retroalimentação não lineares que possibilitam propriedades de auto-organização,

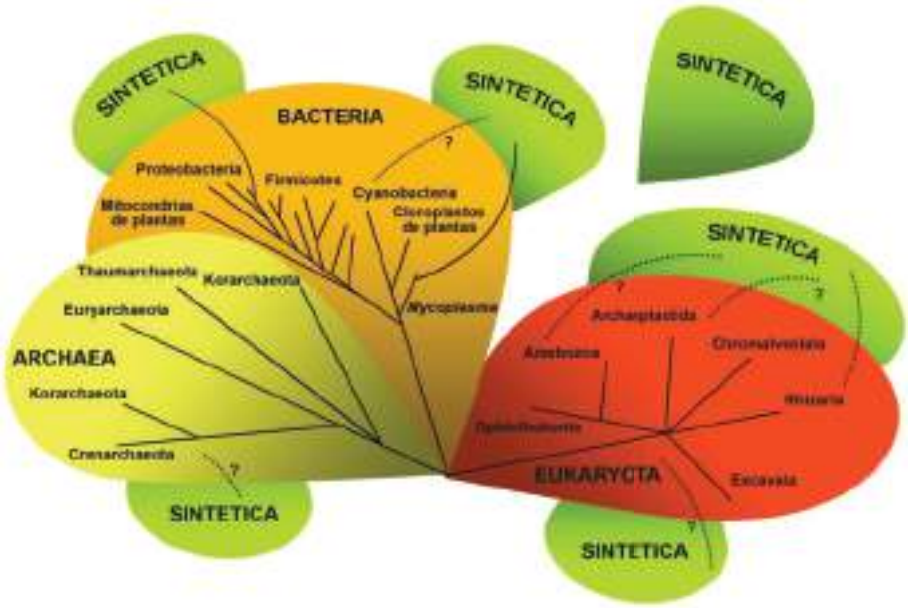


Figura 1. Nova árvore da vida. A Biologia Sintética permite a criação de novas formas de vida. Em seu estágio inicial essa prática ainda é restrita a algumas bactérias como *Mycoplasma* e *Escherichia coli*. Estas servem como "chassis" para a implementação de novos programas genéticos, de interesse industrial/farmacêutico. Novos organismos sintéticos feitos a partir de cianobactérias deverão ser utilizados em breve na produção de biocombustíveis e novas aplicações poderão surgir a partir de diferentes formas de vida naturais e até mesmo completamente sintéticas. A vida sintética é polifilética, ou seja, não tem uma única origem em comum.

evolução, reparo e replicação. Dessa maneira, tais sistemas apresentam um desafio enorme para abordagens racionais de engenharia e se destacam pelo enorme potencial de geração de novas tecnologias mais eficientes, renováveis e/ou de baixo custo. Para possibilitar a criação de sistemas biológicos complexos, foi necessária a criação de uma nova abordagem para a modificação genética denominada BS, que está baseada na:

- Padronização e Abstração. A padronização da maneira como genes ou elementos regulatórios serão clonados, montados e identificados é necessária para garantir que partes individuais ou a combinação das partes que codificam para uma função definida possam ser testadas e caracterizadas a fim de se construir sistemas mais complexos que funcionem de maneira previsível e robusta. Níveis de abstração hierárquica são uma invenção humana para possibilitar a engenharia de sistemas complexos ao ignorar detalhes desnecessários. Atualmente não se faz programação de computadores utilizando códigos binários (0-1), mas com linhagens de programação. Um engenheiro civil não precisa de se preocupar com o desenho de tijolos ou detalhes da instalação elétrica por que já existem especificações definidas para isso. Da mesma maneira, seria quase impossível a engenharia de sistemas biológicos complexos se fosse necessário escrever o sistema utilizando

uma cadeia de nucleotídeos (A-C-T-G) ou se cada parte ou gene precisassem ser caracterizados para a construção de um novo sistema. Dessa maneira, a abstração hierárquica, padronização e linguagens de programação permitem ao designer de um sistema biológico ignorar os detalhes de implementação e focar apenas nos problemas principais da complexidade do sistema.

- Automação do Bio-Design (*Bio-Design Automation*). À medida que bioengenheiros buscam desenhar sistemas com maior complexidade e interações com componentes e novos materiais genéticos, métodos de design serão necessários para gerenciar o desenho e garantir o funcionamento das construções. Para isso, algoritmos, modelos biofísicos e softwares estão sendo desenvolvidos para, por exemplo, (i) o desenho e construção de novas partes e sistemas biológicos, como a plataforma *Genome Compiler*, (ii) o redesenho de sistemas biológicos para novas funções úteis e (iii) combinar estratégias experimentais *in vivo* e *in silico*. Além disso, novos equipamentos de automação de processos (e.g. sequenciamento e tecnologias de larga escala) estão sendo desenvolvidos para o aumento de escala de experimentos visando a redução de custos e tempo.

Dessa maneira a BS procura estabelecer na Biotecnologia um campo conceitual proveniente das Engenharias, baseado nos princípios da padronização, modularidade e modelagem. Disciplinas modernas de engenharia desenvolveram metodologias e processos robustos para lidar com a complexidade crescente de sistemas engenheirados. Esses princípios chave promovem a interoperabilidade entre sistemas, reuso de componentes e o gerenciamento de soluções complexas usando softwares. Uma meta da BS é promover a aplicação desses princípios no desenho de sistemas biológicos (Endy, 2005). Dessa maneira, a BS está interconectada com diversas áreas das ciências biológicas ao criar uma base de conhecimentos destinada a aplicar princípios de engenharia na construção de sistemas biológicos. Essa caixa de ferramentas criada pela BS pode ser utilizada, como será descrito ao longo do capítulo, por um engenheiro computacional interessado em criar dispositivos computacionais em uma bactéria, ou um engenheiro molecular que quer produzir novos químicos renováveis, ou um geneticista buscando construir uma bactéria a partir de simples elementos químicos.

Uma boa analogia para entender os objetivos, os métodos e princípio de abstração da BS pode ser traçado entre a hierarquia da ciência da computação (Andrianantoandro et al., 2006) (Figura 2). Dentro dessa hierarquia, cada parte constituinte está envolvida dentro um sistema mais complexo que define um contexto de atuação. O desenho de um novo sistema que apresente as características desejadas se inicia no fim da cadeia porém a implementação se inicia no início da cadeia (Figura 2). Em seu nível fundamental a BS é baseada em partes biológicas, sendo o DNA a linguagem de programação, as partes biológicas são sequências de dados (AGCTA...) que possuem funções determinadas, por exemplo, uma sequência de DNA que codifica para uma proteína que confere uma cor verde para a célula. As partes padrão de DNA precisam ser caracterizadas e funcionalmente previsíveis, de forma análoga como computadores modernos dependem de componentes eletrônicos confiáveis para seu funcionamento.

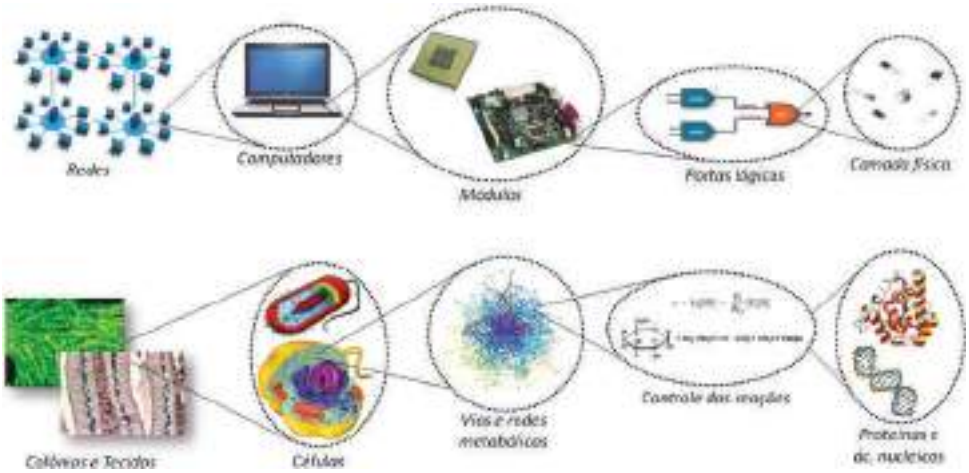


Figura 2. Representação esquemática da analogia entre a Biologia Sintética e a Computação. Ambos os sistemas possuem vários níveis de complexidade para gerar, armazenar e transferir a informação. Modificado a partir de Andrianantoandro et al., 2006.

No segundo nível estão os dispositivos sintéticos que basicamente são compostos por partes biológicas, capazes de processar sinais. Processam *inputs* em *outputs*. Temos como exemplo as reações metabólicas que controlam o fluxo de informação no sistema e manipulam o processo físico. Por exemplo, na presença de um sinal "oxigênio" a célula de levedura é capaz de transformar glicose em dióxido de carbono e água, na ausência do sinal "oxigênio" a levedura transforma glicose em etanol, água e dióxido de carbono. Os biólogos sintéticos estão interessados em utilizar tais sistemas de controle para processar os sinais recebidos (*inputs*) e as funções realizadas pelos dispositivos (*outputs*). O controle do fluxo da informação na célula está relacionado com sinais de reconhecimento de promotores gênicos, reguladores de transcrição, interação de proteínas, etc.; sendo que estes mecanismos de regulação podem ser utilizados para a criação de dispositivos sintéticos equivalentes as portas lógicas baseadas na lógica booleana que realizam funções computacionais como AND e NOR. Em um exemplo básico, uma resposta celular (*output*) que necessite da ativação da expressão de dois genes ao mesmo tempo seria o equivalente ao comando AND; por outro lado, se a resposta celular depender de apenas um dos sinais exclusivamente, teremos uma função NOR.

No terceiro nível temos sistemas sintéticos que são um conjunto de dispositivos sintéticos capazes de capturar sinais, processarem informações e realizarem funções determinadas, como por exemplo, uma célula capaz de captar sinais do ambiente e decidir se irá realizar uma determinada função como combater uma célula tumoral ou produzir determinado metabólito, da mesma forma como um sistema computacional de controle de uma indústria. Por último, existe a arquitetura sintética em que cada célula possui um sistema sintético diferente, sendo necessário uma arquitetura para que esses sistemas possam trabalhar em conjunto para realizar determinada função.

Em outro exemplo, temos células carregando os mesmos sistemas e que necessitam de sincronismo para processarem os sinais de forma eficiente. Nos dois casos, é necessário dominar mecanismos robustos de comunicação célula-célula. Neste sentido, a analogia é traçada com computadores que trabalham em rede para maximizar sua capacidade de processamento de sinais.

Os conceitos da BS podem ser aplicados em praticamente todas áreas da biologia molecular, biotecnologia, medicina e engenharia genética. Porém existem algumas áreas que se destacam como sendo próprias da BS:

- Uma das metas mais emblemáticas da BS é a **reconstrução de células** a partir de simples elementos químicos. Muito já foi feito para a identificação das partes básicas para construção de uma célula e para a síntese de genomas completos. Porém, a construção de microrganismos desenhados especificamente para realizar determinadas tarefas ainda é um objetivo a ser alcançado.
- **Programação genética**: a habilidade de programar células irá possibilitar o controle sobre as condições, da dinâmica e da coordenação das funções celulares. A programação genética consiste na utilização de circuitos genéticos para realizar uma operação computacional, ou seja, na construção de sistemas biológicos para o processamento de sinais e informação. Muito trabalho ainda tem que ser feito para o desenho de circuitos robustos que possam ser facilmente conectados.
- **Construção de novos códigos genéticos**: tem como objetivo desenvolver novas e mais eficientes linguagens de programação para sistemas biológicos. Ou seja, códigos genéticos diferentes do RNA ou DNA. Além disso, por motivos de biossegurança, estes novos códigos genéticos poderiam representar uma barreira entre os sistemas sintéticos e os naturais, evitando a interação entre sistemas engenheirados e organismos naturais (Pinheiro et al., 2012). Além disso, novos códigos podem apresentar características mais adequadas para a construção de novos sistemas biológicos.
- **Desenvolvimento de novas rotas metabólicas**: já estamos vendo o desenvolvimento de diferentes rotas metabólicas para a produção de compostos químicos, biocombustíveis e fármacos a partir de fontes renováveis (Atsumi et al., 2008; Yim et al., 2011; Ro et al., 2006). Porém, muito ainda será feito em direção da substituição de combustíveis fosseis ou no desenvolvimento de novos produtos/processos. Uma outra área de aplicação de novas rotas bioquímicas é os desenvolvimentos de microrganismos capazes de degradar compostos recalcitrantes (biorremediação) na natureza.
- **Desenho auxiliado por computador** (*Computer aided design* – CAD): a escala da engenharia genética está crescendo e novos métodos terão que ser desenvolvidos para integrar diferentes aspectos do desenho celular (Salis et al., 2009). Por exemplo, novos algoritmos terão que ser desenvolvidos para otimizar a síntese de DNA e processos de montagem de DNA, modelos de análise cinética de fluxos metabólicos, expressão gênica e modelos para o desenho racional de enzimas.

Durante este capítulo pretendemos contextualizar e explicar estes diferentes tópicos.

Origens, bases de conhecimento e eventos históricos

Nos últimos 50 anos, descobertas em biologia molecular, genômica e proteômica permitiram a identificação de muitos componentes e processos celulares chave. Podemos considerar o amanhecer da era da Tecnologia do DNA Recombinante as primeiras descrições de **enzimas de restrição** e **DNA ligases** no final da década de 1960. Esses avanços coincidiram com o entendimento do funcionamento de **plasmídeos** e a partir de então era possível manipular qualquer sequência de DNA, removendo-a de um contexto genético e inserindo-a em outro. Não demorou muito para a **clonagem de genes** e introdução de material genético exógeno (**transformação de DNA**) tornarem-se rotina, criando assim protocolos de manipulação genética que foram gradualmente integrados em ferramentas, abrindo as fronteiras para a produção de **proteínas recombinantes** de interesse biotecnológicos, como por exemplo a insulina humana em 1977. Desde a década de 60 também já se sabia da importância do controle da expressão gênica, pioneiramente demonstrada para as enzimas envolvidas no aproveitamento de lactose, conhecido como **operon Lac** (Figura 3). Esse circuito gênico codifica para três enzimas responsáveis pela captação e degradação de lactose. Além das enzimas, outros componentes regulatórios compõem o operon Lac, como o a proteína repressora e outras sequências regulatórias que respondem aos sinais ambientais como a abundância de substratos e produtos. A partir desse exemplo, vários outros sistemas regulatórios foram caracterizados, abrindo o campo de estudo sobre o controle da expressão gênica.

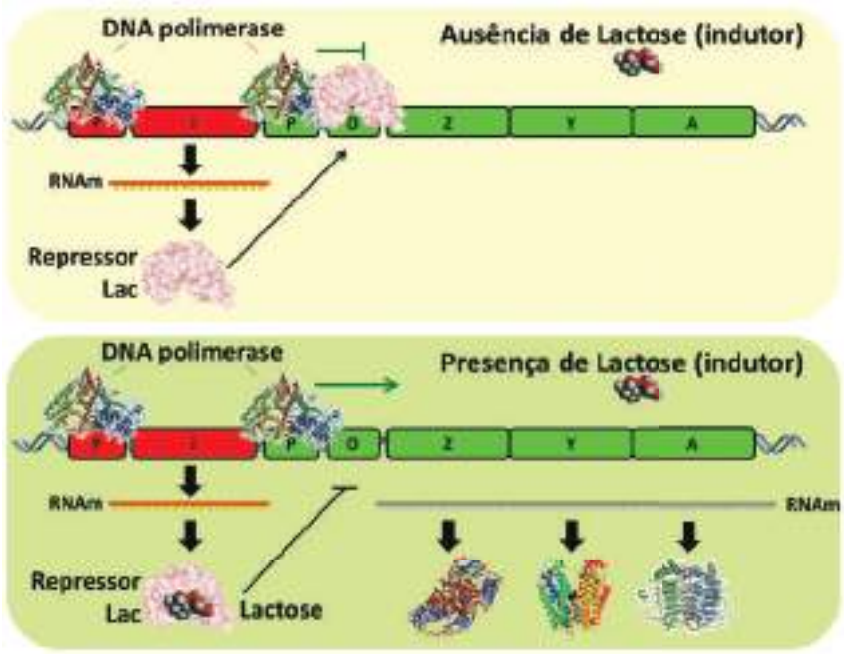


Figura 3. Representação esquemática da regulação do operon lacZ. O repressor Lac (LacI) impede a síntese de enzimas de aproveitamento de lactose (LacZ, LacY e LacA) enquanto houver glicose, causando o uso sequencial de glicose e lactose.

Esse ajuste fino do nível transcricional mostrou-se fundamental e contribuiu de forma decisiva para o desenvolvimento de novos dispositivos e sistemas genéticos que seriam utilizados décadas mais tarde em projetos de BS. Paralelamente, também na década de 60 outros estudos formaram os alicerces dessa área de conhecimento, como a elucidação do **código genético** e os **estudos estruturais de proteínas** que juntos permitiram então o início das atividades de **desenho racional de ligantes** para alvos específicos e todo o desenvolvimento da engenharia de proteínas. Mas somente com o método de **sequenciamento de DNA** desenvolvido por Sanger em 1977 e o advento da **Reação em Cadeia da Polimerase (**PCR**)** em 1983 por Mullis que a Tecnologia do DNA Recombinante realmente deslanchou e tornou-se praticada em laboratórios comuns em todo mundo. As empresas passaram a oferecer **vetores de expressão** (plasmídeos) de proteínas recombinantes e **oligonucleotídeos customizados** para a amplificação das sequências codificadoras e reguladoras, permitindo aos pesquisadores **manipularem sequências e fragmentos de DNA** em um novo patamar de controle. Não tardou muito e a curiosidade dos pesquisadores foi além de simples genes e passaram a estudar a influência de um sobre o outro e seu **funcionamento em conjunto**. Nascia assim a **Biologia de Sistemas** que rumou naturalmente para perguntas em nível genômico. A Biologia de Sistemas, criou um arcabouço de conhecimento para a montagem de modelos matemáticos e biofísicos aplicados a sistemas biológicos. Desde a sua criação, no início dos anos 2000, um progresso significante tem sido feito em tecnologias para a medição de dados celulares e sua análise computacional para a criação de mapas e modelos do funcionamento celular. Esses modelos podem ser utilizados de diversas maneiras, seja para entender o funcionamento de sistemas complexos, simulação de hipóteses (e.g. o efeito da deleção de um gene específico), diagnóstico molecular (e.g. determinado padrão pode indicar a severidade de um câncer ou prever o desenvolvimento de uma doença) e o estudo de células tronco (e.g. para o estudo da transição de estados na diferenciação celular).

No início da década de 90, diferentes projetos de **sequenciamento genômico** foram iniciados, criando um novo marco nessa era, o de conhecimento da sequência de DNA completa de um organismo. Para tal foi necessário treinar **mão de obra qualificada** para gerar e manipular as bibliotecas de DNA a serem sequenciadas. Para se ter uma ideia, em abril de 2003, quando o consórcio público internacional que sequenciou o primeiro genoma humano chegou oficialmente ao fim, a imensa iniciativa tinha consumido 13 anos de trabalho de centenas de cientistas de ao menos 18 países (Brasil inclusive) e estimados USD 2,7 bilhões. O volume de dados e trabalho exigiu o desenvolvimento da **informática aplicada a questões genômicas** para filtragem e montagem dos dados subsequentes até que se chegasse no produto final da sequência de todos cromossomos de um organismo.

Por volta da virada do século as **tecnologias em larga escala** (BOX HTT) já utilizadas pela indústria farmacêutica tornaram-se mais populares e acessíveis e foram incorporadas em muitos projetos de pesquisa, incluindo equipamentos de manuseio e análise de soluções e ensaios celulares versáteis como Citometria de Fluxo e Microfluídica (Figura 4), que estão se tornando pilares tecnológicos em BS, como a síntese e montagem de DNA, devido ao seu emprego em etapas de controle de qualidade e seleção de variantes, por exemplo. As perguntas em nível genômico

tornaram-se acessíveis e o estudo em larga escala da expressão gênica, de proteínas e de metabólitos (**transcritômica**, **proteômica** e **metabolômica**, respectivamente) e seus derivados passaram a ser amplamente empregados, permitindo o entendimento do funcionamento de células e organismos em um outro patamar.

BOX – HTT: tecnologias em larga escala

Uma limitação para a otimização de sistemas biológicos é a falta de completo entendimento de como os dispositivos sintéticos impactam a célula hospedeira. Dessa maneira, métodos experimentais e ferramentas computacionais que permitam analisar o estado celular, incluindo mRNA, proteínas e metabólitos são fundamentais para o desenvolvimento da BS. Por exemplo, um microarranjo de DNA (*DNA microarray*), ou *DNA-chip*, consiste num arranjo ordenado de fragmentos microscópicos de DNA ligados à uma superfície sólida. Cientistas podem utilizar um microarranjo para analisar o nível de expressão de milhares genes simultaneamente, através de hibridização com amostras biológicas (RNAm na forma de cDNA). Utilizando tal técnica é possível distinguir quais genes são mais ou menos expressos em uma célula normal em relação a uma célula cancerosa, por exemplo.

Por outro lado, outros tipos de análise podem indicar o estado fisiológico da célula através da análise de fluxos metabólicos. O metabolismo celular consiste em milhares de genes, enzimas e metabólitos que convertem nutrientes em intermediários celulares e energia. Dessa maneira, através da análise de fluxo utilizando uma fonte de carbono marcado com o isótopo C_{13} é possível quantificar as respostas de alterações experimentais no fluxo das redes metabólicas utilizando espectrofometria de massa para detectar os padrões de marcação em aminoácidos. Além do desenvolvimento de novas técnicas, nos últimos anos aumentou a nossa capacidade de realizar estes experimentos de forma automatizada, padronizada e precisa. Além disso, a diminuição do volume (e.g. de ml para ×l) dos experimentos irá permitir a diminuição dos custos à medida que é necessário menos reagentes, espaço e energia para realiza-los. Os avanços em robótica e sua integração com sistemas de inteligência artificial já permitem o desenvolvimento de equipamentos capazes de varrer de forma organizada e documentada muitos parâmetros/condições em busca de variantes mais eficientes para a função desejada. Essa automação aumentou em várias ordens de grandeza nossa capacidade de selecionar novos agentes antimicrobianos, por exemplo. Os equipamentos são capazes de operar em ampla gama de condições de temperatura e manusear pequenos volumes de maneira precisa, podendo ainda serem acoplados a outros equipamentos analíticos como HPLC e de espectrometria de massa. Mais recentemente tiveram sua funcionalidade ampliada por equipamentos que analisam células, como citômetros de fluxo e microscópios acoplados às câmaras de fluxo. Estas podem ser customizadas de acordo com o tipo e função celular a ser estudada, geralmente construídas de material polimérico sobre um molde de silício obtido por fotolitografia e após a montagem o fluxo é controlado por uma bomba de infusão. Com esses aparatos é possível varrer uma enorme quantidade de mutantes rapidamente e selecioná-los para posterior caracterização.

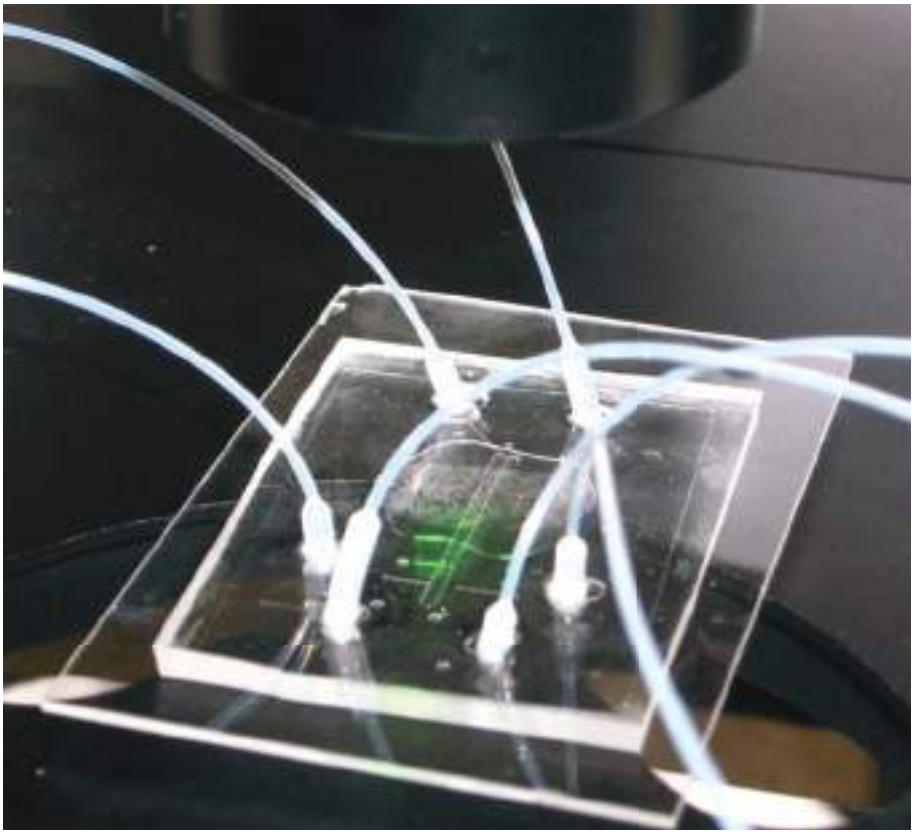


Figura 4. Câmara de microfluídica usada em estudos celulares sob condições de fluxo. Fonte: http://nextbigfuture.com/2007/07/dna-synthesis-costs-and-projections.html

O custo da síntese e sequenciamento de DNA foram lentamente vencidos e a partir do início do século evoluíram em projeção superior à da Lei de Moore, viabilizando a exploração em larga escala de **genomas inteiros** e até mesmo comunidades (**metagenomas**). Em 2014 a empresa Illumina rompeu um marco ao oferecer o sequenciamento de um genoma humano por menos de USD 1.000. Aparatos portáteis com conexão USB como o MinION® prometem análise de 100 milhões de pares de base em 6 horas no conforto da sua casa e utilizando o seu próprio laptop por 900 dólares. Verifica-se dessa maneira que sequenciadores automatizados com velocidade e custos cada vez menores possibilitam oferecer análise genética a um número cada vez maior de consumidores, iniciando a chamada "genômica pessoal", um ramo da genômica que está interessado no sequenciamento e análise de genomas individuais, afim de identificar, por exemplo, perfis individuais de risco a determinadas doenças ou de características ancestrais. O *Personal Genome Project* iniciado pelo Prof. George Church da Universidade de Harvard tem o objetivo de sequenciar e publicar o genoma completo juntamente com o histórico médico de 100.000 voluntários.

BOX – Síntese de DNA

A habilidade de sintetizar grandes quantidades de DNA de uma maneira rápida, barata e acurada terá um grande impacto no tamanho e complexidade de redes de genes sintéticos e irá permitir a síntese de genomas completos com facilidade. Métodos que utilizam PCR para a montagem de pequenos fragmentos têm sido utilizados com sucesso para construir sequencias de mais de 14 kb (Stemmer et al., 1994). Por exemplo, um cluster de genes contínuos de 32 kb contendo os genes de síntese de poliquetídeos de *E. coli* foi construído através da criação de segmentos de 500 pb com PCR (Kodumal & Santi, 2004). Porém, mutações que podem comprometer a expressão gênica podem surgir durante a síntese dos oligonucleotídeos e durante a PCR. Dessa maneira, procedimentos de correções de erros são necessários para construir sequências grandes e de alta fidelidade. Uma estratégia recente utiliza chips de microarranjo foto-programável para construir uma biblioteca de oligonucleotídeos para a construção de grandes sequências de DNA e outra biblioteca de oligonucletídeos para correção de erros (Tian et al., 2004). Os oligos de correção hibridizam com os oligos corretos de construção, enquanto que os oligos incorretos não hibridizam e são lavados do sistema. A combinação dessa estratégia e com a síntese de DNA irá permitir a produção de sequências com megabases de alta fidelidade. Nossa capacidade de síntese de DNA em grande escala progride para viabilizar a fabricação de qualquer desenho de módulo ou sistema gênico sintético.

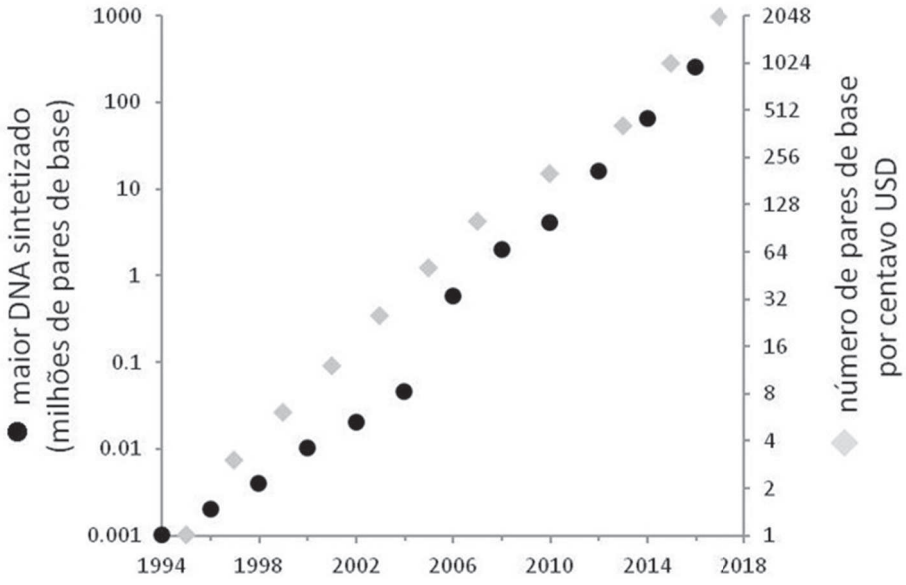


Figura 5. Gráfico da evolução da síntese de DNA. Tanto a capacidade de síntese de sequências cada vez maiores quanto a queda no preço de síntese crescem em ritmos impressionantes. Fonte: http://nextbigfuture.com/2007/07/dna-synthesis-costs-and-projections.html.

Em 2002 foi **sintetizado o primeiro genoma completo** (Cello et al., 2002), do Poliovirus, sendo que, nessa época, o sequenciamento de genomas já era corriqueiro. Hoje os custos de síntese para o consumidor final (empresas e cientistas) estão em torno de 0,20 USD/pb mas os custos reais de síntese caem ainda mais rápido (BOX Síntese de DNA com gráfico da evolução da síntese de DNA).

Finalmente os elementos necessários para o surgimento da BS estavam em mãos e a necessidade de combustíveis, materiais e fármacos de fontes alternativas exerceram grande pressão para o rápido desenvolvimento e popularização da área no início do século XXI. Desde então já era perseguido por diferentes grupos o repertório mínimo de genes necessário para perpetuar e executar a informação genética, o chamado "**genoma mínimo**". A japonesa Kao Corporation, por exemplo, desde 2001 trabalha com "fábricas celulares mínimas (*Minimum Cell Factories*), construídas a partir de *Bacillus subtilis*. Em 2006 começaram com "fábricas celulares refinadas" com funções mais aplicadas para cada projeto. Enquanto isso Craig Venter e seu grupo foram gradualmente removendo partes e sistemas gênicos da bactéria *Mycoplasma genitalium* e após uma década conseguiram reduzir de 482 para 382 o repertório mínimo para essa espécie. Posteriormente, desenharam um **genoma** in silico, sintetizaram esse **genoma** in vitro e conseguiram introduzi-lo intacto em uma célula (chamada de "**chassi**") e fazer com que esse substitui-se o genoma parental, criando a primeira **célula sintética** capaz de se propagar (Lartigue et al., 2007). Lembrando que existe uma certa controvérsia de chamar esse feito de célula sintética uma vez que a parede e maquinaria celular não foram sintetizadas pelos cientistas *ab initio* mas sim aproveitadas da célula recipiente. De qualquer forma, é só o início de uma onda de **organismos sintéticos** que farão parte de nossa vida daqui pra frente. Evidentemente o enorme potencial desse feito trouxe muitas preocupações, como discutiremos adiante no tópico sobre **Bioética e Biossegurança**.

A evolução da base conceitual e domínio técnico já possibilita o desenvolvimento de técnicas complexas como o **MAGE** (*Multiplex Automated Genome Engineering*), pela qual é feita a engenharia genômica múltipla de forma automatizada. Esta técnica é capaz de construir e analisar bilhões de linhagens recombinantes em semanas, fornecendo mutantes com capacidade produtiva ou um *fitness* superior aos obtidos com a tradicional e laboriosa Engenharia Genética tradicional, restrita a um único ou pequeno grupo de genes (Wang et al., 2011). Dessa maneira, se inicia a preencher uma lacuna entre, capacidade analítica, de sequenciamento e análise de genomas com a capacidade de construção de novas linhagens geneticamente modificadas. Torna-se claro a importância dos métodos adotados das Engenharias para construção de uma base conceitual e de técnicas para o desenvolvimento da BS**.** Dessa maneira, embora o termo "biologia sintética" já fosse usado desde 1974 por Szybalsky, sua base conceitual e metodológica precisou avançar bastante para que finalmente fosse posta em prática a partir do início do século XXI.

Biobricks e a secretaria de partes padrão

Um objetivo subjacente da BS é tornar o processo de engenharia de sistemas biológicos mais fácil. Dentro desse processo, tem-se procurado desenvolver e definir

partes biológicas padrão, para garantir que as peças biológicas produzidas possam ser montadas com facilidade e trocadas entre os biólogos sintéticos pelo mundo de uma forma unificada e organizada, garantindo um desenvolvimento mais rápido e transparente da BS.

As **partes biológicas padrão** são sequências de ácidos nucléicos que codificam uma função biológica, que foram refinadas para se adequar a algum padrão técnico de montagem definido. Nesse sentido, um paralelo interessante pode ser feito com um circuito elétrico, em que os seus componentes (ex. diodo, resistência, interruptores, tomadas, etc.) precisam ter o mesmo tipo de conexão para poderem ser montados na placa, apesar de serem produzidos por fabricantes diferentes. O padrão técnico de composição física de montagem de peças biológicas que ganhou muito adeptos inicialmente na comunidade de BS são os *BioBricks*, ou BioTijolos em português (Knight, 2003). A inovação principal dos BioBricks é a padronização do modo de montagem dessas partes padrão: de maneira que a montagem de dois BioBricks resulta em um objeto que também é um BioBrick e pode ser combinado com qualquer outro BioBrick (Figura 6). Detalhes adicionais que definem partes BioBrick são desenvolvidos pela The BioBrick Foundation[1]. O objetivo de ter partes padronizadas é, além de facilitar a montagem e troca de partes, é que as partes individuais ou a combinação das partes que codificam para uma função definida possam ser testadas e caracterizadas a fim de se construir sistemas mais complexos que funcionem de maneira previsível e robusta. Para entender como sintetizar um BioBrick veja mais detalhes na Figura 6. Apesar do padrão BioBrick ter se tornado popular, atualmente continua em uso principalmente por competidores da iGEM (ver adiante) e projetos acadêmicos, enquanto outros padrões de montagem, como Gibson, Golden Gate, SLI e CPEC, têm surgido para possibilitar a construção de sistemas mais complexos de forma eficiente. Uma descoberta que abriu novas perspectivas para a edição de genomas sem deixar marcas de seleção é o sistema CRISPR-Cas, inicialmente descrito como um mecanismo de proteção de células procarióticas contra a invasão de DNA exógeno (Barrangou et al., 2007). Esse sistema foi rapidamente engenheirado para editar objetivamente genomas procarióticos (Jiang et al., 2013) e eucarióticos (DiCarlo et al. 2013). Esse sistema funciona com uma enzima Cas9 capaz de clivar o DNA, que é direcionada ao sítio desejado por uma molécula complementar de RNA, que guia a endonuclease. Dessa forma é teoricamente possível editar qualquer genoma, em qualquer sítio desejado. O enorme potencial biotecnológico dessa ferramenta certamente terá grande impacto na BS.

BOX – Construindo um *BioBrick*

É recorrente a analogia entre os *BioBricks* e as peças de Lego, já que a montagem de duas partes resulta em uma terceira parte que ainda é um BioBrick, da mesma forma que a montagem de duas peças de Lego resulta em um terceiro Lego. Além disso, estas peças podem ser combinadas para construção de dispositivos como contadores e interruptores, da forma que juntamos peças de Lego para a construção de miniaturas de carros, casas e naves espaciais.

Para construir um BioBrick utilizando o método de PCR é necessário possuir algum molde de DNA do qual o BioBrick possa ser amplificado (por ex., um DNA genômico contendo o gene de interesse). Para o desenho dos iniciadores (*primers*) é necessário adicionar na extremidade 5' as sequências referentes aos sítios das enzimas de restrição presentes nos BioBricks, os sufixos e prefixos. A construção de BioBricks contendo sequências codificadoras de proteínas requer um sufixo e um prefixo um pouco mais especializados por duas razões: (i) o prefixo é alterado para garantir o espaçamento entre o sítio de ligação do ribossomo e o códon ATG de início, (ii) BioBricks que codificam proteínas possuem, por padronização, dois códons de sequência TAA de parada. Ao construir os primers, basicamente, é necessário "copiar e colar" determinadas sequências de DNA (31 bp) no fim 5' no seu primer *upstream* (ou iniciador universal) desenhado para o seu fragmento de DNA de interesse. Para mais detalhes, veja o protocolo disponibilizado para a fabricação de BioBricks (*Constructing a BioBrick part via PCR. http://syntheticbiology.org/BioBricks/Part_fabrication.html | BioBricks.org*). Nesse mesmo site, você encontra informações de como produzir pequenas sequências para utilizar como promotores ou sítios de ligação com o ribossomo.

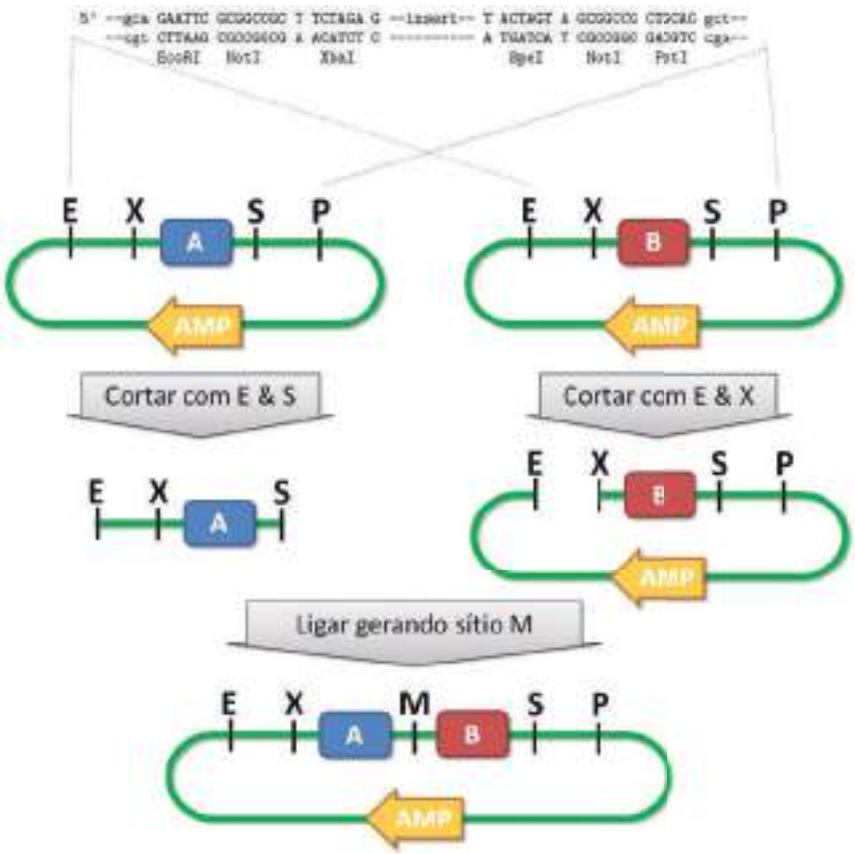


Figura 6. Construindo um BioBrick. A construção de BioBricks contendo sequências codificadoras de proteínas requer um prefixo com sítios das enzimas de restrição EcoRI (GAATTC) e XbaI (TCTAGA) e um sufixo com SpeI (ACTAGT) e PstI (CTGCAG). A ligação do sítio XbaI e SpeI gera o sítio misto M (TCTAGT), que não é palindrômico e não pode ser digerido novamente.

Utilizando apenas 4 enzimas, pode-se montar diferentes componentes em paralelo (*DNA assembly*) gerando extremidades que ainda podem receber outros componentes. Dessa maneira, pode-se juntar genes, reguladores de transcrição e tradução, promotores e qualquer outra peça biológica, para formar novos circuitos biológicos. Dentro desse contexto, foi fundado em 2003, no MIT, a Secretaria de Partes Biológicas Padrão (*Registry of Standard Biological Parts*) para depositar as partes genéticas utilizadas na montagem de dispositivos e sistemas sintéticos. A Secretaria contém milhares de partes que podem ser trocadas por inúmeros laboratórios cadastrados e espera-se que todos contribuam com dados e novas partes para melhorar o repositório. A Secretaria oferece muitos tipos de partes biológicas, incluindo plasmídeos, primers, promotores, domínios de proteínas, sítios de ligação de ribossomos, riborreguladores, genes repórteres, etc. Nesse sentido, o BIOFAB (*biofab.org*) é um projeto recente, em parceria com a empresa de síntese de DNA, *DNA 2.0*, que está desenvolvendo uma plataforma para disponibilizar dezenas de milhares de BioBricks de alta qualidade para fins acadêmicos e comerciais. O portal *OpenWetWare* (*openwetware.org*) é um bom ponto de partida para interagir com a Secretaria de Partes Padrão. Resumindo, a padronização do modo de clonagem e montagem permite não só a troca de partes e uma melhor consolidação do conhecimento, mas também a base para construção de dispositivos e sistemas sintéticos mais complexos, possibilitando também a automação desses processos.

Qualquer laboratório de qualquer universidade do mundo pode se juntar a esta comunidade para utilizar e trocar BioBricks. É uma nova forma de fazer ciência, muito mais colaborativa, transparente e democrática.

Dispositivos sintéticos: interruptores e portas lógicas

A Biologia Sintética nasceu em janeiro de 2000 com duas publicações lado a lado que foram publicadas na *Nature*. Os dois artigos mostravam que os mesmos três genes poderiam ser reconfigurados em diferentes circuitos para produzir diferentes dispositivos: um interruptor (Gardner et al. 2000) e um relógio biológico (Elowitz et al. 2000). Estes estudos ilustraram o princípio que o desenho de sistemas biológicos era possível e razoavelmente previsível.

Utilizando técnicas básicas de biologia molecular, Gardner e colaboradores (2000) construíram um interruptor para regular a transcrição de genes em *E. coli*. Chama-se um interruptor porque possuiu dois pontos de equilíbrio (acesso ou apagado, direita ou esquerda, expressão de X ou expressão de Y). Por exemplo, em uma das construções descritas o repressor 2 (repressor LacI) reprime o promotor 2 (Ptrc-2), e o repressor 1 (repressor Tet) reprime o promotor 1 (PltetO-1). A substância IPTG (Indutor 2) inibe o repressor 2 e a TC (Indutor 1) inibe o repressor 1 (Figura 7).

Dessa maneira, o dispositivo possui dois pontos de equilíbrio: (i) com a presença de TC, o repressor 2 é transcrito e ocorre a repressão do promotor 2, não havendo assim a transcrição do gene repórter (no caso uma proteína luminescente GFP); (ii) e na presença de IPTG, em que ocorre a transcrição do promotor 2, e consequente transcrição de GFP e do repressor 1. Dessa maneira, existem dois estágios: aceso

(transcrição de GFP) e apagado (sem transcrição de GFP). Este sistema é robusto porque funciona de acordo com o modelo proposto (Figura 7).

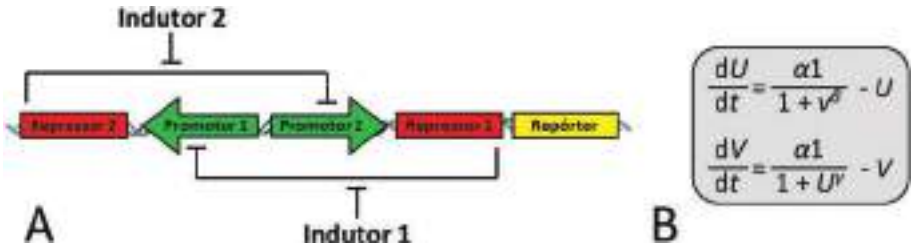


Figura 7. Representação gráfica do interruptor. Esse sistema robusto de controle da expressão gênica possui dois níveis de controle.

Na Figura 7, é mostrado que este dispositivo sintético possui dois pontos de equilíbrio do sistema, o estado 1 e o estado 2 (aceso e apagado, na presença de um dos indutores) e um outro ponto instável de equílibrio que mostra os dois repressores se regulando mutuamente. Seria mais fácil de entender se houvesse, de um lado uma proteína luminescente verde e do outro lado do dispositivo uma proteína luminescente amarela. Os estados de equilibro 1 e 2 representam verde ou amarelo, enquanto o ponto instável de equilíbrio representa a ausência de cor. Este tipo de dispositivo pode ser utilizado na biotecnologia para regular vias metabólicas inteiras, ligando e desligando vias de acordo com um sinal externo; ou na medicina para acionar a resposta a um remédio por exemplo. Funcionam de uma maneira mais eficiente do que o controle via promotores isolados, que podem interagir de maneira promíscua com outros reguladores.

Osciladores e uma bactéria que conta

Um oscilador eletrônico é um circuito elétrico que produz um sinal eletrônico repetitivo, frequentemente uma onda senoidal ou uma onda quadrada. Eles são amplamente utilizados em aparelhos eletrônicos, incluindo aparelhos de transmissão de rádio e televisão.

Stricker e colaboradores (2008) descreveram a construção de um oscilador sintético biológico. Para isso, foi utilizada uma construção genética baseada em um promotor híbrido $P_{lac/ara}$. Este promotor possui dois operadores que respondem a dois sinais: ele é ativado por AraC na presença de arabinose e é reprimido por LacI na ausência de IPTG. Seguido a este promotor foram clonados os genes *araC*, *lacI* e o gene codificador da proteína luminescente GFP (Figura 8). Um pulso de IPTG foi utilizado para sincronizar as células, dessa maneira, AraC é traduzido resultando na ativação do sinal luminescente de GFP. Por outro lado, AraC ativa também a expressão de *lacI* resultando na repressão de *araC* e *gfp*, cessando o sinal luminescente. Como LacI causa a repressão do próprio promotor híbrido de *lacI*, este acaba se reprimindo e um novo ciclo se reinicia.

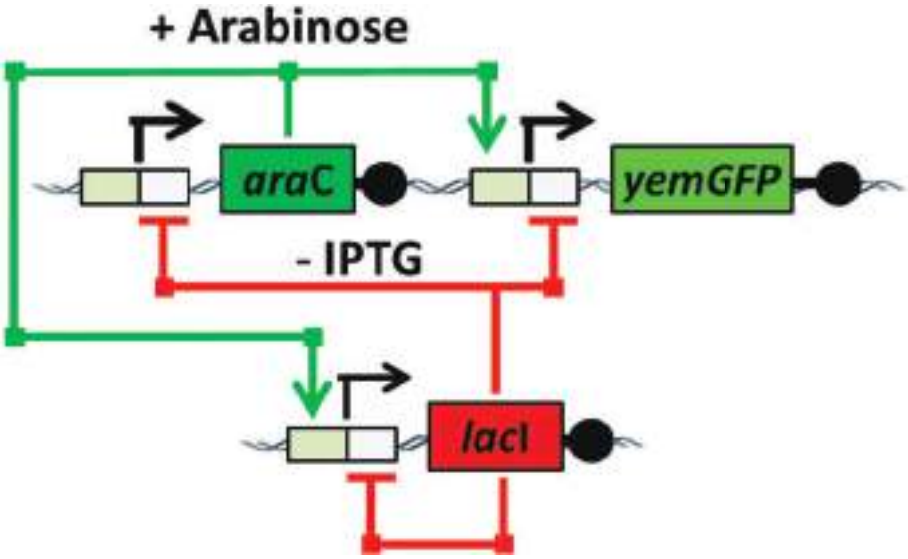


Figura 8. Representação gráfica do oscilador. A ativação por IPTG também leva à repressão da expressão gênica, iniciando novo ciclo.

O período de oscilação pode ser ajustado modificando a concentração do indutor. Na Figura 8 o gráfico apresenta um período de 40 minutos. Interessantemente o tempo de geração das bactérias do experimento varia de cerca 22 a 27 min, dessa maneira foi verificado que a informação e o sincronismo contido no oscilador é transmitido de geração em geração.

Foi verificado que a indução da oscilação foi muito rápida (cerca de 5 min) e inicialmente bem sincronizada, analisando cada célula individualmente. A amplitude da oscilação foi caindo com o progresso do experimento, devido a dessincronização gradual de cada célula da colônia, como já era de se esperar. Desse fenômeno, emerge um próximo tópico da BS, a comunicação célula-célula através do desenvolvimento de mecanismos de percepção de quorum (*quorum sensing*) sintéticos. Assim, as células poderão se comunicar para coordenar e sincronizar o seu comportamento. Esse assunto será abordado no próximo tópico.

Friedland et al. (2009) descrevem a construção de contadores sintéticos para bactérias. Após a construção de um dispositivo aparentemente simples, os autores sofisticaram bastante a construção do regulador para garantir uma resposta (contagem) robusta. O primeiro contador descrito pelo trabalho é o "contador riborregulado por uma cascata transcricional" (*riboregulated transcriptional cascade counter*). Ele conta pulsos de arabinose no meio de cultura utilizando como sinal a proteína luminescente GFP. São riborregulados porque o *taRNA* (controlado por um promotor sensível a arabinose) se liga ao sítio *cr* impedindo a formação de um grampo (pareamento) entre *RBS* e *cr*. Dessa maneira, permite o reconhecimento do *RBS* (sítio de ligação do ribossomo) pelo ribossomo e a tradução do gene (Figura 9). O interessante é que a transcrição do *taRNA* é regulada por um promotor sensível a arabinose. Assim, no primeiro pulso de arabinose, ocorre a transcrição de *taRNA* que permite a tradução da RNA polimerase T7. Após o primeiro pulso, arabinose e *taRNA* são degradados.

No segundo pulso, da mesma maneira, ocorre a transcrição da RNA polimerase T3 através do promotor P_{T7} que é especificamente reconhecido pela RNA polimerase T7. No terceiro pulso, ocorre também através da mediação de *taRNA* e da polimerase T3, a transcrição de GFP. Dessa maneira, a bactéria responde aos 3 pulsos de arabinose com um sinal luminescente.

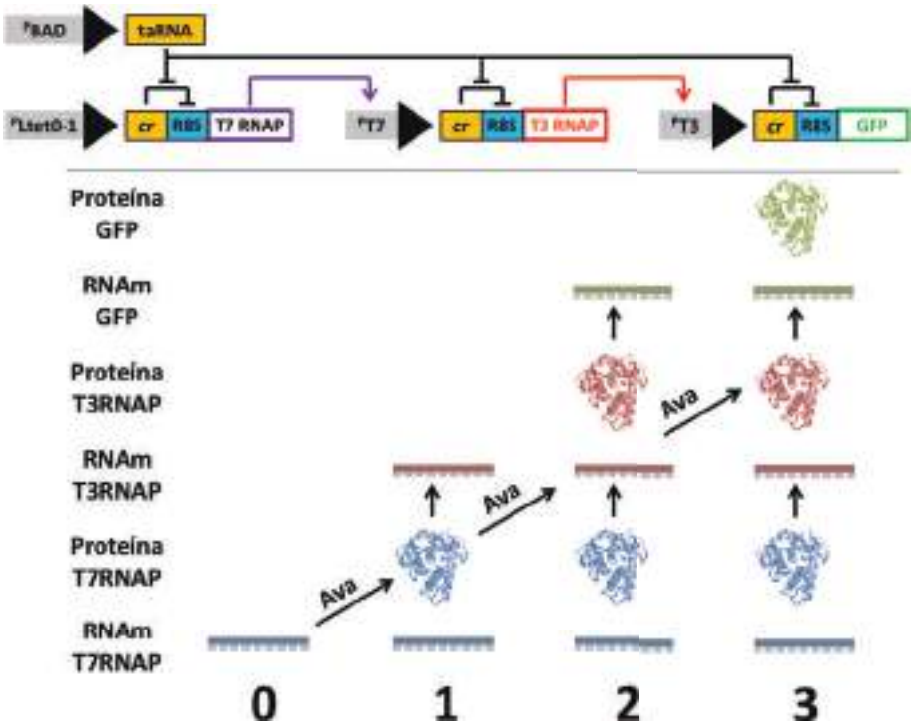


Figura 9. Representação esquemática do contador de cascata. Esse contador usa um sistema riborregulado capaz de contar pulsos de arabinose expressando proteínas fluorescentes diferentes.

Porém foi verificado um vazamento de expressão de GFP nos pulsos intermediários, além disso, demonstrou-se que se o pulso de arabinose não é dado a uma intensidade e frequência corretas. Isto provavelmente ocorre devido aos limites cinéticos intrínsecos envolvendo os tempos de transcrição e degradação de mRNA. Essas constatações reforçam a idéia de que é necessário levar em conta o ruído intrínseco a componentes biológicos, naturais ou sintéticos para a construção de um dispositivo robusto.

Por esses motivos, um segundo contador foi construído, chamado contador de cascada de DNA invertase (*DNA invertase cascade counter*), em que basicamente foi acoplado ao modelo anterior um controle de invertases na região promotora. Uma invertase, enzimas Cre por exemplo, reconhecem determinadas sequências de DNA que flanqueiam uma determinada região de DNA e invertem a orientação dessa região. Nesta construção, essa enzima é utilizada para ativar os promotores através da inversão de sua orientação (Figura 10). Com essa construção foi possível obter

um controle muito maior da expressão de GFP. Este contador agora é muito mais robusto em relação à variação de tempo entre os pulsos de arabinose, podendo contar pulsos realizados em intervalos de tempo de 2 a 12 horas. Este novo dispositivo, além de mostrar o sinal luminescente, grava a informação relacionada aos pulsos no DNA através das recombinações.

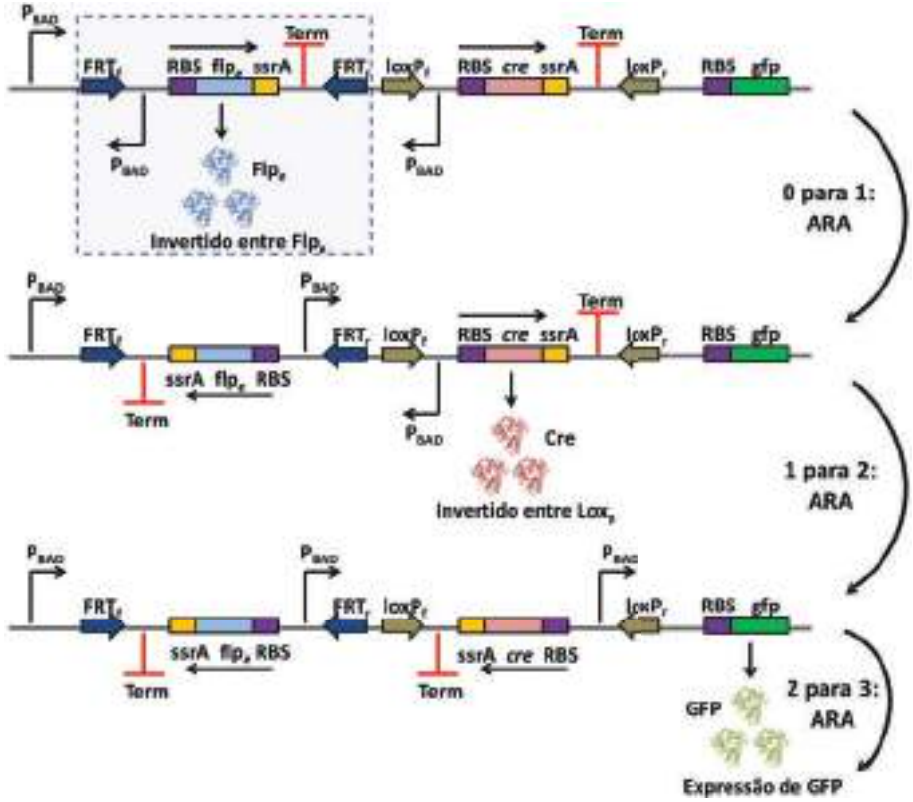


Figura 10. Representação esquemática do contador de cascata de DNA invertase. Esse contador pode ser regulado com maior precisão que o modelo anterior por usar um mecanismo de modificação da orientação do DNA (através da invertase Cre), com a vantagem adicional de gravar o estado de regulação através das recombinações.

Os autores foram ainda mais longe e trocaram o promotor sensível a arabinose por promotores capazes de responder a diferentes sinais. Dessa maneira, o dispositivo pode ser programado para gravar e responder diferentes sinais na ordem desejada. Esse dispositivo genético pode ser programado para diferentes funções a fim de sincronizar diferentes sinais para uma determinada resposta. Dependendo dos sinais acoplados ao dispositivo, estes mecanimos podem ser utilizados para biosensores, biorremediação ou na medicina. Utilizando mecanismos de controle similares, já foram construídos mecanismos de autodestruição celular capazes de responder a diferentes sinais (Pasotti et al., 2011).

Sistemas gênicos multicelulares

Células sintéticas poderão ser criadas facilmente no futuro, porém ainda há uma limitação física para o tipo de função complexa que poderá ser realizada em células isoladas e na confiabilidade do desempenho das células individuais. Ao distribuir redes sintéticas entre múltiplas células utilizando mecanismos de comunicação célula-célula aumenta-se as possibilidades de desenhos de funções mais complexas como também a confiabilidade no sistema.

Mesmo populações geneticamente idênticas possuem fenótipos heterogêneos e seus comportamentos celulares não são idênticos. Dessa maneira, um grupo de células que não se comunicam irão se comportar de uma maneira não sincronizada. Mecanismos de comunicação célula-célula para coordenar o comportamento celular são utilizados para resolver o problema de sincronismos em populações celulares ou em sistemas multicelulares. Para o oscilador sintético, apresentado no item anterior, garantir um sinal robusto ao longo do tempo é necessário que a população celular esteja sincronizada. De outra maneira, uma pequena dessincronização em cada período ao longo do tempo pode representar uma total falta de sincronização da população em pouco tempo.

Um dos esforços pioneiros para engenheirar sistemas multicelulares em BS através da montagem e recombinação de partes foi desenvolvido em *E. coli* para coordenar o comportamento de uma comunidade celular (Weiss, 2000). Os genes responsáveis pela percepção de quorum em *Vibio fischeri* foram compartimentalizados em dois tipos de células: células de envio (*senders*) e células receptoras (*receivers*). Células de envio expressam LuxI, que cataliza a síntese de moléculas pequenas e difusíveis de acil-homoserina lactona (AHL). O sinal AHL se difunde no meio até as células receptoras nas quais formam um complexo com a proteína LuxR e ativam genes regulados por este complexo. Posteriormente, pesquisadores aumentaram a sensitividade de LuxR para uma variedade de derivados de AHL e identificaram os resíduos de LuxR importantes para a ligação aos sinais (Collins et al., 2005).

Em um outro exemplo de como a comunicação celular representa hoje um foco em BS, unindo competências em diferentes frentes incluindo Microfluídica e Computação, o grupo de Hasty descreveu uma rede genética engenheirada para ser acoplada à densidade celular, capaz de gerar oscilações sincronizadas em populações crescentes. Os relógios moleculares sincronizados mostram o caminho para o uso de micróbios na criação de biosensores macroscópicos com resposta oscilatória. Ademais, fornecem um sistema modelo específico para a geração de descrição mecanística para o emergente comportamento coordenado na colônia celular (Danino e Hasty, 2010).

O conhecimento atual da biologia molecular por trás da comunicação intercelular bacteriana e os novos protocolos e plataformas usados para investigar esse fenômeno, em particular os circuitos regulatórios, fornecem a sintax da rede de comunicação bioquímica. Nesse sentido a Engenharia também se mostra presente, contribuindo com princípios necessários para a reprogramação bacteriana através da comunicação química para várias aplicações, desde controle coletivo da expressão gênica em nível populacional até o estudo de interações patógeno-hospedeiro (Mitchell et al., 2011).

Comunicação intercelular por pequenas moléculas difusíveis é usada não somente por formas de vida multicelulares, mas também por organismos unicelulares. Um

amplo conjunto de genes bacterianos é regulado por mudanças no ambiente químico gerado pela densidade populacional local de sua espécie e outras. As moléculas "autoindutoras" dependentes da densidade celular regulam a expressão de genes envolvidos em competência genética, formação de biofilme, persistência, virulência, esporulação, bioluminescência, produção de antibióticos e muitas outras características. Inovações recentes na Tecnologia do DNA Recombinante e Micro/Nanofluídica tornam os circuitos genéticos responsáveis pela comunicação célula-célula acessíveis e maleáveis através de abordagens de BS. O entendimento das diferentes linguagens intercelulares será fundamental para a otimização de processos e ganho de complexidade de sistemas sintéticos. Resumindo toda a parte histórica do desenvolvimento das bases conceituais e marcos importantes, é apresentada uma linha do tempo na Figura 11, que não pretende ser exaustiva, mas serve como referência cronológica para mostrar o crescimento e evolução da complexidade da biologia sintética.

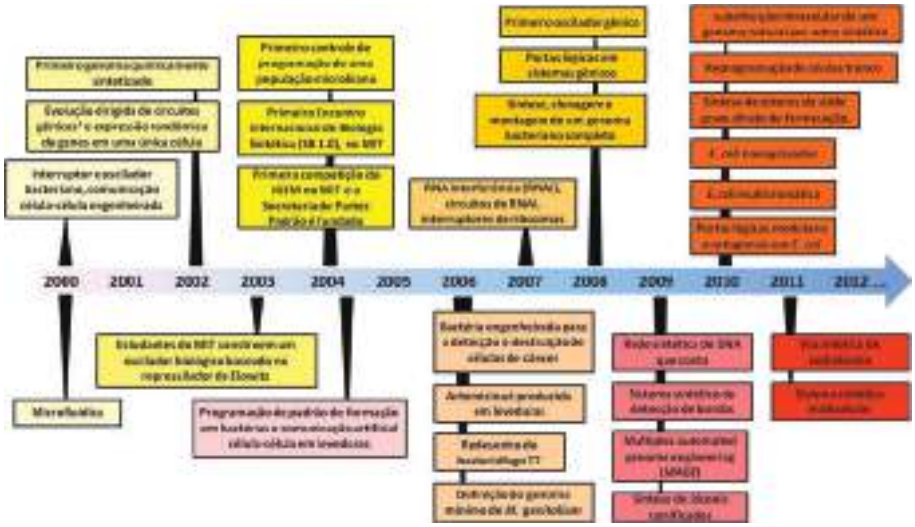


Figura 11. Principais marcos históricos da Biologia sintética. Atualizado a partir de Punick e Weiss, 2009.

Estado da arte da BS: grupos de pesquisa e empresas

O avanço tecnológico aproxima cada vez mais as Ciências da Vida e de Materiais, criando uma crescente interface. Se por um lado cientistas de ponta em centros de pesquisa acadêmicos modernos buscam avidamente expandir a barreira do que é possível; empresas nascem para transformar esses conhecimentos em produtos comerciais e serviços. As novidades surgem rapidamente e, provavelmente, quando esse livro chegar aos leitores novas empresas já estarão surpreendendo com abordagens ainda mais surpreendentes, mas esperamos dar ao leitor a oportunidade de conhecer um pouco melhor o mercado da BS.

O mercado global para BS foi de USD 233,8 milhões em 2008. O segmento de energia e químicos tem a maior participação com USD 80,6 milhões em 2008, porém com crescimento estimado de 80% ao ano O segmento de biotecnologia e fármacos é o segundo maior com USD 80 milhões em 2008 e projetados USD 600 milhões em 2013, com crescimento em torno de 50% ao ano (Bergin et al., 2011). Atrai não somente investimentos de governos e grandes corporações já no ramo de combustíveis, materiais e farmacêuticos, mas também de fundos de capital de risco como, por exemplo, o MLSCF da Malásia.

O número de pesquisadores trabalhando com BS no mundo já ultrapassa 45 mil (Maven Semantic http://bit.ly/tn1jLa), estando mais da metade nos EUA e o restante no Japão, Reino Unido, Alemanha, Itália, Canadá, França, China, Austrália, Espanha, Holanda, e alguns outros menos expressivos incluindo o Brasil (Figura 12). A literatura cresce rapidamente, dobrando a cada 20 meses desde 2004. Os temas mais abordados são controle genético de circuitos regulatórios, biocombustíveis, bioinformática aplicada a estudos em larga escala, biomedicina, aplicações ambientais como biorremediação, genomas mínimos, novos chassis e biosegurança. O número de patentes também se acumula rápidamente atraindo preocupação a respeito da imposição de obstáculos para o desenvolvimento da ciência devido à proteção intelectual.

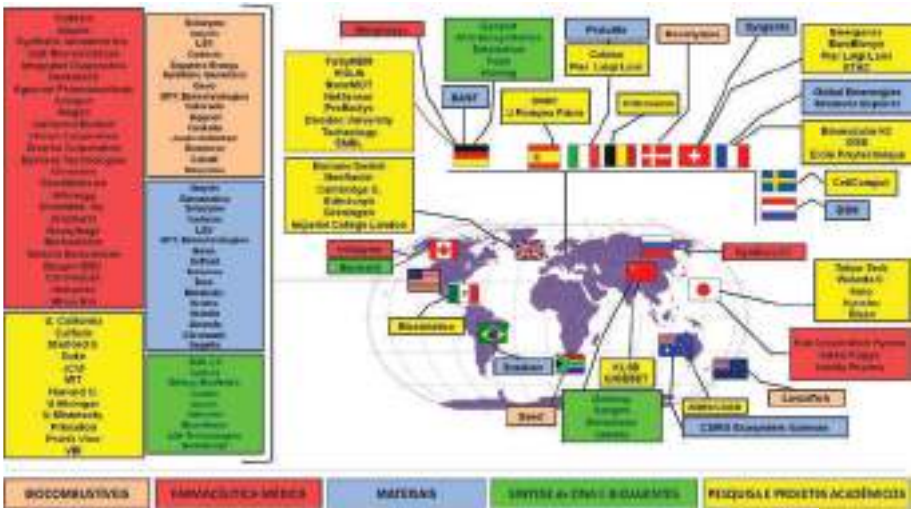


Figura 12. Empresas e grupos de pesquisa líderes em biologia sintética. As caixas estão coloridas de acordo com o setor representado, de acordo com os cinco grupos na parte inferior da figura.

Pesquisas e projetos acadêmicos

Muitos são os contribuintes pioneiros da BS e mais numerosa ainda é a segunda geração que já desponta hoje em vários países do mundo. Por limitações de espaço iremos descrever brevemente somente os principais centros de pesquisa dedicados a BS. Mas lembramos que muitos participantes importantes como Jay Keasling, James

Collins e Tom Knight estão também na iniciativa privada e são criadores de muitas das empresas que vamos citar a diante.

Começando pelos EUA, o Instituto J. Craig Venter (JCVI) foi formado em outubro 2006 através da fusão multidisciplinar de organizações afiliadas focadas em genômica. Com mais de 400 cientistas e técnicos, mais de 23 mil m^2 de instalações laboratoriais e localizado em Rockville e San Diego (EUA), o novo JCVI é um líder mundial em pesquisa genômica. As equipes estão engajadas em algumas das linhas de pesquisa mais emocionantes das ciências biológicas. Publicaram o primeiro genoma diplóide humano (Levy et al., 2007) e resultados iniciais da expedição de amostragem oceânica global que descobriu mais de seis milhões de novos genes e milhares de novas famílias proteicas de organismos encontrados nas águas dos oceanos (Yooseph et al., 2007). Equipes também sequenciaram os genomas da flora microbiana encontrada em ambientes humanos como a vagina, cavidade oral e trato intestinal (Huttenhower et al., 2012). Estão progredindo firmemente no caminho de criar um cromossomo e organismo sintético já tendo transformado com sucesso o genoma de uma espécie bacteriana em outra. Também já sequenciaram genomas de uma variedade de importantes agentes infecciosos e vetores como o mosquito *Aedes aegypti* e estão em busca de entender a evolução de diversos genomas virais como influenza e coronavirus (Nene et al., 2007). Essas pesquisas serão aproveitadas em novas formas de diagnóstico e tratamento de doenças infecciosas no mundo. Essas são apenas algumas de muitas das áreas de pesquisa das equipes do JCVI. Vale lembrar que seu criador está envolvido na criação de empresas como a Synthetic Genomics Inc.

O Centro de Pesquisa e Engenharia em BS (SynBERC) é um consórcio financiado pelo *National Science Foudation* (NSF) norte-americano, e agrega competências de vários laboratórios de BS em centros como o MIT, Harvard University, University of California e Stanford University. Conta com aproximadamente 16 grupos de pesquisa, totalizando uma equipe com aproximadamente 200 cientistas pesquisando ciência básica de fronteira em diversas frentes. Possuem um foco em transferência de tecnologia para a indústria, promovendo parcerias entre as principais industrias farmacêuticas, alimentícias e petroquímicas e uma gama de novas empresas *startups*. Alguns desses centros como o da Universidade da California em Berkeley e o MIT possuem cursos de graduação e pós-graduação em Bioengenharia, permitindo a formação de profissionais com competências em Engenharia e Ciências Biológicas.

Embora o Instituto de Tecnologia da Universidade da California (CalTech) não mantenha um centro denominado como de BS, as pesquisas em diferentes centros como o Departamento de Bioengenharia, Centro de Imagens Biológicas (BIC), de Modelagem de Redes Biológicas (BNMC), de Engenharia Bioinspirada (CBE), de Pesquisa em Energias Sustentáveis (CCSER), o Instituto Jacobs de Engenharia Molecular para a Medicina e o Centro de Fotosíntese Artificial (JCAP) abordam vários temas em BS, desenvolvidos por mais de 25 grupos de pesquisa.

O Instituto de BS (SBI) na Universidade da California em Berkeley foi inaugurado em 2010 para estabelecer os passos para a produção em larga escala de novos sistemas biológicos. Através do esforço conjunto de seus pesquisadores, parceiros e membros da indústria, o SBI está desenvolvendo os padrões e tecnologias necessárias para criar novas aplicações em energia, materiais, fármacos, produtos químicos, alimentícios, segurança e outras indústrias que afetam nossa vida cotidiana. Um instituto

interdisciplinar em sua essência, o SBI conta com pesquisadores de oito departamentos da UC-Berkeley e três divisões do Lawrence Berkeley National Laboratory. A empresa Agilent Technologies, Inc., baseada em Santa Clara, California, é um membro industrial fundador do SBI e novas empresas são incentivadas a desenvolverem parcerias com o Instituto. Os engenheiros e cientistas em Berkeley já se destacam na vanguarda global da pesquisa em BS. Na Universidade da California em San Francisco também existe um núcleo forte em BS.

Como representantes europeus destacamos o Centro para BS (CSB) que reune três Institutos na Universidade de Groningen, Holanda, totalizando cerca de trinta grupos e formando uma massa de centenas de pesquisadores nas diversas áreas, mas com forte ênfase em Biologia Celular. Destacamos também o Instituto de BS e de Sistemas (ISSB) do *Imperial College London* que conta com dezesseis grupos, dos quais muitos participam do *Nest Pathfinder on Synthetic Biology*, uma iniciativa da União Européia que financia 18 projetos com equipes internacionais além do Centro de Tecnologia Federal da Suíca (ETH) com pesquisas na área médica do Prof. Martin Fussenegger.

A Dinamarca recentemente inaugurou o Centro para Biosustentabilidade focado em biologia sintética para a produção de combustíveis, químicos e fármacos com um gerenciamento híbridro entre empresa e centro de pesquisa, buscando por pesquisas orientadas pelo mercado. Para isso, trouxe talentos do mundo todo e possui 50% da equipe formada por estrangeiros e diretores científicos de 6 universidades diferentes da Europa e dos Estados Unidos.

Na Asia se destacam as instituições de pesquisa no Japão, China e Coréia do Sul. No Japão principalmente na Universidade de Tóquio pesquisando alternativas para o ribossomo e código genético. Na China com apoio dos programas federais os pesquisadores estão focando em biocombustíveis e biosensores e na Coréia do Sul principalmente no Instituto Avançado de Ciência e Tecnologia (KAIST) onde pesquisam genoma mínimo. Veja mais grupos de pesquisa individuais na Figura 12.

Fábricas de DNA, biopartes e informática

O ciclo oferta-demanda gerado com o barateamento da síntese de DNA customizado favoreceu a evolução de empresas como a Life Technologies, já no mercado a um bom tempo mas antes restrita ao fornecimento de oligonucleotídeos <100 pb, e favoreceu também o surgimento de muitas empresas como a DNA 2.0 e GeneArt, que além do serviço de síntese de fragmentos >400.000 pb, também utiliza a otimização de códons para aumentar a produção e estabilidade das proteínas recombinantes. O serviço inclui a sequência codificante já inserida no vetor de expressão, agilizando todo o processo. Com esses vetores de expressão otimizados os níveis de proteína recombinante chegam a 100x superiores aos obtidos com vetores tradicionais. São serviços que irão alterar a rotina nos laboratórios acadêmicos e privados. Por exemplo, um gene de 1000 pb custaria USD 200 para sintetizar. E os preços estão caindo pela metade a cada 15 meses.

O volume de dados gerados em projetos larga-escala também é crescente e o desenvolvimento de toda Bioinformática é feito pelas empresas de síntese de DNA e outras especializadas como a italiana Protolife. Ela desenvolve uma ferramenta automatizada de modelagem com inteligência preditiva, capaz de encontrar alvos

otimizados em espaços experimentais gigantescos sem a necessidade de testar cada variante. Essas empresas compõem todo um parque industrial dedicado à essa nova era da BS.

BS aplicada aos combustíveis, químicos e biorremediação

Imagine, uma estação *off-shore* extraindo petróleo em alto-mar a grandes profundidades, uma unidade de craqueamento de petróleo com capacidade de milhões de toneladas por ano para produção de combustíveis e químicos básicos para a produção de solventes, polímeros, lubrificantes, e quase todos os materiais que utilizamos no nosso dia-a-dia. A dependência de energia e materiais provenientes de recursos fósseis trouxe consequências terríveis para o meio ambiente (e.g. acidentes ambientes e acúmulo de resíduos recalcitrantes) e hoje vivemos em uma situação de insegurança ambiental sem saber exatamente quais as consequências da emissão de CO_2 para as mudanças climáticas no futuro. Por outro lado, já temos visto o desenvolvimento de biorefinarias no Brasil e no mundo, nas quais já é amplamente difundida a produção de açúcar, etanol e eletricidade. A vantagem principal é um novo conceito de produção com emissões de CO_2 muito mais baixo, já que o CO_2 produzido, por exemplo, na queima do bagaço para produzir eletricidade ou na combustão do etanol pelos veículos, retorna para lavoura em forma de cana-de-açúcar. O Brasil possui o modelo mais avançado de biorefinaria no mundo, e o etanol brasileiro, diferente do americano produzido a partir do milho, foi considerado, pelo DOE, um combustível avançado devido a sua menor pegada ambiental.

Dentro desse contexto, a biologia sintética é uma das grandes esperanças de transformação para a nossa sociedade dependente de recursos fósseis. Nós últimos anos temos visto o surgimento de novas empresas de biotecnologias, baseadas em conceitos de biologia sintética, e o movimento de grandes corporações petroquímicas em investir e desenvolver novas tecnologias renováveis.

De uma forma geral, as empresas startups de biotecnologia têm estruturado plataformas de tecnologias que possibilitem explorar diferentes produtos ou processos, criando sinergia entre diferentes programas de P&D, enquanto grandes empresas petroquímicas estão interessadas em desenvolver produtos específicos mais focados no seu negócio. Um exemplo interessante é uma empresa pioneira e uma das mais conhecidas na biologia sintética chamada Amyris (Figura 13). Esta empresa se iniciou com uma doação da Fundação Bill e Melinda Gates de USD 40 milhões para o desenvolvimento de uma rota de baixo custo para a produção do ácido artemísico. Porém, utilizando a mesma via metabólica, a via dos isoprenóides, é possível produzir outros compostos de interesse industrial como o isopreno (C_5H_8), monômero para produção de borracha, o farneceno ($C_{15}H_{24}$), um precursor para a produção de biodiesel e outras moléculas como esqualano ($C_{30}H_{50}$), utilizado em cosméticos (Figura 13), além de combustíveis de avião e lubrificantes. Outras possibilidades são a produção de isoprenóides mais complexos como a fragrância linalol ($C_{10}H_{18}O$), muito conhecida por causa do perfume Chanel nº5, carotenóides ou anti-cancerígenos como o taxol (taxadieno).

Outro exemplo de plataforma metabólica é o portfólio de patentes e produtos desenvolvido para empresa LS9, que utiliza os conhecimentos do redesenho da beta-oxidação para a produção de vários produtos a partir de ácidos graxos. Ácidos graxos têm sido utilizados há séculos para a produção de combustíveis, produtos químicos e polímeros, incluindo o biodiesel, surfactantes, solventes e lubrificantes. Porém a demanda crescente e a produção limitada de óleos vegetais têm causado questionamentos sobre o aumento dos preços dos alimentos, sobre a prática de utilização dos solos e os aspectos socioambientais relacionados com a sua produção e destruição de florestas na Ásia. Uma alternativa é a produção desses derivados de ácidos graxos via conversão biológica utilizando microrganismos como levedura e bactérias. A empresa LS9, a partir de rotas de síntese de ácidos graxos por microrganismos, desenvolveu diferente rotas para a produção direta de alcanos para aplicação em combustíveis, alcenos para a produção de polímeros, e álcoois graxos para aplicações em produtos de limpeza. Outro exemplo de destaque são os ésteres de ácidos graxos. Imagine uma plantação enorme de soja no interior do Brasil e sua instalação de extração do óleo do grão da soja, que é composto por vários ácidos graxos. Agora imagine no interior de São Paulo, uma grande usina de etanol cercada por dezenas de quilômetros quadrados de cana-de-açúcar. Imagine também o etanol e óleo de soja sendo transportados até uma usina de biodiesel para, utilizando um catalisador, serem convertidos em um éster (biodiesel) e em glicerina. Muito tempo e dinheiro são gastos nesse transporte, encarecendo o produto e gerando muitos resíduos pelo caminho. Recentemente, foi publicado um trabalho liderado pelo Prof. Jay Keasling demonstrando a possibilidade de se utilizar uma bactéria para produzir biodiesel em apenas uma etapa e utilizando resíduos agroindustriais, como por exemplo, o bagaço de cana-de-açúcar. Essa bactéria modificada geneticamente é capaz de produzir ácidos graxos e etanol e, enzimaticamente, realizar a esterificação desses produtos em biodiesel (Steen et al., 2010). Através de várias modificações genéticas foi possível aumentar a produção de 40 mg/l para quase 700 mg/l de biodiesel. Resultado este ainda baixo para se cogitar uma aplicação a curto-prazo, mas, sem sombra de dúvida, as perspectivas são muito animadoras. Esta tecnologia foi licenciada para empresa LS9 que hoje a explora comercialmente. Outro exemplo de plataforma metabólica baseada da via de produção de ácidos graxos é a desenvolvida para Solazyme, que é capaz de produzir diferentes tipos de ácidos graxos para diferentes aplicações na indústria química e alimentícia.

Recentemente, foi demonstrado a possibilidade de produção de alcóois ramificados e com longas cadeias utilizando a via de Ehrlich. Através da incorporação de uma descarboxilase de amplo espectro de keto-ácidos e uma álcool desidrogenase, microrganismos tem sido engenheirados para a produção desses combustíveis (Atsumi et al., 2008). Utilizando mecanimos enzimáticos semelhantes, duas empresas buscam a produção de isobutanol, a Butamax, uma joint-venture entre a BP e a Dow, e a Gevo. Esses alcoóis são geralmente considerados melhores combustíveis do que o etanol em relação à gasolina e podem ser utilizados para a produção de diferentes *comodities* químicas.

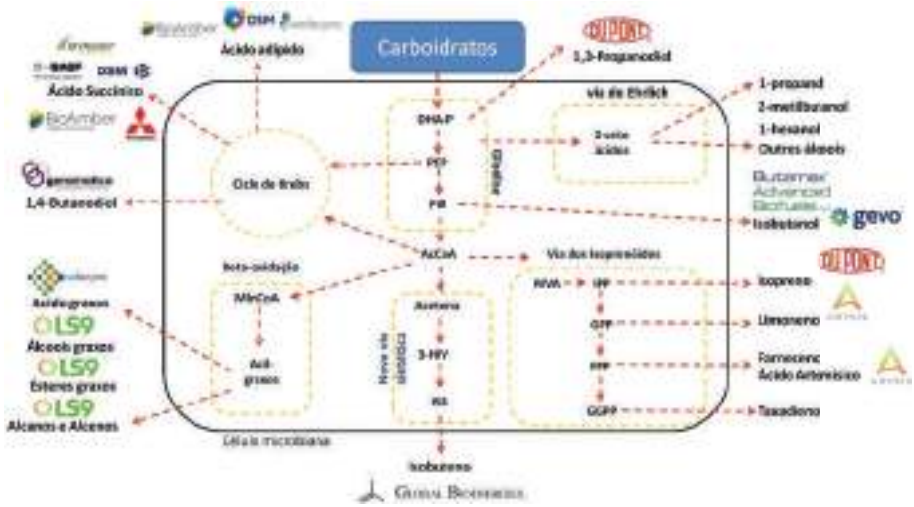


Figura 13. Exemplos de vias metabólicas exploradas por empresas de biotecnologia. Empresas como a Gevo e Butamax pesquisam a produção de biocombustíveis como o isobutanol; a Genomática desenvolve um processo para a produção de 1,4 butanodiol para plásticos e fibras elásticas. Nesse mesmo sentido, a DSM e a Myriant competem para desenvolver rotas para a produção de ácido succínico. Por fim, a empresa Amyris explora a via dos terpenóides para a produção de diferentes compostos como isopreno, farneceno e artemisina.

Um outro ramo interessante é produção de polímeros e materiais verdes. Muitos monômeros para produção de polímeros possuem um processo industrial complexo para sua produção que resulta em maior custo e pegada ambiental. Por outro lado, processos biotecnológicos podem produzir estes monômeros diretamente de açúcar e com alto rendimento, podendo melhorar custos e diminuir os impactos ambientais. Exemplos interessantes são a produção de do monômero ácido adípico para a produção de nylon, ou dos monômeros de ácidos succínico e 1,4-butanodiol (Yim et al., 2011), para a produção de PBS (polibutileno succinato). Nessa mesma linha, a empresa francesa Global Bioenergies foi a empresa pioneira em desenvolver uma rota sintética para oleofinas, desenvolvendo novas enzimas para conversão direta de açúcar a isobuteno, que pode ser utilizado em uma gama de produtos, incluindo polímeros e aditivos de combustível. Essa empresa apresenta um tipo de plataforma biotecnológica diferente das outras apresentadas, ao invés de ter uma plataforma baseada em uma via metabólica, a Global Bioenergies possui uma plataforma baseada em uma enzima. A enzima difosfomevalonato descarboxilase (EC 4.1.1.33), tem sido engenheirada para ampliar a sua atuação em um espectro maior de substratos, possibilitando a produção de diferente alcenos, como isobuteno, propeno e butadieno utilizando microrganismos (Malliere et al., 2009) (Figura 13). Da mesma forma que uma descarbolixase pode atuar em diferentes keto-ácidos, esta descarboxilase pode descarboxilar e desidratar diferentes 3-hidroxialcanoatos.

Outras empresas, como a Algenol, Joule, Aurora Algae, Synthetic Genomics, apostam nas algas para a geração de biocombustíveis. Essa última recebeu aporte de USD 300 milhões da Exxon para o desenvolvimento da nova geração de algas sintéticas para conversão de CO_2 e luz em biocombustíveis. Possíveis pontos positivos

da utilização de algas são a consolidação do processo (há necessidade de uma etapa intermediária para a produção de açúcar), aumento de produtividade e a utilização regiões áridas e água do mar para o cultivo de algas. Porém, os custos de produção de combustíveis por algas ainda são proibitivos.

Até o momento, foram descritas rotas sintéticas para a produção de compostos, porém a BS também avança para desenvolver novas rotas de biorremediação para compostos indesejáveis. Biorremediação já é comum na limpeza de derramamentos de óleo; bactérias como *Rhodococcus* e *Pseudomonas*, entre outras, naturalmente consomem e degradam muitos componentes do petróleo em subprodutos menos tóxicos. Um dos objetivos da BS é criar biosensores e organismos capazes de degradar compostos mais recalcitrantes como dioxinas, pesticidas e até compostos radioativos (Charles e Schmidt, 2010).

Em busca de matérias-primas mais baratas, principalmente caldo de cana-de-açúcar, e um sistema de produção de biorefinarias mais sustentável, boa parte das empresas citadas possuem laboratórios para desenvolvimento de processo no Brasil. Empresas como a Amyris, Butamax, L9S e Solazyme possuem centros de pesquisa e desenvolvimento no interior de São Paulo com infraestrutura de laboratório, plantas piloto e demonstração. Em janeiro de 2013, a Amyris começou a produção comercial de farneceno com uma planta integrada a Usina Paraíso em Brotas, interior de São Paulo. Em maio de 2014 a *joint venture* Solazyme e Bunge iniciaram a produção comercial de óleos renováveis de alto valor agregado em sua planta integrada a Usina Moema também no interior de São Paulo. Uma iniciativa pioneira no Brasil para utilização das ferramentas de BS está sendo realizada pela empresa Braskem para desenvolver uma rota fermentativa para a produção de propeno a partir de cana-de-açúcar. A empresa possui um Centro de Biotecnologia em Campinas que está trabalhando para desenvolver um processo industrial verde para a produção de químicos verdes. Atualmente a empresa possui parcerias com empresas como Genomatica e Amyris para o desenvolvimento de monômeros para a produção de borracha verde visando, principalmente, a produção de pneus. Esses investimentos geram renda e empregos, modernizando a economia brasileira.

Biotecnologia farmacêutica e medicina

A Engenharia Metabólica já é usada há décadas pela Indústria Farmacêutica, em busca de novas enzimas e vias metabólicas para produção de compostos diversos. Através do uso da "nova caixa de ferramentas" fornecida pelas técnicas de BS, a Indústria busca acelerar o desenvolvimento de novos fármacos, mais eficazes, mais baratos e com menos efeitos colaterais. Entre as vantagens estão, por exemplo, o método de otimização de códons, que pode ser usado para alterar características específicas das biopartes. Estas podem ser selecionadas por métodos *in vivo* de varredura em larga escala (BOX bibliotecas). Além disso, os métodos convencionais que dependem fortemente da catálise química, custosa e poluente, potencialmente serão substituídos pela **Biocatálise**, através da qual as reações químicas são catalizadas enzimaticamente. Mas as vantagens não param por aí: com a modificação de enzimas existentes ou através da combinação de diferentes vias metabólicas é possível criar

novos ingredientes farmacêuticos ativos (*APIs*) ou precursores, ainda inexistentes no ambiente natural ou inacessíveis pelas técnicas convencionais. Um exemplo dessa transição tecnológica que chegou ao mercado recentemente pela união da Codexis e Merck é o antidiabético Januvia (*sitagliptin*), um inibidor da dipeptidil peptidase-4 para o tratamento de diabetes tipo II. Pelo desenho e teste de variantes enzimáticos, foi possível identificar uma nova enzima com atividade inicial detectável. Essa enzima foi então aperfeiçoada mais de 25.000 vezes e chegarão em uma variante enzimático altamente ativo, estável e enantioseletivo, a partir de uma atividade que não existia inicialmente no mundo natural (Savile et al., 2010).

BOX – Bibliotecas combinatoriais *in vivo*

Bibliotecas combinatoriais de pequenas moléculas são uma forma eficiente de se buscar novas estruturas chave. Criar essas bibliotecas sinteticamente, no entanto, é custoso e muitos candidatos inicialmente promissores falham em rodadas subsequentes de seleção ou testes clínicos. Produzir bibliotecas de compostos diretamente *in vivo* teria várias vantagens. Primeiro, manter e amplificar tal biblioteca é simples e pode ser feito cultivando células. Segundo, o processo de varredura torna-se facilitado devido a informação genética manter-se diretamente ligada ao composto. Logo, o composto selecionado pode ser identificado analisando a informação genética da célula correspondente. Terceiro, uma seleção genética intracelular pode ser usada diretamente para ensaiar efeitos em atividades enzimáticas, contornando limitações inerentes a ensaios *in vitro*. Quarto, a seleção é feita no contexto de uma célula viva, que requer um nível mais alto de seletividade da droga por seu alvo. Finalmente, potenciais problemas com solubilidade e captação do composto são contornados.

Dessa maneira, a BS desponta com uma plataforma de alta tecnologia para o desenvolvimento de novas terapias e processos de produção de fármacos. Estes aspectos inovadores da BS foram desenvolvidos a partir de décadas de pesquisa e desenvolvimento da indústria farmacêutica que utilizam tecnologias sofisticadas e caras. Por exemplo, o custo total para o desenvolvimento de uma nova droga é cerca de USD 800 milhões de dólares (Kirby & Keasling, 2008). Por isso, tem sido assumido que as tecnologias convencionais seriam muito caras para serem aplicadas na criação de novas drogas que são extremamente necessárias para países em desenvolvimento, para as cujos governos e população não teriam recursos para pagar por estes medicamentos. Porém, o Projeto da Artemisina (Artemisinin Project) desafia esta hipótese. Armados com uma doação da Fundação Bill e Melinda Gates e uma parceria entre a Amyris, o Instituto OneWorld Health e a Universidade de Califórnia, desenvolveu uma nova plataforma de produção de fármacos baseado em leveduras e um processo para produzir uma droga anti-malária de uma forma barata, a partir de açúcares (Kirby & Keasling, 2008). Artemisina é um composto normalmente produzido na planta *Artemisia annua*, comumente conhecida como erva de São João. Nela, um grupo de proteínas trabalham em cadeia para produzir a artemisinina, isolada a partir de folhas e flores da planta. Pesquisadores do grupo do Prof. Jay Keasling em Berkeley introduziram os genes para essas proteínas em levedura, e em seguida otimizaram

a quantidade de proteína produzida a partir de cada gene de forma a sintetizar o precursor da droga apropriadamente (Ro et al., 2006). Utilizando processos robotizados de síntese e montagem de DNA, além de seleção de microrganismos e análise química a Amyris procura otimizar linhagens produtoras de ácidos artemísico. Esse processo demonstrou o grande potencial de reduzir enormemente o custo da artemisinina, uma droga crucial para o tratamento da malária e atraiu investimentos da Sanofi-Aventis, uma das líderes do setor farmacêutico.

Explorando a capacidade de se criar organismos modificados com novas propriedades e capacidades biosintéticas, a indústria farmacêutica já utiliza abordagens de BS para criar, por exemplo, biosensores que auxiliam no desenvolvimento de novos fármacos em diferentes fases do processo. Além disso, novas fontes de compostos bioativos são criados na forma de bibliotecas de pequenas moléculas codificadas geneticamente. A recombinação de partes individuais foi empregada para desenhar proteínas que funcionam como biosensores que são utilizados para identificar e quantificar moléculas de interesse. Integrando partes em dispositivos e sistemas, novas vias biosintéticas podem ser criadas combinando atividades enzimáticas desejadas e a potencial expansão do código genético pode ser usada para introduzir novas funcionalidades em peptídeos e proteínas para aumentar o escopo químico e estabilidade biológica (Neumann et al., 2010). Esses avanços tecnológicos permitem a utilização de matérias-primas mais baratas e na redução de etapas de processo que resultam em menor investimento e custos de produção de fármacos. Este conceito fica claro novo processo de produção do antibiótico cephalexina pela empresa holandesa DSM, que substituiu 13 etapas químicas do método tradicional, danosas ao ambiente, por apenas uma etapa de fermentação e duas enzimáticas feitas por leveduras engenheiradas. Isso gerou economia de 65% de energia e materiais, chegando a um produto final 50% mais barato, sem os danos ambientais do antigo método. A empresa também produz suplimentos alimentares como vitamina B12 através de bactérias dotadas de vias metabólicas sintéticas.

No que diz respeito a biosíntese de compostos farmacêuticos, a tecnologia do DNA recombinante tem sido usada a mais de três décadas para permitir que bactérias produzam moléculas farmacologicamente importantes como insulina e hormônio de crescimento humano. Progresso tem sido feito nos últimos anos na direção de compostos mais complexos, como terpenóides, poliquetídeos, peptídeos não-ribossomais e alcalóides. Abordagens em BS estão sendo usadas para ajustar os níveis e atividades de etapas individuais e componentes. Através do desenho racional e seleção de mRNAs com estrutura secundária definida e ativa, as **Ribozimas**, **Riboswitches** e **Aptâmeros** são pesquisados para uso terapêutico (Famulok et al., 2007) (BOX R). Mesmo antes de sua descoberta na natureza, estes já eram desenhados para controlar a expressão de genes reporter em resposta a ligação de pequenas moléculas, como os fluoróforos da Hoechst ou tetraciclina. Esses exemplos fundamentaram os princípios de como esses elementos podem ser usados para acoplar expressão gênica ao estado metabólico da célula. Estratégias similares podem tornar-se a base de sistemas de varredura em larga escala para identificar cepas engenheiradas capazes de biosintetizar a molécula de interesse.

BOX – Ribozimas, Riboswitches, Aptâmeros

Moléculas de RNA são capazes de desempenhar outros papéis além de transferir a informação genética do genoma para os sítios de tradução nos ribossomos. Dependendo da sequência de nucleotídeos, elas adquirem estruturas tridimensionais que conferem capacidades catalíticas e de ligação diversas. Os RNAs com propriedades catalíticas foram denominados Ribozimas e aqueles com estrutura com alta afinidade de ligação a outros compostos foram denominados de Aptâmeros (do grego "apta", encaixe). Essas foram descobertas feitas a partir da década de 60 e forneceram mais evidências sobre a evolução inicial do que consideramos vida e sistemas biológicos, no chamado "mundo dos RNAs". Desde então nossa compreensão da relação "sequência – estrutura" evoluiu bastante e hoje já é possível desenhar moléculas de RNA com propriedades específicas desejadas. É possivel inclusive criar sistemas autoreplicantes! Em 1990 foi desenvolvido um método para engenheirar ácidos nucleicos através de ciclos sucessivos de mutação e seleção *in vitro*, chamado SELEX (*systematic evolution of ligands by exponential enrichment*) aumentando em muitas ordens de grandeza a afinidade dos ligantes e capacidade catalítica dos elementos "filtrados" após vários ciclos de seleção. O processo tornou-se automatizado gradualmente e recentemente através da técnica de AptaBiD (*aptamer-facilitated biomarker discovery*) é possível selecionar o elemento de RNA em apenas alguns dias e não várias semanas como inicialmente. A indústria farmacêutica busca avidamente por novos fármacos dessa nova classe terapêutica e alguns já se encontram em testes clínicos. Algumas vantagens do uso de RNAs terapêuticos é a ausência de resposta imune e seu curto tempo de vida no organismo uma vez que é rapidamente degradado pelas nucleases endógenas (Berezovski et al., 2008).

Além de catálise e ligação com alta afinidade, os RNAs podem também funcionar como "interruptores", os chamados *Riboswitches*. Esses elementos de RNA tem a capacidade de controlar a tradução de uma fase aberta de leitura de acordo com a presença de ligantes específicos, que alteram a estrutura da porção não traduzida a montante do gene, permitindo ou não sua ligação ao ribossomo e consequêntemente controlando a tradução do mRNA. São outra classe de biopartes que podem ser usados para controlar o metabolismo e desde sua descrição inicial em 2002, já foram encontrados em diferentes Domínios da vida, controlando o metabolismo de diversos compostos como amino ácidos, vitaminas e hormônios de comunicação celular (Winkler et al., 2002).

Na área médica, existe uma forte demanda por novas terapias capazes, por exemplo, de virar o jogo contra a crescente resistência aos antibióticos disponíveis demonstrada por importantes isolados clínicos. O tratamento de câncer é uma outra área que carece de novas terapias, que sejam menos invasivas, mais eficiêntes e produzam menos efeitos colaterais. Nesse sentido, muitos grupos de pesquisa estão avançando de maneira promissora no desenvolvimento de novas terapias baseadas em BS. As estratégias podem focar no desenvolvimento de novos fármacos como também em ampliar a eficácia dos existentes através do enfraquecimentos das defesas naturais dos patógenos. Terapias mais efetivas estão sendo implementadas, na medida em que pesquisadores constroem dispositivos (moléculas, redes genéticas e organismos programáveis, por exemplo) para alterar processos causadores de doenças. As estratégias atuais abordam uma ampla gama desses processos, envolvidos em:

- Combate às doenças infecciosas
- Combate ao câncer
- Desenvolvimento de vacinas
- Engenharia de microbiomas
- Terapia celular e medicina regenerativa

Em um exemplo de como a BS está sendo usada para **combater infecções**, o bacteriófago (vírus que infecta somente bactérias) T7 foi modificado com sequências sintéticas para se replicar mais rapidamente no patógeno alvo e produzir a enzima dispersina B (DpsB), capaz de degradar biofilmes bacterianos (comunidades bacterianas associadas à superfície e encapsuladas por matriz extracelular). Esse duplo ataque foi capaz de matar >99,99% de bactérias em biofilmes, em um ciclo de ataque às células e dispersão da matriz (Lu & Collins, 2007). Um outro estudo de sucesso utilizou adjuvantes sintéticos para aumentar a eficiência de antibióticos já no mercado, através da interferência nos mecanismos de defesa bacterianos, a resposta SOS (Figura 14). O tratamento com esse fago aumentou significativamente a eficácia de três classes de antibióticos; dando um novo fôlego para terapias já existentes. Em ensaios *in vitro*, o uso combinado do fago com quinolina resultou em aumento de 5000x a morte de bactérias resistentes ao antibiótico usado isoladamente (Lu & Collins, 2009). Os vetores de doenças também podem ser atacados, como demonstrado no estudo com o mosquito da Malária, em que genes de endonuclease sintéticos foram usados para controlar o balanço sexual na população, demonstrando a viabilidade de se alterar populações, não somente indivíduos através de sequências auto-replicantes (Windbichler et al., 2011).

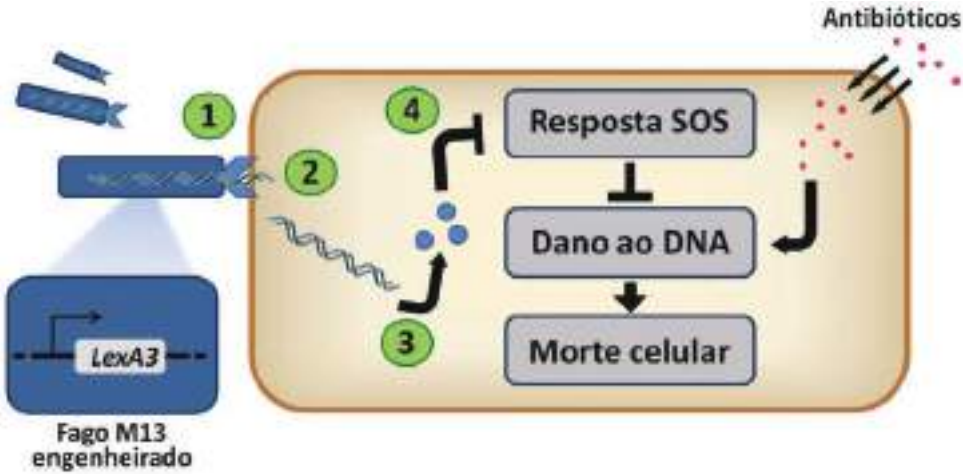


Figura 14. Combatendo infecções através de bacteriófagos engenheirados. Representação esquemática mostrando (1) infecção pelo fago, (2) entrega da construção sintética, (3) produção de LexA3 a partir da construção sintética, (4) Aumento do dano ao DNA e morte celular através da inibição da resposta SOS por LexA3. Modificado de Ruder et al., 2011.

Tratamento de câncer

Apesar do sucesso das terapias modernas, as principais intervenções terapêuticas -- cirurgia, radioterapia e quimioterapia -- ainda resultam frequentemente em dano considerável ao tecido saudável. Pesquisas atuais em BS buscam distinguir com precisão células saudáveis de doentes. Para isso, bactérias foram engenheiradas para identificarem e invadirem células tumorais. Em um estudo, a invasão foi desenhada para ocorrer somente em ambientes específicos relacionados a tumores, através do uso de promotores específicos ativados somente em condições de hipóxia por exemplo, um indicativo do ambiente tumoral (Anderson et al., 2006). Em estudo relacionado, utilizando **interferência de RNA** os invasores bacterianos foram programados para nocautear uma rede genética específica relacionada ao câncer (Figura 15). Nesse caso o cassete sintético codificava para um RNA curto do tipo grampo (*shRNA*), uma proteína invasina e a toxina *lysteriolysin O*, constituindo um coquetel capaz de silenciar genes em células tumorais com alta eficiência e especificidade (Xiang et al., 2006).

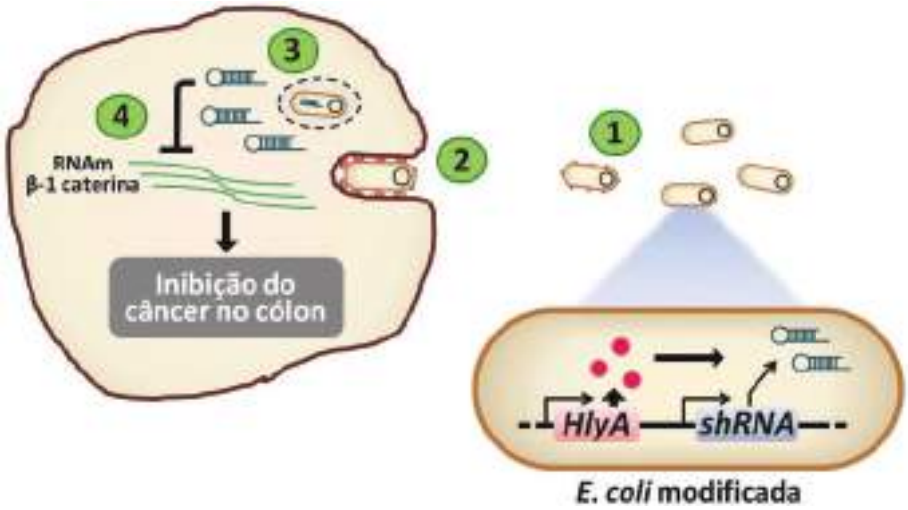


Figura 15. Bactérias programadas para combater o câncer através de interferência de RNA. Representação esquemática mostrando (1) produção de invasina pela *E. coli* engenheirada, (2) invasão bacteriana de uma célula tumoral, (3) liberação do RNA pequeno em forma de grampo, (4) Inibição do câncer de cólon por silenciamento gênico. Modificado de Ruder et al., 2011.

Desenvolvimento de vacinas

O desenvolvimento de novas vacinas é limitado pelos riscos associados com o uso de patógenos atenuados e dificuldades em se alterar a especificidade das vacinas atuais para novos alvos. Pensando nessas dificuldades pesquisadores desenvolveram **lipossomos** – vesiculas sintéticas consistindo em uma bicamada lipídica - para encapsular uma combinação dos componentes necessários para a transcrição e tradução em bactérias e um DNA codificando um antígeno modelo (β-galactosidade). O sistema produziu proteína funcional *in vitro*. Em camundongos vivos, esses

lipossomos expressando o antígeno geraram uma resposta imune humoral mais alta do que as vacinas controle (lipossomos com apenas o antígeno encapsulado, ou apenas com as proteínas de transcrição ou apenas com o DNA do antígeno). Este sistema pode facilmente ser alterado para outros antígenos substituindo o DNA molde e não carrega risco de infecção pelo patógeno atenuado (Amidi et al., 2011).

Avanços adicionais no campo podem vir da combinação de circuitos genéticos com avanços da engenharia genômica para o desenvolvimento de vacinas. Um exemplo disso é a exploração do fenômeno conhecido como viés de códon (do inglês *codon bias*), muito comum em todos organismos. Embora certos codons de DNA sejam sinônimos (diferentes codons podem codificar um único amino ácido), toda espécie tem uma preferência para um conjunto determinado de códons que consegue traduzir eficientemente em uma proteína. Para explorar isso um grupo trocou centenas de códons na sequência gênica que codifica uma proteína de um poliovírus, resultando em uma eficiência de tradução reduzida. O vírus resultante tornou-se atenuado em sua capacidade infectiva porém manteve sua capacidade imunogênica. A novidade veio do fato de que ao invés de fazer as mutações na sequência gênica original, o genoma viral foi sintetizado completamente *in vitro* já contendo as mutações desejadas e inserido em células vivas. A evolução dessa abordagem permitirá "reescrever" o genoma de diferentes vírus visando a obtenção de variantes atenuados (Coleman et al., 2008).

Engenharia de microbiomas

O microbioma humano – a comunidade microbiana associada ao corpo humano – é um ecossistema complexo crescentemente implicado como um regulador de nossa fisiologia. Já foram detectadas mais de 1000 espécies nessa comunidade e estima-se que seu número absoluto seja da ordem de 10 ou 100 vezes o de células humanas. Como os constituintes de um microbioma são tipicamente bem-tolerados (microorganismos comensalistas), são potencialmente excelentes vetores para apresentarem circuitos gênicos sintéticos para combaterem doenças e corrigir condições anômalas. Interações intra- e inter-específicas também desempenham um papel crítico em comunidades microbianas e podem ser exploradas. Nessas linhas, trabalhos atuais já utilizam *E. coli* para previnir a cólera induzindo interações sintéticas entre micróbios intestinais. Durante a infecção, o patógeno *Vibrio cholerae* secreta fatores de virulência, como a toxina da cólera, apenas quando em baixa densidade populacional. Para medir sua própria densidade, *V. cholerae* usa um sistema de percepção de quorum no qual secreta e detecta dois mensageiros (CAI-1 e AI-2) e quando ambos estão em alta concentração cessa a expressão dos fatores de virulência. Uma *E. coli* engenheirada para produzir AI-2 e CAI-1 quando injetada em camundongos foi capaz de aumentar a sobrevida destes drasticamente e reduzir a ligação da toxina ao epitélio intestinal em 80%.

Alternativamente, o microbioma de um paciente poderia ser alterado para fornecer moléculas terapêuticas diretamente ao corpo. Por exemplo, cepas bacterianas comensais foram modificadas geneticamente para secretar moléculas chave para o tratamento de doenças, incluindo insulina para diabetes, um inibidor de fusão do HIV capaz de prevenir a infecção por esse vírus e interleucina-2 para imunoterapia. Apesar de demonstrarem a expressão de importantes moléculas terapêuticas, melhores

resultados poderiam se alcançados utilizando circuitos sintéticos. Colocando-se, por exemplo, a expressão de moléculas terapêuticas sob o controle de sensores celulares que detectem condições aberrantes, a expressão gênica poderia ser ativada e ajustada de maneira coordenada apenas quando intervenções moleculares fossem necessárias, reduzindo a carga metabólica sobre as bactérias e aumentando assim sua habilidade de manutenção no microbioma.

Terapia celular e medicina regenerativa

A introdução de células engenheiradas no corpo humano para o tratamento de doenças é promissora, apesar dos desafios que ainda devem ser superados como a falta de controle do comportamento celular e seu fenótipo pós-implantação. Uma possível solução seria dotar essas células de circuitos sintéticos, permitindo sistemas de controle mais sofisticados. Porém, a grande maioria dos circuitos gênicos sintéticos desenhados até o momento está limitada aos microrganismos. A recente expansão dos circuitos para células de mamíferos, no entanto, abriu caminho para novas e melhores terapias (revisado em Weber & Fussenegger, 2006). Para tal, o controle de genes específicos mostrou-se crucial para que as terapias fossem efetivas. Um exemplo é o sistema de interruptor gênico que liga proteínas inibitórias a um RNA de interferência, através de um shRNA. A expressão gênica é ativada adicionando-se um indutor, que controla os elementos repressores ao nível transcricional, enquanto simultaneamente desliga o componente RNAi para permitir que o transcrito seja retido e traduzido (Figura 16). Esse interruptor é capaz de repressão acima de 99%, bem como de regular a expressão do gene de interesse. Teoricamente algo semelhante pode ser construído para qualquer gene de interesse, bem como o potencial para o uso tecido-específico (através do uso de promotores tecido-específicos). Esse interruptor foi validado em células de camundongo e humano e o controle rigoroso e reversível da expressão celular pode ser usado em diferentes aplicações em terapia celular, bem como para determinar se um fenótipo de doença é decorrente de alterações na expressão gênica.

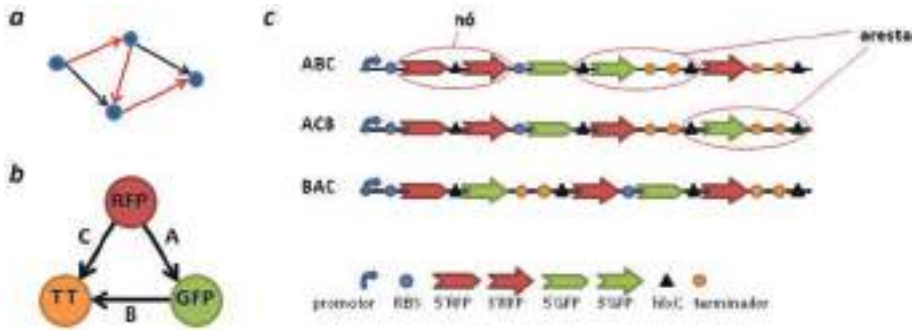


Figura 16. Interruptor genético para controlar a expressão em mamíferos. O circuito genético sintético foi usado para controlar a produção de um pigmento, comprovando o conceito do uso de indutores aliados a RNAs regulatórios. Modificado de Ruder et al., 2011.

O controle não necessariamente precisa ser feito ao nível transcricional, mas pode também ser traducional como no caso em que uma **ribozima** responsiva à uma citocina de crescimento foi capaz de controlar a proliferação de linfócitos T.

A customização de células de acordo com o estado fisiológico de cada paciente será crucial no campo da medicina regenerativa, onde as terapias provavelmente envolverão tecidos criados a partir das próprias células tronco do paciente. Embora o corpo adulto mantenha algumas linhagens de células tronco clinicamente úteis (por exemplo, as hematopoiéticas e adipogênicas), muitas outras são inacessíveis. Com o desenvolvimento de células tronco pluripotentes induzidas (iPSCs), os pesquisadores têm agora potencialmente condições de gerar *in vitro* qualquer tipo celular humano. Isso foi conseguido através da inserção e expressão de apenas quatro genes (*klf4*, *c-myc*, *oct4* e *sox2*) em células maduras de um paciente, um feito que traz enormes potencialidades mas também muitas preocupações. Por exemplo, cópias extras desses genes introduzidas viralmente ficam inseridas permanentemente no genoma celular, o que pode favorecer a formação de tumores. Pensando nessas limitações, uma abordagem de BS foi utilizada para transfectar quimicamente RNAs sintetizados *in vitro* em células e esses funcionaram como transcritos para os quatro genes chave citados acima. Uma vez nas células, os transcritos são traduzidos em proteínas que induzem a pluripotência sem a integração de genes extras no genoma. Usando esse método, os pesquisadores criaram iPSCs com maior rapidez e eficiência do que empregando a transfecção viral. Esse exemplo ilustra bem como abordagens de BS podem trazer mais segurança aos procedimentos terapêuticos.

Biocomputação

De acordo com o verbete da Wikipedia, um computador é uma máquina programável desenhada para, automaticamente, realizar uma sequência de operações aritiméticas ou lógicas. Um computador pode prover-se de inúmeros atributos, dentre eles armazenamento de dados, processamento de dados, cálculo em grande escala, desenho industrial, tratamento de imagens gráficas, realidade virtual, entretenimento e cultura.

Os primeiros computadores analógicos surgiram no século XVII e eram capazes de realizar as funções básica de somar, subtrair, multiplicar e dividir. Mas foi na II Guerra Mundial, em meados do século XX, que realmente nasceram os computadores atuais. A Marinha dos Estados Unidos, em conjunto com a Universidade de Harvard, desenvolveu o computador Harvard Mark I, projetado pelo professor Howard Aiken, com base no calculador analítico de Babbage. O Mark I ocupava 167m^2 e pesada cerca de 30 toneladas aproximadamente, conseguindo multiplicar dois números de dez dígitos em três segundos. Seu funcionamento era parecido com uma calculadora simples de hoje em dia. Nem é preciso citar que esta tecnologia continuou sendo aprimorada, e que algumas décadas depois hoje se discute processores quânticos e computação em nuvem.

Uma plataforma diferente daquela "baseada em silício" que estamos acostumados são os biocomputadores. Em 1994, em um experimento muito elegante, Leonard Adleman desenvolveu o primeiro experimento envolvendo um computador de DNA para resolver o problema do Caminho Hamiltoniano (por exemplo, o problema do

vendedor viajante, em que ele deseja visitar um conjunto de N cidades (vértices), passando por cada cidade exatamente uma vez, fazendo o caminho de menor tamanho possível, como mostrado na Figura 17.

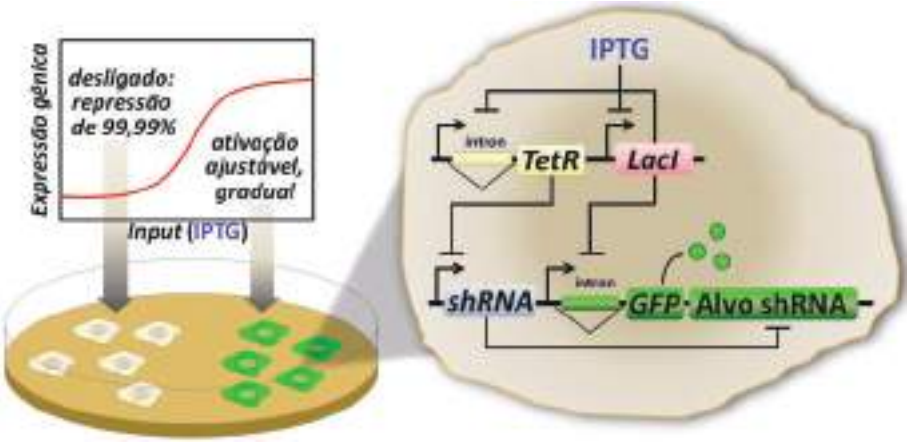


Figura 17. Construção de DNA que codifica um problema Hamiltoniano com três nós. *a*. Um caminho Hamiltaniano é aquele que visita cada ponto somente uma vez. *b*. O grafo contendo o caminho Hamiltoniando começa no nó RFP, procedendo para o nó GFP e terminando no nó TT. *c*. Construção ABC rerepresenta a solução para o problema dos três nós. Os três fragmentos de DNA flanqueados por hixC estão na ordem e orientação corretas, de maneira que os genes GFP e RFP estão intactos. ACB possui o gene RFP intacto, porém o gene GFP está truncado, por fim, a construção BAC não possui nenhum gene intacto.

Existem múltiplas possibilidades de se construir um computador baseado em DNA. A maioria funciona utilizando as portas lógicas (AND, OR, NOT) associadas à lógica digital utilizando como base o DNA, como o exemplo dos contadores bacterianos citados anteriormente. Porém os primeiros computadores moleculares baseados em DNA eram reações *in vitro* utilizando, por exemplo, enzimas de restrição, ligases e DNA (Benenson et al. 2001). Através da mistura desses componentes e reações em cascata de digestão, ligação e hibridização, o *output* final é uma molécula detectável que representa o resultado computacional.

Em 1994, Leonard Aldleman foi capaz desenvolver um computador *in vitro* baseado em DNA para solucionar o problema do Caminho Hamiltoniano, porém apenas em 2009, Baumgardner e colaboradores conseguiram resolver um problema complexo *in vivo* em *E. coli*. Porém, para entender, é necessária uma série de abstrações para tornar sequências de DNA em vértices e arestas de um caminho hamiltoniano (ver Figura 17a). A primeira abstração trata seguimentos de DNA como as arestas de um determinado grafo. As arestas de DNA são flanqueadas por sítios *hixC* que podem ser embaralhados por uma recombinase Hin, criando diversas ordens e orientações randômicas para as arestas do grafo. A segunda abstração está relacionada com os nós, com exceção do nó terminal, em que um nó é um gene dividido ao meio por uma sequência *hixC*. Os autores conseguiram construir enzimas funcionais portando tais sequencias codificadas no DNA. Dessa maneira, a primeira metade (5′) de um nó é

encontrada na aresta de DNA que termina em um nó, enquanto a segunda metade (3′) do gene é encontrado em uma aresta de DNA que se origina no nó. Calma, não é fácil entender mesmo, é preciso pensar e abstrair veja a Figura 17b e 17c.

A Figura 17a mostra o grafo com os 3 nós e as 3 arestas que foram escolhidos para serem codificados no computador bacteriano. O gráfico contém um único caminho hamiltoniano que começa no nó RFP, viajando pela aresta A até o nó GFP, e utilizando a aresta B até alcançar o nó final TT. A aresta C, de RFP até TT é um detrator. A Figura 17b ilustra como as construções de DNA foram utilizadas para solucionar o problema do Caminho Hamiltoniano com um controle positivo e duas configurações sem soluções. Como as soluções precisam originar no nó RFP e terminar no nó GFP, a aresta A de DNA contêm a extremidade 3′ da metade de RFP seguida pela extremidade 5′ de GFP. A aresta B de DNA se origina em GFP e termina em TT, dessa maneira, esse fragmento de DNA possui 3′GFP seguido de um terminador de transcrição duplo. A aresta C se origina em uma metade 3′ de RFP e termina em TT. Finalmente, como os genes codificadores para RFP e GFP estão intactos, com promotores e RBS, e seguidos de um terminador de transcrição, colônias ABC expressam fluorescência vermelha e verde, dessa maneira, possuem aparência amarela.

A programação de bactérias para computar soluções de problemas complexos pode oferecer as mesmas vantagens dos computadores de silício que usamos atualmente, porém, com as seguintes características adicionais: (i) sistemas bacterianos são autônomos, eliminando a necessidade de intervenção humana, (ii) computadores bacterianos podem se adaptar a condições flutuantes, evoluindo para resolver desafios de determinados problemas e (iii) o crescimento exponencial de bactérias continuamente aumenta o número de processadores trabalhando em um problema (Baumgardner et al., 2009).

Comunidade da biologia sintética

A biologia sintética enfrenta um desafio de não ser estigmatizada como foram os alimentos transgênicos no mundo e para evitar isso, a comunidade procura estabelecer relações claras, democráticas e abertas com vários setores da sociedade. Pretendemos, a seguir, dar vários exemplos de como a ciência sai dos laboratórios para não apenas gerar tecnologia mas também para difundir e discutir seus princípios com filósofos, legisladores e o cidadão comum.

Do-It-Yourself Biology e a Ciência Cidadã

Neste momento, em algum lugar dos Estados Unidos, da Inglaterra ou até no Brasil, algum biólogo sintético amador está realizando um experimento na sua cozinha ou garagem. Nos últimos dois anos, entusiastas da biologia molecular têm se juntado para montar organizações de BS amadora, como o DIYBio ou o Biocurious (*biocurious.org*), em que os membros se reúnem em pubs e churrascos para discutir os últimos experimentos realizados nas suas próprias garagens. Inspirados pelos grandes avanços realizados em garagens pelos fundadores de atuais gigantes da informática como Apple

e Dell, os também chamados *biohackers* ou *biopunks* pretendem revolucionar a ciência através de experimentos e ideias não convencionais aplicados a BS.

Este movimento também se caracteriza pela chamada ciência cidadã (tradução livre de *citizen science*), em que os cidadãos ativamente participam no papel de desenvolver a ciência e as novas tecnologias. Além disso, a ciência cidadã estimula o apoio da população à ciência, o desenvolvimento do pensamento científico nas pessoas, além de introduzir novas idéias de diferentes disciplinas ao assunto. Utilizando a Internet como plataforma, um simples projeto de ciências pode envolver dezenas, centenas e milhares de pessoas de diversas formações no mundo dispostas a criar algo novo e interessante. O interesse crescente pela biotecnologia, têm fomentado também o desenvolvimento de laboratórios públicos nos quais, o biologista sintético amador paga uma pequena taxa para utilizar kits e soluções além de toda a infra-estrutura e segurança de um laboratório de Biologia Molecular (*diybionyc.blogspot.com* | *diybio.madlab.org.uk*). Aqui no Brasil, uma iniciativa nesse sentido, para engenheiros amadores, é o Laboratório de Garagem (*labdegaragem.com.br/espaco*). Nos EUA esses grupos de cientistas conseguem apoio financeiro para executar projetos inovadores utilizando o sistema de *crowdfunding*, muitas vezes conseguindo mais do que objetiveram inicialmente como o grupo que está desenvolvendo o *real vegan cheese*.

Porém, junto com o crescimento da ciência cidadã, tem também aumentado a preocupação do governo americano e do FBI a respeito do que os biohackers estão fazendo. Por incrível que pareça, agentes do FBI têm comparecido a reuniões do DIYbio para entender o que as pessoas estão fazendo e qual a possibilidade de utilização das ferramentas para o bioterrorismo (Ledford, 2010; Editorial, 2010). A comunidade DIYbio teme que o foco constante em possíveis atividades terroristas desvie a atenção dos tópicos importantes relacionados com biossegurança: como o descarte de bactérias geneticamente modificadas, normatização/legalização de laboratórios caseiros e equipamentos de segurança mais acessíveis e baratos. Muitas vezes o que tem acontecido é que não existe nenhum tipo de norma ou lei que fale a respeito de laboratórios caseiros para a utilização de bactérias geneticamente modificadas.

É incrível o que está acontecendo neste momento. Não só está ocorrendo uma explosão de conhecimento e técnicas no mundo científico, mas também a população está cada vez mais interessada em fazer parte dessas descobertas e fazer da ciência um exercício cotidiano.

Aspectos legais, bioética e biossegurança

A criação de organismos sintéticos que não são encontrados na natureza nos leva a questões sobre o papel e responsabilidade de seres humanos na criação de novas formas de vida. A possibilidade dessa evolução *in silico* e *in vitro* desafia nossa compreensão do mundo natural e nosso lugar nele. A BS levanta questões difíceis sobre onde traçar a linha do que é "natural" e o que não é ou mesmo se essa linha deve ser traçada a princípio. As inovações surgem mais rápido do que conseguimos entendê-las e controlá-las. Em maio de 2010, após o anúncio do primeiro transplante de um genoma sintético pelo grupo do Craig Venter, o presidente norte-americano Barack Obama

pediu a sua comissão de bioética um estudo sobre os potenciais benefícios e riscos na medicina, meio ambiente e segurança dessas novas tecnologias da BS. Além disso, o presidente pediu a comissão de ética por recomendações para que o Governo Federal possa garantir que o país usufrua da melhor maneira possível dos benefícios desse novo campo da ciência, identificando os limites éticos e minimizando possíveis riscos.

Principalmente no EUA, existe um grande medo a respeito do bioterrorismo. Ferramentas da BS poderiam maximizar o efeito e baratear o desenvolvimento de novas armas químicas. Além disso, existe o medo de que estas novas formas de vida que estão sendo criadas em laboratório possam ser liberadas no meio ambiente gerando impactos imprevisíveis.

Após seis meses de estudos e encontros públicos, a Comissão Presidencial de Bioética lançou 18 recomendações (*bioethics.gov/cms/synthetic-biology-report*) para o governo federal americano declarando, principalmente, que nenhum tipo de moratória ao novo campo de pesquisa é necessário; no momento presente a tecnologia apresenta poucos riscos porque ainda está em sua infância. Dessa maneira os próprios biológicos sintéticos devem regular o desenvolvimento deste campo e para isso, incluíram um treinamento obrigatório de bioética para os pesquisadores da área. Entre as outras recomendações estão, por exemplo, que os microrganismos sintéticos possuam genes suicidas ou possuam a necessidade de algum nutriente laboratorial para a sobrevivência que de alguma maneira limitem a sua sobrevivência fora do laboratório. Outras recomendações, como maior transparência do governo em relação aos projetos de BS financiados com dinheiro público e um maior período para a análise de risco da área.

Além disso, a comunidade DIYbio foi contemplada no relatório sendo que foi considerado que estes não oferecem riscos para a sociedade já que não apresentam infraestrutura para a construção de novos organismos. De acordo com a comissão, não há necessidade de impor normas diferenciadas para esta comunidade. Já os grupos civis ETC e o SBSTTA que também contam com representantes governamentais são bem mais críticos e pedem regulamentação mais severa e até mesmo que a BS seja banida como um todo. Argumentam que a facilidade de acesso a essas tecnologias permite a ação de *biohackers* e bioterroristas. De fato o risco existe mas não pode ser contido proibindo atividades de pesquisa acadêmicas e comerciais. Nesse sentido outros setores da sociedade estão se organizando para regulamentar as atividades que envolvem BS, como por exemplo, a EuropaBio, organização das indústrias de biotecnologia dos países europeus. A ONU também já divulga padrões de segurança a serem adotados pelos países membros. A ideia é tomar ações preventivas e não manter um controle maior do que é feito para evitar o descompasso entre legislação e tecnologia ocorrido no passado, como por exemplo, na época da introdução de plantas geneticamente modificadas.

iGEM e SynbioBrasil

Todo verão no hemisfério Norte, centenas de equipes de estudantes de graduação encaram o desafio de compreender, projetar e implementar novos sistemas biológicos sintéticos utilizando partes-padrão de DNA e operá-los em células vivas, com o

objetivo de solucionar problemas no mundo (ex. alimentação, energia, meio ambiente, medicina, processamento de informações), tendo como base princípios de engenharia. No final do verão, estas equipes se reúnem no MIT para se conhecerem e apresentarem os projetos como parte da competição internacional de BS iGEM (*International Genetically Engineered Machine*). Esta competição tem sido um exemplo de como a BS têm avançado rapidamente e de forma democrática. A competição iniciou-se em 2004 com 5 times e na edição de 2010 já contou com mais de 130 times de diferentes continentes. Devido a este aumento vertiginoso, em 2011 a competição foi dividida em 3 regiões: Europa, Américas e Ásia. Dessa maneira, *Jamborees* regionais em outubro qualificam para a participação da competição mundial no MIT em novembro.

Os projetos são muito interessantes e muitos se destacam por serem ideias simples que possuem um grande impacto na sociedade. Um dos projetos vencedores do iGEM2010 foi "agrEcoli" da Universidade de Bristol. A "agrEcoli" é um dispositivo baseado na bactéria *E. coli* capaz de detectar e sinalizar a presença de nitrato no solo. Dessa maneira, fazendeiros poderiam mapear a concentração de nutrientes no campo e aperfeiçoar o uso de fertilizantes. Este tipo de tecnologia reduz o gasto dos agricultores com fertilizantes além de reduzir os impactos causados na natureza pelo excesso de fertilizantes, entre eles, a eutrofização dos rios. Em uma construção simples, um promotor que responde a concentração de nitratos regula uma proteína GFP que sinaliza a presença de certa quantidade de nitrato. De uma maneira similar, o time da Universidade de Pequim no iGEM de 2010 desenvolveu um biosensor e um bioabsorvente de metais pesados. Outros exemplos mais voltados à Informática tiveram destaque no iGEM2010, como a codificação e armazenamento de dados em *E. coli* (bio-criptografia) e a resolução de um quebra-cabeças Sudoku por *E. coli* (bio-processamento de dados). Mais informações sobre estes e todos os projetos desenvolvidos no iGEM, podem ser encontradas na *wiki* de cada equipe no site do iGEM (igem.org).

No Brasil, o primeiro passo foi dado pela Unicamp que em 2009 representou o Brasil pela primeira vez na competição com o projeto "Microguards", em que criou um dispositivo baseado em *E.coli* e *Saccharomyces cerevisae* para o reconhecimento e destruição de microrganismos contaminantes em biorreatores, como os *Lactobacillus*. Esse dispositivo pode ajudar a evitar o desperdício de açúcar utilizado pelos contaminantes para, por exemplo, aumentar a produção de etanol no país. Desde 2011, alunos da Universidade de São Paulo (USP) tem se organizado para discutir sobre BS no Clube de Biologia Sintética da USP, visando elaborar projetos para o iGEM e divulgar as informações de BS através do site SynbioBrasil (synbiobrasil.org). Um dos principais objetivos do SynbioBrasil é ajudar e estimular a formação de times brasileiros para competir no iGEM. Venha participar!

Perspectivas

A Biologia Sintética se distingue da engenharia e de outras disciplinas científicas tanto em sua abordagem quanto em seu objetivo. Esse novo campo do conhecimento aborda questões científicas da biologia, porém utiliza conceitos de engenharia para tal. Dessa maneira, a BS deveria ser considerada uma disciplina híbrida, que combina

elementos de ciência e engenharia para alcançar o seu objetivo de engenheirar sistemas biológicos. Porém, organismos vivos são muito complexos e atualmente ainda existe uma lacuna de conhecimento sobre o funcionamento destes sistemas. Uma razão para isso, é que sistemas biológicos possuem um alto nível de integração, muito maior do que qualquer outro sistema não-vivo. Além disso, estes sistemas biológicos não foram desenhados por humanos e, tampouco, foram desenhados para serem facilmente compreendidos.

Nossa capacidade de engenheirar sistemas biológicos previsíveis está apenas no início. Podemos realizar modificações genéticas, porém, o efeito na biologia celular ainda é, em geral, imprevisível. Por exemplo, a Amyris possui capacidade de desenhar racionalmente centenas de linhagens geneticamente modificadas por semana, sendo que foi verificado que apenas 0,0016% das linhagens são realmente melhores em performance em um fermentador do que a linhagem parental (Gardner, 2013).

Por muito tempo praticamos uma ciência reducionista que tenta explicar os sistemas vivos a partir de suas pequenas partes. Atualmente, temos uma capacidade impressionante de descrever partes, mas um ínfimo conhecimento de como montar estas partes inanimadas em sistemas vivos artificiais. Ou seja, ainda não sabemos como conferir vida às partes, ainda que elas sejam fisicamente dinâmicas como proteínas e DNA. Por isso, as atividades atuais de BS ainda se dão no interior de um contexto celular existente. Esse fato tem um impacto profundo em nossa capacidade de abstração de componentes biológicos em dispositivos e módulos e a sua utilização no desenho de sistemas artificiais biológicos.

No desenho, fabricação, integração e teste de novos dispositivos celulares, biologistas sintéticos precisam utilizar ferramentas e métodos oriundos da biologia experimental. Entretanto, a biologia experimental ainda não progrediu ao ponto de fornecer uma fundação sólida para a BS da mesma maneira que a física de estados sólidos forneceu para a engenharia elétrica, por exemplo, para o desenvolvimento dos computadores. Em quase todas as áreas de engenharia existem ferramentas de simulação que permitem definir quais componentes são necessários para montar e obter um sistema capaz de realizar a função desejada. Como resultado dessa falta de ferramentas precisas, o desenho de sistemas biológicos sintéticos se tornou um processo interativo de modelagem, construção e testes experimentais até que o sistema possua o comportamento desejado. O processo se inicia com o desenho abstrato de um dispositivo, módulo ou organismo, que é, normalmente, guiado por um modelo matemático. Entretanto, tais tentativas iniciais raramente resultam em uma implementação funcional devido a conhecimentos biológicos limitados. O redesenho racional utilizando modelos matemáticos pode melhorar o comportamento do sistema, porém, a evolução dirigida é uma estratégia complementar que pode resultar em modificações inesperadas e benéficas para o sistema. Finalmente, esse novo sistema é experimentalmente testado e, se necessário, o processo se repete (Andrianantoandro et al., 2006).

Muitos dos sistemas sintéticos biológicos têm sido construídos dessa maneira porque esta é uma metodologia muito tolerante a incerteza e a falta de conhecimento que temos atualmente. No futuro, biologistas sintéticos serão beneficiados por novas tecnologias e métodos para gerenciar a incerteza e a complexidade nesses sistemas. Não podemos esquecer que ferramentas de desenho biológico ainda estão em sua

infância. Porém, modelos metabólicos que incorporam composição celular e regulação genética têm se tornado relativamente preditivos e podem um dia serem capazes de predizer o nível de expressão gênica necessário para alcançar determinado nível de fluxo através de uma reação ou via metabólica.

É possível imaginar que no futuro o microrganismo será feito sob medida para determinada função, seja ela a produção de um produto químico ou o tratamento de doenças ou câncer. Por exemplo, um microrganismo poderá ser construído para produzir um produto químico a partir de determinado material, do mesmo modo que um engenheiro químico constrói uma refinaria e outras indústrias químicas a partir de operações unitárias. As características químicas e físicas do produto e da matéria-prima deverão ser consideradas para o desenho do microrganismo a fim de minimizar os custos de produção e purificação. O envelope celular será desenhado para ser resistente ao determinado produto químico, além disso, a parede celular poderia ser desenhada para o organismo ser tolerante as condições de processamento industrial, maximizando a reciclagem das células nos sistemas. Transportadores celulares específicos serão incorporados na membrana para bombear e manter o produto para fora da célula, além de importar a matéria-prima desejada. A via metabólica será construída a partir da secretária de partes padrão, e desenhada para maximizar o rendimento e produtividade. Caso uma determinada enzima não exista para uma reação particular, será utilizado um software para o desenho auxiliado por computador (CAD) (Keasling, 2010).

Uma vez que a célula tenha sido desenhada no computador, um sistema de controle genético será desenhado para controlar todos os genes, no tempo correto e a níveis apropriados. Biosensores podem ser adicionados para o controle de pH e temperatura do sistema. Redundâncias no sistema de controle genético serão engenheiradas para garantir que os parâmetros programados sejam mantidos independentemente dos estados transientes do processo de produção. Simulações de cenários poderão ser realizadas, incluindo total falha do sistema de controle. A segurança ambiental e dos operadores da planta de produção serão critérios essências para o desenho do microrganismo. Quando todos os controles genéticos forem definidos e testados, o cromossomo(s) final será desenhado e construído. O cromossomo será então encomendado para uma empresa produtora de DNA. Além disso, poderão haver empresas especializadas em diferentes aspectos da síntese, como por exemplo, uma empresa será responsável pelo cromossomo, outra pela membrana e outra pela parede celular e outra responsável por colocar todos os elementos necessários para "ligar" a célula (Keasling, 2010).

Novos grupos de pesquisas e empresas apresentam objetivos e desafios fascinantes. As possibilidades são infinitas. A produção de novos químicos e biocombustíveis tende a utilizar resíduos agroindustriais e domésticos, eletricidade ou diretamente a luz solar, para não competir com a indústria alimentícia. A indústria farmacêutica tende a substituir outros produtos naturais hoje extraídos de plantas por análogos sintéticos com propriedades otimizadas e processos de produção muito mais baratos, utilizando açúcar como matéria-prima. É esperado também que novas terapias mais eficazes e com menos danos colaterais aos pacientes sejam colocadas em prática. Imagine um iogurte contendo um *Lactobacillus* modificado capaz de realizar um *check-up* e fornecer um diagnóstico completo, diariamente, de acordo com as cores

das suas fezes. Através de um sinal de uma doença, como um vírus, uma bactéria ou uma toxina, um *Lactobacillus* modificado seria capaz de acionar a produção de um pigmento específico para sinalizar uma determinada doença. Nesse sentido uma bactéria capaz de responder a diferentes estímulos produzindo diferentes cores foi criada pelo time da Universidade de Cambridge vencedor do iGEM 2009 (http://2009.igem.org/Team:Cambridge)

É difícil encontrar limites para as novas possibilidades da biologia sintética: é o caso de uma criação de peixes fotossintetizantes. O laboratório da Prof. Pamela Silver isolou e transplantou os elementos de produção de glicose e frutose do Ciclo de Calvin em células de mamíferos (Agapakis, 2011). A pesquisa continua, e o peixe ainda precisa de comida, mas imagina-se uma versão no futuro em que não haja necessidade. Mas essa foi a prova de conceito de que animais poderiam ser transformados em fotossintetizantes, Mais ainda, isso poderia ser uma prova de conceito da nossa capacidade de revolucionar a produção de alimentos no futuro.

O desenvolvimento de computadores que, ao invés de microchips de silício, usem microbiochips fotossintetizantes ainda é um sonho distante. Porém, dispositivos genéticos, como contadores, interruptores, osciladores, portas lógicas... já podem ser utilizadas em diversas outras aplicações. Neste novo mundo da biologia sintética, microrganismos podem ser dispositivos que se auto-multiplicam, se auto-regulam e evoluem em escala microscópica, além de possuir uma linguagem fascinante de programação. Nessa nova plataforma, o software pode construir o próprio hardware a partir de açúcar e nutrientes.

Quão rápido as novas tecnologias surgirão ninguém sabe, porém que muitas já fazem parte do nosso cotidiano, é fato. Já sintetizamos DNA para projetos de biotecnologia e já utilizamos dispositivos sintéticos para melhorar a produção de biocombustíveis e no desenvolvimento de novas terapias. Por exemplo, recentemente, a Amyris construiu a primeira planta industrial de farneceno em Brotas, no interior de São Paulo. Quanto mais a sociedade se informar e debater o assunto, estaremos mais preparados para tomar decisões embasadas sobre como regulamentá-las e utilizá-las. Uma grande vantagem da biologia sintética atualmente é a forma transparente e democrática como ela é realizada.

No Brasil, a criação de redes ou centros de pesquisas que consigam integrar diferentes competências necessárias para a formação de grupos de pesquisa em biologia sintética é mais do que urgente. Para garantir competividade internacional, são necessárias políticas públicas para diminuir custos e velocidade de P&D, cursos específicos de formação de recursos humanos em biologia sintética ou bioengenharia (o MIT tem um curso de bioengenharia desde 1998, por outro lado, no Brasil não há nenhum), além de incentivos para a atração e/ou manutenção de talentos brasileiros ou estrangeiros para nossas universidades.

Bibliografias

ADLEMAN, L.M. (1994). Molecular computation of solutions to combinatorial problems. *Science* 266:1021-1024.

AGAPAKIS, C.M., NIEDERHOLTMEYER, H., NOCHE, R.R., et al. (2011). Towards a synthetic chloroplast. *PLoS One* 6:e18877

AMIDI, M., DE RAAD, M., CROMMELIN, D.J., et al. (2011). Antigen-expressing immunostimulatory liposomes as a genetically programmable synthetic vaccine. *Syst Synth Biol* 5:21-31.

ANDERSON, J.C., CLARKE, E.J., ARKIN, A.P., VOIGT, C.A. (2006). Environmentally controlled invasion of cancer cells by engineered bacteria. *J Mol Biol* 355:619-27.

ANDRIANANTOANDRO, E., BASU, S., KARIG, D., WEISS, R. (2006). Synthetic biology: new engineering rules for an emerging discipline *Molecular Systems Biology* 2: 2006.0028.

ATSUMI, S., HANAI, T., LIAO, J.C. (2008). Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. *Nature* 451:86-90.

BAUMGARDNER, J., ACKER, K., ADEFUYE, O., et al. (2009). Solving a Hamiltonian Path Problem with a bacterial computer. *J Biol Eng.* 3:11.

BARRANGOU, R., FREMAUX, C., DEVEAU, H., et al. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315:1709-12.

BENENSON, Y., PAZ-ELIZUR, T., ADAR, R., et al. (2001). Programmable and autonomous computing machine made of biomolecules. *Nature* 414: 430-434.

BEREZOVSKI, M.V., LECHMANN, M., MAK, T.W., KRYLOV, S.N. (2008). Aptamer-facilitated biomarker discovery (AptaBid). *J Am Chem Soc* 130: 9137-43.

BERGIN J. (2011). *Synthetic Biology: Emerging Global Markets.* BCC

CAMERON, D.E., BASHOR, C.J., COLLINS, J.J. (2014). A brief history of synthetic biology. *Nat Rev Microbiol* 12: 381-90.

CHARLES, W., SCHMIDT, M.S. (2010). Synthetic Biology: Environmental Health Implications of a New Field. *Environ Health Perspect*. 118: A118–A123.

CELLO, J., PAUL, A.V., WIMMER, E. (2002). Chemical synthesis of poliovirus cDNA: generation of infectious virus in the absence of natural template. *Science* 297: 1016-8.

COLEMAN, J.R., PAPAMICHAIL, D., SKIENA, S., et al. (2008). Virus attenuation by genome-scale changes in codon pair bias. *Science* 320: 1784-7.

COLLINS, C.H., ARNOLD, F.H., LEADBETTER, J.R. (2005) Directed evolution of *Vibrio fischeri* LuxR for increased sensitivity to a broad spectrum of acyl-homoserine lactones. *Mol Microbiol* 55:712-723.

DANINO, T., MONDRAGÓN-PALOMINO, O., TSIMRING, L., HASTY, J. (2010). *Nature* 463:326-330.

DiCARLO, J.E., NORVILLE, J.E., MALI, P., et al. (2013). Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. *Nucleic Acids Res* 41:4336-43.

DUNLOP, M.J., KEASLING, J.D., MUKHOPADHYAY, A. (2010). A model for improving microbial biofuel production using a synthetic feedback loop. *Syst Synth Biol* 4:95-104.

EDITORIAL (2010). Garage biology *Nature* 467:634-634.

ENDY, D. (2005). Foundations for engineering biology. *Nature*. 438:449-53.

FAMULOK, M., HARTIG, J.S., MAYER, G. (2007). Functional aptamers and aptazymes in biotechnology, diagnostics, and therapy. *Chem Rev*.107:3715-43.

FRIEDLAND, A., LU, T., WANG, X., et al. (2009). Synthetic gene networks that count. *Science* 324:1199-1202.

GARDNER, T.S., CANTOR, C.R., COLLINS, J.J. (2000). Construction of a genetic toggle switch in *Escherichia coli*. *Nature* 403:339-42.

GARDNER, T.S. (2013). Synthetic biology: from hype to impact. Trends in Biotechnology 31:123-125.

GIBSON et al. (2010). Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* 329:52-56.

HUTTENHOWER, C., GEVERS, D., KNIGHT, R., et al. (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486:207-14.

JIANG, W., BIKARD, D., COX, D., et al. (2013). RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat Biotechnol* 31:233-9.

KEASLING, J.D. (2010). Manufacturing molecules through metabolic engineering. *Science* 330:1355-1358.

KIRBYB, J., KEASLING, J.D. (2008). Metabolic engineering of microorganisms for isoprenoid production. *Nat. Prod. Rep.* 25:656-661.

KNIGHT, T.F. (2003). Idempotent Vector Design for Standard Assembly of BioBricks. *Tech. rep., MIT Synthetic Biology Working Group Technical Reports*. http://hdl.handle.net/1721.1/21168

LARTIGUE, C., GLASS, J.I., ALPEROVICH, N., et al. (2007). Genome transplantation in bacteria: changing one species to another. *Science* 317:632-8.

LEON, C.Y., KOSURI, S., ENDY, D. (2005). Refactoring bacteriophage T7. *Mol Syst Biol* 1: 2005-18.

Ledford H (2010). Garage biotech: Life hackers *Nature* 467:650-652.

LEVY, S., SUTTON, G., NG, P.C., *et al* (2007). The diploid genome sequence of an individual human. *PLoS Biol* 5:e254.

LU, T.K., COLLINS, J.J. (2007). Dispersing biofilms with engineered enzymatic bacteriophage. *Proc Natl Acad Sci U S A*. 104:11197-202.

LU, T.K., COLLINS, J.J. (2009). Engineered bacteriophage targeting gene networks as adjuvants for antibiotic therapy. *Proc Natl Acad Sci U S A*. 106:4629-34.

MITCHELL, R.J., LEE, S.K., KIM, T., GHIM, C.M. (2011). Microbial linguistics: perspectives and applications of microbial cell-to-cell communication. *BMB Rep* 44:1-10.

Malliere P (2011). Production of alkenes by decarboxylation of 3-hydroxyalkanoates. Patent number: US 2011/0165644 A1.

NENE, V., WORTMAN, J.R., LAWSON, D., *et al* (2007). Genome Sequence of Aedes aegypti, a Major Arbovirus Vector. *Science* 316:1718-23.

NEUMANN, H., NEUMANN-STAUBITZ, P. (2010). Synthetic biology approaches in drug discovery and pharmaceutical biotechnology. *Appl Microbiol Biotechnol*. 87:75-86.

NIELSEN, J., FUSSENEGGER, M., KEASLING, J., et al. (2014). Engineering synergy in biotechnology. *Nat Chem Biol* 5:319-22.

PASOTTI, L., ZUCCA, S., LUPOTTO, M., *et al* (2011). Characterization of a synthetic bacterial self-destruction device for programmed cell death and for recombinant protein release.*Journal of Biological Engineering* 5:8.

PINHEIRO, V.B., TAYLOR, A.I., COZENS, C., *et al* (2012). Synthetic genetic polymers capable of heredity and evolution. *Science* 336:341-4.

PURNICK, E.M., WEISS, R. (2009). The second wave of synthetic biology: from modules to systems. *Nature Reviews Molecular Cell Biology* 10:410-422.

RO, D.K., PARADISE, E.M., OUELLET, M., et al. (2006) Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* 440:940-943.

RUDE, M.A., BARON, T.S., BRUBAKER, S., et al. (2011). Terminal olefin (1-alkene) biosynthesis by a novel P450 fatty acid decarboxylase from *Jeotgalicoccus sp. Appl Environ Microbiol* 77:1718-27.

RUDER, W.C., LU, T., COLLINS, J.J. Synthetic biology moving into the clinic (2011) *Science* 333:1248-52.

SALIS, H., MIRSKY, E., VOIGT, C. (2009). Automated design of synthetic ribosome binding sites to control protein expression *Nature Biotechnology* 10:946-950.

SAVILE, C.K., JANEY, J.M., MUNDORFF, E.C., et al. (2010) Biocatalytic asymmetric synthesis of chiral amines from ketones applied to sitagliptin manufacture. Science 329:305-9.

STEEN, E.J., KANG, Y., BOKINSKY, G., et al. (2010). Microbial production of fatty-acid-derived fuels and chemicals from plant biomass. *Nature* 463:559-62.

STRICKER, J., COOKSON, S., BENNETT, M.R., et al. (2008). A fast, robust and tunable synthetic gene oscillator. *Nature* 456:516-9.

WANG, H.H., ISAACS, F.J., CARR, P.A., SUN, Z.Z., XU, G., FOREST, C.R., CHURCH, G.M. (2009). Programming cells by multiplex genome engineering and accelerated evolution. *Nature* 460: 894-898.

WEBER, W., FUSSENEGGER, M. (2006). Pharmacologic transgene control systems for gene therapy. *J Gene Med* 8:535-56.

WEISS, R., KNIGHT, T.F. (2000). Engineered communications for microbial robotics. *DNA6: Sixth International Meeting on DNA Based Computers*, pp 1–55.

WINDBICHLER, N., MENICHELLI, M., PAPATHANOS, P.A., et al. (2011) A synthetic homing endonuclease-based gene drive system in the human malaria mosquito. *Nature* 473:212-5.

WINKLER, W., NAHVI, A. BREAKER, R.R. (2002). Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature* 419:952-6.

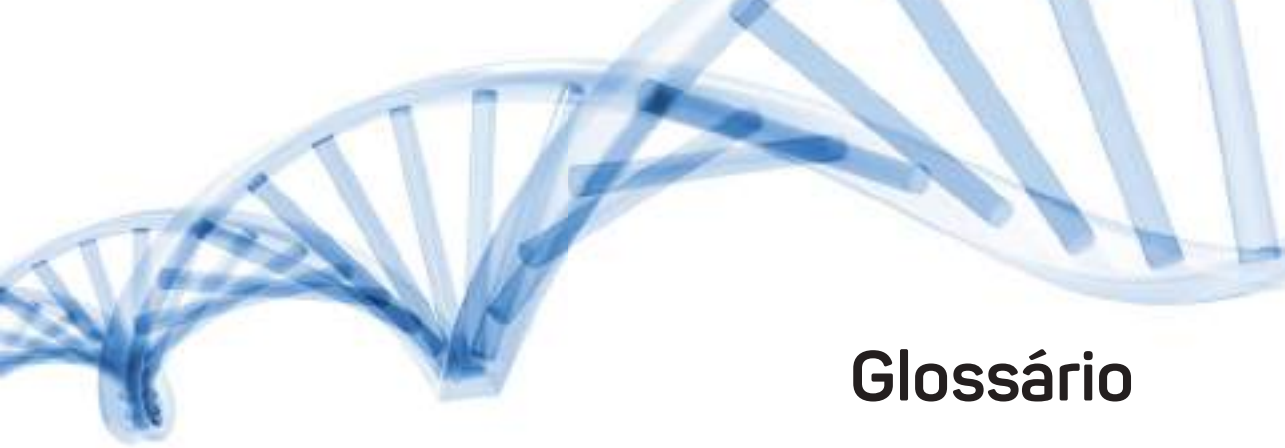
XAVIER, J.B. (2011). Social interaction in synthetic and natural microbial communities. *Mol Syst Biol* 7:483.

XIANG, S., FRUEHAUF, J., LI, C.J. (2006). *Nat. Biotechnol* 24:697.

Xiang S, Fruehauf J, Li CJ (2006). Short hairpin RNA-expressing bacteria elicit RNA interference in mammals. *Nat Biotechnol* 24:69.

YIM, H., HASELBECK, R., NIU, W., et al. (2011). Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol. *Nat Chem Biol* 7:445-52.

YOOSEPH, S., SUTTON, G., RUSCH, D.B., *et al* (2007). The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol* 5:e16.



Glossário

Apomorfia – Qualquer característica evolutivamente mais recente, derivada de outra característica em espécie ancestral. Sinapomorfia, plesiomorfia e autapomorfia são termos correlacionados (ver adiante).

Lineu, Carlos ou Carolus Linnaeus – Cientista sueco (1707 – 1778), um dos criadores da nomenclatura binomial e classificação dos seres vivos, sendo considerado por muitos como um dos criados da taxonomia moderna.

Frederick Sanger – Cientista britânico agraciado com dois prêmios Nobel: um pelo primeiro seqüenciamento completo de uma proteína (insulina) em 1958 e outro em colaboração com Walter Gilbert e Paul Berg em 1980, pela determinação da seqüência de nucleotídeos em uma amostra de DNA, o que levou a ser referenciado na técnica – método de seqüenciamento dideoxy ou simplesmente método de Sanger.

Clado – Representa um conjunto de organismos que supostamente sejam originados de um único ancestral em comum. Quando se elabora uma árvore filogenética, cada um dos ramos criados passa a ser considerado um clado.

Algoritmo: processo lógico (um caminho) para a resolução de um problema.

Analogia: similaridade que não resulta da homologia (ancestralidade comum) entre dois ou mais elementos.

Anotação funcional: Refere-se ao registro de informações (conhecidas ou preditas) relativas à função biológica de um determinado alvo de estudo, por exemplo, um gene ou proteína. (Veja predição funcional).

Árvore filogenética: diagrama representando a história evolutiva de organismos, genes, proteínas ou quaisquer outros elementos. Pode conter ou não uma raiz.

Biogeografia: estudo da distribuição geográfica dos organismos no espaço e tempo geológico.

Bootstrap**:** método estatístico de medida de consistência entre bases de dados no qual novas bases (simuladas) são geradas através de sorteios repetidos, com reposição, dos dados originais.

Cadeia de Markov: procedimento estocástico descontínuo que depende das probabilidades (ao invés das certezas) de mudança de estado e dos estados anteriores.

Clado: representação gráfica de um grupo monofilético.

Cladogênese: processo de separação de espécies de organismos (evento de especiação) ao longo do tempo evolutivo.

Cladograma: representação gráfica que mostra somente a topologia da árvore.

Comprimento de ramo (inglês, branch length**):** comprimento de cada ramo em uma árvore evolutiva de acordo com uma escala que representa certa unidade de medida.

Congruência: medida do quão similar são distintas árvores evolutivas (e.g. táxons, genes, etc.).

Convergência evolutiva: processo pelo qual características (morfológicas, moleculares, etc.) se tornam similares sem que haja ancestralidade comum.

Conversão gênica: processo meiótico no qual ocorre troca não recíproca de informação genética como resultado da formação de um heteroduplex entre cromátides não irmãs.

Datação molecular: estimativa da época de ocorrência de um evento evolutivo baseada em dados moleculares.

Deleção: evento no qual um ou mais nucleotídeos ou resíduos de aminoácidos são removidos de uma cadeia polinucleotídica ou polipeptídica, respectivamente.

Dendrograma (árvore ultramérica): representação gráfica na qual todos os ramos tem o mesmo comprimento.

Duplicação gênica: processo evolutivo de duplicação parcial ou total de um gene originando duas cópias.

Embaralhamento de éxons (inglês, exon shuffling**):** evento(s) de recombinação que combinam ("embaralham") éxons de distintos genes.

Especiação: processo pelo qual se originam uma ou mais espécies diferentes a partir de uma espécie ancestral.

Evolução convergente (convergência evolutiva): processo evolutivo pelo qual características (morfológicas, moleculares, etc.) se tornam similares sem que haja ancestralidade comum.

Evolução divergente (divergência evolutiva): processo evolutivo pelo qual características (morfológicas, moleculares, etc.) acumulam diferenças a partir de um ancestral comum.

Evolução em concerto (inglês, concerted evolution**):** processo evolutivo no qual sequências repetidas em série tendem a se manter homogêneas devido à conversão gênica, *crossing over* desigual, etc.

Família gênica: conjunto de genes que exibem sequências e/ou funções similares que se originaram a partir de evento(s) duplicação gênica.

Filogenética: estudo das relações evolutivas entre diferentes organismos (vivos ou extintos), macromoléculas (e.g. DNA, RNA, proteínas), etc.

Filogenia: história evolutiva de organismos ou genes geralmente representada por uma árvore filogenética.

Filogenômica: interseção entre filogenética e genômica visando promover a predição funcional de genomas genes e seus produtos.

Filogeografia: estudo que relaciona os padrões evolutivos (e.g. genealogia de genes) dos organismos à sua distribuição geográfica.

Filograma (árvore aditiva): representação gráfica na qual o comprimento dos ramos representa o número de mudanças ocorridas entre os nós de uma árvore filogenética.

Grupo monofilético: grupo constituído por um elemento ancestral e todos os seus descendentes. "um grupo de organismos (e.g. espécies, gêneros) ou genes que inclui todos os descendentes de um ancestral comum. Veja "grupo monofilético".

Grupo parafilético: grupo constituído por descendentes de uma única espécie ou gene ancestral, no qual um ou mais descendentes foram excluídos.

Grupo polifilético: grupo constituído por descendentes de mais de uma espécie ancestral.

Heurístico: procedimento algorítmico simplificado e eficiente que não garante a solução exata de um determinado problema.

Homologia: relação de ancestralidade entre dois ou mais elementos (organismos, sequências de DNA, etc.)

In silico**:** analogia dos termos em latim (e.g. *in vivo*, *in situ*, *in vitro*) comumente usados em Biologia para designar os procedimentos realizados com auxílio de computadores.

Indel**:** evento que envolve uma inserção ou deleção de nucleotídeos ou resíduos moleculares identificados pela comparação de sequências moleculares de DNA/RNA ou proteínas, respectivamente.

Inserção: evento no qual um ou mais nucleotídeos ou resíduos de aminoácidos são inseridos (adicionados) a uma cadeia polinucleotídica ou polipeptídica, respectivamente.

Matriz de distância: tabela contendo a distância entre um conjunto de elementos (e.g. sequências moleculares) comparados par-a-par.

Mutação: alteração do material genético herdada pelos descendentes.

Ortologia: relação de homologia a partir de um evento de especiação.

Paralogia: relação de homologia originada a partir de um evento de duplicação gênica.

Predição funcional: Capacidade de prever com precisão a função de um determinado alvo de estudo. Por exemplo, a predição funcional de um gene baseada em sua sequência de nucleotídeos é um procedimento fundamental na genômica e outras áreas.

Politomia: divisão de um nó em vários ramos de uma árvore filogenética.

Pseudogenes: sequências nucleotídicas não funcionais que possuem similaridade com sequências funcionais presentes no genoma.

Ramo: porção de uma árvore evolutiva conectando dois nós.

Sequências homólogas: sequências moleculares (DNA, RNA, proteínas, etc.) que compartilham um ancestral comum independente do grau de proximidade/similaridade entre elas.

Similaridade de sequência: grau de proximidade entre duas ou mais sequências moleculares geralmente expressas em porcentagem (%).

Táxon: conjunto de organismos pertencentes a um dado nível taxonômico (e.g. espécie, gênero, etc.).

Valores de bootstrap**:** representam a porcentagem de vezes que certo clado está presente nas diferentes bases de dados simuladas na análise de *bootstrap*.

Xenologia: relação de homologia originada a partir da transferência horizontal de material genético entre organismos distintos.

Xenólogos: genes ou produtos gênicos que divergiram entre si após um evento de transferência lateral de genes entre organismos distintos.

Apêndice 1

Script CGI Perl que produz um formulário HTML capaz de receber um arquivo contendo uma lista de identificadores de sequências, copiar esse arquivo localmente, buscar no GenBank cada uma das sequências e imprimir seus dados no formato FASTA.

```
#!/usr/bin/perl
use CGI;
use Bio::Perl;
use strict;
my $meuCGI = new CGI;
print $meuCGI->header(-charset=>"UTF-8");
print $meuCGI->start_html(-title=>"Baixando Sequências do Genbank");
print "<h2>Baixando Sequências do Genbank</h2>\n";
print $meuCGI->start_multipart_form;
print "Arquivo com lista de sequências do Genbank: ", $meuCGI->filefield(-name => 'nomeDoArquivo');
print "<br>", $meuCGI->submit(-value => 'Submeter Arquivo');
if (defined($meuCGI->param('nomeDoArquivo'))){
  my $arquivo = $meuCGI->param('nomeDoArquivo');
  my $arquivoDeSaida = "baixados/$arquivo";
  my $buffer;
  my $bytesLidos = 0;
  open (ARQUIVO_DE_SAIDA, "> $arquivoDeSaida") or die "<h3> Não foi possivel abrir salvar o arquivo
'$arquivoDeSaida'</h3>",$meuCGI->end_html;
  while (read($arquivo, $buffer, 1024)) {
    print ARQUIVO_DE_SAIDA $buffer;
    $bytesLidos+=1024;
  }
  close ARQUIVO_DE_SAIDA;
  $bytesLidos = $bytesLidos/1024;
  print "<br>Arquivo '$arquivo' salvo com sucesso ($bytesLidos Kb).\n";
  carregarSequenciasGenbank($arquivoDeSaida);
}
print $meuCGI->end_form;
print $meuCGI->end_html;
exit 0;
sub carregarSequenciasGenbank{
  my $arquivoDeEntrada = shift;
  my $database = "genbank";
  my $format="fasta";
  my $sequence;
  open(IN, "< $arquivoDeEntrada") or die "<h3>Nao foi possivel abrir arquivo de entrada:
'$arquivoDeEntrada'</h3>",$meuCGI->end_html;
  print "<h3>Recuperando Sequências:</h3>";
  print "<pre>";
  while (<IN>){
    chomp;
    $sequence = get_sequence($database, $_);
    write_sequence(">-", $format, $sequence);
    print "\n";
  }
  print "</pre>";
  close IN;
}
```

Apêndice 2

Script CGI Perl que produz um formulário HTML que permite ao usuário escolher qual termo procurar no banco de dados.

```
#!/usr/bin/perl
use CGI;
use DBI;
my $meuCGI = new CGI;
print $meuCGI->header(-charset=>"UTF-8");
print $meuCGI->start_html(-title=>"Consulta ao Banco de Dados");
print $meuCGI->start_multipart_form;
print "Digite parte do gênero buscado: ", $meuCGI->textfield(-name => 'genero', -size => 20, -maxlength => 40);
print $meuCGI->submit(-name => 'botao', -value  => 'Consultar');
if (defined($meuCGI->param('botao'))){
  my $textoConsulta = $meuCGI->param('genero');
  print "<br><hr><br>Resultados da consulta pelo termo: '<b>$textoConsulta</b>'.<br>";
  my $DATABASE = 'livro';
  my $TABLE = 'ResultadosBlast';
  my $CONFIG  = ";mysql_read_default_file=/home/luciano/.my.cnf";
  my $dbh = DBI->connect("DBI:mysql:livro" . $CONFIG, , ) or erro("Não foi possível conectar no bando de
dados '$livro'\n");
  my $SQL = "select QueryAccession, Accession, Genus, MaxScore from ResultadosBlast where upper(Genus)
like upper(\"%$textoConsulta%\");";
  my $sth = $dbh->prepare($SQL);
  if ($sth->execute){
    print "<table border=1><tr bgcolor='#AAAAAA'><td>Query
Acession</td><td>Accession</td><td>Genus</td><td>Score</td></tr>";
    while (my @dados = $sth->fetchrow_array){
      print "<tr><td>$dados[0]</td><td>$dados[1]</td><td><b>$dados[2]</b></td><td>$dados[3]</td></tr>";
    }
    print "</table>";
  }else{
    erro("Não foi possível realizar a consulta.");
  }
}
print $meuCGI->end_form;
print $meuCGI->end_html;
exit 0;
sub erro(){
  my $mensagem = shift;
  print "<h2>ERRO: $mensagem</h2>";
  print $meuCGI->end_form;
  print $meuCGI->end_html;
  exit 0;
}
```

Apêndice 3

Cópia de tela do formulário HTML produzida pelo Script Perl apresentado, em resposta a consulta pela palavra.

Digite parte do gênero buscado: bacillus Consultar

Resultados da consulta pelo termo: '**bacillus**'.

Query Acession	Accession	Genus	Score
CAI64586.1	ZP_08004173.1	**Bacillus**	116
CAI64586.1	ZP_03146815.1	**Geobacillus**	113
CAI64586.1	YP_001124741.1	**Geobacillus**	113
CAI64586.1	ZP_07707498.1	**Bacillus**	112
CAI64586.1	P21543.1	**Paenibacillus**	110
CAI64586.1	YP_146560.1	**Geobacillus**	108
CAI64586.1	YP_003252634.1	**Geobacillus**	108
CAI64586.1	YP_003672373.1	**Geobacillus**	108
CAI64586.1	ADG45817.1	**Geobacillus**	107
CAI64586.1	ACK58047.1	**Geobacillus**	107
CAI64586.1	YP_003949399.1	**Paenibacillus**	107
CAI64586.1	ABL77406.1	**Geobacillus**	106
CAI64586.1	AEQ38578.1	**Anoxybacillus**	103
CAI64586.1	ZP_03225714.1	**Bacillus**	102
CAI64586.1	YP_002316532.1	**Anoxybacillus**	102
CAI64586.1	ZP_01861961.1	**Bacillus**	100

A formação de recursos humanos de qualidade em pesquisa científica requer uma robusta formação teórica. Entretanto o que temos visto ao longo da última década é uma academia preocupada em publicar cada vez mais artigos científicos em revistas especializadas, não necessariamente com qualidade, em detrimento da popularização da ciência e formação de recursos humanos. Neste contexto, alunos de distintos programas de pós-graduação estão se tornando tecnicistas pouco críticos e desarticulados com a evolução do pensamento científico em Ciências Genômicas. Na tentativa de minimizar estas condições, a elaboração deste livro passou por uma criteriosa análise de demanda por conhecimento, estabelecido junto aos alunos de pós-graduação com os quais tive a oportunidade de trabalhar. Isto propiciou uma organização de capítulos que permite ao leitor compreender um pouco da história da genômica no Brasil e no mundo, e os decorrentes avanços oriundos de uma área do conhecimento que, em curto espaço de tempo, torna-se obsoleta pelas novas e fascinantes descobertas associadas. Com a participação de pesquisadores renomados, cada capítulo propicia ao leitor a oportunidade de se envolver a uma fundamentação teórica, ao mesmo tempo em que pincela aprofundamentos em conhecimentos específicos o que torna a obra diferenciada e, possivelmente, uma das principais fontes de consulta nesta área do conhecimento. Assim, esperamos que este livro possa se tornar uma bibliografia obrigatório a professores, pesquisadores e alunos dos mais distintos cursos de graduação e programas de pós-graduação em que as Ciências Genômicas possam se inserir.

Leandro Marcio Moreira
Professor e Pesquisador da Universidade Federal de Ouro Preto,
membro permanente dos programas de pós-graduação em Biotecnologia
(PPGBiotec) e do Mestrado Profissional em Ensino de Ciências (MPEC)

APOIO

CAPES

PROJETO BIGA
Bioinformática, Genômica e Associados
Processo 3385/2013, edital 051/2013

ISBN 9788589265225

9 788589 265225