FILOGENIA MOLECULAR

Daniel Macedo de Melo Jorge

danielmacedo.jorge@gmail.com

SUMÁRIO

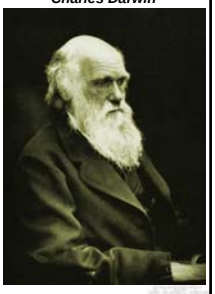
- Filogenia
- Conceitos
- Alinhamentos
- Análises e Métodos
- Principais programas
- Aplicações:
  - ✓Sistemática Molecular
  - ✓Estudo de Famílias Gênicas
  - ✓Desenvolvimento de Novas Drogas
  - ✓Forense

FILOGENIA

- **Definição**:

A relação histórica, resultante da evolução, entre *taxa* terminais, representada em forma de uma árvore ou a representação em forma de árvore da história evolutiva

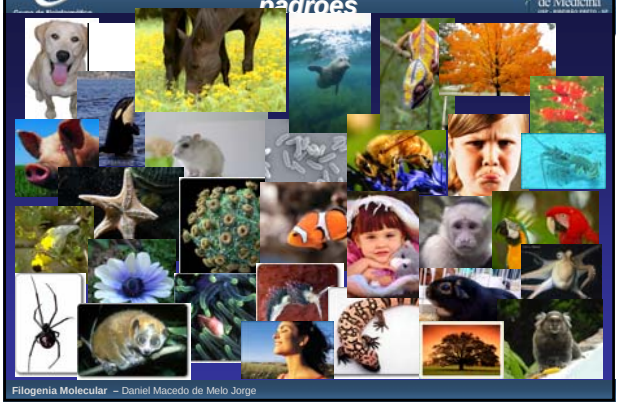
Charles Darwin



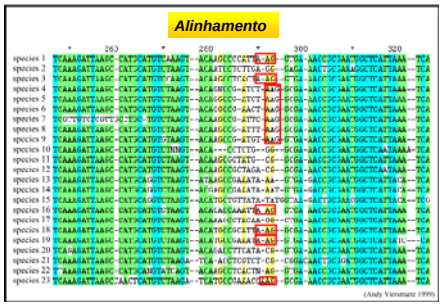
FILOGENIA

- **Objetivos:**
  - Determinar a história evolutiva do gene, da função ou da espécie;
  - Caracterizar ancestrais;
  - Estimar tempo de divergência entre dois organismos desde o último ancestral compartilhado
  - Caracterizar famílias gênicas e protéicas (uso de formas parálogas)

Os organismos possuem padrões



E as moléculas também



CONTRUÇÃO DE ÁRVORES FILOGENÉTICAS

•Como fazer uma árvore?

•Metodologia mais comum...

Apertar o *enter* várias vezes até aparecer uma árvore na tela

CONTRUÇÃO DE ÁRVORES FILOGENÉTICAS

...mas existe outra maneira

1) ***Alinhamento*** – estabelecimento de homologias

2) ***Estimativa da árvore*** – reconhecimento de padrões entre seqüências

3) ***Interpretação da árvore*** – interpretação das relações filogenéticas entre organismos

ALINHAMENTO

- Estabelece as posições homólogas entre as seqüências
- Homologia pode ser confiavelmente inferida a partir de alta similaridade
- Se o alinhamento estiver ruim, toda a análise proveniente dele também o será

CONCEITOS

A ESCOLHA DO GENE

- O segmento deve ter variabilidade compatível com problema filogenético (cheque GenBank)
- Seja criativo, use aminoácidos, terceiras posições, primeiras e segundas, todas elas
- Cheque seu alinhamento

ÁRVORES

Método de representação das relações (não necessariamente evolutivas) entre táxons

Podem ser divididas em:

- *Fenogramas* (baseadas em similaridade)
- *Cladogramas* (baseadas em relações evolutivas)

ÁRVORES

- ***Árvores sem raíz*** → reflete as relações entre os grupos constituintes sem cogitar, necessariamente, o caminho evolutivo dos mesmos

- ***Árvores com raiz*** → escolha de um "outgroup" que seja sabidamente distante de todos as outras entidades (considerado como possuindo um ancestral comum com todas as entidades)
  - Reflete o caminho evolutivo

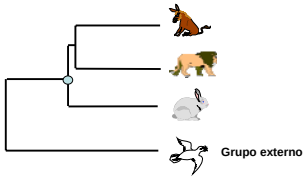
ÁRVORES

A maioria dos métodos filogenéticos produzem árvores sem raiz. Esses métodos geralmente detectam as diferenças entre as seqüências, mas não indicam e não orientam o quanto essas mudanças ocorreram a longo do tempo

- **Informações adicionais:**
  - Grupo externo
  - Hipótese de relógio molecular
  - Saturação dos dados

GRUPO EXTERNO

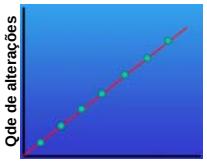
- O grupo externo usa táxons de referência que sabidamente se situam fora do grupo de interesse (o "grupo de dentro").
- Requer conhecimento a priori sobre relacionamentos entre os táxons.



RELÓGIO MOLECULAR

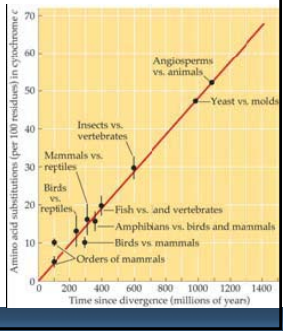
Mas o que é Relógio Molecular?

É um conceito baseado na idéia de que mutações espontâneas se acumulam a uma velocidade constante ao longo do tempo evolutivo em um determinado gene



RELÓGIO MOLECULAR

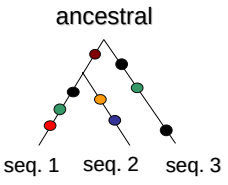
- *Quais genes são bons para serem usado em relógio molecular?*
  - Genes que tenham a mesma taxa de evolução em diferentes linhagens
- Na maioria das vezes deve ter a mesma função em todas as espécies (ex.: citocromo c)
- Relógio molecular conta o numero de gerações e não o numero de anos



SATURAÇÃO DOS DADOS

Perda do sinal filogenético

Quando comparadas seqüências homologas que sofreram uma grande quantidade de mutações ao longo do tempo, é impossível de determinar uma árvore filogenética independente da método ou modelo escolhido



MÉTODOS DE RECONSTRUÇÃO FILOGENÉTICA

-Fenéticos

-Cladísticos

-Probabilísticos

MÉTODOS FENÉTICOS

Árvores calculadas de acordo com a similaridade entre seqüências, sendo baseadas nos ***métodos de distância***. Esses métodos atribuem um valor às diferenças existentes entre duas seqüências.

DISTÂNCIA

- É a medida de diferença entre duas seqüências
- A mais simples é a distância p: onde p = nd/n (nd = número de diferenças, n = número total de sítios)

ATATCCTAGCTTAGATGTTAGG

ACATCCTAGCTTAAATGTTAGG

p = 2/22 = 0.09

DISTÂNCIA

•**Princípios gerais:**

1. Calcular a matriz de distância (a partir do alinhamento)
2. Construção da árvores a partir dos valores da matriz de distância

DADOS: ALINHAMENTO
W: AAGATACAG
X: AAGAGATAA
Y: AACCGATAA
Z: AACCGATAC

→

MATRIZ

	W	X	Y	Z
W		3	5	5
X			2	3
Y				1

→

MÉTODO DE AGRUPAMENTO:
UPGMA
e
NEIGHBOR JOINING

DISTÂNCIA

ESCOLHA DA DISTÂNCIA

- Se p < 0,1 não precisa corrigir
- Se 0,1<p < 0,25 pouca correção, 2 parâmetros no máximo (JC, K2, TajN)
- Se p > 0,25 corrija suas distâncias

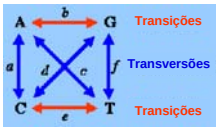
		1	2	3	4	7	8	9	10	Σdn
1	H2 D205B	-	0.17797	0.22824	0.17289	0.44499	0.46042	0.79221	**0.42042**	2.69714
2	H2 ROD-A	0.17797	-	0.20197	0.14244	0.42971	0.44477	0.78892	**0.410895**	2.59667
3	S MAC	0.22824	0.20197	-	0.125	0.46712	0.46687	0.78723	**0.437025**	2.715455
4	S MAN	0.17289	0.14244	0.125	-	0.431	0.44106	0.78538	**0.48319**	2.50096
7	S CFZ	0.44499	0.42971	0.46712	0.431	-	0.41043	0.7929	**0.13298**	3.10913
8	S AGM	0.46042	0.44477	0.46687	0.44106	0.41043	-	0.78327	**0.387845**	3.396665
9	HTLV1	0.79221	0.78892	0.78723	0.78538	0.7929	0.78327	-	**0.739565**	5.469415
10		**0.42042**	**0.410895**	**0.437025**	**0.48319**	**0.13298**	**0.387845**	**0.739565**		2.95136

DISTÂNCIA

MODELOS DE SUBSTITUIÇÃO DE NUCLEOTÍDEOS

•**Jukes & Cantor**

Assume que todos os nucleotídeos têm a mesma probabilidade de mudar para qualquer um dos outros nucleotídeos



•**Kimura-2-Parâmetros**

Assume taxas diferentes entre transições (A-G, C-T) e transversões (A-C, A-T, C-G, G-T).

DISTÂNCIA

MODELOS DE SUBSTITUIÇÃO DE NUCLEOTÍDEOS

•**Outros** (*Tajima Nei (ATGC) Tamura 3 (s/v + GC) Tamura Nei (s1s2/v + GC) HKY (s/v + ATGC)*):

Assumem mais parâmetros, que na média se aproximam mais da realidade, mas têm uma variância maior

•**Variáveis:**

- Correção Gama (Taxa para sítios heterólogos)
- Numero de sítios invariáveis

DISTÂNCIA

UPGMA
(Unweigthed Pair Group Method using arithmetic Averages)

- Sneath & Sokal, 1973
- Agrupamento seqüencial até a raiz
- Assume o relógio molecular
- Complexidade: $O(n^2)$, n quantidade de OTU's

DISTÂNCIA

UPGMA
(Unweigthed Pair Group Method using arithmetic Averages)

Vantagens e desvantagens:

- Quando os dados cumprem a proposta do algoritmo, o resultado fornece uma árvore filogenética ótima
- É um método rápido
- A desvantagem principal é que para dados reais, raramente se cumpre a propriedade ultra-métrica

DISTÂNCIA

EVOLUÇÃO MÍNIMA

- Cavalli-Sfoorza & Edwards, 1967
- Minimiza o somatório dos ramos das árvores
- Busca exaustiva em todas as árvores

Ex.: 4 organismos ou taxa: 3 possíveis árvores

DISTÂNCIA

NEIGHBOR-JOINING

- Saitou & Nei, 1987
- Baseado no princípio de evolução mínima (Cavalli-Sforza & Edwards, 1967), mas é muito mais rápido
- Identifica os vizinhos que sequencialmente minimizam o tamanho total da árvore (S)
- Um dos algoritmos mais usados, eficiente e rápido

DISTÂNCIA

NEIGHBOR-JOINING

- O método começa com uma árvore em forma de estrela
- O primeiro passo é separar o par de OTUs mais próximo (a partir dos valores da matriz de distância), separando dos outros
- Novamente une-se os ramos que apresentam as seqüências mais próximas, a partir da matriz de distância
- Este procedimento é repetido até que todos os ramos são encontrados

DISTÂNCIA

Métodos baseados em distâncias

Vantagens:

- Método simples e muito rápido;
- Pode ser aplicado em bases de dados muito extensas.

Desvantagens:

- Não considera os dados originais, apenas as distancias.

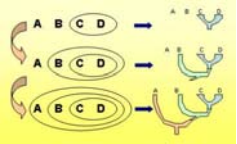
MÉTODOS CLADÍSTICOS

- As árvores são calculadas levando-se em consideração os vários possíveis caminhos da evolução.
- São baseados no métodos de parcimônia.
- Esses métodos usam cada posição do alinhamento como informação evolutiva para construir a árvore.
- Sítios Informativos: aqueles que apresentam pelo menos 2 tipos de caracteres, cada um ocorrendo duas vezes.

```
a. A A G A G T T C A T
b. A G C C G T T C T T
c. A G A T A T C C A T
d. A G A G A T C C T T
                 *   *   *
```


Máxima Parsimonia (MP)

- Supõe que o caminho evolutivo mais provável é o mais simples, que se explica com o menor número de mudanças (substituições).
- Baseado na informação proporcionada pelos caracteres derivados compartidos (SINAPOMORFIAS).



Máxima Parsimonia (MP)

- No alinhamento utiliza somente os sítios informativos
- Funciona melhor com o relógio molecular
- Busca heurística
- Busca *branch & bound*
- Busca exaustiva

Máxima Parsimonia (MP)

- **Vantagens**
  - Método com premissas simples;
  - Deve resultar em dados confiáveis:
  - Homoplasia e substituições múltiplas forem raras ou distribuídas aleatoriamente na topologia.
  - Amostragem densa.
- **Desvantagens**
  - Pode levar a resultados errados se homoplasia for comum ou concentrada em partes específicas da árvore. Exemplo:
  - Desvio de composição de bases;
  - Atração de ramos longos. Mais de uma linhagem acumulando substituições em velocidade diferente das demais linhagens.
  - Exige muito esforço computacional (Tempo).

MÉTODOS PROBABILÍSTICOS

- Usam cálculos probabilísticos para encontrar a árvore que melhor explica a variação dado o conjunto de seqüências
- Máxima verossimilhança (Maximum Likelihood)
- Inferência Bayesiana

Máximo verossimilhança

- O processo de substituição segue um modelo probabilístico onde a expressão matemática, mas não os valores dos parâmetros são conhecidos a priori.
- Os sítios evoluem independentemente
- Todos os sítios seguem o mesmo processo de substituição
- As probabilidades de substituição não mudam com o tempo (elas podem variar entre os ramos)

Máximo verossimilhança

Vantagens:

- Utiliza todos os sítios;
- É o melhor modelo de um ponto de vista teórico;
- Permite considerar modelo evolutivo específico;
- Estimativa acurada dos tamanhos de ramo (substituições/ sítio) mesmo quando ocorre substituições múltiplas;
- Menos sensível a atração de ramos longos se o modelo for adequado.

Desvantagens:

- Valores dos parâmetros são estimados e fixados a partir dos dados e podem não refletir adequadamente as taxas naturais de substituição e freqüências se a amostragem não for significativa;
- Praticamente impossível avaliar todas as possíveis árvores, é feita uma exploração parcial das árvores.
- Exige muito esforço computacional (Tempo).

AVALIANDO A SIGNIFICÂNCIA DA ÁRVORE

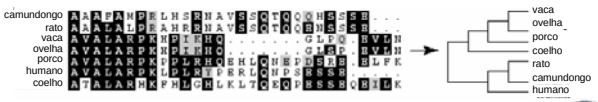
Teste de Confiança:

BOOTSTRAP

Teste estatístico para medir o grau de suporte dos nós nas árvores filogenéticas pelo alinhamento das seqüências

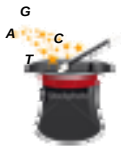
BOOTSTRAP

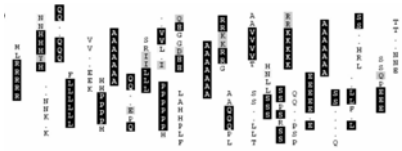
- **Alinhamento inicial e a árvore são formados**
- **O alinhamento inicial e a ávore são feitos da forma usual.**
- **Cada sítio alinhado é então considerado independente**



BOOTSTRAP

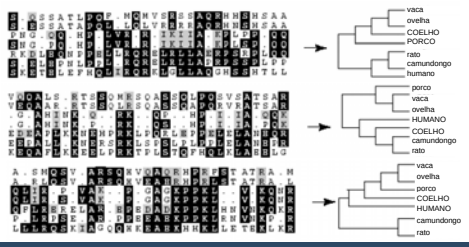
- **Jogue todos os sítios em um chapéu**
- **Amostras aleatórias dos sítios (com reposição) são sorteadas para construir uma nova árvore filogenética com o mesmo método inicial e mesmo tamanho de seqüência**
- **São feitas várias replicatas**





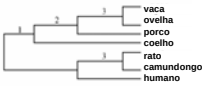
BOOTSTRAP

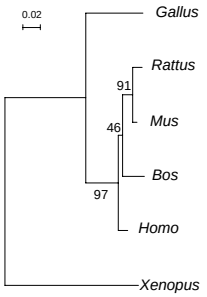
- **O novo alinhamento pode conter alguns sítios múltiplas vezes**
- **Outros sítios podem estar ausentes**



BOOTSTRAP

- **Um valor de bootstrap para cada nó**
- **O valor de bootstrap representa número de vezes que o agrupamento ocorreu nas replicações**

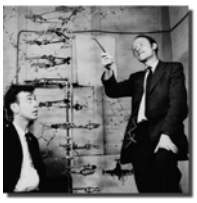




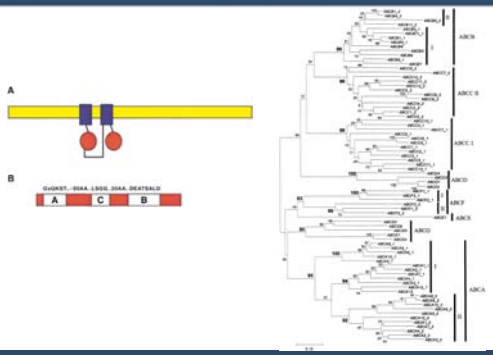
BOOTSTRAP

- O ideal é fazer 1000 ou mais replicatas
- Um agrupamento tem um bom suporte se ocorreu em mais de 80% das árvores (idealmente em mais de 95%)
- Pode ser chamado de nível de confiança da ocorrência para um determinado clado (agrupamento)
- Não avalia a exatidão de uma árvore, somente indica a consistência e estabilidade de clados individuais
- Outros métodos: jacknifing

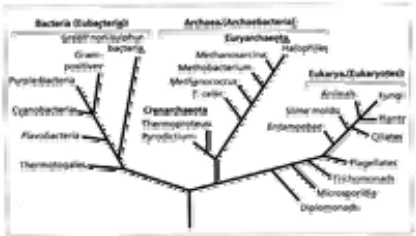
APLICAÇÕES DA FILOGENIA MOLECULAR



SISTEMA DE TRANSPORTE ABC

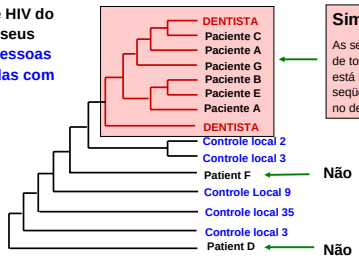


OS CINCO REINOS SÃO TRÊS



FORENSE: O DENTISTA ASSASSINO

Árvore filogenética das seqüências de HIV do DENTISTA, seus pacientes, e pessoas locais infectadas com HIV



Sim:
As seqüências de HIV de todos estes pacientes está no clado das seqüências encontradas no dentista.

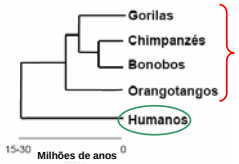
Não

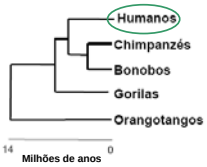
Não

Ou et *al.* (1992) and Page & Holmes (1998)

Freeman and Herron, 1998

A ORIGEM DO *HOMO SAPIENS* REVISTA

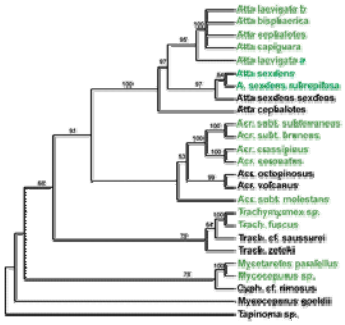




Na era pré-molecular pensava-se que os grande macacos formavam um clado separado dos humanos e que os humanos divergiram dos macacos no mínimo 15-30 milhões de anos.

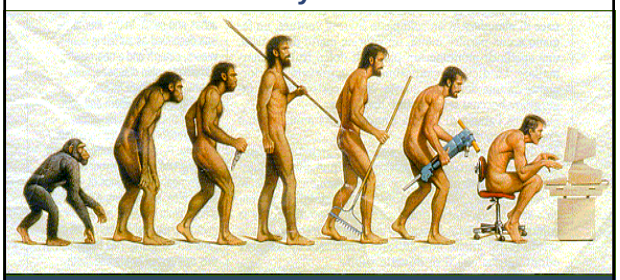
DNA mitocondrial, genes nucleares e hibridação DNA/DNA mostram que os bonobos e chimpanzés são mais relacionados aos humanos que aos gorilas.

Acromyrmex constituem um único gênero?

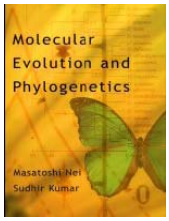


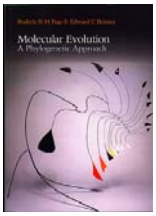
"A classificação por descendência não pode ser inventada por biólogos, ela pode apenas ser descoberta"

Theodosius Dobzansky



LITERATURA RECOMENDADA





WWW recursos para filogenia

- **Compilações**
  - ✓ Lista de sites e recursos: http://www.ucmp.berkeley.edu/subway/phylogen.html
  - ✓ Uma grande quantidade de programas de filogenia http://evolution.genetics.washington.edu/phylip/software.html
- **Databases of rRNA sequences and associated software**
  - ✓ The rRNA WWW Server - Antwerp, Belgium. http://rrna.uia.ac.be
  - ✓ The Ribosomal Database Project - Michigan State University http://rdp.cme.msu.edu/html/
- **Database similarity searches (Blast)** :
  - ✓http://www.ncbi.nlm.nih.gov/BLAST/
  - ✓http://www.infobiogen.fr/services/menuserv.html
  - ✓http://bioweb.pasteur.fr/seqanal/blast/intro-fr.html
  - ✓http://pbil.univ-lyon1.fr/BLAST/blast.html

WWW recursos para filogenia

A. Alinhamento de seqüências
   1. ClustalX : multiple sequence alignment with a graphical interface (for all types of computers). http://www.ebi.ac.uk/FTP/index.html and go to 'software'
   2. ClustalW: (XXXXXXXXXXXX)
   3. MUSCLE (http://www.drive5.com/muscle/)

B. Editores de seqüências
   1. Seqlab (disponível no pacote GCG em socrates)
   2. Jalview (http://www.jalview.org/)
   3. Bioedit (http://www.mbio.ncsu.edu/BioEdit/page2.html)

WWW recursos para filogenia

C. Conversão de formatos
   1. Readseq (http://www.ebi.ac.uk/cgi-bin/readseq.cgi)

D. Visualizador e editor de árvores
   1. Treeview (todas as versões para PC e Mac OS9; versão básica para UNIX/LINUX/; http://taxonomy.zoology.gla.ac.uk/rod/treeview.html)
   2. TreeExplorer in MEGA3.1 (http://www.megasoftware.net/)
   3. Hypertree (for large trees; http://www.kinase.com/tools/HyperTree.html)

E. Software de Filogenia
   1. MrBayes
   Análise Bayesiana (DNA or protein)
   Gratuito
   Muito flexível
   http://morphbank.ebc.uu.se/mrbayes/

WWW recursos para filogenia

E. Software de Filogenia

2. PAUP*

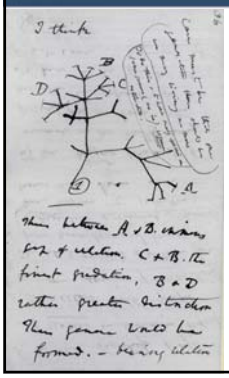
Maximum likelihood (DNA only), parsimony, distance
Pago ($85-$150; incluso upgrades)
O mais flexivel de todos.
http://paup.csit.fsu.edu/

3. Phylip

Maximum likelihood, parsimony, distance (DNA, protein, etc.)
Gratuito
O usuário deve fornecer o valor alpha para a taxa de heterogenicidade
http://evolution.genetics.washington.edu/phylip.html

4. Outros programas:

MEGA 3.1 (http://www.megasoftware.net)
Tree-Puzzle (http://www.tree-puzzle.de/)
PAML (http://abacus.gene.ucl.ac.uk/software/paml.html)



Muito obrigado a todos!!!!!!!!!