



Universidade de São Paulo
Instituto de Química



Análise filogenética para dados moleculares

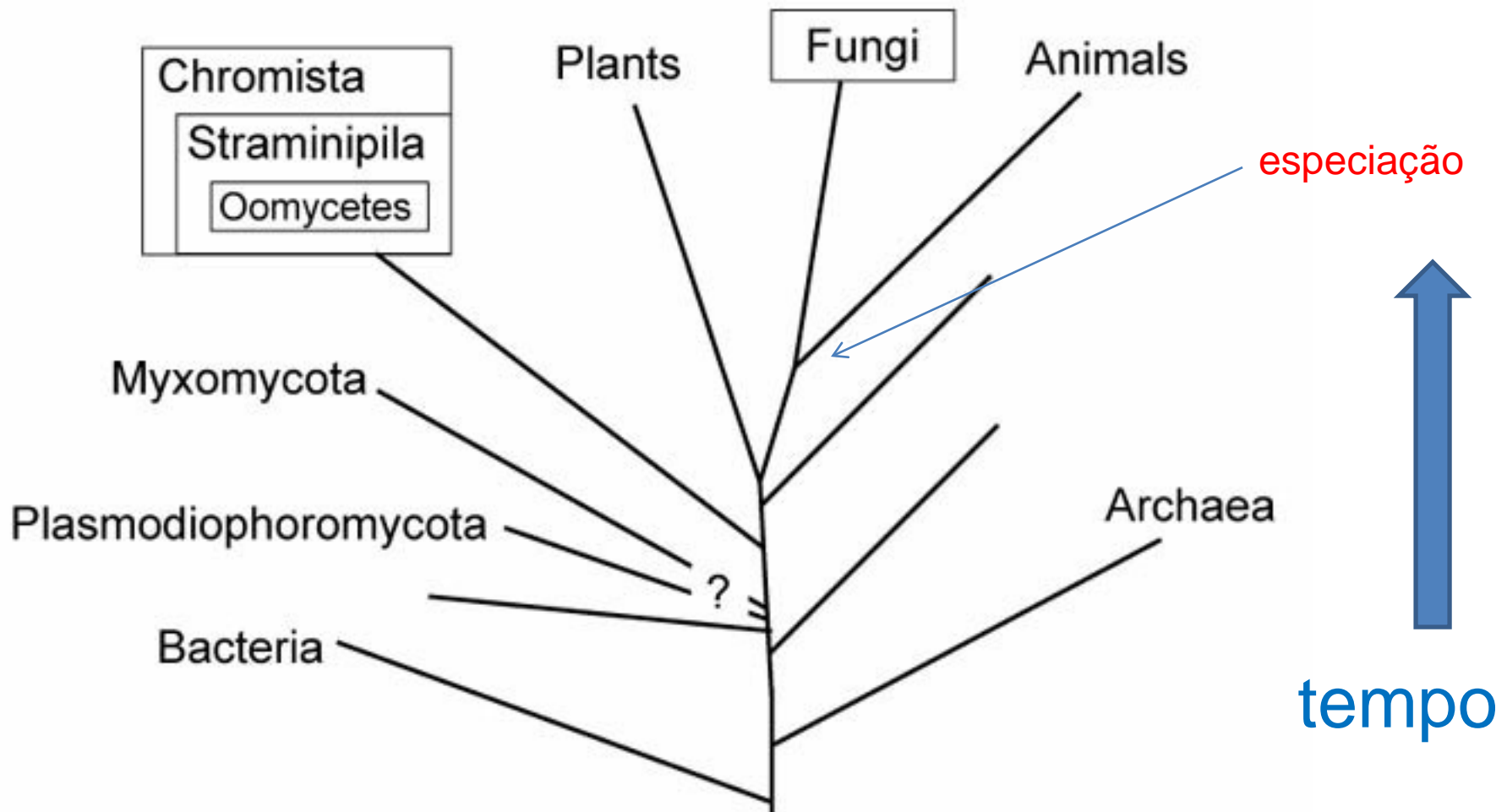
João C. Setubal

2018

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



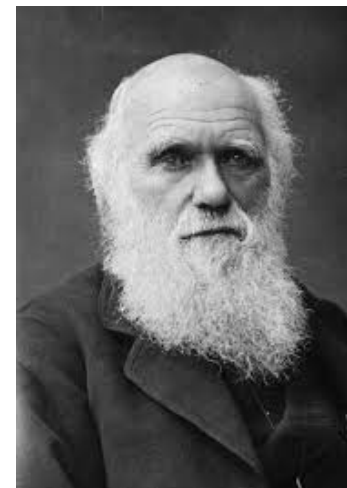
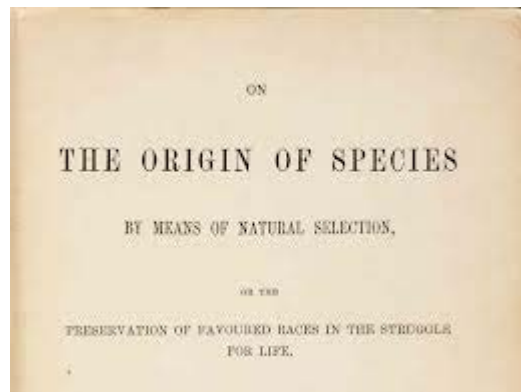
Uma árvore é uma **hipótese** sobre o que ocorreu na evolução

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

— 80 - 100
 — 40 - 80
 - - - 0 - 40
 (bootstrap support)

0.1 —
 (substitutions / site)

Nanoarchaeota
 Crenarchaeota
 Euryarchaeota
 Diplomonadida
 Kinetoplastida
 Chromalveolata
 Plantae
 Amoebozoa
 Fungi

Ciccarelli et al, *Science*, 2006

γ-proteo-
bacteria

β-proteo-
bacteria

α-proteo-
bacteria

ε-proteo-
bacteria

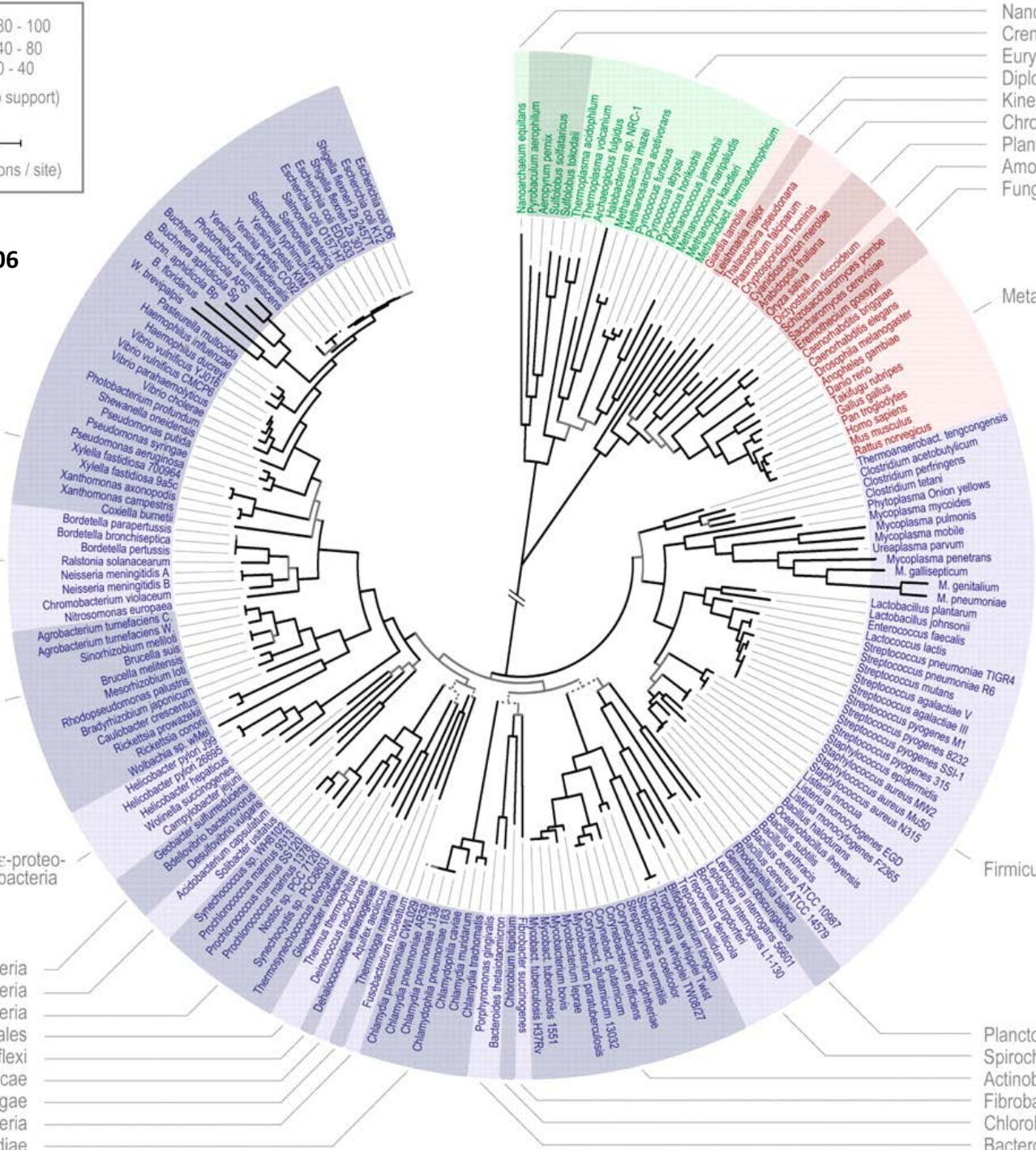
δ-proteo-
bacteria

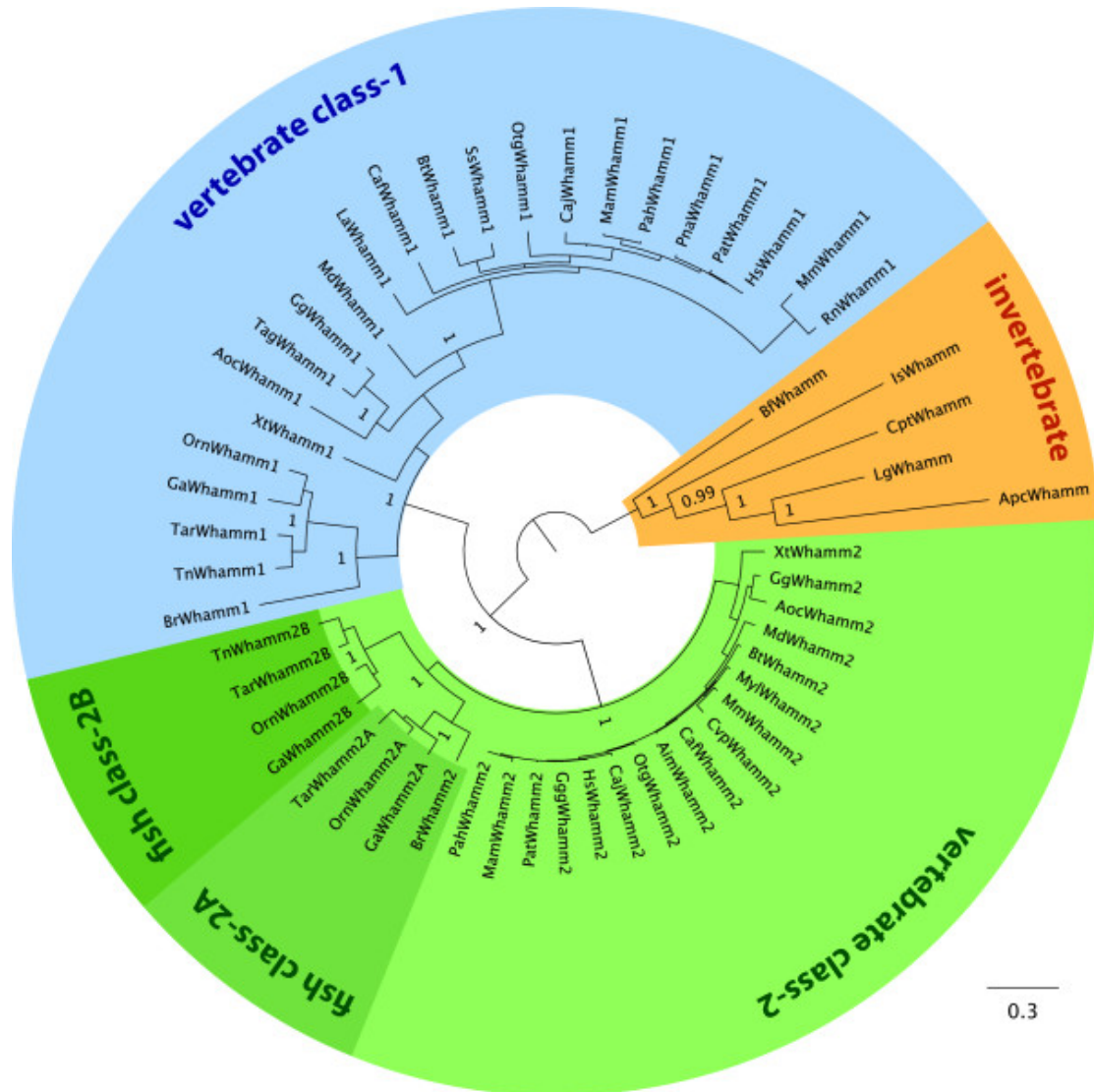
Acidobacteria
 Cyanobacteria
 Deinococcales
 Chloroflexi
 Aquificae
 Thermotogae
 Fusobacteria
 Chlamydiae

Metazoa

Firmicutes

Planctomycetes
 Spirochaetes
 Actinobacteria
 Fibrobacteres
 Chlorobi
 Bacteroidetes



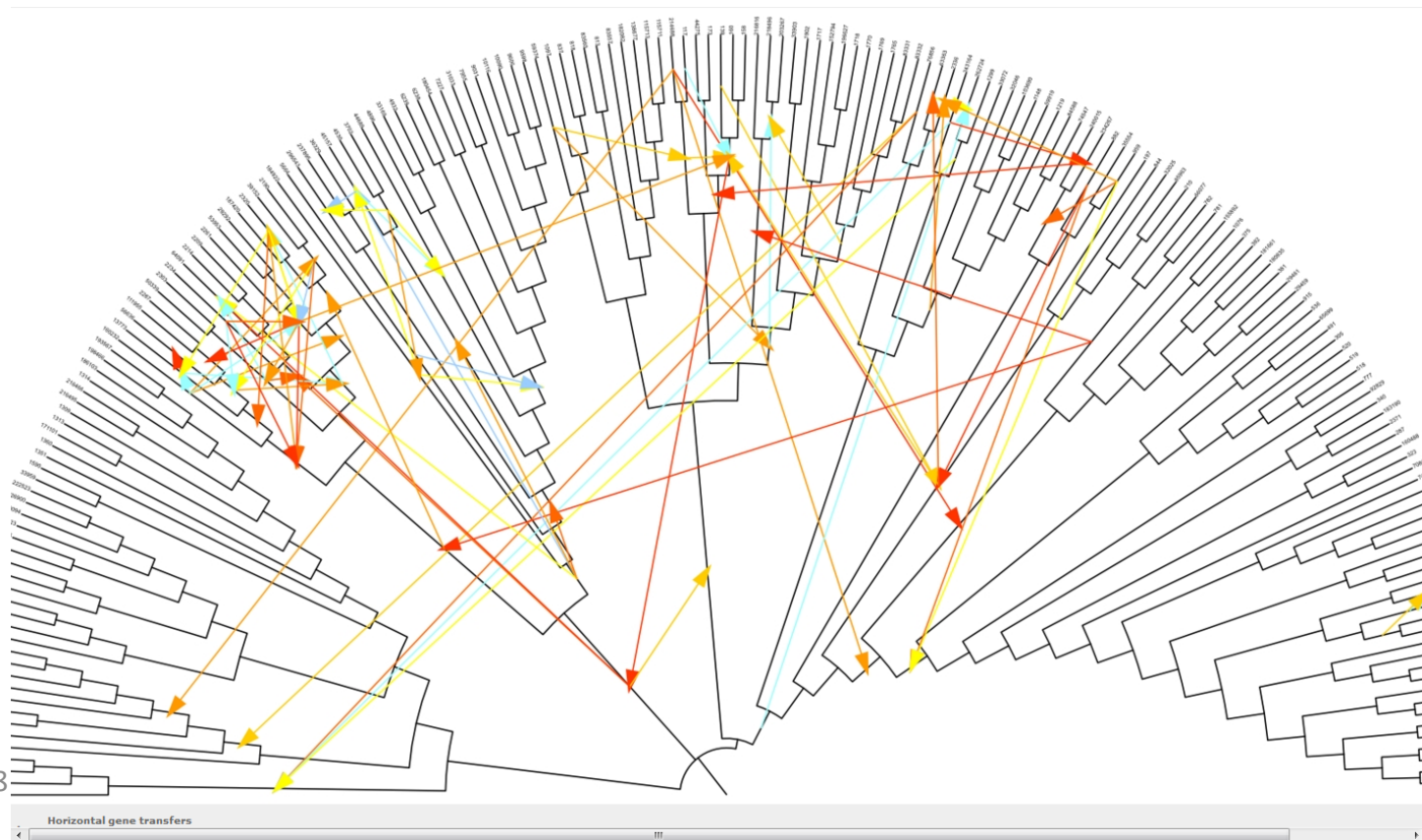


Phylogenetic tree of the WHAMM proteins

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

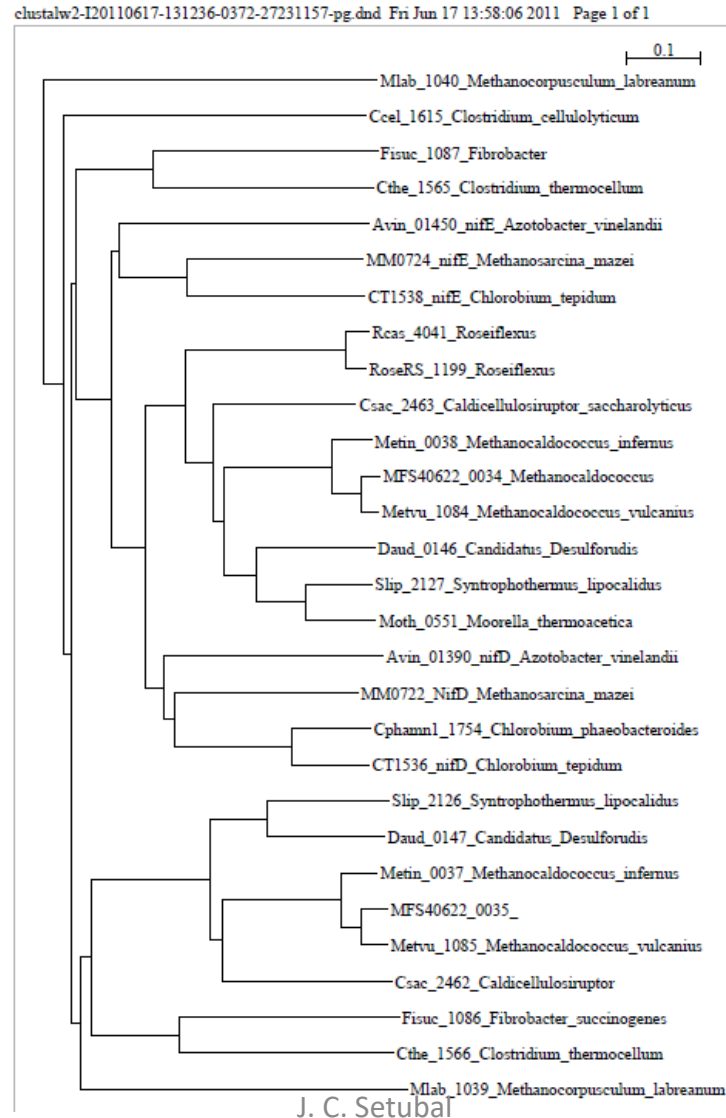
Pipeline

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

Alinhamento múltiplo

Cthe_1566_Clostridium_thermoce	LRDIIENTYK	VLDT-DLOVV	LTGCTAGIVG	DDVDSLSEF	AO-----
Fisuc_1086_Fibrobacter_succino	LDOLIKSTLK	VFDG-DLYVV	LTGCVGGLIG	DDVPSLVNEY	RD-----
Metvu_1085_Methanocaldococcus_	LVEGLRNLVA	RYDP-ELISV	VTTCSSETIG	DDIEAFIRAA	RKKIAS ⁺ EFGE
MFS40622_0035	LVEGLRNLVA	RYDP-DLISV	VTTCSSETIG	DDIEAFIRAA	RKKIAAEFGE
Metin_0037_Methanocaldococcus_	LVEGIRNLVA	RYDP-DLISV	VTTCSSETIG	DDIEAFIRAA	RKKIAKEFGE
Csac_2462_Caldicellulosiruptor	LIEGIRNLVL	RYSPTVIGV	ITTCSETIG	DDIEAFI ⁺ KEA	YKKLSEELSS
Daud_0147_Candidatus_Desulforu	FTEGIRNLVV	RYP-DLITV	VTTCSSEIIG	DDMVSFIKVA	RKRLVSELGP
Slip_2126_Syntrophothermus_lip	VIEGIRNLVV	RYWP-GLIGV	VTTCSSEIMG	DDMVSFLKEA	RARLSREIGR
CT1536_nifD_Chlorobium_tepidum	LKVAIQEAYD	LFHP-KAIAI	FSTCPVGLIG	DDVHAVAREM	KEKLGD----
Cphan1_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	OSECPIGLIG	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_Methanocaldococ	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	EFPNKAVFV	YNTCSTALIG	DDGRDVAKQA	EAIIG----
Rcas_4041_Roseiflexus	LLOSIIEASA	EFPDAKAVFV	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-PAVFV	YNTCVPALIG	DDVDAVCKAA	AERFG----
Cthe_1565_Clostridium_thermoce	LANTIREVYE	RTHA-NAIFV	LTTCAAGIIG	DDVESVCNEA	EEELG----
Fisuc_1087_Fibrobacter	LROTIRDAKE	RFNP-KAIFI	GMACATAIIG	EDIDSIAEEM	EPEVG----
Ccel_1615_Clostridium_cellulol	LVDSLNEVNS	RYNP-KIIAV	LTNCCADIIG	DDVEGCI EGL	PDEIR----
Mlab_1039_Methanocorpusculum_1	LLNKILOECA	SHHP-KFVAI	LGTVPV ⁺ PALIG	CDISGIATEV	FDTTK----
Mlab_1040_Methanocorpusculum_1	LCNAIDELLP	QIQRPKVFLV	YICCVLYLAG	FDEQSTIDEL	KKRNPD----

Filogenia resultante

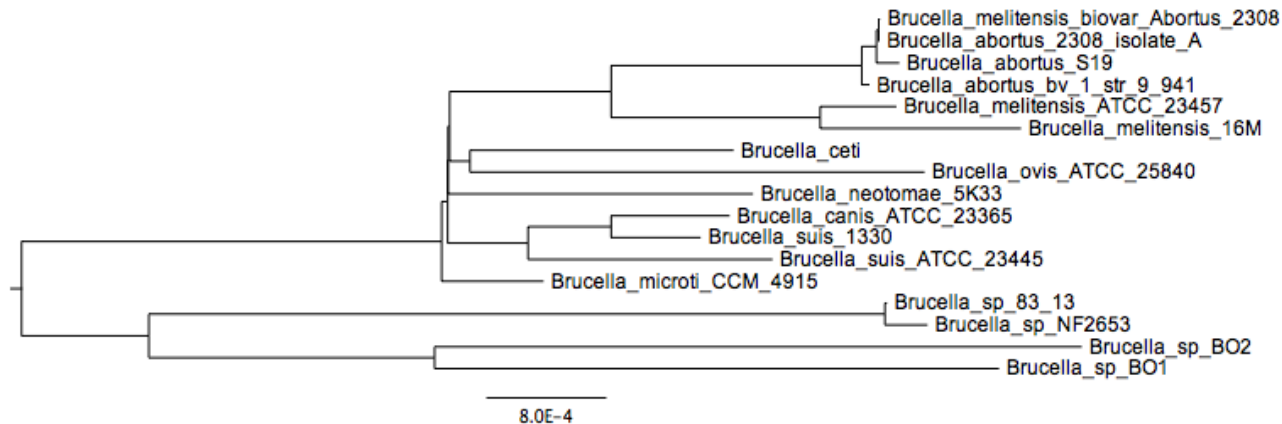


Credit: R. Dixon

Árvores e cladogramas

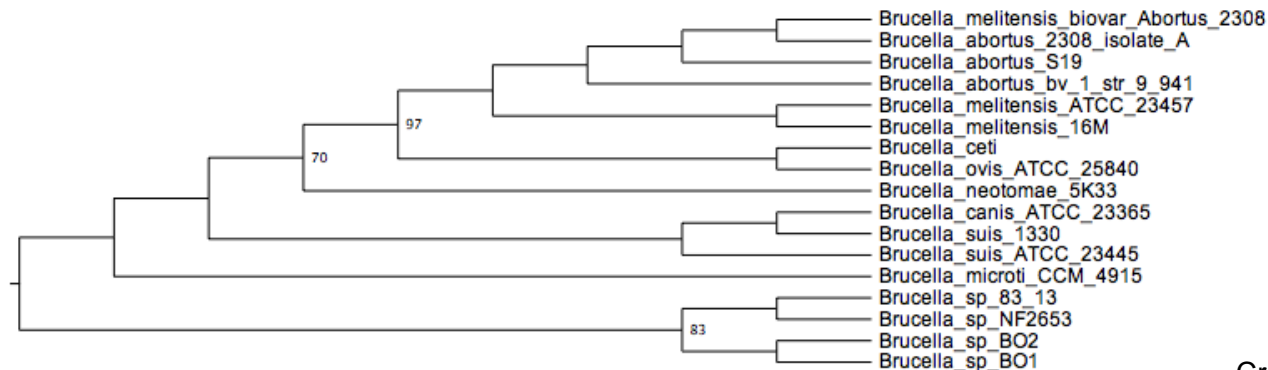
Topologia e comprimento de ramos

A



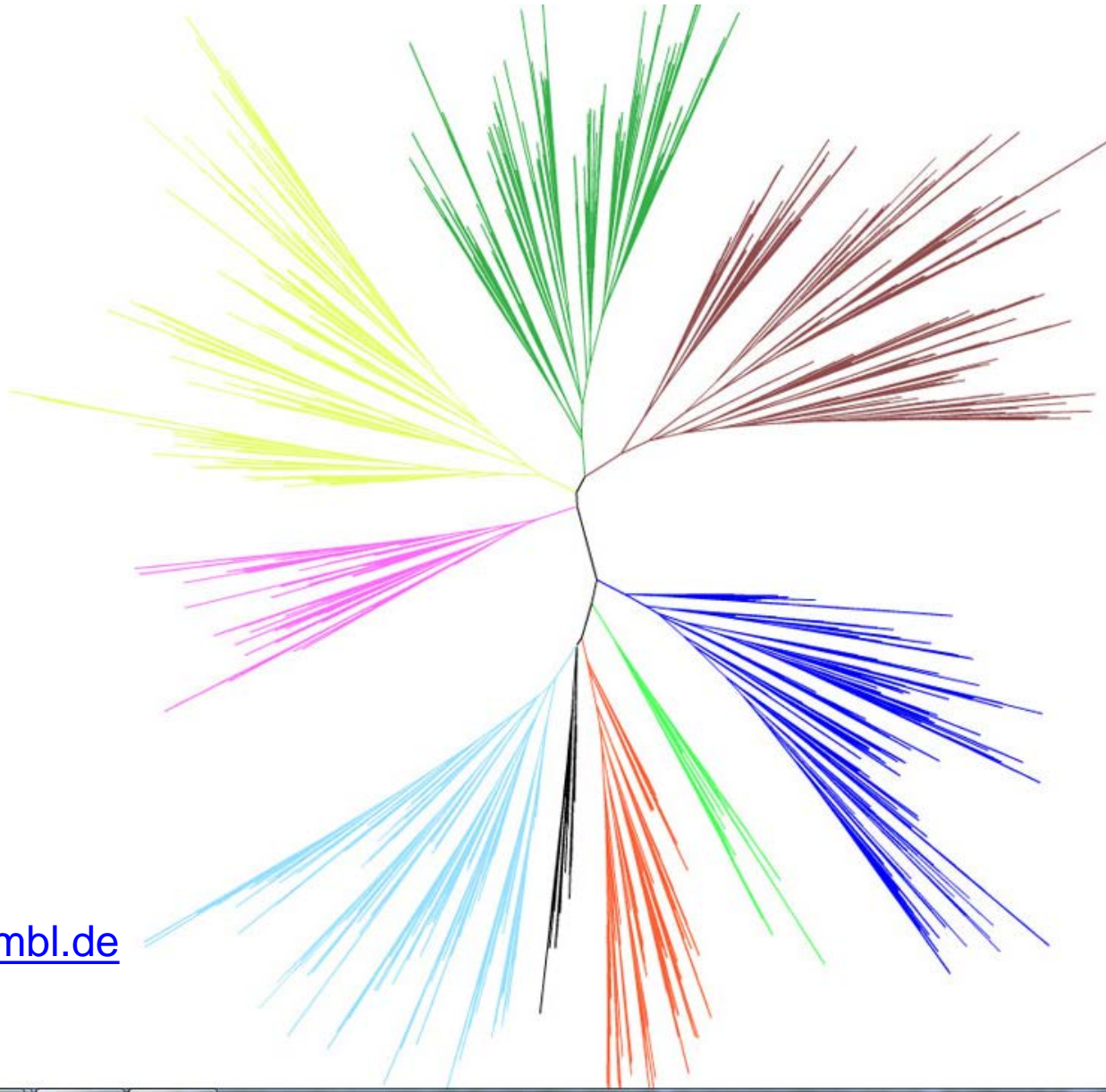
Cladogram version

B



Credit: Wattam et al. 2011

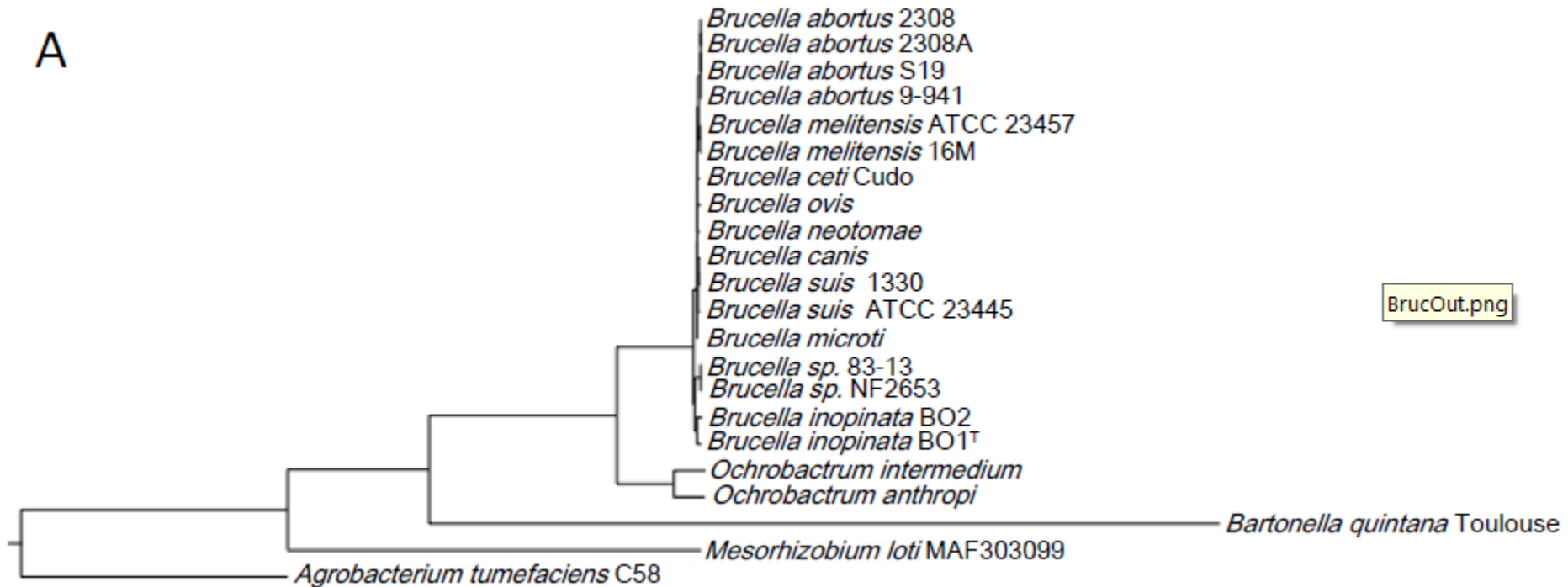
Árvore sem raiz



<http://itol.embl.de>

Árvore com raiz: precisa de um grupo externo

A



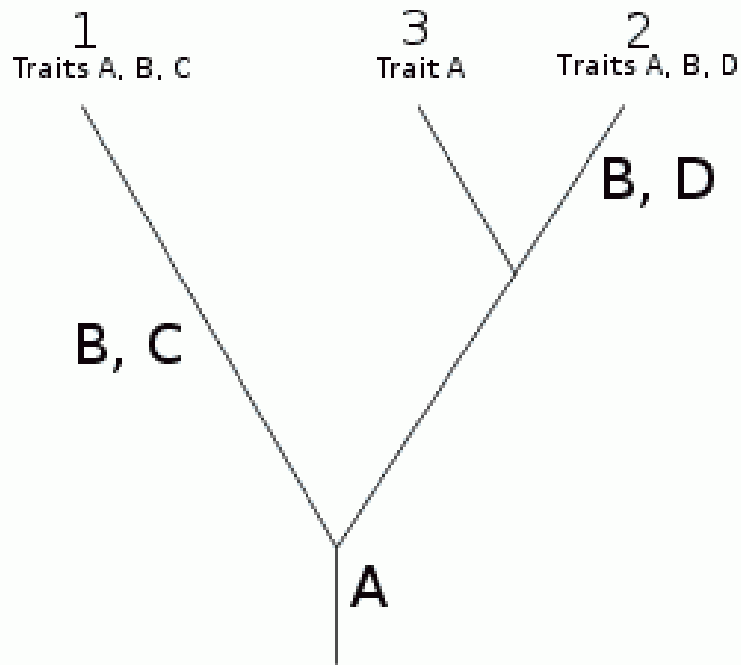
Métodos para reconstrução filogenética

- 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

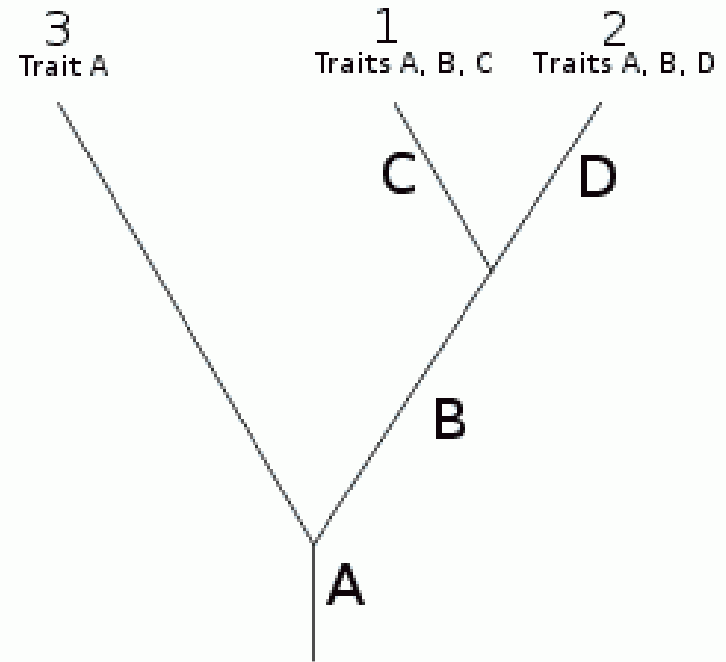
Árvore a partir de uma matriz de distâncias

- Métodos
 - UPGMA
 - Neighbor-joining (NJ)

Parcimônia



Less parsimonious cladogram



More parsimonious cladogram

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

Parcimônia é um princípio muito usado

- 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 **honest**a 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 honest**a?
- 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
- $L(\theta(x)) = P(x \mid \theta)$
- $L(100\text{coroas}) = P(\text{honestidade} \mid 100\text{coroas})$

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

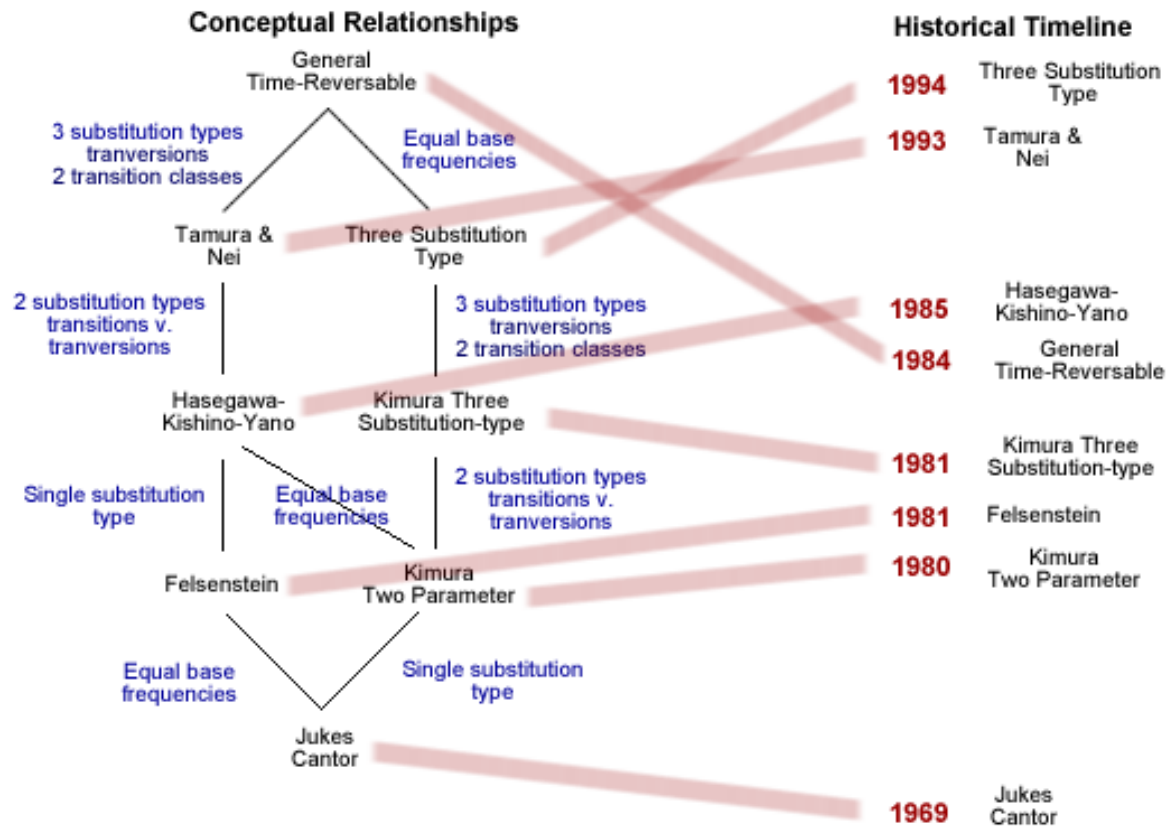
Considerações de tempo de execução

- 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 *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

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)





ProtTest 2.4 server

based on Pal & PhymI

sponsored by

Fundación **BBVA**

Input an alignment and (optionally) a tree

Select a [protein alignment](#)

No file selected.

Tree (in newick format)

Build BioNJ tree User tree

No file selected.

Model Selection options

Model Selection Criterion

AIC

Sample size interpretation (for AICc and BIC)

Total number of characters (alignment 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

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

Programas para inferência filogenética

- **Pacotes**
 - Oferecem vários diferentes programas
 - Diferentes métodos para o mesmo objetivo
 - Podem incluir programas auxiliares
- **Programas individuais**
 - São especializados num métodos

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/>
 - Atualmente na versão 6 (versão 7 beta)

Programas que implementam métodos não-probabilísticos

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

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

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



Xiao Lin

Wageningen UR

Why is MEGA NOT a good program for Phylogenetics?

After taking a PhD course for Phylogenetics, the teachers do not recommend MEGA at all, the reason is MEGA works like a blackbox and the results sometimes will be wrong...We learn a lot of methods and programs like ML, MP, MrBayes, TNT, RAxML, PAUP, BEAST*.....

I know MEGA is the most handy program for beginners and accepted by a lot for publishing aiming, but I want to know more about the drawbacks of this program and why the experts will not select it?

TOPICS

Phylogenetic Tree

MEGA

Phylogenetic Analysis

Oct 24, 2013



Justin C Bagley · University of Brasília

I would agree that the more widely used software programs are RAxML, GARLI, PAUP*, BEAST, MrBayes, phymI, etc. and not MEGA... and you can also scratch TNT off the list. If it is useful at all, it may be that parsimony is only really useful in comparison with the other methods. Anyway, it's interesting to ask why more people don't use MEGA. To answer this directly... I think there is an impression that MEGA uses algorithms that take shortcuts and do not produce the kinds of high quality results and outputs that people expect or are interested in publishing. However, as shown in the Tamura et al. 2011 MBE paper on MEGA5, under at least some circumstances, MEGA5 is faster and produces similar results to phyML and RAxML, two of the most widely used programs for ML analysis!

That said, I also don't rely on MEGA5 for publishable phylogenetics work. But MEGA5 is great for some things. In particular, I find that it is an excellent, easy-to-use starting point for a phylogenetic or phylogeographic analysis. First align your data, and then infer a tree from your data in MEGA. This takes very little time. MEGA is also the best way in my opinion to quickly calculate pairwise genetic distances from sequence data, although you may want to use PAUP* or GARLI to apply a model to your distance calculations (i.e. if you want to do this for making final calculations for publication; however, model-based distances will be higher than those from methods that make fewer assumptions, like calculating p-distances). MEGA works great for this though when you have DNA sequence data. MEGA is also great for web-based acquisition of sequence data (e.g. grabbing sequences from GenBank). And I also think it is good for calculating dn/ds ratios and testing for neutrality, e.g. using the MK test (although I tend to use other software like DnaSP for this, it is sometimes useful to do this in MEGA5 if you've already got your data assigned to different, relevant groups).

Justin Colonial Bagley



Possui bacharelado (2004) e mestrado (2008) em Biologia pela The University of Alabama e doutorado em Biologia Integrativa (2014) pela Brigham Young University. Tem experiência na área de Genética, com ênfase em Evolução, Ecologia e Genética (Animal). My research is broadly focused on the origin and maintenance of biodiversity in the wild. I study micro- and macroevolutionary processes across various scales, drawing on a range of theoretical and analytical frameworks including: phylogeography, phylogenetic systematics, geographic population structure, population history and demography (e.g. gene flow), morphometrics, and life history variation. My work primarily emphasizes freshwater fish communities as ideal model systems for addressing key questions in ecology and evolution. The geographic focus of my research program spans aquatic habitats across ecological settings in (1) Brasil, (2) Central America, (2) the North American Gulf-Atlantic Coastal Plain, and (4) Australia. A running thread of my research is exploring ways of combining inferences from molecular genetic markers and/or field-collected specimens with independent data. In particular, I am interested in incorporating data on current and historical landscapes drawn from GIS, ecological niche modeling, geology, and paleoecology into studies of the evolution of natural populations and communities through space and time. To date, much of my research has involved four areas: phylogeography and historical biogeography, molecular phylogenetic systematics, DNA-based species delimitation, and life-history evolution of freshwater fishes. Research Competencies A. Ichthyology: Field collection, identification, and preservation of fishes; microscopy; dissection; meristic and morphometric analyses of fishes. B. Molecular and Computational Evolutionary Biology: DNA extraction, PCR and clean-up, gel electrophoresis, DNA sequencing, and protocols for DNA purity determination and quantification Sequence editing, (SEQUENCHER, Geneious, PHASE) and alignment (CLUSTALw, MUSCLE, MAFFT, T-COFFEE). Deposition in online repositories (e.g. GenBank, TreeBASE) Population genetic divergence, DNA sequence polymorphism (PAUP*, DnaSP, MEGA5, msBayes) Population structure and differentiation and landscape genetics (BARRIER, GENEPOP, ARLEQUIN, SAMOVA, GENELAND, STRUCTURE; isolation by distance using regression and Mantel tests in IBD, GenAlEx, PASSAGE2). Isolation with migration (θ and effective migration estimation; LAMARC, full-likelihood and full-Bayesian parameter estimation in Migrate-N, IM and IMA2) Phylogenetic model selection (jModelTest, DT-ModSel) Phylogeny reconstruction and nodal support (parsimony, likelihood, Bayesian methods; e.g. TNT, PAUP*, GARLI, RAxML, MrBayes, BEAST) Divergence-dating analysis (PAUP*, r8s, MrBayes, BEAST, *BEAST) Species tree analysis (*BEAST, BEST, JML, Minimize Deep Coalescences in Mesquite) Coalescent simulations and model testing (BEAST, DnaSP, Mesquite, MTML msBayes, DIY-ABC) Historical demography and population expansion (DnaSP, Arlequin, BEAST) Discrete and continuous phylogeographic analysis (BEAST) Comparative phylogeography Phylogenetic comparative methods: phylogenetic signal, correlated evolution, PGLS, phylogenetic logistic regression, ancestral state reconstruction, phylogenetic trait mapping R environment (scripts, packages ADE4, APE, GEIGER, PICANTE, CAPER, GENELAND, SMATR) and graphics Command line for PC and mac OS platforms (DOS, R, UNIX, Linux; e.g., Terminal), grep Statistical analysis of DNA, parametric and nonparametric data C. GIS and Ecological Modeling: Basic cartography skills/competencies in DIVA-GIS, ESRI ArcMap 10 (e.g. ArcToolbox functions, Spatial Analyst extension tools) Data layer manipulation Ecological niche modeling in Maxent: modeling, replication, projection, layer masking, model testing and interpretation

Certificado pelo autor em 05/10/2015

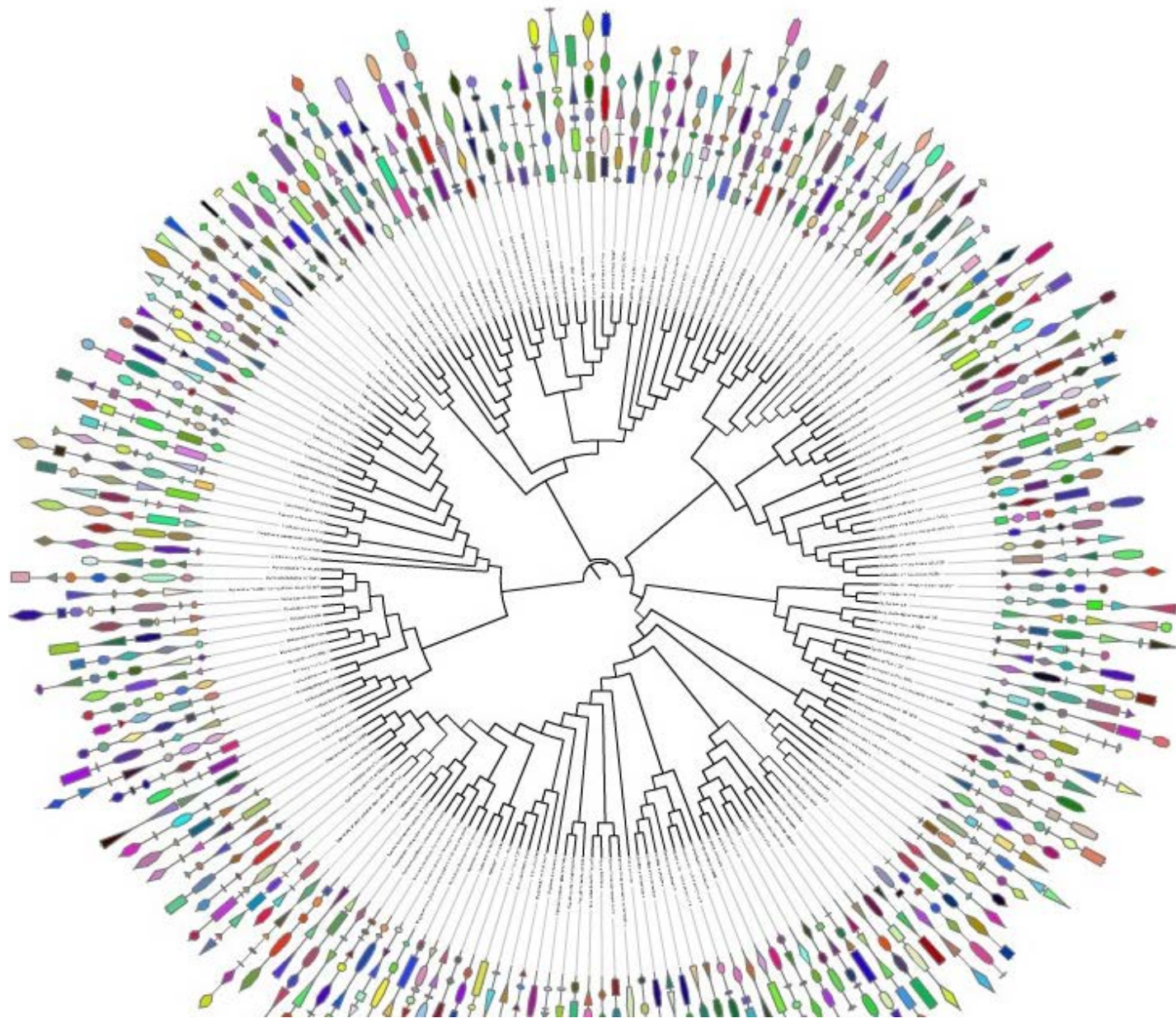
Universidade de Brasília, Campus Darcy Ribeiro, Departamento de Zoologia.

Visualização de árvores: formatos

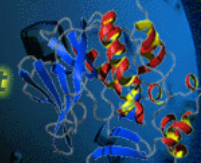
- **Newick, NEXUS**
- (((erHomoC:0.28006,erCaelC:0.22089):0.40998,(erHomoA:0.32304,(erpCaelC:0.58815,((erHomoB:0.5807,erCaelB:0.23569):0.03586,erCaelA:0.38272):0.06516):0.03492):0.14265):0.63594,(TRXHomo:0.65866,TRXSacch:0.38791):0.32147,TRXEcoli:0.57336);
- <http://molecularevolution.org/resources/treeformats>

Visualização de árvores

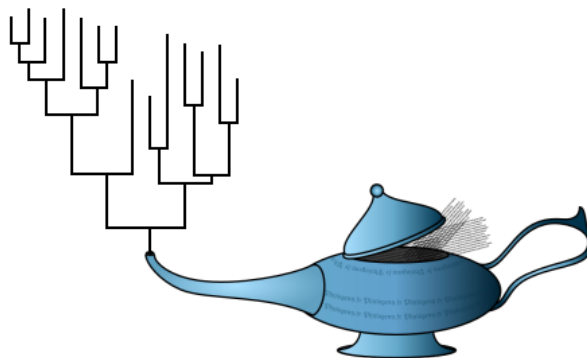
- Interactive Tree of Life <http://itol.embl.de>
- http://en.wikipedia.org/wiki/List_of_phylogenetic_tree_visualization_software



All-in-one: phylogeny.fr



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 Génopoles (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.

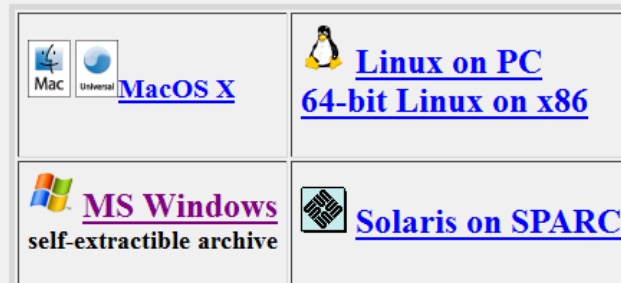
NEW: seaview drives the Gblocks program to select blocks of conserved sites.

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 [dnapars/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



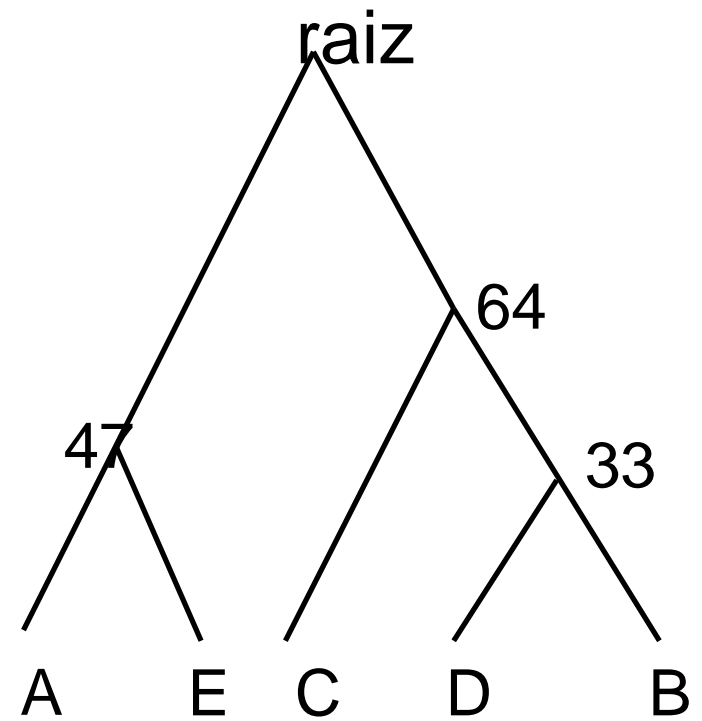
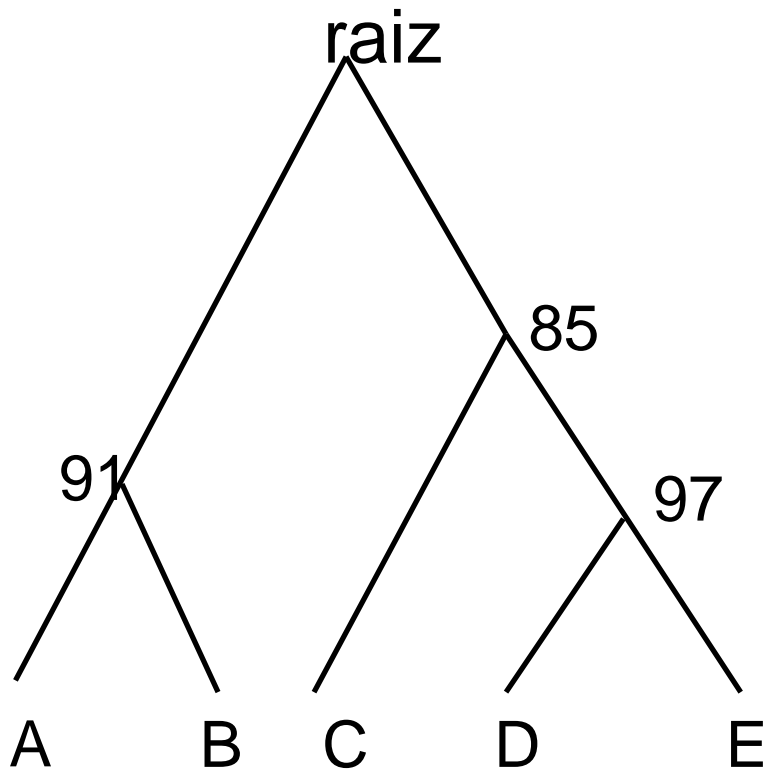
Interpretação

- 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

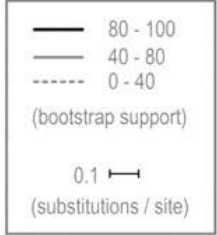
- Valores de **bootstrap**
- Colunas do AM são amostradas aleatoriamente em várias corridas (**replicatas**; geralmente entre 100 e 1000)
- Árvores resultantes são comparadas entre si
- Concordâncias nos clados são calculadas, resultando em número de vezes (ou %) que clados se repetem nas replicatas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

Exemplo de árvores com bootstrap



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!



Nanoarchaeota
 Crenarchaeota
 Euryarchaeota
 Diplomonadida
 Kinetoplastida
 Chromalveolata
 Plantae
 Amoebozoa
 Fungi

Metazoa

Firmicutes

Planctomycetes
 Spirochaetes
 Actinobacteria
 Fibrobacteres
 Chlorobi
 Bacteroidetes

Ciccarelli et al, *Science*, 2006

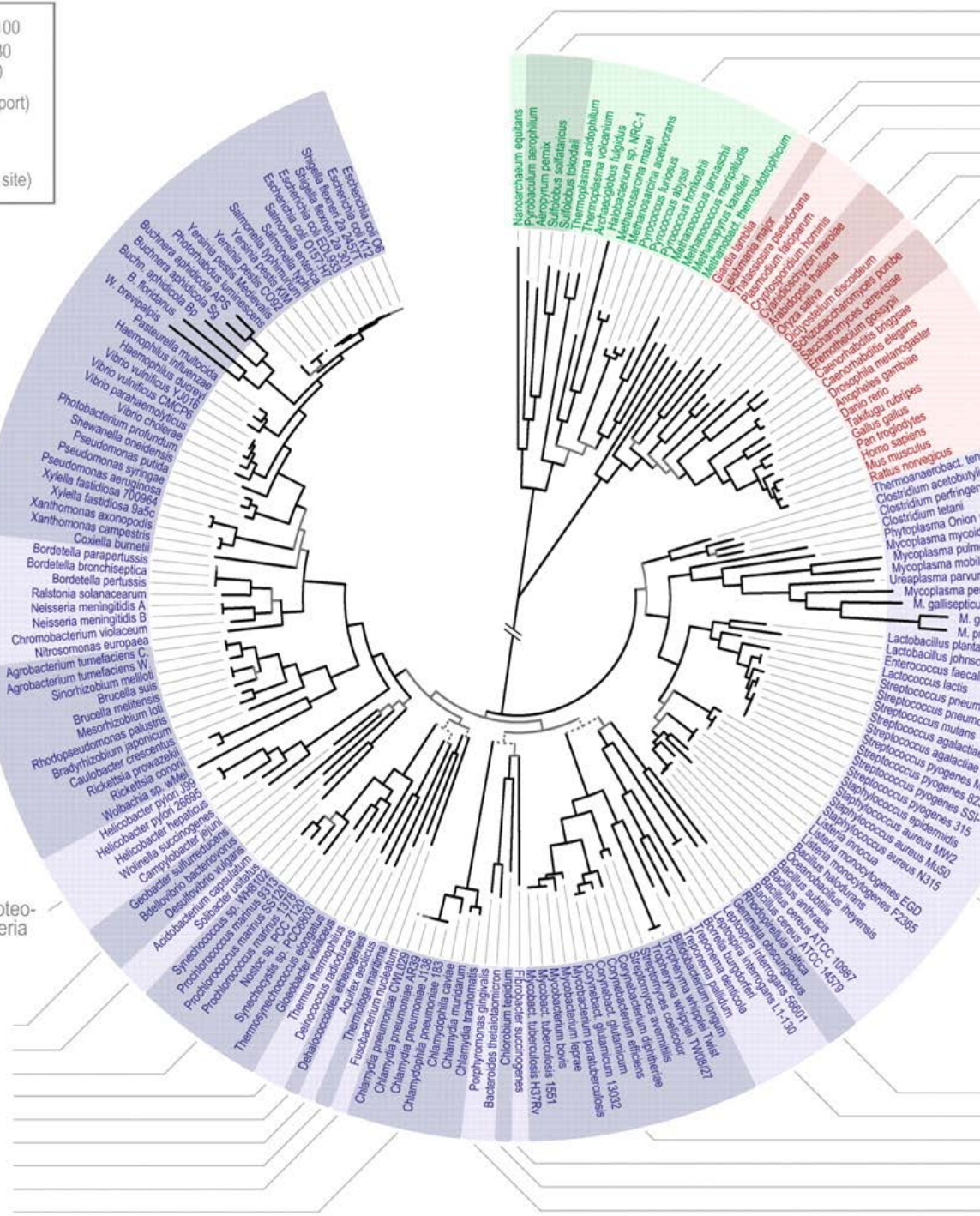
γ -proteo-
bacteria

β -proteo-
bacteria

α -proteo-
bacteria

ϵ -proteo-
bacteria

δ -proteo-
bacteria
 Acidobacteria
 Cyanobacteria
 Deinococcales
 Chloroflexi
 Aquificae
 Thermotogae
 Fusobacteria
 Chlamydiae

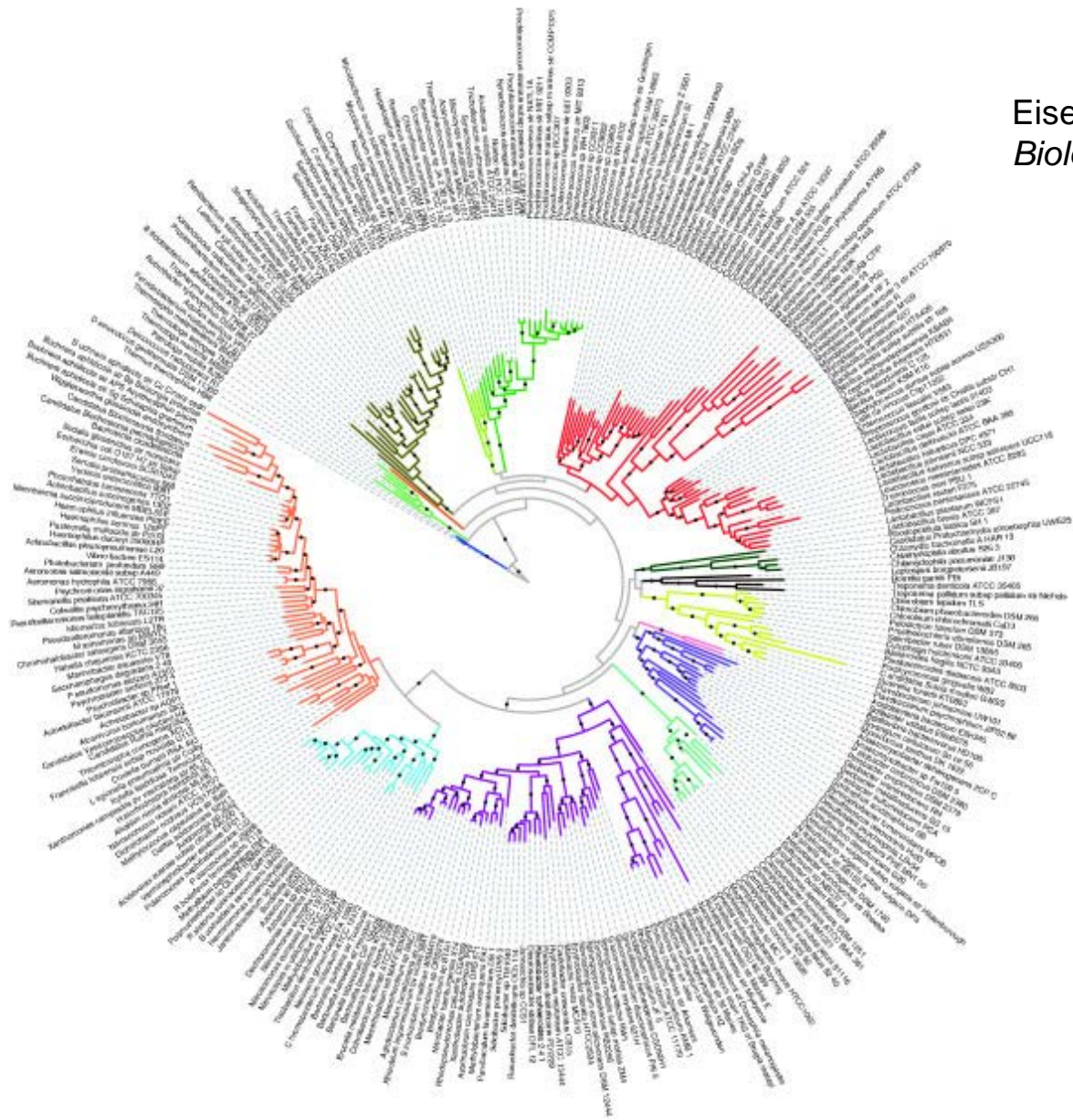


Supermatrizes

- 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
 - Ramos longos (muitas mutações) tendem a ficar artificialmente próximos um do outro (e próximos da raiz)
 - Topologia errada
- 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



Fungos e queijos

- Fabricação de queijos depende da ação de fungos
- Roquefort
 - *Penicillium roqueforti*
- Camembert
 - *P. camemberti*
- Esses fungos vem sendo selecionados e cultivados há séculos

Resultado recém publicado

- Ao comparar diferentes espécies de fungos usados em queijos, descobriu-se
 - *Multiple Recent Horizontal Gene Transfers between Distant Penicillium Species, Flanked by Specific Retrotransposons*

Adaptive Horizontal Gene Transfers between Multiple Cheese-Associated Fungi

Jeanne Ropars,^{1,2,8} Ricardo C. Rodríguez de la Vega,^{1,2,8} Manuela López-Villavicencio,³ Jérôme Gouzy,^{4,5} Erika Sallet,^{4,5} Émilie Dumas,^{1,2} Sandrine Lacoste,³ Robert Debuchy,^{6,7} Joëlle Dupont,³ Antoine Branca,^{1,2,9,*} and Tatiana Giraud^{1,2,9,*}

¹Ecologie, Systématique et Evolution, UMR8079, Univ. Paris-Sud, 91405 Orsay, France

²Ecologie, Systématique et Evolution, UMR8079, CNRS, 91405 Orsay, France

³Institut de Systématique, Evolution, Biodiversité, UMR 7205 CNRS-MNHN-UPMC-EPHE, Muséum national d'Histoire naturelle, Sorbonne Université, CP39, 57 Rue Cuvier, 75231 Paris Cedex 05, France

⁴Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR441, INRA, Castanet-Tolosan 31326, France

⁵Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR2594, CNRS, Castanet-Tolosan 31326, France

⁶Institut de Génétique et Microbiologie, UMR8621, Univ. Paris-Sud, 91405 Orsay, France

⁷Institut de Génétique et Microbiologie, UMR8621, CNRS, 91405 Orsay, France

⁸Co-first author

⁹Co-senior author

*Correspondence: antoine.branca@u-psud.fr (A.B.), tatiana.giraud@u-psud.fr (T.G.)

<http://dx.doi.org/10.1016/j.cub.2015.08.025>

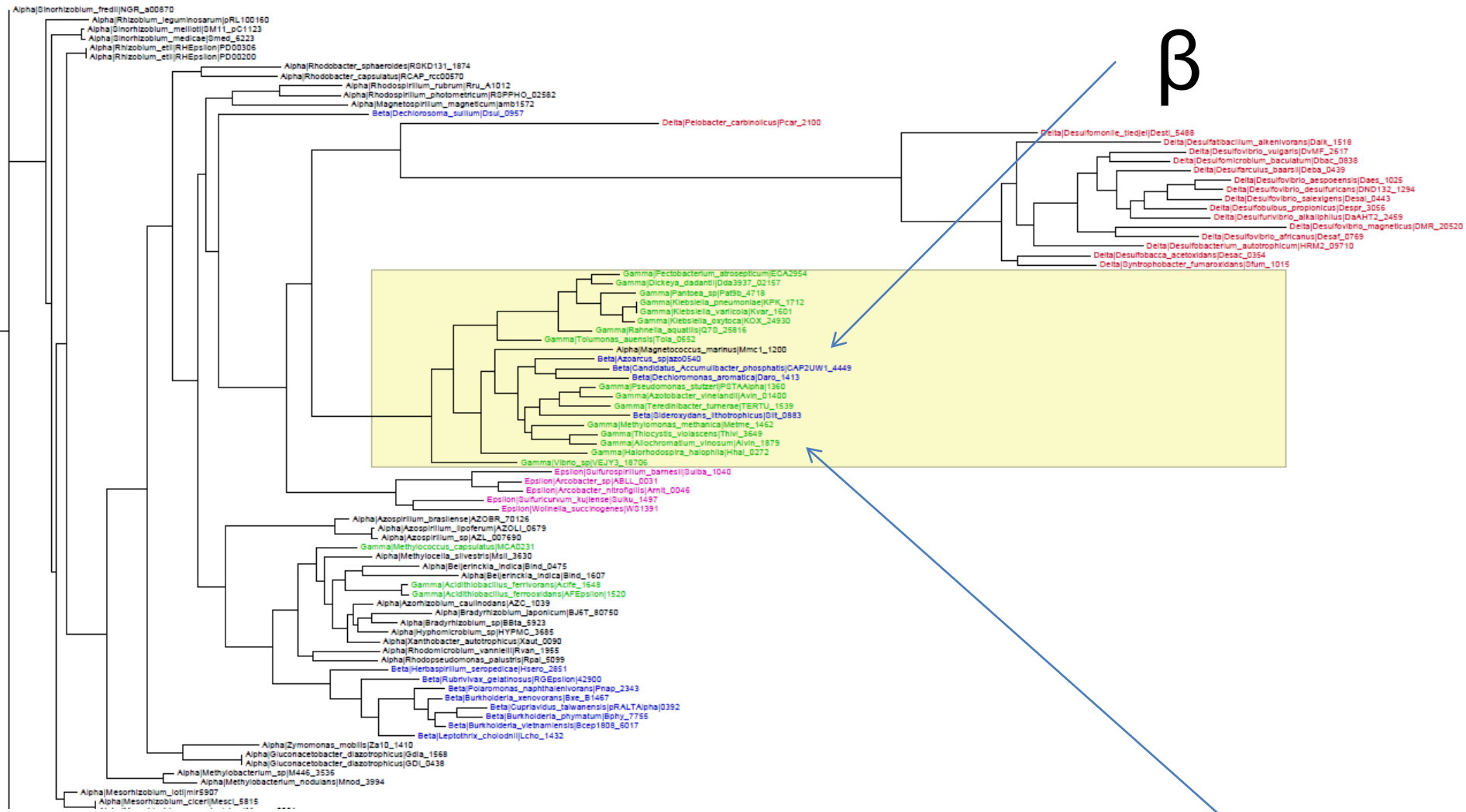
This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

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

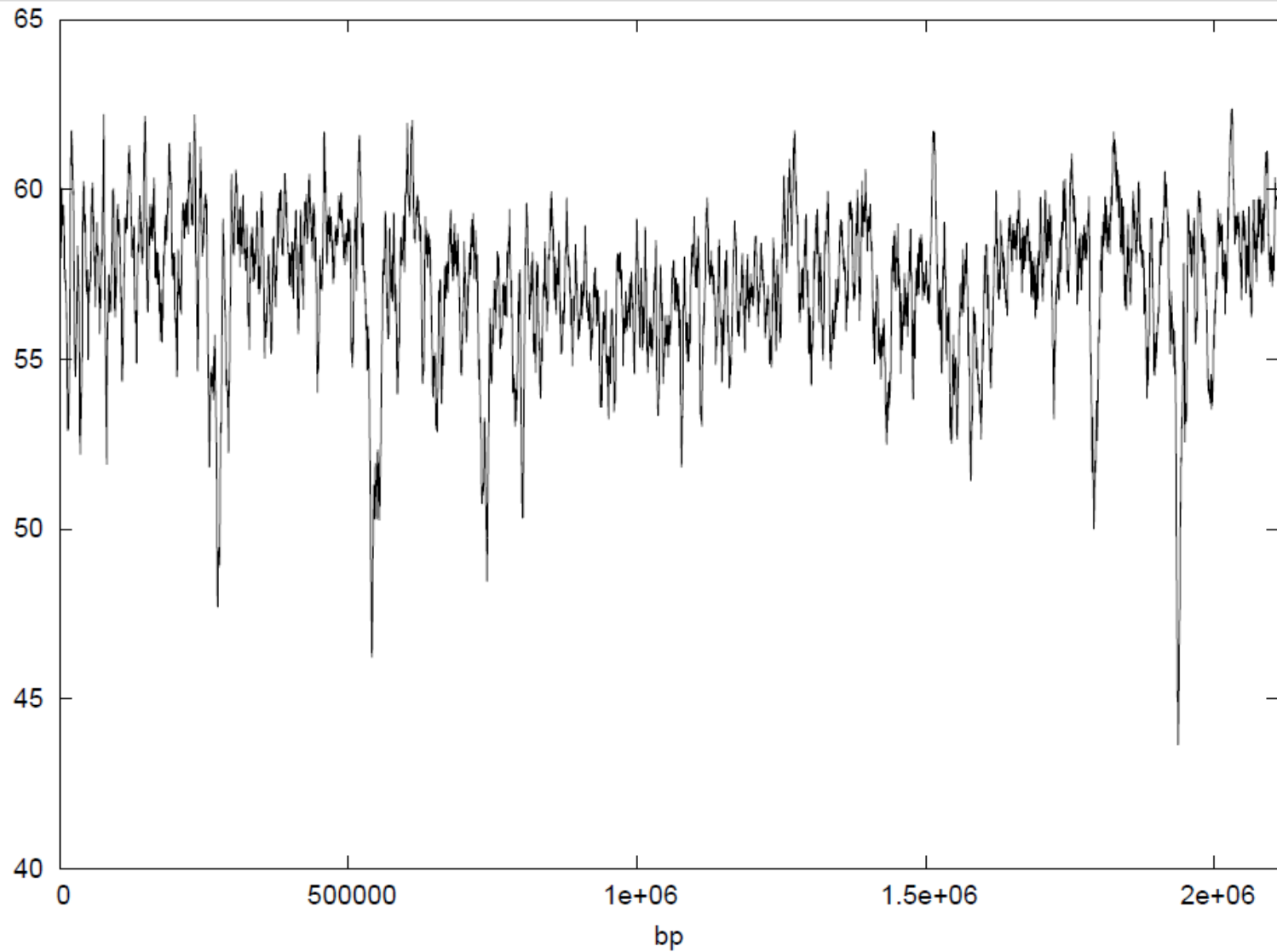


β

γ gamma

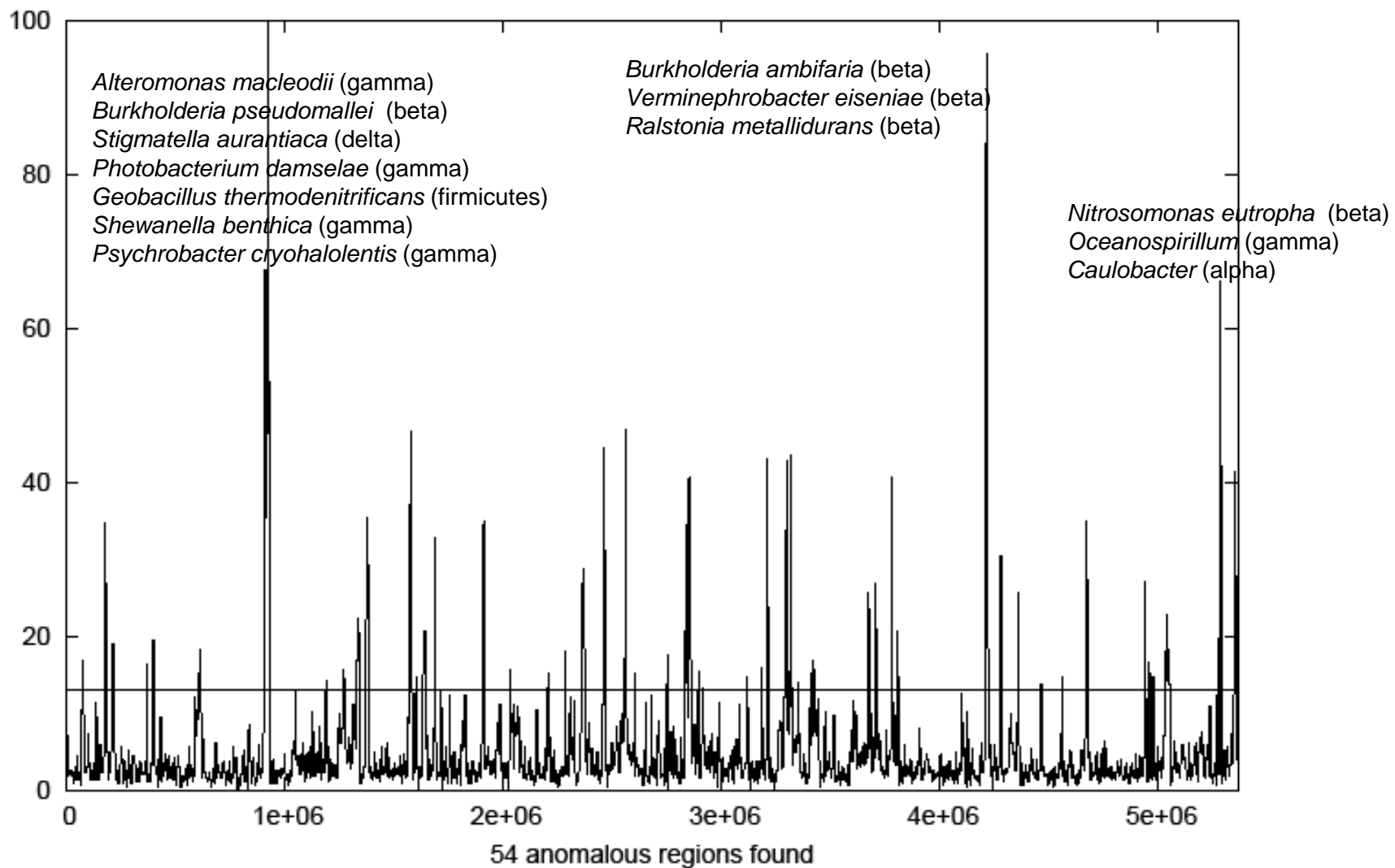
THG recente

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



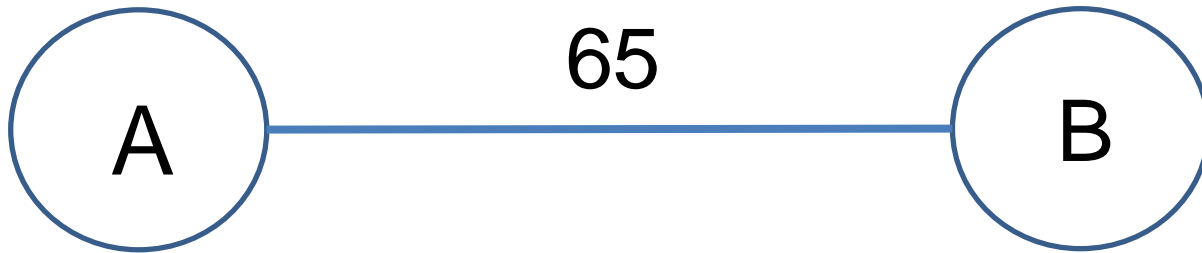
Variação do %GC no cromossomo principal de *Brucella ovis* ATCC25840

Azotobacter vinelandii anomalous regions (Alien_hunter threshold: 13.020)



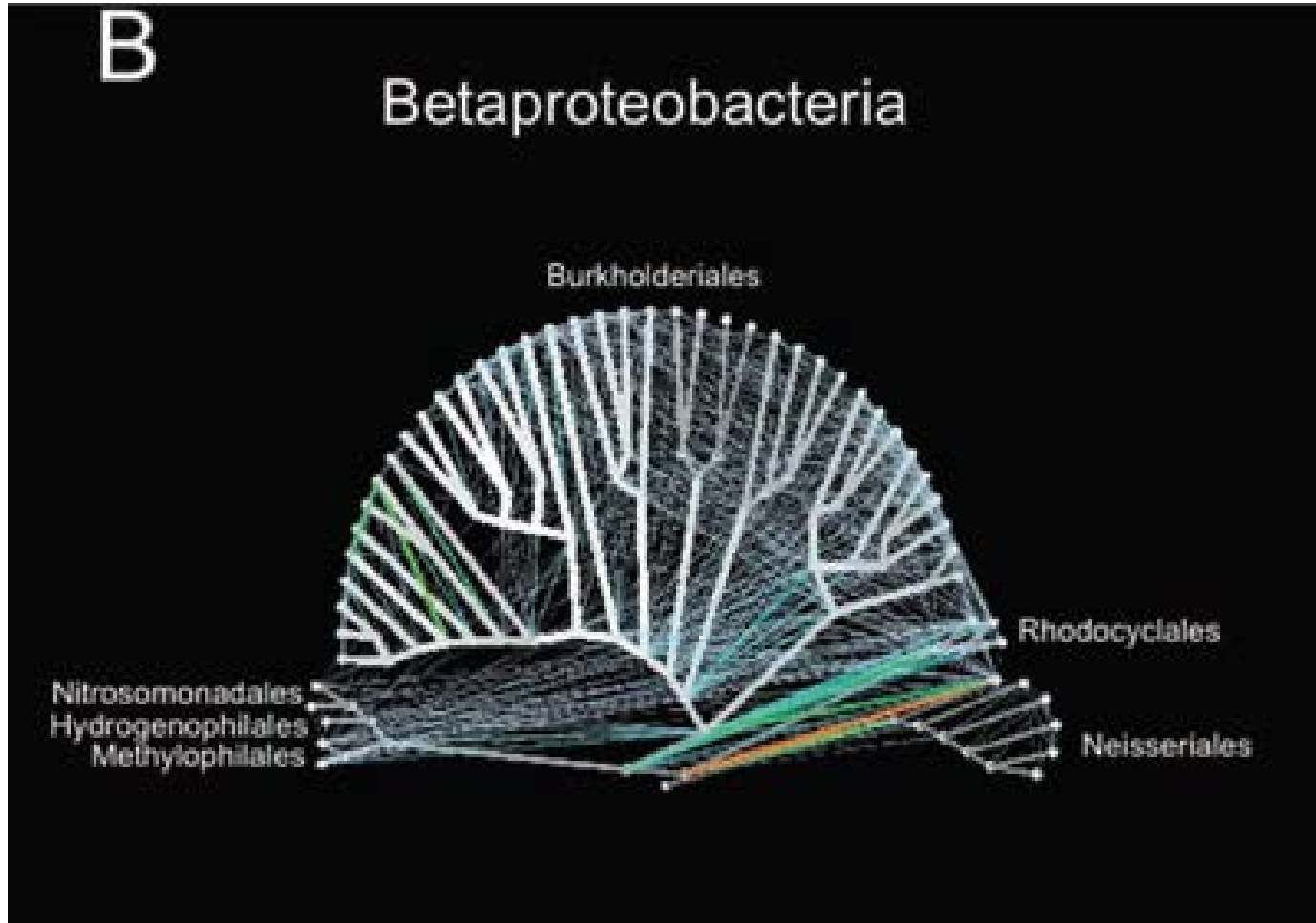
Redes filogenômicas

- 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

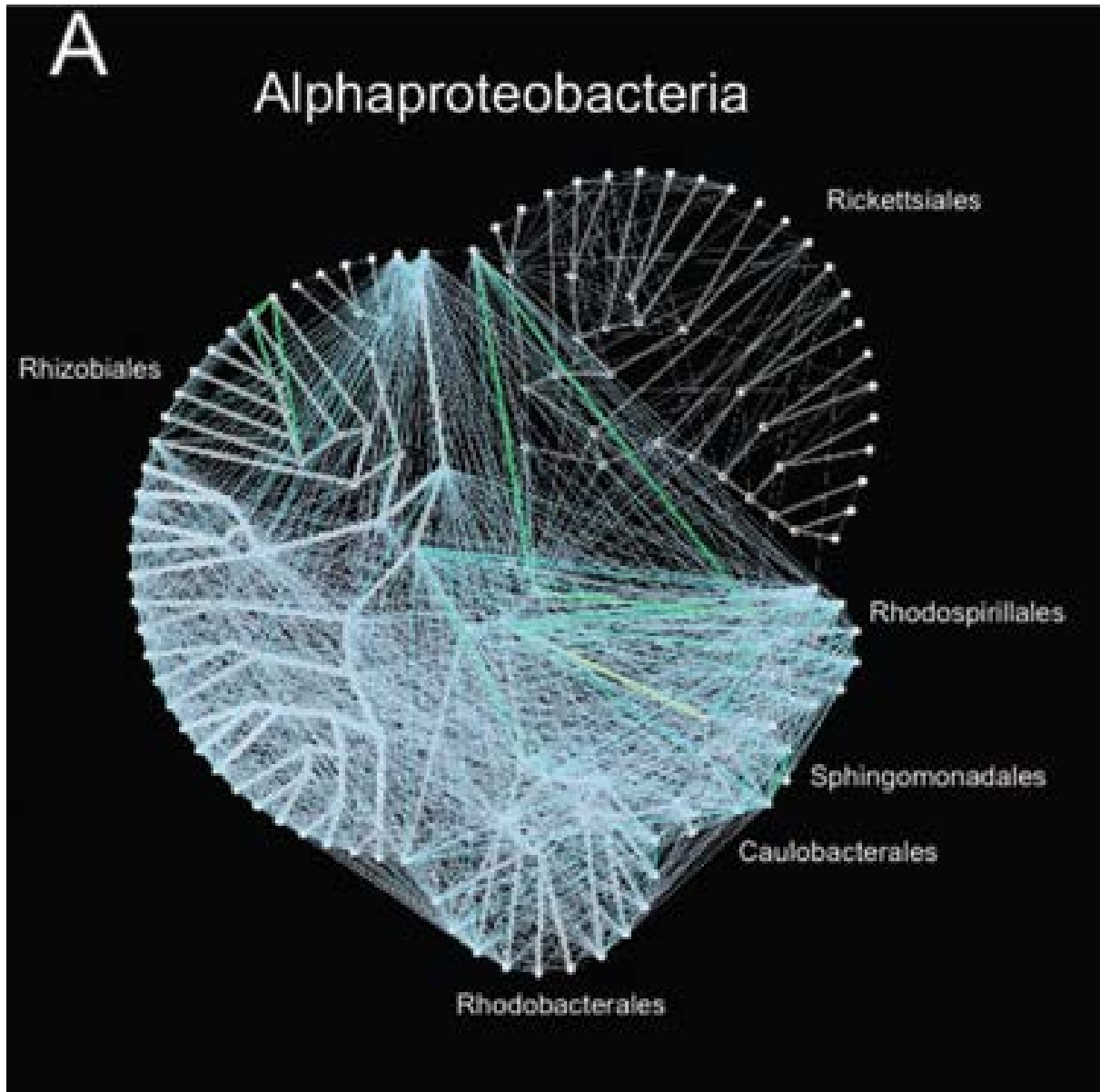


Kloesges et al, *Molecular Biology and Evolution*, 2011

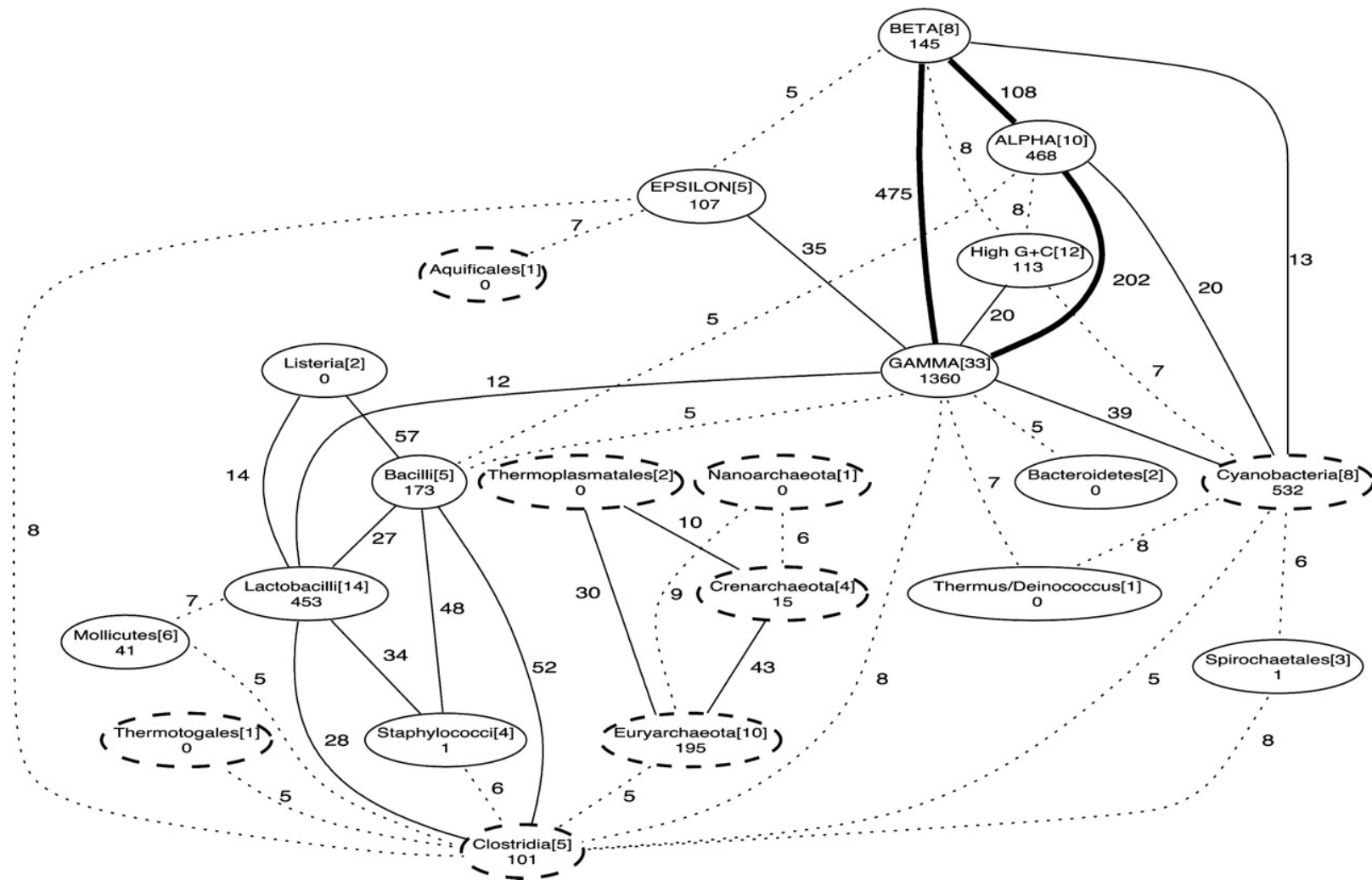
8/23/2018

J. C. Setubal

71



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

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>

Para saber mais

- 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)

Os 100 artigos mais citados

THE TOP 100 PAPERS

Nature explores the most-cited research of all time.

BY RICHARD VAN NOORDEN,
BRENDAN MAHER AND REGINA NUZZO

The discovery of high-temperature superconductors, the determination of DNA's double-helix structure, the first observations that the expansion of the Universe is accelerating — all of these breakthroughs won Nobel prizes and international acclaim. Yet none of the papers that announced them comes anywhere close to ranking among the 100 most highly cited papers of all time.

of carbon nanotubes (number 36) are indeed classic discoveries. But the vast majority describe experimental methods or software that have become essential in their fields.

The most cited work in history, for example, is a 1951 paper² describing an assay to determine the amount of protein in a solution. It has now gathered more than 305,000 citations — a recognition that always puzzled its lead author, the late US biochemist Oliver Lowry. “Although J. Biol. Chem. is not a

to other scientists what kind of work we’re doing”. Another common practice is to ensure that truly foundational discoveries — Einstein’s special theory of relativity, for instance — get fewer citations than they might deserve: they are so important that they quickly enter the textbooks or are inserted into the main text of papers as terms so familiar that they do not need a citation.

Citation counts are riddled with confounding factors. The volume of citations has increased, for example — yet older papers had more time to accrue citations. Engineers tend to cite one another’s work more frequently than, say, physicists. And not all fields have the same number of publications. Molecular biologists therefore recoil from metrics as crude as simply counting citations when they want to measure a paper’s value: instead, they prefer to compare counts for papers of similar age, and in comparable fields.

Nor is Thomson Reuters’ list the only one being compiled. Google Scholar has its own top-100 list for *Nature*. It includes many more citations because the search engine pulls references from a much greater (and poorly characterized) literature base, including books from a large range of publishers. In that list, available at www.nature.com/top100, ecology papers have more prominence. Google Scholar’s list also features books, which Thomson Reuters did not analyse. But among the top papers, many of the same titles show

up. Yet even with all the caveats, the Thomson Reuters hall of fame still has value. If not for anything else, it serves as a reminder of the nature of scientific knowledge. To make exciting discoveries, researchers rely on relatively unsung papers that describe experimental methods, data analysis software.

Here *Nature* tours some of the key papers that tens of thousands of citations have piled up on — to the top of science’s Kilimanjaro — but rarely thrust into the limelight.

Nature, 30/10/2014

Eu falei algo errado

- BLAST é o paper mais citado da história
- Não!
- Lowry, O. H., Rosebrough, N. J., Farr, A. L. & Randall, R. J. Protein measurement with the folin phenol reagent. [*J. Biol. Chem.* **193**, 265–275 \(1951\).](#)
- 305.148 citações
- Watson e Crick, hélice dupla (1953): 5.207

Papers de bioinformática

- 10) clustalW: 40289
- 12) blast1: 38380
- 14) blast2: 36410
- 20) NJ: 30176
- 28) clustalX: 24098
- 41) bootstrap: 21373
- 45) MEGA: 18286
- 76) modelTest: 14099
- 100) MrBayes: 12209

