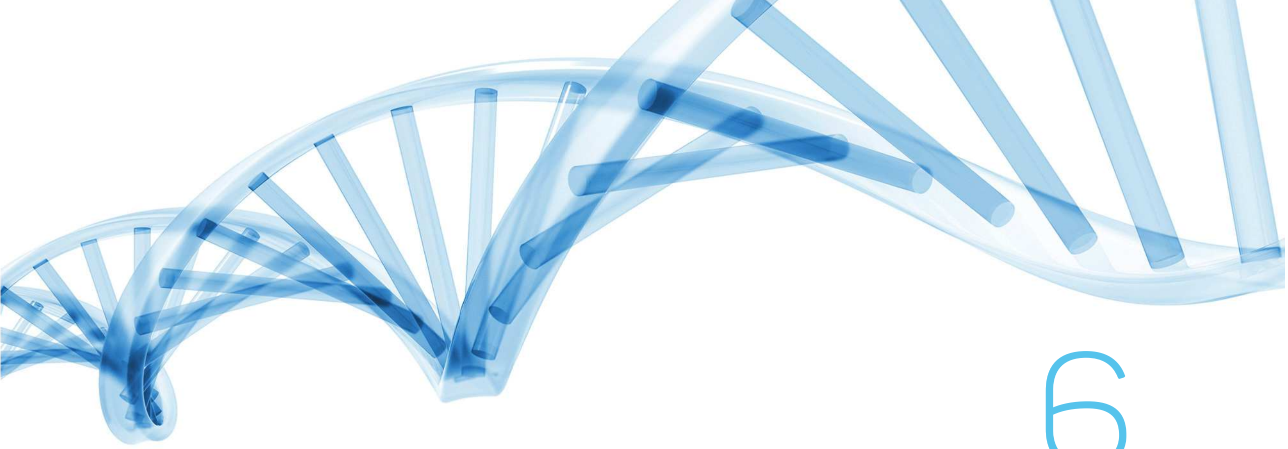

6

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).

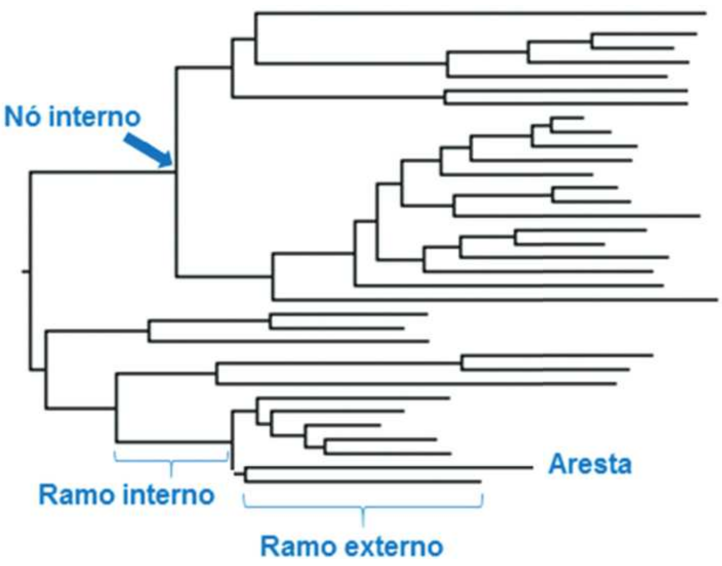

Figura 1. Componentes da árvore evolutiva. Uma árvore evolutiva hipotética mostrando ramos externos e internos, nós internos e arestas (*tips*). As arestas da árvore correspondem aos táxons (OTU) ou macromoléculas (genes, proteínas, introns, etc.).

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,

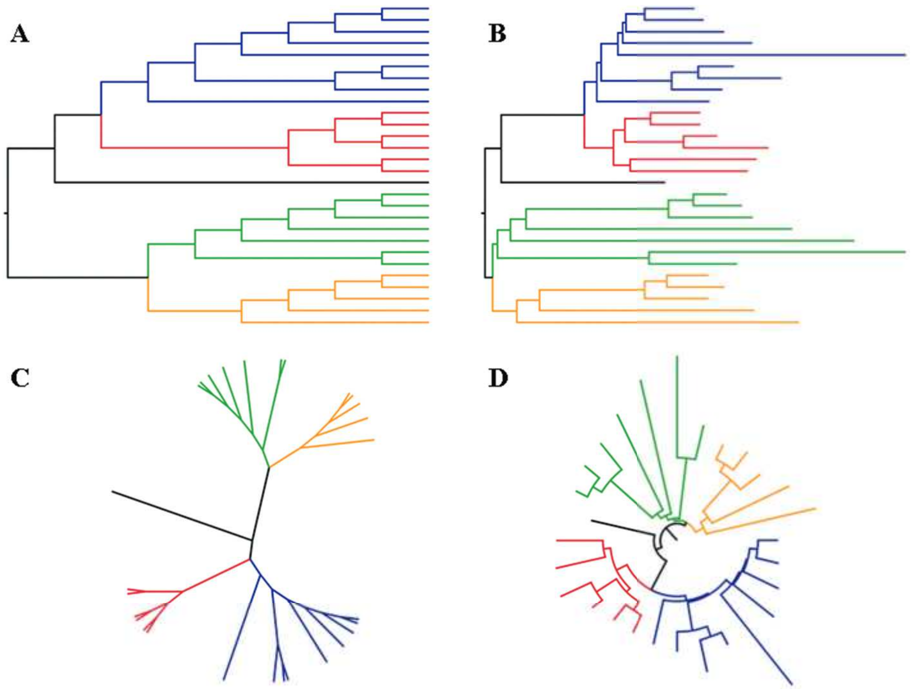

Figura 2. Representações da árvore evolutiva. Cladograma (A), filograma (B, C e D) dispostos nas formas retangular (A e B), radial (C) e polar (D). Imagens geradas a partir de uma árvore hipotética usando o programa FigTree.

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* del*etions – 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 no item 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.

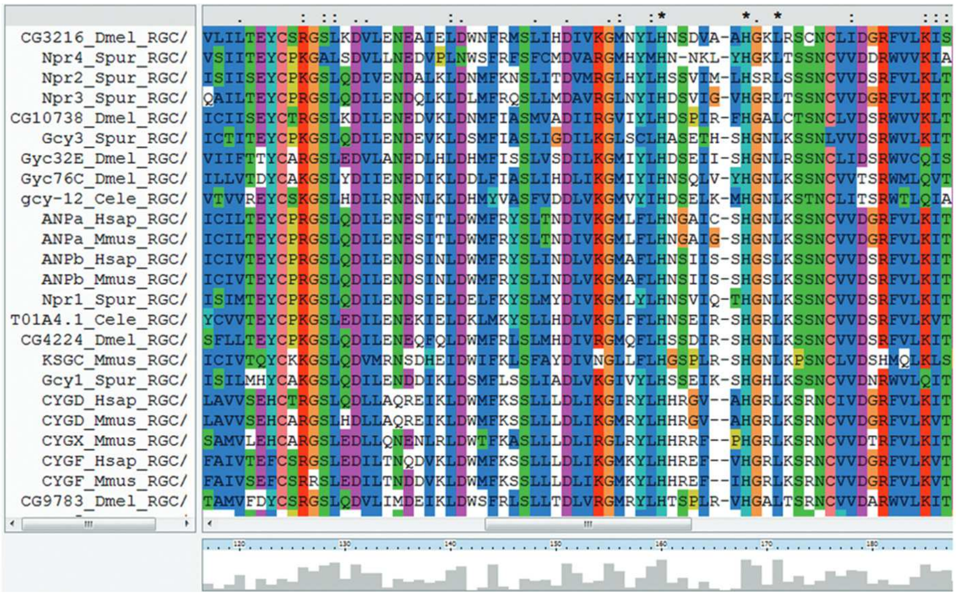

Figura 3. Alinhamento de múltiplas sequências de aminoácidos (proteína) gerado com o programa ClustalX. Os *gaps* (-) no alinhamento correspondem a *indels* nas sequências moleculares.

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.