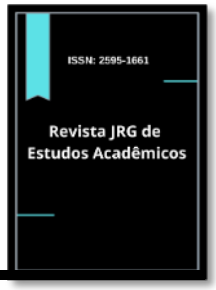
B1 ISSN: 2595-1661

ARTIGO ORIGINAL

Listas de conteúdos disponíveis em Portal de Periódicos CAPES

Revista JRG de Estudos Acadêmicos

Página da revista:
https://revistajrg.com/index.php/jrg



Análises filogenéticas microbianas utilizando dados moleculares: uma revisão

Microbial philogenetic analysis using molecular data: a review

DOI: 10.55892/jrg.v8i18.1790
ARK: 57118/JRG.v8i18.1790

Recebido: 14/12/2024 | Aceito: 05/01/2025 | Publicado *on-line*: 08/01/2025

Elder Cavalcante Fabian[1]
https://orcid.org/0009-0001-7592-5580
http://lattes.cnpq.br/1982428225139183
Instituto Federal Educação, Ciências e Tecnologia, IFMT, Brasil.
E-mail: eldercf@gmail.com

Flávia Tavares Couto Fabian[2]
https://orcid.org/0009-0000-0049-0220
http://lattes.cnpq.br/4192270855591406
Instituto Federal Educação, Ciências e Tecnologia, IFMT, Brasil.
E-mail: tavares.fla@@gmail.com

Resumo

A ideia de desenhar e esquematizar a evolução biológica por meio de árvores filogenéticas provavelmente teve como ponto de partida o desenvolvimento do sistema de classificação por Linnaeus, no século 18, dando origem a filogenia. O estudo das relações evolutivas entre os organismos há muito tem sido explorado pela ciência e durante muito tempo estes processos foram analisados por meio de dados morfológicos e comportamentais. Nas últimas décadas isto tem mudado significativamente graças ao desenvolvimento das técnicas de biologia molecular e o surgimento da filogenética molecular. A partir do ano de 2005, com o desenvolvimento e democratização de técnicas de sequenciamento de DNA conhecidas como Sequenciamento de Nova Geração (NGS), permitiu-se uma nova abordagem de sequenciamento em larga escala acarretando um aumento exponencial do volume de informações disponíveis sobre as características genéticas dos organismos. A capacidade que estas técnicas possuem para resolver questões de grande complexidade tem promovido as análises filogenéticas a ferramentas essenciais para uma diversidade cada vez maior de áreas de pesquisa. Sabidamente, o processamento manual de um volume tão grande de informações se torna inviável, surgindo daí a necessidade de desenvolvimento de ferramentas capazes de trabalhar com grandes bancos de dados. Surge então dessa necessidade a bioinformática, que é um campo interdisciplinar da ciência que combina áreas como a biologia, ciência da computação, estatística, matemática e engenharia para analisar, interpretar e processar dados biológicos. Este estudo apresenta uma revisão narrativa sobre os principais conceitos e metodologias aplicados a modelos evolucionários e de

[1] Graduado em Medicina Veterinária, Mestre em Ciência Animal; Doutor em Ciência Animal.
[2] Graduada em Medicina Veterinária, Mestra em Ciência Animal; Doutoranda em Biociência Animal

reconstrução filogenética, tendo como objetivo contextualizar e dar subsídios para um melhor entendimento sobre os passos de uma análise filogenética a pesquisadores que queiram utilizar esta ferramenta em seus estudos.

Palavras-chave: filogenética molecular; cladograma; relações evolutivas; genes; taxonomia.

Abstract
The idea of designing and scheming biological evolution through phylogenetic trees probably started with the development of the classification system by Linnaeus in the 18th century, giving rise to phylogeny. The study of evolutionary relationships between organisms has long been explored by science and for a long time these processes were analyzed using morphological and behavioral data. In recent decades this has changed significantly thanks to the development of molecular biology techniques and the emergence of molecular phylogenetics. Beginning in 2005, with the development and democratization of DNA sequencing techniques, known as New Generation Sequencing (NGS), a new large-scale sequencing approach was allowed, leading to an exponential increase in the volume of information available on the genetic characteristics of organisms. The capacity that these techniques have to solve issues of great complexity has promoted phylogenetic analysis into essential tools for an increasing diversity of research areas. It is known that the manual processing of such a large volume of information becomes unfeasible, hence the need to develop tools capable of working with huge databases. From that need bioinformatics arises, which is an interdisciplinary field of science that combines biology, computer science, statistics, mathematics and engineering to promote proper analysis, interpretation and processing of these biological data. This study presents a narrative review on the main concepts and methodologies applied to evolutionary models and phylogenetic reconstruction, aiming to contextualize and provide subsidies for a better understanding of the steps of a phylogenetic analysis to researchers who want to use this tool in their studies.

Keywords: *molecular phylogenetics; cladogram; evolutionary relationships; genes; taxonomy.*

1. Introdução

Em meados da década de 1960 o entomólogo alemão Willi Hennig propôs uma nova sistemática para o estabelecimento das relações de parentesco entre os organismos que revolucionou a prática classificatória. Tendo como base principal a teoria da evolução de Darwin e Wallace, e unindo uma objetividade metodológica à perspectiva evolutiva, o método henningiano supunha que os organismos se relacionavam genealogicamente uns com os outros e poderiam ser classificados de acordo com o quão recente era o seu ancestral comum. Mais tarde, o método de Hennig foi denominado de sistemática filogenética e nos dias atuais é chamado de cladística [1].

A cladística é baseada na ideia de que todos os grupos de uma árvore filogenética devem ser monofiléticos. Isso significa que cada grupo deve incluir todas as espécies que descendem de um único ancestral comum. Para reconhecer um grupo monofilético, deve-se identificar sinapomorfias, que são características compartilhadas pelos membros desse grupo [2].

Segundo Sakamoto[3], para a construção destes cladogramas, tradicionalmente utilizam-se dados mais acessíveis como características morfológicas e anatômicas. Já atualmente, graças ao desenvolvimento das técnicas moleculares como o Sequenciamento de Nova Geração **(**NGS), têm crescido a utilização de dados moleculares como sequências de DNA como fonte de dados para os estudos filogenéticos.

De acordo com Pinto [4], eventos genéticos como mutações, reorganização de genomas e recombinações são responsáveis pela biodiversidade e variabilidade genética existentes atualmente. Destes eventos, somente as mutações são consideradas pelos diferentes métodos moleculares de inferência filogenética.

Para que os métodos de filogenética molecular possam ser executados de maneira correta, deve-se considerar a similaridade entre os genes estudados e assumir sua homologia, ou seja, suas semelhanças. Quando se comparam duas seqüências entre si, pode-se sempre calcular o grau de similaridade através da contagem do total de nucleotídeos idênticos entre elas. Quanto maior o grau de similaridade, maior a possibilidade de que estas sequências sejam homólogas [4].

Uma análise cladística tem por finalidade a construção da representação gráfica das relações filogenéticas entre táxons de um determinado grupo por meio de árvores, também chamadas de cladogramas, sendo esta árvore filogenética uma hipótese acerca do relacionamento evolutivo entre um grupo de organismos [5].

Conforme descrito por Buso [6], os métodos cladísticos possuem como característica o fato de indicarem, dentre as árvores calculadas para cada caráter, a que melhor representa-o. Entre os métodos utilizados para esta determinação podemos citar o método da máxima parcimônia, o método da máxima verossimilhança e a inferência bayesiana [3].

2. Conceitos básicos

2.1 O processo evolutivo: anagênese e cladogênese

Dentre as escolas de classificação baseadas em princípios evolutivos, a escola da Filogenética, ou Cladística, tem ganhado força nas últimas décadas por propor uma objetividade metodológica à perspectiva evolutiva. Os princípios que fundamentam a Cladística são a anagênese e a cladogênese [1]. Mazzarolo [7] afirma que o processo evolutivo pode ser definido como uma sequência de repetições intercaladas de dois sub-processos muito importantes: a anagênese e a cladogênese logo, uma novidade evolutiva (Figura 1).

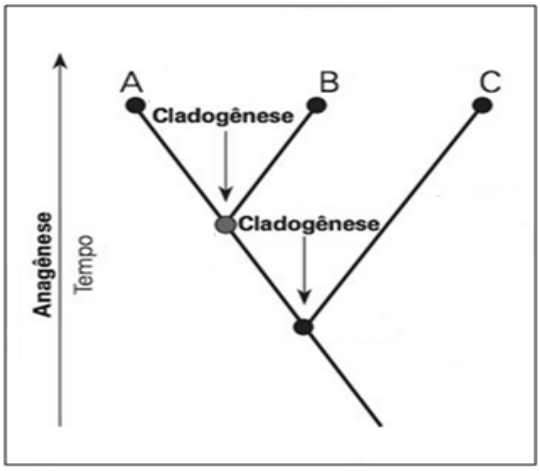


Figura 1. Representação de um cladograma, com os possíveis eventos de anagênese e cladogênese (Adaptado de Moutinho [8].

A anagênese representa a parte do processo de evolução na qual as características de uma determinada espécie podem evoluir ao longo do tempo, ou seja, corresponde a uma evolução progressiva com mudanças nas frequências genéticas de uma população inteira. Caracteriza-se por ser um processo que ocorre apenas no nível de espécie levando a uma alteração das características de um ramo filético sem, no entanto, levar a uma ramificação deste. Esse processo pode ser decorrente de eventos como mutações, recombinações, seleção diferencial de genótipos e deriva genética. A evolução conduzida pela anagênese muitas vezes é dita como microevolução [9].

De maneira distinta, a cladogênese é o processo responsável pela ruptura da coesão inicial de um ramo filético, gerando dois ou mais ramos isolados, que passam a evoluir independentemente. Nesses eventos, que representam processos conhecidos como especiação, há uma interrupção do fluxo gênico entre os indivíduos das espécies descendentes. Os mecanismos que levam à diversificação das categorias superiores à espécie na hierarquia taxonômica constituem a macroevolução [7].

2.2 Entendendo uma árvore filogenética

A grande variabilidade genética e, consequentemente, a grande biodiversidade existente atualmente, são frutos do acúmulo de mudanças ocorridas no conteúdo do DNA, as quais se denominam mutações. Através das mutações no genoma, que são incorporadas e transmitidas ao longo das gerações nas populações, ocorre a evolução das espécies [4].

De acordo com Beneti, Montesinos e Tarfino[10], ao longo do tempo, este acúmulo de mutações, torna as sequências de DNA de diferentes espécies divergentes entre si em maior ou menor grau, embora continuem sendo homólogas, por apresentarem uma origem ancestral comum. Ainda segundo os mesmos autores, diferenças entre as bases em posições homólogas do DNA podem ser caracteres indicativos que permitem gerar hipóteses de relações de parentesco entre as espécies.

Uma árvore filogenética nada mais é do que um grafo que representa as relações evolutivas hipotéticas de ancestralidade entre organismos ou sequências genéticas[5]. Um exemplo genérico de uma árvore filogenética pode ser observado na Figura 2.

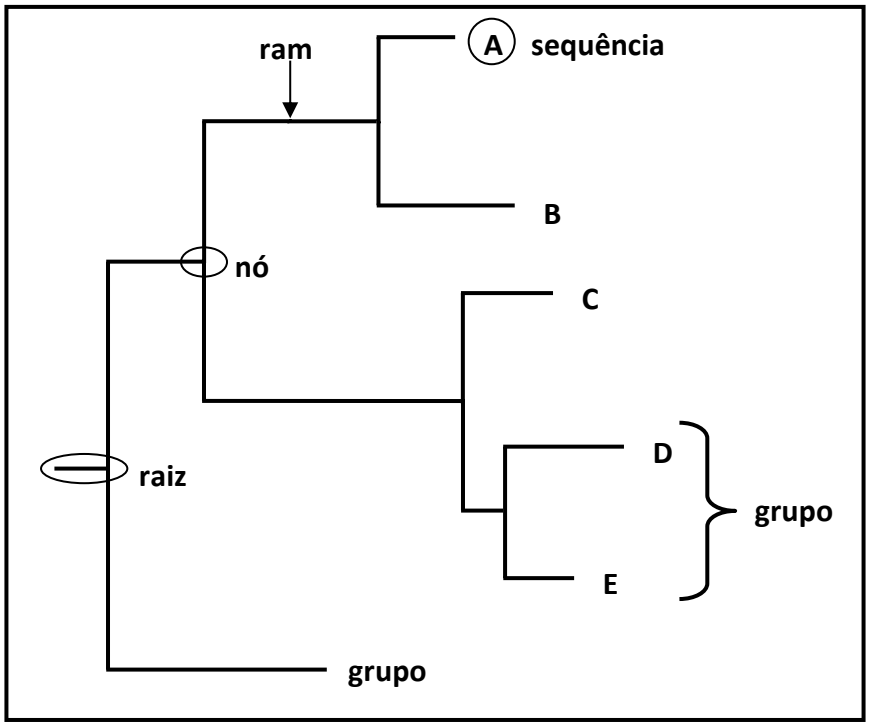


Figura 2. Representação de uma árvore filogenética enraizada (Adaptado de Caldart et al. [5].

Segundo Monteiro e Ursi [11], ao observar-se esta representação, nota-se que a raiz da árvore representa a linhagem ancestral, e as terminações das ramificações, ou nós terminais, representam os descendentes desse ancestral. Portanto, conforme se caminha da raiz para as pontas se está progredindo no tempo. Os nós terminais são denominados unidades taxonômicas operacionais (OTUs) e representam os dados amostrais que foram utilizados para a construção da árvore. Quando se trabalha com filogenia molecular, são as sequências de nucleotídeos ou proteínas que estão nas pontas das ramificações [3].

O enraizamento de uma árvore é um processo muito importante, pois, em árvores não enraizadas, não é possível identificar qual nó representaria o ancestral comum das demais sequências. Para realizar o devido enraizamento deve-se, obrigatoriamente, incluir na análise filogenética um grupo externo, que deve ser formado por uma ou mais OTUs que estão distantemente relacionadas às OTUs de interesse na análise. Ao escolher um grupo externo, deve-se ter em mente que este não pode estar relacionado ao grupo interno nem de forma muito distante nem de forma muito próxima, pois isto prejudicaria, respectivamente, a topologia da árvore e a representatividade do grupo externo. A utilização de mais de um grupo externo é uma opção para a melhoria da topologia da árvore [4].

Os eventos evolutivos podem ser aqueles de especiação, caso a OTU seja população, ou aqueles de duplicação de genes, caso a OTU seja um gene ou proteína, e são representados na árvore através dos nós internos. Os nós representam eventos de divergência ou divisão de um único grupo em dois grupos descendentes e indicam o ancestral comum mais recente de todos os grupos que se originam desse ponto de ramificação. Os nós internos de uma árvore filogenética dão origem aos ramos que, dependendo da forma de seus arranjos podem definir diferentes grupos [12].

Quando da análise do relacionamento entre as OTUs de um cladograma, duas OTUs serão mais relacionadas entre si quando tiverem um ancestral comum mais recente e menos relacionadas se tiverem um ancestral comum menos recente. Para

se encontrar o ancestral comum mais recente de qualquer par ou grupo de OTUs basta iniciar a análise pelas extremidades dos ramos onde estão os alvos da comparação e voltar pelos ramos da árvore até encontrar o ponto em que as linhas das espécies convergem [13].

Conforme Sakamoto [3], uma distinção importante que se deve fazer é acerca dos conceitos de árvore de espécie e árvore de gene. Enquanto a árvore de espécies representa uma árvore de um conjunto de espécies, possuindo uma topologia verdadeira, refletindo os eventos evolutivos ocorridos nas espécies em análise, a árvore de genes representa um conjunto de genes homólogos, inferidos a partir de dados moleculares, demonstrando um evento específico na evolução dos genes.

Deve-se ter em mente que uma árvore de gene demonstra um evento específico na evolução daquele gene em questão e que cada gene por si só pode contar diferentes histórias evolutivas para as espécies em análise. Logo, pode haver discordância entre as árvores inferidas pela análise de diferentes genes assim como entre a árvore de gene e a árvore de espécie. Uma forma de se construir árvores de espécies mais confiáveis, a partir da análise de genes, é a utilização de múltiplos genes dentro de uma análise conjunta [14].

2.3 Construindo uma árvore filogenética com dados moleculares

2.3.1 Selecionando os dados moleculares

Os marcadores moleculares são sequências de DNA que revelam variações entre indivíduos geneticamente relacionados através da identificação de polimorfismos. A seleção do melhor marcador a ser utilizado dependerá dos objetivos da análise [6].

Para complementar, Buso [6] afirma que para a construção de inferências filogenéticas por meio de comparações entre sequências de DNA deve-se utilizar sequências de genes homólogos, ou seja, genes que, por conta da sua ancestralidade em comum, apresentem alto grau de similaridade. Quando uma determinada característica é herdada de um ancestral comum tem-se o que é denominado de homologia. Conforme Caldart et al. [5], genes homólogos compartilhados por diferentes espécies que se originaram de um ancestral comum, sem duplicação ou transmissão horizontal, são chamados ortólogos.

Quando se analisam relações em níveis inferiores, como espécies e gêneros, deve-se dar preferência aos marcadores de rápida evolução. Para questionamentos em níveis taxonômicos elevados como famílias e filos recomenda-se a utilização de marcadores de evolução lenta. Já quando se deseja estudar as relações filogenéticas em diferentes níveis taxonômicos em uma análise conjunta, deve-se utilizar mais de um marcador molecular com diferentes taxas de divergência [10].

Quando se tem por objetivo o estudo da diversidade das comunidades microbianas, frequentemente utilizam-se análises de uma região conservada do genoma bacteriano denominada 16S rRNA (gene 16S rRNA). O gene 16S rRNA codifica para a subunidade ribossômica menor, que é parte do sítio de ocorrência da síntese protéica, estando presente em todas as bactérias. Esse gene se caracteriza por apresentar alto nível de conservação ao longo da evolução podendo ser utilizado como marcador das relações evolutivas dos microrganismos [15].

De acordo com Amann, Ludwing e Schleifer [16], características como, grande quantidade de rRNA na maioria das células, aparente falta de transferência lateral de genes, comprimento adequado (cerca de 1500 nucleotídeos) e, principalmente, a disponibilidade de robustos bancos de dados públicos, contribuíram muito para que o gene 16S rRNA tenha se tornado opção recorrente em estudos filogenéticos.

2.3.2 Conhecendo o gene 16S rRNA

O ribossomo é constituído de duas subunidades não idênticas, sendo uma subunidade menor (*30S*) e uma subunidade maior (*50S*). A subunidade menor consiste em uma molécula de 16S rRNA e 21 proteínas diferentes, enquanto que a subunidade maior contém uma molécula de 5S rRNA, uma molécula de 23S rRNA e 31 proteínas diferentes [17].

Como se pode observar na Figura 3, uma característica importante que o gene 16S rRNA possui, que o torna útil na determinação de relações filogenéticas, é que ele é constituído por regiões de sequências conservadas intercaladas com sequências variáveis que incluem 9 regiões hipervariáveis (V1-V9) [18].

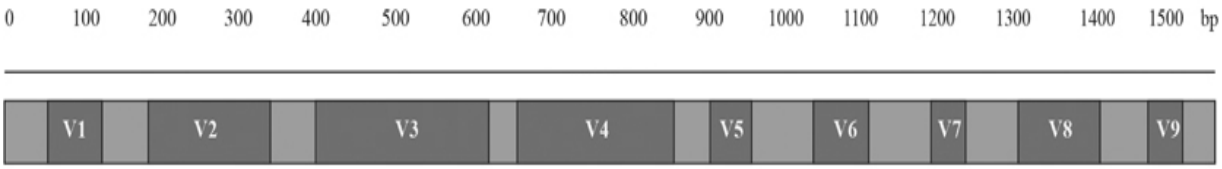


Figura 3. Representação conceitual do gene 16S rRNA com regiões conservadas (cinza claro) e hipervariáveis (cinza escuro) (Adaptado de Singer et al. [19]).

Segundo Machado [20], estas regiões variáveis podem ser usadas para se estabelecer relações evolutivas entre dois indivíduos muito próximos, como cepas dentro de espécies particulares, ao passo que as regiões conservadas podem ser usadas para elucidar relações mais antigas como a identificação de espécies.

2.3.3 Alinhando as sequências de nucleotídeos

O principal objetivo para a construção de um cladograma, utilizando sequências de nucleotídeos de organismos de interesse, é o de propor uma hipótese filogenética para o grupo. Estas sequências de nucleotídeos podem ser geradas por análises durante o estudo em questão ou serem obtidas de bancos de dados, devendo ser sistematizadas em uma única matriz de dados [10].

Segundo Caldart et al. [5], o primeiro pa sso para a organização destes dados é a realização do alinhamento entre estas sequências, permitindo, desta forma, a comparação entre os sítios homólogos das bases presentes nas mesmas. A Figura 4 apresenta uma matriz a partir do software Mega com as sequências dispostas nas linhas e os sítios nas colunas.

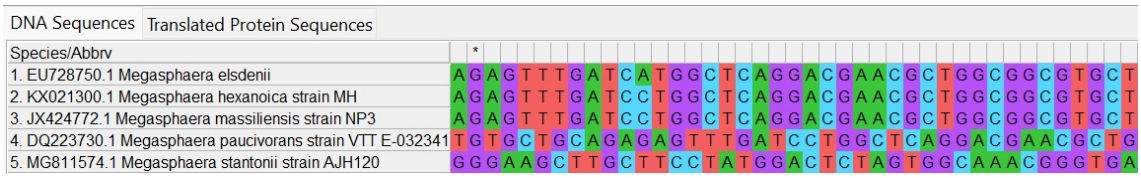


Figura 4. Comparação de sequências homólogas. (Fonte: Autores).

Durante o alinhamento de regiões homólogas pode-se identificar a existência de substituições de nucleotídeos (*cinza claro*) e de alguns eventos evolutivos como inserções ou deleções de nucleotídeos nas sequências (*cinza escuro*).

Embora esses eventos possam ser identificados, é difícil saber se uma das sequências ganhou o nucleotídeo ou se a outra sequência o perdeu. Devido a esta incerteza, estes eventos, denominados "*indels*", são normalmente representados por um traço na matriz. Estes traços, denominados *gaps,* são adicionados nas sequências

com a finalidade de ajustar o alinhamento entre sítios idênticos ou similares, minimizando a diferença entre eles [10].

A obtenção dos alinhamentos corretos entre as seqüências nucleotídicas é crucial para a construção de uma árvore filogenética confiável. Este alinhamento é conseguido através da utilização de programas de computadores que usam algoritmos particulares que comparam a similaridade entre as sequências, alinhando-as a partir das que apresentam maior similaridade adicionando as outras sequências progressivamente [4].

2.3.4 Métodos de inferência da árvore filogenética

Existem diversos métodos que podem ser utilizados para a construção de uma árvore filogenética e cada um deles possui suas particularidades. Tendo em mente que a árvore filogenética que será construída a partir do alinhamento de sequências de nucleotídeos é uma inferência, ela pode ou não coincidir com a árvore filogenética verdadeira [4].

Horiike [21] cita que podemos separar os métodos de inferência de árvores filogenéticas em dois grupos principais, de acordo com o tipo de dados que são utilizados para a inferência, sendo eles os métodos baseados em matrizes de distância e os métodos baseados em estados de caracteres.

De acordo com Gainett, Dias e Montesino [22], os métodos baseados em distâncias buscam reconstruir as relações filogenéticas em função da medida de dissimilaridade entre os pares de sequências presentes no alinhamento. As sequências são comparadas par a par e a distância global entre os táxons é estimada matematicamente. Estes métodos têm por característica serem computacionalmente rápidos e gerarem apenas uma árvore. São exemplos deste tipo de metodologia o método da média aritmética não ponderada (UPGMA) e o método de agrupamento de vizinhos (NJ) [3].

Métodos de distância em geral não apresentam robustez e podem de maneira frequente gerar árvores não confiáveis. Atualmente estes métodos não são considerados como os ideais para se estimar filogenias, embora, sejam úteis em outros contextos [22].

Já os métodos que utilizam algoritmos baseados em estados de caracteres, trabalham gerando e avaliando inúmeras topologias hipotéticas de árvores filogenéticas e, através de um sistema de pontuação, escolhe aquela que melhor se ajusta aos dados do alinhamento. Atualmente estes métodos têm sido os mais utilizados em estudos filogenéticos e, entre os métodos que utilizam esta abordagem, pode-se citar o método da máxima parcimônia (MP), o método da máxima verossimilhança (MV) e a inferência bayesiana (IB) [6].

Vale destacar que a escolha do método, assim como dos programas computacionais, é um processo que depende dos objetivos do autor, não existindo um método ou programa mais ou menos indicado. No entanto, a comparação entre árvores obtidas por diferentes métodos de inferência não deve ser realizada, visto que seus pressupostos analíticos não são equiparáveis [22].

Na tabela 1 estão listados os principais métodos de inferência de árvores filogenéticas e os softwares mais utilizados para a realização desses procedimentos.

Tabela 1. Lista de métodos para inferência de árvores filogenéticas

Método	Grupo	Algoritmo	Software
UPGMA	Matriz de distância	Agrupamento para a menor distância evolutiva	MEGA 7
NJ	Matriz de distância	Agrupamento para o menor comprimento total da ramificação	PHYLIP, Clustal X, MEGA 7
MP	Estado de caracteres	Busca pela árvore com menor número total de alterações no estado do caractere	PHYLIP, MEGA 7
MV	Estado de caracteres	Busca pela árvore com maior probabilidade	PHYLIP, PhyML, RAxML, FastTree, MEGA 7, TOPALi v2
IB	Estado de caracteres	Busca pela árvore com maior probabilidade posterior	MrBayes, TOPALi v2

UPGMA, método da média aritmética não ponderada; NJ, método de agrupamento de vizinhos; MP, método da máxima parcimônia; MV, método da máxima verossimilhança; IB, inferência bayesiana. Adaptado de Horiike [21].

Método da média aritmética não ponderada (UPGMA)

O método UPGMA [23] é o método original usado para reconstruir árvores filogenéticas usando matriz de distância evolutiva. Segundo Buso [6] o agrupamento é feito procurando-se, entre todas as OTUs estudadas as duas com a menor distância evolutiva na matriz de distância. O cluster recém-formado substitui as OTUs que representa na matriz de distância e são calculadas as distâncias entre o cluster recém-formado e cada uma das OTUs restantes. Este processo é repetido até que todas as OTUs estejam agrupadas.

De acordo com Horiike [21] o grande demérito da metodologia UPGMA é que esta assume que a taxa evolutiva do nó das duas OTUs em cluster para cada uma das duas OTUs seja idêntica, logo, todo o processo se baseia na pressuposição de que a taxa evolutiva é igual em todos os ramos, o que na maioria das vezes não é verdadeiro.

Método de agrupamento de vizinhos (NJ)

O método NJ [24] baseia-se na busca de pares de OTUs vizinhas que minimizem o comprimento total de ramificação, e consequentemente da árvore, em cada estágio do agrupamento de OTUs. Este método não se preocupa em agrupar as OTUs mais proximamente relacionadas mas sim em inferir o menor ramo possível.

De acordo com Pinto [4], o algoritmo se inicia com uma árvore semelhante a uma estrela, sem ramos internos. No primeiro momento, introduz o primeiro ramo interno e calcula o tamanho da árvore resultante. O algoritmo continua ligando os possíveis pares de OTUs e, no final, junta o par que leva à menor árvore. O processo é assim repetido sempre juntando dois pares vizinhos de OTUs baseando-se no menor ramo interno possível.

O mérito do método NJ é que ele é rápido sendo, portanto, prático para análises que contenham um grande conjunto de dados. Caso o número de táxons não seja grande, outros métodos que não se baseiam na distância evolutiva substituem o método NJ. Sabe-se também que a precisão do método NJ é menor no caso de sequências de DNA muito curtas [21].

Método da máxima parcimônia (MP)

A MP é a pioneira dos métodos baseados em caracteres e foi desenvolvido por Henning [25] usando dados morfológicos, sendo mais tarde adaptada para a utilização de dados de nucleotídeos por Fitch [26]. A MP baseia-se no conceito de que as hipóteses mais simples são mais adequadas do que as mais complexas. O algoritmo busca, dentre as diferentes topologias possíveis das árvores, encontrar aquela que, para um grupo de seqüências alinhadas, possa ser explicada com o mínimo de substituições de caracteres. Este processo é feito através do cálculo da probabilidade da esperança de cada nucleotídeo no nó ancestral [3].

Para Horiike [21] a principal desvantagem da MP é o fato dela supor que um caractere comum seja obrigatoriamente de um ancestral comum, o que acaba por subestimar a divergência real entre táxons distantemente relacionados.

Método da máxima verossimilhança (MV)

O método MV é um método estatístico baseado em caracteres capaz de inferir um grande número de árvores diferentes e estimar, para cada uma delas, a probabilidade de representarem a árvore filogenética verdadeira [27]. Este processo é realizado através de cálculos que utilizam modelos de substituição sendo que, o algoritmo determinará a árvore que apresenta maior valor de probabilidade, ou seja, supõe-se que a inferência com maior probabilidade de refletir os dados observados é preferível àquela com menor probabilidade [6].

Gainett, Dias e Montesinos [22] cita que esse método se destaca por incluir modelos evolutivos previamente conhecidos, o que maximiza o conhecimento da realidade e aumenta consideravelmente a possibilidade de optar-se por aquela árvore que represente de maneira mais fiel o processo evolutivo dos dados observados. A maior desvantagem deste método segundo Pinto [4], é que, quanto maior for o número de sequências incluídas na análise, maior será o tempo e a atividade computacional requeridas para executá-la, o que muitas vezes pode tornar o método inviável.

Inferência bayesiana (IB)

Desde que o reverendo e estudioso Thomas Bayes propôs o seu teorema no século XVIII, os métodos bayesianos são largamente utilizados em uma série de estudos. A IB começou a ser utilizada com maior frequência em estudos de Sistemática Filogenética somente a partir da década de 1990 [22].

A IB é um método de estado de caráter que se baseia em probabilidades posteriores sob o melhor modelo estimado, a fim de fornecer um conjunto de árvores filogenéticas plausíveis que permite a escolha daquela que possui maior probabilidade. Para isso, este método utiliza um conjunto de observações, um modelo de evolução para essas observações e as demais probabilidades associadas, que devem ser previamente estabelecidos (*priors*). As probabilidades posteriores e as diferentes topologias que a árvore pode assumir são obtidas utilizando-se o algoritmo Markov Chain Monte Carlo (MCMC) [3].

De acordo com Horiike [21], a abordagem bayesiana tornou-se um dos métodos mais utilizados atualmente devido aos avanços na velocidade e capacidade de processamento computacional. Mesmo assim a IB enfrenta algumas críticas relacionadas principalmente aos *priors* e ao possível impacto negativo que podem ocasionar no resultado geral da análise se estes forem estabelecidos de maneira inadequada.

2.3.5 Teste de Confiança em Topologias

Deve-se ter em mente que o processo de reconstrução filogenética se caracteriza por ser uma inferência, uma estimativa pontual da filogenia, ou seja, não se tem plena certeza de quão robustos são os dados que dão suporte às relações evolutivas representadas no cladograma. Desta forma, faz-se necessária a avaliação da confiança no suporte de cada nó da topologia escolhida [5].

O método mais comumente empregado para esta finalidade é conhecido como *bootstrapping* [28]. De acordo com Pinto [4] a análise *bootstrap* é uma técnica que estima os erros estatísticos em ocasiões em que não se conhece a distribuição na amostra original ou esta é de difícil derivação analítica mediante repetição da análise filogenética sobre réplicas do alinhamento original.

Para cada dado de alinhamento artificial, uma réplica da árvore é construída e a proporção de cada ramificação interna para todas as árvores artificiais é calculada. Após este processo realiza-se uma comparação entre cada uma das réplicas e a árvore original, sendo que o valor de *bootstrap* obtido (Figura 5) corresponde ao percentual de vezes em que os grupamentos das árvores artificiais coincidem com os da árvore original [21].

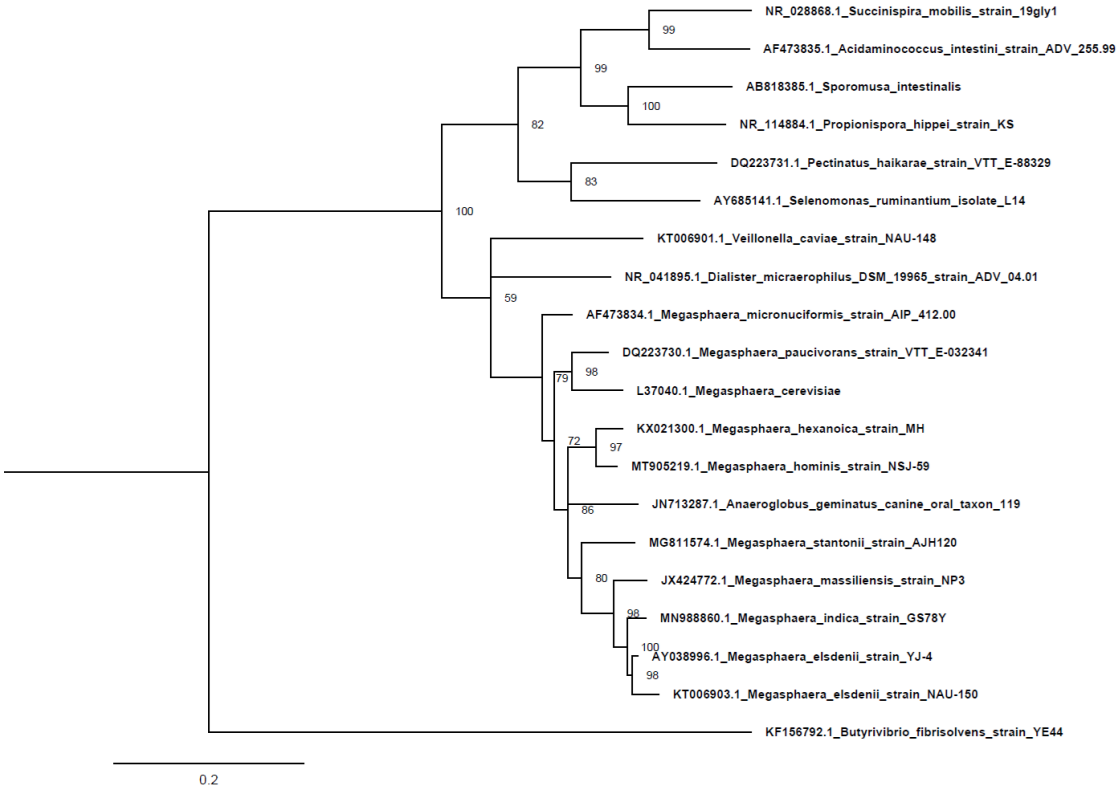


Figura 5. Exemplo de uma árvore filogenética com valores de bootstrap (Fonte: Autores)

Ainda não existe um consenso sobre o valor mínimo aceitável da proporção de *bootstrap,* mas pode-se dizer que quanto maior for a proporção maior será a confiança na topologia daquele nó. Trabalhos como os de Zharkikh e Li [29] e Hillis e Bull [30] concluíram que um ponto de corte razoável seria de 70%. Já segundo Horiike [21], empiricamente, quando o valor de *bootstrap* é igual a 90% ou mais, o ramo interno pode ser considerado confiável.

4. Conclusão

É inegável que haja crescente importância na utilização de dados moleculares como fonte de informações para a construção de inferências filogenéticas. O número de sequências gênicas depositadas em inúmeros bancos de dados aumenta a cada dia e mesmo microrganismos que ainda não foram cultivados pelos métodos tradicionais possuem informações moleculares disponíveis. Este cenário abre caminho para uma infinidade de possibilidades de estudos e análises. Porém, o conhecimento das ferramentas de processamento e interpretação desses dados é vital para o sucesso de todo este processo.

Muito se engana aquele que pensa que a inferência filogenética é um processo simples que pode ser resumido a apenas alguns comandos computacionais. Deve-se ter em mente que, na construção de árvores filogenéticas, todas as ações necessitam ser muito bem planejadas e avaliadas, desde a escolha do grupo a ser estudado e do melhor método de inferência, até a definição de quais algoritmos e programas estarão envolvidos nas análises, tendo como objetivo principal obter hipóteses que expliquem, de forma coerente e precisa, as evidências observadas.

Referências

1. SANTOS, C.M.D.; CALOR, A.R. Ensino de biologia evolutiva utilizando a estrutura conceitual da sistemática filogenética – I. **Ciência & Ensino**, v.1, p.1-8, 2007.

2. LOPES, S.G.B.C.; CHOW, F. Noções básicas de sistemática filogenética. In: LOPES, S. G. B. C.; CHOW, F.; LAHR, D. J. G.; TURRINI, P. **Diversidade biológica, História da vida na Terra e Bioenergética.** São Paulo, USP/Univesp/Edusp, p.54-67, 2014.

3. SAKAMOTO, T. **Ferramentas para análise filogenética e de distribuição taxonômica de genes ortólogos.** 2016. 112 p. Tese (Doutorado em Bioinformática) – Programa de Pós-Graduação em Bioinformática, Universidade Federal de Minas Gerais, Belo Horizonte, 2016.

4. PINTO, J.F.C. **Epidemiologia molecular do vírus da Imunodeficiência humana do tipo i: métodos de Inferência filogenética.** 2004. 69 p. Dissertação (Mestrado em Saúde Pública) – Escola Nacional De Saúde Pública Sérgio Arouca, Rio de Janeiro, 2004.

5. CALDART, E.T.; MATA, H.; CANAL, C.W., RAVAZZOLO, A.P. Análise filogenética: conceitos básicos e suas utilizações como ferramenta para virologia e epidemiologia molecular. **Acta Scientiae Veterinariae**, v.44, p.1-20, 2016.

6. BUSO, G.S.C. **Marcadores moleculares e análise filogenética.** Brasília: Embrapa Recursos Genéticos e Biotecnologia, 2005. 22 p.

7. MAZZAROLO, L.A. Conceitos básicos de sistemática filogenética. 2005. Disponível em: <http://www.mzufba.ufba.br/WEB/Ensino_Arquivos/Mazzarolo_Apostila.pdf>, Acesso em 15/04/2020.

8. MOUTINHO, W.T. Sistemática Filogenética. 2020. Disponível em: <https://www.coladaweb.com/biologia/reinos/sistematica-filogenetica>. Acessado em: 17 de abril de 2020.

9. COLLEY, E.; FISCHER, M.L. Especiação e seus mecanismos: histórico conceitual e avanços recentes. **História, Ciências, Saúde – Manguinhos**, v.20, p.1671-1694, 2013.

10. BENETI, J.S.; MONTESINOS, R.; TARFINO, M. Sistemática filogenética baseada em dados moleculares. In: BENETI, J.S.; MONTESINOS, R.; GIOVANNETTI, V. (Org.). **Tópicos de Pesquisa em Zoologia.** 1ed. São Paulo: Instituto de Biociências, v.1, p.84-102, 2017.

11. MONTEIRO, E.C.; URSI, S. Introdução à Filogenética para Professores de Biologia. 2011. Disponível em: <http://www2.ib.usp.br/index.php?option=com_docman&task=doc_download&gid=59&Itemid=98>, Acesso em: 08 de abril de 2020.

12. BAUM, D. Reading a phylogenetic tree: The meaning of monophyletic groups. **Nature Education**, v.1, p.190, 2008.

13. KHAN ACADEMY. Phylogenetic trees. Disponível em: < https://www.khanacademy.org/science/high-school-biology/hs-evolution/hs-phylogeny/a/phylogenetic-trees>, Acesso em 08 de abril de 2020.

14. MADDISON, W.P. Gene trees in species trees. **Systematic Biology**, v.46, p.523-536, 1997.

15. SILVEIRA, E.L. **Identificação de comunidades bacterianas de solo por seqüenciamento do gene 16S rRNA.** 2004. 83p. Dissertação (Mestrado em Microbiologia) – Programa de Pós-Graduação em Microbiologia Agropecuária, Universidade Estadual Paulista, Jaboticabal, 2004.

16. AMANN, R.I.; LUDWING, W.; SCHLEIFER, K.H. Ribossomal RNA-targeted nucleic acid probes for studies in microbial ecology. **FEMS Microbiological Reviews**, v.24, p.555-565, 2000.

17. JESUS, R.B. **Diversidade bacteriana ruminal em bovinos nelore.** 2014. 33 p. Dissertação (Mestrado em Zootecnia) – Programa de Pós-Graduação em Zootecnia, Universidade Estadual Paulista, Jaboticabal, 2014.

18. PETROSINO, J.F.; HIGHLANDER, S.; LUNA, R.A.; GIBBS, R.A.; VERSALOVIC, J. Metagenomic pyrosequencing and microbial identification. **Clinical Chemistry**, v.55, p.856-866, 2009.

19. SINGER E.; BUSHNELL, B.; COLEMAN-DERR, D.; BOWMAN, B.; BOWERS, R.M.; LEVY, A.; GIES, E.A.; CHENG J.; COPELAND, A.; KLENK, H.; HALLAM, S.J.; HUGENHOLTZ, P.; TRINGE, S.G.; WOYKE, T. High-resolution phylogenetic microbial community profiling. **The ISME Journal**, v.10, p.2020-2032, 2016.

20. MACHADO, J.B. A. **Uso da biblioteca genômica RNAr 16S como ferramenta para o estudo da microbiota fecal humana.** 2013. 81 p. Dissertação (Mestrado em Farmácia) – Programa de Pós-Graduação em Farmácia, Universidade de São Paulo, São Paulo, 2013.

21. HORIIKE, T. An introduction to molecular phylogenetic analysis. **Reviews in Agricultural Science**, v.4, p.36- 45, 2016.

22. GAINETT, G.; DIAS, P.H.S.; MONTESINOS, R. Metodologia da inferência filogenética. In: BENETI, J.S.; MONTESINOS, R.; GIOVANNETTI, V. (Org.). **Tópicos de Pesquisa em Zoologia.** 1ed. São Paulo: Instituto de Biociências, v.1, p.67-83, 2017.

23. SOKAL, R.; MICHENER, C. A statistical method for evaluating systematic relationships. **University of Kansas Science Bulletin**, v.38, p.1409-1438, 1958.

24. SAITOU, N.; NEI, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. **Molecular Biology and Evolution**, v.4, p.406-25, 1987.

25. HENNIG, W. **Phylogenetic Systematics.** Urbana, University of Illinois Press, 1966. 263 p.

26. FITCH, W.M. Toward defining the course of evolution: minimum change for a specified tree topology. **Systematic Zoology**, v.20, p.406-416, 1971.

27. FELSENSTEIN, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. **Journal of Molecular Evolution**, v.17, p.368-376, 1981.

28. EFRON, B. Bootstrap methods: another look at the jackknife. **The Annals of Statistics,** v.7, p.1-26, 1979.

29. ZHARKIKH, A.; LI, W.H. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. **Journal of Molecular Evolution,** v.9, p.1119-1147, 1992.

30. HILLIS, D.M.; BULL, J.J. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. **Systematic Biology,** v.42, p.182-192, 1993.