Evidências da evolução: filogenia molecular

AULA 23

Meta da aula

Apresentar e desenvolver os conceitos e métodos utilizados na filogenia molecular.

objetivos

Esperamos que, após o estudo do conteúdo desta aula, você seja capaz de:

- Definir filogenia e sistemática molecular.
- Listar métodos para inferir relações de ancestralidade entre um grupo de seqüências alinhadas.

Pré-requisitos

Para acompanhar esta aula, é importante que você reveja os conceitos de Filogenia, Taxonomia e Biogeografia (Aulas 3, 4 e 17 da disciplina Diversidade dos Seres Vivos); estude novamente o impacto da Sistemática Filogenética (Aula 13 da disciplina Grandes Temas em Biologia); e, principalmente, releia as Aulas de 2 a 11 da disciplina Introdução à Zoologia.

INTRODUÇÃO

Nesta aula, vamos falar sobre as interações entre as disciplinas Evolução, Biologia Molecular, Genética de Populações e Filogenia Molecular. Você vai aprender a utilizar seqüências de nucleotídeos ou aminoácidos como caracteres para estimar relações entre organismos, ou em outras palavras, para construir filogenias.

A LIGAÇÃO ENTRE SISTEMÁTICA MOLECULAR, FILOGENIA MOLECULAR E EVOLUÇÃO

Sistemática Molecular

É a detecção, descrição e explicação da diversidade biológica em nível molecular; ela analisa tanto a variação que ocorre entre as espécies quanto a que ocorre dentro das espécies.

A **Sistemática Molecular** é uma disciplina que utiliza **marcadores genéticos** para inferir processos e **filogenias** populacionais.

Marcador genético ou marcador molecular

É um DNA polimórfico ou a seqüência de uma proteína derivada de uma única localização cromossômica (loco), usado no mapeamento genético e na identificação individual ou de determinado táxon.

Filogenia

É a história evolutiva de populações de organismos relacionados.

!

Volte à Aula 8 desta disciplina e reveja com detalhes o papel dos marcadores moleculares no estudo da evolução!

O estudo da Evolução Molecular envolve duas grandes áreas:

- Área 1 = Evolução das macromoléculas. Investiga as taxas e padrões de mudança do material genético (seqüências de DNA) e dos produtos por ele codificados (proteínas) no tempo evolutivo, além dos mecanismos responsáveis por essas mudanças;
- Área 2 = Reconstrução da história evolutiva dos genes e organismos, Filogenia Molecular ou Filogenética Molecular. Investiga a história evolutiva dos organismos e das macromoléculas, segundo inferência a partir de dados moleculares.

As áreas 1 e 2 são fortemente relacionadas: o conhecimento filogenético é essencial para a determinação da ordem das mudanças nos caracteres moleculares estudados, o que caracteriza, geralmente, o primeiro passo na inferência causal da mudança; o conhecimento acerca do padrão e da taxa de mudança de uma dada molécula é crucial para as tentativas de reconstrução da história evolutiva de um grupo de organismos.

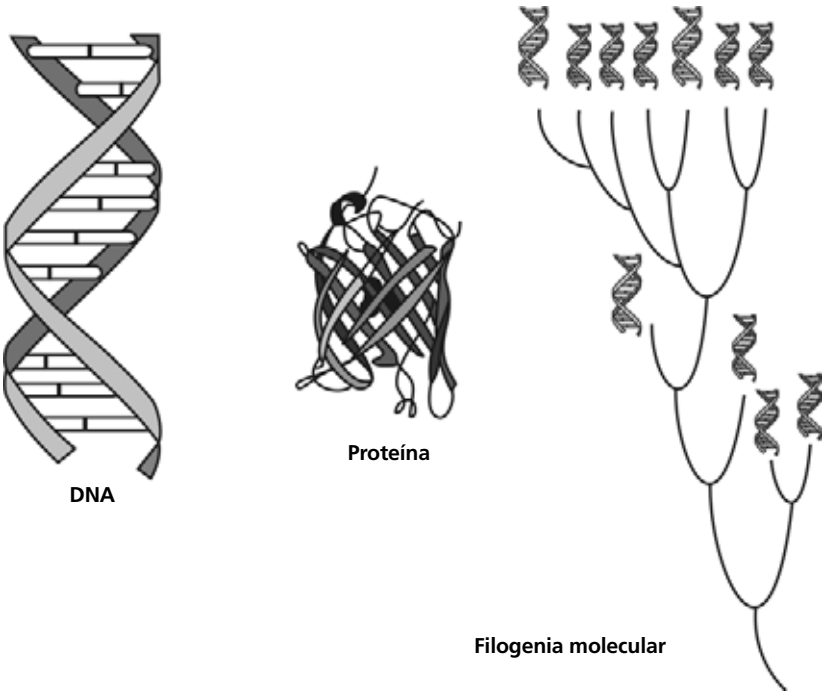


Figura 23.1: Moléculas utilizadas na construção de filogenias moleculares e exemplo de uma árvore filogenética.

!

Árvore filogenética. A representação gráfica de reconstrução filogenética geralmente é constituída em forma de "árvore", com uma topologia específica, seja ela enraizada ou não. Você já viu o conceito de árvore filogenética em diversos momentos do seu curso de Biologia! A primeira vez foi na Aula 11 de Grandes Temas em Biologia. Volte à Aula 2 desta disciplina e reveja, também, a Aula 17 de Diversidade dos Seres Vivos. As características gerais de uma árvore filogenética serão apresentadas a seguir.

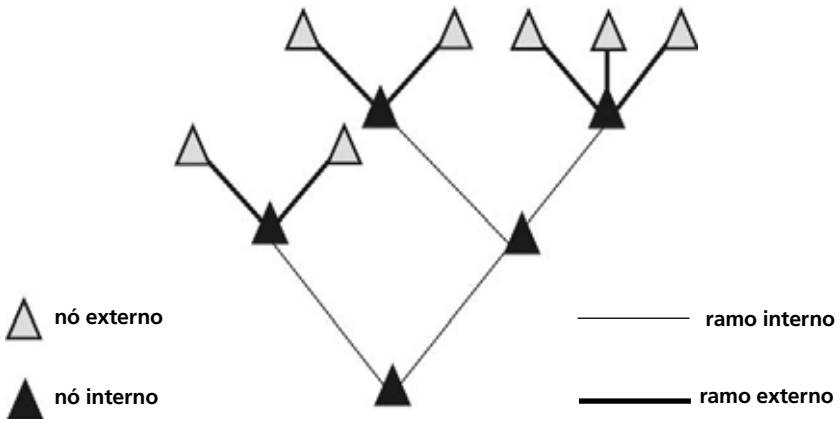


Figura 23.2: Componentes de uma árvore filogenética.

A disciplina Evolução Molecular soma os dados empíricos gerados pelas técnicas de Biologia Molecular com a fundamentação teórica da disciplina Genética de Populações.

A análise das moléculas, principalmente DNA e proteínas, cria um substancial banco de dados comparativos que constitui a matéria-prima para os estudos da disciplina Evolução Molecular. Esta matéria é responsável por avaliar taxas, processos e limitações da mudança molecular através do tempo.

A Sistemática Molecular inclui ambas as variações intraespecíficas, que constitui o campo de trabalho da Genética de Populações; e a diversidade interespecífica, que é, tradicionalmente, o campo de trabalho da Filogenética.

O uso de genealogias alélicas, baseado na taxa de substituição de nucleotídeos, torna possível predizer os efeitos da deriva genética, mutação, migração e seleção, em tempos predeterminados, como na ancestralidade comum de determinados alelos.

ATIVIDADE 1

Em que se diferencia a Filogenética da Filogenética Molecular?

_____
_____
_____
_____
_____
_____
_____
_____
_____
_____

RESPOSTA COMENTADA

Ambas as disciplinas utilizam caracteres para inferir relações de ancestralidade entre táxons. A Filogenética Molecular difere apenas quanto ao tipo de caracteres que utiliza, ou seja, ela dispõe de dados gerados pela análise de moléculas de proteínas e ácidos nucléicos. A Filogenética clássica utiliza caracteres morfológicos, ecológicos, embriológicos etc.

Caráter e caracteres (plural)

Ao longo desta aula, vamos utilizar este termo para os traços distintivos gerados no estudo de filogenias. Na verdade, essa palavra é traduzida do inglês *character* e é bastante utilizada na literatura especializada, em português.

HISTÓRICO E EVOLUÇÃO DA SISTEMÁTICA MOLECULAR E DA BIOLOGIA MOLECULAR

No século XVIII, Carolus Linnaeus ou Carl Linné (veja a Aula 9, Introdução à Zoologia) estabeleceu um critério para a descrição e categorização da diversidade biológica. Esse sistema hierárquico era inicialmente independente da Teoria Evolutiva; alguns dos primeiros evolucionistas, como George-Louis Leclerc, o Comte de Buffon (veja a Aula 13, Introdução à Zoologia), se opunham ao Sistema Lineano e ao essencialismo aristoteliano nele embutido. No entanto, o Sistema de Linnaeus prevaleceu e os evolucionistas posteriores, como Jean-Baptiste Lamarck, Charles Darwin (veja a Aula 3, Evolução) e Ernst Heirich Haeckel (veja a Aula 7, Introdução à Zoologia), simplesmente adaptaram o sistema para produzir uma classificação baseada nas relações filogenéticas.

Os primeiros esforços para reconstruir a história filogenética eram baseados em poucos critérios objetivos, e as estimativas de filogenia eram pouco mais que suposições plausíveis, geradas por peritos em grupos taxonômicos particulares. Durante a maior parte da primeira metade do século XX, os sistematas estavam mais envolvidos com problemas de espécies, especiação e variação geográfica do que com problemas de filogenia.

Essa situação começou a mudar durante as décadas de 1930, 1940 e 1950, por meio do esforço de pesquisadores, como o botânico Walter Zimmermann e o zoólogo Willi Hennig (veja a Aula 2, Introdução à Zoologia). Eles começaram a definir métodos objetivos para a reconstrução da história evolutiva, com base em caracteres compartilhados por organismos vivos e fósseis.

Na década de 1960, esses métodos foram refinados e transformados em critérios explícitos para a estimativa de filogenias. Vários algoritmos baseados nesses critérios foram implementados em programas de computador, o que permitiu a análise de um grande e complexo conjunto de dados. Os últimos 30 anos continuaram a representar avanços conceituais e operacionais na estimativa de filogenias, assim como na análise de mudanças microevolutivas; agora, os estudos de filogenia não mais se limitam a aplicações na classificação biológica. Na verdade, estudos de filogenia permearam quase todas as subdisciplinas da Biologia, e biólogos comparativos de todos os tipos reconhecem a importância de métodos filogenéticos na interpretação de padrões e processos biológicos.

O século XX teve duas grandes eras em relação às moléculas: a Era das Proteínas e a Era do DNA. A Era das Proteínas teve seu clímax na década de 1960, enquanto a Era do DNA viveu o apogeu nas décadas de 1980 e 1990. Mais recentemente, temos vivido, na Biologia Molecular, as Eras da Genômica e da Proteômica.

Na década de 1950, os estudos evolutivos incorporaram os métodos do seqüenciamento de proteínas, análise de padrões de fragmentos trípticos, eletroforese em gel de amido e técnicas imunológicas mais apuradas. Poucos anos depois, Frederick Sanger e colaboradores (1953) determinaram a primeira seqüência completa da proteína insulina.

!

Eletroforese em gel de amido

Desde a origem do gel de eletroforese de amido, da visualização histoquímica das enzimas nos géis e dos estudos clássicos de H. Harris, J. L. Hubby e R. C. Lewontin (veja Aula 8, Evolução), uma importante revolução ocorreu no entendimento de processos micro e macroevolutivos. A eletroforese de proteínas – migração de proteínas sob influência de um campo elétrico – é um dos métodos mais baratos e eficazes na investigação de fenômenos genéticos no nível molecular.

Várias investigações de variabilidade genética em populações naturais, fluxo gênico, hibridização entre espécies, reconhecimento de limites entre espécies e relações filogenéticas utilizaram e utilizam proteínas e enzimas. A principal suposição que os biologistas evolutivos fazem no uso de dados de isozimas é a de que mudanças na mobilidade das enzimas sob um campo elétrico refletem alterações na seqüência de DNA que as codifica. Assim, se o padrão de bandas de dois indivíduos é diferente, supõem-se que essas diferenças possuem base genética e são herdáveis.

Apesar de consideravelmente menos precisa que o seqüenciamento de proteínas, a eletroforese dessas macromoléculas consome muito menos tempo, e foi amplamente utilizada no estudo de relações filogenéticas entre populações ou espécies relativamente próximas evolutivamente. O uso da eletroforese desencadeou o desenvolvimento de medidas de distância genética, e o Índice de Nei (NEI, 1972) facilitou muito o estudo das relações evolutivas entre populações ou espécies próximas evolutivamente. Adicionalmente, foram também extensamente utilizadas técnicas de imunossistemática, tais como a fixação de microcomplementos e de hibridização de DNA.

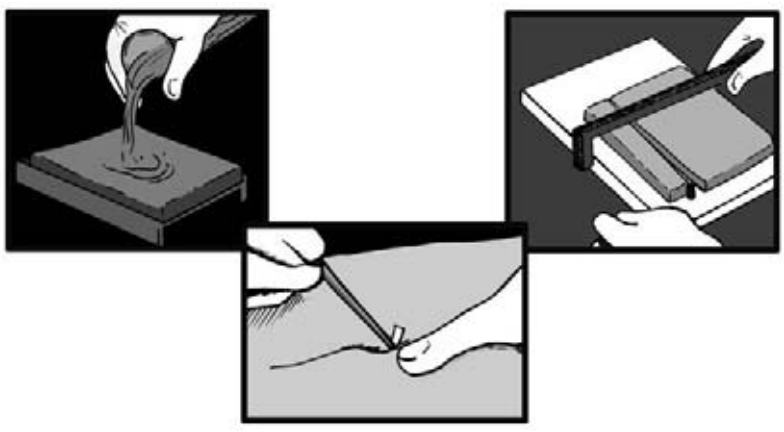


Figura 23.3: Eletroforese em gel de amido: preparo do gel; aplicação, em um corte na origem do gel, de pequenos recortes de papel-filtro embebidos nas amostras; após a corrida, corte do gel em fatias para ensaio enzimático.

Já se sabia, em 1956, que as substituições de aminoácidos ocorriam de maneira não aleatória entre as diferentes partes de uma proteína (comparando-se a insulina de boi, ovelha, porco, cavalo e baleia, constatava-se que as mudanças só ocorriam nas posições de 8 a 10 da cadeia alfa); sabia-se também que a maioria das substituições de aminoácidos das mesmas proteínas, embora de espécies diferentes, parecia não ter efeito notório em sua atividade biológica. Por outro lado, pequeno número de substituições de aminoácidos podia causar considerável diferença na atividade biológica de proteínas diferentes, porém relacionadas (por exemplo: a vasopressina e a oxitocina do boi diferem em apenas dois aminoácidos).

Nas décadas de 1960 e 1970, o acúmulo de seqüências de proteínas (na época, mais fáceis de analisar do que os ácidos nucléicos) forneceu, pela primeira vez, dados adequados para o estudo de evolução, especialmente o das relações evolutivas entre ordens, classes, filos e reinos. Esses dados estimularam a construção de árvores filogenéticas e o desenvolvimento de diversos métodos para a construção dessas árvores.

A árvore construída a partir de seqüências de uma única proteína, o citocromo c, era similar à árvore conhecida, baseada em caracteres não- moleculares (morfológicos, ecológicos, reprodutivos, ontogenéticos etc.) para diversas espécies de vertebrados e invertebrados; isso revelou o potencial da filogenética molecular. Tal acúmulo de dados gerou também grande interesse na metodologia de alinhamento de seqüências.

E. Zuckerkandl e L. Pauling propuseram, em 1965, a **Teoria do Relógio molecular** com base em dados conhecidos para hemoglobinas e citocromo c – a taxa de substituição de aminoácidos nessas proteínas era, aproximadamente, a mesma dentre diversas linhagens de mamíferos. Essa teoria suscitou grande interesse no uso de macromoléculas em estudos evolutivos (se proteínas evoluem a taxas constantes, elas podem ser utilizadas para a determinação do tempo de divergência entre espécies e para a reconstrução das relações filogenéticas entre organismos). O Relógio Molecular gerou muita controvérsia, pois nos níveis morfológico e fisiológico as taxas evolutivas pareciam ser bem mais erráticas (sem rumo).

O advento de várias técnicas para estudos de DNA, a partir de 1970, tais como análise de restrição, clonagem de genes, Reação em Cadeia da Polimerase (PCR) e técnicas de seqüenciamento, acarretou uma explosão de conhecimento em Biologia Molecular e o estabelecimento de uma nova era no estudo da Evolução Molecular.

Teoria do Relógio molecular

Decorre da regularidade, como em um relógio, da mudança ocorrida em uma molécula ou em um genótipo através do tempo geológico. É a teoria de que as moléculas evoluem em proporção direta ao tempo, de forma que diferenças entre seqüências homólogas de DNA ou proteínas podem ser usadas para estimar o tempo decorrido, desde a última vez em que as duas moléculas (ou os táxons que as contêm) possuíram um ancestral comum.

Técnicas de DNA Recombinante

No início da década de 1970, uma nova maneira de explorar as principais moléculas constituintes de uma célula começou a ser posta em prática. Essas metodologias inovadoras foram coletivamente chamadas "Tecnologia do DNA Recombinante", "Técnicas de Clonagem Molecular" ou de "Engenharia Genética". O DNA era considerado o componente celular mais difícil de ser isolado e analisado, devido a seu tamanho (os genes são parte de uma enorme molécula de DNA condensada no cromossomo) e constituição quimicamente monótona (quatro tipos de nucleotídeos). Graças às novas técnicas, genes específicos podem ser isolados em quantidade, redesenhados e devolvidos às células e organismos.

As **Técnicas de DNA Recombinante** e de clonagem gênica permitem que os cientistas isolem e caracterizem qualquer gene ou outra seqüência de qualquer organismo. Essas técnicas tornaram-se viáveis com a descoberta das enzimas de restrição, que reconhecem e quebram seqüências específicas no DNA. Seqüências de DNA de interesse são inseridas em pequenas moléculas de DNA auto-replicantes, chamadas vetores de clonagem. Tais moléculas recombinantes são amplificadas por meio de replicação *in vivo*, após serem introduzidas por transformação em bactérias. Bibliotecas genômicas podem ser construídas em vetores contendo um jogo completo de seqüências de DNA genômico ou cópias de DNA feitas em um organismo, a partir do RNAm (cDNA – DNA complementar ao RNA mensageiro, ou seja, sem os íntrons). Genes específicos podem ser isolados dessas bibliotecas por complementação genética e por hibridização com sondas de ácidos nucléicos, marcados radiativamente e contendo seqüências de DNA de função conhecida.

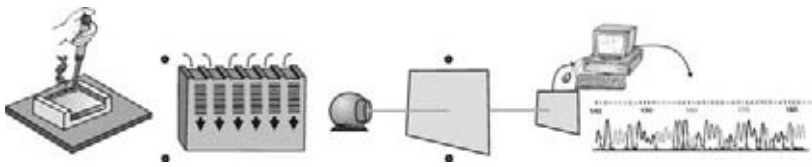


Figura 23.4: A técnica de seqüenciamento de nucleotídeos do DNA revolucionou o estudo da Evolução.

O acelerado progresso no Estudo da Evolução molecular foi grandemente facilitado pelo desenvolvimento de computadores de alta velocidade; cada vez mais, sua rapidez e baixo custo permitem o uso por número crescente de pesquisadores de métodos progressivamente sofisticados.

Junto aos avanços da biotecnologia, ocorreram melhorias na análise da variação molecular dentro de uma mesma espécie e entre espécies diferentes. A habilidade para se obter árvores gênicas dentro de uma mesma espécie encorajou o desenvolvimento da **Teoria Coalescente** (HUDSON, 1991) e da análise da Filogeografia (AVISE, 1994). Novos

Teoria Coalescente ou da Coalescência

Teoria baseada em velocidades de divergência, determinadas pelo Relógio Molecular, para inferir o tempo de separação de dois táxons relacionados desde a linhagem do seu ancestral comum; processo evolutivo que é observado voltando no tempo, de modo que a diversidade alélica é acompanhada através das mutações até os alelos ancestrais. Essa teoria pode ser utilizada para fazer previsões sobre o tamanho efetivo das populações, idades e freqüência dos alelos, seleção, velocidade de mutação ou tempo decorrido até que se identifique o ancestral comum.

métodos de análise relacionam não somente a geração de hipóteses filogenéticas, mas também o teste de hipóteses sobre Biogeografia, Ecologia, comportamento, Fisiologia, desenvolvimento, Epidemiologia e praticamente todo e qualquer aspecto da Biologia. Mais sofisticação na análise de dados evolutivos melhorou nossa habilidade de investigar as particularidades dos caracteres moleculares em relação aos modelos e processos de evolução.

Filogeografia

É o estudo da Biogeografia revelado pela comparação de filogenias de populações ou espécies com sua distribuição geográfica. É também o campo de estudo dos princípios e processos que governam a distribuição geográfica de linhagens genealógicas dentro das espécies, com ênfase em fatores históricos, integrando conhecimentos de Genética Molecular, Genética de Populações, Filogenética, Demografia e Geografia Histórica.

!

Filogeografia

O termo "filogeografia" foi criado, em 1987, por John C. Avise e colaboradores (AVISE *et al.*, 1987). Nos primeiros grandes levantamentos moleculares de linhagens de DNA mitocondrial (mtDNA ou DNAmit) em populações naturais, frases complicadas foram empregadas para resumir uma observação direta: os ramos de árvores gênicas intraespecíficas apresentam, comumente, um padrão geográfico. Em outras palavras, o componente genealógico tipifica a distribuição espacial dos genótipos dentro de organismos relacionados e entre eles.

Após a criação do termo filogeografia, várias relações entre genealogias gênicas e geografia puderam ser descritas simplesmente como padrões filogeográficos. Os estudos filogeográficos podem ser relacionados com a demografia de populações e a Teoria da Coalescência. O que nasceu como um mero termo útil tornou-se uma disciplina "adolescente" com ricas conexões com a Biologia, a Paleontologia e a Geografia Histórica. As perspectivas filogeográficas revolucionaram conceitual e empiricamente as interpretações dos processos microevolutivos na Natureza.

Simplificando, podemos dizer que a Filogeografia é a disciplina que relaciona as genealogias gênicas com a Filogenética e a Geografia.

A análise e a interpretação da distribuição das linhagens requerem usualmente importação de dados oriundos da Genética Molecular, da Genética de Populações, da Etologia, da Demografia, da Biologia Filogenética, da Paleontologia, da Geologia e da Geografia Histórica. Assim, a Filogeografia é um esforço integrativo que atua no entroncamento de diversas disciplinas nas áreas de micro e macroevolução.

ATIVIDADE 2

Qual foi a molécula pioneira na história da Biologia Molecular?

_____

_____

_____

_____

RESPOSTA COMENTADA

A molécula pioneira foi a proteína. O seqüenciamento de aminoácidos e a eletroforese de alozimas (reveja este termo nas Aulas 8 e 21, de Evolução) foram a base da Teoria do Relógio Molecular e de todas as filogenias geradas nas décadas de 1950 a 1980.

INFERÊNCIAS EVOLUTIVAS INTRA-ESPECÍFICAS OU INFERÊNCIAS GENEALÓGICAS

Quando trabalhamos com populações de uma dada espécie, estamos de fato abordando as metas gerais da Genética de Populações, as quais procuram caracterizar e explicar a variação genética intra-especificamente. Essa variação é a matéria bruta para futuras mudanças evolutivas, e os diferentes níveis de variação em populações atuais distintas podem evidenciar mudanças evolutivas ocorridas no passado.

A análise de variações alélicas intra-específicas, mais do que fornecer a possibilidade de propormos uma genealogia (ou árvore de genes e/ou alelos), permite uma série de análises estatísticas, englobando fluxo gênico, tamanho populacional, tamanho efetivo da população, divergências populacionais, histórias demográfica e mutacionais, freqüências alélicas, genotípicas e fenotípicas.

GENÉTICA MOLECULAR DE POPULAÇÕES

Em tempos passados, a genética de populações era um assunto puramente teórico. Seu foco constituía relações entre estrutura de populações, sistemas de acasalamento, mutação, migração, seleção e deriva genética, desde que estes pudessem ser deduzidos *a priori* de dados de herança mendeliana e processos darwinianos.

As freqüências alélicas eram as variáveis fundamentais da genética de populações, mas nenhum método experimental de utilização geral estava disponível para detectar diferenças de alelos entre organismos presentes em populações naturais.

Não existem dados de freqüências alélicas disponíveis para aplicar as teorias de genética de populações, com exceção de alguns casos especiais, como, por exemplo, as inversões cromossômicas em *Drosophila*, que podem ser estudadas citologicamente.

A genética de populações é mais importante hoje do que foi em qualquer outro tempo. Essa importância se deve à descoberta das diferenças genéticas (polimorfismos) entre organismos, o que tornou obsoleto o estudo genético focado em organismos mutantes que manifestam diferenças fenotípicas visíveis, como ervilhas, que são lisas ou enrugadas, ou moscas-de-fruta, com olhos vermelhos ou brancos e os cruzamentos controlados.

A genética de populações estuda as diferenças que ocorrem naturalmente entre os organismos. As diferenças da mesma espécie são chamadas polimorfismos genéticos. Divergências genéticas são as diferenças que se acumulam entre espécies. Define-se genética de populações como o estudo de polimorfismos e divergências.

O estudo direto de genes e seus produtos (proteínas), sem necessidade dos cruzamentos, significa que a análise genética detalhada não está mais restrita a animais domésticos, plantas cultiváveis e ao pequeno número de organismos experimentais que podem ser cultivados em laboratório. A análise genética é possível para qualquer organismo.

POLIMORFISMOS MOLECULARES

Um dos atributos universais das populações naturais é a diversidade fenotípica. Entre os indivíduos de qualquer população, muitos diferentes fenótipos podem ser encontrados para a maior parte dos caracteres. Variação genética, na forma de alelos múltiplos de vários genes, existe na maioria das populações naturais.

Atualmente, dados sobre as diferenças genéticas entre organismos são obtidos pela análise direta de moléculas de DNA ou proteínas.

O estudo de polimorfismos moleculares é baseado em seqüências de nucleotídeos ou aminoácidos. Os resultados consistem na forma de seqüências alinhadas, ou seja, arrumadas umas em relação às outras, de forma que cada posição corresponda à mesma posição na molécula do ancestral comum, a partir do qual todas as seqüências evoluíram. As seqüências podem derivar de indivíduos dentro de uma mesma espécie ou de indivíduos representando duas ou mais espécies.

POLIMORFISMOS DE DNA

Os métodos de manipulação do DNA (digestão com enzimas de restrição, hibridização com sondas, amplificação por PCR, eletroforese) podem ser usados em várias combinações para analisar o DNA de genomas amostrados a partir de populações naturais.

A literatura moderna apresenta grande quantidade de métodos de detecção da variabilidade genética. Cada abordagem possui vantagens e limitações. Os mais importantes tipos de métodos de análise das variações em nível de DNA foram descritos na Aula 8 desta disciplina.

SUBSTITUIÇÕES NÃO-SINÔNIMAS

São trocas de nucleotídeos no DNA codificante (ou gênico) que resultam em um novo códon que especifica um aminoácido diferente. Por exemplo: o códon GCA, que corresponde ao aminoácido alanina, sofre mutação, em que o G é alterado para C, formando o códon CCA, que corresponde ao aminoácido prolina. Essas mutações são ditas conservativas quando resultam na substituição de um aminoácido por outro quimicamente semelhante; e não-conservativas, quando o novo aminoácido possui cadeia lateral diferente. Substituições sinônimas ou silenciosas são trocas de nucleotídeos no DNA gênico que resultam em um novo códon que especifica o mesmo aminoácido. Por exemplo: o códon GCA, que corresponde ao aminoácido alanina, sofre mutação, em que o A é alterado para U, formando o códon GCU, que corresponde ao mesmo aminoácido alanina. Tais substituições freqüentemente ocorrem na posição da terceira base de um códon que, devido à degeneração do código genético, muitas vezes não implica alteração do aminoácido. Reveja a Aula 26 de Biologia Molecular em que foram apresentadas detalhadamente as características do código genético.

POLIMORFISMOS DE PROTEÍNA

As moléculas de proteínas podem ser separadas por eletroforese. Na técnica de eletroforese de isozimas, a posição da migração de uma enzima em uma matriz de eletroforese é identificada por meio da reação com um substrato específico acoplado a um corante que se precipita no local. Desta forma, a posição de uma enzima no gel é marcada pelo surgimento de uma banda escura.

A eletroforese de enzimas identifica um grupo de **SUBSTITUIÇÕES NÃO-SINÔNIMAS** de nucleotídeos, já que a troca de aminoácidos vai refletir em uma alteração na carga da molécula e, conseqüentemente, na migração da molécula no gel.

Polimorfismos desse tipo são chamados de alozimas. Existe menor quantidade de polimorfismos de proteínas do que de DNA, visto que a detecção do polimorfismo de alozimas requer diferença na seqüência de aminoácidos.

O polimorfismo de alozimas é demonstrado na **Figura 23.5**, que resume os resultados de experimentos de eletroforese em populações de 243 espécies. Os números entre parênteses constituem a quantidade de espécies examinadas em cada tipo de organismo. "Polimorfismo ou P" refere-se à proporção estimada de genes que são polimórficos; "Heterozigosidade ou H" refere-se à proporção estimada de genes codificantes de enzimas que se espera encontrar em heterozigose em um indivíduo médio.

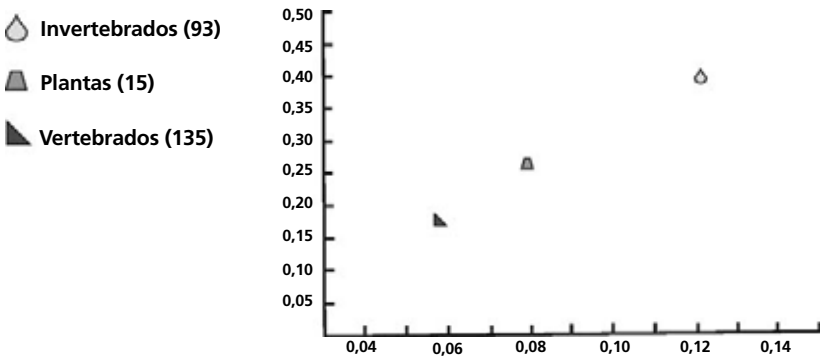


Figura 23.5: Níveis estimados de Heterozigosidade e Proporção de locos polimórficos derivados de estudos de alozimas para vários grupos de plantas e animais. Entre parênteses está o número de espécies estudadas.

Note, na **Figura 23.5**, que os invertebrados apresentam valores mais altos de genes polimórficos e de heterozigosidade do que as plantas e os vertebrados. Uma possível explicação para essa distribuição de valores seriam as limitações introduzidas nos sistemas orgânicos mais complexos pelos processos evolutivos, impedindo nesses organismos o acúmulo de mutações. Em outras palavras, quanto mais complexo o organismo, mais conservadas, menos variável ou polimórficas seriam suas proteínas.

ATIVIDADE 3

Qual a importância das técnicas da Biologia Molecular no estudo dos polimorfismos genéticos?

_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____

RESPOSTA COMENTADA

As técnicas de Biologia Molecular permitiram o estudo de polimorfismos genéticos em qualquer organismo, desde bactérias até baleias jubarte. Antes do advento da Biologia Molecular, só era possível estudar plantas cultiváveis, animais domésticos e organismos com tempos de geração pequenos (como camundongos, Drosophila e leveduras).

O CONTEÚDO INFORMATIVO DAS SEQÜÊNCIAS MOLECULARES

As seqüências podem fornecer muita informação. Para isso, devemos analisar alguns conceitos-chave que podem ser ilustrados com um exemplo. Os resultados da tabela a seguir compreendem 500 pares de bases (pb) da seqüência codificante de cinco alelos que ocorrem naturalmente no gene da rodopsina 3 (Rh3) de *Drosophila simulans*. Note que somente os sítios variáveis (polimórficos) estão apresentados no **Quadro 23.1**.

Quadro 23.1: Sítios polimórficos no gene da rodopsina 3 (Rh3) de *Drosophila simulans*

Sítios polimórficos (ocorrem outros 484 sítios monomórficos)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
f	T	C	T	A	C	C	T	C	C	T	C	G	G	T	T	A
g	T	C	C	T	A	C	C	T	C	C	T	G	G	T	T	T
h	C	T	C	C	C	C	C	T	C	T	T	T	G	C	T	A
i	C	T	C	C	C	C	C	T	T	C	T	G	A	C	T	T
j	C	T	C	C	C	T	C	T	T	T	T	G	G	C	C	A

Diferença pareada

6	6	4	7	4	4	4	4	6	6	4	4	4	6	4	6

Configuração amostral

(3,2) (3,2) (4,1) (3,1) (4,1) (4,1) (4,1) (4,1) (3,2) (3,2) (4,1) (4,1) (4,1) (3,2) (4,1) (3,2)

Sítios filogeneticamente informativos, onde S = sim e N = não

S	S	N	N	N	N	N	N	S	S	N	N	N	S	N	S

Vários tipos de sítios de nucleotídeos podem ser distinguidos:

1) Sítio segregante constitui uma posição que é polimórfica na amostra. Neste exemplo, são os 16 sítios apresentados. Eles estão numerados em seqüência, mas na realidade encontram-se espalhados ao longo dos 500pb, separados por distâncias que variam entre 2 e 104pb. A amostra contém 484pb que não variam, sítios monomórficos ou não-segregantes. A proporção de sítios segregantes (S) é de 16 dividido por 500 = 0.0320.

2) Diferença de pares ou pareada entre quaisquer duas amostras é um sítio no qual a seqüência difere. A proporção ou diferença pareada em uma amostra é obtida pela comparação das seqüências em todos os possíveis pares, fazendo a média do número das diferenças. No exemplo do gene Rh3, existem 5 seqüências que podem ser pareadas de 10 diferentes maneiras. Em n seqüências existem n(n-1)/2 possíveis comparações de pares. O número de diferenças de pares em cada sítio polimórfico está listado na parte inferior da tabela. Por exemplo, o sítio 1 possui 2 T e 3 C, de forma que 2 x 3 = 6 combinações em comparação de pares. Outro exemplo é o sítio 4, que possui 1 A, 1 T e 3 C, de forma que (1 x 3) + (1 x 3) + (1 x 1) = 7 combinações em comparação de pares.

3) A configuração amostral de um sítio é o conjunto de números fornecendo, em ordem decrescente, quantos elementos de cada tipo diferente estão presentes em um determinado sítio da amostra. O sítio 1, dos dados do gene Rh3, possui a configuração (3, 2, 0, 0), mas normalmente os zeros

são omitidos e a configuração é escrita (3, 2). A representação (3, 2) significa que o sítio amostrado inclui 3 seqüências com um nucleotídeo majoritário (neste exemplo o C) e 2 seqüências com um nucleotídeo diferente (neste caso, o T). O sítio 2 também possui configuração amostral (3, 2), embora, neste caso, os nucleotídeos majoritário e minoritário estejam invertidos. Isso significa que a configuração amostral é indiferente à identidade do nucleotídeo em um sítio, dependendo exclusivamente dos números relativos de tipos. Quando ocorre um empate, ambos os números são listados. Por exemplo, o sítio 4 tem configuração (3, 1, 1), em que cada 1 representa um *singleton*, tipo que ocorre uma única vez no sítio. Todos os 484 sítios monomórficos possuem a configuração (5), mas normalmente escrevemos (5, 0) para enfatizar que os sítios são invariáveis.

4) Uma amostra de seqüências alinhadas também contém sítios que fornecem informações sobre a genealogia ou relações de ancestralidade entre essas seqüências. Um sítio polimórfico de nucleotídeos é dito filogeneticamente informativo se ao menos uma minoria de nucleotídeos não forem *singletons*. Esses sítios permitem que as seqüências sejam divididas em dois grupos, cada qual contendo dois ou mais membros, sendo os membros de cada grupo mais similares entre si do que a membros de qualquer outro grupo. Por exemplo, o sítio 1 nos dados do gene Rh3 é filogeneticamente informativo, porque a configuração (3, 2) separa a amostra em dois grupos: o primeiro possuindo C no sítio e tendo três membros e o segundo possuindo T no sítio e tendo dois membros. A implicação é a de que em um tempo anterior, na história evolutiva, esse sítio devia ter sido monomórfico para C ou T, e uma substituição de nucleotídeos criou uma segunda linhagem com o sítio ocupado pelo nucleotídeo alternativo. Tal suposição é justificada desde que cada tipo de substituição de nucleotídeo em um sítio possa acontecer apenas uma vez e que não ocorra mutação reversa que possa restaurar o nucleotídeo original.

O ALINHAMENTO DE MACROMOLÉCULAS

Aqui chegamos a uma etapa da mais extrema importância em análises de sistemática molecular: o alinhamento das seqüências. Evidentemente, um alinhamento errado vai comprometer todo o resto

das análises. O mais comum é colocar as macromoléculas seqüenciadas no computador e deixar que um dos inúmeros programas feitos para alinhamento faça o resto.

!

Quase nunca as seqüências são corretamente alinhadas pelo computador. Claro que esses programas podem constituir um passo inicial, mas a forma mais correta e segura de alinhamento é manual!

Um programa computacional simplesmente procura "juntar igual com igual", sem qualquer "preocupação" com os processos biológicos. Para isso, é importante que, após uma primeira alternativa de alinhamento "proposta" pelo computador, olhemos para cada uma das bases e procuremos arranjá-las, utilizando os nossos conhecimentos.

Para tal, preciso conhecer os diferentes tipos de mutações e de substituições.

Quadro 23.2: Tipos de mutações possíveis

AAATCGATCCGATTA	seqüência original
G**AA**C**CGAT**TC**A**T**TA**	transições
AAATCGATCCGATTA	seqüência original
T**AA**AGT**AT**A**CCAG**TC	transversões
AAATCGATCCGATTA	seqüência original
AAAT/CCGATTA	deleção CGAT
AAATCGATCCGATTA	seqüência original
AAATCGATCTCCTA**CGATTA**	inserção
AAATCGATCCGATTA	seqüência original
AAATCCGATCG**ATTA**	inversão

!

Tipos de mutações

Você viu na aula de mutação e reparo de DNA (Aula 13 de Biologia Molecular) e reviu na Aula 9 de Evolução que as mutações de ponto podem ser classificadas como: (1) transições, quando ocorrem substituições de nucleotídeos de purina (dois anéis químicos) por purina ou de pirimidina (um anel químico) por pirimidina; e (2) transversões, quando ocorrem substituições de purina por pirimidina ou de pirimidina por purina (um anel químico por dois e vice-versa). Quando a mutação não é pontual, pode envolver deleção (perda) ou inserção (ganho) de vários nucleotídeos ou, ainda, a inversão da ordem de vários nucleotídeos no cromossomo (recorde as alterações estruturais dos cromossomos – Aula 19, Genética Básica).

Quadro 23.3: Tipos de mutações possíveis

Seqüência original	**ATA AAG GCA CTG GTC CTG**
	Ile Lys Ala Leu Val Leu
Sinônima	**ATA AAG GCA CTG GTA CTG**
	Ile Lys Ala Leu Val Leu
Não-sinônima	**ATA AAG** CCA **CTG GTC CTG**
	Ile Lys Pro **Leu Val Leu**
Sem sentido	**ATA** TAG **GCA CTG GTA CTG**
	Ile parada

!

Tipos de substituição

Veja os significados dos termos substituições sinônimas e não-sinônimas nos verbetes do início desta aula. Mutações sem sentido são trocas de nucleotídeos no DNA gênico que resultam em um códon de término (*stop codon*), que não especifica nenhum aminoácido e sinaliza para a interrupção da síntese do polipeptídeo que está sendo sintetizado. Por exemplo: o códon AAG que corresponde ao aminoácido lisina sofre mutação, em que o A é alterado para U, formando o códon UAG, que não tem correspondência para aminoácidos.

Agora imagine, por exemplo, como um computador "interpretaria" o alinhamento de regiões com deleções e insersões (também chamadas *indels* em inglês), o que é comum no caso de genes ribossomais e regiões repetitivas do genoma. O alinhamento manual tende a ser um trabalho demorado e muitas vezes cansativo, mas é imprescindível!

ATIVIDADE 4

Alinhe manualmente as seguintes seqüências obtidas para dois organismos:

AAATTGTTAACCCCTTGGAAAACCTTTGGG

AAAGTTAACCCCGGCTCTTTGGG

RESPOSTA COMENTADA

As seqüências foram alinhadas com base nas regiões conservadas (invariáveis). Foi necessária a inserção de lacunas (em inglês gaps) quando uma seqüência não apresentava correspondência com a outra. Em negrito estão as regiões variáveis.

AAATTGTTAACCCCTTGGAAAACCTTTGGG

AAA--GTTAACCCC—-GGCT---CTTTGGG

INFERÊNCIAS EVOLUTIVAS INTERESPECÍFICAS OU INFERÊNCIAS FILOGENÉTICAS

Muita gente acha que, uma vez tendo as seqüências alinhadas, basta usar um programa de computador e a "árvore" sairá pronta. Infelizmente, ainda há muitos que tratam as análises filogenéticas como uma "caixa-preta" e não usam conhecimentos científicos para analisar os dados. Vamos tentar desenvolver um quadro conceitual para que possamos entender – na teoria e na prática – pelo menos um pouco dos princípios e das metodologias de que se dispõe hoje para inferências filogenéticas.

As técnicas de sistemática molecular produzem fundamentalmente dois tipos de informações:

- dados de distância: quando as diferenças entre moléculas são medidas como uma só variável;
- dados de caráter: quando as diferenças entre moléculas são medidas como uma série de variáveis descontínuas, sendo cada uma delas do tipo multiestado.

Dados de caráter podem ser convertidos em dados de distância, mas dados de distância nunca podem ser convertidos em dados de caráter.

Mas o que são variáveis do tipo multiestado? Imagine uma sequência de DNA qualquer que tenha 20 nucleotídeos, como no exemplo a seguir:

ACTTTCGATGCTAAGCTAAT

Cada uma das bases ocupa uma posição distinta na sequência. No nosso exemplo, a primeira adenina (A) ocupa a posição 1; a citosina seguinte ocupa a posição 2 e assim por diante, como está representado a seguir:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
A	C	T	T	T	C	G	A	T	G	C	T	A	A	G	C	T	A	A	T

Dessa forma, cada posição, ou sítio, ocupada na seqüência da macromolécula será considerada um caráter independente dos caracteres (posições) que a precedem ou sucedem. A presença de uma adenina na posição 1 de nosso exemplo vai constituir o estado do caráter denominado "primeira posição na seqüência"; a citosina será o estado de caráter da segunda posição, e assim por diante. Agora, embora no nosso exemplo tenhamos uma adenina, imagine que poderíamos ter, na primeira posição, uma citosina, uma guanina ou uma timina (ver exemplo a seguir); esse mesmo tipo de raciocínio é aplicável a todas as outras posições.

1	ou	1	ou	1	ou	1
A		C		G		T

Assim, os caracteres podem assumir pelo menos quatro estados diferentes; daí dizermos que, quando analisamos dados de caráter, temos como ferramenta "uma série de variáveis descontínuas, sendo cada uma delas do tipo multiestado".

INFERIR FILOGENIAS: FILOGENÉTICA MOLECULAR

O alinhamento de seqüências de aminoácidos ou nucleotídeos pode ser utilizado para formular suposições acerca das relações ancestrais entre indivíduos ou grupos taxonômicos. A filogenética molecular ou sistemática molecular é a disciplina que formula essas suposições ou inferências.

Cada alinhamento de seqüências resulta em uma árvore gênica. Essa árvore não é necessariamente congruente com uma árvore de espécies, devido à maneira pela qual os polimorfismos nas espécies ancestrais tornam-se dispersos nas espécies descendentes.

A **Figura 23.6** apresenta a árvore para sete espécies, S1-S7, e um sítio nucleotídico que é polimórfico para A e C no ancestral comum S1. O polimorfismo é retido na espécie S2, mas ocorre fixação em todas as outras espécies. Devido à maneira como a fixação ocorreu, este sítio nucleotídico sugere que a espécie S4 é relacionada mais proximamente com S6 e S7 do que com S5. Mas a verdade de fato é o oposto! Esse tipo de problema é o mais crítico para espécies proximamente relacionadas.

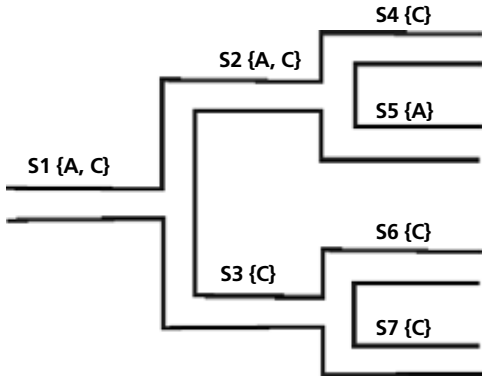


Figura 23.6: Árvore gênica com sete espécies. Um sítio de nucleotídeos é polimórfico para A e C no ancestral comum (espécie S1).

Para espécies que não são proximamente relacionadas, ocorre um outro tipo de problema, ou seja, duas ou mais mutações independentes ocorrem no mesmo sítio (*multiple hits*, golpe ou acerto múltiplo). Devido ao golpe múltiplo, dois sítios que diferem podem ter sofrido mais de uma mudança. Há também possibilidade de homoplasia, que, no contexto da filogenética molecular, refere-se a sítios de nucleotídeos ou de aminoácidos que são idênticos, não por causa de identidade por descendência a partir de um ancestral comum, mas por mutação de um dos seguintes tipos:

- mutações paralelas no mesmo sítio (por exemplo, duas substituições C→T independentes);
- mutações convergentes no mesmo sítio (por exemplo, C→T em uma seqüência e A→T em outra);
- mutações reversas no mesmo sítio (por exemplo, C→T e mais tarde T→C).

O número de diferenças entre duas seqüências alinhadas pode estar, na verdade, subestimado, devido ao efeito dos golpes múltiplos. Alguns dos métodos para corrigir esses efeitos serão examinados a seguir.

MODELOS DE EVOLUÇÃO DE SEQÜÊNCIAS

Em primeiro lugar, é preciso levar-se em conta que, a partir da comparação de seqüências atuais, não é possível reconhecer todas as substituições realmente ocorridas durante a evolução das seqüências, pelo fato de poder haver mais de uma substituição em uma única posição.

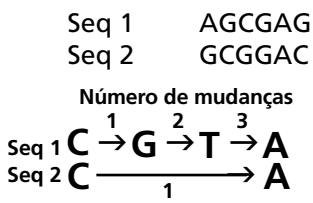


Figura 23.7: Duas seqüências que atualmente apresentam o nucleotídeo A na quinta posição, apesar de possuírem um ancestral comum com um nucleotídeo C na quinta posição. A história evolutiva dessas seqüências diferiu porque a seqüência 1 sofreu quatro mudanças e a seqüência 2 sofreu apenas uma.

Para se lidar com esse problema, é necessário que as distâncias sejam corrigidas, de acordo com algum modelo.

De modo geral, os modelos de evolução de seqüências baseiam-se no processo de Markov, em que cada mudança de um nucleotídeo para outro apresentará uma taxa específica. Assim, supõe-se que as substituições obedeçam a uma distribuição de Poisson e as taxas dessas substituições possam ser arranjadas em uma matriz geral. Nessa matriz, as taxas de substituição serão especificadas pelos parâmetros associados aos 12 possíveis tipos de mudanças (os 4 tipos de transição e os 8 tipos de transversão) e à freqüência de bases, assumindo 4 possibilidades (A, C, T ou G). Assim, a matriz será do tipo "4 por 4" e os diferentes modelos de substituição serão simplesmente casos especiais da matriz geral.

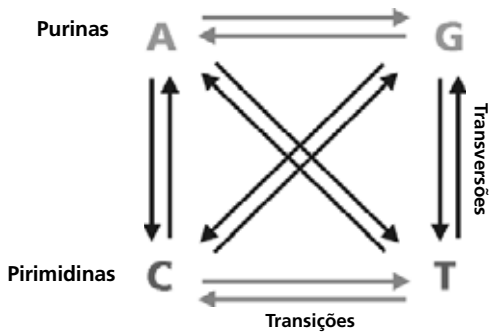


Figura 23.8: Tipos possíveis de mudanças de nucleotídeos.

ATIVIDADE 5

Identifique as mutações abaixo como transições (S) ou transversões (V):

( ) A → G	( ) T → G
( ) A → C	( ) T → C
( ) A → T	( ) T → A
( ) C → G	( ) G → C
( ) C → A	( ) G → A
( ) C → T	( ) G → T

RESPOSTA COMENTADA

Você lembrou que transições são trocas de purinas por purinas ou pirimidinas por pirimidinas? E que as purinas têm dois anéis químicos e as pirimidinas apenas um? Essa lembrança ajuda a resolver a atividade. Vejamos: as purinas são A e G; as pirimidinas, C e T. Trocas A → G e C → T serão sempre transições. Já as transversões são trocas de bases do tipo purina (dois anéis) por pirimidinas (um anel) e vice-versa.

Assim, temos:

(S) A → G	*(V) T → G*
(V) A → C	*(S) T → C*
(V) A → T	*(V) T → A*
(V) C → G	*(V) G → C*
(V) C → A	*(S) G → A*
(S) C → T	*(V) G → T*

O mais simples modelo de evolução ou de substituição de nucleotídeos foi desenvolvido por Jukes e Cantor (1969).

Para \ De	A	T	C	G
A	$1-3\alpha$	α	α	α
T	α	$1-3\alpha$	α	α
C	α	α	$1-3\alpha$	α
G	α	α	α	$1-3\alpha$

Figura 23.9: Matriz de substituição com um parâmetro, segundo Jukes e Cantor (1969).

O modelo de Jukes e Cantor é o de um parâmetro e assume que: 1) todas as mudanças têm probabilidades iguais (25%) de ocorrência; 2) todos os sítios podem ser alterados; 3) eles fazem isso na mesma velocidade.

Existem outros modelos, mais realistas, que levam em conta o fato de existir uma fração dos nucleotídeos que nunca é substituída; de que as transições (substituições entre pirimidinas ou entre purinas) são mais freqüentes que as transversões (substituições de purina para pirimidina ou vice-versa); de que as taxas de substituições entre os sítios são heterogêneas; e de que as proporções entre as bases são diferentes etc. O princípio continua o mesmo, ou seja, verificamos nas seqüências atuais um número menor de substituições do que o que realmente ocorreu na evolução. Com os modelos, pretendemos saber o valor de distância mais adequado para a reconstrução da árvore.

Os modelos de substituição estão relacionados uns aos outros, partindo de um mais simples em direção a modelos mais complexos, isto é, mais ricos em parâmetros (veja um exemplo na **Figura 23.10**).

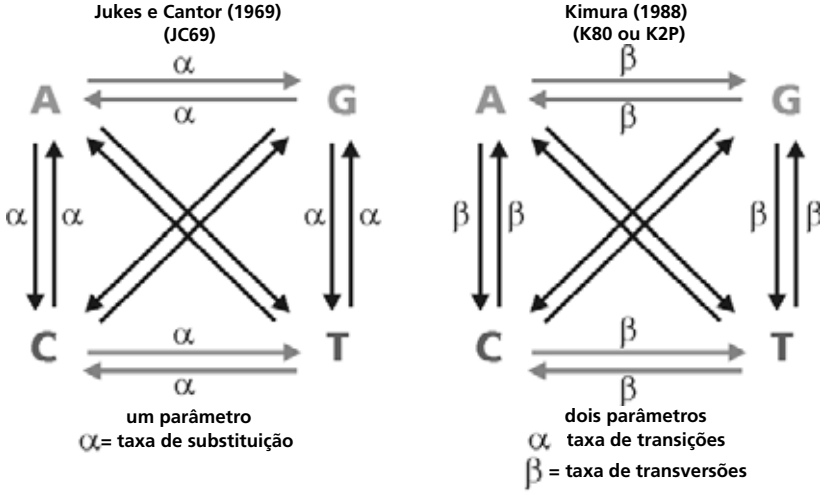


Figura 23.10: Modelos com um e dois parâmetros.

Em resumo, os modelos apresentam um entrelaçamento espacial de acordo com o aumento ou diminuição dos respectivos parâmetros, tornando-se os modelos mais simples casos particulares dos mais complexos.

Mas, na prática, como decidir qual modelo de substituição seria o mais adequado a um determinado conjunto de dados? Perguntas do tipo "será que ao acrescentarmos novos parâmetros aos modelos estaremos melhorando os valores de verossimilhança?" são bastante comuns.

Em teoria, o ideal seria estimar os valores de verossimilhança para um conjunto de dados, utilizando todos os diferentes modelos e, então,

escolher o melhor deles; ou seja, escolher o que apresentou a maior probabilidade de explicar a origem evolutiva das seqüências estudadas para a inferência filogenética. Trata-se de um trabalho e tanto, já que contamos atualmente com mais de 5 dezenas (50!) de modelos descritos na literatura! Felizmente, já existem programas computacionais que realizam esse tipo de teste, chamado Teste de Razão de Verossimilhança (ou LRT, do inglês *Likelihood Ratio Test*), bastante conhecido na estatística clássica.

MÉTODOS DE INFERÊNCIA FILOGENÉTICA

A Sistemática Molecular revolucionou as abordagens na classificação biológica dos organismos, pois ela utiliza dados que são independentes da Morfologia. As relações de ancestralidade entre organismos inferidas a partir de seqüências moleculares usualmente dão suporte àquelas inferidas a partir de caracteres morfológicos.

Muitos métodos têm sido desenvolvidos para inferir relações de ancestralidade entre um grupo de seqüências alinhadas. Elas podem ser comparadas por análises de árvores filogenéticas, obtidas por meio de simulações computacionais de evolução de seqüências, ou por meio dos próprios organismos, quando uma verdadeira filogenia é conhecida, por exemplo, a partir de experimentos. Os métodos diferem quanto:

- à eficiência no uso do tempo de processamento computacional e no número de sequências que podem ser analisadas;
- à capacidade de identificar a árvore correta para uma dada quantidade de dados;
- à consistência da árvore correta; com crescente probabilidade, conforme aumenta a quantidade de dados;
- à robustez da árvore correta, mesmo que algumas concepções do método estejam equivocadas.

Não surpreendentemente, todos os métodos atuam razoavelmente bem, se os dados estiverem de acordo com as concepções delineadas pelo método e se houver dados suficientes. O fator mais importante parece ser a precisão da correção adotada para os "golpes múltiplos" (*multiple hits*). Ainda assim, a maioria dos métodos apresenta uma decepcionante performance quando a taxa de evolução de um ramo para o seguinte varia dramaticamente.

Como nenhum dos métodos é superior em cada critério sob todas as condições, uma grande variedade de métodos coexiste. Muitos autores optam por analisar seus dados utilizando múltiplos métodos, na esperança de que as árvores resultantes difiram, no máximo, em detalhes não-essenciais. Uma discussão aprofundada dos métodos e seus respectivos méritos e deficiências está além da abordagem desta aula.

Os métodos mais comumente usados podem ser classificados em três grandes vertentes:

Métodos de Distância

São fundamentados em diferenças pontuais/locais entre seqüências (*pairwise differences*), corrigidas para os golpes múltiplos. Estes incluem:

- *Unweighted pair-group method with arithmetic mean* (UPGMA). Esse método considera a árvore filogenética aditiva e que todos os táxons estão igualmente distantes da raiz. Ele tem caído em desuso, pois assume uma taxa constante de evolução em cada ramo e atua de maneira pobre quando esta concepção é violada (o que constantemente ocorre);
- *Minimum evolution* (ME) ou Evolução Mínima. Nesse método, estima-se, para cada árvore alternativa possível, o comprimento de cada braço ou ramo, a partir das distâncias evolucionárias entre os táxons, computando-se a somatória de todos os comprimentos de braços (S). O critério de ME é a árvore que apresenta o menor valor da somatória dos comprimentos dos braços (S). Resumindo: ele examina todas as árvores possíveis e seleciona aquela que apresenta o menor comprimento total dos ramos. Essa abordagem é computacionalmente intratável quando lidamos com um grande número de seqüências, porque há muitas árvores possíveis;
- *Neighbor joining* (NJ) ou Agrupamento de Vizinhos. Esse método é baseado no princípio da Evolução Mínima. Ele não examina todas as topologias possíveis, mas procura encontrar, seqüencialmente, vizinhos que minimizem o comprimento total da árvore. Agrupa seqüencialmente os pares de seqüências mais intimamente relacionados. Esse método é extremamente eficiente em termos computacionais; usualmente apresenta árvores bastante próximas àquelas apresentadas pelo método anterior.

Métodos de Parcimônia

É baseado na suposição de que a árvore mais provável é a que requer o menor número de mudanças para explicar toda a variação observada na matriz de caracteres (por exemplo: seqüências alinhadas). O método baseia-se no princípio da homologia, ou seja, se dois táxons compartilham uma característica, é porque esta foi herdada do último ancestral comum a ambos. Ainda que a evolução possa não ser sempre estritamente parcimoniosa, o método assume que o critério de parcimônia leva ao maior número total de acertos da árvore verdadeira, quando se minimiza nela o número de passos evolutivos aceitos. Assim, são métodos que sistematicamente buscam, dentre as árvores possíveis, aquela com o menor número de passos mutacionais.

- *Unweighted parsimony* (UP) trata cada tipo de mudança (por exemplo, transição ou transversão) como igualmente informativo;
- *Weighted parsimony* (WP) atribui maior importância a alguns tipos de mudanças (usualmente transversões) ao selecionar a melhor árvore. Esse método usualmente atua de melhor maneira que o anterior.

Método de Máxima Verossimilhança ou *Maximum likelihood* (ML)

É um método que assume um modelo de substituição de nucleotídeos ou aminoácidos e, baseado nesse modelo, identifica a árvore que maximiza a probabilidade de se obter as seqüências observadas. Intuitivamente apelador, porém computacionalmente "pesado", esse método é bastante tolerante com a violação de suas concepções e atua muito bem, mesmo quando as taxas de substituição são moderadamente diferentes entre os ramos.

ATIVIDADE 6

Em que os métodos de Máxima Verossimilhança e de Parcimônia são similares entre si e distintos dos Métodos de Distância?

_____
_____
_____
_____
_____
_____
_____

RESPOSTA COMENTADA

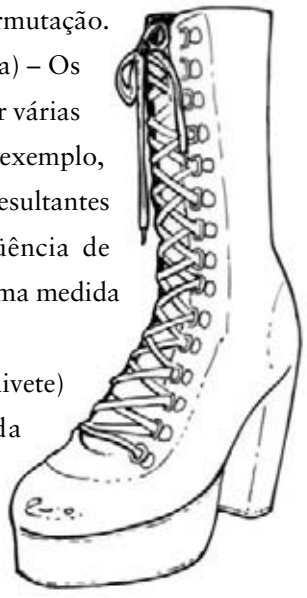
Os Métodos de Distância são baseados em matrizes de distâncias; ou seja, a matriz de caracteres é transformada em uma matriz de distâncias. Já os métodos de Máxima Verossimilhança e de Parcimônia são baseados em análises de estados de caracteres (os caracteres são analisados diretamente).

PROCURANDO A ÁRVORE ÓTIMA: CONFIANÇA NAS ÁRVORES OBTIDAS OU CONFIANÇA EM HIPÓTESES FILOGENÉTICAS

Os resultados das inferências filogenéticas devem ser testados. As árvores geradas pelos distintos métodos são analisadas por métodos que atribuem valores de confiança nos nós. Esses métodos são classificados em: 1) métodos de reamostragem de caracteres (*Bootstrap, Jack-knife*), 2) de análise de decaimento, e 3) testes de permutação.

• *Bootstrap* (tradução livre do inglês: cadarço de bota) – Os caracteres são reamostrados com realocação para criar várias matrizes replicadas; as réplicas são analisadas (por exemplo, por parcimônia), e a concordância entre as árvores resultantes é resumida em um consenso de maioria. A freqüência de ocorrência dos grupos (Proporções do *Bootstrap*) é uma medida de sua confiabilidade;

• *Jack-knifing* (tradução livre do inglês: passar o canivete) – Uma proporção dos caracteres é selecionada aleatoriamente e apagada, criando-se várias matrizes replicadas menores; as réplicas são analisadas (por exemplo, por parcimônia), e a concordância entre as árvores resultantes é resumida em um consenso de



Bootstrap

maioria. A freqüência de ocorrência dos grupos (Proporções do *Jack-knife*) é uma medida de sua confiabilidade – os resultados são muito parecidos aos do *Bootstrap*, mas o método não é tão disponível nem tão utilizado;

• Análise de Decaimento – Avalia-se o número de passos (Índice de Decaimento ou Suporte de Bremer) entre a árvore de máxima parcimônia e a primeira árvore subótima, em que determinado clado não mais apareça. Ele indica o número necessário de mutações para quebrar um determinado arranjo. Quanto maior esse número, maior a confiança em seus resultados. O Suporte Total para uma árvore é a soma dos Índices de Decaimento de cada clado;

• Teste de Permutação – A idéia é verificar a força de agrupamento de um determinado clado em uma árvore filogenética. Ele compara a melhor árvore com árvores forçadas a serem compatíveis com uma árvore restringida. Para isso, é estimada a diferença entre escores da árvore de Máxima Parcimônia contendo o referido clado e de árvores sem o referido clado. A significância é dada pelo valor da probabilidade (p = 0,01 é significante em nível de 1%).

CONFIABILIDADE DOS MÉTODOS FILOGENÉTICOS

Os métodos filogenéticos podem ser avaliados quanto: 1) a sua consistência (quanto mais dados, mais próximos da verdade); 2) a sua eficiência (quão rápidos com determinado número de dados), e 3) a sua robustez (quão sensíveis às violações dos pressupostos).

A maior parte dessas avaliações foi conduzida com muito poucos táxons (na maioria, apenas quatro).

Amplas simulações com quatro táxons mostraram que: 1) métodos baseados em modelos têm bom desempenho quando o modelo é preciso; 2) violação dos pressupostos leva todos os métodos a inconsistências (**Zona de Felsenstein**) quando os comprimentos dos ramos ou taxas de mutação forem muito desiguais; 3) métodos de máxima verossimilhança são bastante robustos frente a violações dos pressupostos do modelo; 4) parcimônia com pesos diferenciados pode ter desempenho superior à parcimônia tradicional (pesos iguais), ou seja, pode ter uma Zona de Felsenstein menor.

Zona de Felsenstein

É uma região no espaço paramétrico de inconsistência para um determinado método de inferência filogenética, sob determinado modelo evolutivo.

Não se sabe quão generalizáveis são as conclusões obtidas com quatro táxons, já que simulações com muitos táxons sugeriram que a parcimônia pode ser bastante precisa e eficiente. Portanto, necessita-se de mais estudos para auxiliar na escolha do método de preferência.

CONCLUSÃO

Atenção! Apenas depois de se ter levado a efeito todos os procedimentos adequados para inferências filogenéticas, ou seja, após utilizar quaisquer das metodologias aqui discutidas, é que começa um dos trabalhos mais sérios do pesquisador: com os resultados em mãos, chegou o momento de olhar para eles e interpretá-los à luz do conhecimento científico! É com os resultados em mãos que devemos considerar a biologia dos organismos estudados associada aos padrões e processos evolutivos. Um computador não pensa nem considera nada; segue apenas algoritmos específicos. Os cérebros pensantes somos nós, e apenas nós poderemos contribuir cientificamente, e não o resultado que sai pronto do computador!

RESUMO

A Filogenia Molecular é o estudo da história evolutiva de populações de organismos relacionados a partir de dados moleculares.

A obtenção desses dados só foi possível com o advento da Biologia Molecular, com suas sucessivas eras de estudos de proteínas e DNA.

Atualmente, é possível realizar uma análise molecular em qualquer organismo; os dados sobre as diferenças genéticas entre organismos são obtidos, pois, pela análise direta de moléculas de DNA ou proteínas.

O estudo de polimorfismos moleculares é baseado em seqüências de nucleotídeos ou aminoácidos. Os resultados consistem na forma de seqüências alinhadas, arrumadas umas em relação às demais, de maneira que cada posição corresponda a uma outra na molécula do ancestral comum, a partir do qual todas as seqüências evoluíram.

O alinhamento de seqüências de aminoácidos ou nucleotídeos pode ser utilizado para formular suposições acerca das relações ancestrais entre indivíduos ou grupos taxonômicos. Cada alinhamento de seqüências resulta em uma árvore gênica.

Os modelos de evolução molecular são simplificações que simulam quantas e quais substituições de nucleotídeos ocorreram durante a evolução das seqüências.

Muitos métodos têm sido desenvolvidos para inferir relações de ancestralidade entre um grupo de seqüências alinhadas. Os três principais são: Métodos de Distância, Parcimônia e Máxima Verossimilhança. O primeiro baseia-se em dados de distância e os outros dois em dados de caracteres.

Os resultados das inferências filogenéticas devem ser testados por métodos que atribuem valores de confiança nos nós. Tais métodos são classificados em: 1) método de reamostragem de caracteres (*Bootstrap, Jack-knife*), 2) método de análise de decaimento, e 3) testes de permutação.

ATIVIDADES FINAIS

1. O que você entende por Teoria do Relógio Molecular?

_____

_____

_____

_____

_____

RESPOSTA

A Teoria do Relógio Molecular determina que as moléculas evoluem em proporção direta ao tempo. Assim, as diferenças entre seqüências homólogas de DNA ou proteínas podem ser usadas para estimar o tempo transcorrido, uma vez que as duas moléculas divergiram.

2. Que tipos de dados são utilizados nas inferências filogenéticas?

_____

_____

_____

_____

_____

RESPOSTA

Utilizamos dois tipos de dados nas inferências filogenéticas: dados de distância e dados de caracteres. Os primeiros medem as diferenças entre moléculas na base de uma só variável (presença ou ausência, por exemplo). Os dados de caracteres medem diferenças como uma série de variáveis descontínuas, sendo cada uma do tipo multiestado (por exemplo: zero, um, dois ou três espinhos por pata de um inseto, totalizando quatro variáveis de estados distintos).

3. Para que servem os modelos de evolução molecular e como eles diferem entre si?

_____

_____

_____

_____

RESPOSTA

Os modelos de evolução molecular são simplificações que simulam quantas e quais substituições de nucleotídeos ocorreram durante a evolução das seqüências. Com os modelos de substituição, desde os mais simples, com apenas um parâmetro, até os modelos mais complexos, com múltiplos parâmetros, pretendemos saber qual é o valor de distância mais adequado para a reconstrução da árvore. O melhor será o que apresentar a maior probabilidade de explicar a origem evolutiva das seqüências estudadas.

4. Quais são os principais métodos de inferência filogenética e em que são baseados?

_____

_____

_____

_____

_____

RESPOSTA

São três: Métodos de Distância, Parcimônia e Máxima Verossimilhança. O primeiro é baseado em dados de distância e os outros dois em dados de caracteres.

5. Como e por que testamos a confiança de uma árvore filogenética?

_____

_____

_____

_____

_____

RESPOSTA

Testamos a confiança de uma árvore filogenética por meio da utilização de métodos de reamostragem de caracteres (como o Bootstrap), análise de decaimento ou testes de permutação. Os métodos filogenéticos devem ser testados, principalmente, para verificar sua consistência (proximidade com a verdade) e robustez (sensibilidade às violações dos pressupostos).

AUTO-AVALIAÇÃO

Você estudou nesta aula como extrair informações de seqüências de nucleotídeos e de aminoácidos. Quem diria que os 4 nucleotídeos e os 20 aminoácidos poderiam nos contar tantas histórias de vida! Para entender como as seqüências são informativas foi necessário introduzir termos e métodos de análise. Você conseguiu acompanhar? Se conseguiu, que bom! Então, passe para a parte prática na próxima aula. Não? Vamos rever a Aula 26, Tradução ou Síntese de Proteínas, da disciplina de Biologia Molecular, para que as mutações e suas conseqüências no DNA codificante sejam esclarecidas. Não se detenha às definições dos métodos de inferência filogenética. Neste curso, o mais importante é que você saiba que eles existem, mesmo que não consiga descrevê-los em detalhes (muitos pesquisadores que utilizam esses métodos desconhecem as definições).

INFORMAÇÕES SOBRE A PRÓXIMA AULA

Na próxima aula, você vai aprender a utilizar os bancos de dados genéticos da internet, inclusive como usar programas de busca e identificação de seqüências de ácidos nucléicos e proteínas.