

Análise filogenética para dados moleculares

João C. Setubal

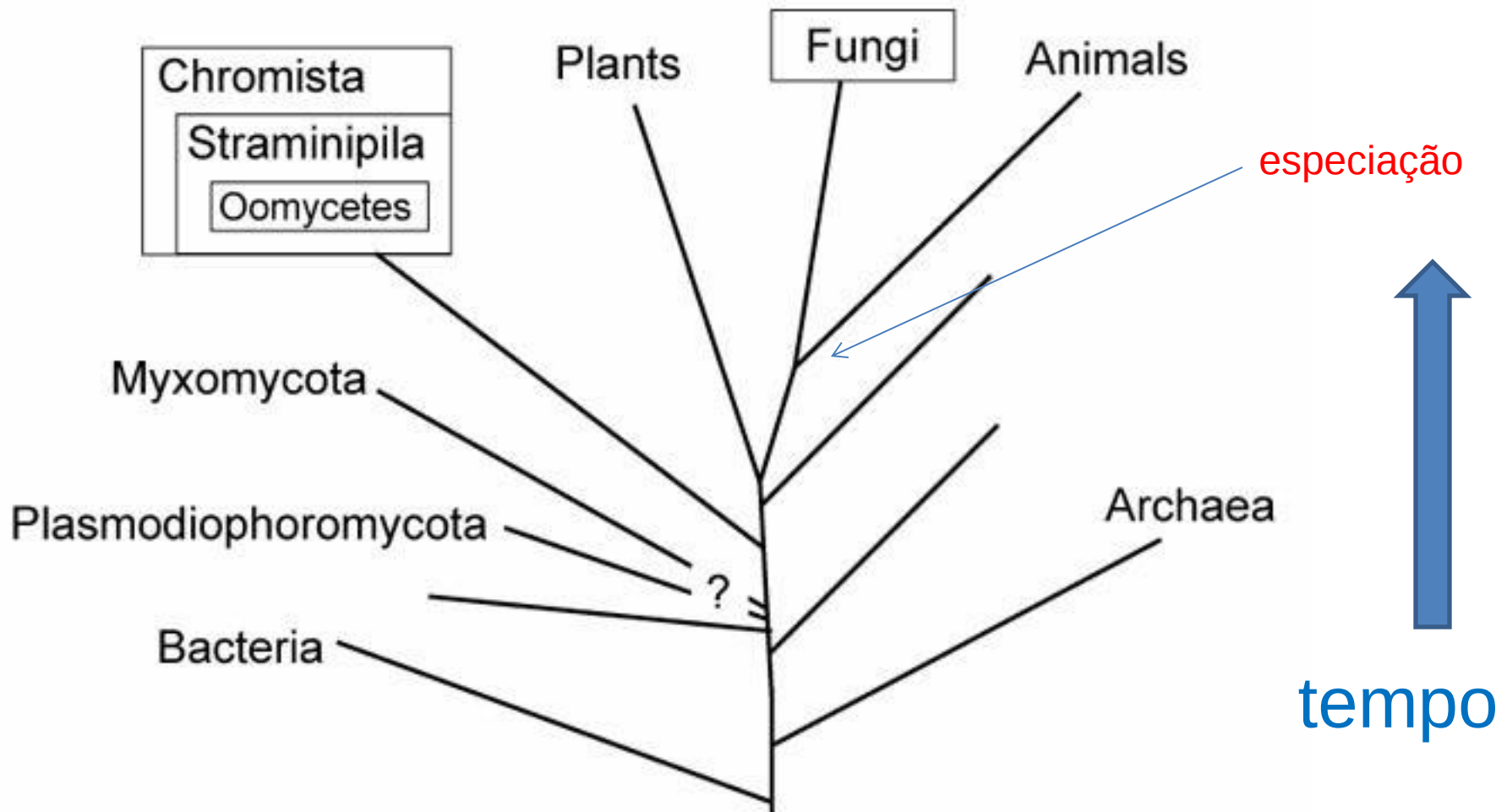
USP

2014

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



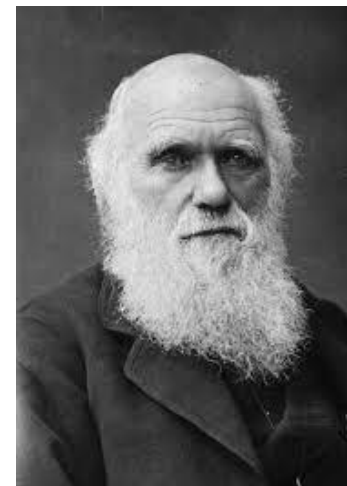
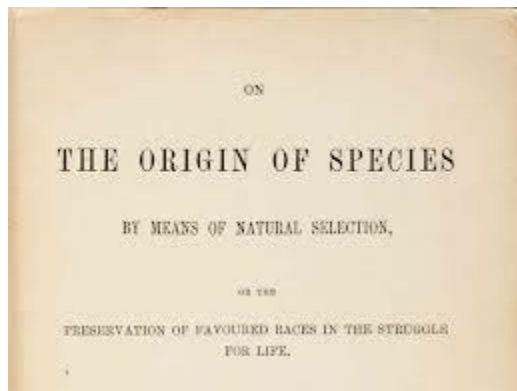
Uma árvore é uma **hipótese** sobre o que ocorreu na evolução

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)



Evolução e tempo

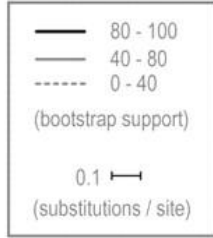
- Árvores com dados moleculares são hipóteses sobre quantas (e quais) mudanças ocorreram nas sequências
- Não são hipóteses sobre o tempo decorrido
- A menos que haja uma ligação entre mudança nas sequências e tempo
 - O relógio molecular

Problemas

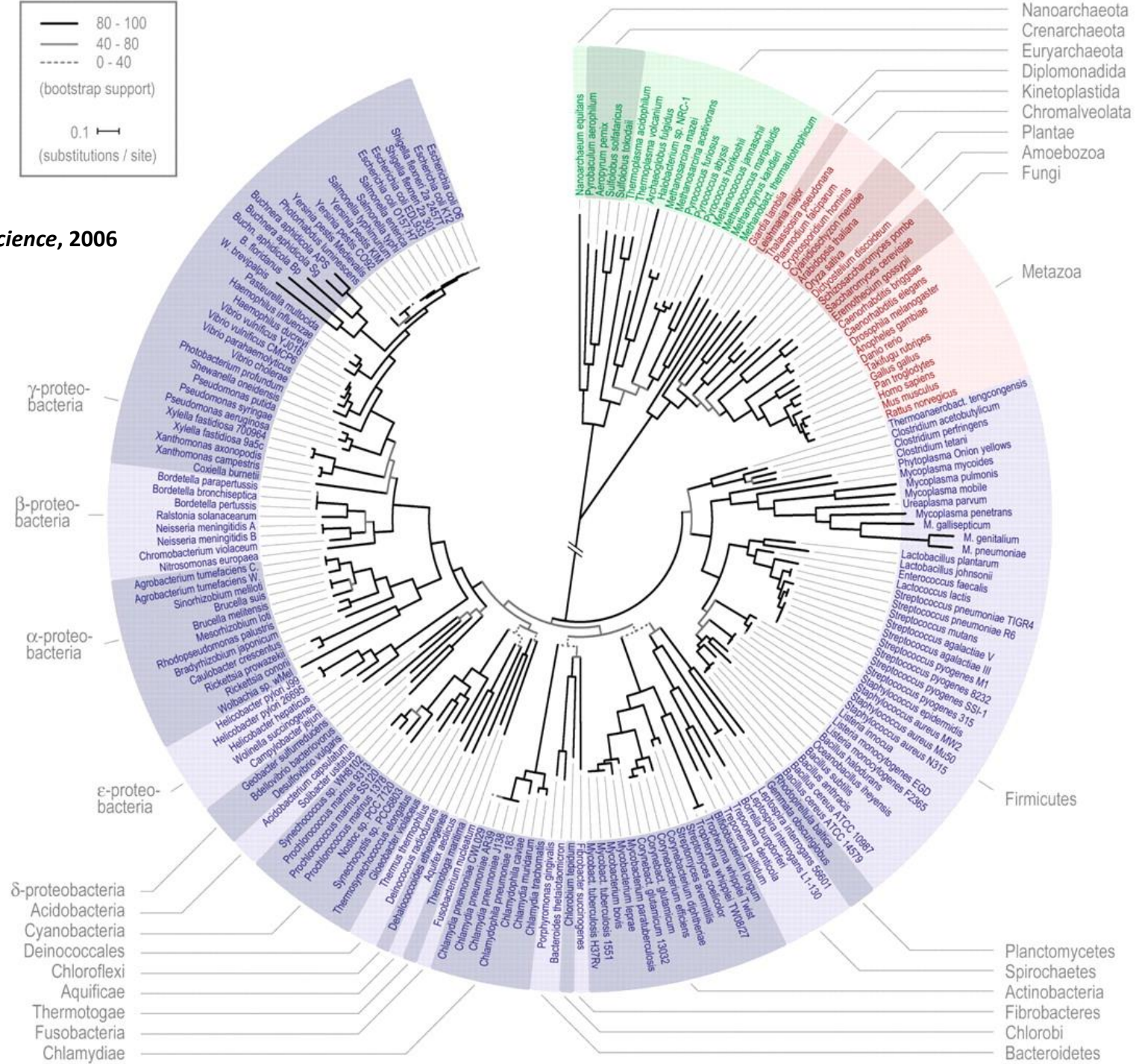
- Evolução não é uniforme no tempo
- Ritmos diferentes
 - Espécies
 - Genes
 - Sítios de genes
- Relógio molecular supõe que existe uniformidade

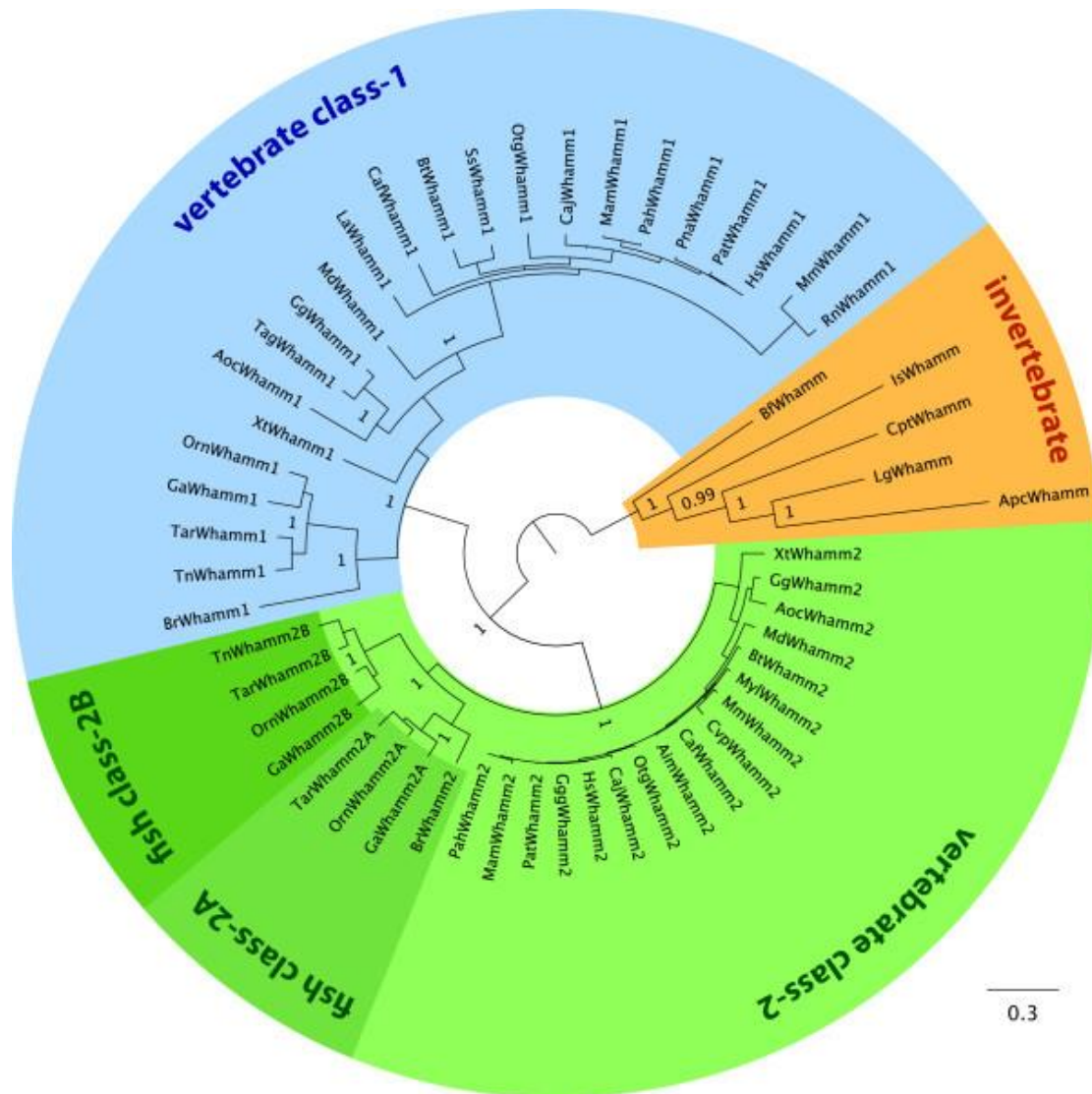
Exemplos de perguntas

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



Ciccarelli et al, *Science*, 2006



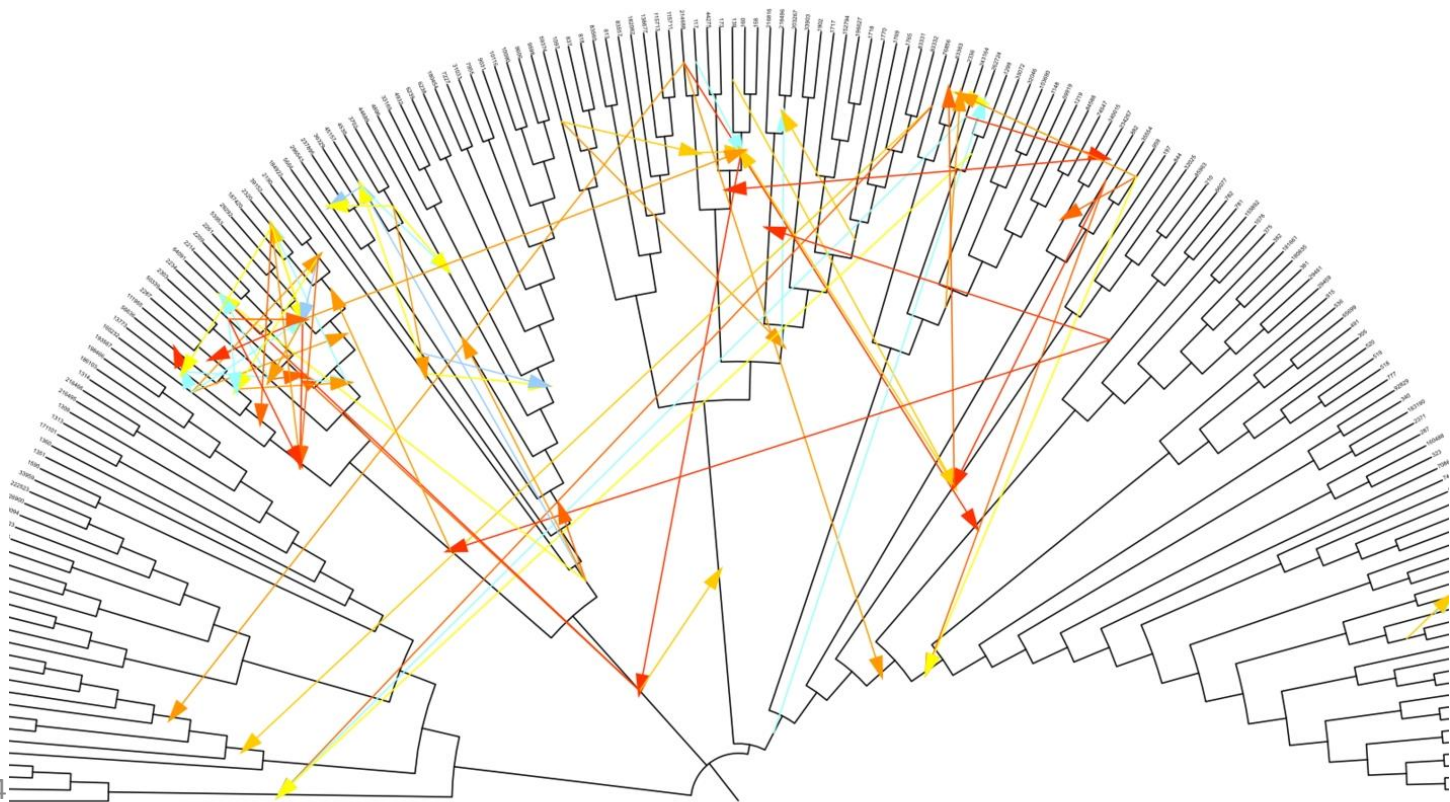


Phylogenetic tree of the WHAMM proteins

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

Transferência horizontal de genes

- É a principal razão que explica discrepâncias entre árvore de espécies e árvores de genes



História de populações

- Epidemiologia forense
 - Surto
 - *Salmonella*
 - Ebola
 - Antraz

Taxonomia não é filogenia

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

Phytophthora



Sequências de entrada

- Devem ser **homólogas**
- O problema do ovo e da galinha
- Similaridade (BLAST) pode usada para recuperação inicial de possíveis sequências homólogas

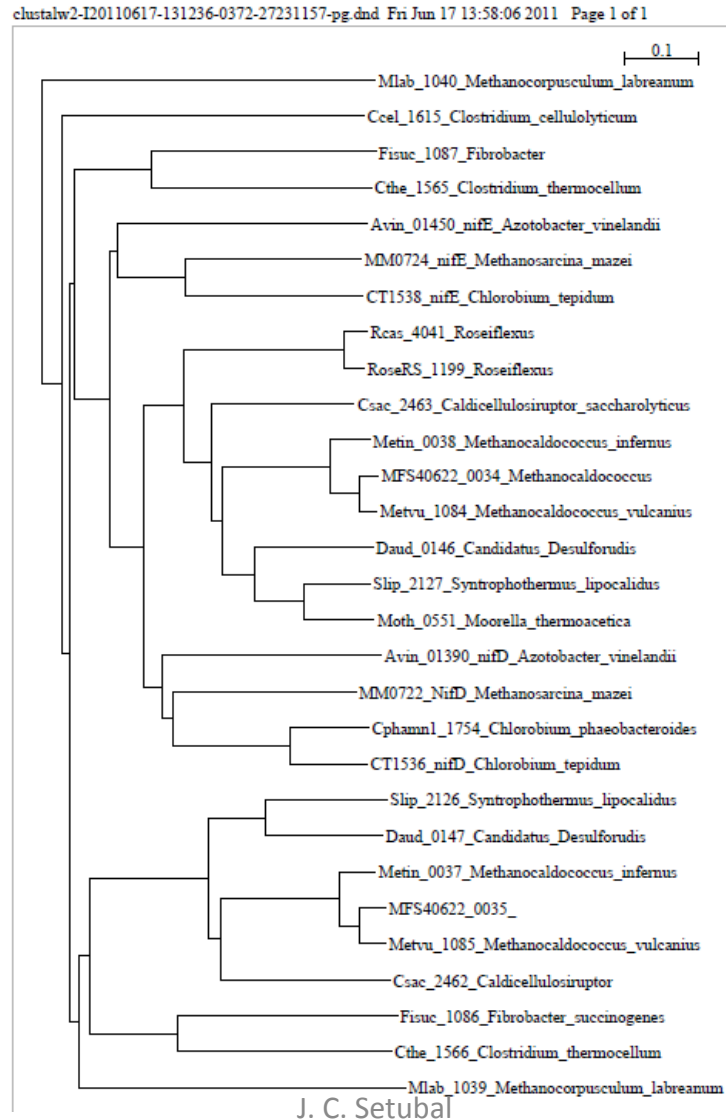
Pipeline

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

Alinhamento múltipo

Cthe_1566_Clostridium_thermoce	LRDIIENTYK	VLDT-DLOVV	LTGCTAGIVG	DDVDSLVSSEF	AQ-----
Fisuc_1086_Fibrobacter_succino	LDOLIKSTLK	VFDG-DLYVV	LTGCVGGLIG	DDVPSLVNEY	RD-----
Metvu_1085_Methanocaldococcus_	LVEGLRNLVA	RYDP-ELISV	VTTCSSETIG	DDIEAFIRAA	RKKIASSEFGE
MFS40622_0035	LVEGLRNLVA	RYDP-DLISV	VTTCSSETIG	DDIEAFIRAA	RKKIAAEFGE
Metin_0037_Methanocaldococcus_	LVEGIRNLVA	RYDP-DLISV	VTTCSSETIG	DDIEAFIRAA	RKKIAKEFGE
Csac_2462_Caldicellulosiruptor	LIEGIRNLVL	RYSP-TVIGV	ITTCSETIG	DDIEAFIKEA	YKKLSEELSS
Daud_0147_Candidatus_Desulforu	FTEGIRNLVV	RYRP-DLITV	VTTCSSEIIG	DDMVSVFIKVA	RKRLVSELGP
Slip_2126_Syntrophothermus_lip	VIEGIRNLVV	RYWP-GLIGV	VTTCSSEIMG	DDMVSVFLKEA	RARLSREIGR
CT1536_nifD_Chlorobium_tepidum	LKVAIQEAYD	LFHP-KAIAI	FSTCPVGLIG	DDVHAVAREM	KEKLGD----
Cphamnl_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	LEOACLEAIR	LFPEAKGLII	FTTCTTGLIG	DDVOAVARSV	EKKTG----
Slip_2127_Syntrophothermus_lip	LKASCLEAFR	LFPEARGMII	FTTCTTGLIG	DDVOGVAROV	EKEVG----
Daud_0146_Candidatus_Desulforu	LLKSALEAVR	LFPEATGIIM	YTTCTTGLIG	DDIGSVAKOI	ERETG----
Metvu_1084_Methanocaldococcus_	LEKACLEAAA	EFPEAKGIII	YATCTTGLIG	DNLGAVAKKV	EEKIG----
MFS40622_0034_Methanocaldococc	LEKACLEAAA	EFPOAKGIII	YATCTTGLIG	DNLEAVARKV	EEKIG----
Metin_0038_Methanocaldococcus_	LEKACIEAAE	EFPEAKGIFI	YATCPTALIG	DNLEAVARKV	EEKIK----
Csac_2463_Caldicellulosiruptor	LYNAIIEANO	EFPEAKAVFI	YATCPTALIG	DDLEAVAKKA	SKAIG----
RoseRS_1199_Roseiflexus	LLOSIIEANA	EFNNAKAVFV	YNTCSTALIG	DDGRDVAKQA	EAIIG----
Rcas_4041_Roseiflexus	LLOSIIEASA	EFPDAAKAVFV	YNTCSTALIG	DDGRDVAKQA	EAIIG----
CT1538_nifE_Chlorobium_tepidum	LYKSLIELID	OYOP-NAAFI	YSTCIIGLIG	DDIDAVCKKV	AKEKG----
MM0724_nifE_Methanosarcina_maz	LSNAIDELAG	IYRP-PVIFV	YSTCIVGIIG	DDLEAVCKTA	SKKHN----
Avin_01450_nifE_Azotobacter_vi	LFHAIRQAVE	SYSP-PAVAV	YNTCVPALIG	DDVDVAVCKAA	AERFG----
Cthe_1565_Clostridium_thermoce	LANTIREVYE	RTHA-NAIFV	LTTCAGGIIG	DDVESVCNEA	EEELG----
Fisuc_1087_Fibrobacter	LROTIRDAKE	RFNP-KAIFI	GMACATAIIG	EDIDSIAEEM	EPEVG----
Ccel_1615_Clostridium_cellulol	LVDSLNEVNS	RYNP-KIIAV	LTNCCADIIG	DDVEGCIEGL	PDEIR----
Mlab_1039_Methanocorpusculum_1	LLNKILOECA	SHHP-KFVAI	LGTPVPALIG	CDISGIATEV	FDTTK----
Mlab_1040_Methanocorpusculum_1	LCNAIDELLP	QIQRPKVFLV	YICCVLYLAG	FDEQSTIDEL	KKRNP----

Filogenia resultante

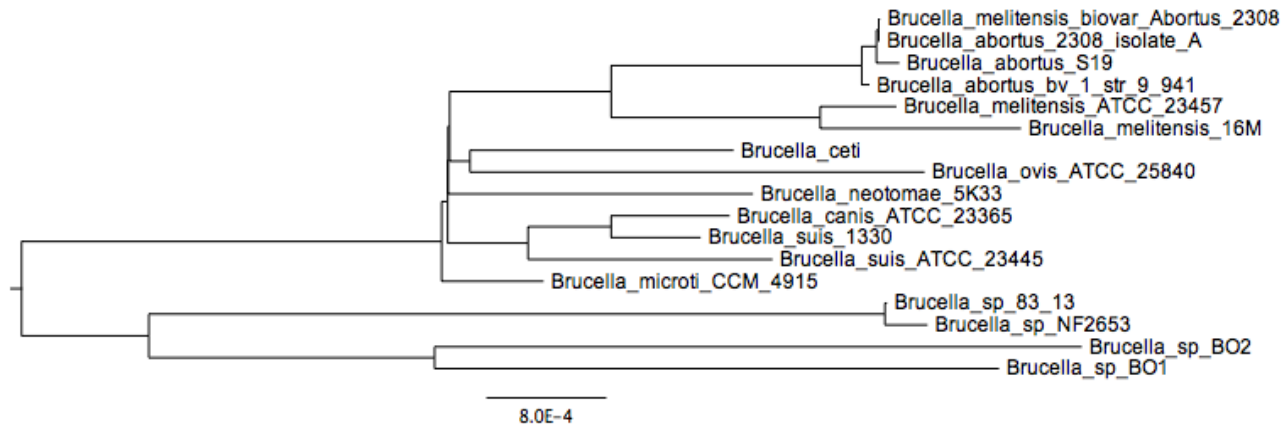


Credit: R. Dixon

Árvores e cladogramas

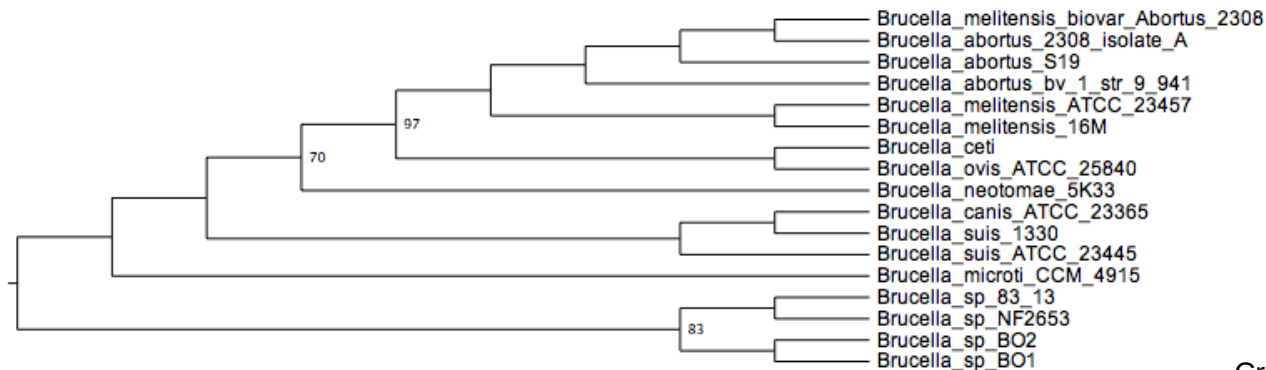
Topologia e comprimento de ramos

A



Cladogram version

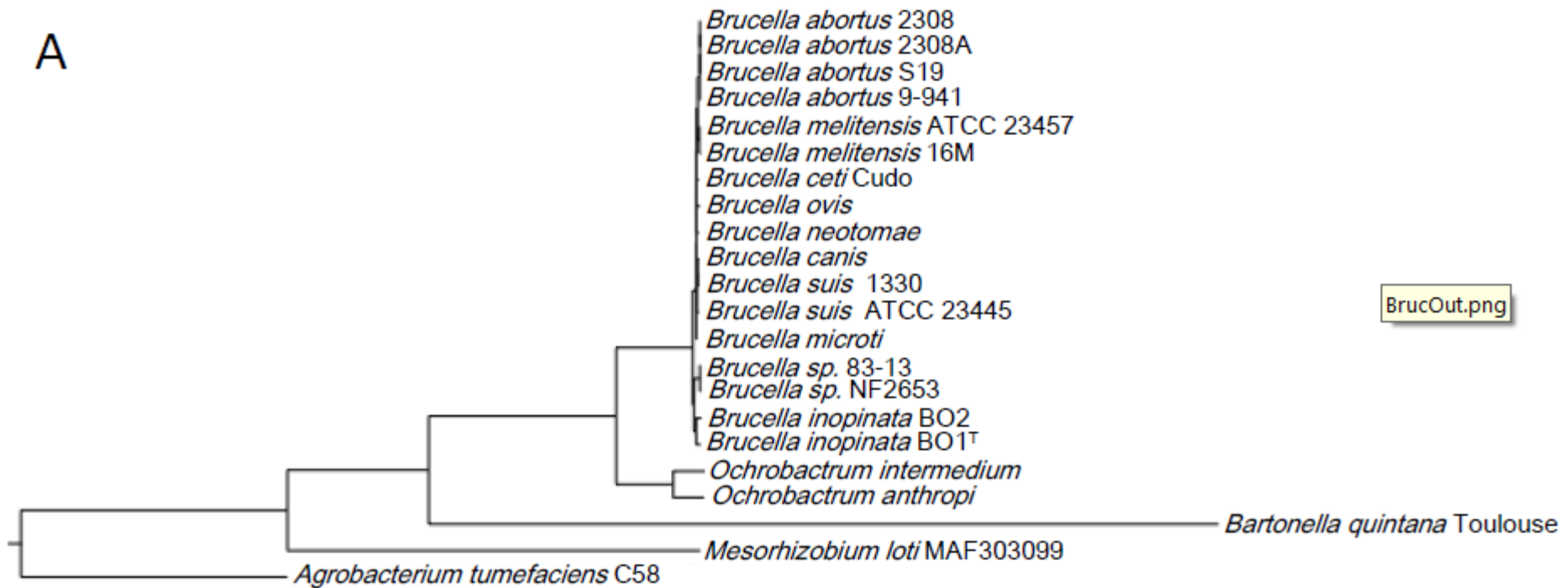
B



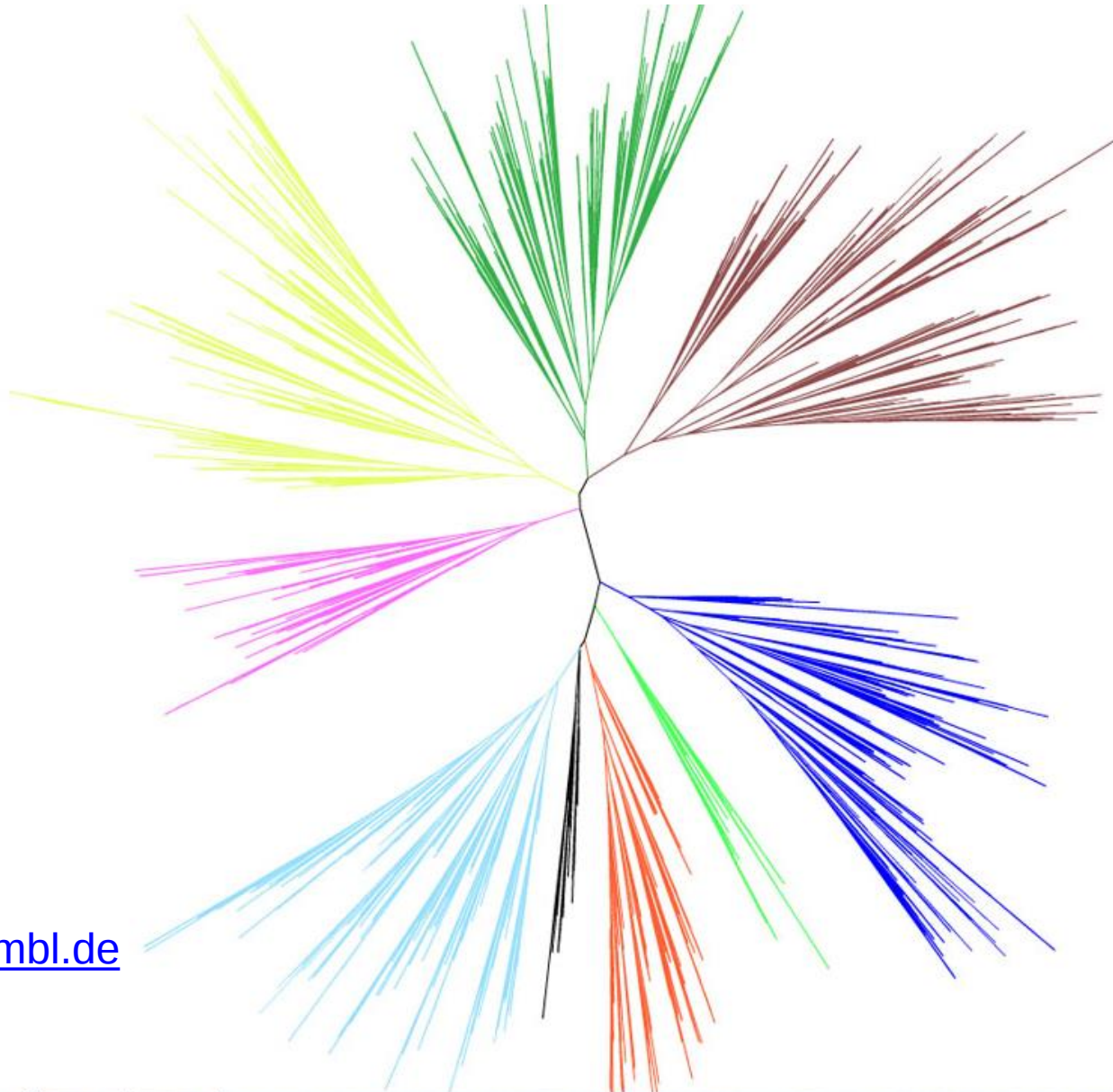
Credit: Wattam et al. 2011

Árvore com raiz: precisa de um grupo externo

A



Árvore sem raiz (sem grupo externo)



<http://itol.embl.de>

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

Uma matriz de distâncias genômicas

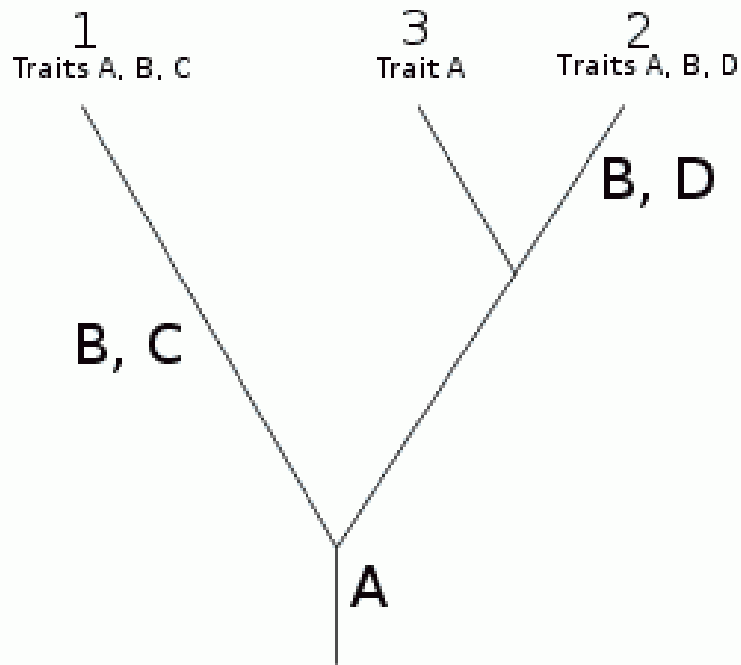
MUMi results for the rodent strain NF 2653 genome along with six *Brucella* sp. genomes computed using the concatenated contigs of each incomplete genome^a

Strain	MUMi value ^b					
	83-13	BO2	NF 2653	BO1	<i>B. suis</i>	<i>B. microti</i>
<i>B. neotomae</i> 5K33	0.145	0.168	0.146	0.168	0.017	0.022
<i>Brucella</i> sp. strain 83-13		0.172	0.009	0.175	0.147	0.150
<i>Brucella</i> sp. strain BO2			0.168	0.107	0.169	0.169
<i>Brucella</i> sp. strain NF 2653				0.172	0.147	0.146
<i>B. inopinata</i> BO1					0.169	0.167
<i>B. suis</i> biovar 3 686						0.020

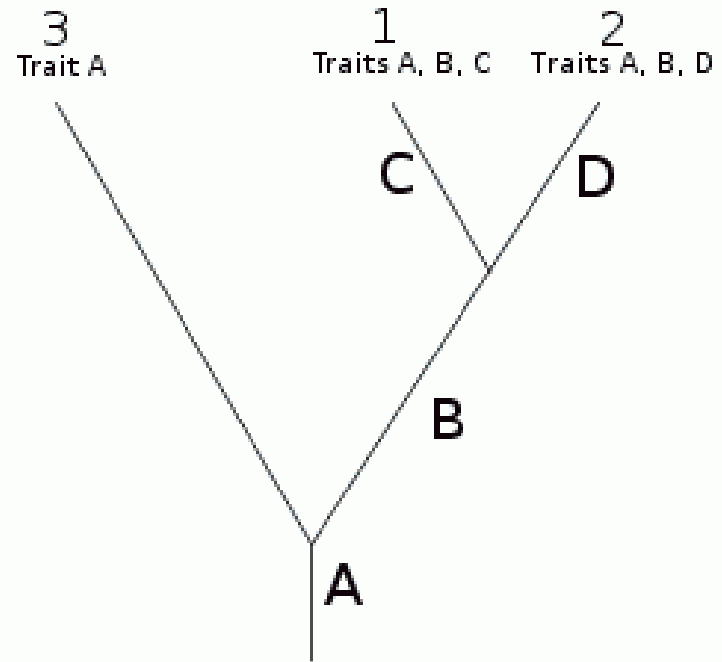
↵^a Only the sequence of *B. microti* was complete, and its two chromosome sequences were concatenated. Concatenation was done by inserting a string with 100 nucleotides between contigs.

↵^b The minimum MUMi value is 0, and the maximum MUMi value is 1.

parcimônia



Less parsimonious cladogram



More parsimonious cladogram

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

Parcimônia é um princípio muito usado

- 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 (*outcome*)
- A verossimilhança de um conjunto de valores de parâmetros dadas as observações é igual à probabilidade dessas observações dados esses valores
- $L(\theta(x)) = P(x \mid \theta)$

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

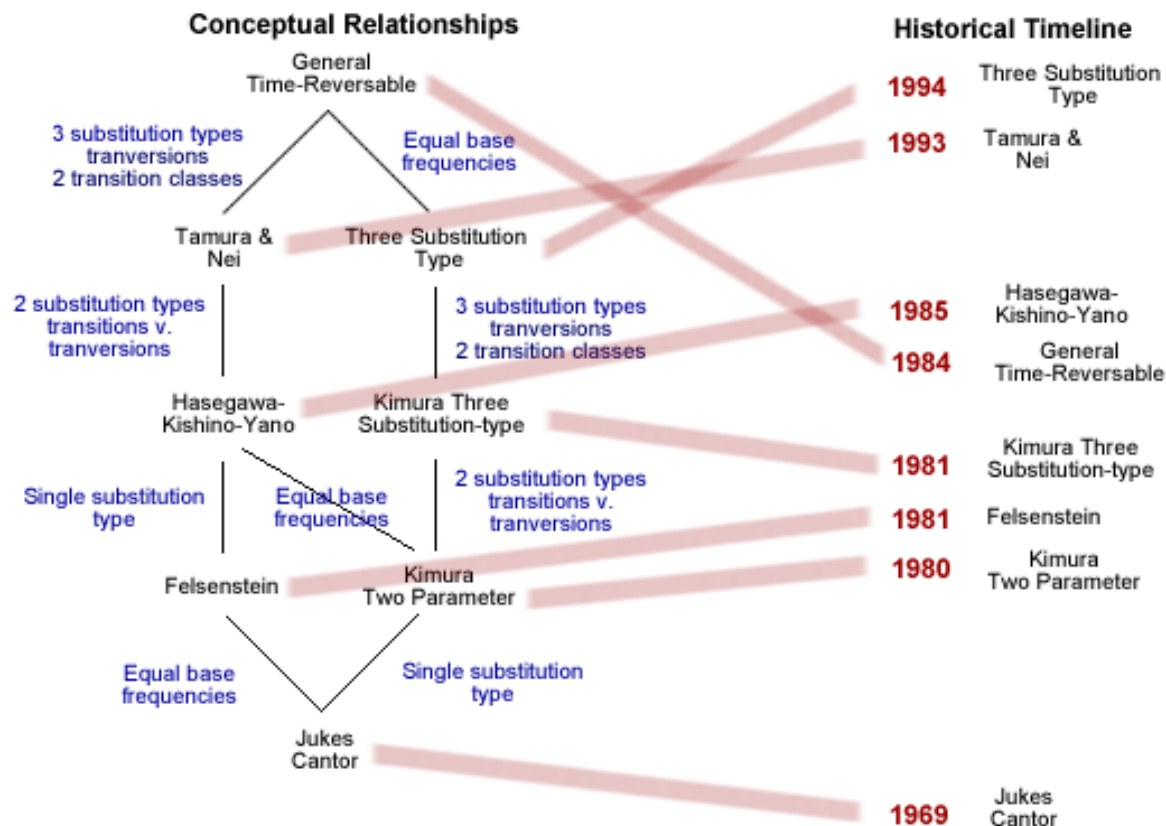
Considerações de tempo de execução

- Até o ano 2000 (aprox.) distância e parcimônia eram os métodos mais usados
 - os outros eram muito lentos
- Agora **máxima verossimilhança** se tornou “padrão”

Modelos de evolução

- Exceto por 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, Blosom62, mtREV, rtREV, cpREV, DCMut, VT, mtMAM, custom

- Todos eles supõem que cada *site* evolui de forma independente

Para escolher modelos

- ModelTest: para nucleotídeos
- ProtTest: para aminoácidos

Substituições sinônimas e não-sinônimas

- Código genético é degenerado
- Glicina: GGA, GGC, GGG, GGU
- Mutação na terceira base **não altera** o aminoácido
 - Sinônima (silenciosa)
- Mutação na primeira base altera o aminoácido
 - Não-sinônima

Razão Ka/Ks

- Ka/Ks ou dN/dS
- Razão entre o número de subs. não-sinônimas (Ka) e o número de subs. sinônimas (Ks)
- Usado para inferir a direção e magnitude de seleção natural agindo em genes codificadores de proteínas
- $Ka/Ks > 1$: seleção positiva ou Darwiniana
- $Ka/Ks < 1$: seleção purificadora ou estabilizadora
- $Ka/Ks = 1$: não há seleção (neutra)

Para calcular Ka/Ks

- Hurst, L. (2002). "The Ka/Ks ratio: diagnosing the form of sequence evolution". *Trends in Genetics* **18**: 486–489
- <http://services.cbu.uib.no/tools/kaks>

Programas para inferência filogenética: Pacotes

- PHYLIP

- Joe Felsenstein

- <http://evolution.genetics.washington.edu/phylip.html>

- PAUP

- David Swofford

- <http://paup.csit.fsu.edu/>

- MEGA

- Sudhir Kumar, Koichiro Tamura & Masatoshi Nei

- <http://www.megasoftware.net/>

Programas que implementam métodos não-probabilísticos

- **Distância**
 - Pacotes
 - Neighbor-joining
 - UPGMA
- **Parcimônia**
 - pacotes

O problema da caixa preta

- Idealmente: todo usuário de um método e respectivo programa deveria entender os princípios do método
- No caso de métodos de filogenia
 - Estatística não trivial

Máxima verossimilhança

- RaXML
 - A. Stamatakis
 - <http://www.exelixis-lab.org/>
- phyML
 - O. Gascuel et al. Systematic Biology, 59(3):307-21, 2010
 - <http://www.atgc-montpellier.fr/phyml/>
- fastTree
 - Morgan N. Price in Adam Arkin's group
 - <http://www.microbesonline.org/fasttree/>
 - “FastTree can handle alignments with up to a million of sequences in a reasonable amount of time and memory”

Um resultado de desempenho pontual

- Criação de uma árvore ML para 500 sequências de proteínas com aprox. 300 aa
- Computador desktop “normal” (4 GB de RAM)
- RAxML or PHYml levaram aprox. 10 horas
- Fasttree levou menos do que 1 hora

Inferência bayesiana

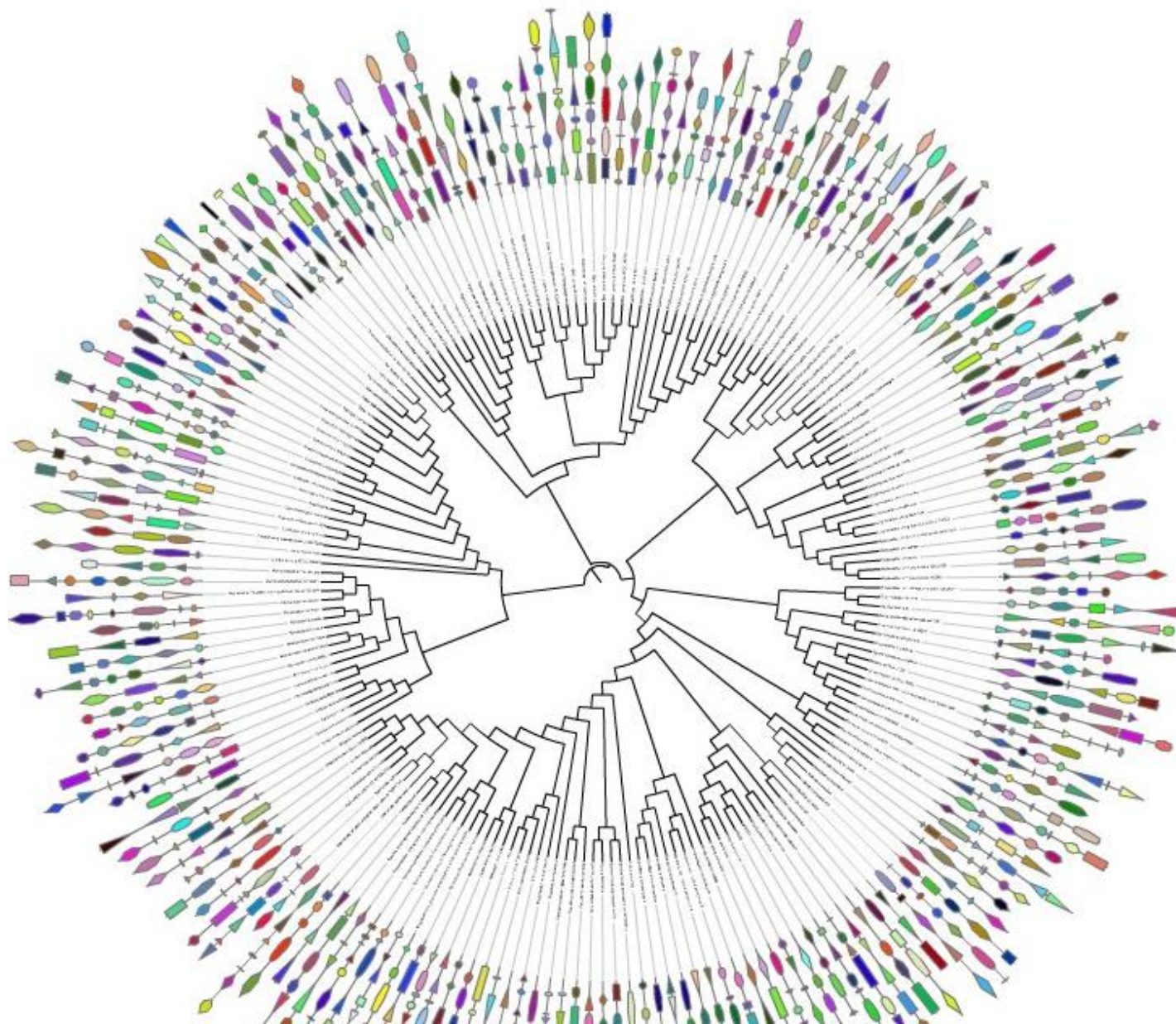
- MrBayes
- Ronquist and Huelsenbeck. Bioinformatics. 2003 19(12):1572-4.
- <http://mrbayes.sourceforge.net/>
- Mais lento comparado a RAxML e phyML
- Resultados não são conclusivamente melhores do que ML

Visualização de árvores: formatos

- **Newick, NEXUS**
- (((erHomoC:0.28006,erCaelC:0.22089):0.40998,(erHomoA:0.32304,(erpCaelC:0.58815,((erHomoB:0.5807,erCaelB:0.23569):0.03586,erCaelA:0.38272):0.06516):0.03492):0.14265):0.63594,(TRXHomo:0.65866,TRXSacch:0.38791):0.32147,TRXEcoli:0.57336);
- <http://molecularevolution.org/resources/treeformats>

Visualização de árvores

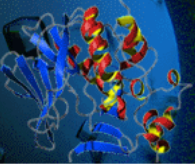
- Interactive Tree of Life <http://itol.embl.de>
- http://en.wikipedia.org/wiki/List_of_phylogenetic_tree_visualization_software



All-in-one: phylogeny.fr

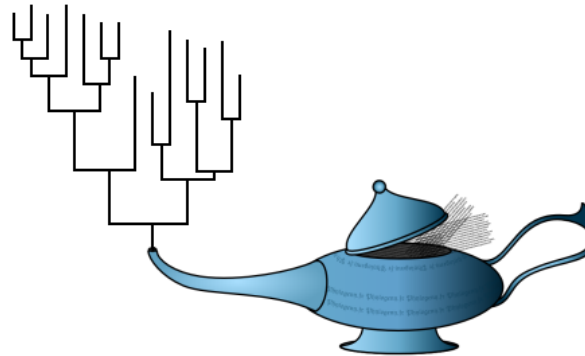


Information
Genomique et
Structurale



Home Phylogeny Analysis Blast Explorer Online Programs Your Workspace Documentation Downloads Contacts

Phylogeny.fr Robust Phylogenetic Analysis For The Non-Specialist



Phylogeny.fr is a free, simple to use web service dedicated to reconstructing and analysing phylogenetic relationships between molecular sequences.

Phylogeny.fr runs and connects various bioinformatics programs to reconstruct a robust phylogenetic tree from a set of sequences.

If you use this site, please cite:

Dereeper A., Guignon V., Blanc G., Audic S., Buffet S., Chevenet F., Dufayard J.-F., Guindon S., Lefort V., Lescot M., Claverie J.-M., Gascuel O. *Phylogeny.fr: robust phylogenetic analysis for the non-specialist* Nucleic Acids Research. 2008 Jul 1; 36 (Web Server Issue):W465-9. Epub 2008 Apr 19. ([PubMed](#))

Phylogeny.fr (2)

● Phylogeny analysis

"One Click"

Paste your set of sequences and let the software make decisions on your behalf (Each step is optimized for your data).

"Advanced"

Manually set parameters for the various steps.

"A la Carte"

Create your own phylogeny workflow using more programs available.

● Explore your sequence neighbors

Paste your single sequence, run Blast and explore its homologous sequences.

● Online phylogeny programs

Direct access to the individual tools available on this server.

Multiple Alignment:

MUSCLE
T-Coffee / 3D-Coffee
ClustalW
ProbCons

Phylogeny:

PhyML
TNT
BioNJ
MrBayes

Tree viewers:

TreeDyn
Drawgram
Drawtree
ATV

Utilities:

Gblocks
Jalview
Readseq
Format converter

This project is funded by the Réseau National des Génomiques (RNG).

This project is managed in a GForge project, which aims to help collaboration and development management (using Subversion).

 RSS Feed  Mailing-list  Mentions légales



Building your tree locally: SeaView

SeaView

Version 4.3.5

NEW: seaview computes and draws parsimony, distance and PhyML phylogenetic trees.

NEW: seaview prints trees and outputs them in scalable vector graphics (SVG) format.

NEW: seaview drives the Gblocks program to select blocks of conserved sites.

SeaView is a multiplatform, graphical user interface for multiple sequence alignment and molecular phylogeny.

- SeaView reads and writes various file formats ([NEXUS](#), MSF, CLUSTAL, FASTA, PHYLIP, [MASE](#), Newick) of DNA and protein sequences and of phylogenetic trees.
- SeaView drives programs [muscle](#) or [clustalw](#) for multiple sequence alignment, and also allows to use any external alignment algorithm able to read and write FASTA-formatted files.
- Seaview drives the [Gblocks](#) program to select blocks of evolutionarily conserved sites.
- SeaView computes phylogenetic trees by
 - parsimony, using PHYLIP's [dnaphars/protpars](#) algorithm,
 - distance, with [NJ](#) or [BioNJ](#) algorithms on a variety of evolutionary distances,
 - maximum likelihood, driving program [PhyML](#) 3.0.
- SeaView prints and draws phylogenetic trees on screen, SVG, PDF or PostScript files.
- SeaView allows to download sequences from EMBL/GenBank/UniProt using the Internet.

Screen shots of the main [alignment](#) and [tree](#) windows. On-line [help](#) document. Old [seaview version 3.2](#)

Download SeaView



Interpretação

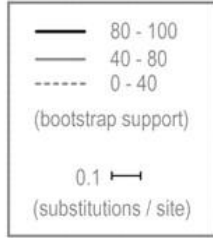
- Árvores são apenas hipóteses
- GIGO: garbage in, garbage out
- Os métodos em geral (menos distância) fornecem uma árvore com **nota (score)**
 - Parcimônia: **número mínimo de mutações**
 - ML: **valor da verossimilhança logarítmica**
 - Bayesiano: **probabilidade posterior**
- A árvore de melhor nota pode não ser a árvore “verdadeira”
- Para avaliar a qualidade da árvore
 - Confiabilidade de sua topologia

Confiabilidade da topologia

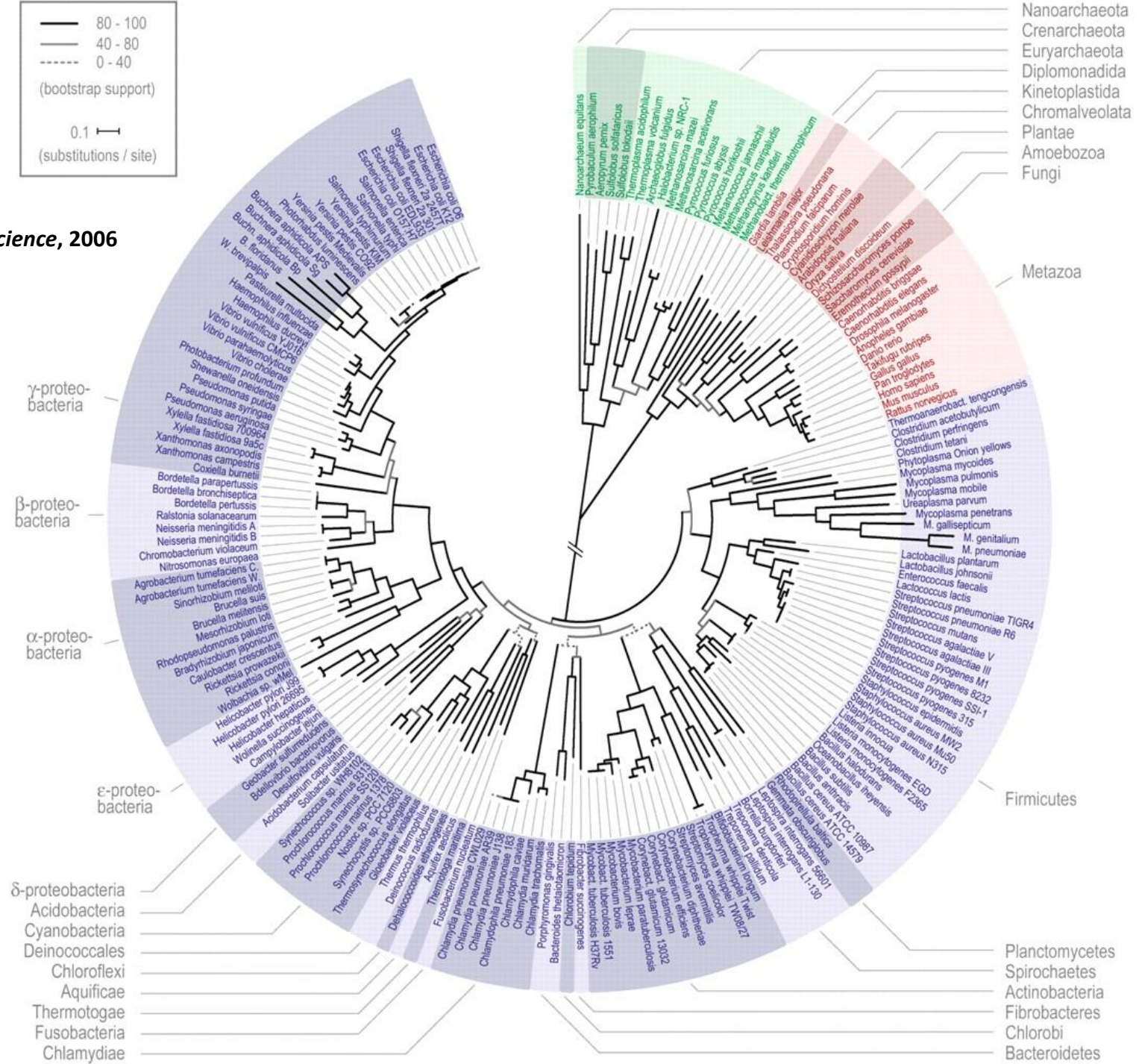
- Valores de **bootstrap**
- Colunas do AM são amostradas aleatoriamente em várias corridas (**replicatas**; geralmente entre 100 e 1000)
- Árvores resultantes são comparadas entre si
- Concordâncias nos clados são calculadas, resultando em número de vezes (ou %) que clados se repetem nas replicatas
- Valores bons são considerados aqueles maiores do que 0.7 (70%)
- Custosos para calcular
- PhyML fornece valores aproximados (ALRT) muito mais rapidamente

Como lidar com todas essas incertezas?

- Aprenda mais sobre evolução e inferência filogenética
- Se a filogenia é crucial para seus resultados
 - Use mais de um método!



Ciccarelli et al, *Science*, 2006

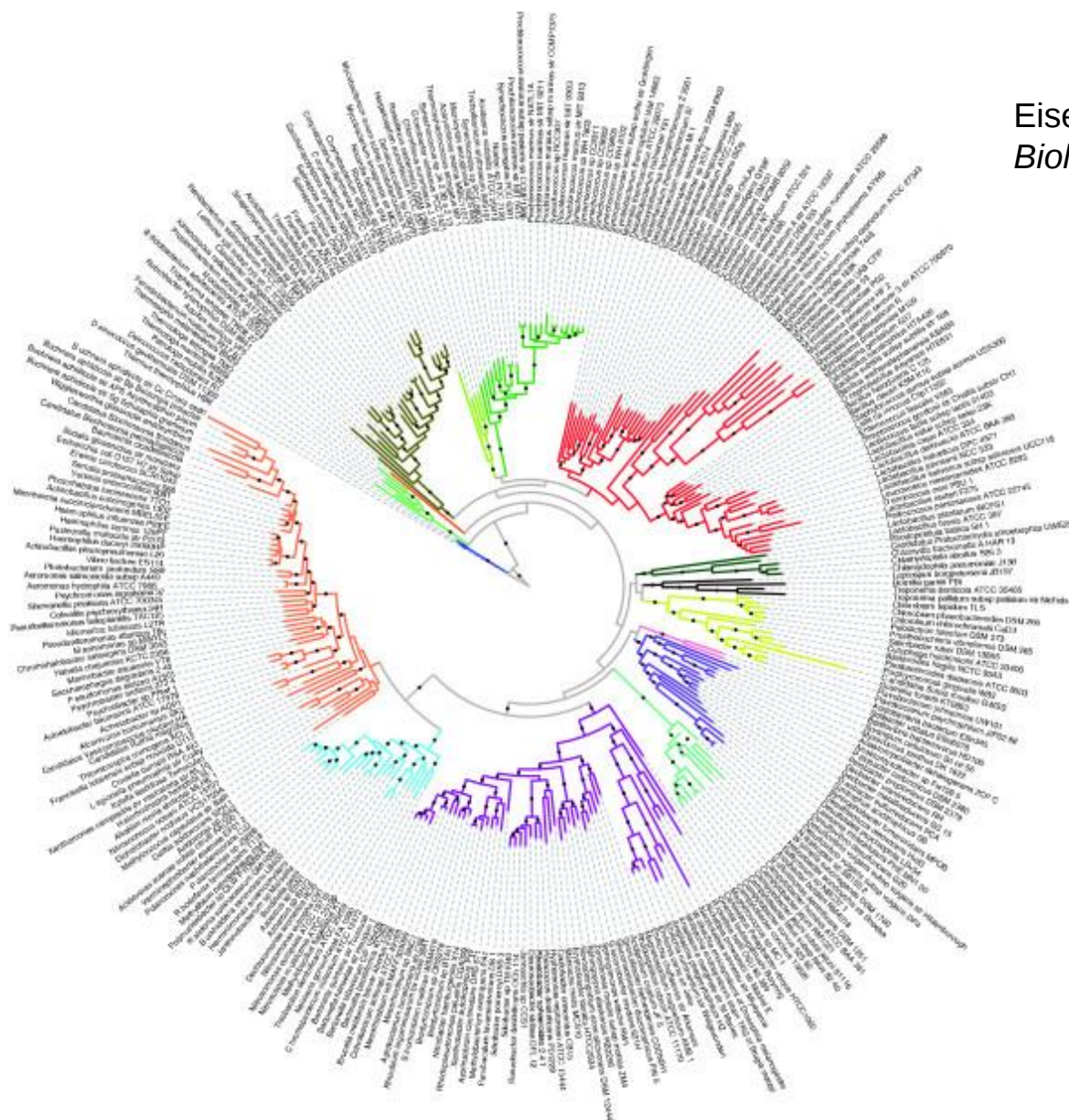


Supermatrizes

- Método bom para obter árvores robustas de espécies quando genomas completos ou quase completos estão disponíveis
- Determinar famílias de proteínas para os genomas de interesse
- Determinar quais famílias tem exatamente um representante de cada genoma
- AM para cada família
- Concatenar todos os Ams (“a supermatriz”)
- Construir árvore com base no AM concatenado

Problemas

- Diferentes taxas de evolução
- Long branch attraction
- O problema de HGT



- | | | | |
|-------------------------|-----------------------------|------------------|-----------------------|
| ■ Gammaproteobacteria | ■ Acidobacteria | ■ Cyanobacteria | ■ Deinococcus/Thermus |
| ■ Betaproteobacteria | ■ Bacteroidetes/Chlorobi | ■ Chloroflexi | |
| ■ Alphaproteobacteria | ■ Spirochaetes | ■ Actinobacteria | |
| ■ Epsilonproteobacteria | ■ Chlamydiae/Planctomycetes | ■ Aquificae | |
| ■ Deltaproteobacteria | ■ Firmicutes | ■ Thermotogae | |

Transferência Horizontal de Genes

- Material genético é passado de uma célula (doadora) para outra (receptora)
- O doador pode ser completamente diferente do receptor
- Exemplo: humanos e bactérias

Exemplo de HGT

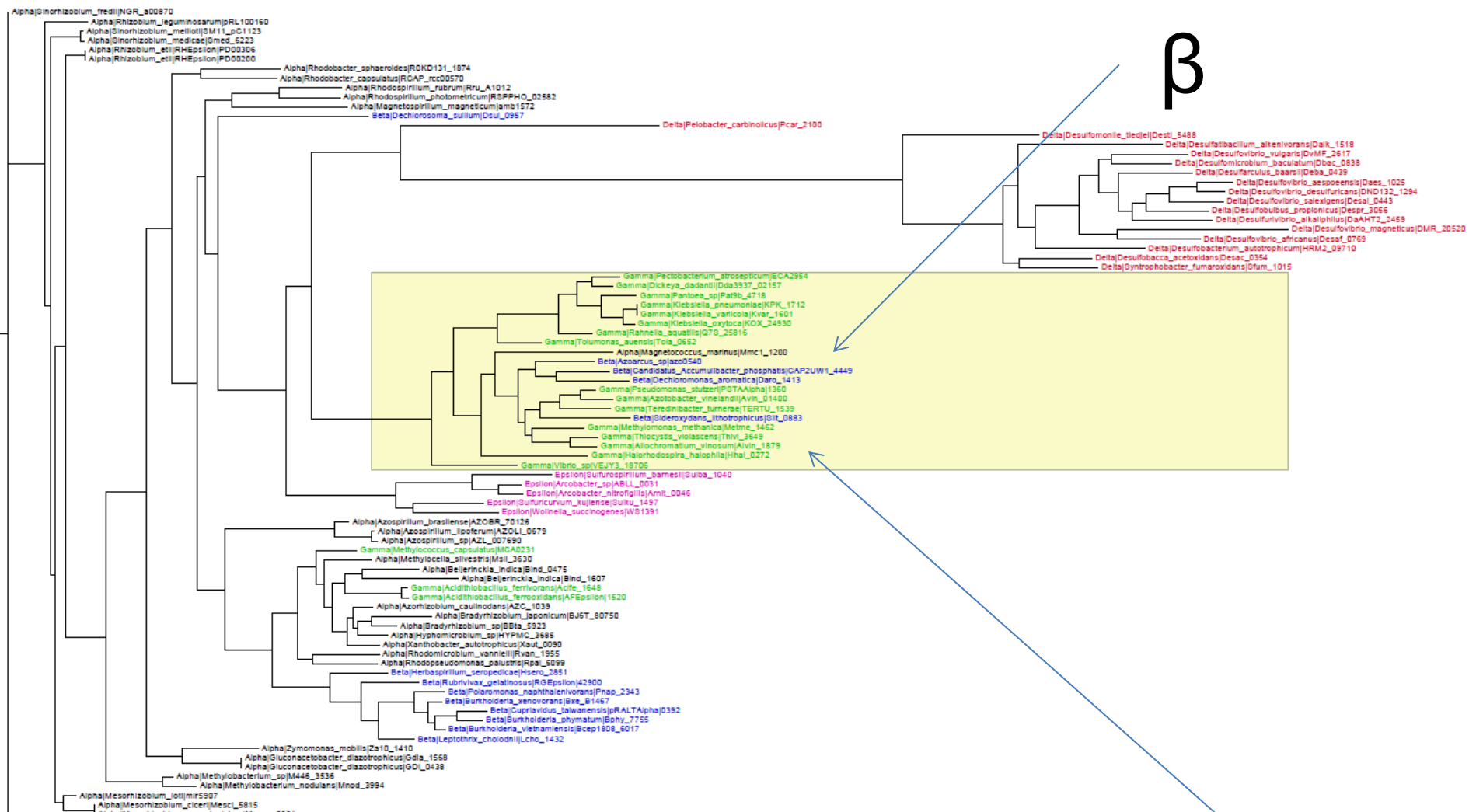
- Humanos e bactérias

Transferência Horizontal de Genes

- Atrapalha a construção de árvores de espécies
- Como detectar?
- THG antiga
- THG recente

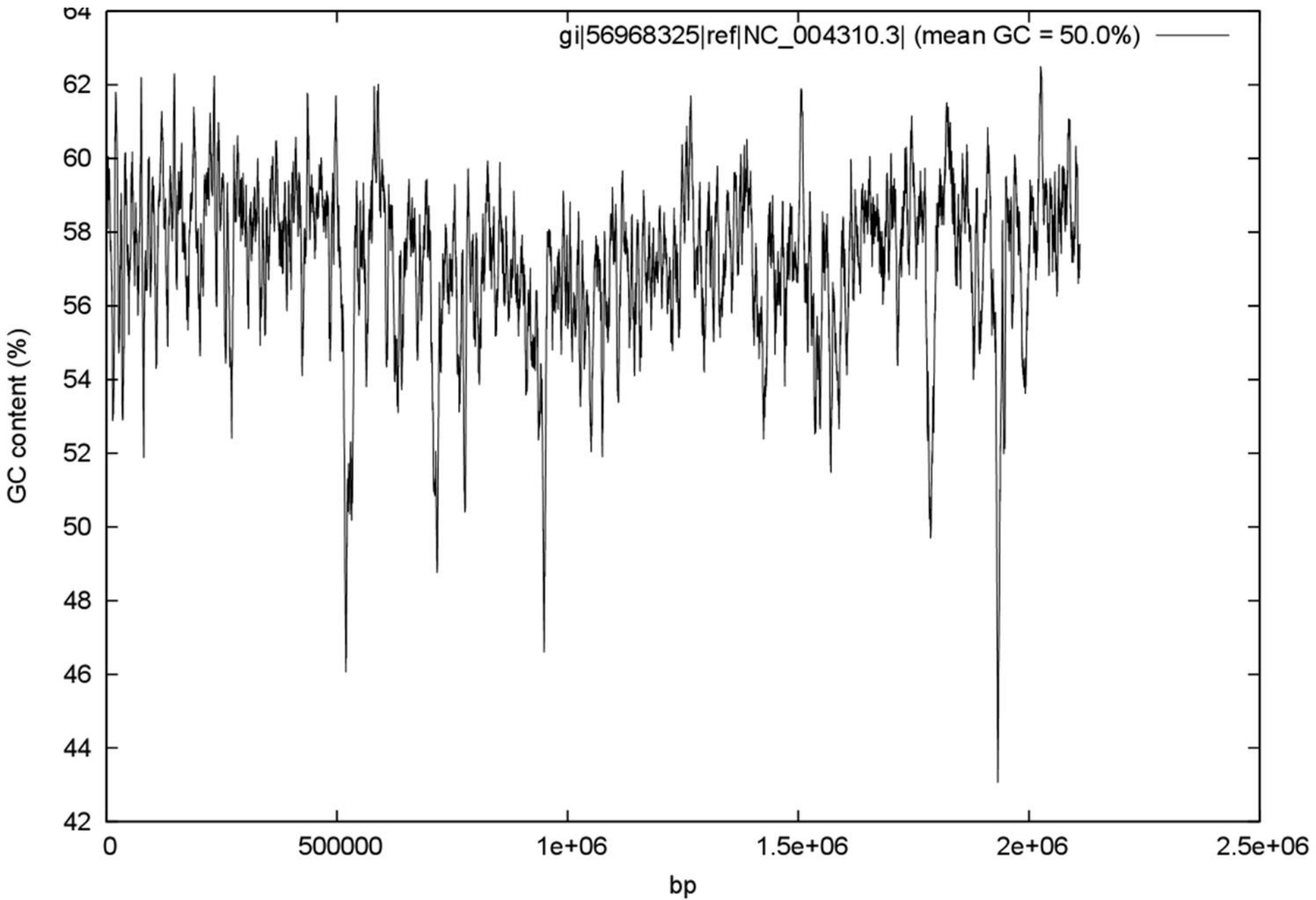
THG antiga

- Incongruência de árvores
 - Quando a árvore de um gene difere da árvore (robusta) de espécies

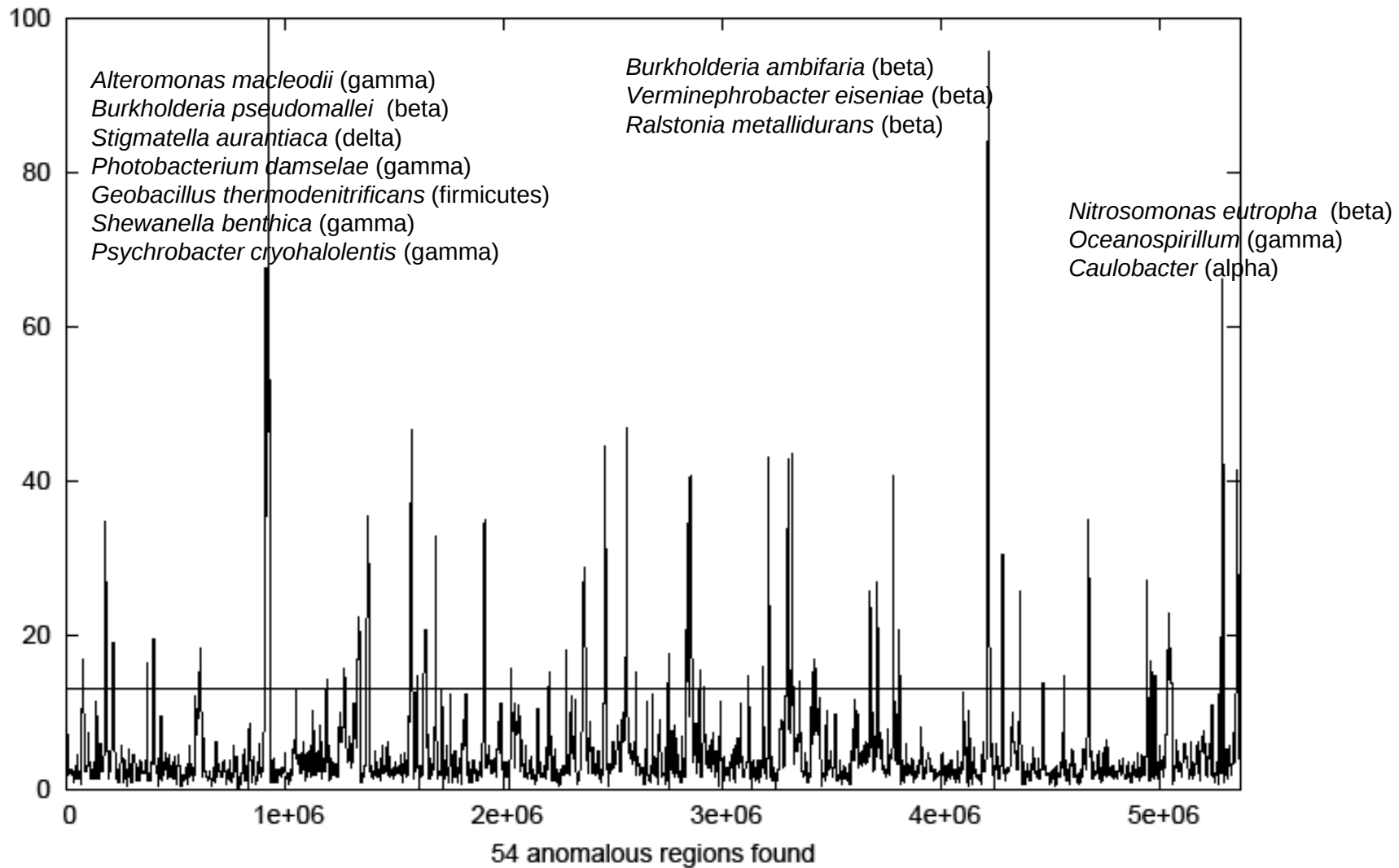


THG recente

- Incongruência de árvores
- Outros métodos
 - Desvios na composição (%GC, dinucleotídeos, uso de codons) da sequência
 - Ilhas genômicas

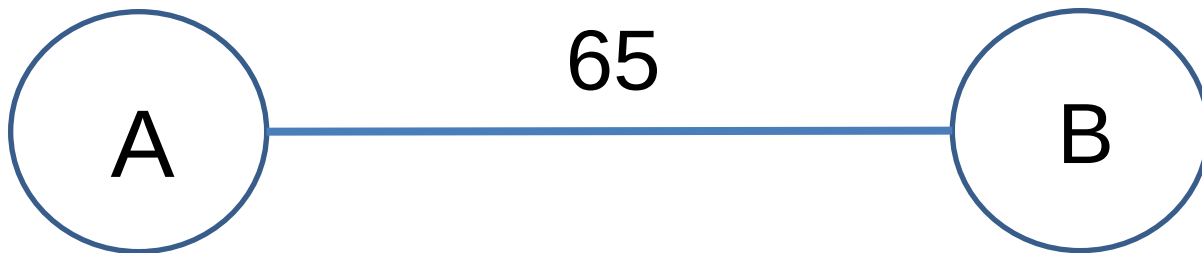


Azotobacter vinelandii anomalous regions (Alien_hunter threshold: 13.020)



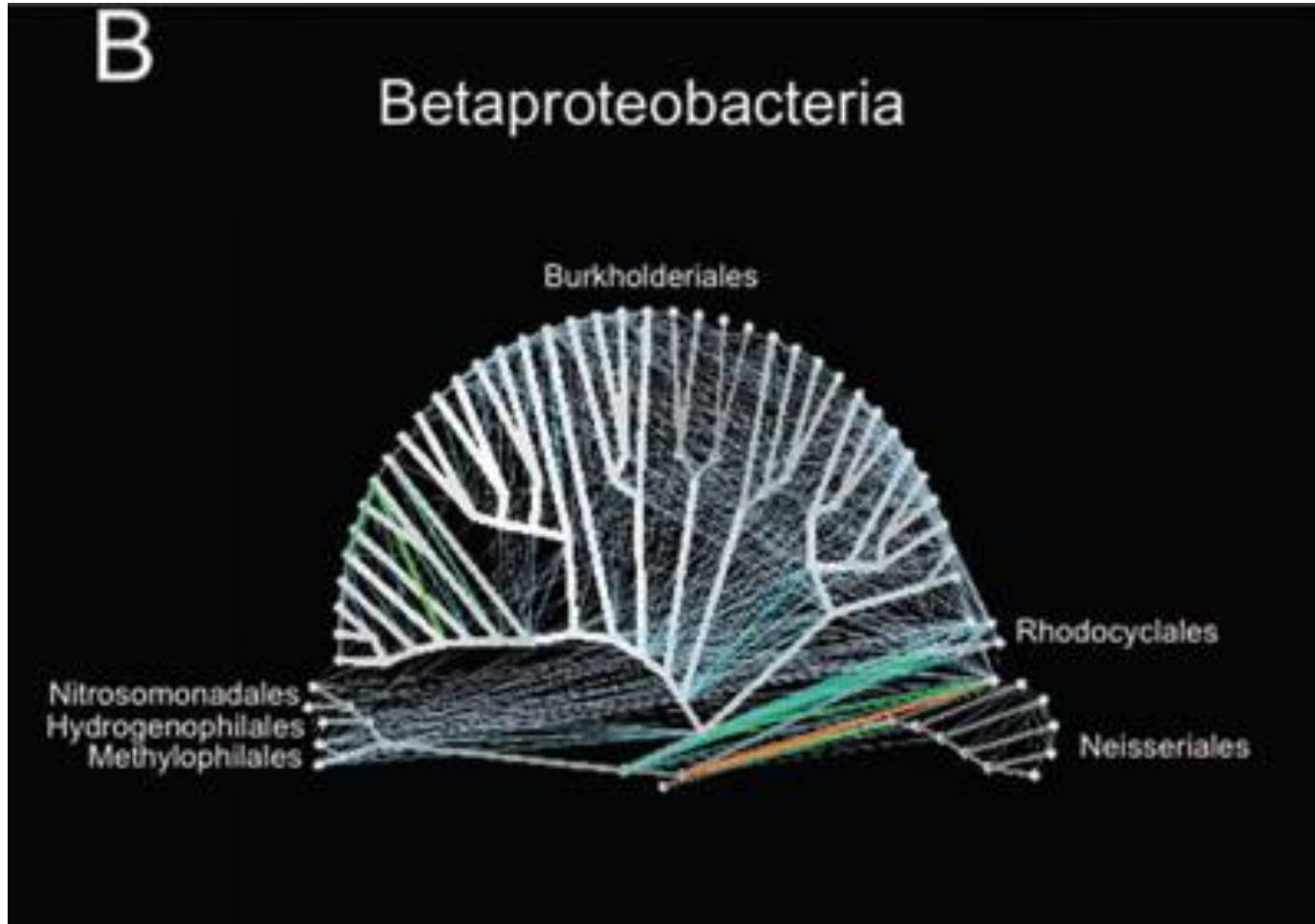
Redes filogenômicas

- Redes que mostram compartilhamento de genes



A superposição de uma **árvore de espécies** numa tal rede mostra possíveis eventos de **transferência horizontal**

Uma rede filogenômica

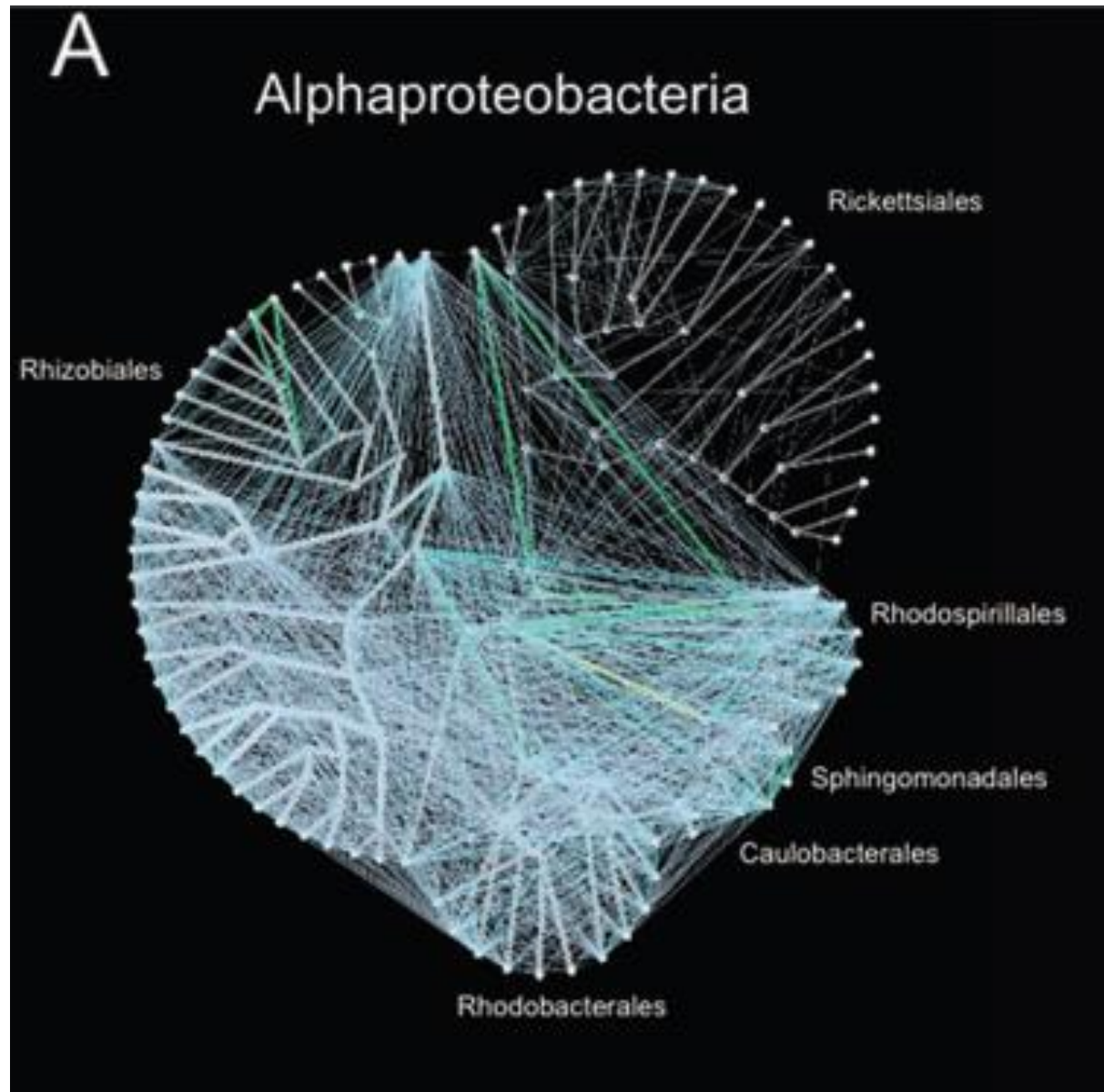


Kloesges et al, *Molecular Biology and Evolution*, 2011

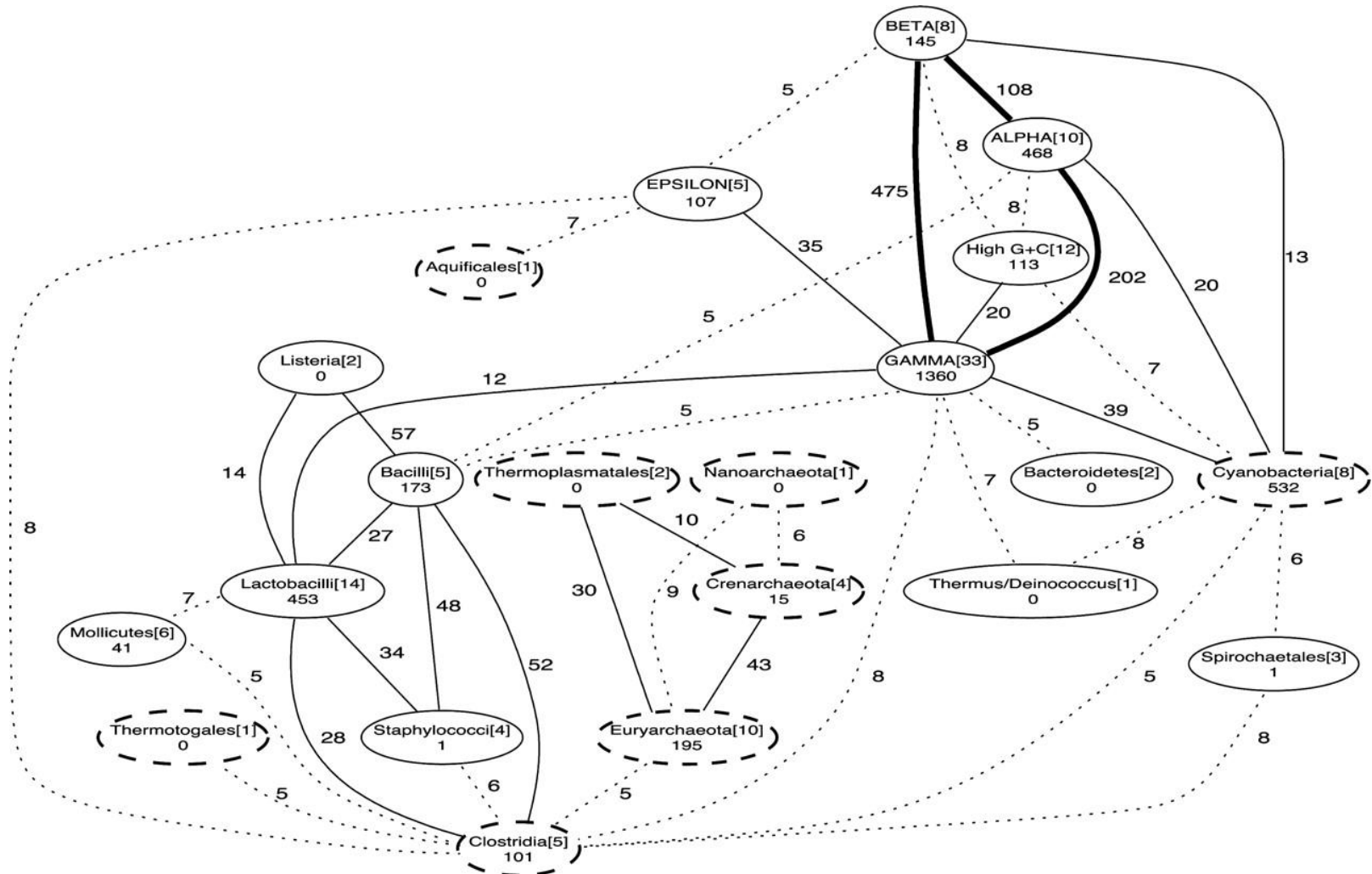
11/7/2014

J. C. Setubal

64



Highways of obligate gene transfer within and among phyla and divisions of prokaryotes, based on analysis of the 22,348 protein trees for which a minimal edit path could be resolved.



Beiko R G et al. PNAS 2005;102:14332-14337

Literatura

- Yang e Rannala. Molecular phylogenetics: principles and practice. *Nature Reviews Genetics*, 13:303-314, 2012
- ***Bioinformatics***. Baxevanis and Ouellette (Eds.) Wiley-Interscience, 2005 (3rd edition), **ch. 14**
- D. Mount. ***Bioinformatics***. CSHL Press, 2004 (2nd edition), **ch. 7**
- ***The phylogenetic handbook***. Lemey, Salemi and Vandamme (Eds.) Cambridge University Press, 2009 (2nd edition)