Princípios de Sistemática Molecular

- Ciências teóricas e sistemática biológica ✓
- DNA, genes, código genético e mutação ✓
- Alinhamento de seqüências ✓
- Mudanças evolutivas em seqüências de nucleotídeos ✓
- Otimização em espaços contínuos e discretos ✓
- Árvores filogenéticas – Introdução ✓
- Inferência filogenética – Critério de matrizes de distâncias
- Inferência filogenética – Critério de parcimônia
- Inferência filogenética – Critério de verossimilhança
- Busca
- Confiabilidade
- Discussão de trabalhos em sistemática molecular

Bioinformática

- Nada em biologia faz sentido, exceto à luz da evolução. [Theodosius Dobzhansky (1900-1975)]

- Nada em Bioinformática faz sentido, exceto à luz da Biologia.

Evolução

- Uma grande parte da bioinformática se faz através de biologia comparativa;
- A biologia comparativa está baseada em relações evolutivas entre entidades comparáveis;
- As relações evolutivas são geralmente expressas na forma de árvores filogenéticas.

Distância × Similaridade

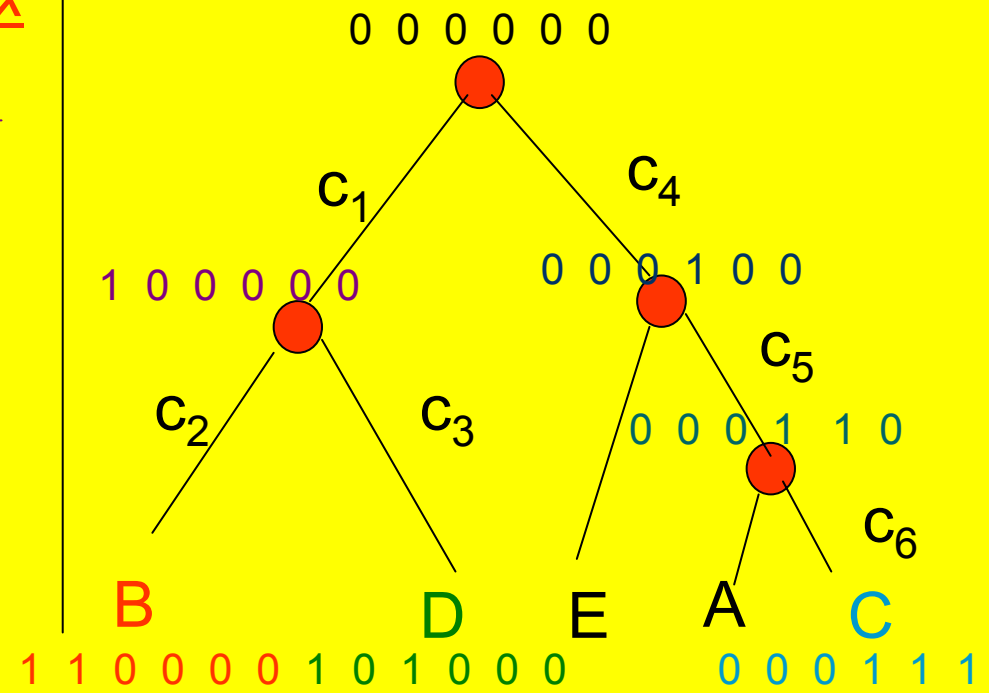
- Distância (D): número de posições distintas em seqüências de atributos;
- Similaridade (S): número de posições coincidentes em seqüências de atributos;
- Se $D \in [0,1]$, então $S = 1 - D$

Métricas de Distância

- Distância baseada em algum conjunto de atributos relevantes;
- Sistemática molecular
  - Distância baseada em DNA:
    - ✓ Número de bases distintas (distância de Hamming);
    - ✓ Número mínimo de operações (deleções, inserções, substituições) para converter uma seqüência em outra (distância de Levenshtein);
    - ✓ Com correção × Sem correção.
  - Distância baseada em proteína:
    - ✓ PAM matrix;
    - ✓ BLOSUM matrix.

A character state matrix

Taxon	c_1	c_2	c_3	c_4	c_5	c_6
A	0	0	0	1	1	0
B	1	1	0	0	0	0
C	0	0	0	1	1	1
D	1	0	1	0	0	0
E	0	0	0	1	0	0



Dangers in Molecular Phylogenies

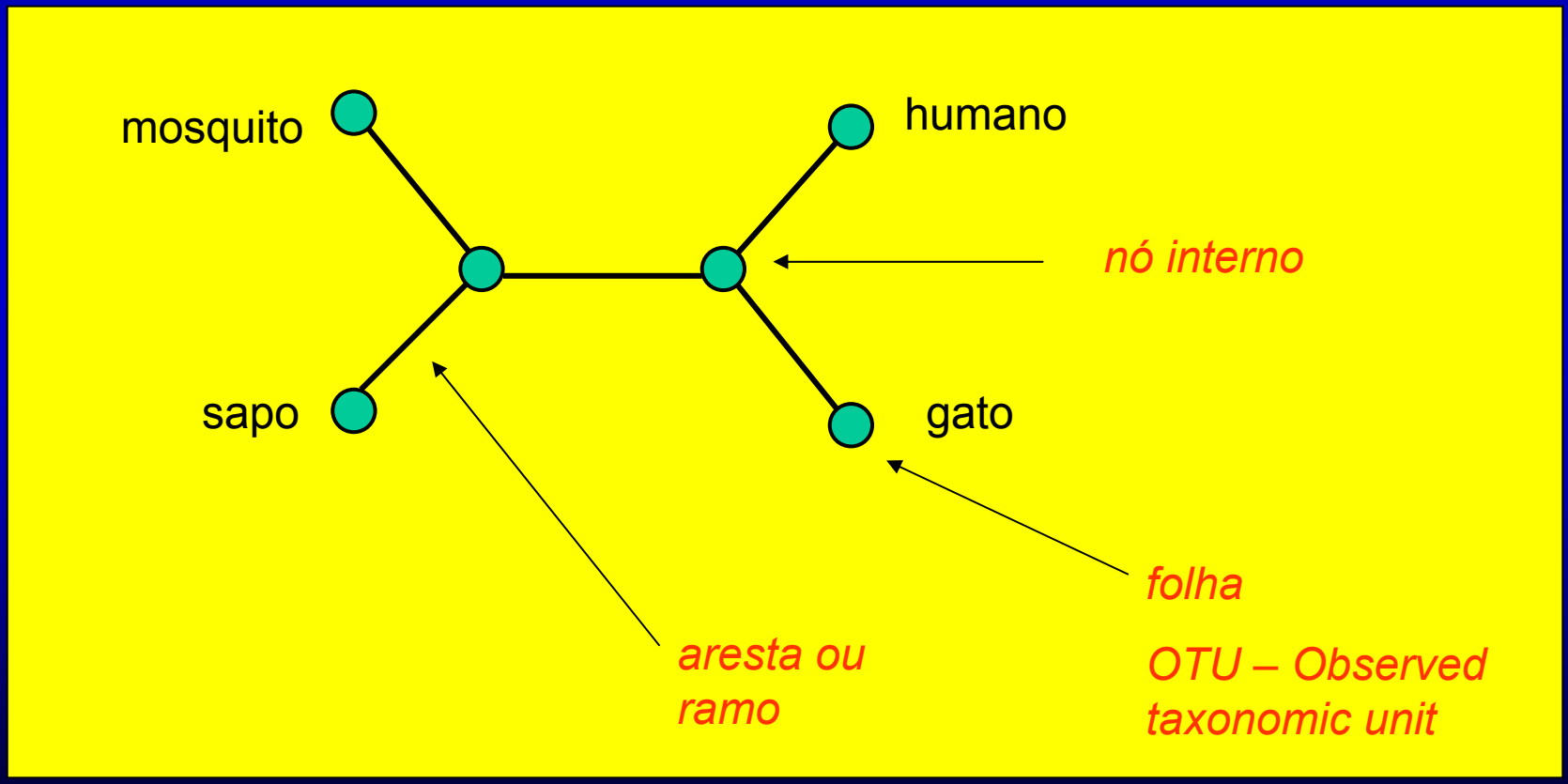
We have to remember that gene/protein sequence can be homologous for different reasons:

- **Orthologs -- sequences diverged after a speciation event**
- **Paralogs -- sequences diverged after a duplication event**

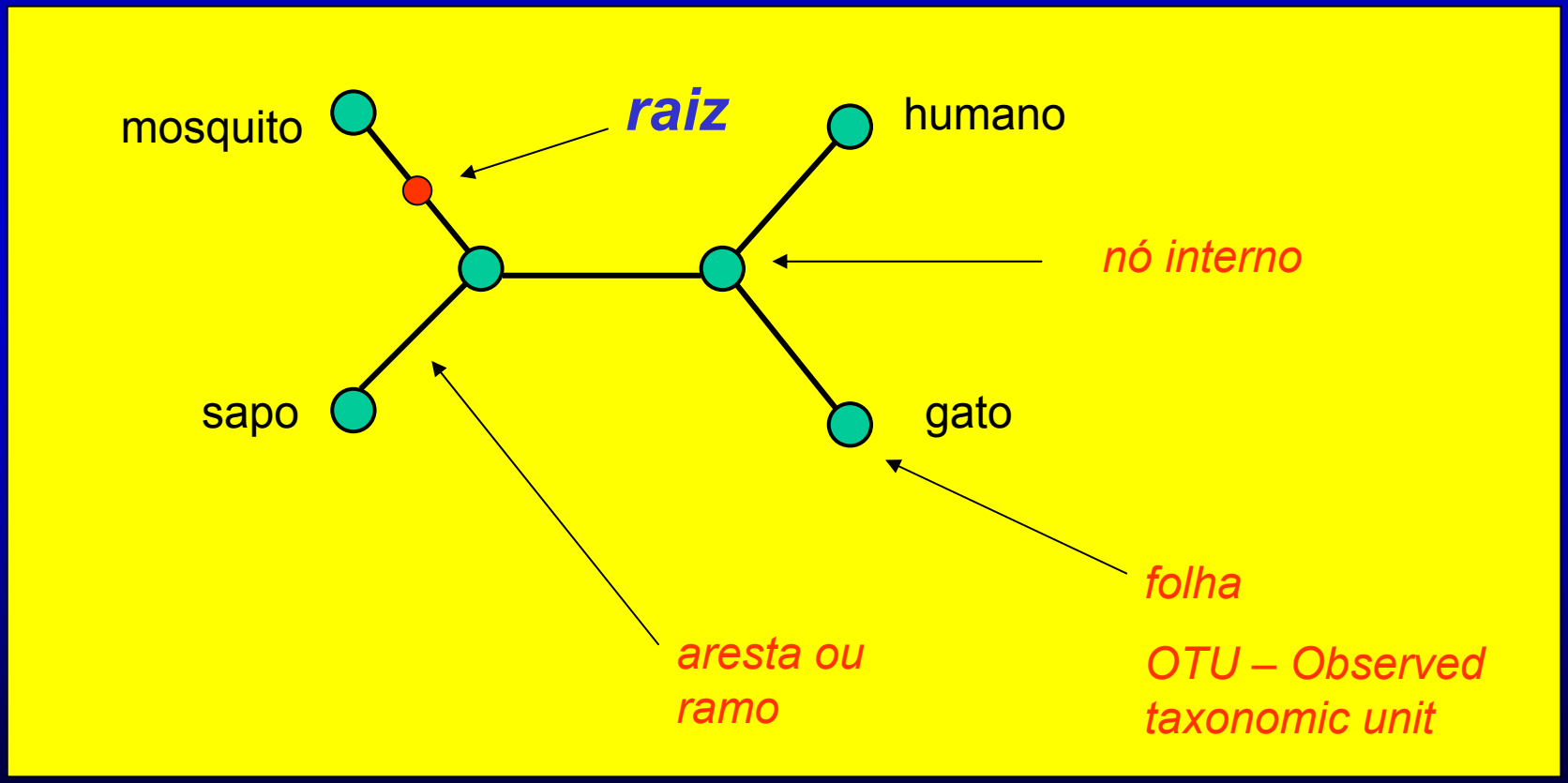
Homology: identical character due to shared ancestry (evolutionary *signal*)

Homoplasy: identical character due to evolutionary convergence or reversal (evolutionary *noise*)

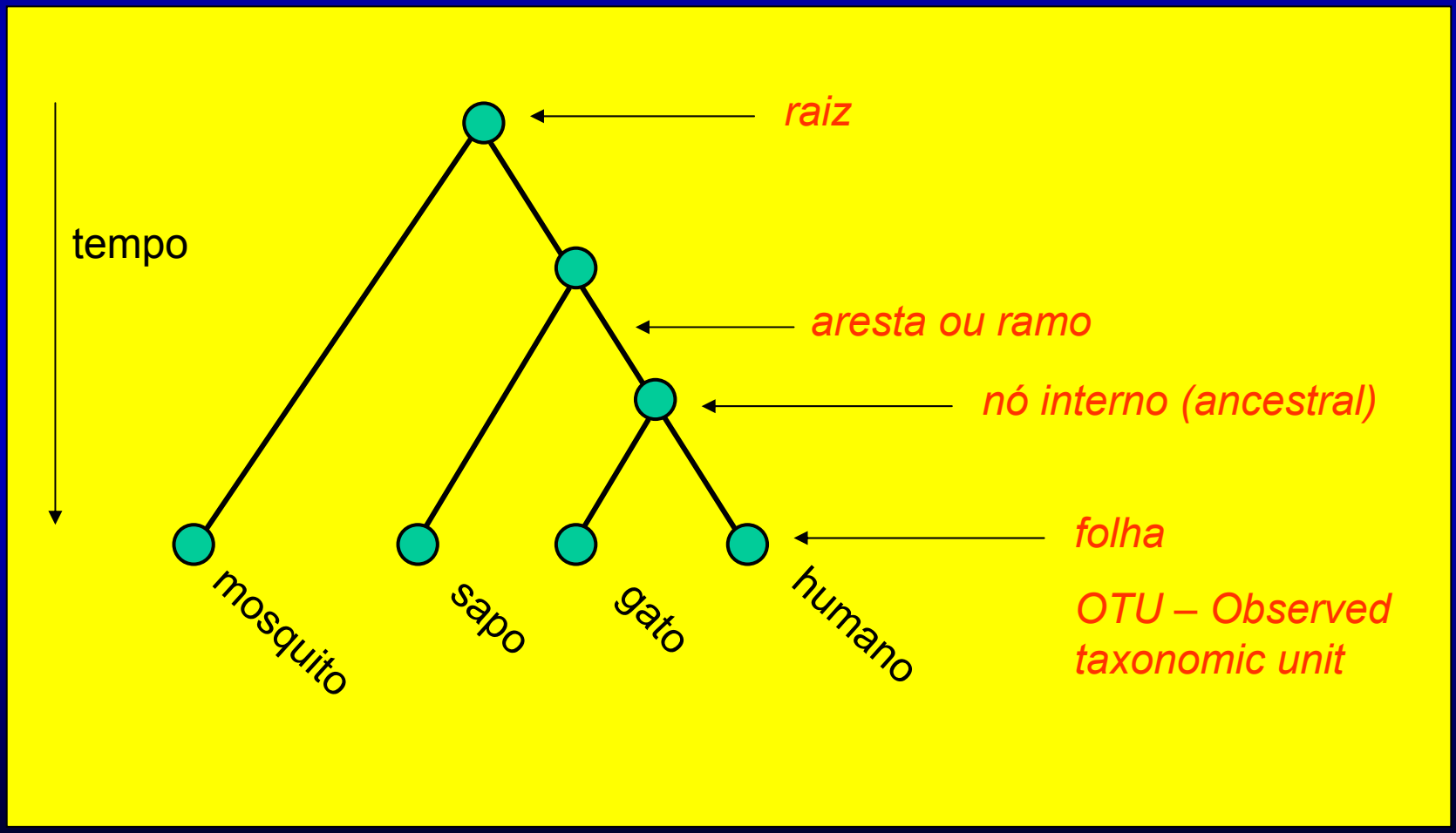
Árvore filogenética sem raiz



Árvore filogenética com raiz



Árvore filogenética com raiz



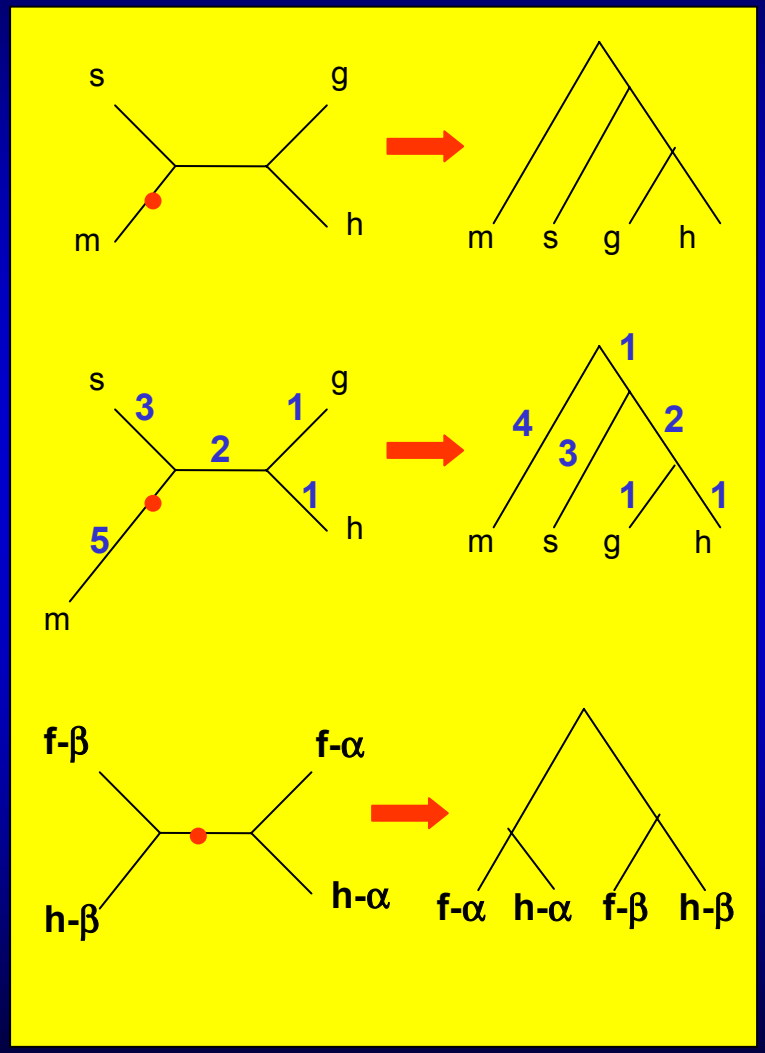
- Árvore sem raiz:
  - ✓Reflete apenas as relações de similaridade;
- Árvore com raiz:
  - ✓Indica também a direção da evolução;

Como enraizar uma árvore?

Outgroup – posicione a raiz entre a OTU mais distante e as demais OTUs.

Midpoint – posicione a raiz no ponto intermediário do caminho (soma do comprimento dos ramos entre duas OTUs) mais extenso.

Gene duplication – posicione a raiz entre cópias de genes parálogos.



Explosão combinatória

# seqüências	# árvores sem raiz	# árvores com raiz
2	1	1
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10.395
8	10.395	135.135
9	135.135	2.027.025
10	2.027.025	34.459.425

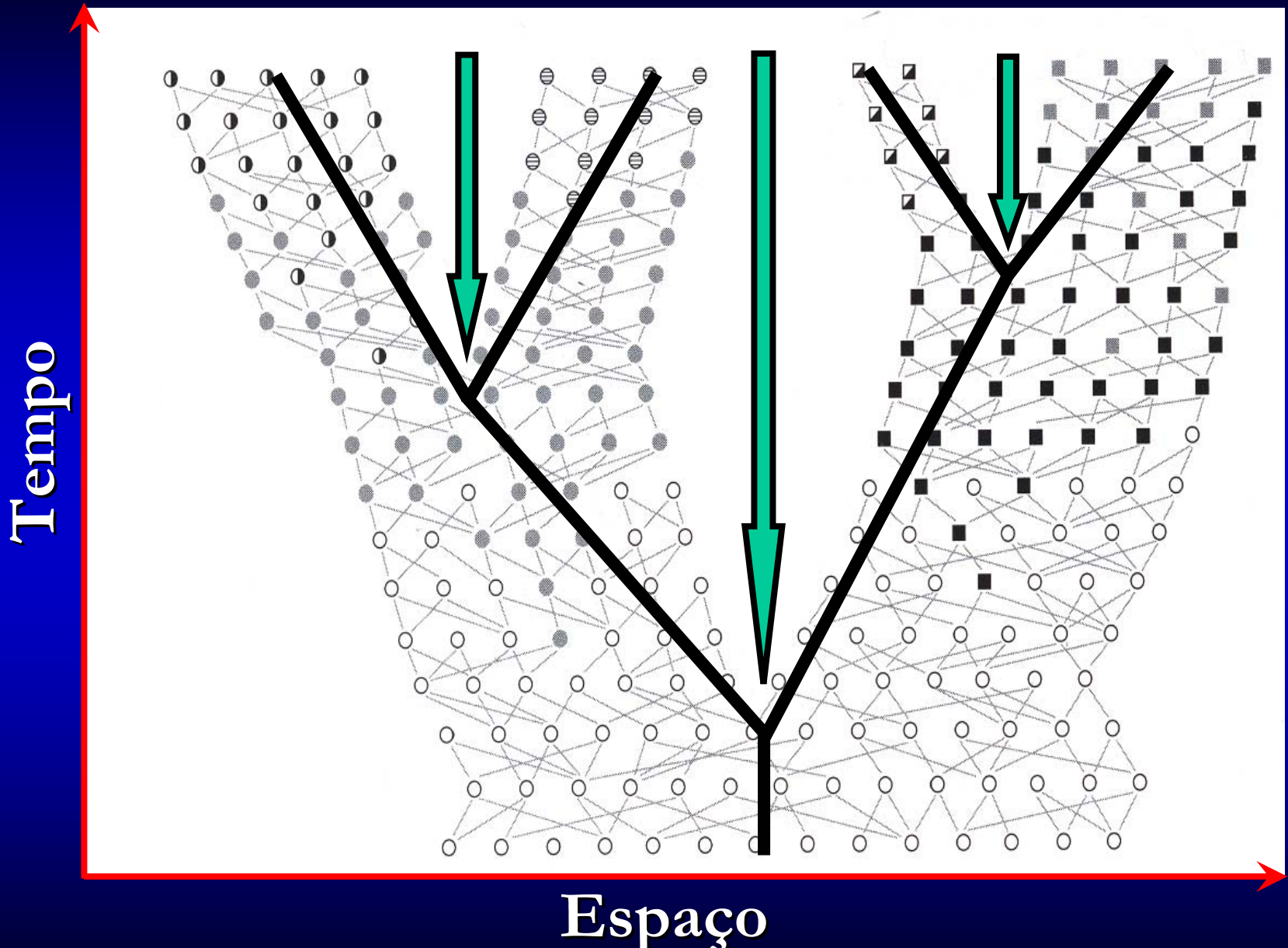
Explosão combinatória

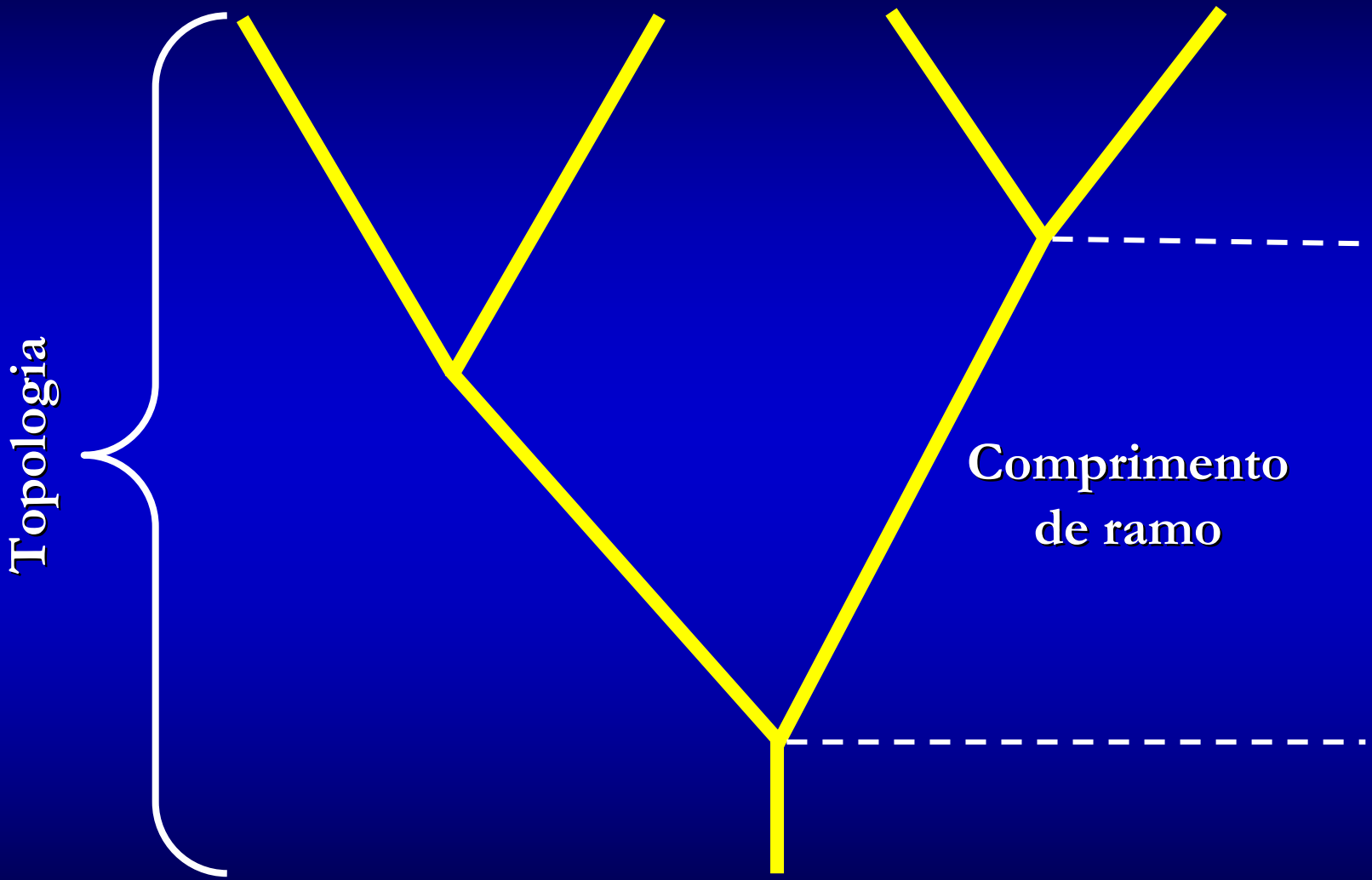
- 5 seqüências: 105 árvores candidatas
- 15 seqüências: 213.458.046.676.875 árvores candidatas
- 20 seqüências: 8.200.794.532.637.891.559.375 árvores candidatas
- *n* seqüências:

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

- a situação "melhora" quando se toma árvores sem raiz:

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


Tempo
Espaço


Topologia
Comprimento
de ramo

- **Reconstrução filogenética**
  - Inferência do padrão de ancestralidade comum mais recente
    - **Topologia**
  - Inferência do número de mudanças desde a divergência do ancestral comum mais recente
    - **Comprimento**

Reconstrução filogenética

Métodos:

- Baseados em distância: comparações par-a-par para se obter a topologia (*são os mais rápidos computacionalmente*);
- Parcimônia: propõe a topologia que indica a menor quantidade de eventos evolutivos;
- Máxima verossimilhança: partindo de um modelo evolutivo e das diferenças apresentadas pelas OTUs, propõe a topologia que melhor explica as OTUs.
- ***Obs****: todos eles também sugerem os comprimentos de ramos para cada topologia candidata, mas sob paradigmas distntos.*

Reconstrução filogenética

Métodos:

- Baseados em distância: abordagens fenéticas que envolvem clusterização, ou seja, agrupamento por similaridade (princípio da descendência de um ancestral comum);
- Parcimônia e máxima verossimilhança: abordagens cladísticas, baseadas em genealogia.
- ***Obs**: As abordagens cladísticas são superiores às abordagens fenéticas, mas requerem mais computação.*

Reconstrução filogenética

- métodos fenéticos ou não-baseados em modelo evolutivo: são aqueles que consideram o estado corrente das seqüências de atributos, não importando a história evolutiva, ou seja, a dinâmica dos passos intermediários. A árvore que melhor explica os relacionamentos entre as seqüências de atributos é denominada fenograma.

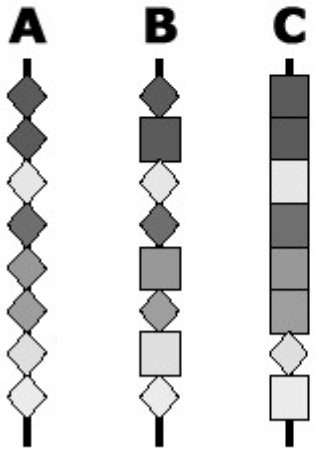
- métodos cladísticos ou baseados em modelo evolutivo: são aqueles que consideram as possibilidades de resultado de um processo evolutivo, importando a dinâmica dos passos intermediários, e adotam a árvore que melhor explica os relacionamentos entre as seqüências de atributos resultantes, sempre com base em uma hipótese evolutiva. Esta hipótese evolutiva pode estar baseada em algum modelo evolutivo ou em algum critério de otimalidade. A árvore que melhor explica os relacionamentos entre as seqüências de atributos é denominada cladograma. Na árvore adotada, o comprimento dos ramos pode ser informativo (quando a hipótese está baseada em um modelo evolutivo), resultando em um filograma.

Reconstrução filogenética

MÉTODOS FENÉTICOS

Análise de vários caracteres

(fenótipos, sítios polimórficos, freq. alélicas, etc.)

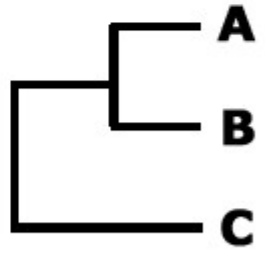


Matriz de dissimilaridade em base ao número de diferenças

	A	B	C
A	0		
B	3	0	
C	7	6	0

Construção de árvore filogenética

Fenograma

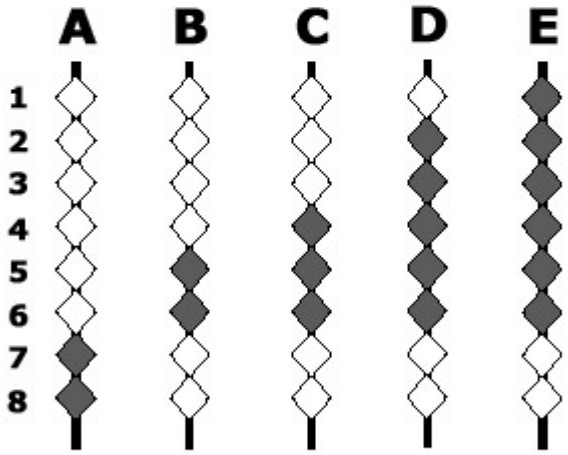


Relações de similaridade

Reconstrução filogenética

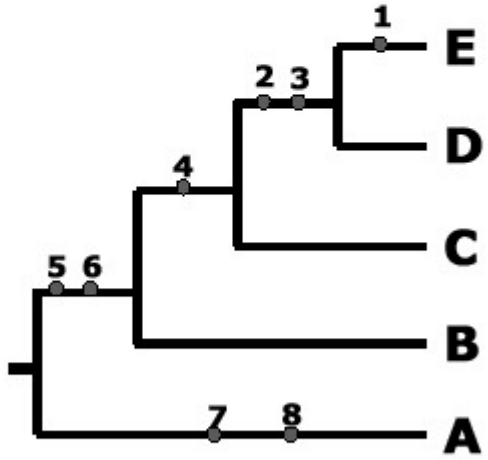
MÉTODOS CLADÍSTICOS

Análise de vários caracteres com estados ancestrais (0) e derivados (1)



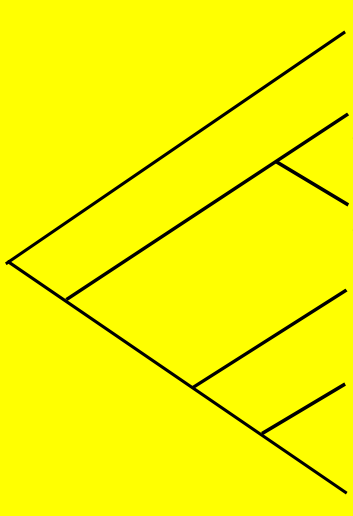
Construção de árvore filogenética

Cladograma

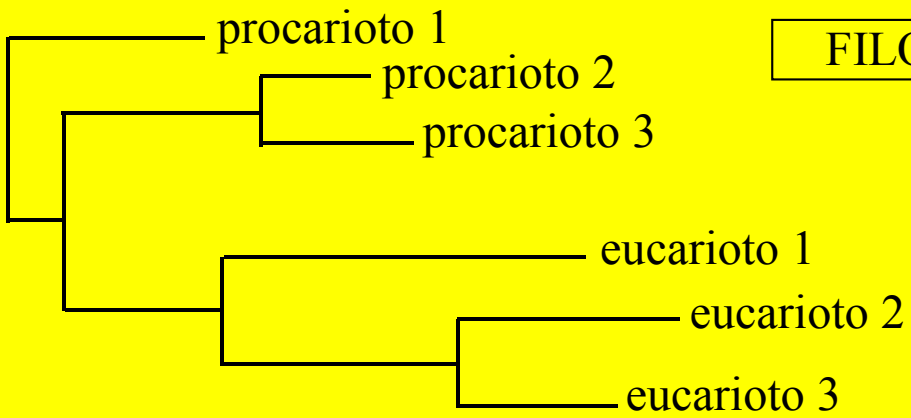


Relações de ancestralidade comum

Reconstrução filogenética



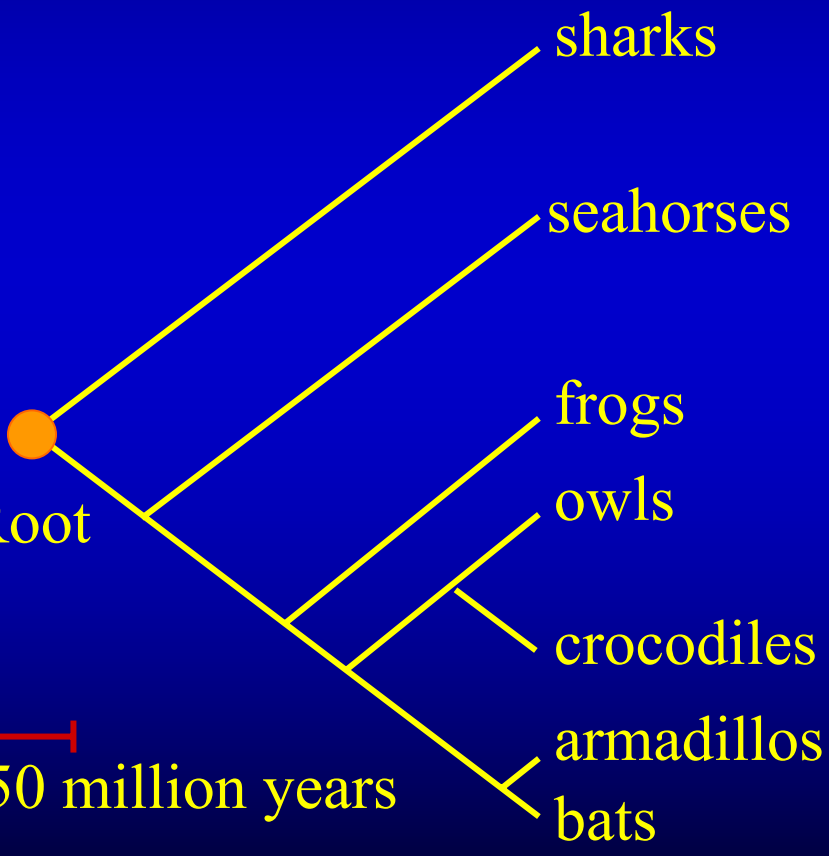
CLADOGRAMA



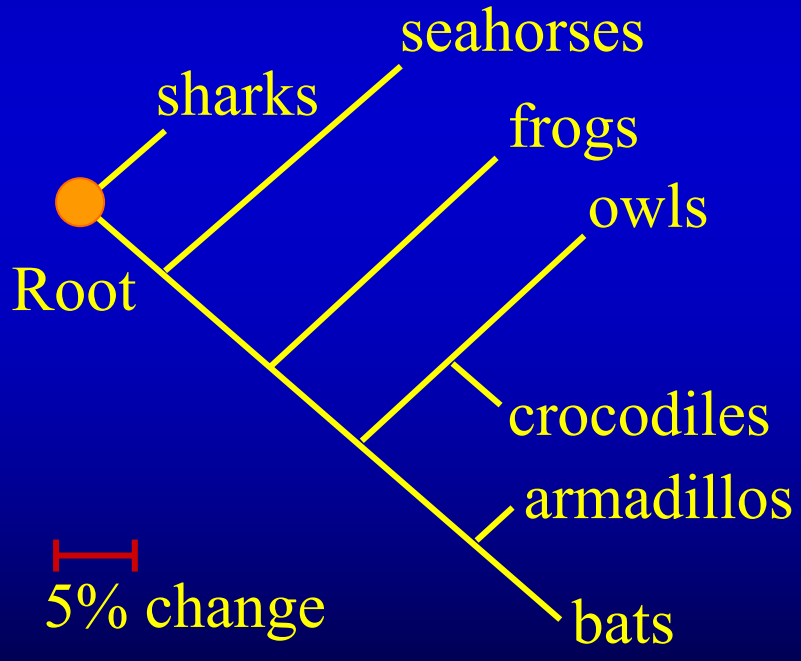
FILOGRAMA

Tree Types

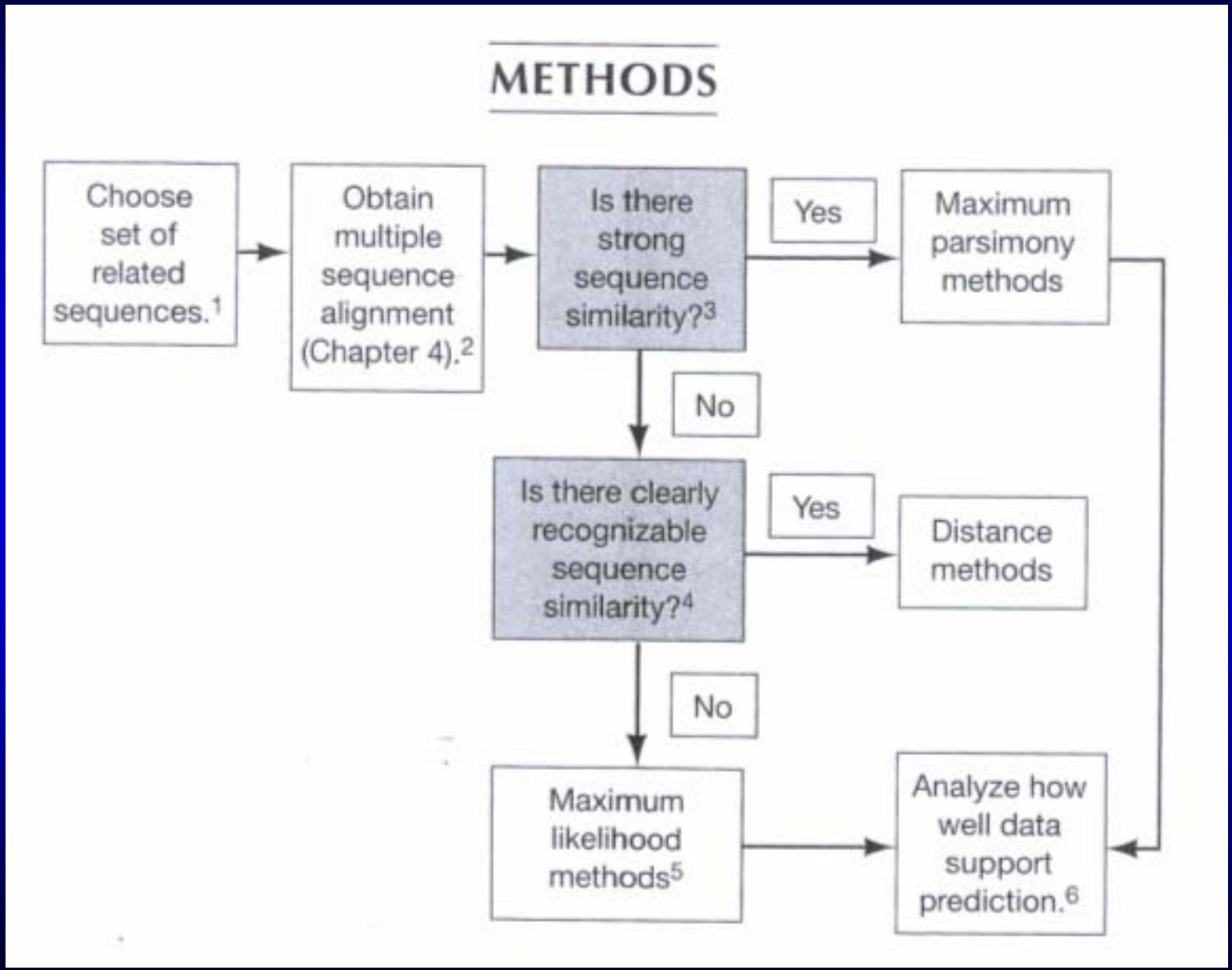
Cladograms measure *time*.



Phylograms measure *change*.



Tony Weisstein, http://bioquest.org:16080/bedrock/terre_haute_03_04/phylogenetics_1.0.ppt


METHODS
Choose set of related sequences.[1]
Obtain multiple sequence alignment (Chapter 4).[2]
Is there strong sequence similarity?[3]
Yes
Maximum parsimony methods
No
Is there clearly recognizable sequence similarity?[4]
Yes
Distance methods
No
Maximum likelihood methods[5]
Analyze how well data support prediction.[6]

Voltando aos métodos baseados em distância ...

- # Reconstrução filogenética

  - ✓ Métodos baseados em distância

    - **Objetivo**

      – Ajustar uma árvore a uma matriz de distâncias genéticas

    - **Passos**

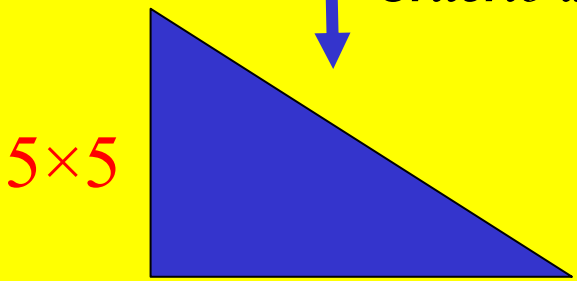
      1. Estimativa de distâncias genéticas a partir de dados de seqüências moleculares (DNA, proteínas)

      2. Agrupamento das distâncias

Métodos baseados em distância

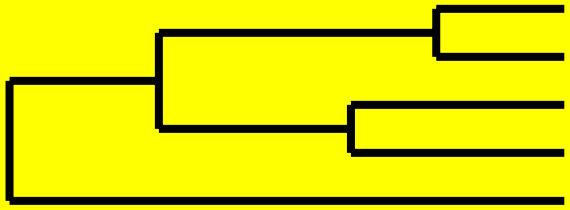
Alinhamento Múltiplo

Critério de similaridade



Matriz de distâncias

Método de agrupamento



Árvore filogenética

Métodos de agrupamento (Clustering)

- UPGMA
- WPGMA
- Fitch-Margoliash
- Evolução mínima
- *Neighbor-Joining* (NJ)

UPGMA

- Unweighted-Pair Group Method with Arithmetic means (Sokal and Michener, 1958)

WPGMA

- Weighted-Pair Group Method with Arithmetic means

- Hipótese do relógio molecular: as taxas de evolução são as mesmas ao longo dos vários ramos da árvore filogenética.

Matriz de distâncias genéticas estimadas

(não-corrigidas)

	Humano	Chimpanzé	Gorila	Orangotango	Gibão
Humano	—	*0,014*	*0,043*	*0,130*	*0,174*
Chimpanzé	1	—	*0,029*	*0,116*	*0,159*
Gorila	3	2	—	*0,087*	*0,159*
Orangotango	9	8	6	—	*0,159*
Gibão	12	11	11	11	—

também para WPGMA

- Conversão das distâncias estimadas em distâncias ***evolutivas*** (esperadas)
  - Correção para eventos de substituições múltiplas
    - Modelo de Jukes-Cantor

$$\hat{d}_{JC} = -\frac{3}{4}\ln\left(1-\frac{4}{3}\hat{p}\right)$$

Matriz de distâncias pareadas entre seqüências corrigidas pelo modelo de Jukes-Cantor (1 parâmetro)

	Humano	Chimpanzé	Gorila	Orangotango	Gibão
Humano	—	0,015	0,045	0,143	0,198
Chimpanzé		—	0,030	0,126	0,179
Gorila			—	0,092	0,179
Orangotango				—	0,179
Gibão					—

ambém para WPGMA

- Reconstrução filogenética
  - UPGMA — Solução única (produz uma única topologia)
    - Passos
      - Agrupar as seqüências ou grupos de seqüências mais semelhantes entre si — seqüências com a menor distância genéticas
        - » **Topologia**
      - Definir o ponto de ramificação entre as seqüências como a média entre as distâncias entre seqüências ou grupos de seqüências
        - » **Comprimento do ramo**

também para WPGMA

1. Qual par de seqüências será agrupado?
   - No método ***UPGMA*** o par de seqüências (ou grupo de seqüências) a ser agrupado primeiro é aquele que apresenta a ***menor*** distância entre todos os pares (ou grupos) de seqüências — **Topologia**

2. Qual o ponto de ramificação entre o par de seqüências agrupado?
   - No método ***UPGMA*** o ponto de ramificação é definido como a média da distância entre o par de seqüências (ou grupo de seqüências) agrupado — **Comprimento de ramo**

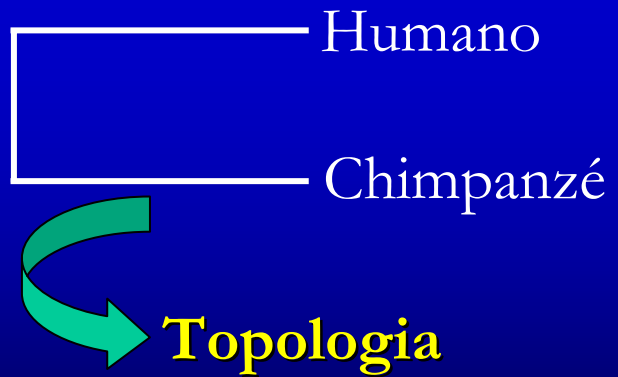
ambém para WPGMA

	Humano	Chimpanzé	Gorila	Orangotango	Gibão
Humano	—	**0,015**	**0,045**	**0,143**	**0,198**
Chimpanzé		—	**0,030**	**0,126**	**0,179**
Gorila			—	**0,092**	**0,179**
Orangotango				—	**0,179**
Gibão					—

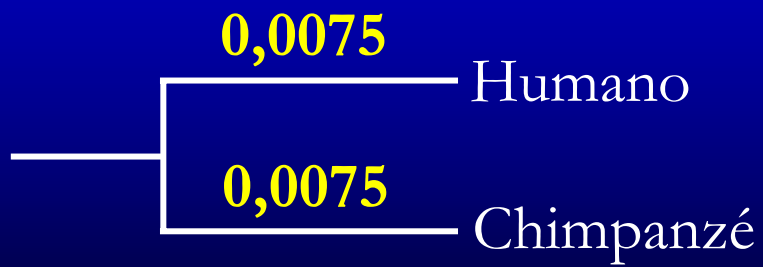
- Qual par de seqüências será agrupado?
  - No método ***UPGMA*** o par de seqüências (ou grupo de seqüências) a ser agrupado primeiro é aquele que apresentar a ***menor*** distância entre todos os pares (ou grupos) de seqüências

ambém para WPGMA

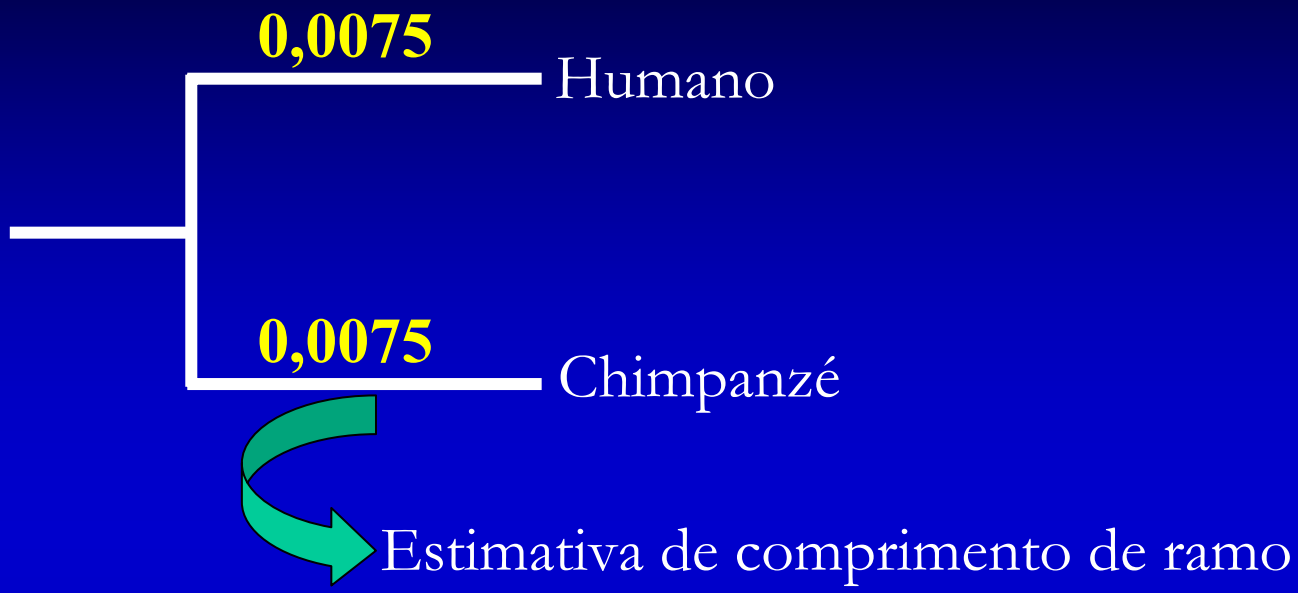
- Qual par de seqüências será agrupado?
  - O par Humano – Chimpanzé
    - Distância de Jukes-Cantor = **0,015**



- Qual o ponto de ramificação entre o par de seqüências Humano – Chimpanzé?
  - No método ***UPGMA*** o ponto de ramificação é definido como a média da distância entre o par de seqüências (ou grupo de seqüências) agrupado
    - Distância Humano – Chimpanzé = **0,015**
      - ✓ Ponto de ramificação ⇒ **0,015/2 = 0,0075**



...também para WPGMA



- Comprimento de ramo ⇒ número de mudanças ocorridas entre os nós de uma árvore
  - O método UPGMA ***simultaneamente*** estima a topologia da árvore e os comprimentos de ramo

- As seqüências Humano – Chimpanzé formam agora um agrupamento
  - Qual será a distância das demais seqüências para este agrupamento?
    - A distância média das seqüências aos membros do agrupamento

Cálculo das distâncias entre o agrupamento Humano – Chimpanzé e as demais seqüências

Matriz original $N \times N$

	Humano	Chimpanzé	Gorila	Orangotang o	Gibão
Humano	—	0,015	0,045	0,143	0,198
Chimpanzé		—	0,030	0,126	0,179
Gorila			—	0,092	0,179
Orangotang o				—	0,179
Gibão					—

$$\hat{d}_{(hu-ch),go} = \frac{\left(\hat{d}_{hu,go} + \hat{d}_{ch,go}\right)}{2}$$

$$\hat{d}_{(hu-ch),go} = \frac{(0{,}045 + 0{,}030)}{2} = 0{,}037$$

$$\hat{d}_{(hu-ch),gi} = \frac{\left(\hat{d}_{hu,gi} + \hat{d}_{ch,gi}\right)}{2}$$

$$\hat{d}_{(hu-ch),gi} = \frac{(0{,}198 + 0{,}179)}{2} = 0{,}189$$

$$\hat{d}_{(hu-ch),or} = \frac{\left(\hat{d}_{hu,or} + \hat{d}_{ch,or}\right)}{2}$$

$$\hat{d}_{(hu-ch),or} = \frac{(0{,}143 + 0{,}126)}{2} = 0{,}135$$

Matriz original $N \times N$

	Humano	Chimpanzé	Gorila	Orangotang o	Gibão
Humano	—	0,015	0,045	0,143	0,198
Chimpanzé		—	0,030	0,126	0,179
Gorila			—	0,092	0,179
Orangotang o				—	0,179
Gibão					—

$$\hat{d}_{(hu-ch),go} = \frac{\left(\hat{d}_{hu,go} + \hat{d}_{ch,go}\right)}{2}$$

$$\hat{d}_{(hu-ch),go} = \frac{(0{,}045 + 0{,}030)}{2} = 0{,}037$$

$$\hat{d}_{(hu-ch),or} = \frac{\left(\hat{d}_{hu,or} + \hat{d}_{ch,or}\right)}{2}$$

$$\hat{d}_{(hu-ch),or} = \frac{(0{,}143 + 0{,}126)}{2} = 0{,}135$$

$$\hat{d}_{(hu-ch),gi} = \frac{\left(\hat{d}_{hu,gi} + \hat{d}_{ch,gi}\right)}{2}$$

$$\hat{d}_{(hu-ch),gi} = \frac{(0{,}198 + 0{,}179)}{2} = 0{,}189$$

	Hu – Ch	Go	Or	Gi
Hu - Ch	0,000	0,037	0,135	0,189
Go		0,000	0,092	0,179
Or			0,000	0,179
Gi				0,000

Matriz reduzida $(N-1) \times (N-1)$

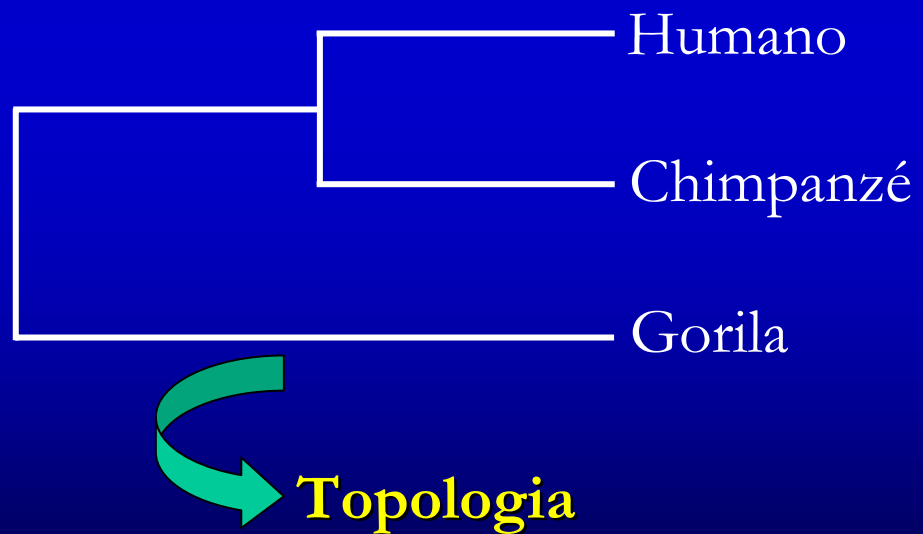
	Hu – Ch	Go	Or	Gi
Hu - Ch	0,000	0,037	0,135	0,189
Go		0,000	0,092	0,179
Or			0,000	0,179
Gi				0,000

Matriz reduzida $(N-1) \times (N-1)$

- Qual par de seqüências será agrupado?
  - No método ***UPGMA*** o par de seqüências (ou grupo de seqüências) a ser agrupado primeiro é aquele que apresentar a ***menor*** distância entre todos os pares (ou grupos) de seqüências

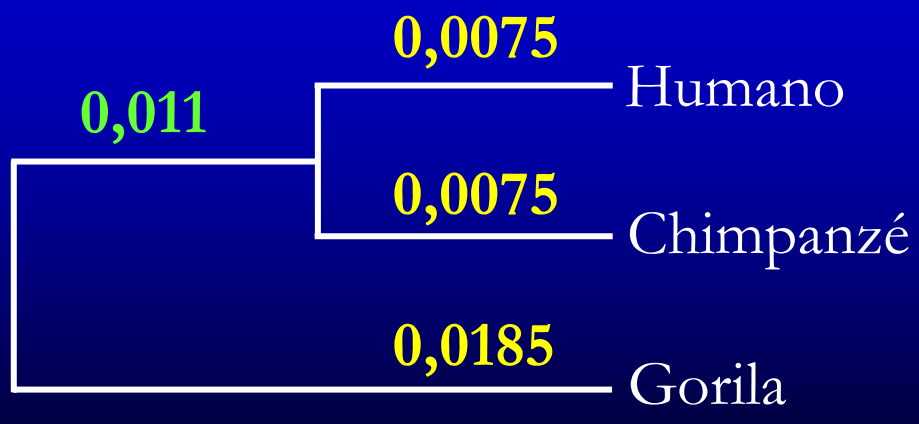
ambém para WPGMA

- Qual par de seqüências será agrupado?
  - O par (Humano – Chimpanzé), Gorila
    - Distância de Jukes-Cantor = **0,037**

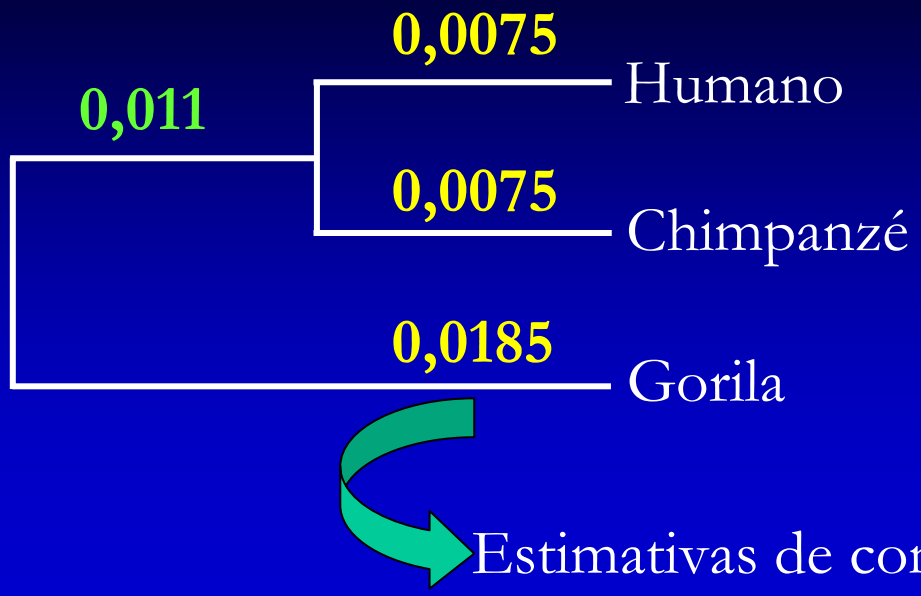


ambém para WPGMA

- Qual o ponto de ramificação entre o par de seqüências (Humano – Chimpanzé), Gorila?
  - No método ***UPGMA*** o ponto de ramificação é definido como a média da distância entre o par de seqüências (ou grupo de seqüências) agrupado
    - Distância (Humano – Chimpanzé), Gorila = **0,037**
      - ✓ Ponto de ramificação ⇒ **0,037/2 = 0,0185**



também para WPGMA



- Comprimento de ramo ⇒ número de mudanças ocorridas entre os nós de uma árvore
  - O método UPGMA ***simultaneamente*** estima a topologia da árvore e os comprimentos de ramo

- As seqüências (Humano – Chimpanzé), Gorila formam agora um agrupamento
  - Qual será a distância das demais seqüências para este agrupamento?

UPGMA:

- A distância média entre as seqüências que pertencem aos ramos que sofrerão agrupamento

WPGMA:

- A distância média entre os ramos que sofrerão agrupamento

Matriz original $N \times N$

	Humano	Chimpanzé	Gorila	Orango	Gibão
Humano	—	0,015	0,045	0,143	0,198
Chimpanzé		—	0,030	0,126	0,179
Gorila			—	0,092	0,179
Orango				—	0,179
Gibão					—

	Hu – Ch	Go	Or	Gi
Hu - Ch	0,000	0,037	0,135	0,189
Go		0,000	0,092	0,179
Or			0,000	0,179
Gi				0,000

Matriz reduzida $(N-1) \times (N-1)$

Cálculo das distâncias entre o agrupamento (Humano – Chimpanzé – Gorila) e as demais seqüências

Matriz original $N \times N$

	Humano	Chimpanzé	Gorila	Orango	Gibão
Humano	—	0,015	0,045	0,143	0,198
Chimpanzé		—	0,030	0,126	0,179
Gorila			—	0,092	0,179
Orango				—	0,179
Gibão					—

$$\hat{d}_{(hu-ch-go),or} = \frac{\left(\hat{d}_{hu,or} + \hat{d}_{ch,or} + \hat{d}_{go,or}\right)}{3}$$

$$\hat{d}_{(hu-ch-go),or} = \frac{(0{,}143 + 0{,}126 + 0{,}092)}{3} = 0{,}121$$

$$\hat{d}_{(hu-ch-go),gi} = \frac{\left(\hat{d}_{hu,gi} + \hat{d}_{ch,gi} + \hat{d}_{go,gi}\right)}{3}$$

$$\hat{d}_{(hu-ch-go),gi} = \frac{(0{,}198 + 0{,}179 + 0{,}179)}{3} = 0{,}185$$

Cálculo das distâncias entre o agrupamento (Humano – Chimpanzé – Gorila) e as demais seqüências

	Hu – Ch	**Go**	**Or**	**Gi**
Hu - Ch	0,000	0,037	0,135	0,189
Go		0,000	0,092	0,179
Or			0,000	0,179
Gi				0,000

Matriz reduzida $(N-1) \times (N-1)$

$$\hat{d}_{(hu-ch-go),or} = \frac{\left(\hat{d}_{(hu-ch),or} + \hat{d}_{go,or}\right)}{2}$$

$$\hat{d}_{(hu-ch-go),or} = \frac{(0{,}135 + 0{,}092)}{2} = 0{,}114$$

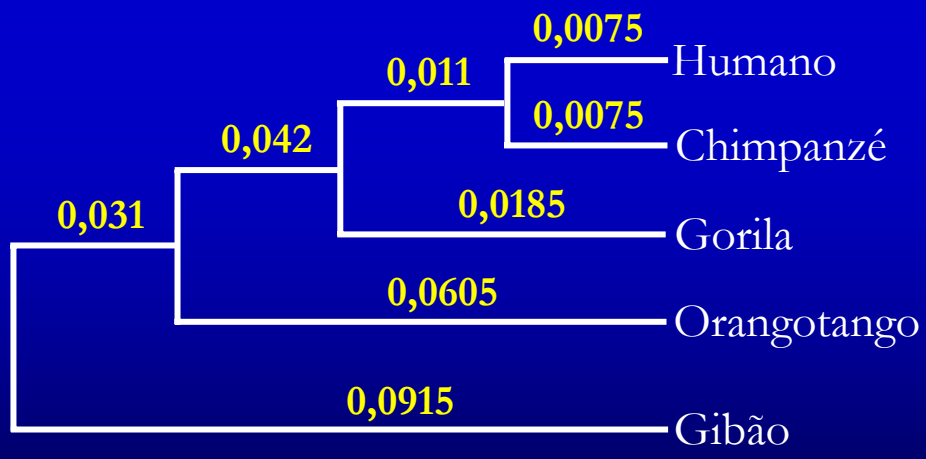
$$\hat{d}_{(hu-ch-go),gi} = \frac{\left(\hat{d}_{(hu-ch),gi} + \hat{d}_{go,gi}\right)}{2}$$

$$\hat{d}_{(hu-ch-go),gi} = \frac{(0{,}189 + 0{,}179)}{2} = 0{,}184$$

Assim por diante ...

Matriz de distâncias genéticas corrigidas pelo modelo de Jukes-Cantor

	Humano	Chimpanzé	Gorila	Orango	Gibão
Humano	—	0,015	0,045	0,143	0,198
Chimpanzé	0,015	—	0,030	0,126	0,179
Gorila	0,037	0,037	—	0,092	0,179
Orango	0,121	0,121	0,121	—	0,179
Gibão	0,183	0,183	0,183	0,183	—

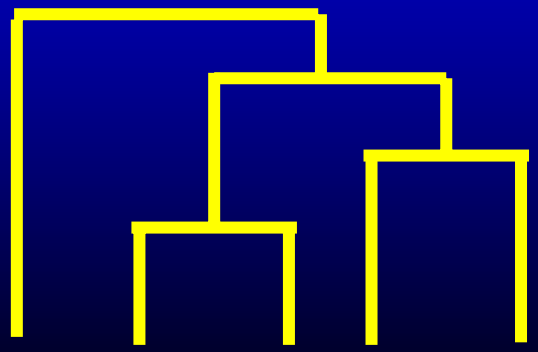


Estimativas de comprimentos de ramos
Distâncias "patrísticas"

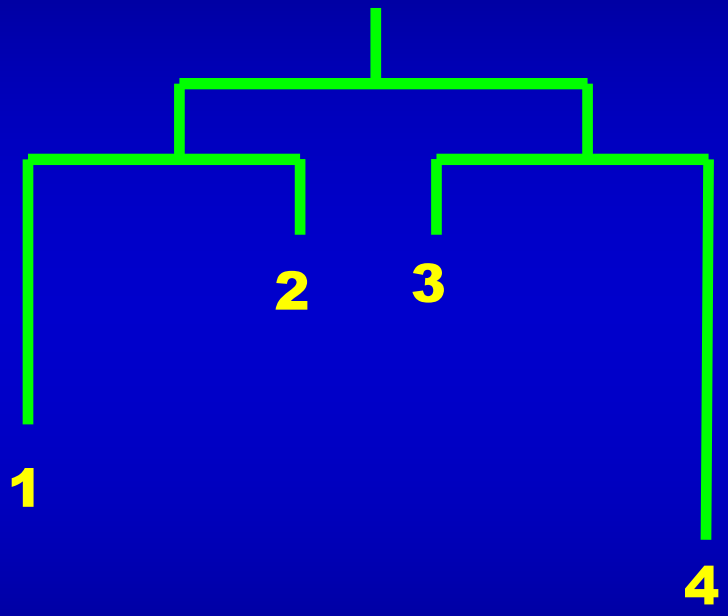
Ultrametricidade (árvore com raiz)

- A distância das folhas à raiz é a mesma, ou seja, a distância é proporcional ao tempo evolutivo;
- Uma árvore ultramétrica é caracterizada pela condição dos 3 pontos: dadas as 3 distâncias par-a-par entre 3 seqüências, pelo menos 2 delas são idênticas, sendo que a terceira é idêntica ou menor.

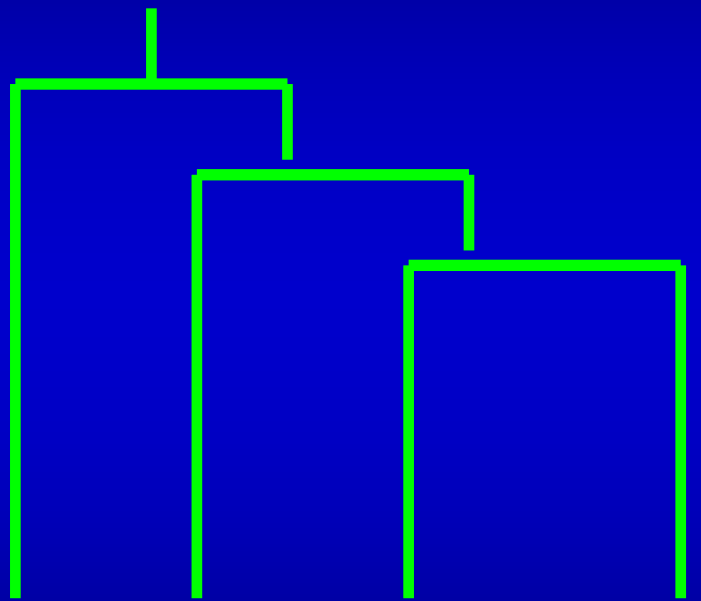
- $d_{AC} \leq \max(d_{AB}, d_{BC})$



Quando os ramos não têm a mesma distância à raiz UPGMA e WPGMA podem falhar espetacularmente.



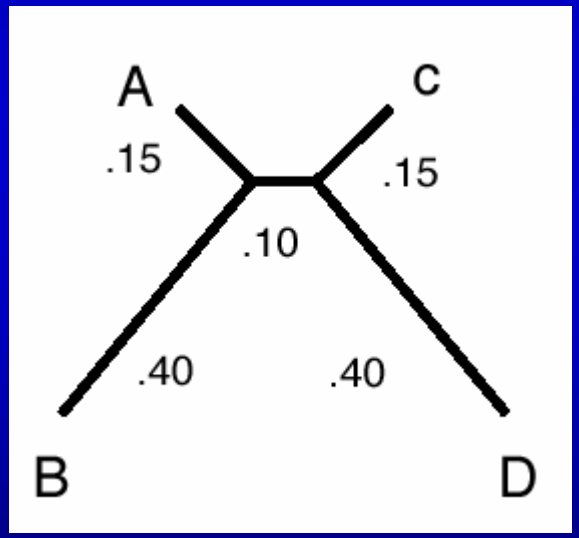
Uma árvore real ...



... reconstruída incorretamente por UPGMA ou WPGMA

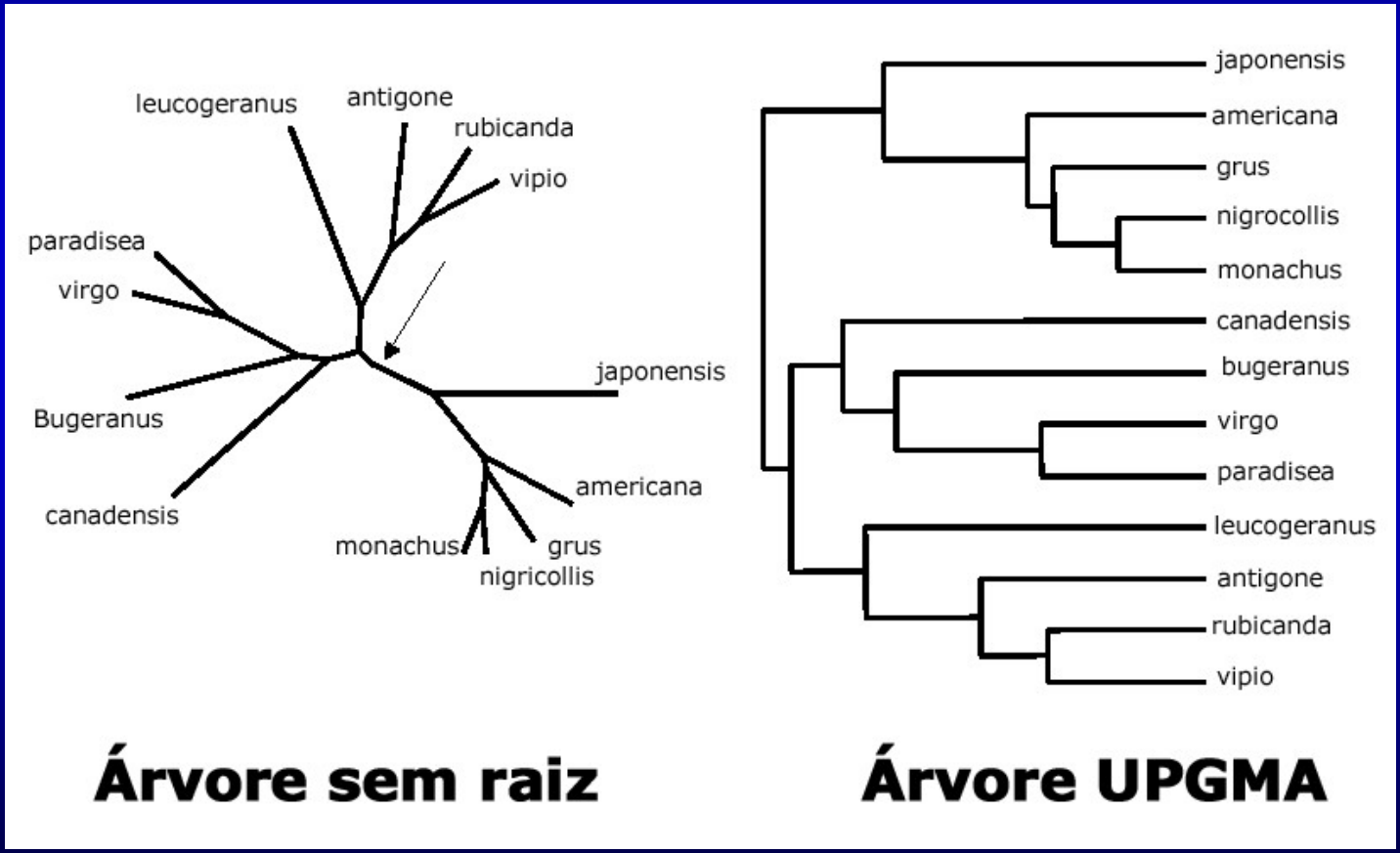
Problemas com UPGMA e WPGMA

- Só produz a árvore correta se todos os ramos têm a mesma distância à raiz.



	A	B	C
B	0.55		
C	0.40	0.65	
D	0.65	0.90	0.55

Sem ultrametricidade e sem raiz



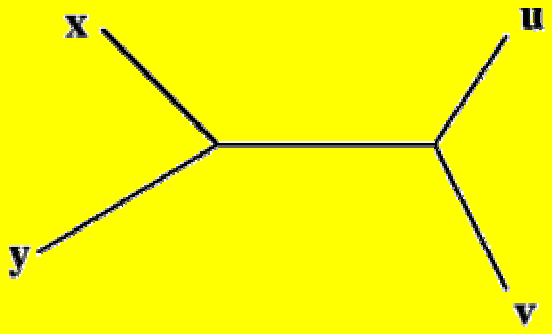
Usar UPGMA e WPGMA?

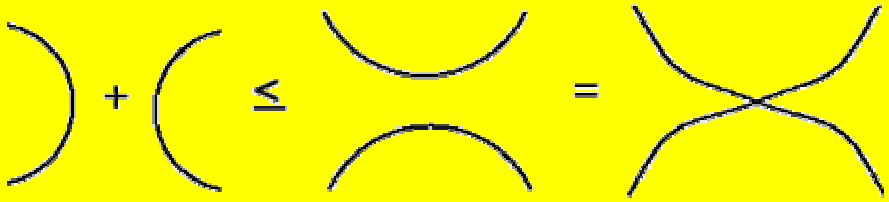
Não.
Conheça a técnica e evite empregá-la.

Aditividade (árvore sem raiz)

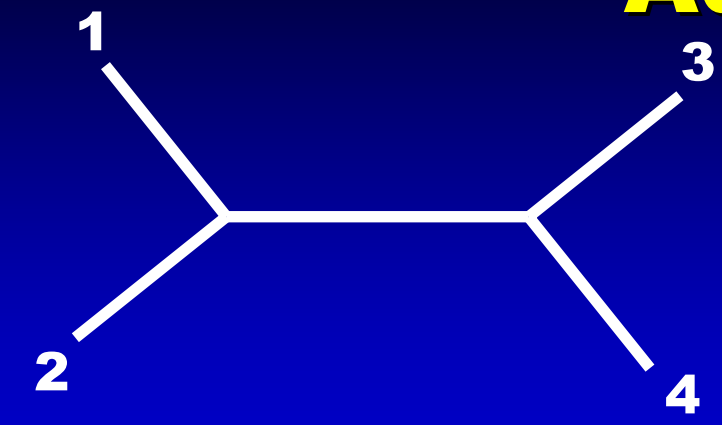
Uma árvore aditiva é caracterizada pela condição dos 4 pontos. Quaisquer 4 folhas da árvore em qualquer ordenação admitem a seguinte relação:

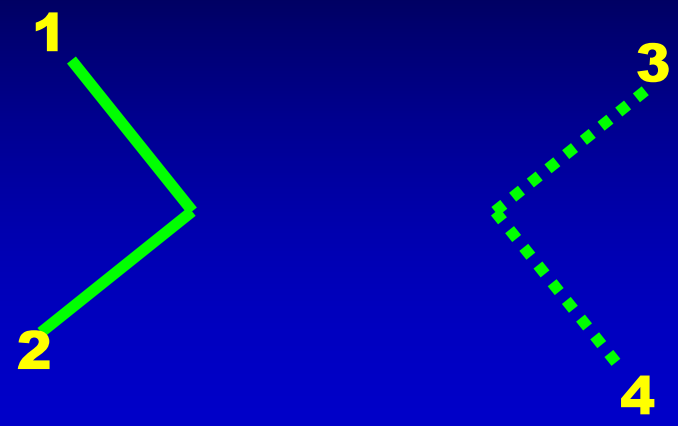
$$d(x,y) + d(u,v) \leq \max\{[d(x,u) + d(y,v)], [d(x,v) + d(y,u)]\}$$

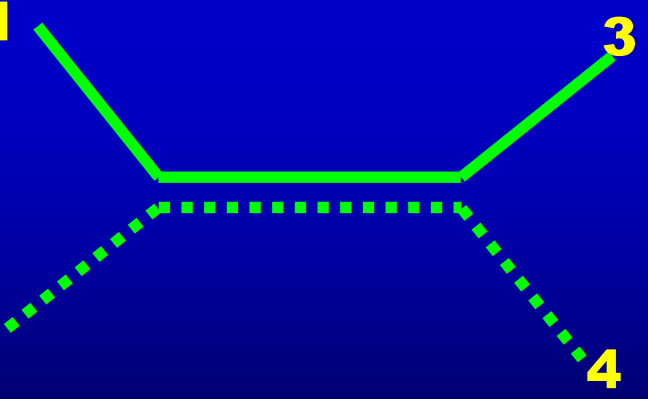


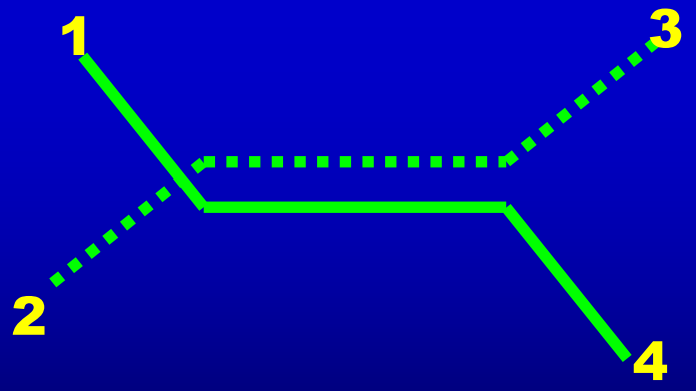


Aditividade



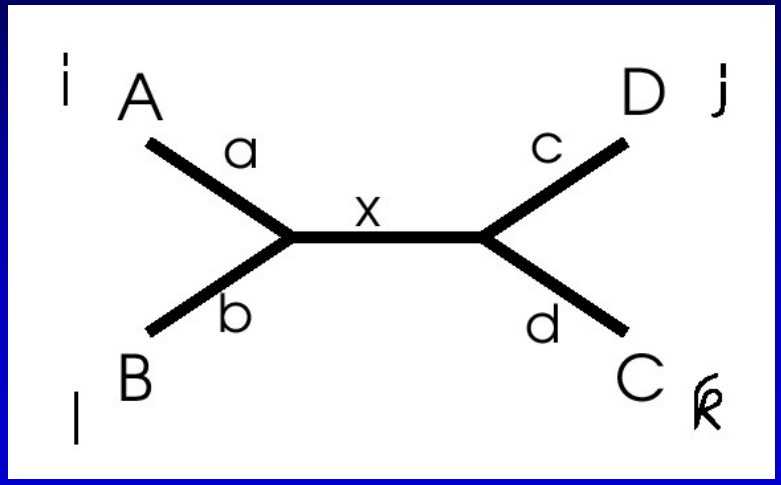






$$d_{12} + d_{23} < d_{13} + d_{24} = d_{14} + d_{23}$$

Entendendo a Aditividade



$$d_{ij} + d_{kl} = d_{ik} + d_{jl} \geq d_{il} + d_{jk}$$

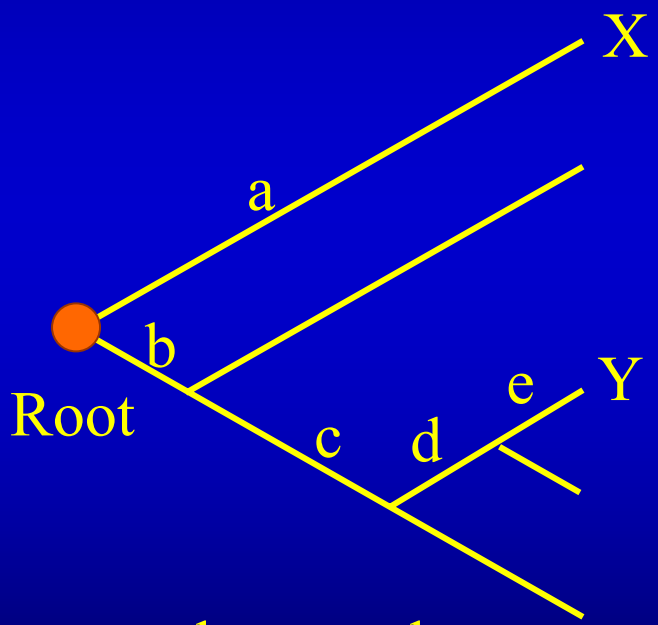
$$\begin{aligned} d_{ij} + d_{kl} &= a + b + c + d + 2x \\ d_{ik} + d_{jl} &= a + b + c + d + 2x \\ &= d_{il} + d_{jk} + 2x \end{aligned}$$

$$\begin{aligned} d_{ik} &= a + x + d \\ d_{jl} &= b + x + c \\ d_{ij} &= a + x + c \\ d_{kl} &= b + x + d \\ d_{il} &= a + b \\ d_{jk} &= c + d \end{aligned}$$

Tree Properties

Ultrametricity

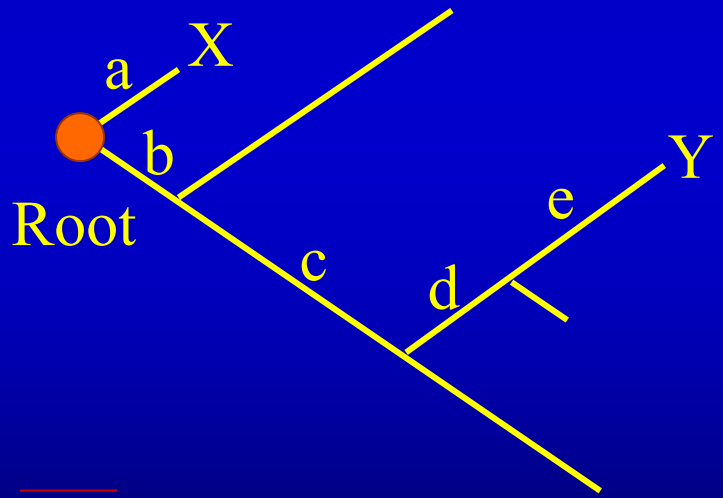
All tips are an equal distance from the root.



$a = b + c + d + e$

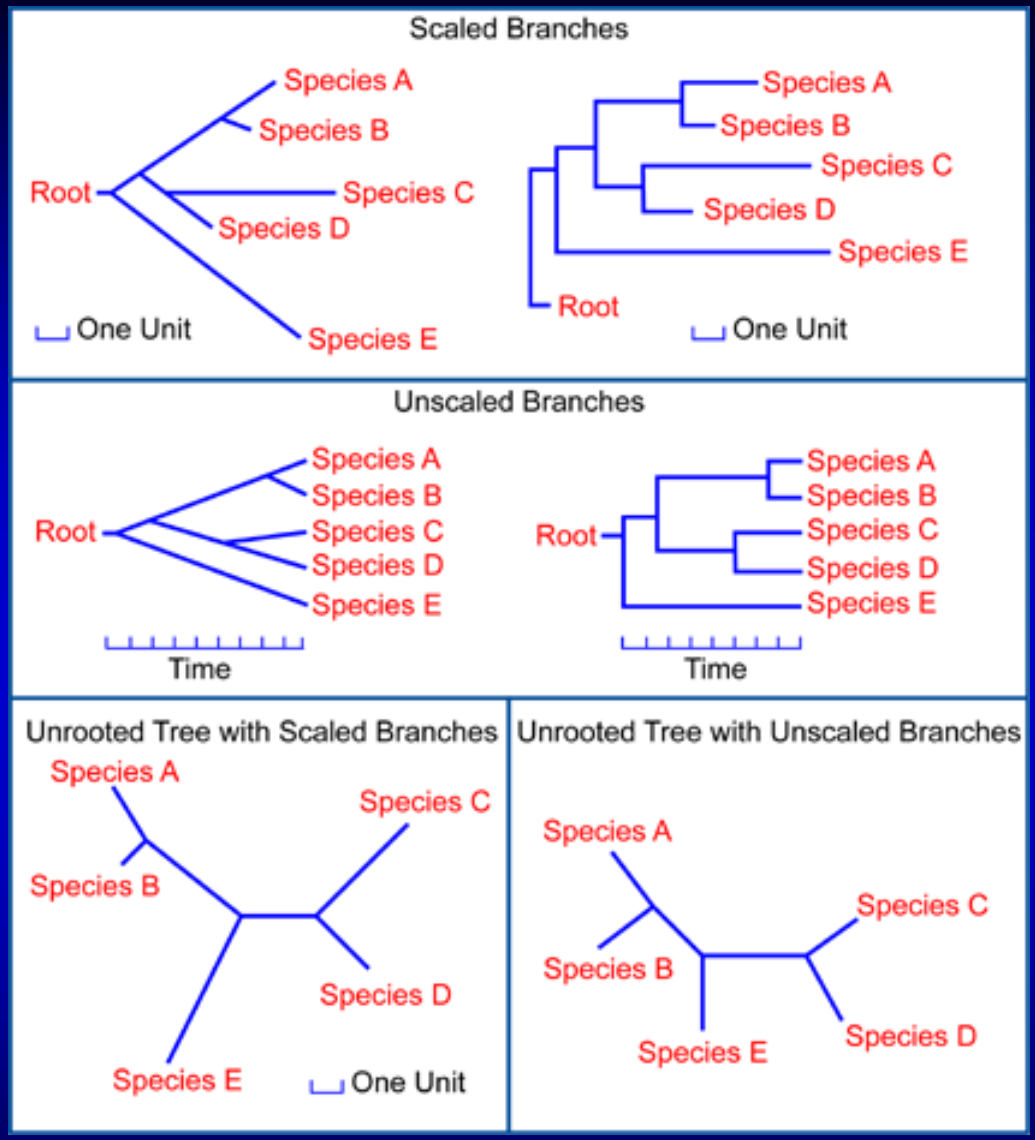
Additivity

Distance between any two tips equals the total branch length between them.



$\overline{XY} = a + b + c + d + e$

Tony Weisstein, http://bioquest.org:16080/bedrock/terre_haute_03_04/phylogenetics_1.0.ppt



- Image: http://www.ncbi.nlm.nih.gov/About/primer/phylo.html

Métodos de agrupamento (Clustering)

- UPGMA
- WPGMA
- Fitch-Margoliash
- Evolução mínima
- *Neighbor-Joining* (NJ)

Fitch-Margoliash – Preliminares

- **Se a matriz de distâncias original é aditiva, há algoritmos para encontrar a árvore cuja matriz de distâncias patrísticas é idêntica à matriz de distâncias original.**
- **O custo computacional destes algoritmos é $O(N^2)$.**
- **No entanto, se a matriz de distâncias original não for aditiva, a matriz de distâncias patrísticas não pode ser idêntica a ela e não há mais algoritmos tratáveis para se encontrar a árvore que melhor aproxime estas duas matrizes.**

Fitch-Margoliash – Preliminares

- M_{ij}: **distâncias originais**
- d_{ij}: **distâncias patrísticas**
- **Problema de otimização 1 (critério de Cavalli-Sforza & Edwards)**

$$\sum_{i,j} (M_{ij} - d_{ij})^2$$

- **Problema de otimização 2 (critério de Fitch-Margoliash)**

$$\sum_{i,j} \frac{(M_{ij} - d_{ij})^2}{M_{ij}^2}$$

Fitch-Margoliash – Preliminares

- **Repare que as distâncias patrísticas d_{ij} dependem da topologia e dos comprimentos dos ramos.**
- **Os dois problemas de otimização são intratáveis, pois requerem que todas as topologias de árvore sejam avaliadas.**

Exemplo passo-a-passo em material complementar.

Métodos de agrupamento (Clustering)

- UPGMA
- WPGMA
- Fitch-Margoliash
- Evolução mínima
- *Neighbor-Joining* (NJ)

Evolução Mínima – Preliminares

- **Minimize simultaneamente o critério de Fitch-Margoliash e o comprimento total dos ramos da árvore (princípio de evolução mínima).**
- **Como implementar?**

Algoritmo de Neighbor-Joining (NJ)