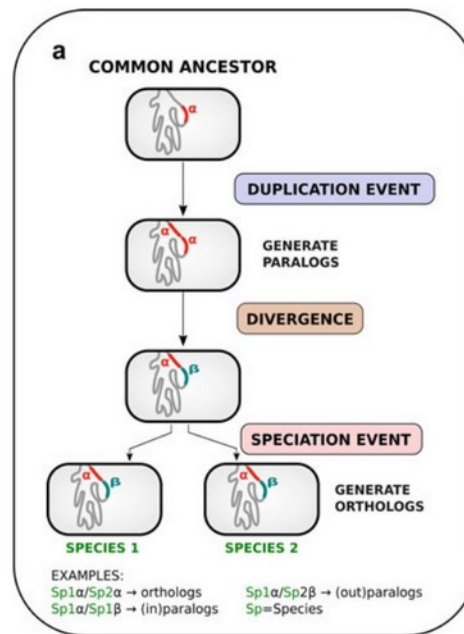


Phylogenetic analysis

Leonardo Magalhães Cruz

Especiação vs. Duplicação gênica



Cruz LM et al. Ch5 – Protein Function Prediction. Springer 2017 doi: 10.1007/978-1-4939-7231-9_5

Eventos de duplicação gênica podem gerar cópias de um gene em um genoma. Estas cópias do gene podem evoluir “independentemente”, dando origem a genes diferentes (ex., que codificam para proteínas com funções diferentes). Como estas cópias do gene surgiram a partir de um único gene ancestral, são homólogas e, por surgirem por duplicação gênica, são **homólogas parálogas**. Eventos de especiação geram novas espécies a partir de uma espécie ancestral. Se genes parálogos existem no genoma ancestral, serão herdados pelas novas espécies e os mesmos genes em ambas as espécies são **homólogos ortólogos**.

Como podemos dizer se duas sequências são homólogas?

A homologia frequentemente se manifesta através da **similaridade** significativa entre as sequências de nucleotídeos e proteínas ou através das estruturas tridimensionais das proteínas

Similaridade

Extensão pela qual sequências de nucleotídeos ou proteínas são relacionados (usando um critério de similaridade arbitrário)

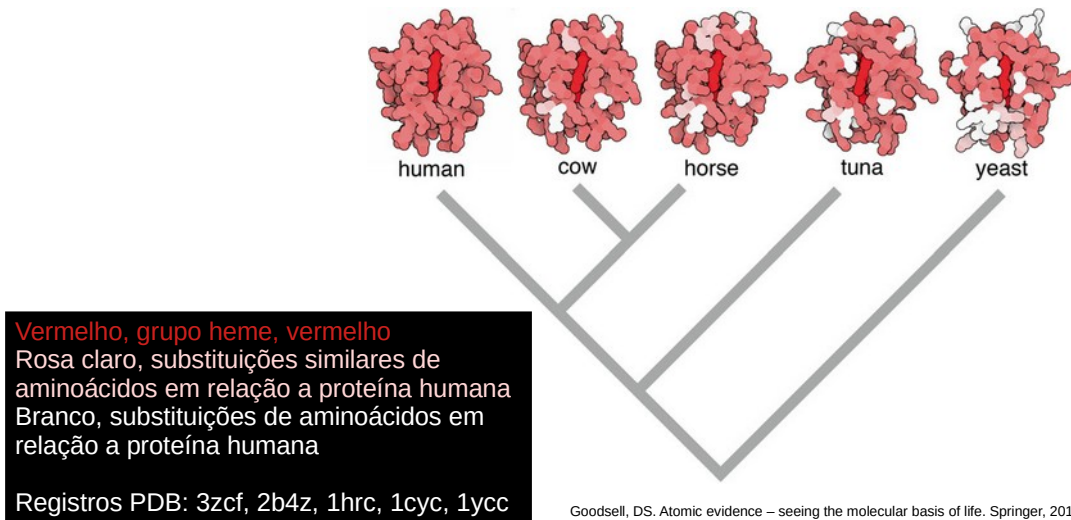
Identidade

Extensão pela qual duas sequências (nucleotídeos ou aminoácidos) são invariáveis

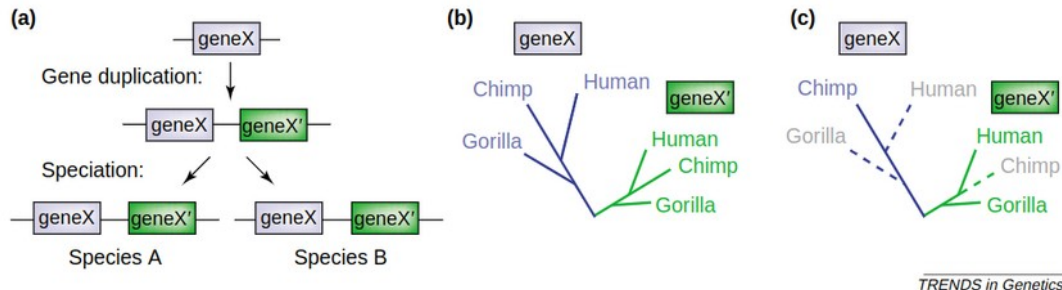
A homologia entre proteínas é mais facilmente detectada através da estrutura tridimensional, o aspecto de uma molécula mais próximo da sua função.

Evolução do citocromo c

A estrutura de proteínas é mais conservada que a sequência de aminoácidos



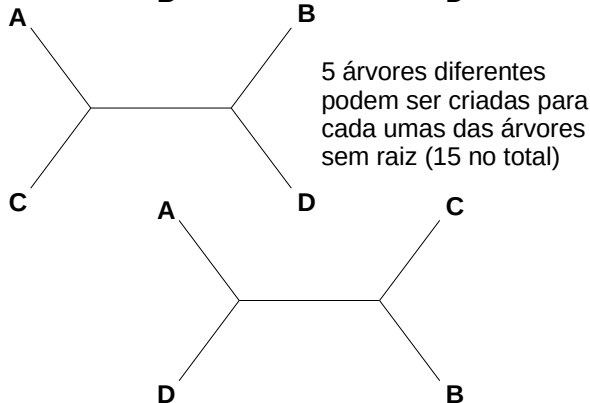
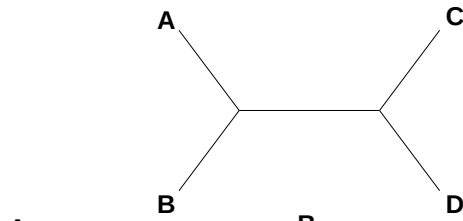
Como a homologia pode interferir na filogenia



Da mesma forma que eventos de duplicação gênica e especiação podem ocorrer ao longo da evolução, eventos de **perda de genes** também ocorrem. Estes eventos podem tornar a filogenia de um grupo de organismos, genes ou proteínas mais complexo, dificultando a identificação de genes ortólogos e parálogos.

Número de árvores possíveis

Para 4 taxa há 3 árvores sem raiz possíveis



5 árvores diferentes
podem ser criadas para
cada uma das árvores
sem raiz (15 no total)

nº. de árvores **sem** raiz possíveis

$$\frac{(2s-5)!}{2^{s-3}(s-3)!}$$

nº. de árvores **com** raiz possíveis

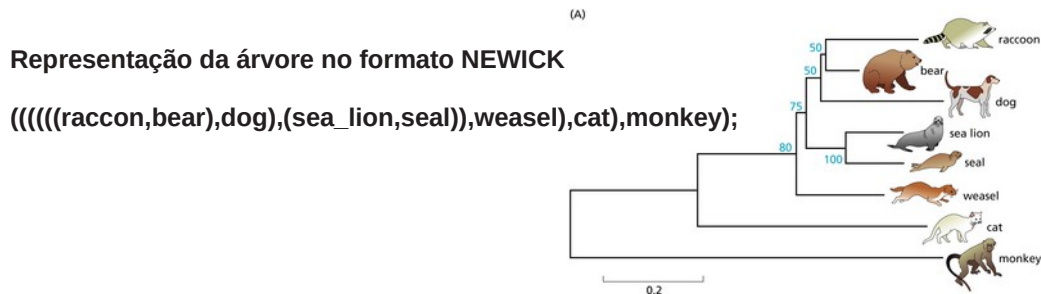
$$\frac{(2s-3)!}{2^{s-2}(s-2)!}$$

s = nº. de taxa

Taxa	sem raiz	com raiz
4	3	15
8	10.395	135.135
10	2.027.025	34.459.425
22	3×10^{23}	
50	3×10^{74}	
100	2×10^{182}	

Alguns métodos para análise filogenética tentam identificar qual é a “melhor” árvore, dentre todas aquelas possíveis para o conjunto em estudo. Entretanto, se o número de organismos, genes ou proteínas é grande, não é possível analisar todas as árvores. A alternativa é usar critérios que otimizem a análise do maior número de árvores possíveis. Os algoritmos que fazem isso são conhecidos como heurísticos.

Representação de uma Árvore Filogenética

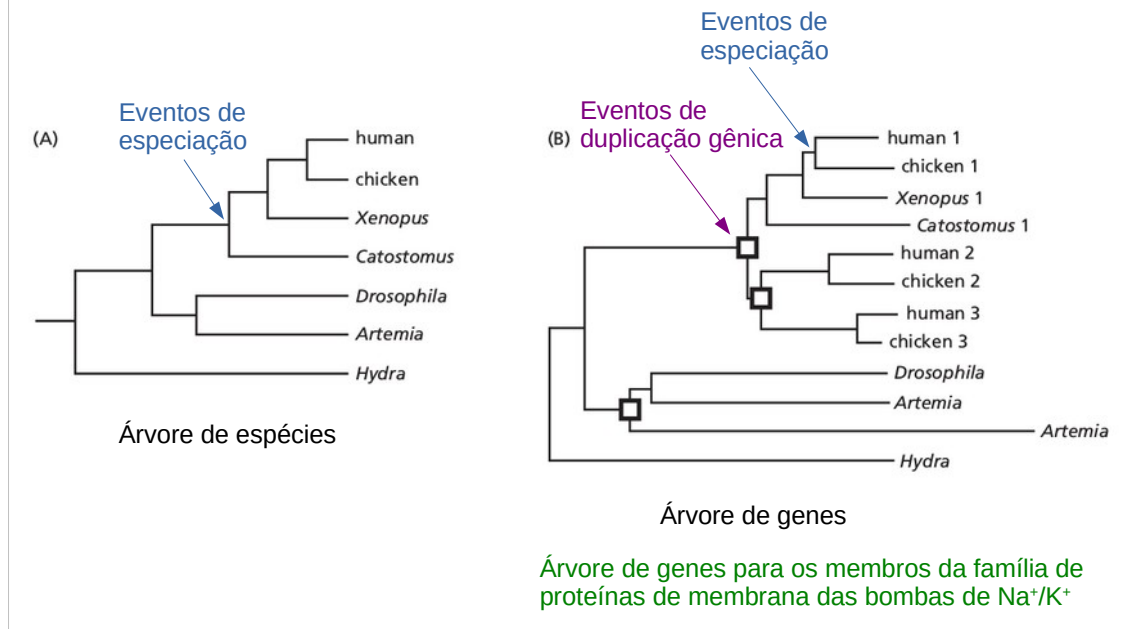


Valores para os comprimentos de ramos também podem ser incluídos

```
(((((raccon:0.20,bear:0.07):0.01,dog:0.25):0.03,  
(sea_lion:0.12,seal:0.12):0.08):0.03,weasel:0.18):0.02,cat:0.47):0.3,monkey:0.8);
```

O formato **newick** representa uma árvore filogenética usando parênteses para aninhar os grupos. Cada nó na árvore é bifurcado e os dois grupos presentes nesta bifurcação são unidos por um parênteses.

Árvore de espécies vs árvore de genes

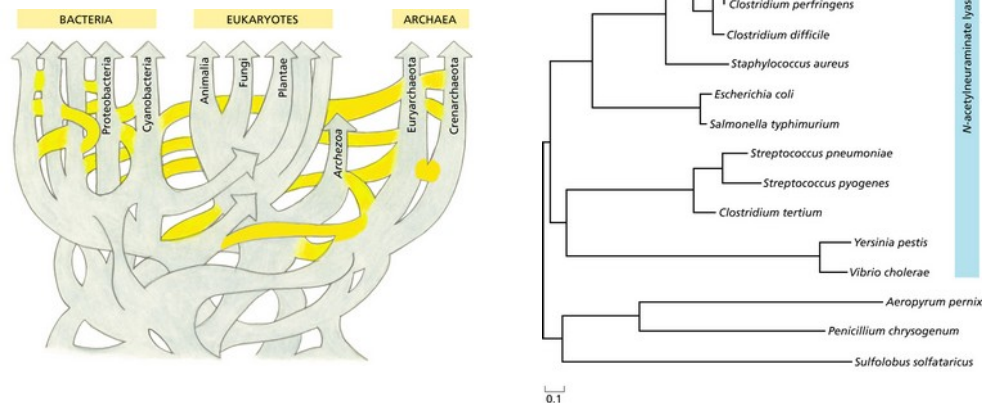


Uma árvore filogenética gerada para inferir a história evolutiva de um grupo de organismos irá representar somente eventos de especiação partindo de cada nó (que representa um organismo ancestral).

Por outro lado, uma árvore filogenética gerada para inferir a história evolutiva de um grupo de genes/proteínas irá conter tanto nós representando eventos de especiação, quanto nós representando eventos de duplicação gênica.

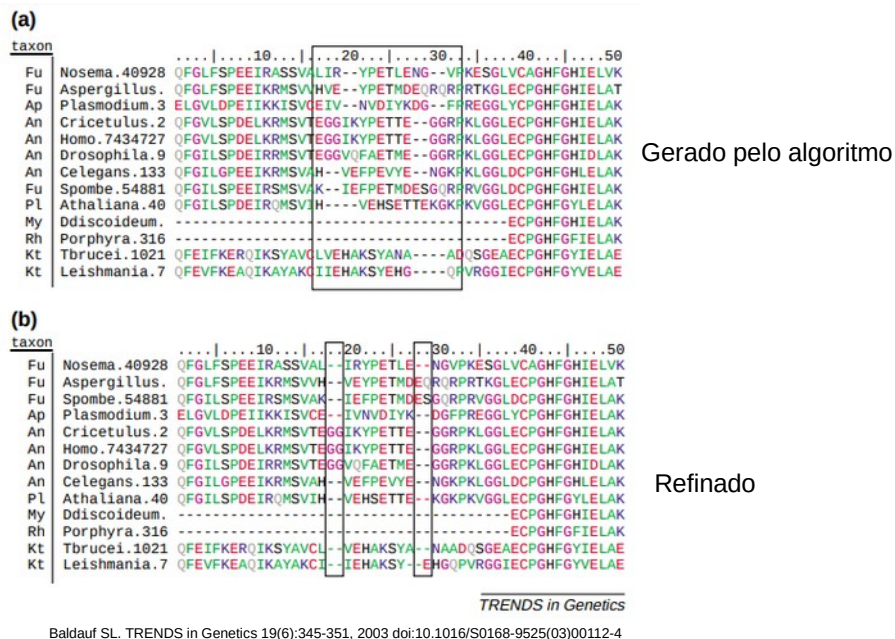
Transferência Horizontal de Genes

HGT – Horizontal Gene Transfer
LGT – Lateral Gene Transfer



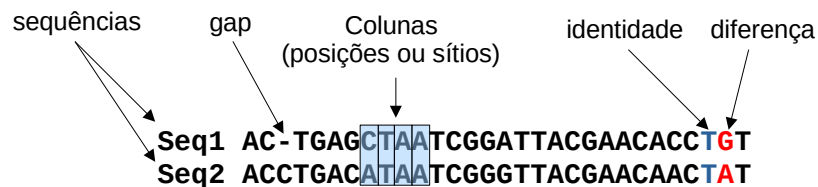
A HTG dificulta a interpretação da árvore filogenética porque, neste caso, a transmissão entre genes/proteínas não ocorreu através da descendência a partir de um ancestral comum.

A filogenia começa no alinhamento



Para um conjunto de sequências de nucleotídeos ou aminoácidos, podem existir regiões mais e outras menos conservadas. Nas regiões menos conservadas, há mais substituições e inserções/deleções entre as sequências comparadas, tornando o alinhamento entre as sequências mais difícil e mais sujeito a erros. É importante sempre avaliar o alinhamento gerado por um programa e editá-lo ou remover sítios duvidosos, caso se julgue necessário.

Alinhamento – comparando sequências



Um *gap* representa um deslocamento de uma sequência em relação a outra, incluído pelo programa para ajustar as sequências no alinhamento. Indica que houve uma deleção de base na **Seq1** ou então uma inserção de base na **Seq2**, ao longo do tempo de evolução das duas sequências. Entretanto, não é possível determinar se o evento ocorrido foi um deleção ou uma inserção e, neste caso nos referimos como sendo uma **indel**.

Cada coluna de bases no alinhamento representa um sítio contendo bases homólogas em todas as sequências.

Calculando similaridade/dissimilaridade

Quão semelhantes (próximas) são as sequências?

```
          10      20      30
          |      |      |
Seq1 AC-TGAGCTAATCGGATTACGAACACCTGT
Seq2 ACCTGACATAATCGGGTTACGAACAACATAT
      ** *** ***** ***** ** *
```

Identidade = $25/29 = 86\%$

Posições com *gaps* ignoradas

Quão diferentes (distantes) são as sequências?

A FILOGENIA USA DISTÂNCIAS!

```
          10      20      30
          |      |      |
Seq1 AC-TGAGCTAATCGGATTACGAACACCTGT
Seq2 ACCTGACATAATCGGGTTACGAACAACATAT
      *      *      *      *
```

Diferenças = $5/29 = 17\%$

Posições com *gaps* ignoradas

A partir do alinhamento entre sequências é possível determinar o número de bases idênticas (identidade) entre duas sequências ou o número de bases diferentes (distância ou dissimilaridade) entre duas sequências. As análises filogenéticas, usam as diferenças entre as sequências.

Distância p

Meio mais simples de estimar a distância evolutiva

Fração das posições não idênticas em um alinhamento entre duas sequências

Posições com *gaps*
são ignoradas

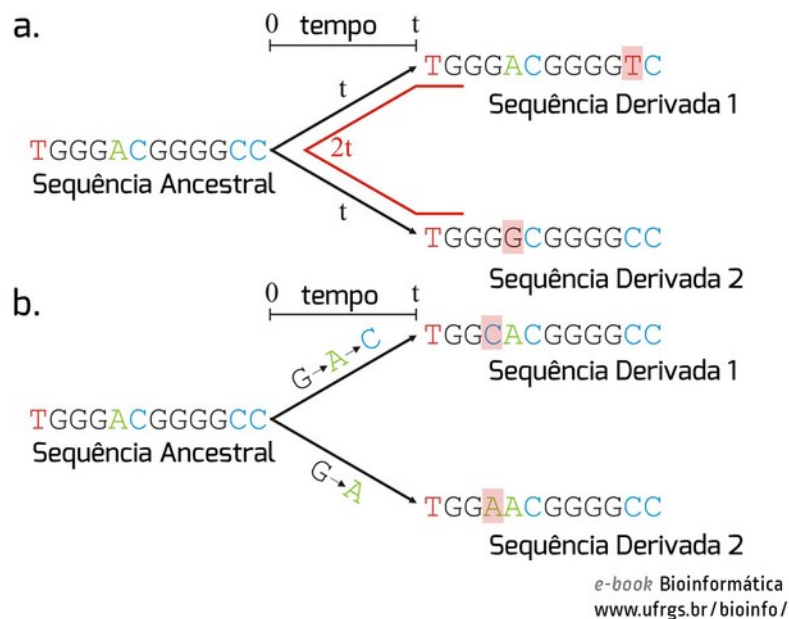
		10		20		30
Seq1	AC -	TGAGCTAATCGGATTACGAACACCTGT				
Seq2	AC	CTGACATAATCGGGTTACGAACA				ACTAT
		**		*		* *

$$\text{Distância } p = 5 / 29 = 0,172$$

Devido a possibilidade de ocorrer mutações sobrepostas, a distância p quase sempre subestima o número real de mutações

Correções são necessárias para melhorar a estimativa da distância evolutiva

Múltiplas substituições



Na figura **a**, a partir da **sequência ancestral**, é possível identificar uma única mutação para a **sequência derivada 1** e também para a **sequência derivada 2** (em posições diferentes). Isso é perceptível quando comparamos diretamente as duas sequências derivadas em um alinhamento.

Entretanto, na figura **b**, duas mutações ocorreram entre a **sequência ancestral** e a **sequência derivada 1** e na mesma posição que na **sequência derivada 2**, mas quando comparamos as duas sequências derivadas em um alinhamento, só é possível identificar uma única alteração entre elas.

Modelos de Substituição

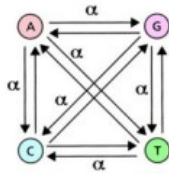
- **Modelos de substituição ou evolutivos**
 - Corrigem a distância p
 - Descrevem padrões de substituições de nucleotídeos/aminoácidos que ocorreram ao longo do processo evolutivo das sequências

Modelos de Substituição

- **Para substituições de nucleotídeos os modelos consideram**
 - Corrigem a distância p para múltiplas substituições
 - Diferenças nas taxas de transição (TS) e transversão (TV)
 - Frequência de bases
 - Condição de reversibilidade
- **Variação de taxa entre sítios**
- **Sítios invariáveis**

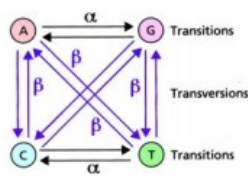
Modelos de Substituição

Jukes-Cantor
(1 parâmetro)



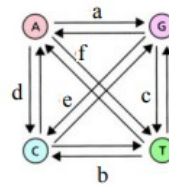
1. A frequência dos nucleotídeos não é considerada
2. Todas substituições têm a mesma taxa

Kimura 2P
(2 parâmetros)



1. A frequência dos nucleotídeos não é considerada
2. Transições e transversões têm diferentes taxas

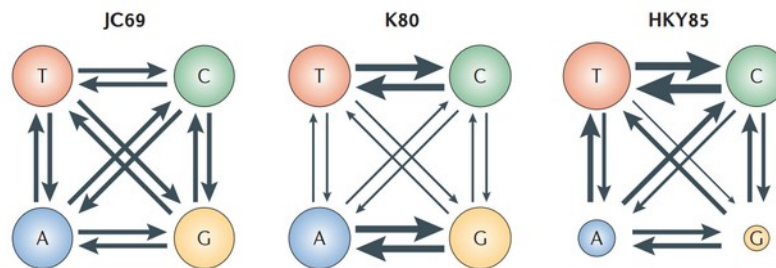
GTR (General Time Reversible)
(6 parâmetros)



1. A frequência dos nucleotídeos é considerada
2. Transições e transversões têm diferentes taxas, assim como os tipos de mudanças nucleotídicas

Identificação de modelos comumente usados: Jukes-Cantor JC ou JC69), Felsenstein 81 (F81), Kimura 2 parâmetros (K2P ou K80), HKY85, Tamura-Nei (TN), Kimura 3 parâmetros (K3P ou K81), SYM
Os modelos podem considerar ainda: a) variação na taxa de mutações ao longo dos sítios no alinhamento e b) presença de sítios invariáveis.

Modelos de Substituição



Yang, Z. & Rannala, B. Nature Reviews Genetics 13:303-314, 2012. Doi: 10.1038/nrg3186

Espessura da linha = taxas de substituição

Tamanho dos círculos = frequência dos nucleotídios

Devido a variação na taxa de mutação local e na restrição seletiva, sítios diferentes na sequência de DNA ou proteína frequentemente evoluem a taxas diferentes

Esta variação nas taxas são acomodadas assumindo uma distribuição gama (Γ) das taxas para os sítios no alinhamento

JC69 + Γ , HKY85 + Γ , GTR + Γ

Modelos de Substituição

Comparação entre modelos de substituição de nucleotídeos comumente usados em análise filogenética

Model name	Base composition	Different transition and transversion rates	All transition rates identical	All transversion rates identical	Reference
JC (JC69)	1:1:1:1	No	Yes	Yes	Jukes and Cantor (1969)
Felsenstein 81 (F81)	Variable	No	Yes	Yes	Felsenstein (1981)
K2P (K80)	1:1:1:1	Yes	Yes	Yes	Kimura (1980)
HKY85	Variable	Yes	No	No	Hasegawa et al. (1985)
Tamura-Nei (TN)	Variable	Yes	No	Yes	Tamura and Nei (1993)
K3P (K81)	Variable	Yes	No	Yes	Kimura (1981)
SYM	1:1:1:1	Yes	No	No	Zharkikh (1994)
REV (GTR)	Variable	Yes	No	No	Rodriguez et al. (1990)

Principais métodos para análise filogenética

- **UPGMA (Unweighted Pair Group Method Using Arithmetic Averages)**
 - Método de Agrupamento de Pares não Ponderados Usando Médias Aritméticas
 - Método de agrupamento
- **ME (Minimum Evolution)**
 - Busca pela árvore com a menor soma dos ramos
- **NJ (Neighbor-Joining)**
 - Aproximação de vizinhos
 - Pode ser visto como uma variação do método ME
 - Método de agrupamento(?)
 - **BIONJ** – otimização do método NJ

Principais métodos para análise filogenética

- **MP (*Maximum Parsimony*)**

- Máxima Parcimônia
- Busca pela árvore com o número mínimo de mutações

- **ML (*Maximum Likelihood*)**

- Máxima Verossimilhança
- Probabilístico
- Busca pela árvore que maximiza a verossimilhança do alinhamento para um modelo de substituição

- **BI (*Bayesian Inference*)**

- Inferência Bayesiana
- Probabilístico
- Variante do método ML
- Busca pela árvore com maior verossimilhança para o alinhamento

Abordagens para reconstrução filogenética

Métodos baseados em distâncias evolutivas



Converte o alinhamento em uma matriz de distâncias (UPGMA e NJ)

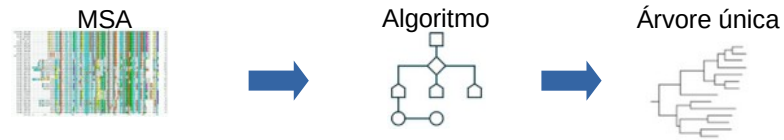
Métodos baseados em caracteres



O alinhamento é usado para construção da árvore (MP, ML e BI)

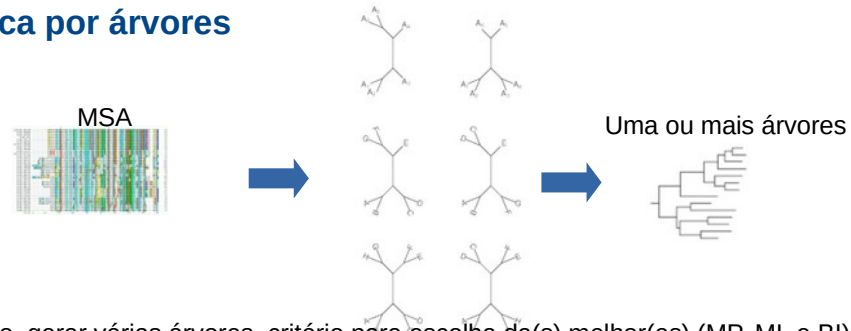
Abordagens para reconstrução filogenética

Algorítmica



Rápido, produz uma única árvore (UPGMA e NJ)

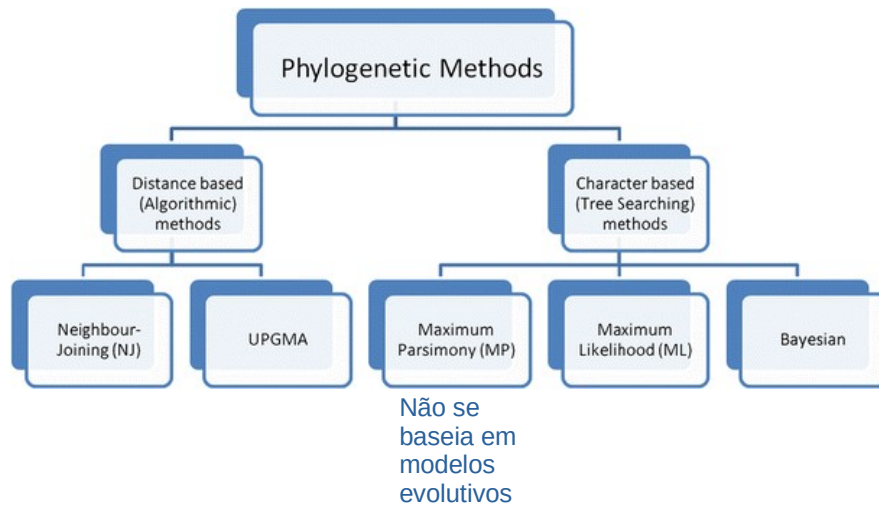
Busca por árvores



Lento, gerar várias árvores, critério para escolha da(s) melhor(es) (MP, ML e BI)



Métodos para construção de árvores

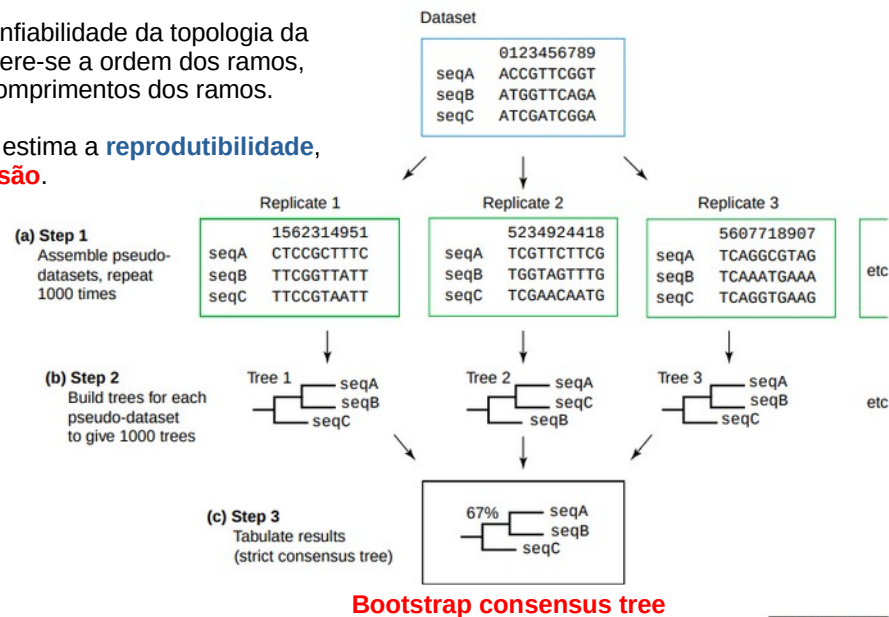


Sleator, R.D. *Microbial Ecology* 66(1):1-4, 2013. DOI: 10.1007/s00248-013-0236-x

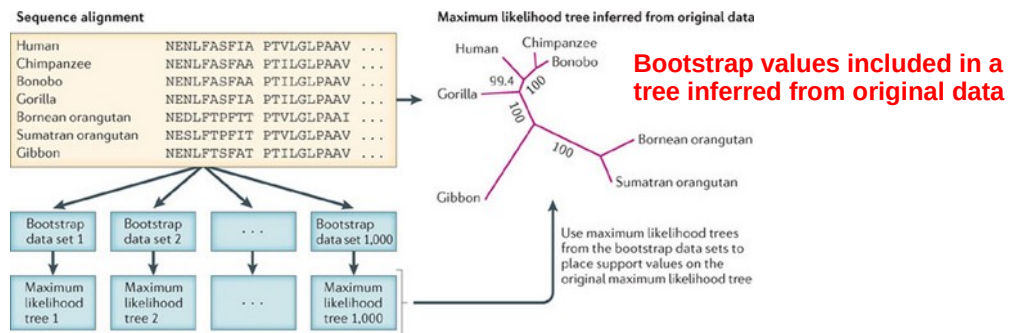
Teste de *bootstrap*

Testa a confiabilidade da topologia da árvore: refere-se a ordem dos ramos, não aos comprimentos dos ramos.

O método estima a **reprodutibilidade**, **não precisão**.

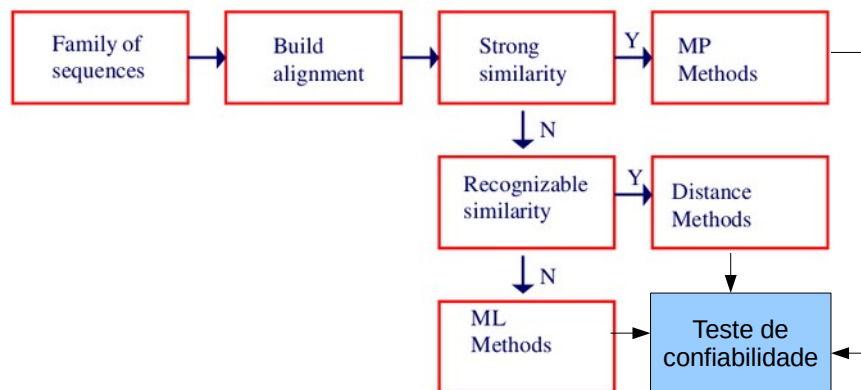


Sampling error in the estimated tree and bootstrap analysis



Yang, Z. & Rannala, B. Nature Reviews Genetics 13:303-314, 2012. Doi: 10.1038/nrg3186 Nature Reviews | Genetics

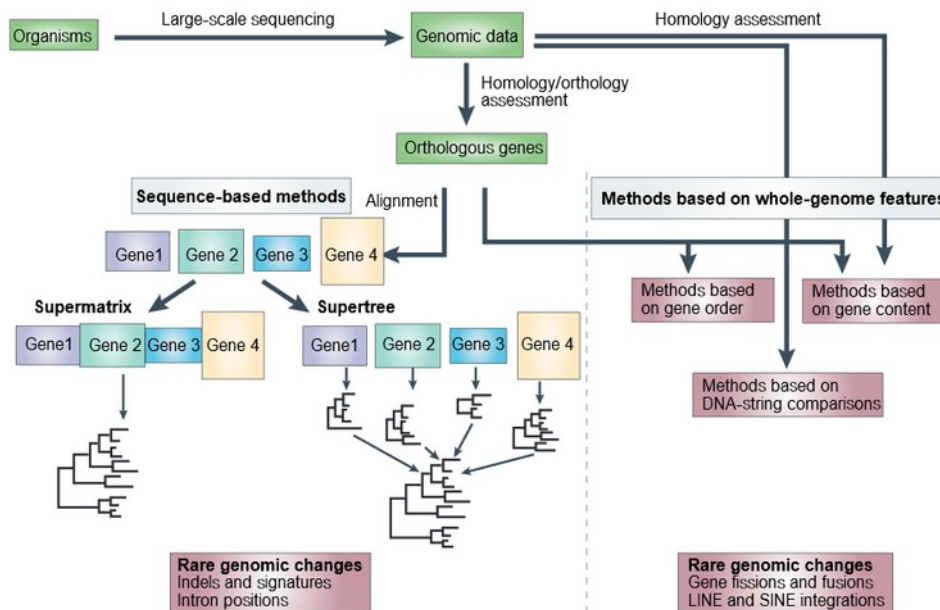
Phylogeny Flowchart



(Mount, *Bioinformatics*)

Este fluxograma pode ser utilizado para guiar aproximações iniciais na análise filogenética. Não necessariamente, o método MP sempre fornecerá os melhores resultados quando as sequências em um alinhamento de sequências múltiplo possuem alta identidade/similaridade.

Métodos para inferência filogenômica



Delsuc, F. et al. Nature Reviews Genetics 6:361-375, 2005. Doi:10.1038/nrg1603

A partir de sequências genômicas é possível utilizar um conjunto de diferentes genes para a análise filogenética. Em geral a árvore gerada para um conjunto de genes é mais robusta que a(s) árvore(s) gerada(s) para o(s) gene(s) separadamente.

Uma etapa crítica na filogenômica é a identificação dos genes homólogos ortólogos nos diferentes genomas.