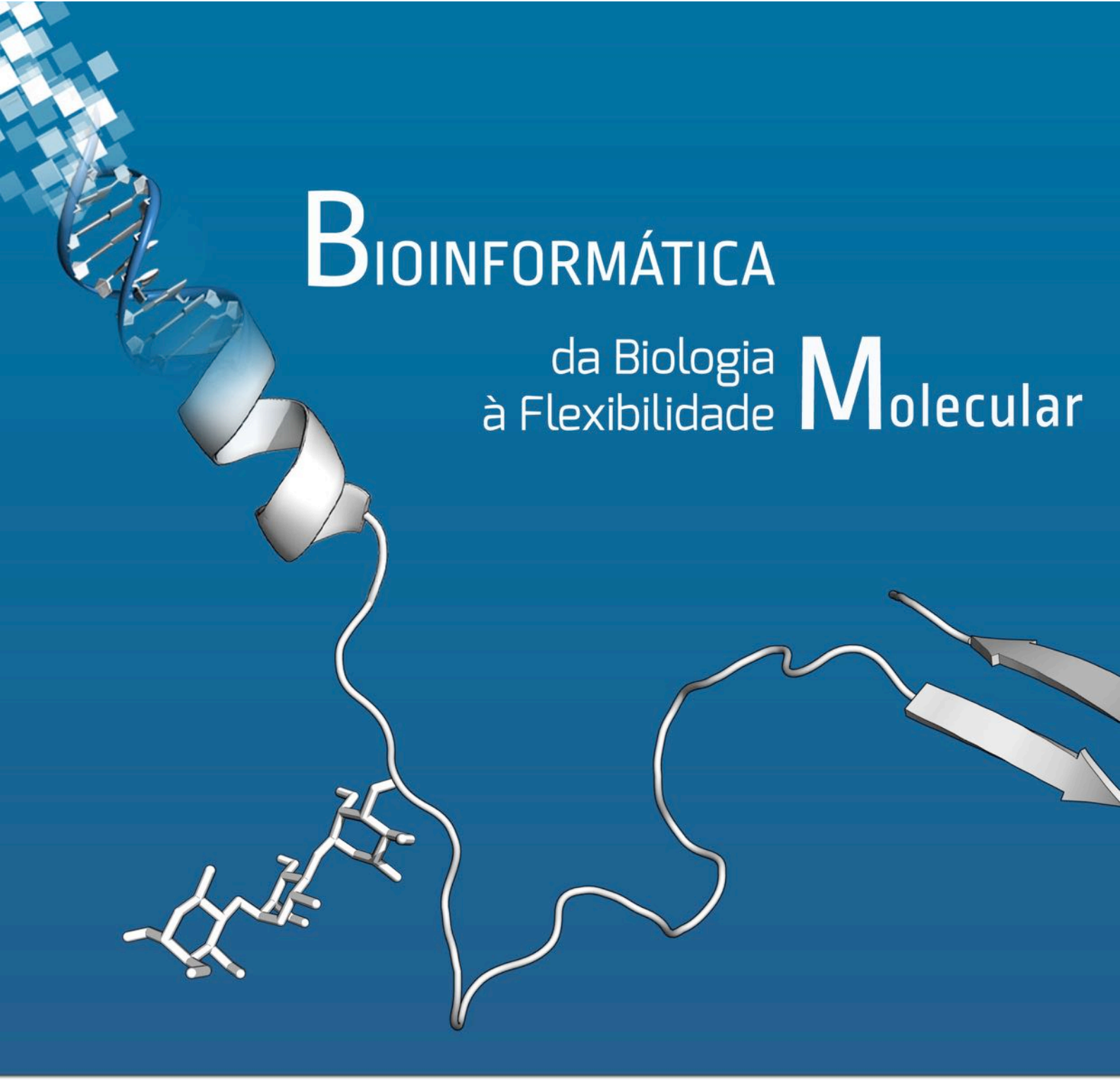


Bioinformática
da Biologia à Flexibilidade Molecular

Hugo Verli (Org.)

1ª edição
São Paulo, 2014

ISBN 978-85-69288-00-8
9 788569 288008

Sociedade Brasileira de Bioquímica e Biologia Molecular — SBBq

Apoio:
CNPq
CAPES
FAPERGS
UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Hugo Verli Organizador

Bioinformática: da Biologia à Flexibilidade Molecular

1ª Edição

São Paulo

Sociedade Brasileira de Bioquímica e Biologia Molecular - SBBq

2014

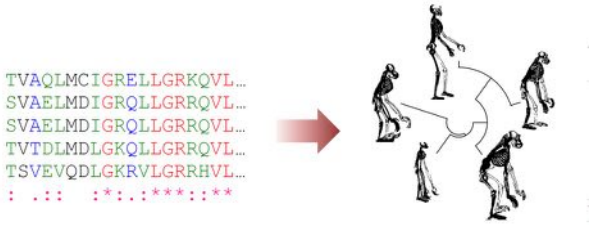
Ficha catalográfica elaborada por Rosalia Pomar Camargo CRB 856/10

B615 Bioinformática da Biologia à flexibilidade molecular / organização de Hugo Verli. - 1. ed. - São Paulo : SBBq, 2014. 282 p. : il.

1. Bioinformática 2. Biologia Molecular

CDU 575.112
ISBN 978-85-69288-00-8

5. Filogenia Molecular



Estabelecimento de relações evolutivas a partir de sequências de aminoácidos ou nucleotídeos.

Rodrigo Ligabue Braun
Dennis Maletich Junqueira
Hugo Verli

5.1. Introdução

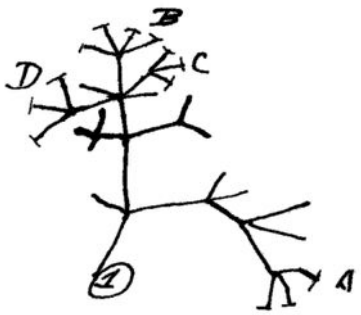
Desde seus primórdios, a humanidade se mostrou inclinada a organizar e classificar o mundo à sua volta com o objetivo de facilitar o entendimento e a comunicação. Em relação ao mundo natural, diferentes sistemas foram empregados para compor métodos de organização e classificar os organismos, utilizando critérios naturais ou artificiais.

Um dos sistemas de maior influência no período pré-Darwiniano foi a Escala Natural de Platão. Neste sistema, do fogo ao ser humano, diferentes níveis eram organizados à maneira de uma escada. A ideia de ascensão estava associada à perfeição, representada em sua forma plena pelo homem. O sistema classificatório de Lineu, por sua vez, se baseava em características visíveis, arbitrariamente selecionadas para classificar os seres vivos (por exemplo, número de patas ou de pétalas), sendo o ser humano o organismo do topo da cadeia. Sistemas como este são considerados sistemas artificiais, pois estão sujeitos à tendência de seu autor em considerar um caractere em detrimento de outro(s), conforme sua vontade ou necessidade. Entretanto, como o próprio Lineu reconheceu, tais sistemas foram absolutamente necessários para a fase inicial (descritiva) da biologia, servindo de base para o sistema natural de classificação e para as hipóteses de similaridade que surgiriam a seguir.

Ao final do século XVIII e início do século XIX, surgem os sistemas naturais de classificação. Estes buscavam refletir sobre a ordem natural dos seres vivos através de poucas características intrínsecas, geralmente associadas à forma. No entanto, com o objetivo de tornar a classificação mais racional, tomaram lugar debates sobre a real necessidade de haver um sistema hierárquico de organização dos organismos. Opositores da ideia consideravam que a classificação era, muitas vezes, inadequada e desnecessária, e que não deveria ser um fim em si mesma, senão um método para o levantamento de novas perguntas à Biologia.

Em 1818, a introdução do conceito de homologia por E.G. Saint-Hillaire causa uma revolução nas ciências biológicas. Para ele e seus colegas, partes homólogas correspondiam às partes de animais diferentes com uma estrutura essencialmente semelhante, mesmo com forma ou função distintas. Por

exemplo, as asas de um morcego, as nadadeiras de uma baleia e os braços de um macaco, segundo esta lógica, são considerados órgãos homólogos e podem servir como critério para agrupar morcegos, baleias e macacos em um mesmo grupo. Assim, a homologia serviria como critério principal para uma classificação natural dos organismos.



A primeira árvore filogenética moderna (esboço de Darwin no manuscrito de A Origem das Espécies)

A partir da famosa publicação de Darwin, "A Origem das Espécies", em 1859, a classificação dos organismos passou a ser não apenas natural, mas também a apresentar uma condição essencial de ancestralidade comum. Segundo este pensamento, os organismos são derivados uns dos outros, desde o surgimento da vida na terra. Darwin representou este padrão através de um esquema de ramificação, onde os galhos representam o tempo entre o organismo ancestral e o novo organismo, e os nós representam os próprios organismos. Mais tarde, esta viria a ser a primeira árvore filogenética utilizada para representar processos evolutivos.

Com influência direta da teoria evolutiva de Darwin (e colaborações de Wallace e Lamarck), desenvolve-se a Taxonomia Evolutiva. Este sistema de classificação incorporou o vetor tempo (caráter temporal normalmente inferido por meio de fósseis) e, além disto, adicionou uma quantificação da divergência estrutural entre os grupos (a chamada distância patrística). Já em meados do século XX, inicia-se a Fenética (taxonomia numérica ou neodansoniana). Esta escola buscava incluir na classificação dos organismos o máximo possível de características, atribuindo-lhes o mesmo peso na tentativa de eliminar qualquer subjetividade ou arbitrariedade. Seu impacto, entretanto, foi limitado devido às dificuldades em traduzir os índices (valores) obtidos em informações relevantes do ponto de vista biológico (como a separação de espécies, por exemplo). Na mesma época, surge a Cladística (ou sistemática filogenética), liderada pelo entomólogo alemão Willi Hennig. Na proposta de Hennig (1950), organismos que compartilhassem características derivadas (apomórficas) poderiam ser considerados descendentes do organismo ancestral, na qual a característica em seu estado primitivo (ou plesiomórfico) passou para o estado derivado.

Desde a origem dos sistemas de classificação até a Cladística, os métodos baseavam-se essencialmente no fenótipo dos organismos, ou seja, em suas características físicas claramente discerníveis. Entretanto, com o advento dos métodos de sequenciamento, tanto protéico quanto genômico, cada vez mais os dados moleculares foram se tornando importantes nas análises evolutivas de ancestralidade. Neste sentido, a ciência passa de um ponto de vista macroscópico a um ponto de vista molecular de análise.

O método de sequenciamento de aminoácidos, iniciado por Sanger em 1954, abriu caminho para que proteínas de uma mesma classe, em diferentes organismos, pudessem ser comparadas quanto às suas origens evolutivas. Da mesma forma, ao decodificar a primeira longa sequência de DNA, em 1977, Sanger deu início à explosão do sequenciamento de ácidos nucleicos, permitindo a comparação de genes em larga escala. É importante destacar que as sequências moleculares podem tanto ser comparadas entre si, buscando conhecer a história evolutiva de um gene ou proteína (por exemplo, relações entre hemoglobinas de diferentes mamíferos), quanto podem ser associadas a outros dados na reconstrução da história evolutiva de organismos (por exemplo, associando as relações obtidas por comparação de DNA ribossomal de aves com datação de fósseis, buscando estabelecer relações de ancestralidade).

No entanto, ao lidar com sequências moleculares, diferentes questões podem surgir. Por exemplo, o conceito de gene é di-

nâmico e mudou muito desde sua primeira definição. Além disso, genes podem sofrer diferentes processos evolutivos que alteram sua estrutura e/ou função, como mutações e rearranjos, ou ainda duplicações e perdas de função. Esses fatores fazem com que a relação 1:1 entre gene e organismo seja perdida. Por exemplo, uma mesma leguminosa pode possuir duas cópias do gene para a proteína leghemoglobina (genes parálogos). Além disso, muitas sequências do genoma não chegam à etapa de tradução, podendo conter elementos regulatórios ou transponíveis. Tais variações aumentam a complexidade e dificultam a interpretação das relações de descendência.

5.2. Aplicações

Ao classificarmos os organismos, atribuímo-lhes uma história evolutiva. Essa história, entretanto, é frequentemente desconhecida. Sendo assim, é necessário inferir a sequência de mudanças que levaram ao surgimento de um novo organismo ou proteína. Contudo, existe apenas uma história verdadeira, que talvez jamais seja conhecida. Assim, ao empregarmos as técnicas filogenéticas, o objetivo é coletar e analisar dados capazes de fornecer a melhor estimativa para chegarmos à filogenia verdadeira. De certa forma, a obtenção de filogenias lembra a atuação de um historiador. Baseando-se em dados disponíveis no presente (tais como organismos vivos, fósseis e sequências moleculares), tenta-se obter uma imagem de como teria sido o passado.

Quando analisamos sequências de nucleotídeos ou aminoácidos para inferir uma filogenia, utilizamos informações derivadas das taxas evolutivas para determinar a sequência de eventos que levaram ao surgimento de novos organismos. A taxa de evolução molecular refere-se à velocidade na qual os organismos acumulam diferenças genéticas ao longo do tempo. Essa taxa é frequentemente definida pelo número de substituições por sítio (ou posição no alinhamento de sequências) por unidade de tempo e, portanto, são usadas para descrever a dinâmica das mudanças em uma linhagem ao longo de várias gerações.

As taxas evolutivas são empregadas quando se buscam estimativas temporais para datação de eventos evolutivos. Normalmente, se assume que as mudanças nas sequências se acumulam a uma taxa mais ou menos constante ao longo do tempo. Esse conceito é chamado de Hipótese do Relógio Molecular. Entretanto, é conhecido que as taxas evolutivas são dependentes de vários fatores, tais como o tempo de geração, o tamanho da população e do próprio metabolismo, o que normalmente viola o modelo estrito de relógio molecular. Com base nestas informações, diversos modelos foram propostos para lidar com desvios no comportamento temporal de diferentes linhagens moleculares e, hoje em dia, são referidos como relógios moleculares relaxados.

Atualmente, a inferência filogenética é um campo de pesquisa à parte das outras ciências. Tornou-se uma ferramenta complementar para diversas áreas e indispensável para outras. Apesar de ter sido idealizada para desvendar apenas as relações evolutivas entre organismos, atualmente a filogenética molecular é aplicada a problemas muito mais diversos que este. Com o advento do relógio molecular estrito, foi possível aplicar a estimativa de tempo às filogenias e datar surgimento de espécies, disseminação de organismos e, até mesmo, entender grandes eventos biológicos que ocorreram no passado. Com a abordagem relaxada do relógio molecular, iniciou-se a utilização de modelos de dinâmica populacional que comportam os eventos coletivos de grupos específicos. Ainda, com o avanço da capacidade de processamento computacional, vem sendo possível criar algoritmos capazes de reconstruir genomas ancestrais. Também a partir da filogenética molecular desenvolveu-se o campo da filogeografia. Segundo esta área do conhecimento, as filogenias podem ser utilizadas para verificar a distribuição geográfica de indivíduos. Neste contexto, outras técnicas, além das filogenias, são incorporadas às aná-

lises, incluindo a estruturação de genes, as análises de redes e as análises de haplótipos.

A filogenia molecular busca inferir a história evolutiva de organismos ou outras entidades biológicas (como proteínas e genes) a partir de sequências de ácidos nucleicos ou aminoácidos. Ao investigar as relações entre diferentes espécies, análises de genes ribossomais são comumente empregadas, pois independentemente da espécie ou do organismo, os indivíduos possuirão genes codificantes de RNA ribossômico. Em contrapartida, quando se busca compreender as relações entre diferentes enzimas de uma mesma família é necessário utilizar sequências de aminoácidos, e não de nucleotídeos. Em determinadas situações, o genoma completo pode ainda ser utilizado para inferir a filogenia. Este é o caso de diversos vírus, especialmente quando se busca compreender a origem de novas variantes ou a disseminação de uma cepa. O alvo de estudo (isto é, sequência de nucleotídeos ou aminoácidos, gene ou genoma) depende, exclusivamente, do objetivo da análise e é um dos principais fatores a ser definido primariamente pelo pesquisador.

Atualmente, as filogenias funcionam como importantes ferramentas para diferentes áreas do conhecimento, incluindo as áreas de evolução, genética, epidemiologia, microbiologia, virologia, parasitologia, botânica e zoologia, dentre outras. Adicionalmente, de maneira inédita, a inferência filogenética foi utilizada como evidência para a resolução de crime e principal prova durante um impasse internacional envolvendo diferentes países. Em resumo, dependendo do objetivo, os métodos de construção de filogenias (inferência filogenética) são a base para diversas áreas e importantes objetos para o avanço computacional na análise de dados biológicos.

5.3. Representação de árvores

A Filogenética (termo obtido por união dos termos gregos para tribo e origem) é a ciência que busca reconstruir a história evolutiva dos organismos, levando em conta as sequências de nucleotídeos ou aminoácidos. As hipóteses sobre a história evolutiva são o resultado dos estudos filogenéticos e se chamam Filogenia.

As filogenias ou árvores filogenéticas representam o contexto evolutivo dos organismos de forma gráfica. São formadas por nós (pontos) ligados por diversos ramos (linhas) (Figura 1-5). Os nós terminais, mais externos na filogenia, identificam os indivíduos, genes ou proteínas que foram amostrados e incluídos na análise filogenética. Geralmente representam o alvo de estudo do pesquisador e estão ligados aos nós mais internos na filogenia através de traços horizontais, chamados de ramos terminais (Figura 1-5).

Os nós internos, pelo contrário, representam indivíduos não amostrados. Eles identificam uma inferência evolutiva do ancestral comum mais recente dos ramos derivados daquele nó e se ligam a nós cada vez mais internos, através dos ramos internos. Por exemplo, na Figura 1-5, os grupos de nós terminais representados em verde possuem como ancestral comum o nó laranja, mais interno, enquanto os nós terminais azuis possuem como ancestral comum o nó lilás. Da mesma forma, o nó vermelho é a representação do indivíduo, gene ou proteína mais ancestral da filogenia que, através de processos evolutivos, deu origem aos nós laranja e lilás.

O tamanho dos ramos horizontais pode ter diferentes significados, dependendo do método para inferência da filogenia, conforme

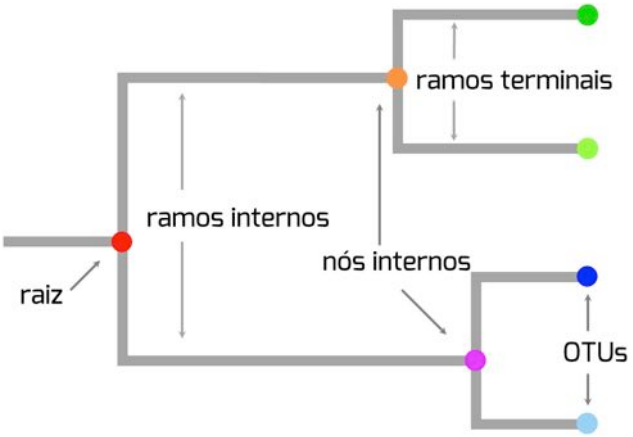


Figura 1-5: Nomenclatura associada a árvores filogenéticas.

veremos a seguir. No entanto, os ramos representados na vertical (Figura 1-5) não expressam qualquer significado, e seu tamanho não altera em nada a idéia filogenética. Como a análise pode ser feita em diferentes níveis, utilizando dados moleculares de genes, proteínas, indivíduos, espécies, gêneros, famílias, ou qualquer outro taxon, os nós terminais são amplamente denominados OTUs (*operational taxonomical units*), ou unidades taxonômicas operacionais (também chamados de folhas, Figura 2-5). A ordem e disposição exata das OTUs em uma filogenia é denominada topologia.

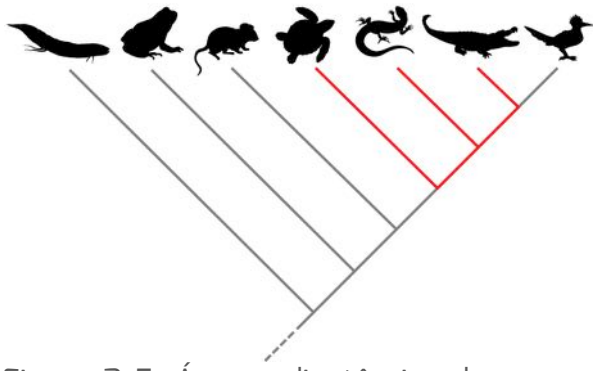


Figura 2-5: Árvore dicotômica dos grupos de vertebrados. As OTUs (nós terminais) estão representadas por ícones (peixes pulmonados, anfíbios, mamíferos, tartarugas, lagartos e serpentes, crocodilos e aves). Observe que o grupo dos répteis é parafilético (destacado em vermelho). O grupo seria considerado monofilético se incluísse as aves.

Além da forma gráfica, as árvores filogenéticas podem também ser descritas na forma textual. Em vez do diagrama com linhas e pontos, as relações evolutivas são representadas por notações com parênteses. A estrutura da árvore da Figura 2-5, por exemplo, pode ser descrita linearmente como (Peixes pulmonados, (Anfíbios, (Mamíferos, (Tartarugas, (Lagartos, (Crocodilos, Aves)))))) ou (Peixes pulmonados + (Anfíbios + (Mamíferos + (Tartarugas + (Lagartos + (Crocodilos + Aves)))))). Estas notações foram desenvolvidas para utilização computacional da informação filogenética. Algoritmos e programas que realizam análises moleculares necessitam da informação na forma textual e, quando necessário, fornecem a saída para o usuário na forma gráfica.

Partindo do princípio de derivação evolutiva, onde um organismo dá origem a outro (ou outros), podemos reconhecer dois principais processos na representação de filogenias: derivação dicotômica e derivação politômica. No primeiro caso, cada nó interno dá origem a apenas dois ramos. Para espécies, por exemplo, a ramificação de um ancestral comum em dois ramos evidencia o processo de especiação. No segundo caso, três ou mais ramos surgem de um mesmo nó interno.

Apesar de árvores dicotômicas serem mais comuns e normalmente esperadas, em alguns casos, como a dispersão explosiva do HIV e do HCV, árvores politômicas representam melhor o processo evolutivo. Casos como estes, onde um ancestral comum origina simultaneamente várias linhagens descendentes, são chamadas de politomias verdadeiras (*hard polytomies*). Por outro lado, as politomias falsas (*soft polytomies*) são casos onde a topologia não foi bem resolvida por não haver certeza do padrão de ancestralidade, tornando múltipla uma divisão que se esperaria ser formada por uma série de divisões dicotômicas.

Assim, ao agruparmos as OTUs segundo a sua ancestralidade, podemos reconhecer diferentes padrões: grupos monofiléticos, parafiléticos e polifiléticos (Figura 2-5). Os grupos monofiléticos incluem todos os membros descendentes de um único ancestral, assim como o próprio ancestral. Na Figura 2-5, por exemplo, as aves e os crocodilos são considerados um grupo monofilético, pois compartilham o mesmo ancestral comum. Da mesma forma, as aves, os crocodilos e os lagartos também podem ser considerados um grupo monofilético, pois se originaram de um mesmo ancestral. A análise das relações entre os grupos, neste caso, dependerá do objetivo do pesquisador. Adicionalmente, os grupos monofiléticos podem ser denominados clados por agruparem duas ou mais sequências que são descendentes de um mesmo ancestral (Figura 3-5a e b). A organização da topologia em que um clado está contido em outro é comumente chamada de clados aninhados ou clados embutidos (Figura 3-5c).

Os grupos parafiléticos, por sua vez, se

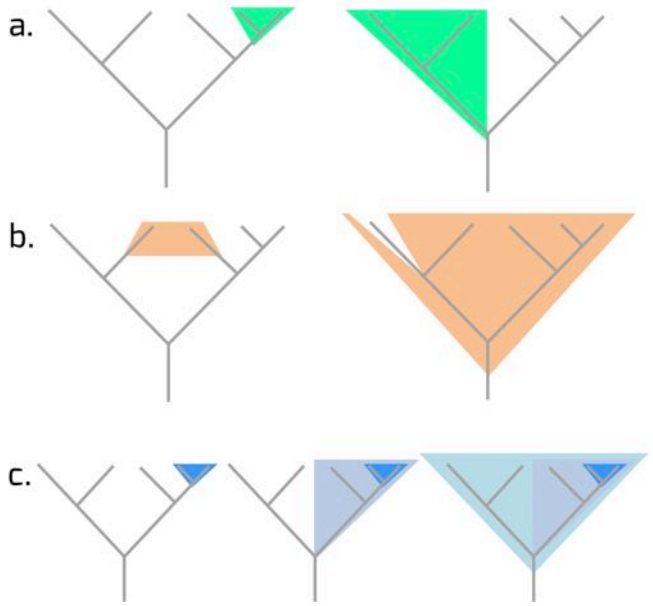


Figura 3-5: (a) Exemplos de clados destacados em verde. (b) Exemplos de organizações da topologia que não caracterizam a existência de um clado, destacados em laranja. (c) Diferentes níveis de clados que podem estar embutidos em um clado de maior ordem. Observe que os clados de diferentes ordens, quando embutidos, formam clados monofiléticos.

originam de um único ancestral, mas nem todos os organismos derivados deste ancestral fazem parte do grupo. Na Figura 2-5, os répteis são um grupo formado pelas tartarugas, lagartos e crocodilos, e seu ancestral comum está na base do ramo que dá origem às tartarugas. No entanto, este ancestral comum também deu origem às aves e, por isso, os répteis não podem ser considerados um grupo monofilético, mas um grupo parafilético.

Finalmente, os grupos polifiléticos provêm de dois ou mais ancestrais diferentes. Nestas relações se encontram OTUs que apresentam características comuns, mas que possuem diferentes ancestrais comuns. Por exemplo, a condição endotérmica (animais que mantém a sua temperatura corporal constante) é apenas apresentada por aves e mamíferos. Por este critério, poderíamos agrupar estes dois grandes grupos sem, no entanto, compartilharem o mesmo ancestral comum direto (Figura 2-5). A organização destes grupos permite descrever características resultantes de convergência evolutiva, pois uma mesma característica se desenvolveu independentemente em diferentes grupos.

Sabendo das relações evolutivas entre os táxons e da existência de ancestrais comuns, as árvores podem ser representadas de maneira a evidenciar o ancestral mais antigo (árvore com raiz ou enraizada), ou apenas destacar as relações evolutivas entre os táxons, sem destacar qual a OTU mais ancestral (árvore sem raiz ou não enraizada) (Figura 4-5).

A raiz da filogenia é a espécie ou sequência ancestral a todo o grupo que está sob análise. Quando presente, a raiz aplica uma direção temporal à árvore, permitindo observar o sentido das mudanças evolutivas da raiz (mais antigo) aos ramos terminais (mais modernos). Uma árvore não enraizada, pelo contrário, reflete apenas a topologia estabelecida entre as OTUs, sem indicar o ancestral do grupo. Árvores não enraizadas podem ser confusas, e sua interpretação requer mais cuidado devido à facilidade em cometer erros de análise (Figura 4-5).

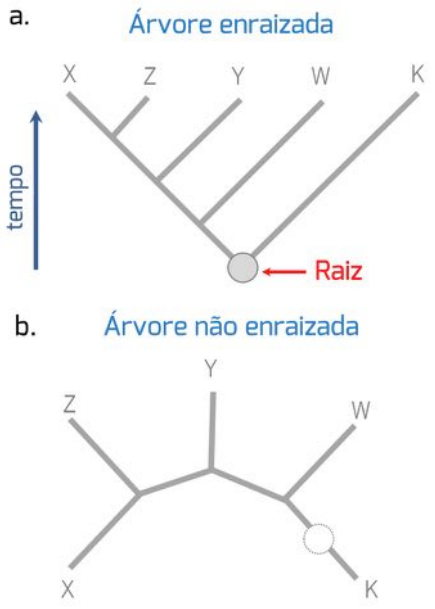


Figura 4-5: Comparação de árvores (a) enraizadas e (b) não enraizadas. No primeiro caso, é possível definir a direção das mudanças evolutivas, devido à presença do vetor tempo dado pela presença da raiz.

A identificação de uma raiz nas filogenias geralmente requer a inclusão de uma ou diversas OTUs que representem grupos externos. Os grupos externos devem ser ancestrais comuns das OTUs em estudo, já conhecidos, que indicarão caracteres presentes em organismos mais próximos aos ancestrais, provendo um direcionamento para a interpretação dos processos evolutivos. Para o caso do estudo de HIV, por exemplo, é comum que os vírus da imunodeficiência de símios (SIV) sejam utilizados como grupo externo nas filogenias, pois sabidamente estes vírus deram origem ao HIV.

A adição de grupos externos aumenta o número de topologias diferentes que uma filogenia pode assumir. O número de árvores possíveis varia com o número de OTUs e com a presença ou ausência de raiz. Para mais de duas OTUs, a quantidade de possíveis árvores com raiz é sempre maior que o número de árvores sem raiz. A possibilidade de inferência de diferentes topologias para os mesmos dados moleculares ressalta a extrema variabilidade de cenários possíveis na busca do verdadeiro evento evolutivo. É importante também ressaltar que, assim como a complexidade, o tempo computacional envolvido na construção das filogenias aumenta exponencialmente com o aumento de OTUs.

Em relação à topologia das árvores, a inversão de ramos derivados de um mesmo nó não altera a relação evolutiva apresentada pela árvore (Figura 5-5). Nesse sentido, a árvore filogenética pode ser comparada a um móbile: cada peça suspensa é livre para girar em seu eixo, ficando mais próxima ou mais distante espacialmente das outras peças, sem alterar a estrutura geral do objeto. Independentemente da posição destas OTUs, após o giro dos ramos, o mesmo ancestral comum será identificado e, por isso, não há qualquer alteração no significado da filogenia.

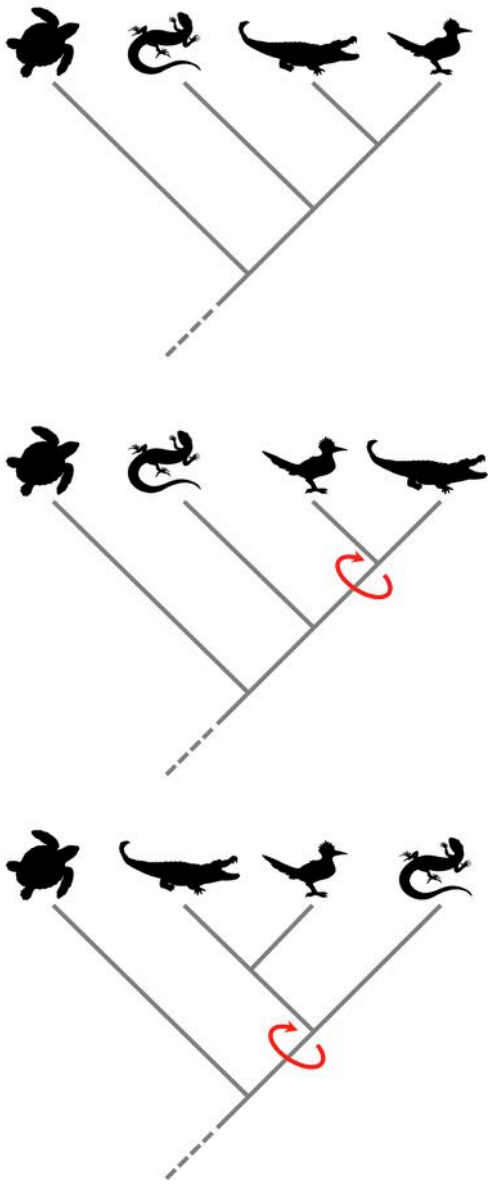


Figura 5-5: A porção terminal da árvore dos vertebrados (representada na Figura 2-5) foi rearranjada de diferentes maneiras (as setas indicam o ponto de rotação). Conforme a analogia de um móbile, todas elas representam a mesma relação evolutiva.

Quanto à nomenclatura de árvores filogenéticas, diferentes termos são empregados, tais como cladogramas, filogramas e dendrogramas (Figura 6-5). Um cladograma é uma árvore simples, que retrata as relações entre os nós terminais. Pelo contrário, uma árvore aditiva (árvore métrica ou filograma) apresenta informações adicionais, pois o comprimento dos ramos é proporcional a algum atributo, como quantidade de mudança. Por sua vez, uma árvore ultramétrica (ou dendrograma) constitui um tipo especial de filogenia devido aos seus ramos serem equidistantes da raiz. Os dendrogramas podem, desta forma, retratar o tempo evolutivo. É importante ressaltar que alguns autores denominam qualquer filogenia como cladograma, o que pode ser confuso.

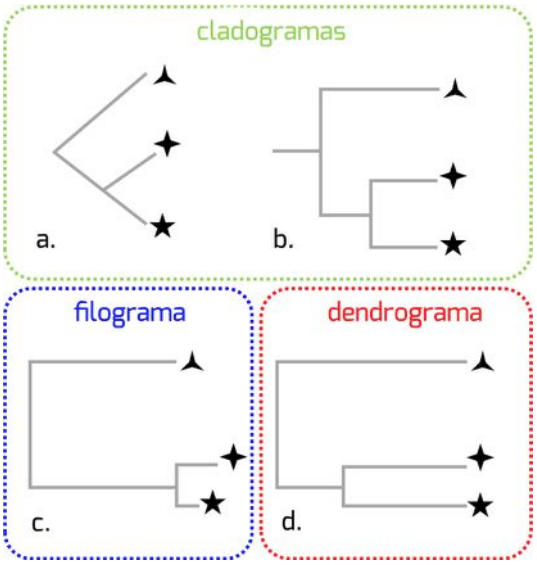


Figura 6-5: Nomenclatura de árvores filogenéticas. Observe que os cladogramas *a* e *b* são equivalentes, mas o filograma *c* e o dendrograma *d* não o são.

O tipo de dado molecular a ser empregado nas análises também deve ser levado em conta. Sequências de aminoácidos são mais conservadas que sequências de ácidos nucleotídeos em decorrência da degeneração do código genético. São, portanto, úteis em análises de produtos de genes ou espécies que visam entender fenômenos que aconteceram há amplos períodos de tempo evolutivo. Além disso, por formarem um conjunto de pelo menos 20 membros (contra quatro membros presentes em DNA ou RNA), sua variação pode ser mais significativa.

A despeito desta diferença no volume de informação, com a popularização do sequenciamento de ácidos nucleicos, especialmente DNA, sequências de nucleotídeos passaram a ser as mais empregadas em estudos de filogenia. Ácidos nucleicos são mais propensos a alterações, podendo sofrer transições (quando ocorre a troca de uma purina por outra purina, ou de uma pirimidina por outra pirimidina) e transversões (quando ocorre a troca de uma purina por uma pirimidina ou vice-versa), além de inserções ou deleções de pares de base que interferem no quadro de leitura. Essa variabilidade pode ser interessante no estudo de eventos mais recentes do ponto de vista evolutivo.

É preciso, assim, conhecer o caso de estudo e o tipo de pergunta que se busca responder com cada filogenia. Ao lidarmos com genes de diferentes espécies, por exemplo, é importante saber da existência e disposição de íntrons, da necessidade de lidar com o gene inteiro ou apenas parte dele ou da necessidade de incluir regiões regulatórias para a análise.

Um exemplo recente da aplicação de análises filogenéticas está no caso da identificação da origem da linhagem do vírus influenza H1N1, envolvido no surto de gripe de 2009. Para tanto, Smith e colaboradores empregaram genomas completos de influenza isolados de diferentes localidades e hospedeiros, e construíram árvores filogenéticas para cada uma das oito regiões do genoma buscando identificar a fonte de cada rearranjo presente no vírus envolvido no surto. Por meio das árvores obtidas, foi possível rastrear a contribuição genética dos vírus isolados de aves, suínos e humanos (Figura 7-5). Assim, o emprego da filogenia neste trabalho permitiu não apenas caracterizar o vírus do ponto de vista molecular, como também reconstruir a história evolutiva do agente etiológico de uma pandemia.

5.4. Distância genética

A formulação de modelos evolutivos é uma maneira de descrever matematicamente os processos que moldam as mudanças nas sequências de nucleotídeos ou aminoácidos dos organismos ao longo do tempo. Do ponto de vista molecular, estas mudanças podem ser resultado de diferentes forças evolutivas que reorganizam a sequência e a própria estrutura dos genes.

Um modelo geral para descrever de maneira eficaz estas alterações evolutivas deveria considerar os processos de substituição, inserção, deleção e duplicação, bem como ocorrência de transposição ou até mesmo de retrotransposição. Contudo, apesar de estes fenômenos serem claros agentes na modelagem dos genomas, matematicamente

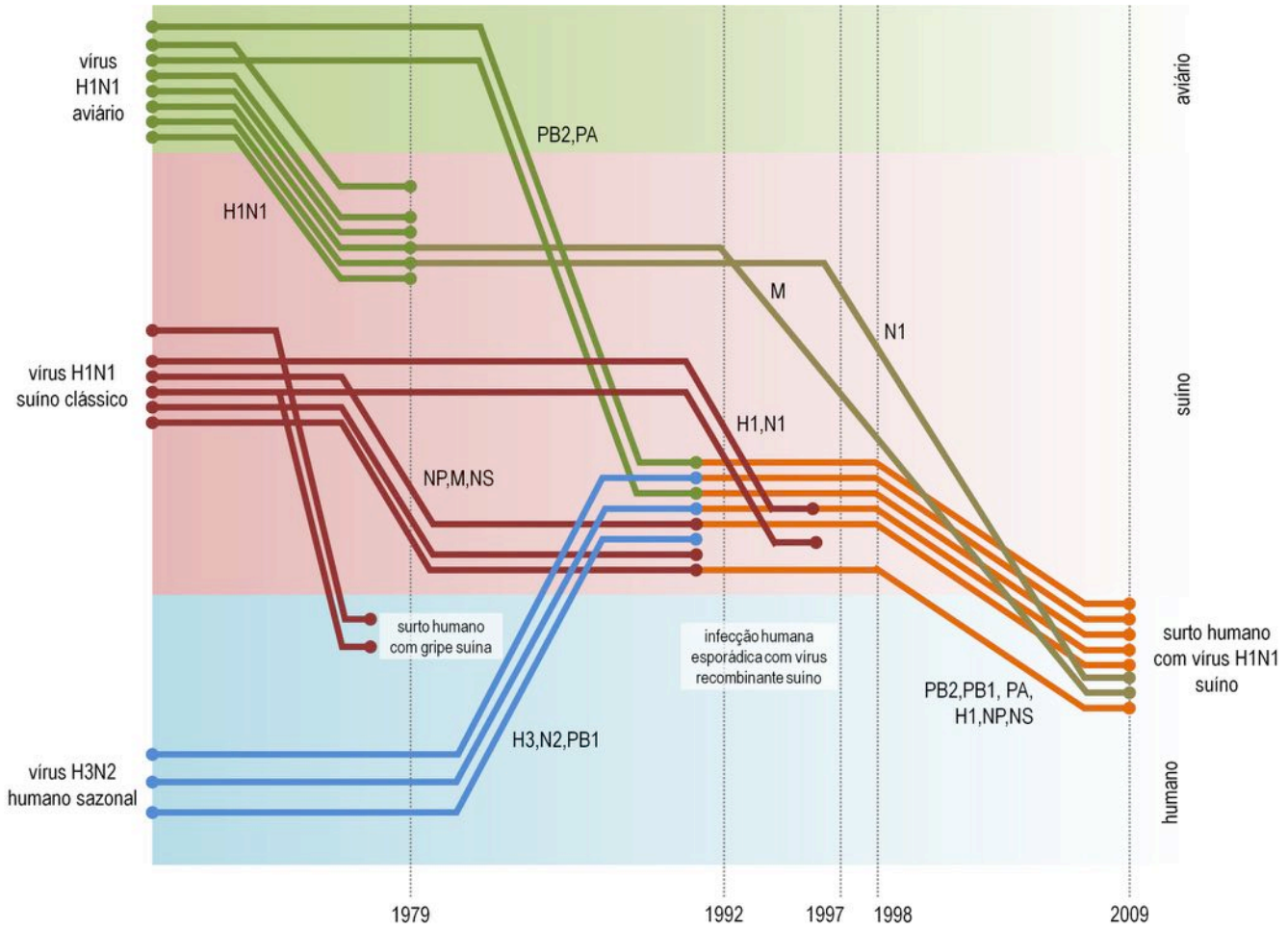


Figura 7-5: Representação esquemática das recombinações que originaram o vírus Influenza envolvido no surto de gripe suína em 2009. Diferentes linhas representam diferentes regiões do genoma do vírus. Observe a interação entre vírus de origens aviária, suína e humana em eventos que datam, pelo menos, desde 1990. Os eventos de recombinação e as análises temporais foram baseadas em análises filogenéticas (Adaptado de Smith e colaboradores, *Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic*. *Nature*, *459*, 1122-1125, **2009**).

ainda não é factível colocá-los como componentes de modelos que expliquem inteiramente o processo evolutivo.

Assim, devido à grande relevância dos mecanismos de substituição para a evolução dos genomas em diferentes organismos e da disponibilidade de modelos de probabilidade estatística que expliquem este processo, as trocas têm sido o principal alvo para o desenvolvimento de modelos matemáticos e compõem a base de diversos métodos de inferência filogenética.

Após a divergência de duas sequências a partir de seu ancestral comum, de forma dicotômica, fenômenos evolutivos garantirão as mudanças nas sequências de nucleotídeos de forma independente (Figura 8-5). Uma medida tradicional para expressar o número de substituições de nucleotídeos que se acumularam nas sequências desde a divergência é chamada de distância genética. Esta informação é uma medida quantitativa da dissimilaridade genética entre diferentes OTUs, e permite estabelecer uma estimativa relativa da quantidade de mudanças que ocorreram desde a divergência.

A distância é também um importante conceito na construção de filogenias, pois está diretamente relacionada com a relação evolutiva entre duas OTUs: uma menor distância

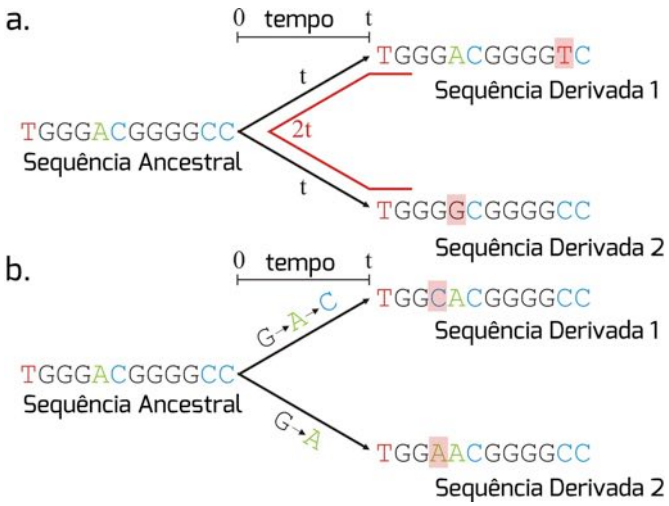


Figura 8-5: Após a divergência de dois organismos a partir de seu ancestral comum, seus genomas acumularão diferenças independentemente. (a) A medida da dissimilaridade genética entre duas sequências homólogas ao longo do tempo é chamada de distância genética, e a relação temporal entre duas sequências divergentes é dada por $2t$. (b) A ocorrência de múltiplas substituições ao longo do tempo na divergência de sequências homólogas pode mascarar as verdadeiras diferenças entre as sequências. Apesar de ocorrerem dois eventos de mutação na sequência derivada 1, apenas o último evento é observado, pois ocorreram no mesmo sítio. Os quadrados em vermelho evidenciam as diferenças em relação às sequências ancestrais.

genética indica uma relação evolutiva mais próxima, enquanto que um valor maior sugere uma derivação evolutiva proporcionalmente maior. Tipicamente, a informação da distância genética é incorporada à inferência filogenética na definição do tamanho dos ramos. No entanto, além desta informação é necessária uma escala de distância que especifique o número de mudanças que ocorreram ao longo do ramo.

O método mais simplista para avaliar a distância genética entre duas sequências é conhecido como distância p. Este método é baseado na contagem das diferenças dividida pelo número total de sítios do alinhamento. Se oito sítios são diferentes entre duas sequências homólogas com tamanho de 100pb, a distância p obtida será 0,08. Este resultado reflete a porcentagem de sítios diferentes em relação ao tamanho total da sequência, e geralmente é utilizado na especificação da escala de distância das filogenias (Figura 8-5).

A variação genética em um determinado sítio pode decorrer de diferentes processos e resultar em mais de uma substituição. As múltiplas substituições, ou *multiple hits*, ocorrem naturalmente e podem subestimar o verdadeiro número de mudanças no cálculo da distância p, já que "escondem" as diversas trocas de nucleotídeos ou aminoácidos. Na Figura 8-5b, por exemplo, apesar de ocorrerem duas substituições no mesmo sítio ao longo de um dos ramos, aparentemente a sequência derivada parece ter sofrido somente um evento evolutivo. Sendo assim, a relação entre as diferenças nas sequências e o tempo decorrido da divergência nem sempre é linear, especialmente devido à ocorrência das múltiplas substituições em um mesmo sítio.

Devido à ineficácia da distância p em efetivamente estimar a distância genética entre duas sequências, diferentes modelos probabilísticos foram desenvolvidos para descrever as mudanças entre os nucleotídeos e corrigir a distância observada. Tais modelos implicam no uso de diversas suposições simples a respeito das probabilidades de substituição de um nucleotídeo por outro, mas garantem uma aproximação da realidade quando sustentadas por uma taxa de mutação fidedigna.

Estas técnicas de correção são comumente conhecidas por modelos de substituição (ou matrizes de substituição), e garantem a conversão da distância observada em medidas de distâncias evolutivas próximas da realidade, permitindo reconstruir a história evolutiva dos organismos.

Diversos modelos de substituição foram propostos para explicar as trocas de nucleotídeos em sequências de DNA, reduzindo a complexidade do processo evolutivo a um padrão de mudança simples que consegue ser explicado através de poucos parâmetros. Todos estes modelos, no entanto, de alguma forma são interrelacionados, diferindo principalmente no número de

parâmetros utilizados para explicar estas substituições. Devido à influência do modelo de substituição na inferência de filogenias, a escolha de um método particular deve ser justificada. A estratégia mais simples é utilizar os modelos que comportam o maior número de variáveis, embora a complexidade não esteja diretamente relacionada à melhor qualidade de análise das sequências. Com o aumento de parâmetros, o sistema se torna mais complexo, aumentando a probabilidade de erro e exigindo um maior processamento computacional. Assim, é necessário verificar os alinhamentos caso-a-caso para atribuir o melhor modelo de substituição na inferência filogenética.

A substituição de nucleotídeos ou aminoácidos em uma sequência é usualmente modelada sob a forma de um processo quase aleatório. Devido ao caráter dinâmico desta aleatoriedade, é necessário enquadrar as substituições, seguindo certos pressupostos. Assim, as substituições são descritas por um processo de Markov homogêneo, onde a probabilidade de substituição de um nucleotídeo X pelo Y não depende do estado prévio do nucleotídeo X.

As probabilidades de mudança de um nucleotídeo para outro (ou de um aminoácido para outro) são especificadas através de uma matriz 4x4 das taxas de substituição (ou 20x20 no caso dos aminoácidos) que especificam com qual taxa cada um dos nucleotídeos ou aminoácidos poderá mudar para outro. É necessário assumir também que os eventos de substituição sejam independentes ao longo dos sítios das sequências, e ainda, possuam um caráter reversível. Além disso, devem especificar a frequência estacionária dos nucleotídeos, ou frequência de equilíbrio, onde será atribuída a provável proporção de cada um dos caracteres na sequência.

Para sequências de nucleotídeos, o modelo de substituição mais simples foi proposto por Jukes e Cantor em 1969 (JC69). Segundo este modelo, as mudanças entre os nucleotídeos podem ocorrer com a mesma probabilidade, assumindo uma frequência estacionária igual para todos (cada nucleotídeo tem 25% de chance de ocorrer na sequência).

Com o advento da publicação das primeiras sequências de genoma mitocondrial, na década de 1980, se observou que as transições eram muito mais comuns que as transversões. Devido à uniformidade do método proposto por Jukes e Cantor, foi necessário criar um modelo que acomodasse essas diferenças.

Assim, o modelo proposto por Kimura (K80 ou K2P) cria as variáveis α e β para representar, respectivamente, as taxas de transição e de transversão. Apesar da inclusão de dois parâmetros, as frequências de equilíbrio se mantêm constantes em ¼ para cada nucleotídeo. Em 1981, Kimura adiciona um terceiro parâmetro (γ) ao modelo já proposto, passando a ser identificado como K3P. A atualização do modelo permitiu dividir as taxas de transversão em duas variáveis.

Alguns genomas apresentam uma grande quantidade de guaninas e citosinas em relação a timinas e adeninas. Se algumas bases são mais frequentes que outras, será esperado que algumas substituições ocorram com mais frequência que outras. O modelo criado por Felseinstein (F81) acomoda essas observações e permite que as proporções individuais de cada nucleotídeo (frequência estacionária) sejam diferentes de ¼. É importante ressaltar que este modelo considerará a mesma proporção de bases em todas as sequências envolvidas no alinhamento. Se diferentes sequências possuem diferente composição de bases, a pressuposição principal do modelo será violada.

O modelo HKY85, proposto por Hasegawa, Kishino e Yano, essencialmente mistura os modelos K2P e F81. Além de supor que a frequência das bases é variável, este modelo permite que transições e transversões ocorram com taxas diferentes.

Posteriormente, o modelo GTR (*generalised time-reversible*), o mais complexo dos modelos aqui apresentados, foi desenvolvido a partir do HKY85 com o intuito de acomodar diferentes taxas de substituição e diferentes frequências de bases. Este modelo requer seis parâmetros para taxa de substituição e quatro parâmetros para a frequência das bases, misturando todos os modelos aqui descritos.

Atualmente, além destes mais de 200 modelos de substituição podem ser aplicados a alinhamentos de nucleotídeos. Alguns programas, como Modeltest e Jmodeltest, são capazes de selecionar o modelo de substituição que melhor se ajusta a um dado alinhamento.

Uma importante extensão desses modelos de substituição incorpora a possibilidade de variação nas taxas evolutivas entre os sítios, permitindo ao modelo mais realismo. Assim, para cada sítio no DNA será atribuída uma probabilidade de evolução a uma taxa contida em um intervalo discreto de probabilidades. O método que garante a heterogeneidade de taxas evolutivas é modelado através de uma distribuição gama (Γ), que considera um número específico de taxas de

evolução para os sítios do DNA.

A aplicabilidade deste modelo nas inferências filogenéticas é facilitada pela simplicidade do método, já que apenas um único parâmetro (α) controla a forma da distribuição gama. Quando $\alpha < 1$, existe um grande número de taxas de evolução entre os sítios das sequências em análise, ou seja, quanto maior α, menor a heterogeneidade. Algumas vezes, uma proporção de sítios invariáveis (I), no qual uma determinada proporção de sítios é assumida como incapaz de sofrer substituição, pode também ser usada para modelar a heterogeneidade entre os sítios.

Ao contrário dos modelos de substituição de nucleotídeos, os modelos que explicam as trocas de aminoácidos são tradicionalmente empíricos. A partir da análise de alinhamentos de proteínas com identidade mínima de 85% Dayhoff, em 1970, desenvolveu uma série de matrizes de probabilidade que explicavam as mudanças de aminoácidos ao longo do tempo.

As matrizes PAM, como ficaram conhecidas, correspondem a modelos de evolução nos quais os aminoácidos são substituídos aleatoriamente e independentemente, de acordo com uma probabilidade predefinida que depende do próprio aminoácido.

Em 1992, um novo modelo de substituição de aminoácidos é criado por Henikoff e Henikoff. A análise de sequências de proteínas distantes evolutivamente, possibilitada pelo modelo de Henikoff-Henikoff, estabeleceu as bases para a criação das matrizes BLOSUM. As matrizes desta série foram identificadas por números (por exemplo, BLOSUM62) que se referem à porcentagem mínima de identidade dos blocos dos aminoácidos utilizados para construir o alinhamento. Matrizes similares, como GONNET e JTT, surgiram na mesma época.

Em 1996, foi proposto um modelo de substituição específico para proteínas codificadas pelo DNA mitocondrial, onde foi observado desvio de transições entre aminoácidos em relação às proteínas codificadas pelo material genético nuclear. Essa matriz, criada por Adachi e Hasegawa, foi chamada de mtREV.

Finalmente, em 2001, Whelan e Goldman propõem a matriz WAG, baseada em combinação e ampliação de vários modelos de substituição anteriores. Tal matriz é considerada superior às suas antecessoras para descrever filogenias de proteínas globulares.

5.5. Inferência filogenética

A reconstrução filogenética, ou seja, a reconstrução da história evolutiva de organismos, é um complexo processo que envolve uma série de etapas. O alinhamento, além de ser o primeiro passo, é um importante ponto para a inferência de filogenias (ver capítulo 3). Um alinhamento preciso, além de garantir maior confiabilidade nas análises posteriores, é requerido por todos os métodos de inferência filogenética para construção da árvore.

Depois que o alinhamento foi proposto, diversos métodos podem ser usados para estimar a filogenia das sequências estudadas. Podemos dividir estes métodos em dois principais grupos: métodos quantitativos e métodos qualitativos (Tabela 1-5). Estes grupos diferem na forma como os dados são tratados, refletindo diretamente como os dados do alinhamento serão inicialmente processados.

Os métodos quantitativos se baseiam na quantidade de diferenças entre as sequências do alinhamento para calcular uma árvore final. Já os métodos qualitativos constroem diversas filogenias que são classificadas seguindo uma determinada qualidade (critério). A filogenia que obtiver o maior valor associado à tal qualidade será a filogenia resultante.

Os métodos quantitativos compreendem os métodos de distância. Estes métodos convertem o alinhamento em matrizes de distância par-a-par para todas as sequências incluídas. Dentro destes algoritmos destacam-se dois métodos principais: UPGMA e aproximação dos vizinhos. Devido à grande eficiência computacional, estes métodos geralmente são utilizados para construção de uma filogenia inicial, que posteriormente é submetida a algum método do grupo qualitativo. Como principal ponto negativo, estes métodos apresentam apenas uma filogenia como resultado final (ver adiante).

Idealmente, todas as possíveis árvores para um dado alinhamento deveriam ser analisadas para garantir a escolha da melhor filogenia. Para isso, é necessário atribuir certos parâmetros que avaliem, dentre todas as ár-

Tabela 1-5: Comparação entre os tipos de métodos para inferência de filogenias.

Tipo	Método	Princípio	Programa
Métodos Quantitativos	UPGMA	Agrupa sequencialmente as OTUs com menor distância evolutiva entre si	Geneious MEGA
	Aproximação dos vizinhos	Busca a árvore com a menor soma total de ramos	MEGA Geneious HyPhy
Métodos Qualitativos	Máxima Parcimônia	Busca a filogenia com menor número de eventos evolutivos	PAUP MEGA Mesquite
	Máxima Verossimilhança	Busca a árvore com o valor de maior verossimilhança entre todas as filogenias construídas	PAUP PAML phyML MEGA
	Estatística Bayesiana	Amostra um número representativo de filogenias a partir do espaço amostral total de árvores e busca a mais provável	Mr. Bayes BEAST BAMBE

vores, aquela que explica as relações evolutivas de forma mais precisa.

Assim, os métodos qualitativos envolvem algoritmos que atribuem um critério de otimização para escolher a melhor filogenia. Nestes métodos, diversas filogenias são construídas e, seguindo um critério definido pelo algoritmo utilizado, uma filogenia será identificada como a que melhor explica a relação evolutiva entre os OTUs. O critério é utilizado para atribuir um valor a cada filogenia e ordená-las segundo este valor.

Estes métodos têm a vantagem de requerer uma função explícita para escolha das filogenias, sendo portanto independente da escolha do operador. No entanto, devido ao caráter de sua análise, são métodos mais refinados e intrinsecamente mais demorados computacionalmente. Três critérios de otimização são tradicionalmente empregados na inferência de filogenias: (a) Máxima Parcimônia, (b) Máxima Verossimilhança e (c) Inferência Bayesiana.

Por se tratarem de métodos que buscam uma única filogenia entre diversas árvores, os métodos qualitativos exigem algoritmos que vasculhem o maior número possível de filogenias em busca da melhor árvore. Dois grupos de algoritmos são destacados: os algoritmos exatos e os algoritmos heurísticos. Atualmente, devido ao tempo e à exigência computacional, os métodos heurísticos são preferidos aos exatos. No entanto, qualquer um deles pode ser aplicado aos métodos qualitativos de inferência filogenética. Como desvantagem dos métodos qualitativos, repetidos processos de procura em um mesmo conjunto de sequências podem levar a resultados diferentes, dependendo da árvore que é construída inicialmente pelo algoritmo.

Os métodos exatos buscam todas as filogenias possíveis para um grupo de sequências. O funcionamento destes métodos geralmente envolve a seleção aleatória inicial de três OTUs para a construção de uma árvore filogenética não enraizada. Por tentativa, um a um, novas OTUs, também tomadas aleatoriamente do alinhamento, são inseridas em diferentes posições na árvore. Esse procedimento é repetido até todos os táxons serem inseridos, garantindo que todas as filogenias possíveis para o alinhamento dado sejam geradas.

A partir da aplicação de um critério de otimização (dado pelo método qualitativo) para classificar as filogenias e ordená-las segundo este valor, é possível organizar um espaço virtual que contém todas as filogenias possíveis para o alinhamento empregado. É importante lembrar que, tomando poucas sequências, milhões de árvores podem ser geradas. Este conjunto total de filogenias é comumente chamado de espaço amostral. Como exemplo, podemos organizar o espaço amostral de filogenias originadas a partir de um alinhamento de dez sequências em um gráfico bidimensi-

onal baseado no valor atribuído pelo critério de otimização a cada árvore (Figura 9-5). Nestas condições, será possível observar que algumas árvores possuem valores maiores que outras, formando picos que agrupam as melhores filogenias. Da mesma forma, entre diferentes picos existem vales representados por árvores com valores menores e, portanto, menos consistentes.

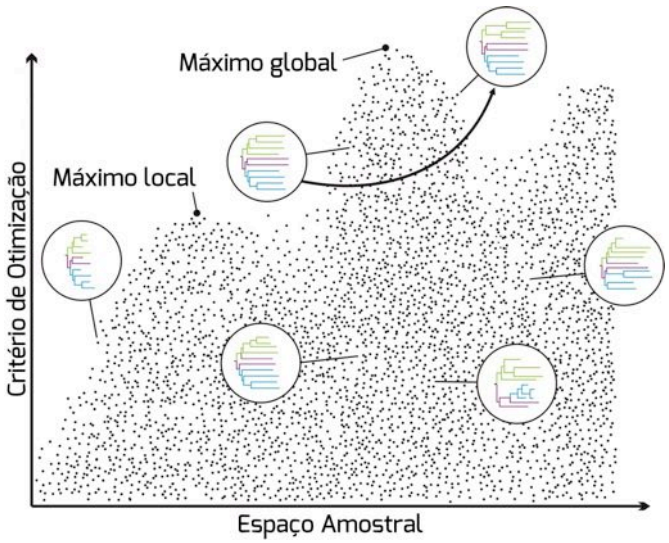


Figura 9-5: Descrição de parte do espaço amostral das possíveis filogenias para um determinado sistema, ordenadas segundo um valor atribuído pelo critério de otimização. Cada ponto no gráfico representa uma topologia diferente inferida a partir de um conjunto de dez sequências homólogas. O espaço amostral, neste caso, é definido por 2.027.025 filogenias e apresenta, segundo o critério de otimização, dois máximos locais e um máximo global, que contém as melhores filogenias. Em destaque, algumas filogenias exemplificando as possibilidades de arranjo dos ramos. A seta indica a mudança de topologia da filogenia e o consequente aumento de seu valor dado pelo critério de otimização.

Os métodos de busca exaustiva construirão um espaço amostral de árvores através de métodos específicos de modificação das filogenias. Por acumularem um grande número de resultados, estes métodos exigem um tempo computacional muito elevado, por vezes tornando-se proibitivos.

Os algoritmos de busca heurística procuram pela melhor filogenia em um subconjunto de todas as filogenias possíveis. Apesar de serem muito mais rápidos computacionalmente, estes métodos não garantem que a filogenia correta seja encontrada, pois apenas algumas árvores do espaço amostral total serão consideradas. Ainda assim, estes métodos tem mostrado grande eficiência.

Atualmente, os principais métodos qualitativos de inferência filogenética incorporam algoritmos de busca heurística para amostrar as filogenias do espaço amostral virtual. Usualmente, estes algoritmos de busca são executados em dois passos. Primeiramente, diferentes árvores são construídas e, após encontrar a melhor árvore guiada por um critério de otimização, aplica-se um algoritmo para modificar aleatoriamente o arranjo dos ramos. Este método permite testar se outros arranjos são ou não mais consistentes.

Devido ao grande número de métodos para inferência filogenética, a decisão quanto ao uso de cada um é de grande importância para a interpretação do resultado final: a filogenia. Ao escolher um método, é fundamental verificar o poder (tamanho e quantidade de sequências necessária para resolver a filogenia), a eficiência (habilidade de estimar a filogenia correta com um número limitado de dados), a consistência (habilidade de estimar a filogenia correta com um número de dados ilimitado) e a robustez (habilidade de estimar a filogenia correta quando certos pressupostos da análise são violados).

Até o momento, não existe um método que apresente todas estas características simultaneamente e garanta a reconstrução filogenética correta. É importante, sobretudo, conhecer a biologia do organismo (ou dos organismos) em questão para que a escolha do método tenha, além de tudo, uma justificativa biológica.

5.6. Abordagens quantitativas

UPGMA

O método baseado em distâncias UPGMA (*unweighted pair-group method using arithmetic averages*, ou método de agrupamento par a par usando médias aritméticas não ponderadas) foi proposto por Sneath e Sokal, em 1973, e é o método mais simples para reconstrução filogenética. O UPGMA

parte do pressuposto de que todas as linhagens evoluem a uma taxa constante (hipótese do relógio molecular).

No UPGMA, uma medida de distância evolutiva é computada para todos os pares de sequências utilizando um modelo evolutivo. Após, estas distâncias são organizadas na forma de uma matriz, conforme ilustrado abaixo:

Sequências	*1*	*2*	*3*	*4*
2	$d_{1,2}$			
3	$d_{1,3}$	$d_{2,3}$		
4	$d_{1,4}$	$d_{2,4}$	$d_{3,4}$	
5	$d_{1,5}$	$d_{2,5}$	$d_{3,5}$	$d_{4,5}$

O agrupamento das sequências é iniciado pelo par com menor distância. Supondo que $d_{1,2}$ seja a menor distância no exemplo acima, as sequências *1* e *2* são agrupadas com um ponto de ramificação na metade dessa distância ($d_{1,2/2}$). As sequências *1* e *2* são então combinadas em uma entidade composta, agora denominada *y*, e a distância entre esta entidade *y* e as outras sequências é computada (observe abaixo).

Sequências	$y_{(1,2)}$	*3*	*4*
3	$d_{y,3}$		
4	$d_{y,4}$	$d_{3,4}$	
5	$d_{y,5}$	$d_{3,5}$	$d_{4,5}$

Supondo que $d_{y,3}$ seja a menor distância, *y* e *3* são combinados em uma nova entidade composta, digamos, *z*. Seu ponto de ramificação é calculado levando em conta a distância de cada membro de *y* (*1* e *2*) em relação a *3* e dividindo por *2*, ou seja, $(d_{1,3}+d_{2,3})/2$. O mesmo procedimento se repete, calculando a menor distância entre *z* e outra sequência (suponhamos que seja a sequência *4*). Calculam-se a distância de cada membro de *z* até *4*, divide-se o somatório das distâncias por dois e cria-se uma nova sequência composta. O mesmo procedimento é repetido até que existam apenas duas sequências a serem agrupadas (comumente, uma sequência simples e uma entidade composta).

Ao empregar sequências de DNA ou proteína proximamente relacionadas, o UPGMA pode construir duas ou mais "árvores empatadas" (*tie trees*). Essas árvores surgem quando dois ou mais valores de distância na matriz se mostram idênticos. É possível representar todas as árvores empatadas, mas essa abordagem é pouco útil, uma vez que tais árvores são muito semelhantes e surgem por erros de estimativa das distâncias. Para tais casos, sugere-se apresentar uma única árvore, geralmente a árvore consenso do *bootstrap* (ver seção 5.8).

Por se basear na hipótese do relógio molecular, o UPGMA pode levar à obtenção de topologias falsas quando tal hipótese não for satisfeita pelos dados. Sabe-se que o método é muito sensível a variações nas taxas evolutivas entre linhagens, fato este que levou a proposição de métodos onde as variações são ajustadas para a obtenção de sequências que satisfaçam o relógio molecular. Apesar disso, devido ao surgimento de métodos mais robustos e mais eficientes em lidar com dados não uniformes, o UPGMA encontra-se praticamente abandonado como alternativa para reconstrução filogenética.

Aproximação dos Vizinhos

O método de aproximação dos vizinhos (*neighbor joining* ou NJ) foi proposto por Saitou e Nei em 1987. Este método se baseia em um aceleramento dos algoritmos de evolução mínima que existiam até então. Em sua versão original, estes algoritmos buscavam a árvore com menor soma total de ramos, de maneira que todas as árvores possíveis precisavam ser construídas para que se verificasse qual delas apresentava a menor soma. O algoritmo de NJ facilitou esse processo, tendo o princípio de evolução mínima implícito no processo e produzindo apenas uma árvore final.

Para construir a filogenia, o NJ começa por uma árvore totalmente não resolvida (topologia em estrela) (Figura 10-5). Tendo como base uma matriz de distâncias (semelhante à matriz inicial construída pelo método de UPGMA) entre todos os pares de sequências, construída a partir da aplicação de um modelo de substituição (conforme descrito na seção 5.4), o par que apresentar a menor distância é identificado, unido por um nó (que representará o ancestral comum deste par de sequências) e incorporado na árvore (na Figura 10-5, f e g são unidos pelo nó u). As distâncias de cada sequência do par são recalculadas em relação ao novo nó u, assim como as distâncias de todas as outras sequências são recalculadas em relação ao novo nó u. O algoritmo reinicia, substituindo o par de vizinhos unidos pelo novo nó e usando as distâncias calculadas no passo anterior.

Quando duas somatórias de ramos são iguais, a decisão sobre quais ramos unir depende do programa empregado. Alguns optam pela primeira sequência apresentada no arquivo de dados, enquanto outros escolhem aleatoriamente qual dos pares deve ser unido primeiro. Árvores empatadas (*tie trees*) são raras com o uso de NJ, e recomenda-se o emprego da árvore consenso do *bootstrap* (ver seção 5.8) para evitá-las. Uma variação do algoritmo NJ, o BIONJ tem se mostrado ligeiramente melhor que o NJ em casos pontuais; no entanto, conserva o mesmo princípio do algoritmo.

5.7. Abordagens qualitativas

Parcimônia

O princípio de parcimônia foi proposto por Guilherme de Occam (ou *William of Ockham*) no século XVII. Occam defendia que a natureza é por si só econômica e opta por caminhos mais simples. O pensamento se espalhou por diversas áreas do conhecimento e, atualmente, seu princípio é conhecido como Navalha de Occam.

Historicamente, a parcimônia teve um papel muito importante no estabelecimento da disciplina de filogenética molecular. Desde 1970, foi o critério de otimização mais utilizado para inferência de filogenias.

Contudo, atualmente a máxima parcimônia foi substituída por outros métodos, como máxima verossimilhança e inferência Bayesiana devido, principalmente, às simplificações nos processos evolutivos assumidas pelo método e, sobretudo, nas limitações de seu uso. Apesar disso, a máxima parcimônia ainda está integrada ao campo da inferência filogenética por ser um método rápido e, em alguns casos, muito efetivo.

A aplicação do princípio de máxima parcimônia nas reconstruções filogenéticas é conceitualmente simples: dentro de um conjunto de filogenias, aquela filogenia que apresentar o menor número de eventos evolutivos (substituições) deve ser a mais provável para explicar os dados do alinhamento.

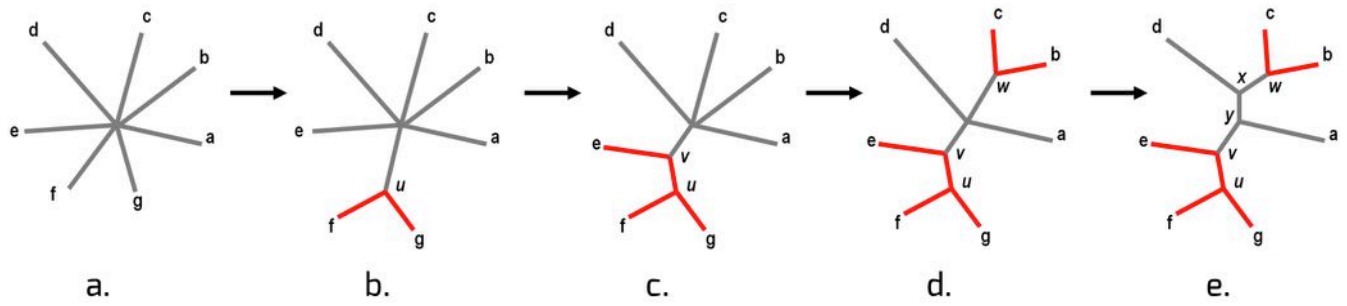


Figura 10-5: Começando com uma árvore em estrela (a), a matriz de distâncias é calculada para identificar o par de nós a ser unido (nesse caso, f e g). Estes são unidos ao novo nó u (b). A porção em vermelho é fixada e não será mais alterada. As distâncias do nó u até os nós a-e são calculadas e usadas para unir o próximo vizinho. No caso, u e e são unidos ao recém criado nó v (c). Mais duas etapas de cálculo levam à árvore em (d) e então à árvore em (e), que está totalmente resolvida, encerrando o algoritmo.

Metodologicamente, o critério de parcimônia deve determinar a quantidade total de mudanças na filogenia, descrevendo o tamanho dos ramos. Adicionalmente, a parcimônia guia a busca, entre todas as árvores possíveis, daquela filogenia que minimiza os passos evolutivos de forma máxima sendo, portanto, a filogenia de máxima parcimônia.

Assim que uma determinada filogenia é proposta, o método calculará as probabilidades de mudanças dos nucleotídeos desde os ramos terminais até os ramos mais ancestrais da árvore. Por se tratar de um método qualitativo, a parcimônia considera cada sítio do alinhamento individualmente e calcula as probabilidades de ocorrência dos quatro nucleotídeos nos táxons ancestrais.

Devido ao caráter probabilístico do método, é necessário que certas pressuposições sejam estabelecidas para especificar o custo de substituição dos nucleotídeos. A forma mais simples do método (Parcimônia de Wagner) assume que as substituições de nucleotídeos tem custo 1, enquanto que a não alteração não é penalizada (Figura 11-5a). No entanto, esquemas um pouco mais complexos que levam em consideração as questões biológicas envolvidas no processo evolutivo foram propostas. Um esquema comum de matriz com custo desigual, proposto para especificar as transições e as transversões, leva em consideração a diferença na probabilidade de mudança entre purinas e pirimidinas (Figura 11-5b). Comumente, a matriz é especificada sem que constem os respectivos nucleotídeos, no entanto, por convenção são atribuídos nas linhas e colunas em ordem alfabética (A, C, G e T).

Para o método de parcimônia, apenas sítios variáveis são considerados informativos. Estes sítios devem apresentar dois caracteres diferentes presentes em, no mínimo, dois indivíduos (Figura 12-5b). Aqueles sítios que não apresentam variação ou apresentam autapomorfias (caracter diferente presente em apenas um indivíduo) serão descartados automaticamente das análises.

Devido ao tamanho dos alinhamentos e ao número de OTUs incluídas para a inferência de filogenias, foi necessário que algoritmos fossem desenvolvidos para acelerar os cálculos na busca pela árvore de máxima parcimônia. Algoritmos de programação dinâmica são capazes de lidar com a atribuição de custos e realizar os devidos cálculos para escolha da filogenia com o menor custo. Diversos algoritmos foram desenvolvidos, embora a parcimônia de Sankoff, desenvolvida em 1975, tenha se tornado uma das mais populares.

a.

Matriz de custo igual =

	A	C	G	T
A	0	1	1	1
C	1	0	1	1
G	1	1	0	1
T	1	1	1	0

b.

Matriz de custo desigual =

	A	C	G	T
A	0	4	1	4
C	4	0	4	1
G	1	4	0	4
T	4	1	4	0

Figura 11-5: Matrizes de custo aplicadas ao método de máxima parcimônia para penalizar as substituições de um nucleotídeo por outro. (a) Matriz de custos iguais para todas as mudanças entre nucleotídeos. (b) Matriz de custo desigual, considerando a maior probabilidade de ocorrência de transições em relação às transversões ao longo do processo evolutivo.

Após a atribuição de uma matriz de custo e a proposição de uma filogenia, o algoritmo utilizará cada um dos sítios informativos do alinhamento independentemente para cálculo dos custos (Figura 11-5).

Considere a matriz desigual da Figura 11-5b e a filogenia inicialmente proposta na Figura 12-5a. O esquema demonstra que para cada sítio informativo será construída uma filogenia com a mesma topologia da árvore proposta em 12-5a (ver adiante).

Tomando, por exemplo, o sítio 28, identificamos a presença de três ancestrais não amostrados que, no entanto, para o cálculo dos custos, terão que ter seus caracteres inferidos. Segundo o algoritmo de Sankoff, os cálculos devem iniciar tomando os clados mais derivados (isto é, mais recentes). Em 12-

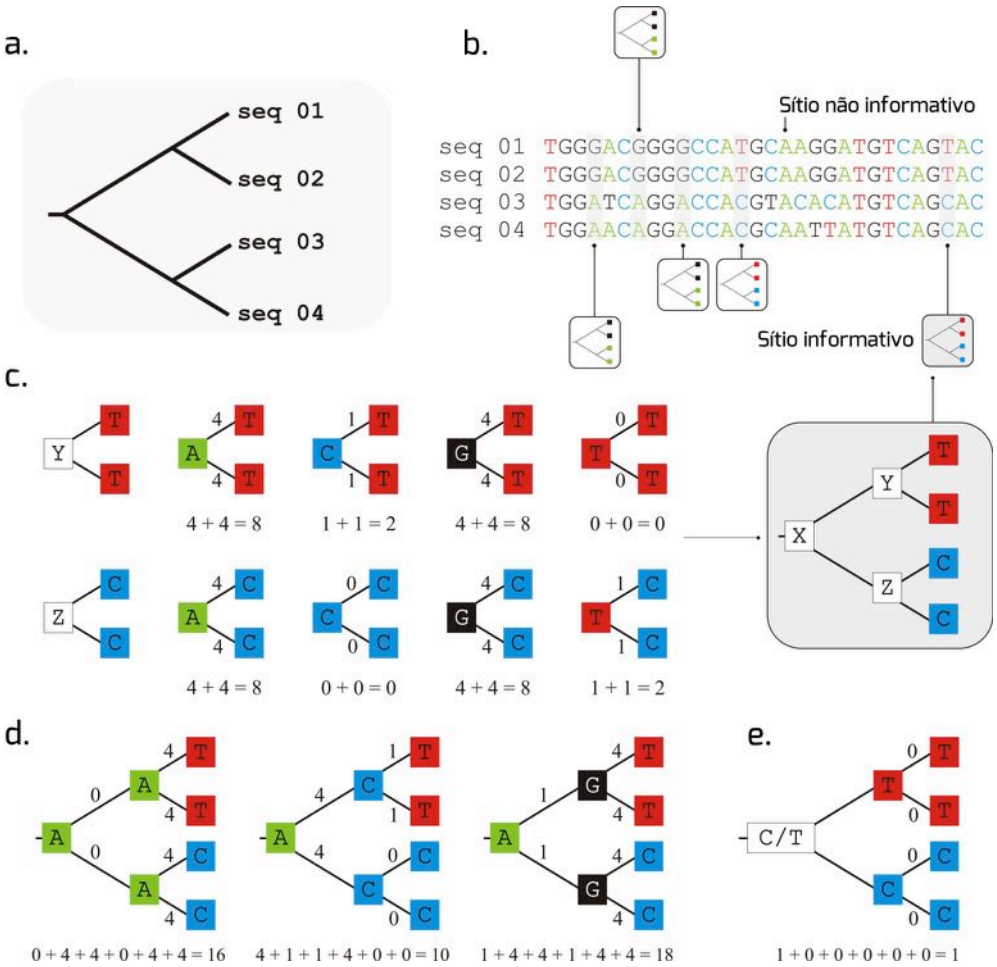


Figura 12-5: Determinação dos custos de substituição pelo método de parcimônia para um sítio do alinhamento de nucleotídeos. (a) Topologia da filogenia proposta para quatro táxons (ver adiante). (b) Alinhamento de nucleotídeos de quatro sequências homólogas. Destacados em cinza estão os sítios informativos para o método de parcimônia. Os demais sítios são considerados não informativos e serão descartados durante os cálculos. (c) Cálculo dos custos para os dois clados presentes na filogenia proposta em "a". O método supõe que a posição "Y" possa ser ocupada por qualquer um dos quatro nucleotídeos. (d) Exemplo do procedimento adotado pelo método, supondo que a posição "X" na filogenia foi ocupada pelo nucleotídeo A. É necessário considerar todas as possibilidades de caracteres nos sítios ancestrais e calcular os respectivos custos. (e) Arranjo de menor custo para a posição 28 do alinhamento de nucleotídeos.

5c, a posição "Y" da filogenia necessariamente foi ocupada por um dos quatro nucleotídeos. Em cada uma das proposições (A, C, G ou T), o custo associado à substituição é consultado na matriz. No primeiro caso, a hipótese para ocupação da posição "Y" é A. O custo da substituição em cada um dos ramos deve ser verificado e somado. Por exemplo, a substituição de A por T possui custo 4. Como a mesma substituição ocorreu em dois ramos diferentes, somamos o custo total, que totaliza 8. O mesmo procedimento será repetido considerando os outros três nucleotídeos na posição "Y".

Após o cálculo dos custos para as posições "Y" e "Z", é necessário verificar os custos de substituição de "X" para "Y" e "X" para "Z". A Figura 12-5d apresenta a primeira hipótese para ocupação da posição "X": o nucleotídeo A. Aqui, o algoritmo somará os custos de substituição de todos os ramos, novamente considerando cada um dos quatro

nucleotídeos na posição "X", mas também considerando a variação nas posições "Y" e "Z". A Figura 12-5e identifica a filogenia com o menor custo para o sítio 28. Note que o caractere mais ancestral pode ser tanto o nucleotídeo T quanto C. Os mesmos cálculos serão realizados para todos os sítios do alinhamento, tomando a topologia dada em 12-5a e, ao final, os menores custos para cada sítio serão somados para encontrar o tamanho dos ramos da árvore. A árvore que possuir os ramos mais parcimoniosos será tomada como a árvore de máxima parcimônia.

Computacionalmente, o cálculo dos tamanhos de ramos mais parcimoniosos não é um problema. O desafio da maioria dos métodos de reconstrução filogenética está na inferência da topologia. Assim como no método de máxima verossimilhança, discutido a seguir, o método de máxima parcimônia contará com algoritmos heurísticos para arranjo das topologias. A filogenia é então proposta pelo algoritmo, e o critério de parcimônia avalia a árvore. A partir de perturbações realizadas nesta topologia, uma nova topologia é proposta e novamente o critério qualifica a filogenia.

Apesar de velozes, os métodos de parcimônia falham ao estimar a relação evolutiva entre um grande número de táxons, especialmente se diferentes linhagens possuem taxas evolutivas variáveis ou taxas evolutivas muito rápidas. Nestes casos, é comum que o método agrupe incorretamente os táxons com maiores taxas de evolução, levando à inferência da filogenia errada (atração de ramos longos).

Ainda, por não ter um modelo de substituição especificado, o método de parcimônia é incapaz de considerar mutações reversas ou múltiplas substituições. Métodos que geram diferentes hipóteses a partir do alinhamento, considerando as observações biológicas na seleção do modo de substituição dos nucleotídeos e, assim, lidam com eventos aleatórios de probabilidade, substituíram o uso da máxima parcimônia e, atualmente, são os principais métodos utilizados para a inferência de filogenias.

Máxima Verossimilhança

Idealmente, os métodos de inferência filogenética devem resgatar o máximo de informações contidas em um dado conjunto de sequências homólogas, buscando desvendar a verdadeira história evolutiva dos organismos.

Quando um grande número de mudanças evolutivas em diferentes linhagens é demasiadamente desigual, o método de máxima parcimônia tende a inferir filogenias inconsistentes, proporcionalmente convergindo à árvore errada quanto maior o número de sequências no alinhamento. Assim, abre-se espaço para uma técnica de inferência filogenética mais robusta, que alie as informações do alinhamento a um modelo estatístico capaz de lidar com a probabilidade de mudança de um nucleotídeo para outro de maneira mais completa.

Dentro do campo da filogenética computacional, o método de máxima verossimilhança primeiramente ocupou este espaço e, desde então, tem sido amplamente utilizado devido à qualidade da abordagem estatística empregada.

A implementação de uma concepção estatística para a máxima verossimilhança, originalmente desenvolvida para estimar parâmetros desconhecidos em modelos probabilísticos, se deu entre 1912 e 1922 através dos trabalhos de A. R. Fisher.

Apesar de utilizado para dados moleculares na década de 1970, o método de máxima verossimilhança só se tornou popular na área da filogenética a partir de 1981, com o desenvolvimento de um algoritmo para estimar filogenias baseadas no alinhamento de nucleotídeos. Atualmente, diversos programas implementam este método para realizar a inferência filogenética, incluindo PAUP, MEGA, PHYLIP, fastDNAml, IQPNNI e METAPIGA, dentre outros (Tabela 1-5).

O objetivo principal do método da máxima verossimilhança é inferir a história evolutiva mais consistente com relação aos dados fornecidos pelo conjunto de sequências. Neste

modelo, a hipótese (topologia da árvore, modelo de substituição e comprimento dos ramos) é avaliada pela capacidade de predizer os dados observados (alinhamento de sequências homólogas). Sendo assim, a verossimilhança de uma árvore é proporcional à probabilidade de explicar os dados do alinhamento. Aquela árvore que com maior probabilidade, entre as outras árvores possíveis, produz o conjunto de sequências do alinhamento, é a árvore que reflete a história evolutiva mais próxima da realidade, mais verossímil e, por isso, de máxima verossimilhança.

É importante ressaltar que diferentes filogenias podem explicar um determinado conjunto de sequências, algumas com maior probabilidade e, outras, com menor probabilidade. No entanto, a soma das verossimilhanças de todas as árvores possíveis para um determinado conjunto de sequências nunca resultará em 1, pois não estamos lidando com as probabilidades de que estas filogenias estejam corretas, mas avaliando a probabilidade de explicarem o alinhamento que foi fornecido.

Se, por exemplo, aplicássemos o método de máxima verossimilhança para inferir a árvore filogenética de um grupo de sequências homólogas que incluem porções recombinantes, encontraríamos uma árvore filogenética com um determinado valor de verossimilhança. A utilização do método, por si só, garantiria como resultado a inferência de uma filogenia. No entanto, sabemos que esta árvore, apesar de ser a mais plausível para explicar o alinhamento dado, não tem qualquer relação com a realidade evolutiva do organismo, já que eventos de recombinação aconteceram no decorrer do tempo e impedem a explicação sob a forma dicotômica de uma filogenia.

A aplicação do método de máxima verossimilhança exige a construção de uma filogenia inicial, geralmente obtida por métodos quantitativos. Como exemplo, considere a árvore filogenética proposta inicialmente e o respectivo alinhamento de nucleotídeos da Figura 13-5. Para calcularmos a verossimilhança desta filogenia será necessário utilizar um modelo evolutivo, que será importante para atribuir valores e parâmetros às substituições e ajudará no cálculo da probabilidade de que uma sequência X mude para uma sequência Y ao longo de um segmento da árvore.

Dado um determinado modelo evolutivo (JC69, K2P, F81, HKY ou GTR, por exemplo), e assumindo que cada sítio do alinhamento evolui de maneira independente dos demais, podemos calcular o valor de verossimilhança para cada um destes sítios e, posteriormente, multiplicar os valores de cada sítio para encontrar a verossimilhança da árvore dada (Figura 13-5 e a Figura 14-5). Sítios que apresentam deleções serão eliminados da análise.

Como os nós internos destas árvores, geradas a partir de cada sítio do alinhamento, são a representação de OTUs não amostrados (isto é, ancestrais) e, por conseguinte, não se conhecem suas sequências de nucleotídeos, será necessário considerar a ocorrência de todos os nucleotídeos (A, T, C e G) nestas posições da árvore (Figura 13-5c).

Por certo, alguns cenários são mais prováveis que outros; no entanto, todos devem ser considerados durante os cálculos de verossimilhança, pois apresentam alguma probabilidade de terem gerado as sequências dadas no alinhamento. Adicionalmente, além de calcular a probabilidade de todas as mudanças possíveis para cada um dos sítios do alinhamento (Figura 13-5c), a expressão matemática da verossimilhança ainda incluirá o tamanho dos ramos, dentre outros elementos do modelo de substituição, como um fator determinante para o cálculo (Figura 13-5d).

A probabilidade de ocorrência de cada um dos quatro nucleotídeos no nó mais interno da árvore será igual à respectiva frequência estacionária dada pelo modelo de substituição, já que este parâmetro especifica a proporção esperada de cada um dos quatro nucleotídeos. No modelo de Jukes e Cantor, por exemplo, assume-se que os quatro nucleotídeos ocorrem em proporções iguais de 25%.

Conforme o exemplo da Figura 13-5d, a equação utilizada para calcular a verossimilhança da filogenia

proposta no sítio 28, inicialmente, leva em consideração a frequência estacionária do nucleotídeo G, já que este é o nucleotídeo que está sendo considerado como presente no nó mais ancestral da árvore. A probabilidade de este G ser substituído por um A (P_{GA}), ou permanecer G (P_{GG}) será dada pelo modelo de substituição escolhido. Da mesma forma, serão os casos P_{GT}, P_{AC} (repetido duas vezes cada pelo fato de existirem dois ramos terminais com o mesmo nucleotídeo).

O tamanho dos ramos entre dois nós será multiplicado pelas probabilidades de substituição dos nucleotídeos, levando em conta variações em parâmetros do modelo de substituição. Apesar da dificuldade de cálculo computacional, os algoritmos aplicados à inferência filogenética (baseados no princípio de Pulley) automaticamente estimarão o tamanho de cada ramo de modo que este maximize o valor da verossimilhança da árvore filogenética em construção. Nestes casos, o algoritmo atribui diversos valores de distância para um ramo e, a cada valor, verifica a verossimilhança da árvore, buscando aqueles valores que resultam na filogenia com a maior verossimilhança.

A probabilidade de observar os dados em um sítio particular é a soma das probabilidades de todos os possíveis nucleotídeos que poderiam ser observados nos nós internos da árvore (Figura 13-5c). O número de

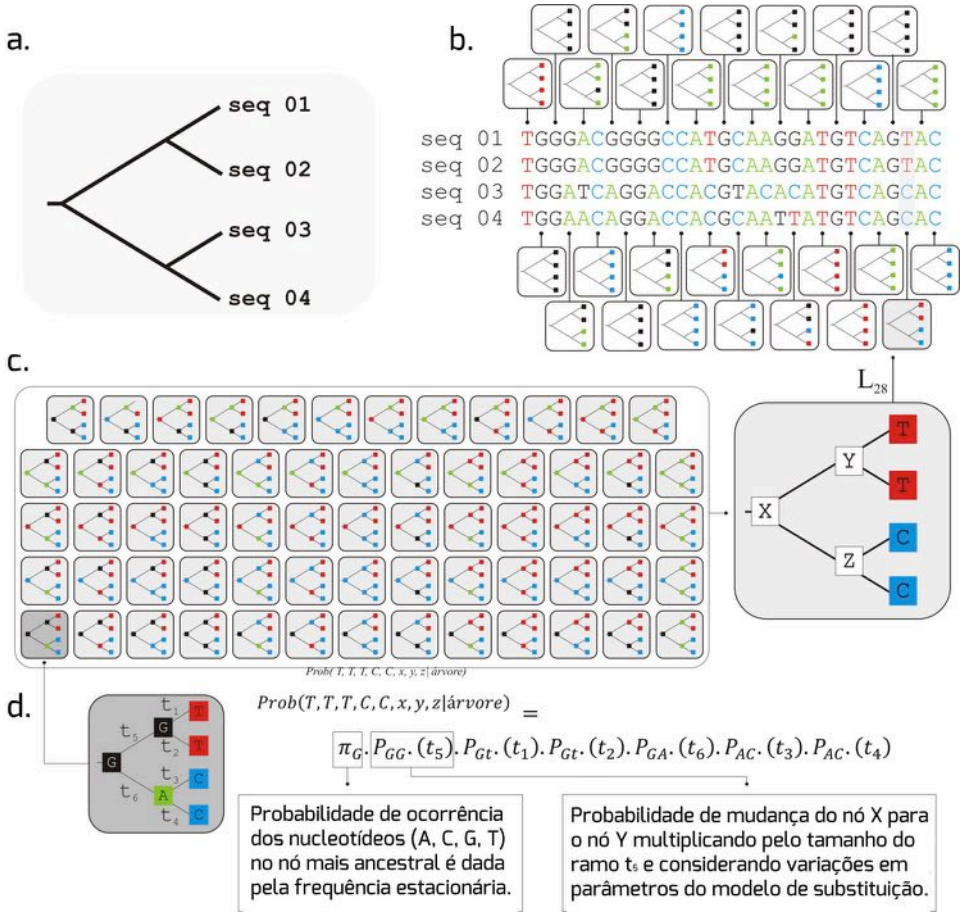


Figura 13-5: Esquema do cálculo da verossimilhança para uma filogenia e seu respectivo alinhamento de nucleotídeos. (a) Árvore filogenética proposta inicialmente para o alinhamento em "b". (b) Para cada posição do alinhamento é destacada a organização dos quatro sítios do alinhamento na árvore proposta em "a". Como exemplo, apenas o sítio do alinhamento destacado em cinza será considerado para o cálculo da verossimilhança. Os quadrados pretos, azuis, verdes e vermelhos nos ramos terminais das filogenias representam, respectivamente, os nucleotídeos guanina, citosina, adenina e timina. (c) Probabilidade de cada uma das 64 possíveis combinações de nucleotídeos nos nós internos da árvore, já que estes representam os sítios de táxons ancestrais não amostrados (P_{XY}, P_{YT}, P_{XZ}, P_{ZC}). (d) O esquema para o cálculo da máxima verossimilhança leva em conta a multiplicação do tamanho dos ramos (t_1, t_2, t_3, t_4, t_5 e t_6) pelas respectivas probabilidades de transição (P_{GG}, P_{GT}, P_{GA} e P_{AC}), além da frequência estacionária dos quatro nucleotídeos no nó mais ancestral (π_X).

nós internos rapidamente se torna muito grande com o aumento do número de OTUs. Felizmente, através de um algoritmo criado por Felseinstein (algoritmo de "poda"), que se aproveita da própria topologia da filogenia, esses cálculos podem ser realizados de uma maneira computacionalmente eficiente.

Neste processo, propõe-se que os cálculos da verossimilhança de uma determinada árvore sejam feitos a partir de sub-árvores dos ramos terminais em direção aos nós internos, semelhante ao algoritmo usado para o cálculo da parcimônia. No entanto, quando aplicado este método à inferência por máxima verossimilhança é necessário garantir que os modelos de substituição, não presentes no método de máxima parcimônia, sejam reversíveis, ou seja, que a probabilidade de mudança de A para T (P_{AT}) seja a mesma que T para A (P_{TA}). A introdução deste método permitiu que as análises de verossimilhança pudessem ser aplicadas a grandes conjuntos de sequências, de forma mais rápida e efetiva.

Ao final, multiplicamos os valores de verossimilhança de todos os sítios e encontramos o valor de verossimilhança da árvore (Figura 14-5):

A expressão matemática acima indica que a verossimilhança (L) é igual à multiplicação (Π) das probabilidades de cada sítio i (D^i, calculado conforme Figura 13-5), dada a árvore filogenética (topologia, modelo evolutivo e tamanho dos ramos). Aquela árvore que tiver o maior valor de verossimilhança entre todas as árvores possíveis para um determinado alinhamento de sequências será a árvore que melhor explica o alinhamento e, por isso, a árvore de máxima verossimilhança. Por fim, é importante ressaltar que, apesar de estarmos avaliando nucleotídeos neste exemplo, o mesmo raciocínio poderia ser aplicado para a inferência filogenética para um alinhamento de aminoácidos.

Até o momento vimos, em linhas gerais, como realizar o cálculo de verossimilhança para uma dada filogenia (Figura 13-5). No entanto, outra função importante dos métodos computacionais de inferência filogenética é apontar a topologia e encontrar a árvore de máxima verossimilhança entre todas as árvores possíveis para o conjunto de dados. Infelizmente, não existem algoritmos que garantam a localização da árvore real devido ao grande espaço amostral de árvores possíveis (Figura 9-5).

Após uma árvore ser construída, é necessário calcular sua verossimilhança e comparar este valor com todas as árvores já construídas. Como é impossível testar a verossimilhança para todas as filogenias possíveis, os algoritmos de máxima verossimilhança incluirão buscas heurísticas para solucionar este problema (estes métodos construirão diferentes filogenias a partir do mesmo conjunto de dados do alinhamento).

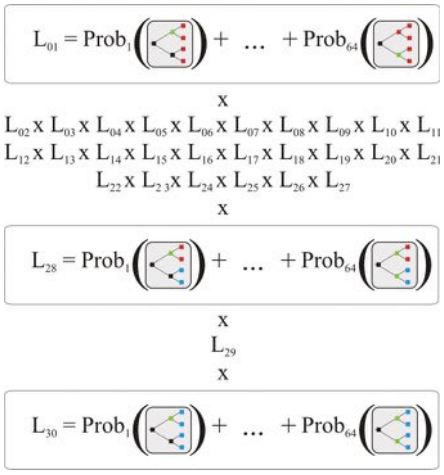


Figura 14-5: Cálculo da máxima verossimilhança de uma dada filogenia, considerando seu respectivo alinhamento de nucleotídeos contendo quatro táxons e 30 bases (Figura 13-5b). Para cada sítio (L_{01}, L_{02}, ..., L_{30}) será calculado um valor de probabilidade que envolve a consideração de todos os quatro nucleotídeos em cada um dos ramos ancestrais da filogenia. Posteriormente, os valores de verossimilhança de cada sítio serão multiplicados para encontrar a verossimilhança total da filogenia.

Na problemática das filogenias, diferentes programas têm proposto as mais diversas alternativas para avaliar o maior número de árvores do espaço amostral total e encontrar aquela com o maior valor de verossimilhança. No entanto, como regra geral, a maioria dos programas de máxima verossimilhança segue alguns passos comuns:

i) Uma filogenia preliminar com determinada topologia é construída (geralmente são utilizadas árvores construídas pelo método de aproxima-

ção de vizinhos);
ii) Os parâmetros para esta árvore são modificados buscando maximizar a verossimilhança (em alguns casos, a filogenia vai sendo construída pela adição de novos táxons aleatoriamente). Para a modificação da filogenia, os algoritmos podem implementar técnicas de rearranjos de ramos, conforme descrito em 5.4;
iii) O valor de máxima verossimilhança para esta árvore é armazenado;
iv) Outras topologias são construídas e seus parâmetros também são avaliados;
v) Finalmente, a filogenia que possuir o valor de máxima verossimilhança será a melhor estimativa evolutiva para o dado conjunto de sequências.

Embora estes processos simplifiquem os verdadeiros fenômenos biológicos que governam a evolução de uma sequência, apresentando assim dificuldades em identificar a árvore com o maior valor de verossimilhança, eles são normalmente robustos o bastante para estimar as relações evolutivas entre táxons.

Como estes métodos implicam em encontrar a árvore com o valor máximo de verossimilhança entre todas as árvores amostradas, o resultado final sempre fornecerá apenas uma filogenia, ao contrário dos métodos Bayesianos que serão vistos a seguir. Cabe ressaltar que, devido ao uso de diferentes algoritmos, na prática, um mesmo conjunto de sequências submetido a diferentes programas para inferência filogenética por máxima verossimilhança dificilmente resultará na mesma árvore. Por isso, é necessário ser cauteloso ao interpretar árvores geradas pelo método de máxima verossimilhança.

Análises Bayesianas

A estatística Bayesiana nasceu com a publicação de um ensaio matemático do reverendo Thomas Bayes, em 1793. Nesta publicação, o reverendo apresenta o desenvolvimento de um método formal para incorporar evidências prévias no cálculo da probabilidade de acontecimento de determinados eventos.

Inicialmente, este método foi aplicado apenas no campo da matemática e, só a partir de 1973, passa a ser incorporado no pensamento biológico e na inferência filogenética. Com o advento de diversos programas de acesso livre para realizar a inferência de filogenias por estatística Bayesiana, o método se difundiu e, atualmente, tornou-se um campo de estudo específico dentro da filogenética computacional.

A inferência Bayesiana engloba o método de máxima verossimilhança (Tabela 2-5) mas, adicionalmente, inclui o uso de informações dadas *a priori*. Estas informações refletem características a respeito da filogenia, do alinhamento ou dos táxons, que o pesquisador sabe de antemão.

Entre os principais parâmetros que podem ser conhecidos antes da reconstrução filogenética pode-se destacar a taxa evolutiva, tipo de relógio molecular, parâmetros do modelo de substituição, datas de coleta das amostras, datas para calibração da filogenia (achados fósseis, datação por carbono-14, aproximações arqueológicas, etc.), distribuição geográfica, organização monofilética de um grupo de indivíduos ou, até mesmo, parâmetros de dinâmica populacional.

Os valores atribuídos *a priori* são incorporados à estatística Bayesiana na forma de probabilidades e comporão o termo chamado de probabilidade anterior (*prior probability*). Se sabemos de antemão que um determinado grupo de organismos é ancestral em relação a outro, podemos atribuir uma maior probabilidade àquelas filogenias que relacionam estes organismos da maneira como sabemos *a priori*.

Qualquer informação útil, que é fornecida pelo pesquisador antes da própria reconstrução da filogenia, poderá ser convertida em uma probabilidade anterior para ser inserida nas análises de inferência Bayesiana. No entanto, as informações cedidas *a priori* devem

Tabela 2-5: Comparação entre os métodos de máxima verossimilhança e inferência Bayesiana.

Método	Vantagens	Desvantagens
Máxima Verossimilhança	Captura totalmente a informação dos sítios do alinhamento para construção das filogenias	Comparativamente ao método Bayesiano, o algoritmo para reconstrução por máxima verossimilhança é mais lento
Estatística Bayesiana	Tem grande ligação com a máxima verossimilhança, sendo, no entanto, geralmente mais rápida. Modelos populacionais podem ser incluídos para inferência das filogenias	Os parâmetros para as probabilidades anteriores devem ser especificados e pode ser difícil especificar quando as análises são satisfatórias

ser distribuições de números prováveis (mínimo e máximo), e não números exatos. Quando estes valores não são conhecidos ou quando, por exemplo, não se quer atribuir maior probabilidade a uma determinada topologia, o parâmetro terá uma distribuição uniforme de probabilidades.

Na maioria dos aplicativos que lidam com inferência Bayesiana existem distribuições uniformes associadas às probabilidades anteriores que assumem que todos os valores possíveis são dados pela mesma probabilidade.

Além das probabilidades anteriores, a inferência Bayesiana é baseada nas probabilidades posteriores de um parâmetro como, por exemplo, a topologia. Através da probabilidade posterior é possível verificar a probabilidade de cada uma das hipóteses (árvores filogenéticas). Sendo assim, ao final das análises, é possível estabelecer uma estimativa da probabilidade dos eventos retratados por uma determinada filogenia, ou seja, a probabilidade de cada filogenia. As probabilidades posteriores são calculadas utilizando a fórmula de Bayes:

$$\mathrm{L}(H \mid D) = \frac{\mathrm{L}(H)\,\mathrm{L}(D \mid H)}{\mathrm{L}(D)}$$

O termo $\mathrm{L}(H \mid D)$ é chamado de distribuição de probabilidades posteriores, e é dado pela probabilidade da hipótese (topologia da árvore, modelo de substituição e comprimento dos ramos) a partir dos dados disponíveis (alinhamento de sequências). O termo $\mathrm{L}(D \mid H)$ descreve o cálculo de máxima verossimilhança, enquanto o multiplicador $\mathrm{L}(H)$ é a probabilidade anterior. Para o termo que envolve a função de máxima verossimilhança, é ainda necessário considerar também todos os tópicos já discutidos na seção anterior. O denominador $\mathrm{L}(D)$ é uma integração sobre todas as possibilidades de topologias, tamanhos de ramo e valores para os parâmetros do modelo evolutivo, o que garante que a soma da probabilidade posterior para todos eles seja 1. O denominador atuará como um normalizador para o numerador. Reescrevendo, temos:

$$\mathrm{L}(filogenia|alinhamento) = \frac{\mathrm{L}(filogenia)\,\mathrm{L}(alinhamento|filogenia)}{\sum_{\mathrm{H}} \mathrm{L}(filogenia)\,\mathrm{L}(alnhamento|filogenia)}$$

onde o termo filogenia descreve a topologia da árvore, o modelo de substituição e o comprimento dos ramos. Assim, através da multiplicação das probabilidades anteriores pela verossimilhança, divididos pelo fator de normalização, o método busca a hipótese (topologia da árvore, o modelo de substituição e o comprimento dos ramos) em que a probabilidade posterior é máxima.

O objetivo da inferência Bayesiana é calcular a probabilidade posterior para cada filogenia proposta. No entanto, para cada árvore diversos parâmetros devem ser especificados pelo usuário, incluindo topologia, tamanho dos ramos, parâmetros do modelo de substituição, parâmetros populacionais, relógio molecular, taxa evolutiva, etc. Dada uma filogenia, todos os parâmetros terão sua probabilidade posterior calculada. Se dadas 1000 filogenias, teremos 1000 valores de probabilidade posterior para cada parâmetro.

Devido à impossibilidade de construção de todas as filogenias possíveis para a maioria dos alinhamentos, a análise Bayesiana se aproveita de técnicas de amostragem para estimar os valores esperados de cada parâmetro.

Neste sentido, os métodos de inferência

Bayesiana utilizam as Cadeias de Markov Monte Carlo (MCMC, *Monte Carlo Markov Chain*) para aproximar as distribuições probabilísticas em uma grande variedade de contextos. Esta abordagem permite realizar amostragens a partir do conjunto total de filogenias, relacionando cada filogenia a um valor probabilístico. Sem a aplicação de um método que obtenha amostras do espaço de possíveis filogenias, como o modelo de MCMC, a estimativa de todos os parâmetros se tornaria analiticamente impossível nos atuais computadores.

Um dos métodos de MCMC mais usados na inferência filogenética é uma modificação do algoritmo Metropolis, chamado de Metropolis-Hastings. A ideia central deste método é causar pequenas mudanças em uma filogenia (topologia, tamanho dos ramos, parâmetros do modelo de substituição, etc.) e, após a modificação, aceitar ou rejeitar a nova hipótese de acordo com o cálculo de razão das probabilidades. Este método garante que diversas árvores sejam amostradas do espaço total de filogenias, amostrando filogenias com probabilidade posterior mais alta (Figura 15-5):

i) Inicialmente, o algoritmo MCMC gera uma filogenia aleatória X, arbitrariamente escolhendo o tamanho dos ramos para dar início à cadeia;

ii) O valor de probabilidade associado a esta filogenia é calculado (probabilidade posterior calculada através da fórmula de Bayes);

iii) Perturbações aleatórias são realizadas nesta filogenia inicial X (mudanças na topologia, no tamanho dos ramos, nos parâmetros do modelo de substituição, etc.) e geram uma filogenia Y;

iv) A probabilidade posterior é calculada para a filogenia Y;

v) A filogenia Y é tomada ou rejeitada para o próximo passo baseado na razão R (probabilidade posterior de Y dividida pela probabilidade posterior de X). Se R é maior que 1, a filogenia Y é tomada como base para o próximo passo. Se R é menor que 1, um número entre 0 e 1 é tomado aleatoriamente. Se R é maior que o número aleatório gerado, a filogenia será tomada, no entanto se for menor, a filogenia Y é rejeitada;

vi) Se a nova proposta Y for rejeitada, retorna-se ao estado X e novas modificações serão realizadas nesta filogenia;

vii) Supondo que a proposta Y tenha sido aceita, ela sofrerá uma nova perturbação a fim de gerar uma nova filogenia;

viii) Todas as árvores amostradas são armazenadas para posterior comparação. Os pontos visitados formam uma

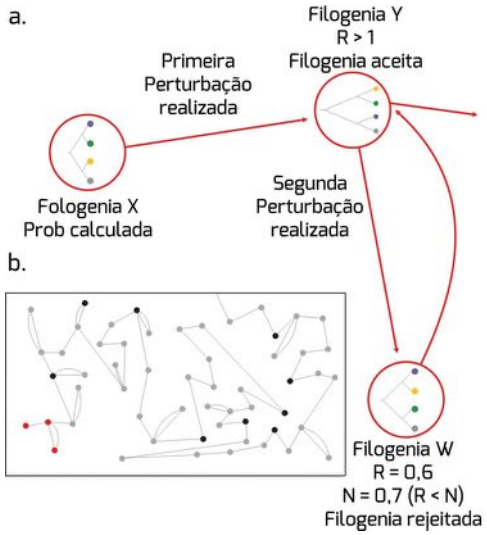


Figura 15-5: Esquema de amostragens MCMC aplicada à inferência filogenética pelo método Bayesiano utilizando o algoritmo de Metropolis-Hastings. (a) Após a proposição de uma filogenia inicial X, perturbações aleatórias são realizadas para gerar a filogenia Y. Devido à razão R>1, a nova filogenia é aceita. Nova perturbação é realizada para gerar a filogenia W e, devido a razão de probabilidades R resultar em um número menor que 1, um número aleatório N é sorteado. Sendo R<N, a nova proposição é rejeitada e a cadeia retorna à filogenia Y. (b) Andamento da cadeia na amostragem de filogenias. Cada círculo destaca uma nova filogenia que é proposta após a perturbação. As linhas conectando os círculos evidenciam a direção do andamento da cadeia. Apesar de a cadeia percorrer muitos passos, apenas alguns serão registrados para análise final (círculos pretos). Os círculos em vermelho são aqueles evidenciados em (a).

espécie de cadeia ao longo do espaço amostral total de filogenias.

O principal objetivo da cadeia é amostrar filogenias com probabilidades crescentes. No entanto, é importante que o algoritmo utilizado para tal permita que algumas árvores com menor probabilidade sejam amostradas para evitar que a cadeia fique "presa" em picos de máximo local (Figura 9-5).

Sendo assim, o cálculo da razão R considerando um valor aleatório entre 0 e 1 garantirá que, em determinados momentos, uma filogenia com menor probabilidade seja aceita. Por este método, é possível amostrar filogenias da região de um vale passando, por exemplo, de um pico de ótimo local para o pico de ótimo global (Figura 9-5).

A proposta de novas árvores na cadeia de Markov é uma etapa crucial para uma boa amostragem de filogenias. Na abordagem Bayesiana, uma boa amostragem inclui um grande número de filogenias, suficientemente diferentes entre si. Se filogenias muito diferentes são propostas, serão rejeitadas com muita frequência, pois é provável que tenham menor probabilidade posterior. Pelo contrário, se filogenias muito similares forem geradas, o espaço amostral não será varrido adequadamente e a cadeia deverá "correr" por muitos passos (amostrar um maior número de filogenias), aumentando o tamanho da cadeia e o tempo computacional.

Estimar o quanto a cadeia deve percorrer para amostrar um número suficiente de filogenias para as sequências dadas (espaço de árvores) é um fator fundamental para obter bons resultados em uma análise Bayesiana. Na maioria dos programas que utilizam estatística Bayesiana para inferir filogenias, o usuário deve especificar o tamanho da cadeia. Esse número é de grande subjetividade, e depende diretamente da distribuição das probabilidades anteriores, do número de táxons incluídos na filogenia e da relação evolutiva entre eles.

A Figura 16-5 exemplifica o andamento da amostragem da MCMC em um espaço de filogenias. Supondo que os quadrados em a, b e c representam um espaço amostral de filogenias, semelhante ao apresentado na Figura 15-5b, e que os pontos pretos sejam as filogenias que vão sendo amostradas com o desenvolvimento da MCMC vemos que, ao final do processo, depois de empregados 100 mil passos (Figura 16-5c), um grande número de filogenias foi amostrado.

Ainda, na região delimitada por um círculo, assumimos que estão as filogenias com maior probabilidade de explicar a história evolutiva de um grupo de organismos, ou seja, as filogenias reais. Note que quanto maior o número de passos percorridos pela cadeia, maior a amostragem do espaço de filogenias e maior o número de amostras dentro da região com filogenias de alta probabilidade.

Ao final, após o término da cadeia, a distribuição das probabilidades posteriores de todos os parâmetros deve ser verificada. No

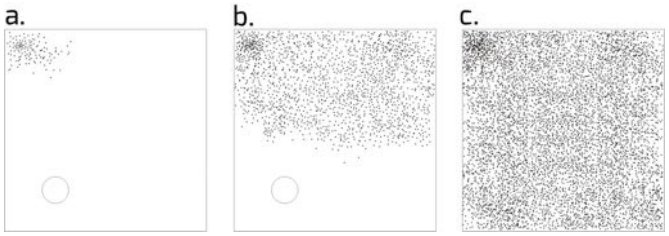


Figura 16-5: Espaço de possíveis árvores analisadas pela MCMC. Considerando que os quadrados descrevem o espaço amostral de todas as filogenias possíveis para um dado conjunto de sequencias, os pontos pretos representam as filogenias que foram amostradas ao longo da cadeia. Os círculos presentes no canto esquerdo inferior representam a região de máximo global (isto é, maior probabilidade) neste espaço amostral. O andamento da cadeia neste exemplo é o mesmo apresentado na Figura 15-5b (a) cento e trinta passos percorridos pela cadeia; (b) trinta mil passos percorridos pela cadeia; (c) cem mil passos percorridos pela cadeia. Nota-se que quanto maior o número de passos percorridos, maior a amostragem de filogenias no espaço. Da mesma forma, aumenta a probabilidade de a cadeia amostrar aquelas filogenias de máximo global.

entanto, as amostras tomadas no início da cadeia são tipicamente descartadas, pois estão sob forte influência do local de início da cadeia. As filogenias do início da cadeia estão muito longe de pontos máximos no espaço amostral e, por isso, é provável que todas as novas filogenias sugeridas subsequentemente sejam tomadas para o próximo passo (qualquer árvore proposta será mais provável que as árvores iniciais semelhantes àquela gerada aleatoriamente).

Esta fase inicial é conhecida como período de *burn in* (Figura 17-5). Conforme a cadeia avança, espera-se que a probabilidade das árvores amostradas aumente e, quando um número suficiente de filogenias for amostrado, chegue a uma distribuição estacionária. Em termos Bayesianos, espera-se que a cadeia atinja a convergência.

Um dos primeiros indicativos de que a cadeia convergiu para a distribuição correta está na estabilidade dos valores de probabilidade dos parâmetros da cadeia (cada parâmetro da filogenia poderá ter uma distribuição independente). Portanto, a representação gráfica dos valores das probabilidades e dos respectivos passos da cadeia (*trace plot*) é uma importante ferramenta para monitorar o desempenho da MCMC (Figura 17-5).

Devido ao aumento brusco de probabilidade das filogenias que são visitadas pelo andamento da cadeia, os gráficos necessariamente incluirão os valores medidos em escala logarítmica ($\ln L$, Figura 17-5). Em estatística Bayesiana, é comum que seja atribuído um intervalo de credibilidade de 95% para os parâmetros amostrados. Estes valores são obtidos através da eliminação de 2,5% dos valores mais baixos e de 2,5% dos valores mais altos para um determinado parâmetro. Um intervalo de credibilidade contém o valor correto com 95% de probabilidade; no entanto, não se trata de um intervalo de confiança.

Adicionalmente, outros métodos são úteis para diagnosticar a convergência da cadeia, tais como o exame do tamanho amostral efetivo (ESS) e a comparação de amostras resultantes de diferentes cadeias (várias cadeias de MCMC são aplicadas para o mesmo conjunto de dados). Apesar de ser computacionalmente intensiva, a última alternativa parece ser a mais confiável para verificar a convergência. Contudo, o exame de ESS é, ainda hoje, o método mais utilizado. O tamanho amostral efetivo é uma estimativa para verificar o número de amostras independentes existentes na cadeia, ou seja, quantas amostras não similares foram tomadas. Atualmente, um ESS maior que 200 é um indicativo de que a cadeia convergiu adequadamente.

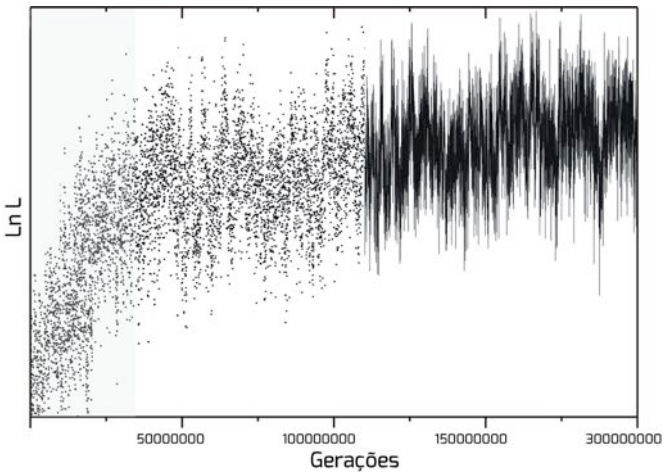


Figura 17-5: Representação gráfica das probabilidades das filogenias na cadeia ao longo de 300 milhões de amostragens. O esquema demonstra duas visualizações possíveis: à esquerda, são mostrados apenas os pontos referentes às amostras tomadas ao longo da cadeia e, à direita, as amostragens sucessivas são ligadas umas as outras para facilitar a visualização do comportamento da cadeia. Em cinza, a fase inicial de *burn in* da Cadeia de Markov Monte Carlo.

A técnica de *Metropolis Coupling*, conhecida como MCMCMC ou $(MC)^3$, através da introdução da corrida simultânea de duas cadeias, pode ajudar na amostragem de máximos globais e beneficiar na convergência da cadeia. Nesta técnica uma cadeia, chamada de quente (*hot chain*), permite aproximar os valores de máxima e mínima probabilidade das amostras para que a cadeia possa, de forma mais rápida, "saltar" entre picos de probabilidade, especialmente de máximos locais para máximos globais. O aquecimento da cadeia é dado pelo parâmetro β e visa diminuir a altura dos picos locais no espaço amostral. Uma segunda cadeia simultânea, chamada de fria (*cold chain*), utiliza as informações destes saltos da cadeia quente para melhorar a sua

amostragem e garantir a convergência.

Os métodos Bayesianos de inferência filogenética ainda têm a vantagem de aplicar modelos que envolvem diferentes tipos de relógios moleculares.

As distâncias genéticas, depois de "tratadas" pelos modelos de substituição, não tem qualquer significado sozinhas quando se deseja estimar, por exemplo, a idade do ancestral comum mais recente de duas OTUs. Esta e outras questões podem ser avaliadas quando aplicamos uma medida de tempo nas inferências, a fim de calibrar as taxas evolutivas. Sequenciamentos de amostras isoladas em diferentes épocas podem fornecer a calibração adequada para inferências temporais, pois se assume uma taxa evolutiva constante ao longo de um tempo t para todos os ramos de uma filogenia (relógio molecular estrito).

As taxas evolutivas dependem de diversos fatores e podem variar, nem sempre seguindo a constância proposta por este modelo. Após a introdução de um tipo específico de relógio molecular relaxado, as taxas de evolução podem variar ao longo da árvore para diferentes grupos e não são correlacionadas, ou seja, grupos evolutivamente próximos não necessariamente terão taxas de evolução semelhantes (relógio molecular relaxado não correlacionado).

Complexos modelos de dinâmica populacional podem ser analisados sob uma perspectiva Bayesiana. Quando o conjunto de sequências submetido às analises são isolados de uma população homogênea, os parâmetros de história demográfica podem ser usados para modelar as mudanças populacionais ao longo do tempo. Desta forma, através da estatística Bayesiana é possível, além da inferência filogenética, refinar as análises e datar filogenias e ramos específicos (Figura 18-5), inferir caracteres ancestrais e analisar a dinâmica populacional sob uma ótica evolutiva.

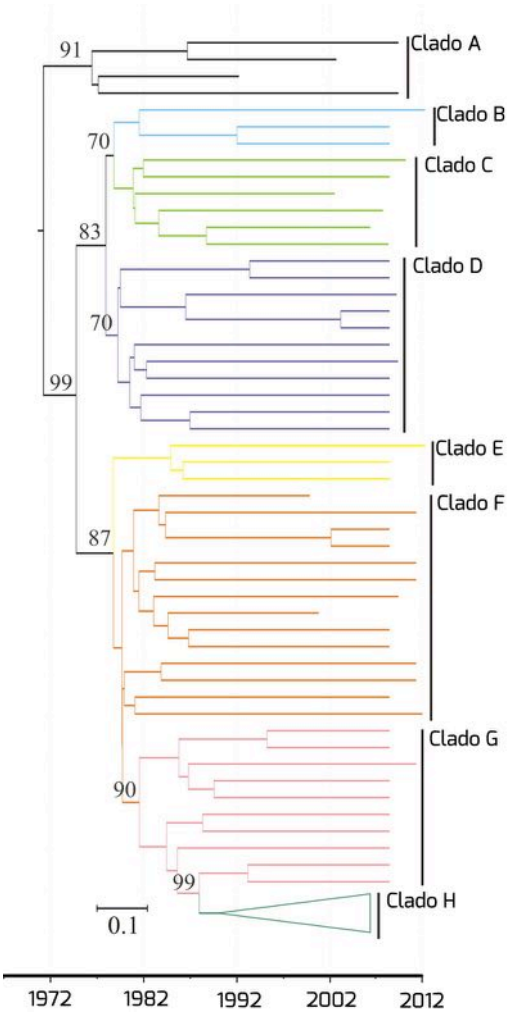


Figura 18-5: Árvore filogenética consenso gerada por inferência Bayesiana para 70 sequências de nucleotídeos. As cores nos ramos representam diferentes clados (B-H). O grupo externo está identificado como clado A. O Clado H foi agrupado para facilitar a representação. Nos nós estão especificados os valores de probabilidade posterior acima de 70. Abaixo, é apresentada a escala temporal inferida a partir da utilização de um relógio molecular relaxado.

5.8. Confiabilidade

O papel principal das técnicas de inferência filogenética é desvendar as relações evolutivas reais através de dados moleculares, buscando garantir que esta reconstrução seja fidedigna. Além da inferência das relações evolutivas entre os táxons, é igualmente importante que a filogenia possua precisão. Esta característica está relacionada ao número de filogenias que podem ser excluídas, a partir do conjunto total de filogenias, por não serem "verdadeiras". Quanto maior o número de filogenias excluídas neste processo, mais preciso é o método.

Em geral, na maioria dos casos de reconstrução filogenética, a falta de precisão das filogenias está relacionada ao conjunto de dados que está sendo fornecido no alinha-

mento. O gene considerado, o tamanho das sequências, o número de indivíduos e o grupo externo são atribuições fundamentais para uma reconstrução filogenética precisa e dependem, especialmente, do objetivo do estudo e da própria disponibilidade de informação.

Em muitos casos, o pesquisador é ainda dependente do número de amostras e do sucesso de coleta em campo, sobretudo, quando seu objeto de estudo se trata de uma espécie rara ou de indivíduos de difícil amostragem. No entanto, apesar de toda a informação relacionada ao conjunto de dados, a dificuldade de amostragem de indivíduos parece ser, sem dúvida, o principal problema relacionado a precisão das filogenias, pois a falta de dados de variabilidade genética compromete a inferência de história evolutiva coerente.

Como é possível saber se a amostragem foi suficiente e a filogenia é confiável? Usualmente, a resposta para esta questão consiste na reamostragem de dados. Se novas amostras forem tomadas e a mesma filogenia for reproduzida, a filogenia proposta tem seu valor reforçado. No entanto, na maioria dos casos, a reamostragem de dados da forma usual (coletas de novos espécimes, reamostragens em campo, achado fóssil diferente, etc) não é factível. Assim, algoritmos que produzem diferentes amostragens utilizando o mesmo conjunto de dados foram desenvolvidos para possibilitar a verificação da confiabilidade nos clados das filogenias. Destaca-se entre estes algoritmos o método de *bootstrap*.

Bootstrap é um método de reamostragem utilizado para realizar comparações da variabilidade das hipóteses filogenéticas, oferecendo medidas de confiabilidade aos clados propostos. A reamostragem é realizada a partir do mesmo conjunto de dados, e novas amostras fictícias com o mesmo tamanho serão geradas.

Segundo este método, cada sítio do alinhamento será tratado de forma independente. Conforme a Figura 19-5, inicialmente o algoritmo reconstruirá a filogenia a partir do alinhamento dado e, posteriormente, diversas replicatas serão reconstruídas. As colunas, representando os sítios do alinhamento, serão aleatoriamente tomadas (amostradas) pelo algoritmo e, em seguida, serão agrupadas uma ao lado da outra de maneira a formar um novo alinhamento (com o mesmo número de sítios do alinhamento original, Figura 19-5).

Por este método, é possível que um mesmo sítio seja amostrado mais de uma vez e, portanto, alguns sítios não serão selecionados para o novo alinhamento. Um número fornecido pelo usuário especificará o número de pseudoreplicatas (novos alinhamentos) que serão construídas. Assim que uma pseudoreplicata for criada, o algoritmo constrói a filogenia correspondente.

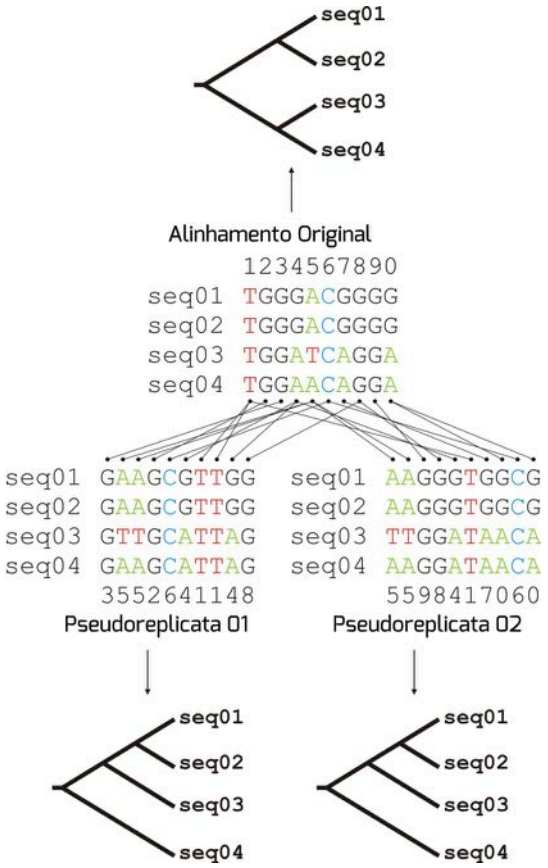


Figura 19-5: Método de *boostrap* para filogenias. A partir do alinhamento original, as colunas que representam os sítios serão aleatoriamente amostradas para construir pseudoreplicatas (um mesmo sítio pode ser sorteado diversas vezes). Estas, por sua vez, serão utilizadas para a inferência de filogenias, da mesma forma que o alinhamento original.

É importante ressaltar que a inferência destas filogenias será realizada pelo método de construção especificado pelo usuário, seja aproximação de vizinhos, máxima parcimônia ou máxima verossimilhança (para árvores bayesianas, veja adiante). Ao final, o algoritmo analisará os clados e automaticamente verificará a presença de determinados agrupamentos em todas as filogenias construídas. Se, por exemplo, encontramos as sequências 1 e 2 formando um clado em 70% das filogenias construídas, atribuiremos a confiabilidade de 70 ao clado formado por estas duas sequências. Comumente, o valor de confiabilidade dos clados é colocado próximo ao ancestral comum do clado (Figura 18-5).

A partir dos resultados de confiabilidade dos clados é possível também construir filogenias baseando-se na árvore consenso gerada pela regra da maioria (*majority-rule consensus tree*). Neste método, o algoritmo tabulará todos os clados formados em todas as replicatas geradas. Aqueles clados que mais aparecerem servirão para montar a filogenia consenso.

Ao contrário dos métodos de aproximação de vizinhos, máxima parcimônia e máxima verossimilhança, a confiabilidade de filogenias construídas através de estatística Bayesiana é inerente ao processo. Como diversas filogenias são amostradas ao longo do desempenho da Cadeia de Markov, não é necessário nenhum método para simular reamostragens do mesmo conjunto de dados. As amostras serão resumidas a partir da distribuição posterior de filogenias como frequência de clados individuais e serão identificadas por um número próximo ao ancestral comum daqueles clados (Figura 18-5). Portanto, o valor de probabilidade posterior de um clado representa uma inferência a respeito da probabilidade daquele clado.

A comparação dos valores de *bootstrap* e de probabilidade posterior dos clados para filogenias construídas a partir do mesmo alinhamento utilizando máxima verossimilhança e o método Bayesiano, respectivamente, leva a conclusão de que o método Bayesiano superestima a confiança aos clados. A confiança atribuída pela probabilidade posterior é geralmente maior que aquela atribuída pelo método de *bootstrap*. Por isso, enquanto uma confiança acima de 70 é considerada sustentada para o *bootstrap*, apenas valores acima de 90 podem ser considerados relevantes para os métodos Bayesianos.

5.9. Interpretação de filogenias

Árvores filogenéticas são diagramas que denotam a história evolutiva de diferentes OTUs a partir de seu ancestral comum. Mais do que isso, as filogenias moleculares são ferramentas que ajudam no entendimento dos diversos processos evolutivos que moldam o genoma dos organismos. Desta forma, a interpretação das implicações evolutivas associadas a um, ou a um conjunto de táxons, está diretamente relacionada à disposição dos ramos internos e externos de uma árvore. Independentemente do método de inferência, ou da forma como a árvore é apresentada, a interpretação dos resultados será baseada nos mesmos pressupostos, ainda que métodos diferentes possam originar filogenias diferentes.

Inicialmente, é necessário observar a presença de uma raiz. Como já discutido, o método de enraizamento pelo grupo externo é o mais comum e utiliza organismos sabidamente relacionados ao grupo em evidência, servindo para orientar o algoritmo em relação às características mais ancestrais do grupo. O grupo externo ajudará a evidenciar o tempo evolutivo. Na Figura 20-5, por exemplo, o grupo externo é dado pelo orangotango, pois este compartilha o mesmo ancestral comum que o restante do grupo. No caso de filogenias sem raiz, é necessário ter cautela nas interpretações, pois este tipo de diagrama apenas revela a relação entre os táxons.

Depois de encontrada a raiz da filogenia, é preciso avaliar os ramos. Dependendo do método, os ramos podem ter significados diferentes. Na Figura 18-5, os ramos evidenciam o tempo real, apresentando OTUs amostradas no passado. Pelo contrário, na Figura 20-5, os ramos evidenciam apenas um

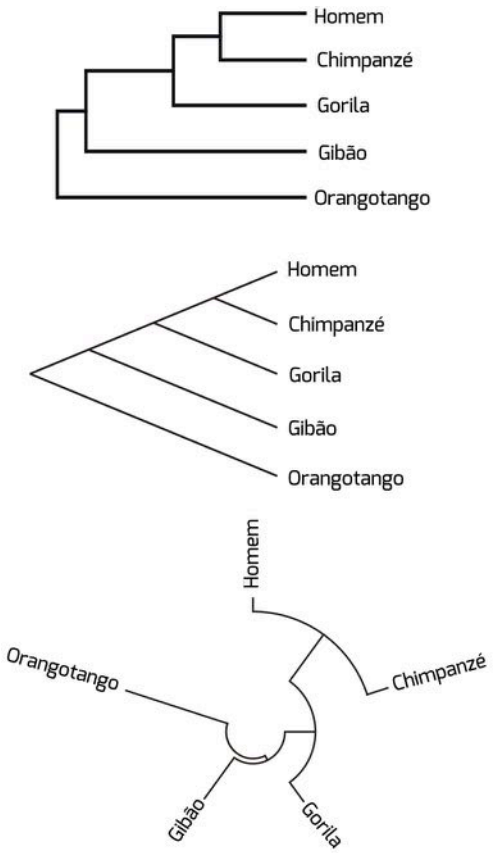


Figura 20-5: Diferentes representações da filogenia dos primatas.

tempo evolutivo representado pelo número de modificações genômicas, desde o organismo ancestral até os ramos terminais. Além disso, deve-se perceber a escala na qual os ramos foram representados, pois estes indicam o número de substituições que provavelmente ocorreram ao longo do processo evolutivo e podem ajudar na interpretação das taxas evolutivas.

Conclusões evolutivas baseadas em árvores filogenéticas devem ser sustentadas em árvore confiáveis e, por isso, a medida de confiabilidade dos ramos deve ser denotada. Inicialmente, é necessário verificar o método utilizado para reconstrução da filogenia e, quando necessário, verificar o algoritmo utilizado para gerar a confiabilidade dos clados. Ramos com maiores valores de confiabilidade gerarão conclusões mais confiáveis, enquanto que clados com baixos valores deverão ser interpretados com maior cuidado. No entanto, não é necessário negar totalmente conclusões baseadas em filogenias com baixa confiabilidade nos ramos. O tipo de método, a forma de amostragem e o número de OTUs podem ser fatores de interferência e, assim, podem prejudicar a valorização dos ramos.

O padrão de organização dos ramos de uma filogenia denota o padrão de ancestralidade. As filogenias não são escadas, onde alguns organismos são "mais evoluídos" que outros, mas uma representação da história da derivação de OTUs. Na Figura 18-5, por exemplo, é possível observar que os clados B, C, D, E, F e G possuem um ancestral comum que compartilha um outro ancestral com o clado A. Já o clado H, representado por um triângulo para evidenciar um grande número de táxons naquele ponto da filogenia, teve um ancestral comum dentro do clado G. Este padrão sugere que o clado H se originou a partir do clado G. Da mesma forma, podemos observar a disposição do clado G em relação ao F e concluir que o primeiro se originou a partir do segundo.

No caso da Figura 20-5, observamos que humanos e chimpanzés tiveram um mesmo ancestral comum. Com base nestes dados, é incorreto pensarmos que humanos são derivados de chimpanzés, ou que humanos são mais evoluídos que chimpanzés. Estes organismos estão apenas formando um mesmo clado dentro da filogenia dos primatas.

Por último, é fundamental saber o objetivo do estudo filogenético a ser realizado. Árvores filogenéticas devem ser construídas para responder uma determinada questão, que pode envolver apenas um, ou diversos organismos.

Quando possível, é importante reconstruir a filogenia utilizando diferentes métodos de inferência e compará-las entre si. A conclusão desta forma será melhor sustentada. Além disso, atualmente, a história retratada em uma filogenia não é por si só satisfatória. Outras ferramentas podem ser utilizadas para complementar e sustentar a interpretação de uma filogenia, incluindo análises de recombinação, pressão seletiva e estruturação populacional, verificação de coespeciação, construção de redes filogeográficas, compa-

ração com dados de fósseis, eventos geológicos, dados históricos e, até mesmo, análises de dados comportamentais.

Um exemplo da combinação de análises filogenéticas com dados históricos veio na confirmação da origem e disseminação humana a partir da África. Através da utilização de dados histórico-antropológicos (como vestígios materiais de hominídeos ancestrais), fósseis de hominídeos e análises de DNA mitocondrial de representantes de diferentes etnias, os pesquisadores puderam traçar as rotas de disseminação humana a partir da África.

Outro exemplo está na solução de um enigma que perturbou zoólogos por um longo período: a posição taxônomica do panda-gigante entre os mamíferos carnívoros. Apesar de esta espécie ser fisicamente muito similar a um urso, outras características, como dentição e anatomia das patas, levaram à proposição de uma hipótese antes não imaginada.

Tal hipótese propunha que o panda-gigante (*Ailuropoda melanoleuca*) seria proximamente relacionado ao o panda-vermelho (*Ailurus fulgens*), um mamífero de pequeno porte, semelhante ao guaxinim. Com o emprego de diferentes dados, incluindo fósseis, anatomia de mamíferos atuais, distribuição geográfica, sequências de DNA de diferentes porções do genoma, sequências de aminoácidos de diferentes proteínas e mapeamento cromossômico, foi possível estabelecer uma história evolutiva plausível, capaz de descrever a origem evolutiva do panda-gigante (Figura 21-5).

Por meio dessa análise combinada de dados, se propôs que o panda-gigante, um urso, derivou do ancestral comum dos ursos há cerca de 24 milhões de anos, muito antes das derivações que originaram todos os outros ursos existentes hoje. Além disso, observou-se que os ursos e os procionídeos (grupo que inclui o guaxinim e o panda-vermelho) possuem um ancestral comum que deu origem às duas linhagens há aproximadamente 30 milhões de anos.

A filogenia molecular é uma ferramenta útil quando empregada isoladamente, mas que pode se beneficiar de diferentes tipos de dados para propor uma história evolutiva. Em última análise, a decisão sobre que tipos de

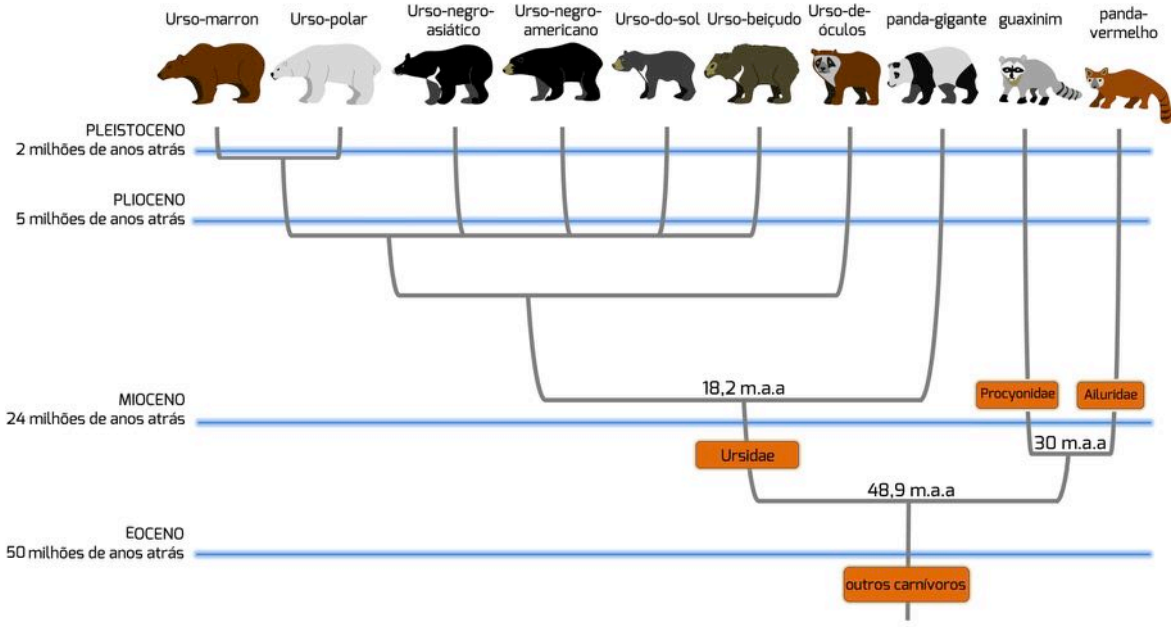


Figura 21-5: Posição filogenética do panda-gigante, baseada na combinação de diferentes tipos de dados. Baseado em BININDA-EMONDS, Olaf R.P. *Phylogenetic position of the giant panda*. Em: LINDBURG, D.G. & Baragona, K. *Giant pandas: Biology and conservation*. Berkeley: University of California Press, 2004; e em EIZIRIK, Eduardo e colaboradores: *Pattern and timing of diversification of the mammalian order Carnivora inferred from multiple nuclear gene sequences*. *Mol Phylogenet Evol*, 56, 49, 2010.

dados (além dos moleculares) serão empregados na análise filogenética dependerá da pergunta a ser respondida com essa técnica. Não existem regras pré-estabelecidas, e as estratégias analíticas precisam ser propostas caso a caso.

5.10. Conceitos-chave

Ancestral: organismo ou sequência que originou novo(s) organismo(s) ou sequência(s). Em alguns casos pode ser considerado o mesmo que primitivo.

Apomórfico: refere-se a um caractere novo adquirido ao longo do processo evolutivo, uma inovação. Uma apomorfia pode servir de diagnóstico para separação de clados.

Aproximação dos vizinhos: *neighbor joining* (NJ), método de inferência filogenética quantitativo baseado em distância genética.

Autapomorfias: apomorfias específicas e restritas a um clado.

Bootstrap: método de reamostragem que permite verificar a confiabilidade dos ramos de uma filogenia.

Cadeias de Markov Monte Carlo: método utilizado pela estatística Bayesiana para amostrar as probabilidades de distribuição de diferentes parâmetros das filogenias.

Clado: grupo formado por um ancestral e todos seus descendentes, um ramo único em uma árvore filogenética.

Derivado: que se originou de um ancestral e é mais recente no tempo evolutivo (nota: deve-se evitar o termo "mais evoluído" e, em seu lugar, empregar "derivado").

Distância Genética: medida quantitativa da divergência genética entre organismos.

Espaço Amostral de Filogenias: espaço teórico que inclui todas as filogenias possíveis (com raiz ou sem raiz) para um determinado alinhamento.

Frequência de equilíbrio: ponto em que não existe mais alteração nas frequências dos alelos.

Grupos irmãos: clados que dividem um ancestral comum.

Homologia: similaridade originada por ancestralidade comum.

Inferência filogenética Bayesiana: método qualitativo de inferência filogenética baseado na estatística Bayesiana. Através da Cadeia de Markov Monte Carlo este método buscará as árvores mais prováveis dentro das filogenias amostradas.

Máxima Parcimônia: método qualitativo de inferência filogenética que busca a árvore que minimiza o número total de substituição de nucleotídeos.

Máxima Verossimilhança: método qualitativo de inferência filogenética que busca a árvore com a máxima verossimilhança.

Monofilia: associação entre o ancestral comum e todos os seus descendentes, formando um clado monofilético.

Múltiplas Substituições: eventos múltiplos de substituição de nucleotídeo localizado em um mesmo sítio do DNA.

Modelos de Substituição: modelos matemáticos utilizados para descrever o processo evolutivo ao longo do tempo, podendo ser aplicados ao alinhamento de nucleotídeos ou aminoácidos.

Ortólogo: genes homólogos em diferentes organismos e que mantém a mesma função.

OTU: unidade taxonômica operacional, folha ou nó terminal em uma árvore filogenética.

Parafilia: associação entre o ancestral comum e apenas parte de seus descendentes, formando um clado parafilético.

Parálogo: genes homólogos de um mesmo organismo que divergiram após duplicação.

Plesiomórfico: dotado de características do ancestral que são conservadas nos descendentes.

Polifilia: associação entre diferentes OTUs sem a necessidade de um único ancestral comum, frequentemente originada por convergência evolutiva.

Primitivo: diz-se de características ou organismos ancestrais, anteriores no tempo evolutivo a organismos ou características mais recentes.

Probabilidades Anteriores: distribuição dos valores de um parâmetro filogenético que é sabido de antemão pelo pesquisador.

Probabilidades Posteriores: conjunto da distribuição dos valores de parâmetros filogenéticos resultantes do método de inferência Bayesiana.

Sistemática: estudo da diversificação das formas vivas e suas relações ao longo do tempo.

Taxonomia: estudo que busca agrupar os organismos com base em suas características e nomear os grupos obtidos, classificando-os em alguma escala.

Taxon: grupo (de qualquer nível hierárquico) proposto pela taxonomia.

Topologia: descreve a ordem e a disposição exata das OTUs em uma filogenia.

UPGMA: *unweighted pair-group method using arithmetic average*, método de inferência filogenética quantitativo baseado em distância.

5.11. Leitura recomendada

FELSENSTEIN, Joseph. ***Inferring Phylogenies***. Sunderland: Sinauer, 2004.

GREGORY, T. Ryan: ***Understanding Evolutionary Trees***. Evo. Edu. Outreach, 2008, 1,121-137.

LEMEY, Philippe; SALEMI, Marco; Vandamme, Anne-Mieke (Org.). ***The Phylogenetic Handbook***. 2.ed. Cambridge: Cambridge University Press, 2009.

MATIOLI, Sergio Russo; FERNANDES, Flora M.C. (Org.). ***Biologia Molecular e Evolução***. 2.ed. Ribeirão Preto: Holos, 2012.

NEI, Masatoshi; KUMAR, Sudhir. ***Molecular Evolution and Phylogenetics***. Nova Iorque: Oxford University Press, 2000.

PABÓN-MORA, Natalia; GONZÁLEZ, Favio. A classificação biológica: de espécies a genes. In: ABRANTES, Paulo C. (Org.), ***Filosofia da Biologia***. Porto Alegre: Artmed, 2011.

SCHNEIDER, Horacio. ***Métodos de Análise Filogenética: Um Guia Prático***. 3.ed. Ribeirão Preto: Holos, 2007.