Universidade de São Paulo
Instituto de Química

USP

Análise filogenética para dados moleculares – aula 1

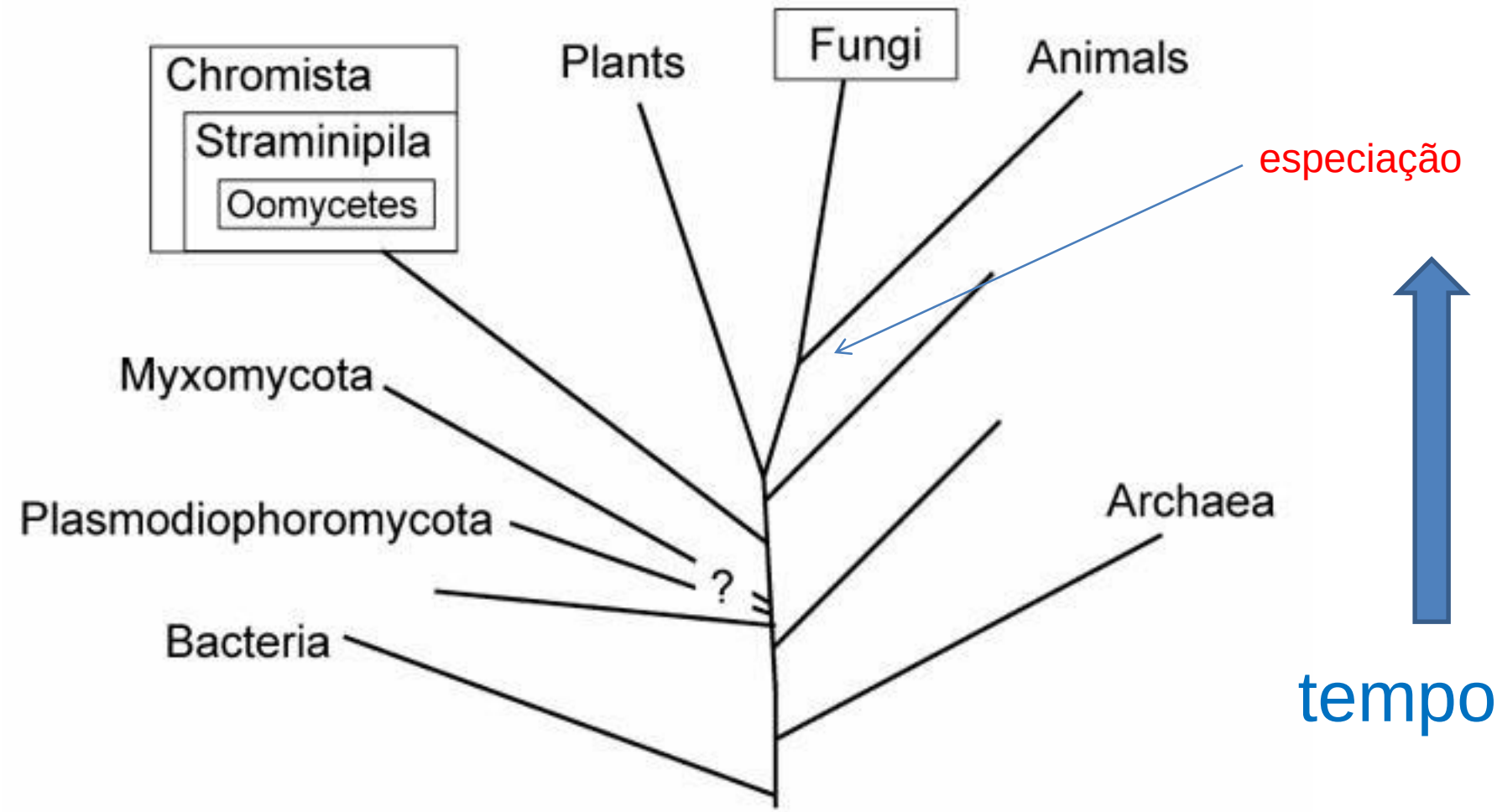
João Carlos Setubal

2021

Sumário

1. Conceitos básicos
2. Qual é a pergunta biológica?
3. Que sequências de entrada devem ser usadas?
4. Pipeline de análise: passos e componentes
5. Visualização da saída
6. Interpretação da saída

Uma filogenia é uma árvore



Credit: www.apsnet.org

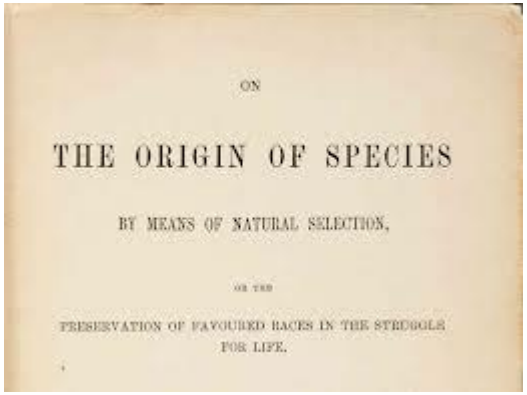
Uma árvore é uma hipótese sobre o que ocorreu na evolução

Pressupõe aceitação da idéia de que as espécies e as sequências de DNA evoluem ao longo do tempo

Para sequências de DNA a evolução é um **fato**

Para certas espécies e tempos geologicamente curtos a evolução é um **fato**

Somente para tempos geologicamente longos a evolução das espécies é uma **teoria** (Charles Darwin – seleção natural)

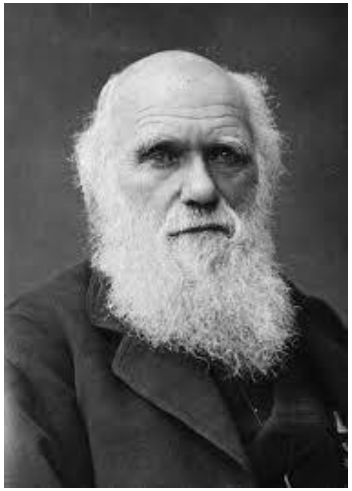

ON

THE ORIGIN OF SPECIES

BY MEANS OF NATURAL SELECTION,

OR THE

PRESERVATION OF FAVOURED RACES IN THE STRUGGLE FOR LIFE.



Evolução e tempo

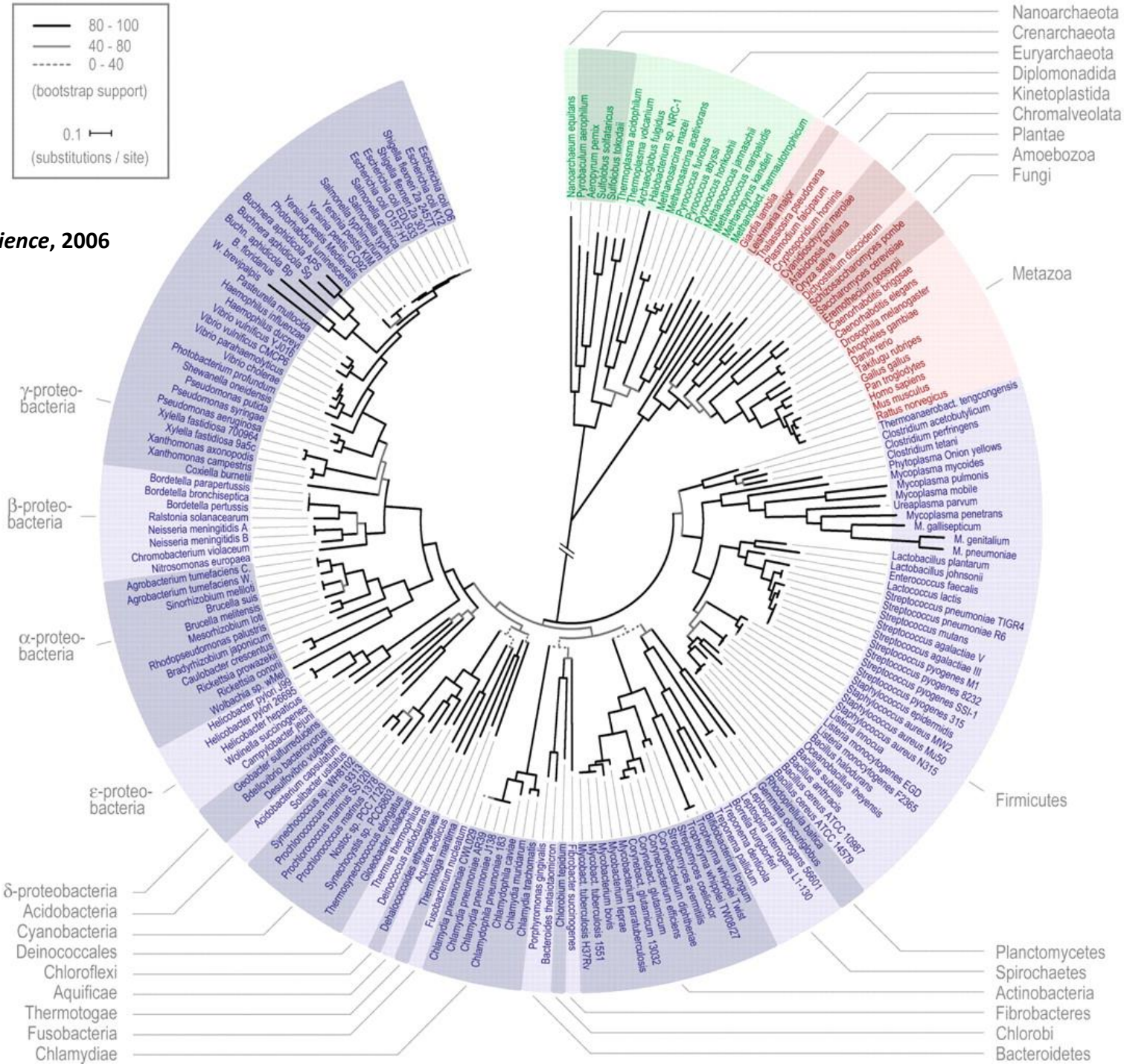
- Árvores com dados moleculares são hipóteses sobre quantas (e quais) mudanças ocorreram nas sequências
- Não são hipóteses sobre o tempo decorrido
- A menos que haja uma ligação entre mudança nas sequências e tempo
  - O relógio molecular

Problemas

- Evolução não é uniforme no tempo
- Ritmos diferentes
  - Espécies
  - Genes
  - Sítios de genes
- Relógio molecular supõe que existe uniformidade

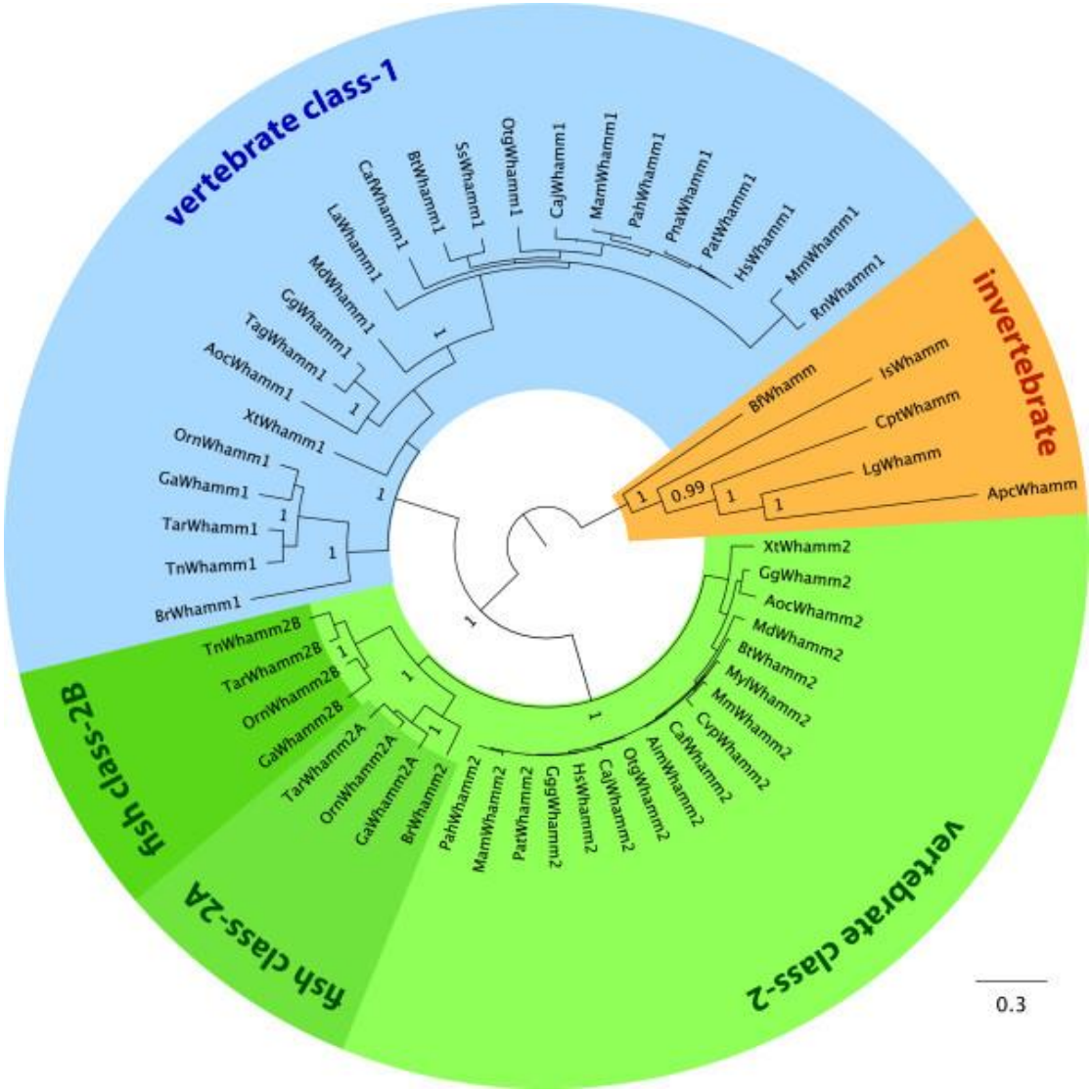
Exemplos de perguntas

- Como as espécies de interesse se relacionam evolutivamente?
- Qual é a história evolutiva de genes específicos?
  - Árvores de genes X árvores de espécies
- Próximo slide: exemplo de árvore de espécies



Ciccarelli et al, *Science*, 2006

Próximo slide: exemplo de árvore de genes



Phylogenetic tree of the WHAMM proteins

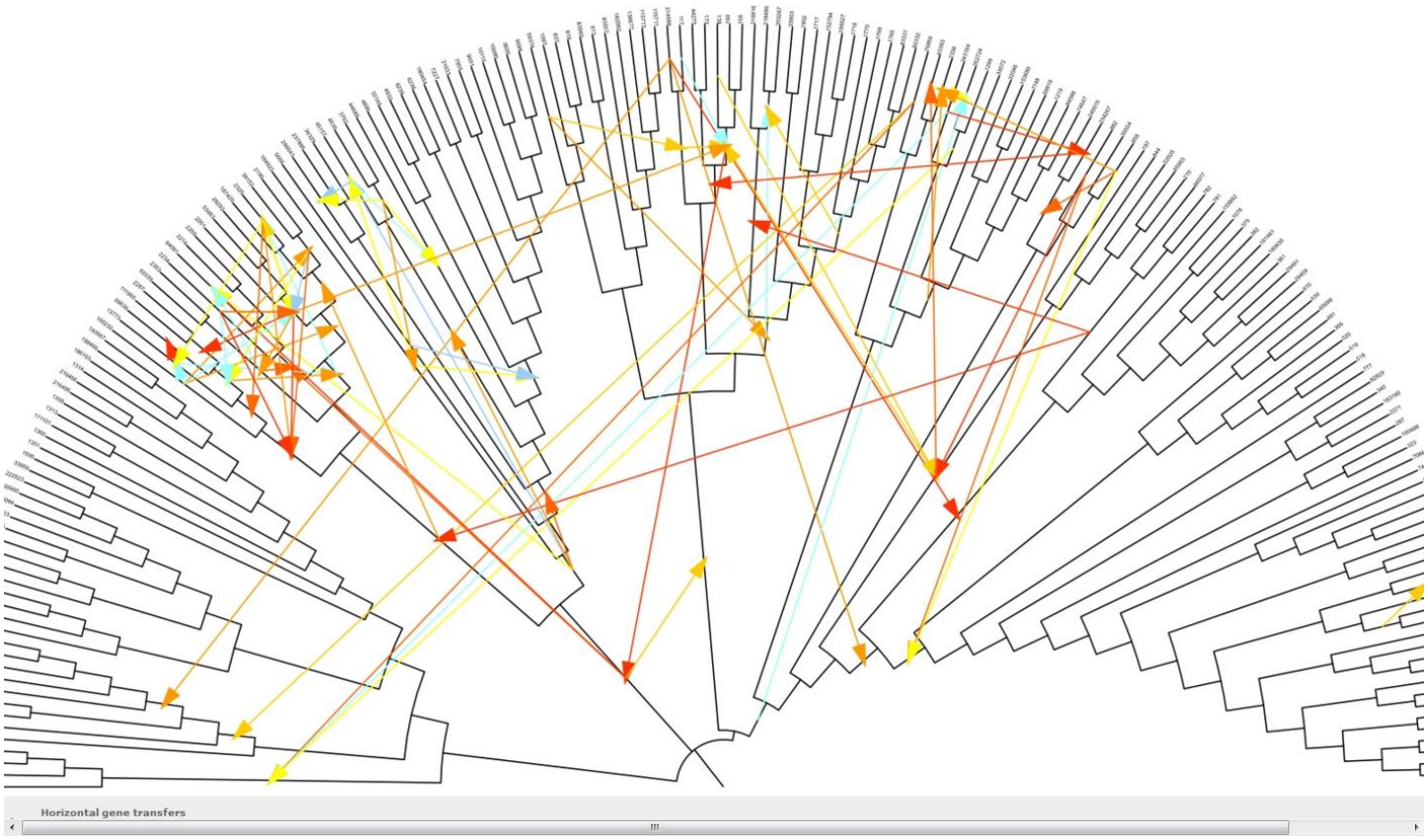
Kollmar *et al. BMC Research Notes* 2012 **5**:88 doi:10.1186/1756-0500-5-88

Por que a árvore de espécies pode ser diferente da árvore de genes?

- Principalmente por causa do fenômeno Transferência horizontal (ou lateral) de genes, geralmente abreviado por HGT
- A árvore de espécies supõe que estamos estudando o sinal filogenético vertical (de pai para filho)
- HGT ocorre quando material genético de um organismo é incorporado por outro
- ocorre em todos os organismos, inclusive humanos!
- mais comum em procariotos

Ilustração de transferência horizontal de genes

Esta figura mostra como poderíamos mapear eventos de HGT numa árvore de espécies



Aula 2 tem exemplos concretos de HGT

Árvores filogenéticas são importantes também ao nível de populações

- É um tipo de análise muito comum em epidemiologia
- especialmente para entender a origem de surtos
    - *Salmonella*
    - Ebola
    - SARS-Cov-2

É importante lembrar que taxonomia não é filogenia

- Kingdom: Chromalveolata
- Phylum: Heterokontophyta
- Class: **Oomycota**
- Orders (& families)
- Lagenidiales
  - Lagenidiaceae
  - Olpidiosidaceae
  - Sirolpidiaceae
- Leptomitales
  - Leptomitaceae
- Peronosporales
  - Albuginaceae
  - Peronosporaceae
  - Pythiaceae
- Rhipidiales
  - Rhipidaceae
- Saprolegniales
  - Ectrogellaceae
  - Haliphthoraceae
  - Leptolegniellaceae
  - Saprolegniaceae
- Thraustochytriales

Ao lado temos a hierarquia taxonômica para a classe Oomycota. Dentro de cada classe temos as ordens. Isso é taxonomia. Não há nenhuma hipótese sobre a história evolutiva das ordens e das famílias (ou seja: como as ordens se relacionam entre si? como as famílias dentro de uma ordem se relacionam entre si? essas perguntas não são respondidas pela hierarquia ao lado)

Sequências de entrada para inferência de uma filogenia

- Devem ser homólogas
- O problema do ovo e da galinha
  - devo fazer filogenia apenas para sequências homólogas
  - mas como saber se as sequências são homólogas sem fazer filogenia?
- Lidamos com isso usando similaridade (BLAST): recuperação inicial de possíveis sequências homólogas

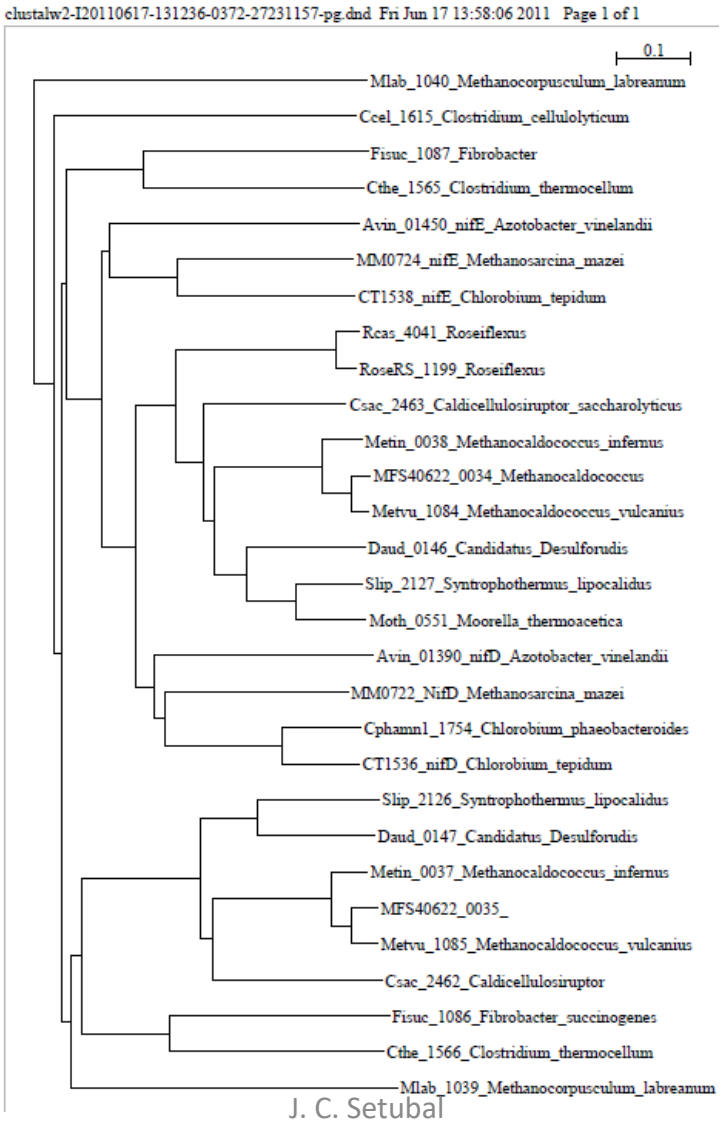
Etapas de processamento para se chegar a uma árvore

1. Alinhamento múltiplo
2. Edição do alinhamento
3. Reconstrução filogenética (inferência)
4. Visualização da árvore

Alinhamento múltipo

```
Cthe_1566_Clostridium_thermoce  LRDIIENTYK VLDT-DLQVV LTGCTAGIVG DDVDSLVSEF AQ--------
Fisuc_1086_Fibrobacter_succino  LDQLIKSTLK VFDG-DLYVV LTGCVGGLIG DDVPSLVNEY RD--------
Metvu_1085_Methanocaldococcus_  LVEGLRNLVA RYDP-ELISV VTTCSSETIG DDIEAFIRAA RKKIASEFGE
MFS40622_0035_                  LVEGLRNLVA RYDP-DLISV VTTCSSETIG DDIEAFIRAA RKKIAAEFGE
Metin_0037_Methanocaldococcus_  LVEGIRNLVA RYDP-DLISV VTTCSSETIG DDIEAFIRAA RKKIAKEFGE
Csac_2462_Caldicellulosiruptor  LIEGIRNLVL RYSP-TVIGV ITTCSSETIG DDIEAFIKEA YKKLSEELSS
Daud_0147_Candidatus_Desulforu  FTEGIRNLVV RYRP-DLITV VTTCSSEIIG DDMVSFIKVA RKRLVSELGP
Slip_2126_Syntrophothermus_lip  VIEGIRNLVV RYWP-GLIGV VTTCSSEIMG DDMVSFLKEA RARLSREIGR
CT1536_nifD_Chlorobium_tepidum  LKVAIQEAYD LFHP-KAIAI FSTCPVGLIG DDVHAVAREM KEKLGD----
Cphamn1_1754_Chlorobium_phaeob  LKEAIQEAYD IFRP-KAIGI FSTCPVGLIG DDVHAVAREM KEKLGD----
MM0722_NifD_Methanosarcina_maz  LKKAIDEVVK IFNP-EAVTI CATCPVGLIG DDIEAVSREA EKEHG-----
Avin_01390_nifD_Azotobacter_vi  LAKLIDEVET LFPLNKGISV QSECPIGLIG DDIESVSKVK GAELS-----
Moth_0551_Moorella_thermoaceti  LEQACLEAIR LFPEAKGLII FTTCTTGLIG DDVQAVARSV EKKTG-----
Slip_2127_Syntrophothermus_lip  LKASCLEAFR LFPEARGMII FTTCTTGLIG DDVQGVARQV EKEVG-----
Daud_0146_Candidatus_Desulforu  LLKSALEAVR LFPEATGIIM YTTCTTGLIG DDIGSVAKQI ERETG-----
Metvu_1084_Methanocaldococcus_  LEKACLEAAA EFPEAKGIII YATCTTGLIG DNLGAVAKKV EEKIG-----
MFS40622_0034_Methanocaldococc  LEKACLEAAA EFPQAKGIII YATCTTGLIG DNLEAVARKV EEKIG-----
Metin_0038_Methanocaldococcus_  LEKACIEAAE EFPEAKGIFI YATCPTALIG DNLEAVARKV EEKIK-----
Csac_2463_Caldicellulosiruptor  LYNAIIEANQ EFPEAKAVFI YATCPTALIG DDLEAVAKKA SKAIG-----
RoseRS_1199_Roseiflexus         LLQSIIEANA EFPNAKAVFV YNTCSTALIG DDGRDVAKQA EAIIG-----
Rcas_4041_Roseiflexus           LLQSIIEASA EFPDAKAVFV YNTCSTALIG DDGRDVAKQA EAIIG-----
CT1538_nifE_Chlorobium_tepidum  LYKSLIELID QYQP-NAAFI YSTCIIGLIG DDIDAVCKKV AKEKG-----
MM0724_nifE_Methanosarcina_maz  LSNAIDELAG IYRP-PVIFV YSTCIVGIIG DDLEAVCKTA SKKHN-----
Avin_01450_nifE_Azotobacter_vi  LFHAIRQAVE SYSP-PAVFV YNTCVPALIG DDVDAVCKAA AERFG-----
Cthe_1565_Clostridium_thermoce  LANTIREVYE RTHA-NAIFV LTTCAAGIIG DDVESVCNEA EEELG-----
Fisuc_1087_Fibrobacter          LRQTIRDAKE RFNP-KAIFI GMACATAIIG EDIDSIAEEM EPEVG-----
Ccel_1615_Clostridium_cellulol  LVDSLNEVNS RYNP-KIIAV LTNCCADIIG DDVEGCIEGL PDEIR-----
Mlab_1039_Methanocorpusculum_l  LLNKILQECA SHHP-KFVAI LGTPVPALIG CDISGIATEV FDTTK-----
Mlab_1040_Methanocorpusculum_l  LCNAIDELLP QIQRPKVFLV YICCVLYLAG FDEQSTIDEL KKRNPD----
```

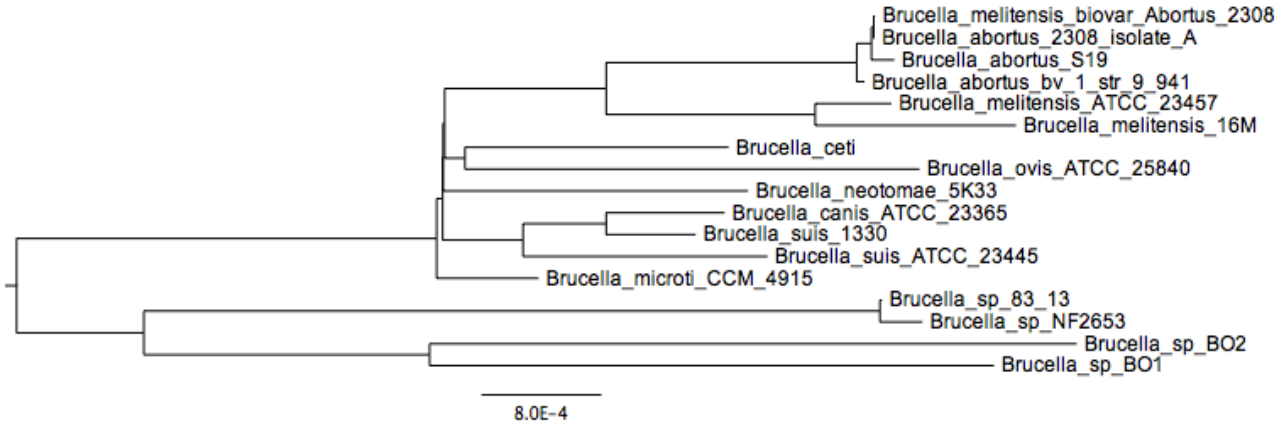
Filogenia resultante



Credit: R. Dixon

Uma árvore supõe que o comprimento de ramos significa distância evolutiva (p.ex. número de mutações). Se o comprimento dos ramos não tem esse significado, temos um cladograma

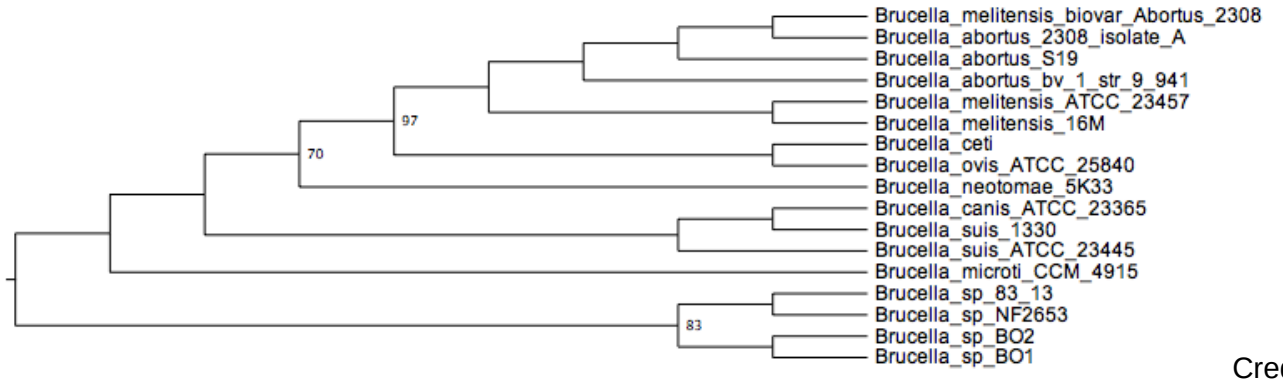
A



árvore

Cladogram version

B

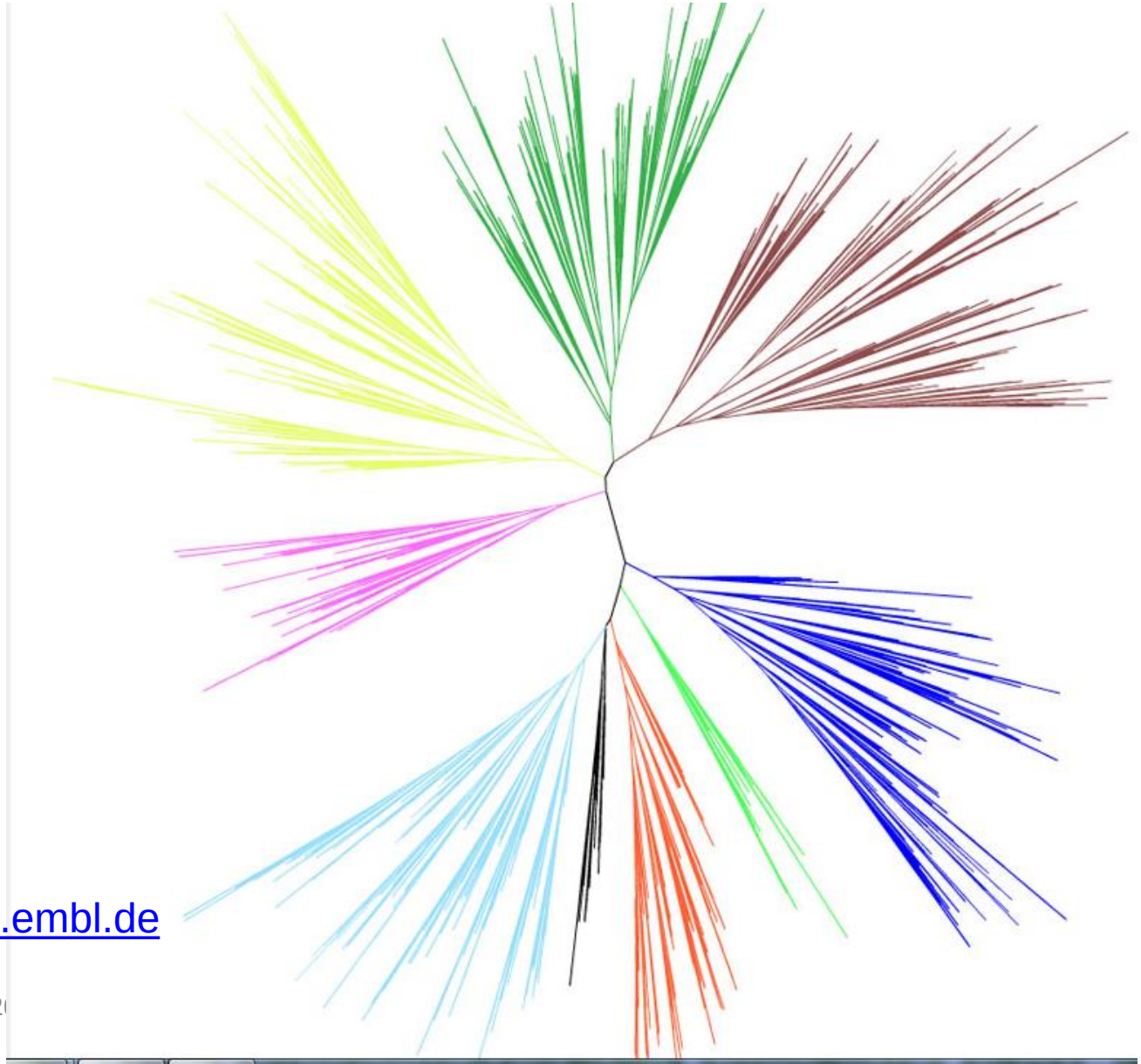


Credit: Wattam et al. 2011

Raiz de árvores filogenéticas

- Ao se inferir uma árvore filogenética, o resultado pode ser uma árvore com raiz ou sem raiz
- A raiz é o ponto da árvore que indica onde estaria o ancestral comum das espécies que estamos estudando
- na ausência de indicações específicas, as árvores inferidas não terão raiz

Árvore sem raiz

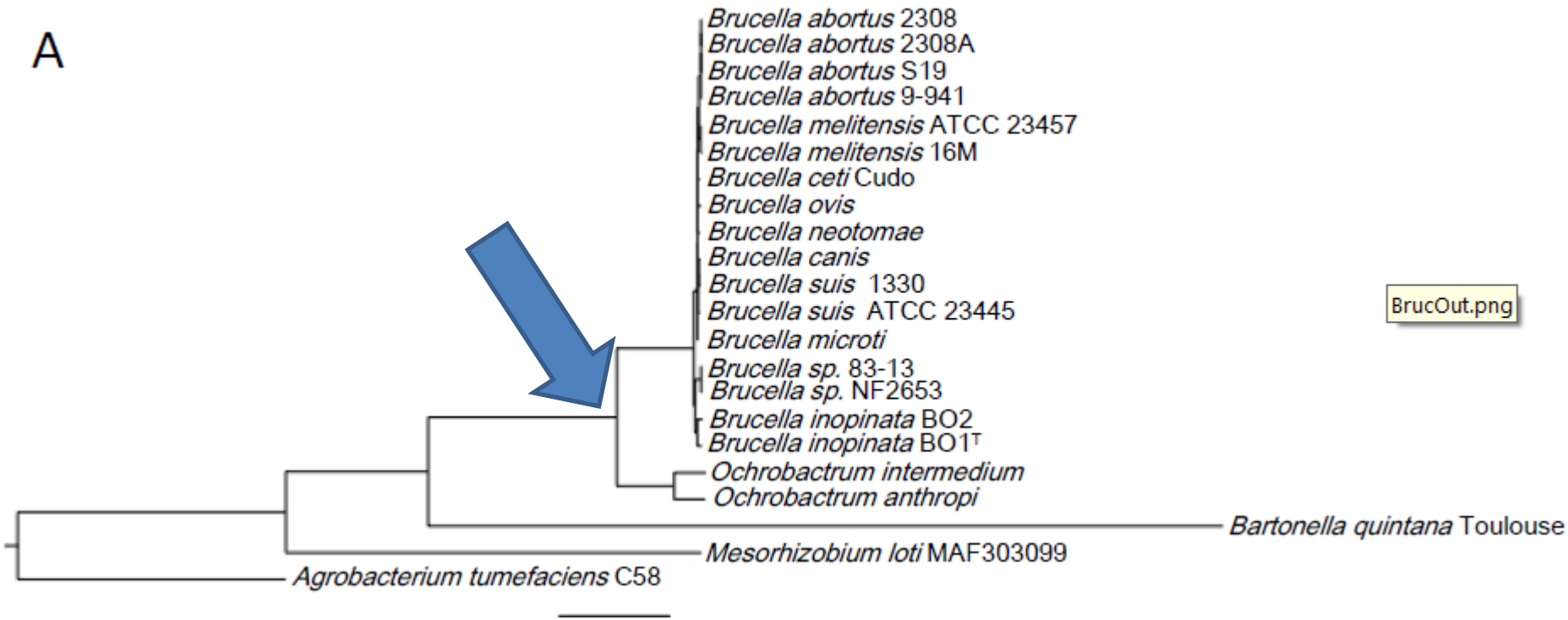


http://itol.embl.de

Inclusão de raiz

- É possível se enraizar uma AF por meio de "grupos externos"
- Neste caso, usamos conhecimento prévio de que certas espécies não pertencem ao grupo de espécies que estamos estudando. O ponto de junção do grupo de interesse com esses grupos externos nos dá a raiz da nossa árvore

Árvore com raiz



Neste caso, o grupo de interesse é Brucella + Ochrobactrum. Portanto Bartonella, Mesorhizobium, e Agrobacterium estão sendo usados como grupos externor

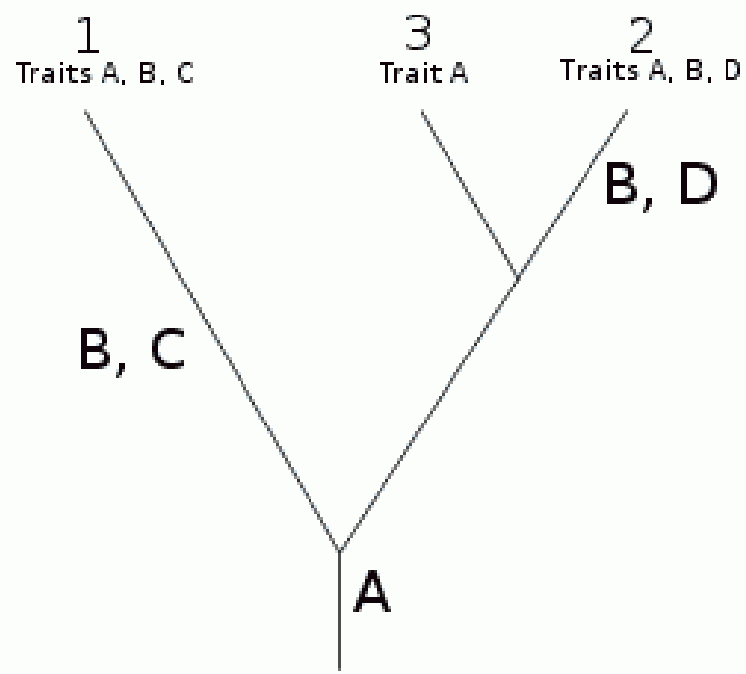
Métodos para reconstrução filogenética

- Distância
  - Matriz de distâncias
- Parcimônia
  - Minimizar as mutações ao longo dos ramos
- Máxima verossimilhança (likelihood)
  - Busca a árvore mais verossímil supondo um modelo probabilístico de evolução
- Inferência bayesiana
  - Também probabilístico, mas a abordagem é bayesiana

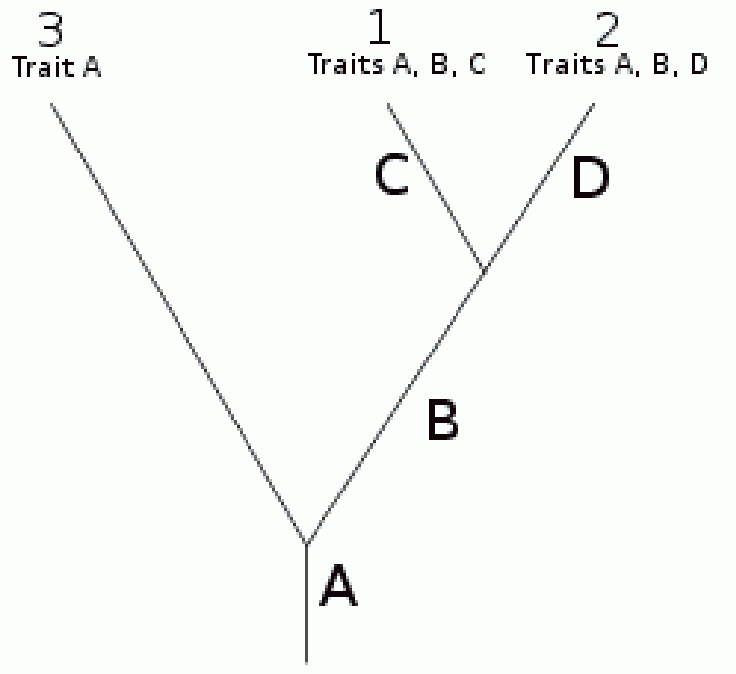
Árvore a partir de uma matriz de distâncias

- Métodos
  - UPGMA
  - Neighbor-joining (NJ)

Parcimônia



Less parsimonious cladogram



More parsimonious cladogram

http://palaeos.com/phylogeny/glossary.html

Parcimônia é um princípio muito usado em ciência

- As explicações mais simples são as mais próximas da “verdade”

ML e Bayesiano

- ML
  - Probabilidade (dados | modelo)
- Bayesiano
  - Probabilidade (modelo | dados)
- Dados são as sequências observadas
- Modelo ⇨ a árvore
- Bayesiano permite tratamento de incertezas nos dados

Probabilidade e verossimilhança (*likelihood*)

- Qual é a probabilidade de que uma moeda honesta jogada 100 vezes tenha como resultado “coroa” todas as vezes?
- Se uma moeda é jogada 100 vezes e resulta em coroa todas as vezes, qual é a verossimilhança de que a moeda seja honesta?
- Verossimilhança = função de um parâmetro (honestidade da moeda) dada uma observação (100 coroas consecutivas, ou *outcome*)
- A verossimilhança de um conjunto de valores de parâmetros dadas as observações é igual à probabilidade dessas observações dados esses valores
- L(theta(x)) = P(x | theta)
- L(100coroas) = P(honestidade|100coroas)

ML para inferência filogenética

- Avalia a probabilidade de que o modelo de evolução escolhido gerou os dados observados: P(D|H)
- Por exemplo, todos os nucleotídeos são igualmente prováveis
- O programa testa todos os possíveis nucleotídeos em cada nó interno da árvore e calcula a probabilidade de que essas escolhas teriam gerado os dados observados (as sequências das folhas)
- As probabilidades de todas as possíveis reconstruções são somadas para determinar a verossimilhança para cada site
- A verossimilhança da árvore é o produto das verossimilhanças para todas as posições do alinhamento

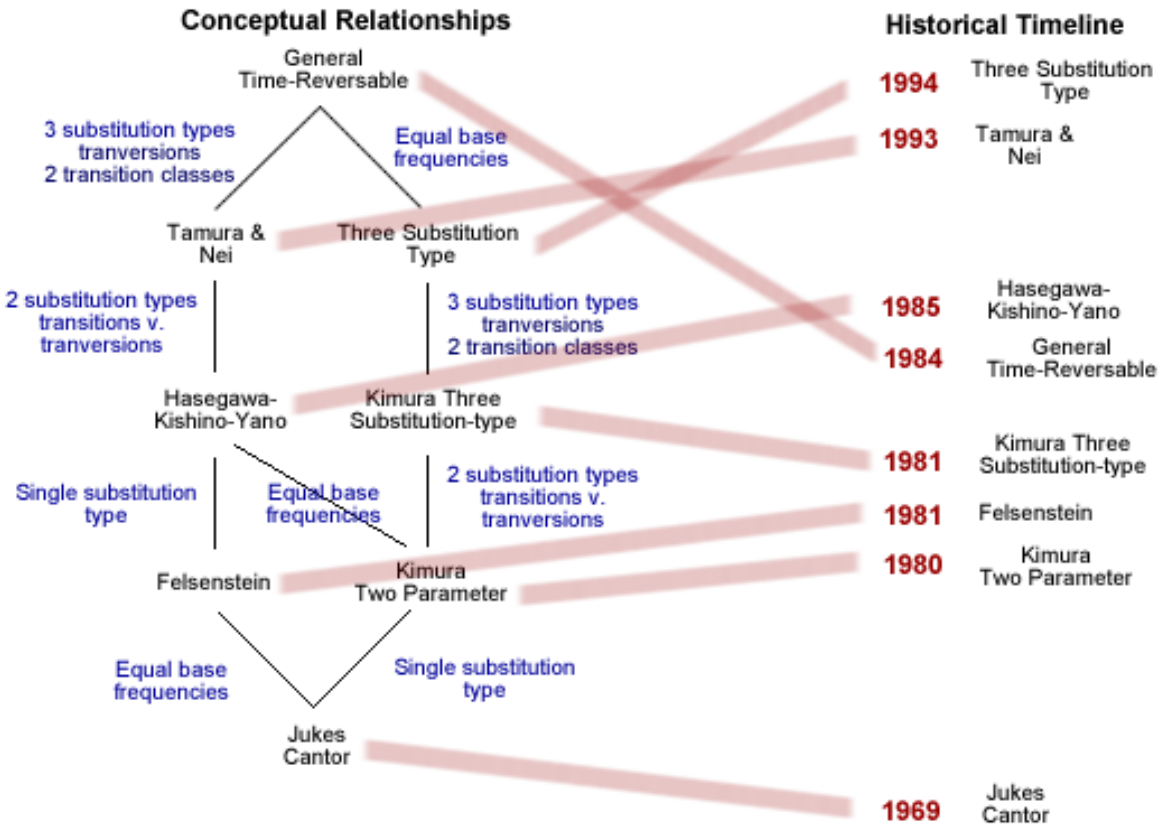
Considerações de tempo de execução

- Até aproximandamente o ano 2000, distância e parcimônia eram os métodos mais usados
  - os outros eram muito custosos computacionalmente
- Já faz uns 10 anos máxima verossimilhança se tornou “padrão”
- salvo situações específicas, não há desculpa para não usar ML na inferência

Modelos de evolução

- Exceto *distância*, todos os outros métodos dependem de modelos de evolução

Modelos de evolução para DNA



http://authors.library.caltech.edu/5456/1/hrst.mit.edu/hrs/evolution/public/models/sequence.html

Evolução de proteínas

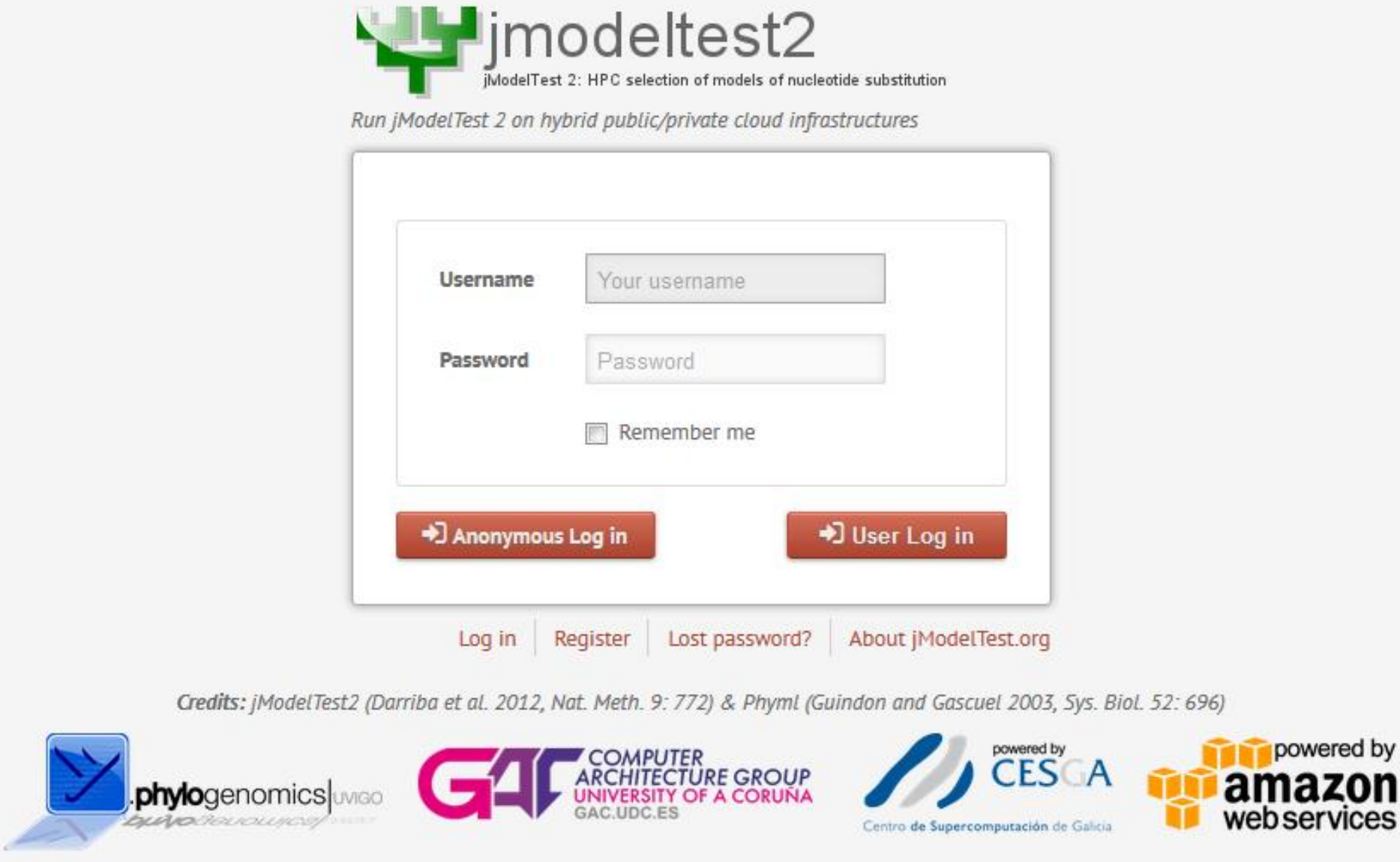
- Matrizes de substituição de aminoácidos
  - PAM
  - BLOSUM
  - WAG
    - Whelan and Goldman (2001) Mol. Biol. Evol. 18, 691-699

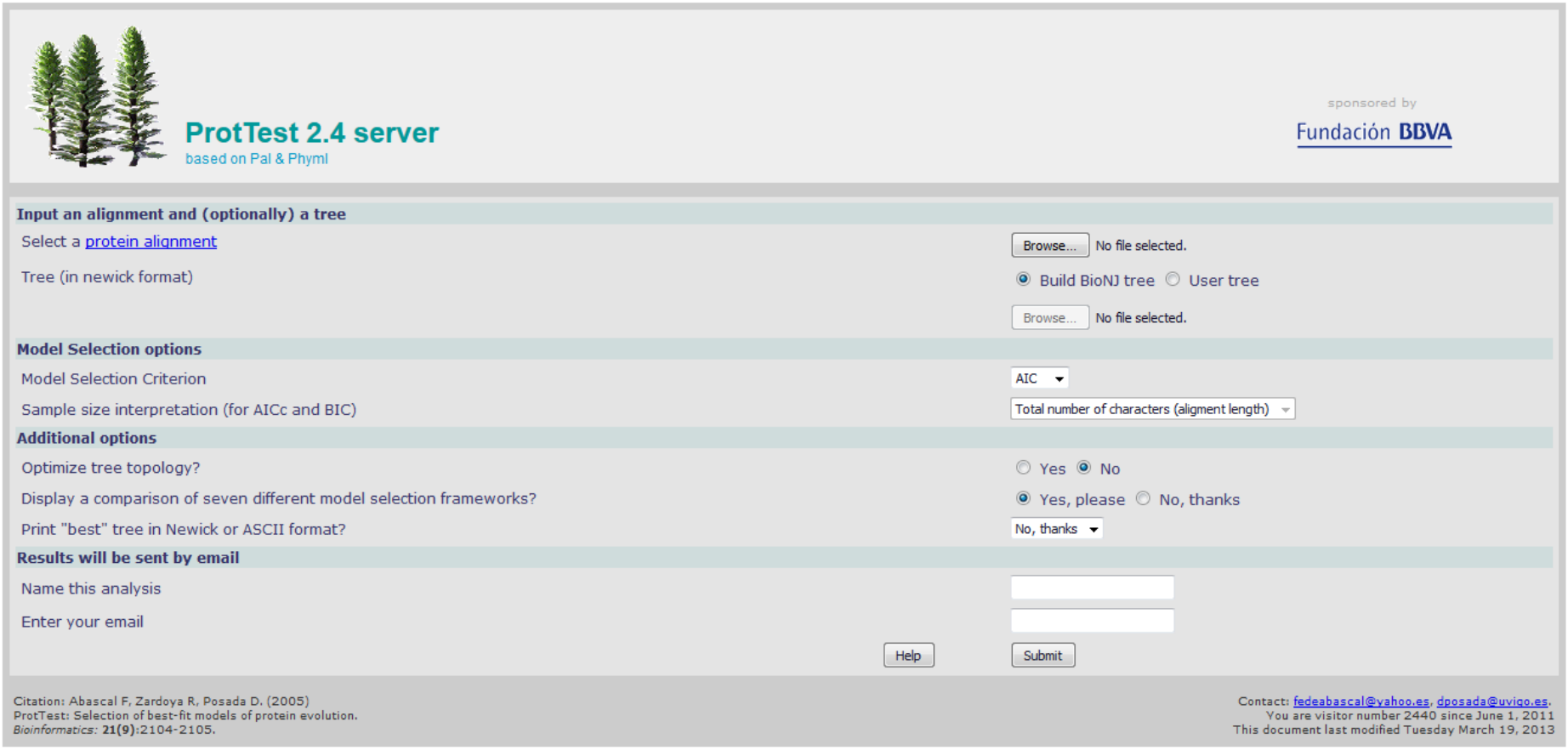
Modelos em PhyML

- **DNA**
  - JC69, K80, F81, F84, HKY85, TN93, GTR, custom
- **Aminoácidos**
  - LG, WAG, Dayhoff, JTT, Blosum62, mtREV, rtREV, cpREV,DCMut, VT, mtMAM, custom
- Todos eles supõem que cada *site* evolui de forma independente

Para escolher modelos

- Programas desenvolvidos pelo grupo de David Posada (Universidade de Vigo, Galicia, Espanha)
- ModelTest: para nucleotídeos
- ProtTest: para aminoácidos


jmodeltest2
jModelTest 2: HPC selection of models of nucleotide substitution
Run jModelTest 2 on hybrid public/private cloud infrastructures
Username
Your username
Password
Password
Remember me
Anonymous Log in
User Log in
Log in
Register
Lost password?
About jModelTest.org
Credits: jModelTest2 (Darriba et al. 2012, Nat. Meth. 9: 772) & Phyml (Guindon and Gascuel 2003, Sys. Biol. 52: 696)
phylogenomics UVIGO
COMPUTER ARCHITECTURE GROUP
UNIVERSITY OF A CORUÑA
GAC.UDC.ES
powered by CESGA
Centro de Supercomputación de Galicia
powered by amazon web services


ProtTest 2.4 server
based on Pal & Phyml
sponsored by
Fundación BBVA
Input an alignment and (optionally) a tree
Select a protein alignment
Browse... No file selected.
Tree (in newick format)
Build BioNJ tree
User tree
Browse... No file selected.
Model Selection options
Model Selection Criterion
AIC
Sample size interpretation (for AICc and BIC)
Total number of characters (aligment length)
Additional options
Optimize tree topology?
Yes
No
Display a comparison of seven different model selection frameworks?
Yes, please
No, thanks
Print "best" tree in Newick or ASCII format?
No, thanks
Results will be sent by email
Name this analysis
Enter your email
Help
Submit
Citation: Abascal F, Zardoya R, Posada D. (2005)
ProtTest: Selection of best-fit models of protein evolution.
Bioinformatics: 21(9):2104-2105.
Contact: fedeabascal@yahoo.es, dposada@uvigo.es.
You are visitor number 2440 since June 1, 2011
This document last modified Tuesday March 19, 2013