Universidade de São Paulo
Instituto de Química

Análise filogenética para dados moleculares

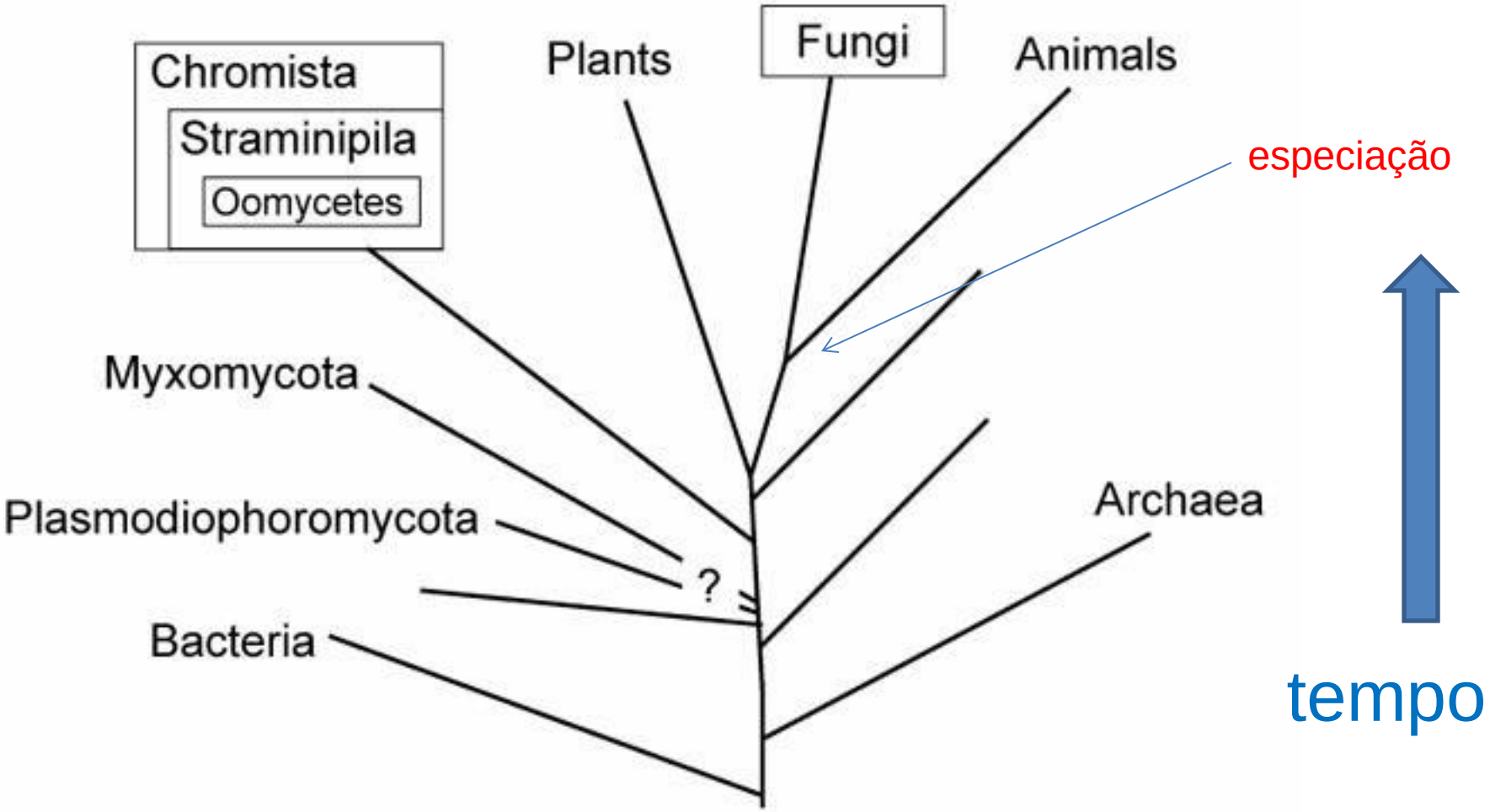
João C. Setubal

2017

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

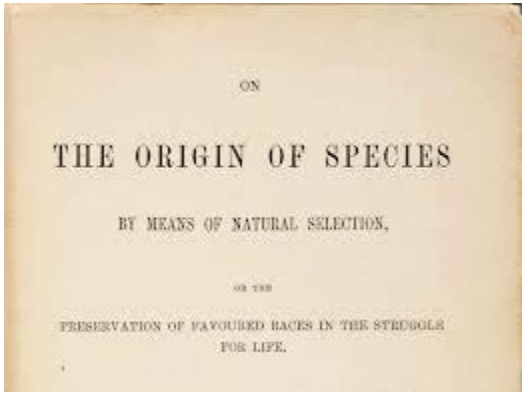
Uma árvore é uma hipótese sobre o que ocorreu na evolução

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)

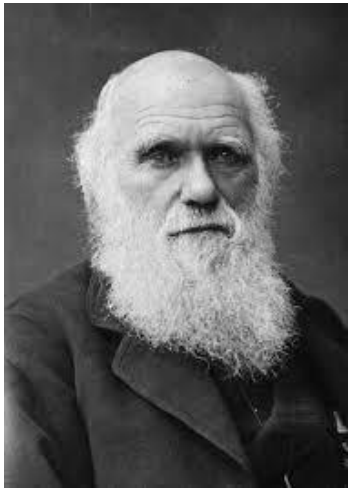

ON

THE ORIGIN OF SPECIES

BY MEANS OF NATURAL SELECTION,

OR THE

PRESERVATION OF FAVOURED RACES IN THE STRUGGLE FOR LIFE.



Evolução e tempo

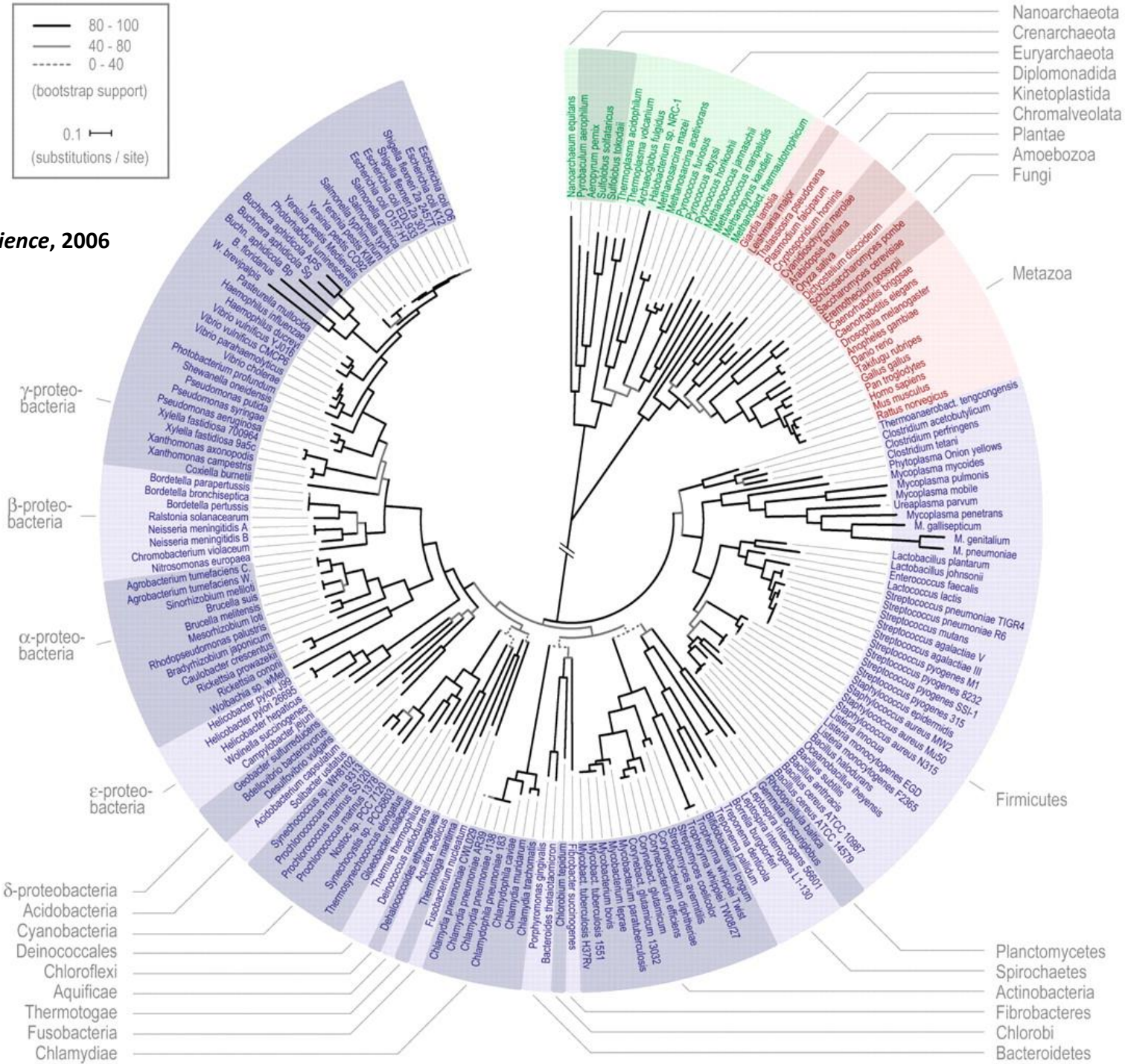
- Árvores com dados moleculares são hipóteses sobre quantas (e quais) mudanças ocorreram nas sequências
- Não são hipóteses sobre o tempo decorrido
- A menos que haja uma ligação entre mudança nas sequências e tempo
  - O relógio molecular

Problemas

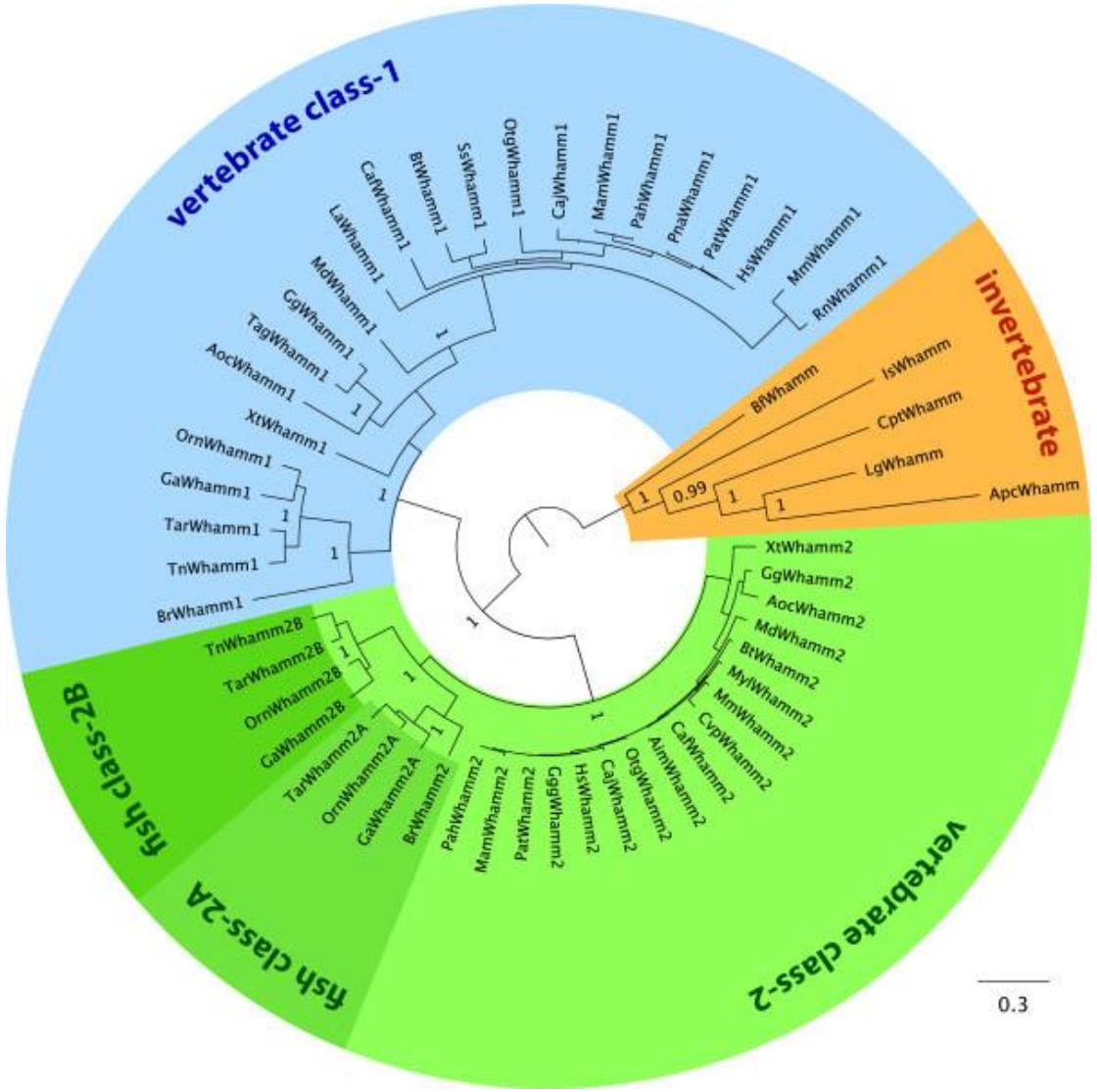
- Evolução não é uniforme no tempo
- Ritmos diferentes
  - Espécies
  - Genes
  - Sítios de genes
- Relógio molecular supõe que existe uniformidade

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



Ciccarelli et al, *Science*, 2006

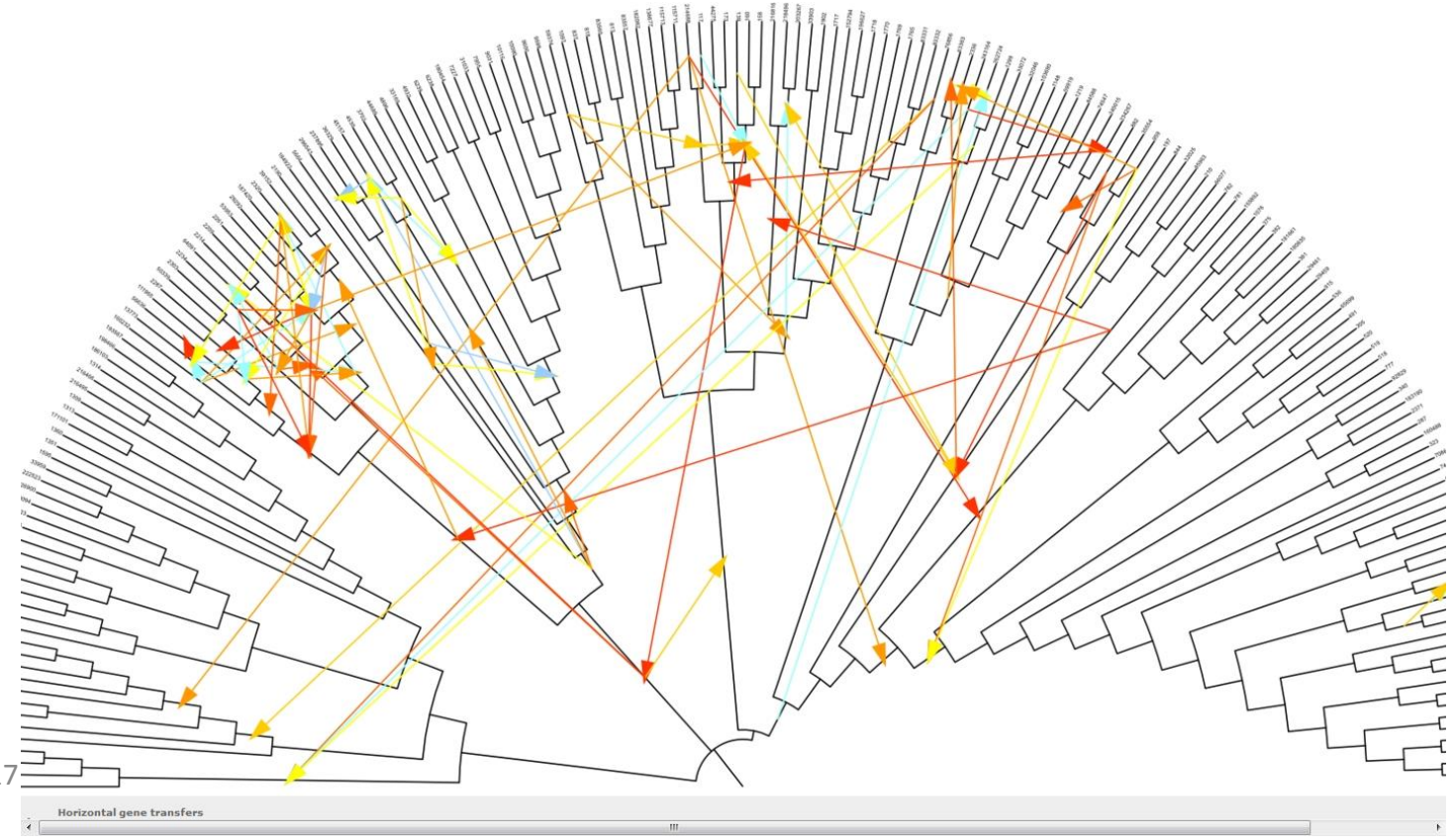


Phylogenetic tree of the WHAMM proteins

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
  - Surtos
    - *Salmonella*
    - Ebola
  - Antraz

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

Phytophthora

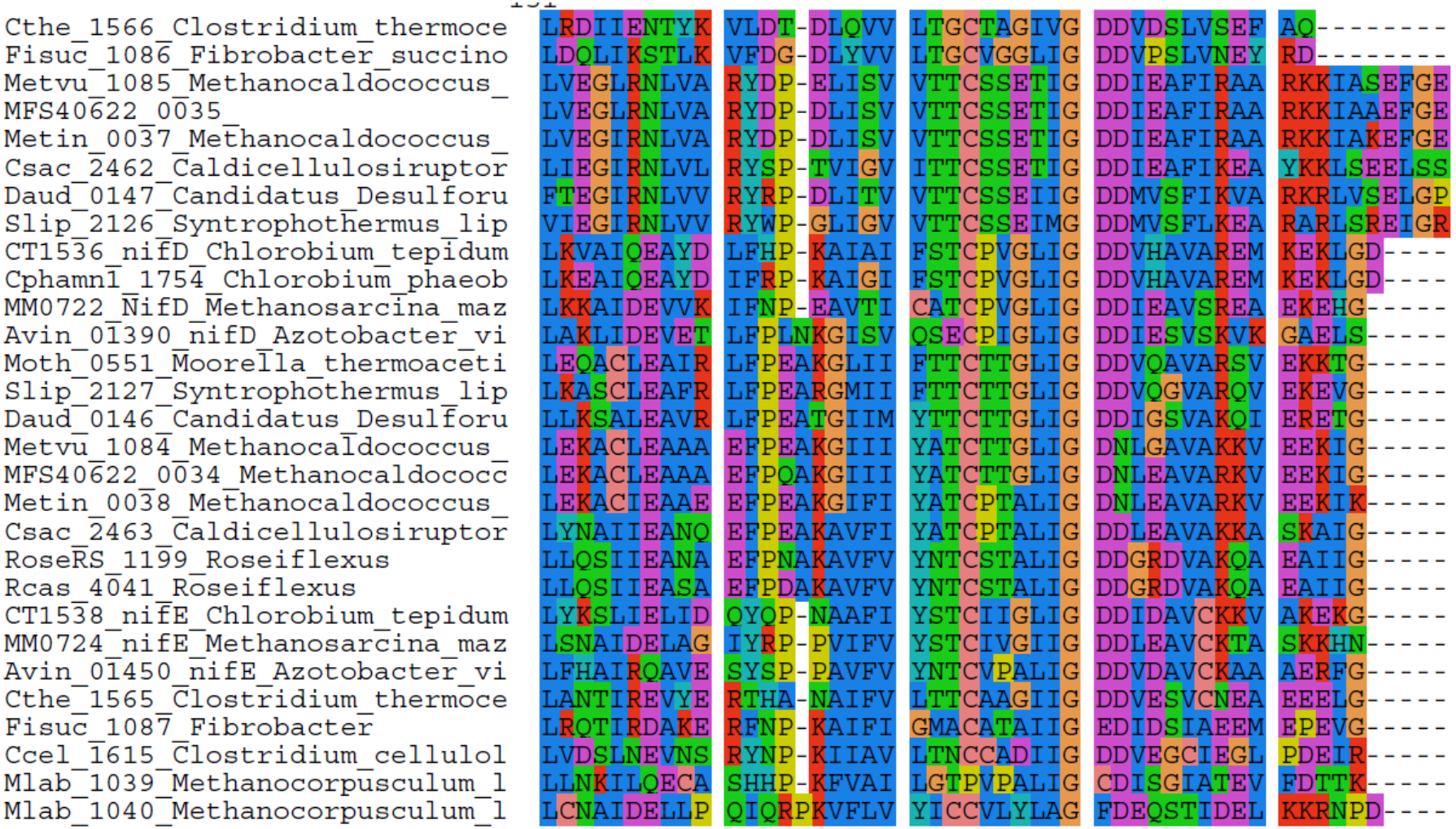
Sequências de entrada

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

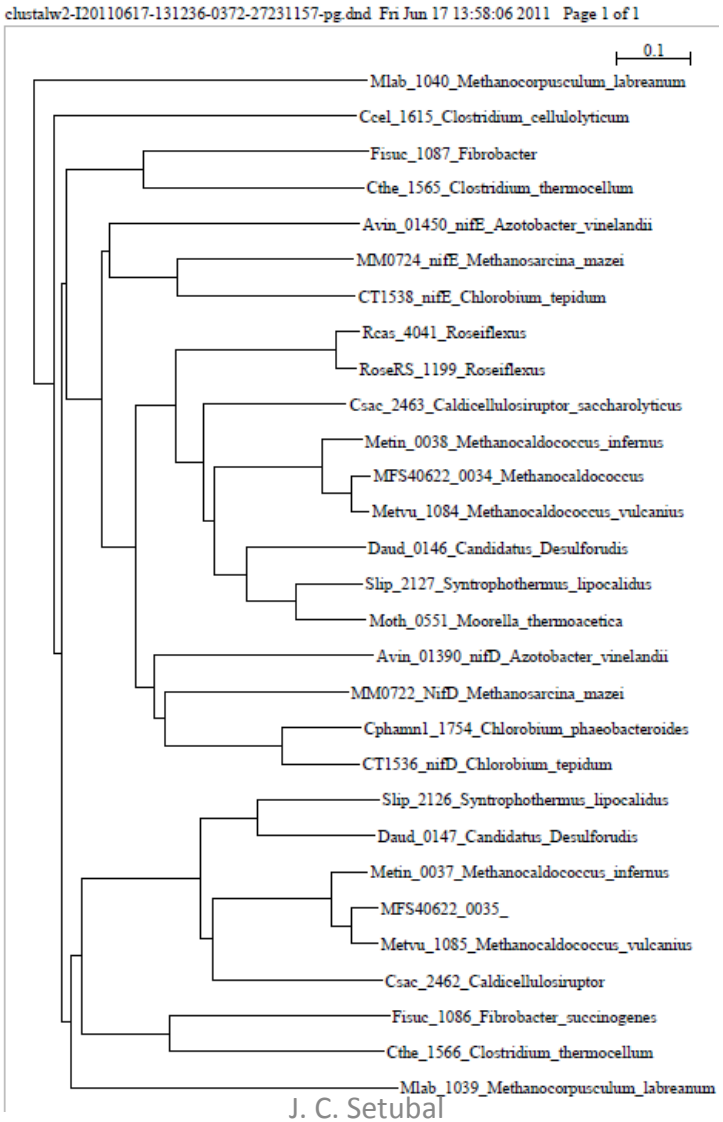
Pipeline

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

Alinhamento múltipo



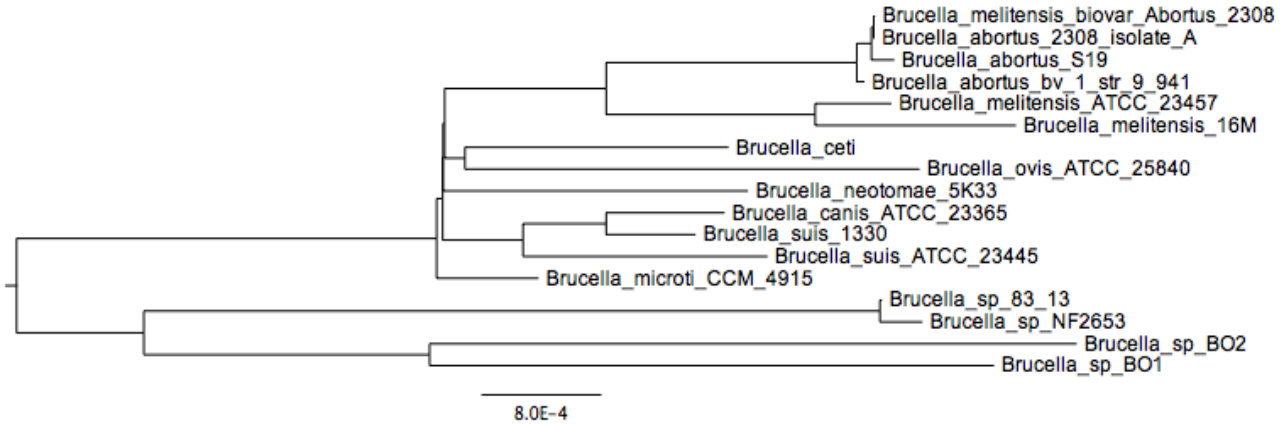
Filogenia resultante



Credit: R. Dixon

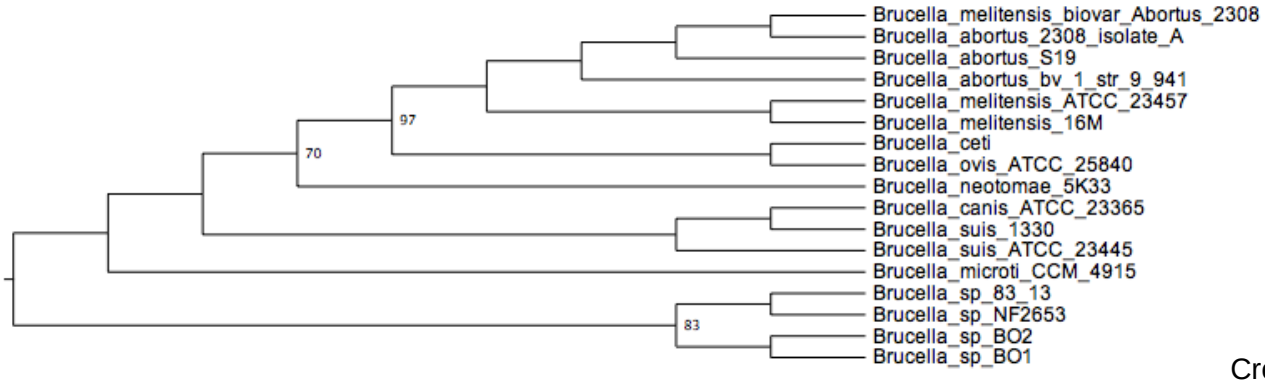
Árvores e cladogramas
Topologia e comprimento de ramos

A



Cladogram version

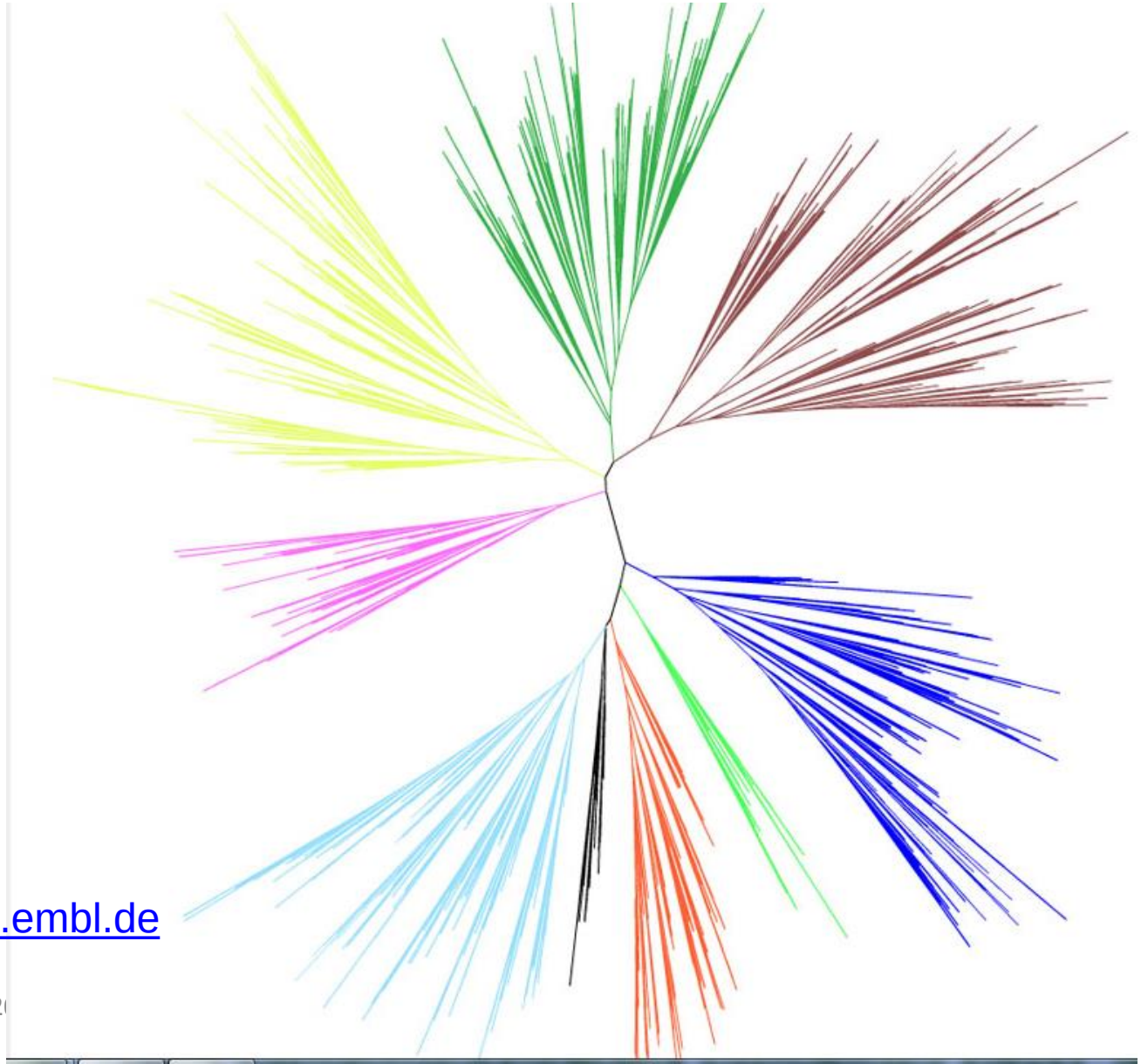
B



Credit: Wattam et al. 2011

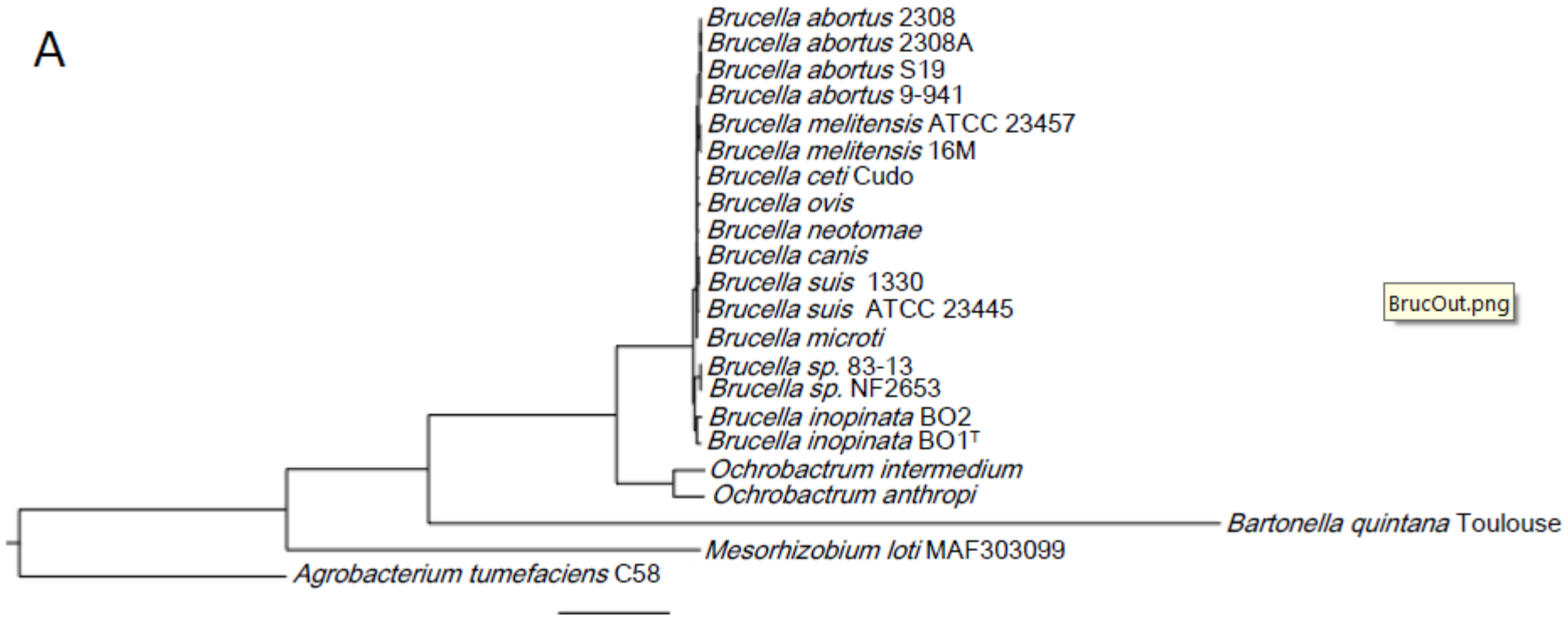
Comprimento dos ramos: número de substituições por coluna

Árvore sem raiz



http://itol.embl.de

Árvore com raiz: precisa de um grupo externo



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

Distância e similaridade

- São conceitos muito parecidos
- Em particular *distância de edição*
- Como transformar sequência *s* em sequência *t*
- Operações
  - Substituição do caracter *a* por *b* (custo = 1)
  - Inserção ou Remoção de um caracter (custo = 2)
- O algoritmo de PD já visto resolve esse problema

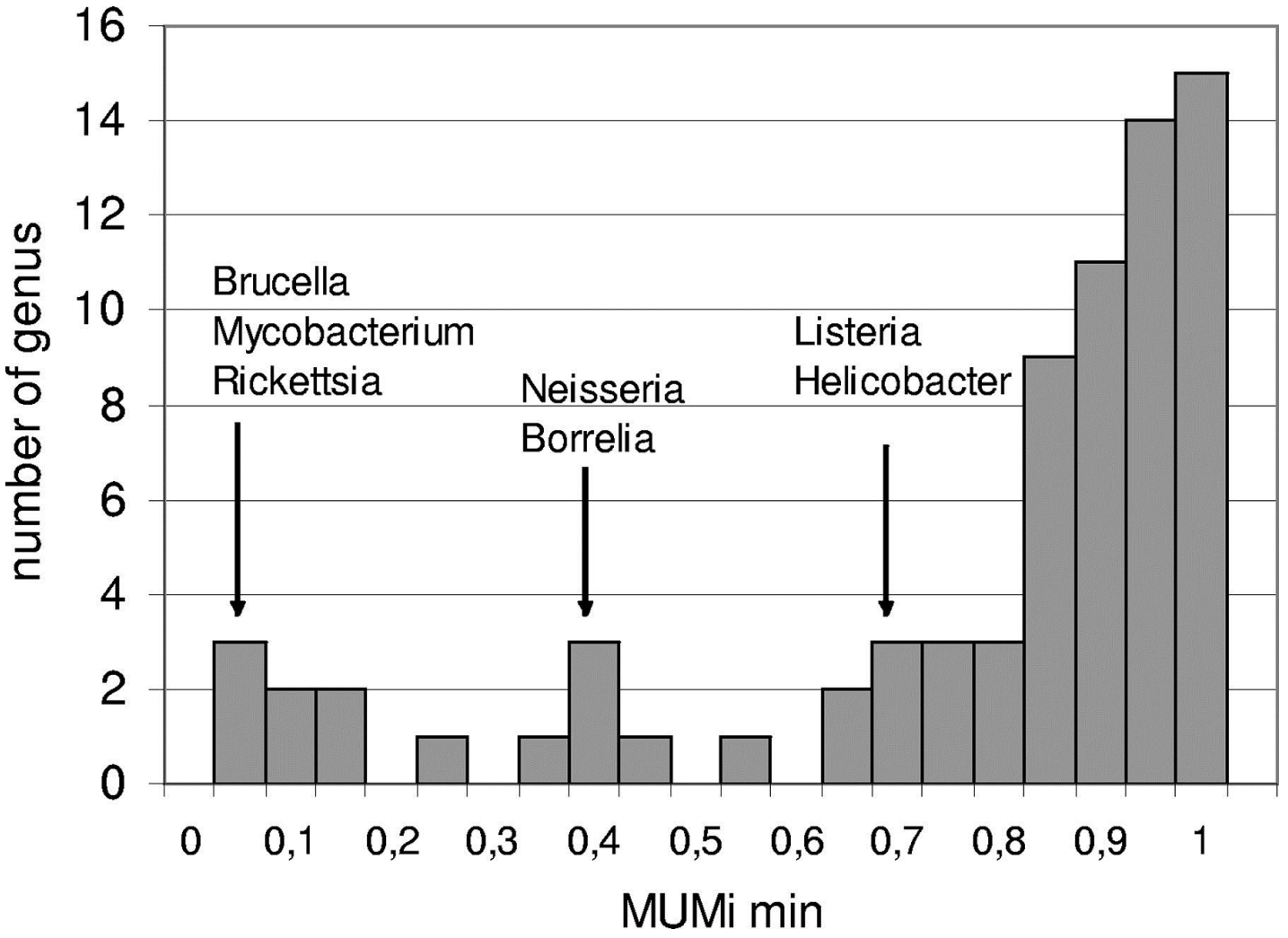
Uma fórmula de distância genômica

- MUMi = MUM index
- Baseado em MUMmer
- Deloger et al. 2009
- MUMi = $1 - L_{mum}/L_{av}$
- L_{mum} = soma dos comprimentos de todos os MUMs que não tem sobreposição
- L_{av} = comprimento médio dos 2 genomas sendo comparados
- Para obter MUMi, basta rodar MUMmer com um script perl desses autores

Distâncias MUMi e taxonomia

- Ilustra bem a diferença entre filogenia e taxonomia
- Qual é a distância que separa espécies de gêneros?

Distribution of all minimal MUMi values per genus.



Marc Deloger et al. J. Bacteriol. 2009;191:91-99

Journal of Bacteriology

Conclusão

- Não dá para comparar distâncias MUMi entre diferentes gêneros

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	*B. suis*	*B. microti*
B. neotomae 5K33	0.145	0.168	0.146	0.168	0.017	0.022
Brucella sp. strain 83–13		0.172	0.009	0.175	0.147	0.150
Brucella sp. strain BO2			0.168	0.107	0.169	0.169
Brucella sp. strain NF 2653				0.172	0.147	0.146
B. inopinata BO1					0.169	0.167
B. suis biovar 3 686						0.020

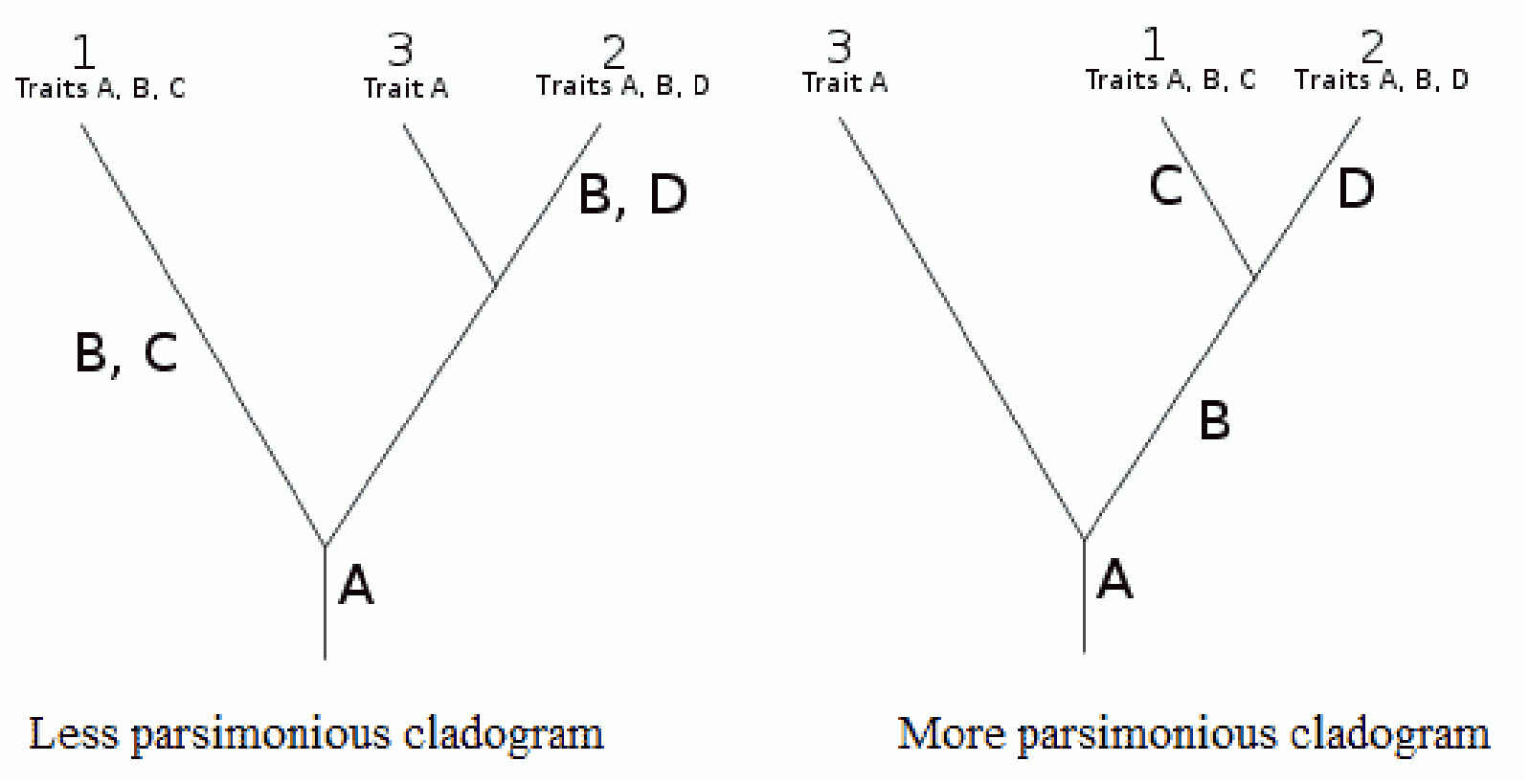
[a] Only the sequence of *B. microti* was complete, and its two chromosome sequences were concatenated. Concatenation was done by inserting a string with 100 nucleotides between contigs.

[b] The minimum MUMi value is 0, and the maximum MUMi value is 1.

Árvore a partir da matriz

- Métodos
  - UPGMA
  - Neighbor-joining (NJ)

Parcimônia



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 (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

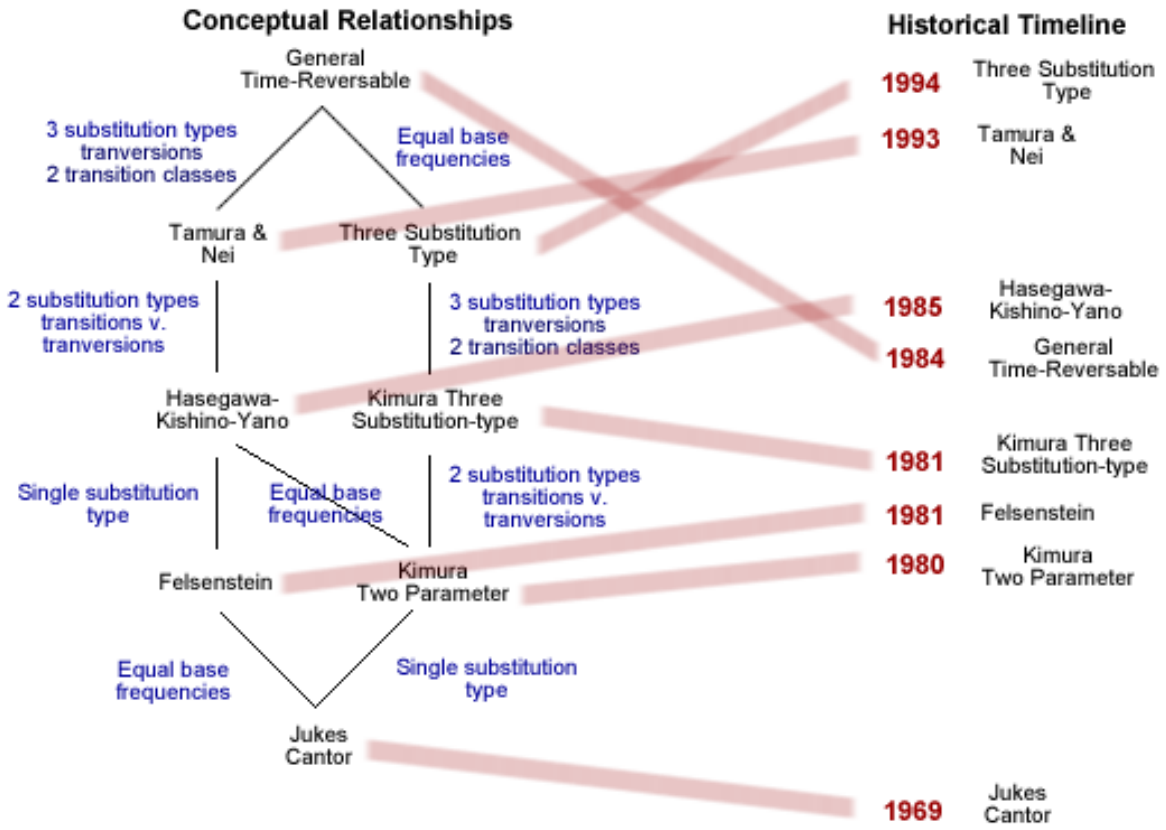
Considerações de tempo de execução

- Até 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

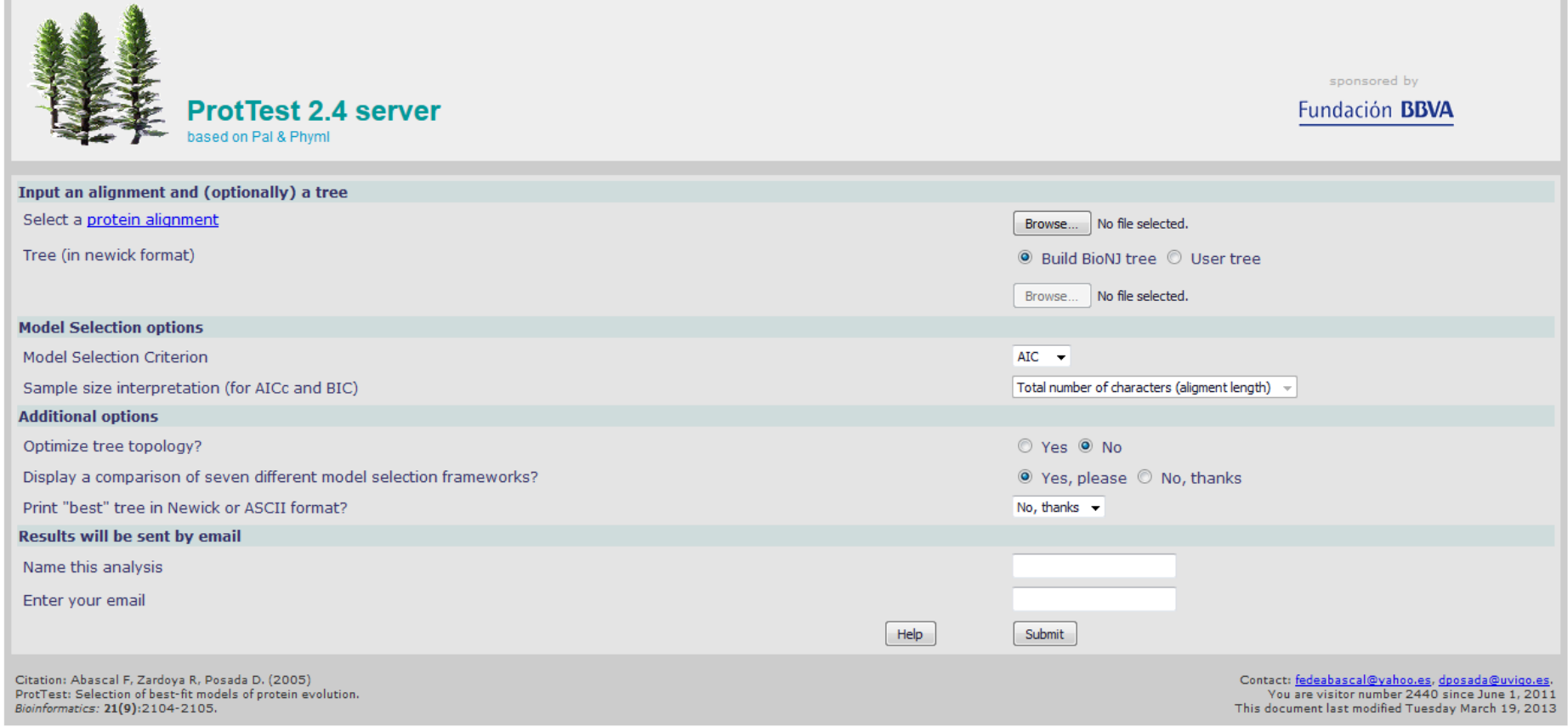
Modelos em PhyML

- **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
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

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

ResearchGate

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, phyml, 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 pariwise 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 usefult 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.

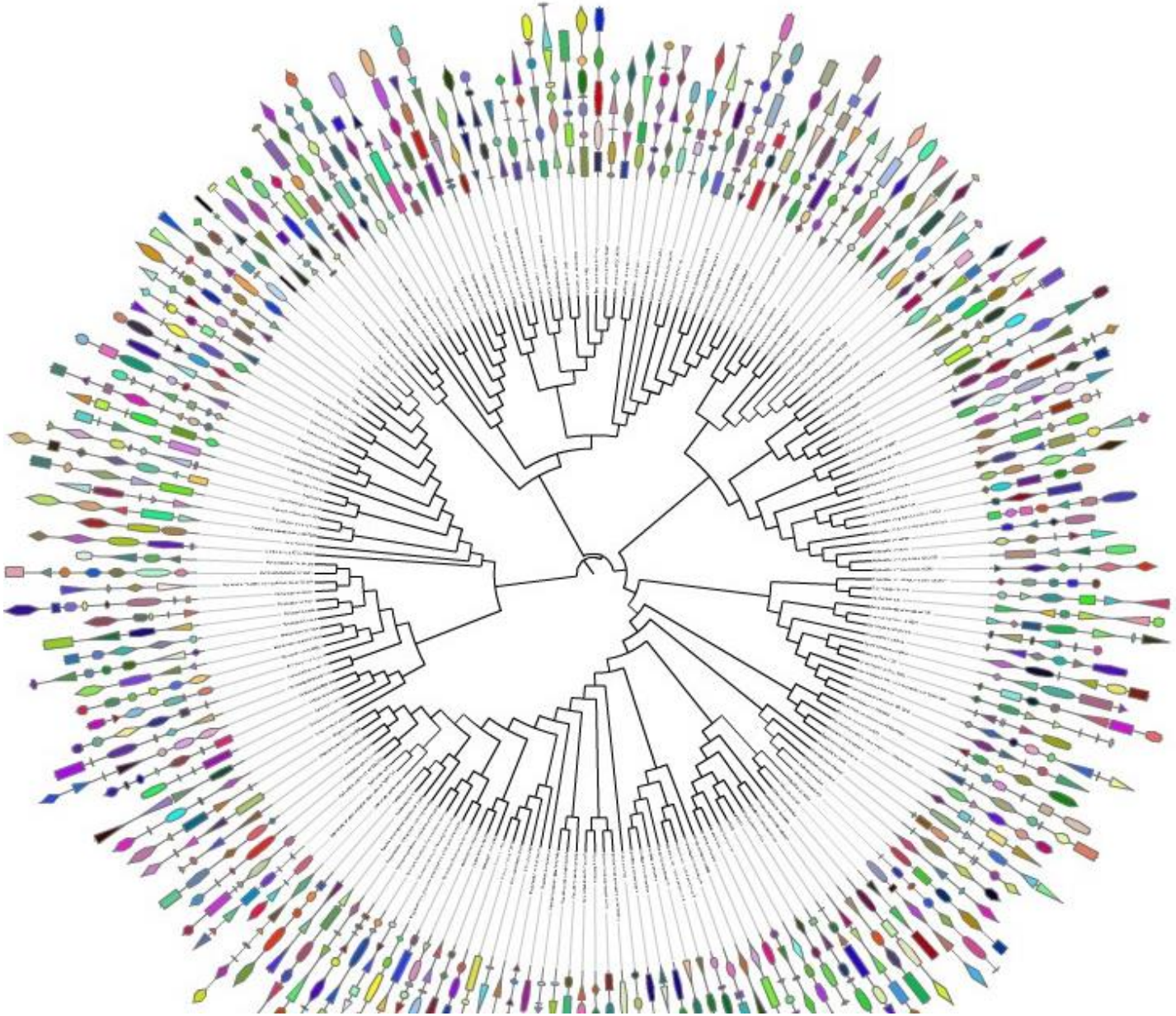
Visualização de árvores: formatos

- **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**

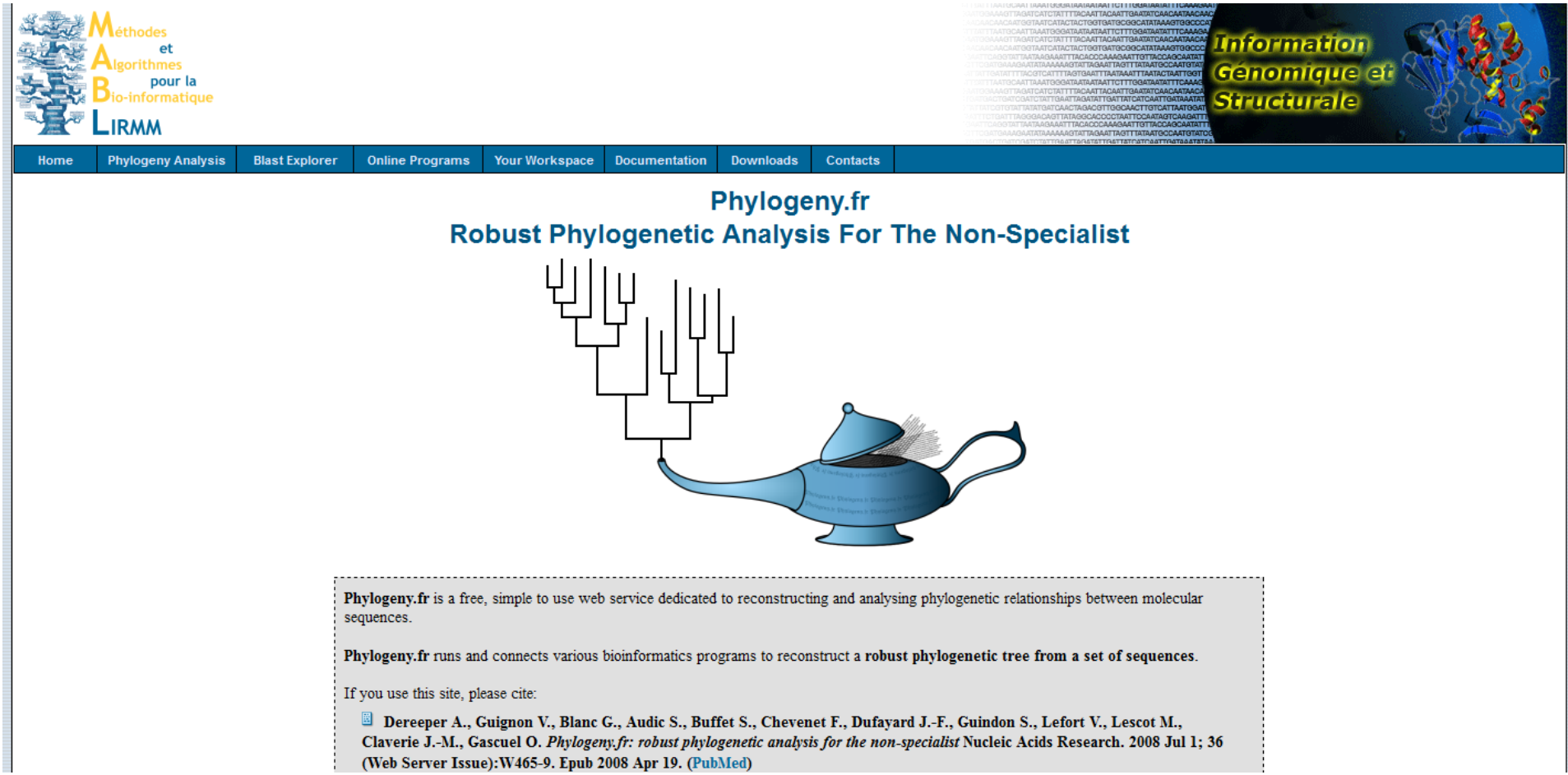
Visualização de árvores

- 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 (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:	Phylogeny:	Tree viewers:	Utilities:
MUSCLE	PhyML	TreeDyn	Gblocks
T-Coffee / 3D-Coffee	TNT	Drawgram	Jalview
ClustalW	BioNJ	Drawtree	Readseq
ProbCons	MrBayes	ATV	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

Marseille - Nice genopole®

cnrs Centre National de la Recherche Scientifique
dépasser les frontières

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

Mac Universal **MacOS X**	**Linux on PC** **64-bit Linux on x86**
MS Windows **self-extractible archive**	**Solaris on SPARC**

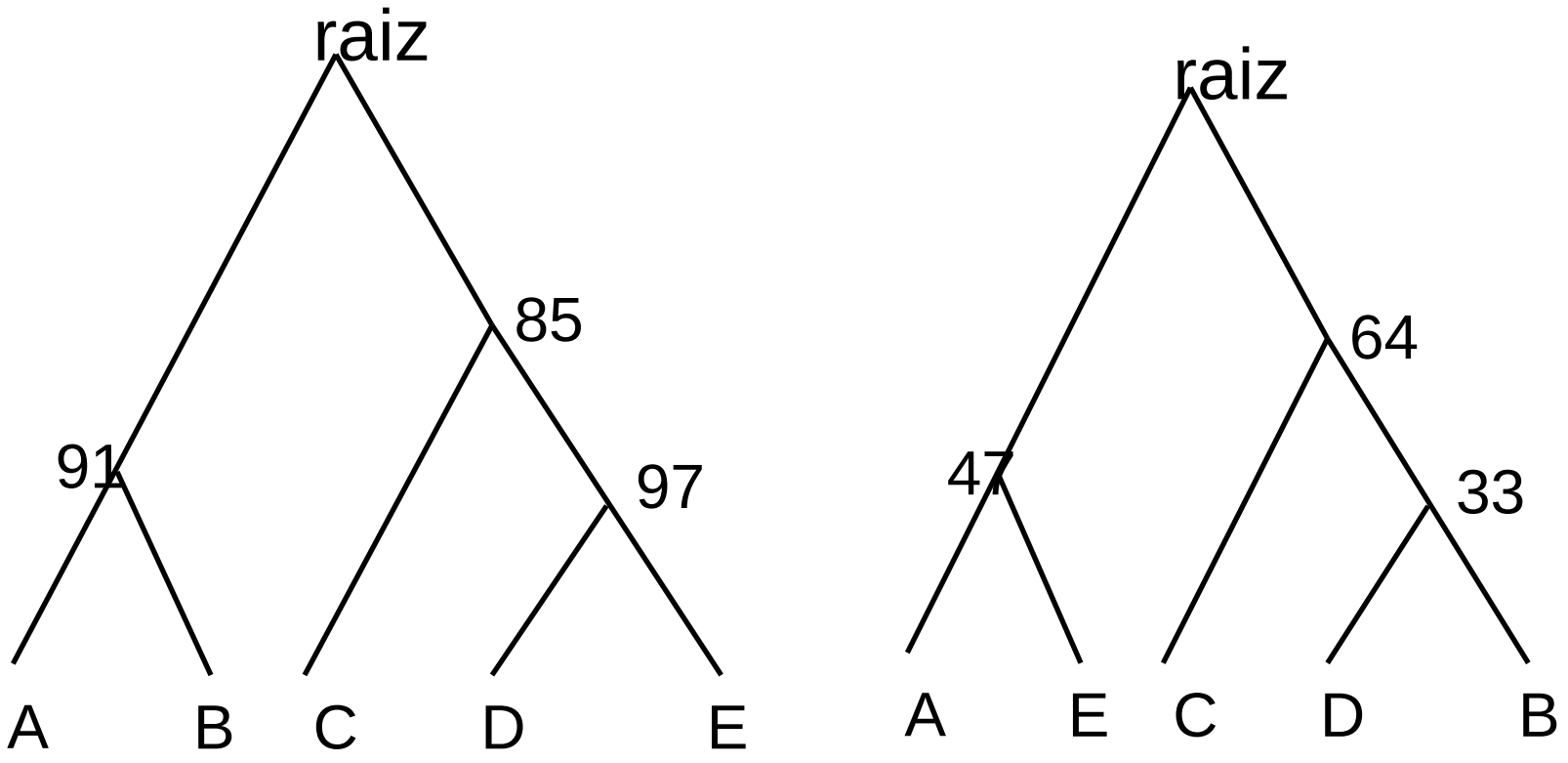
Interpretação

- 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

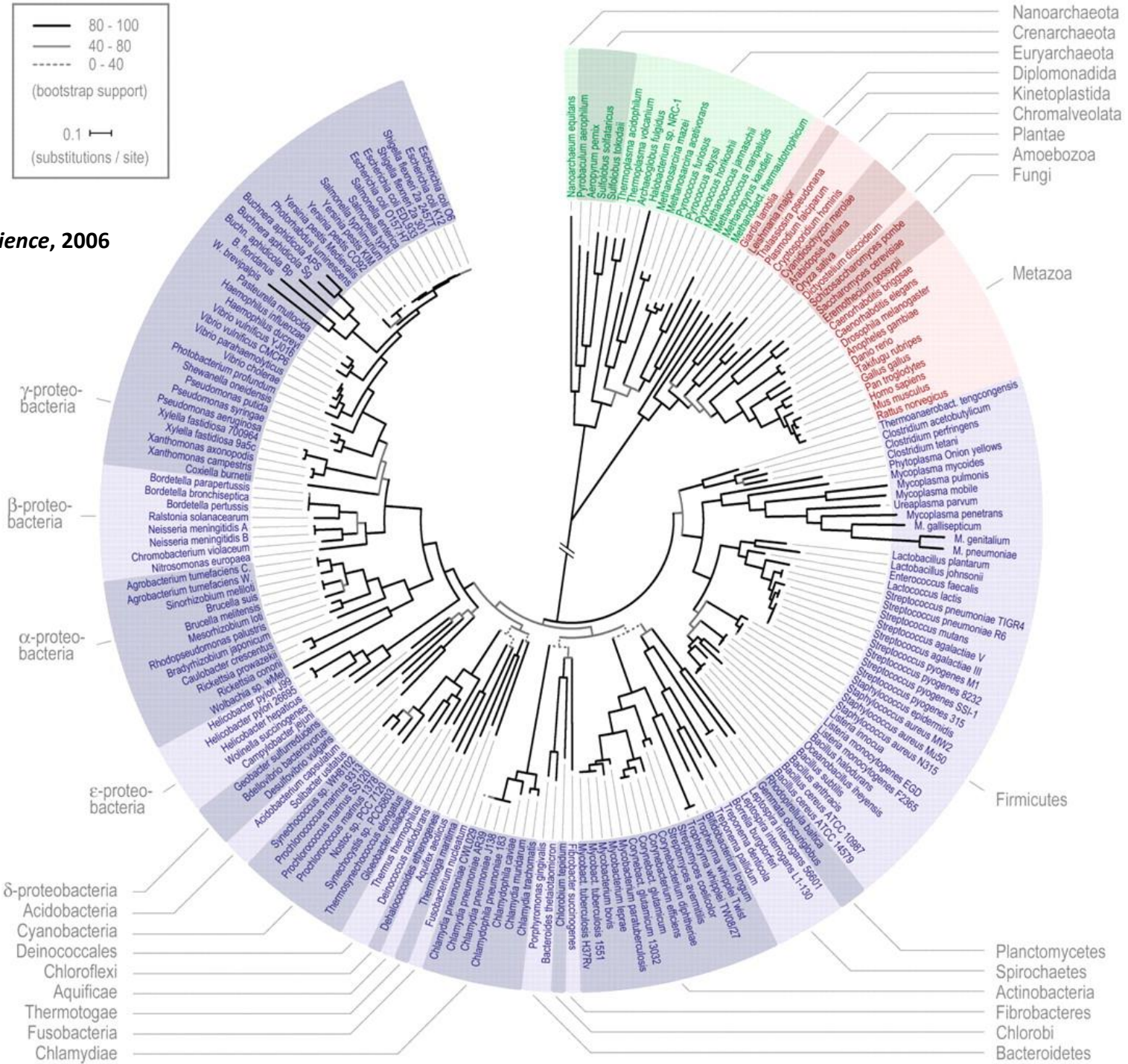
- Valores de bootstrap
- Colunas do AM são amostradas aleatoriamente em várias corridas (replicatas; geralmente entre 100 e 1000)
- Árvores resultantes são comparadas entre si
- Concordâncias nos clados são calculadas, resultando em número de vezes (ou %) que clados se repetem nas replicatas
- Valores 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!



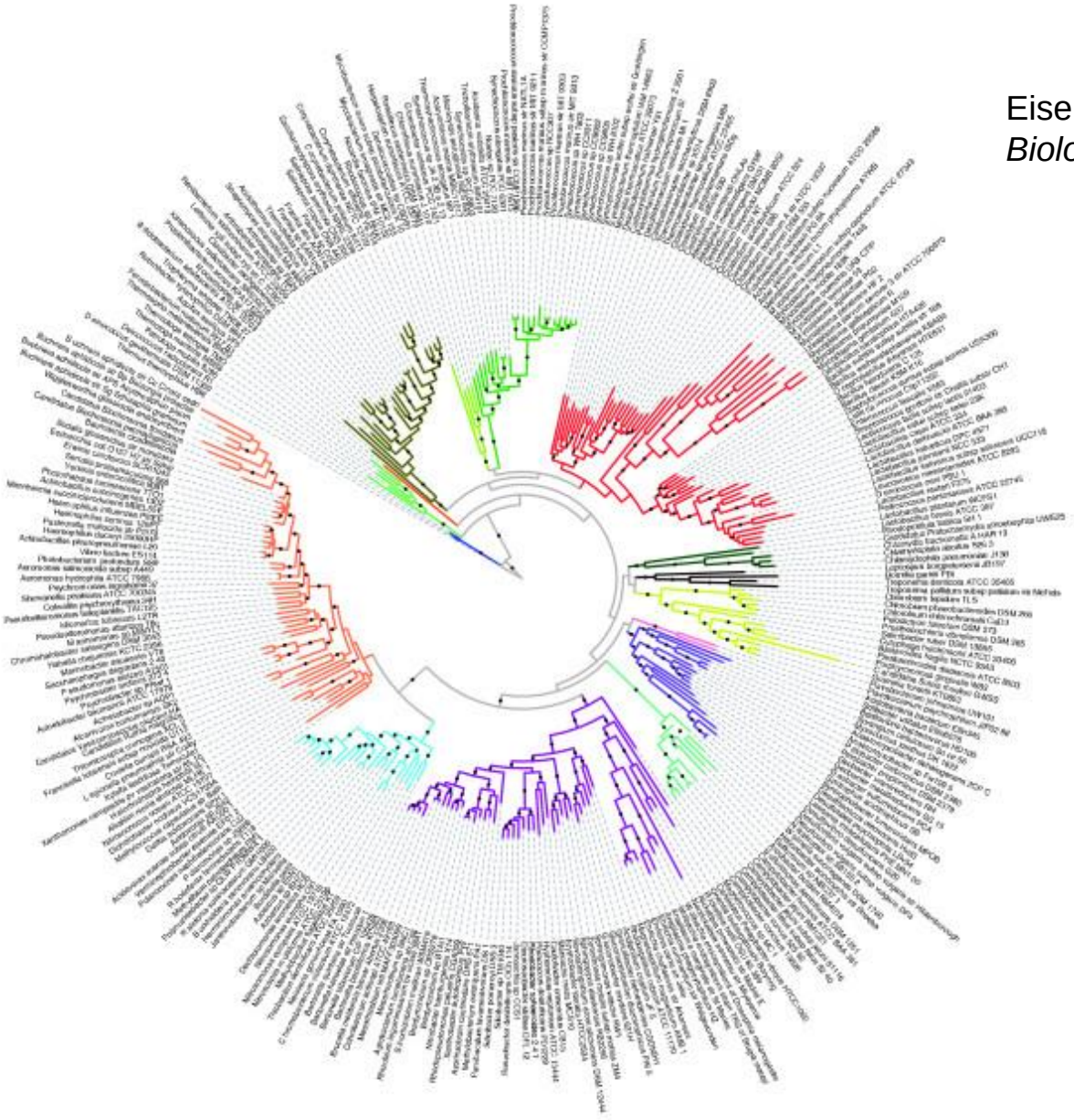
Ciccarelli et al, *Science*, 2006

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

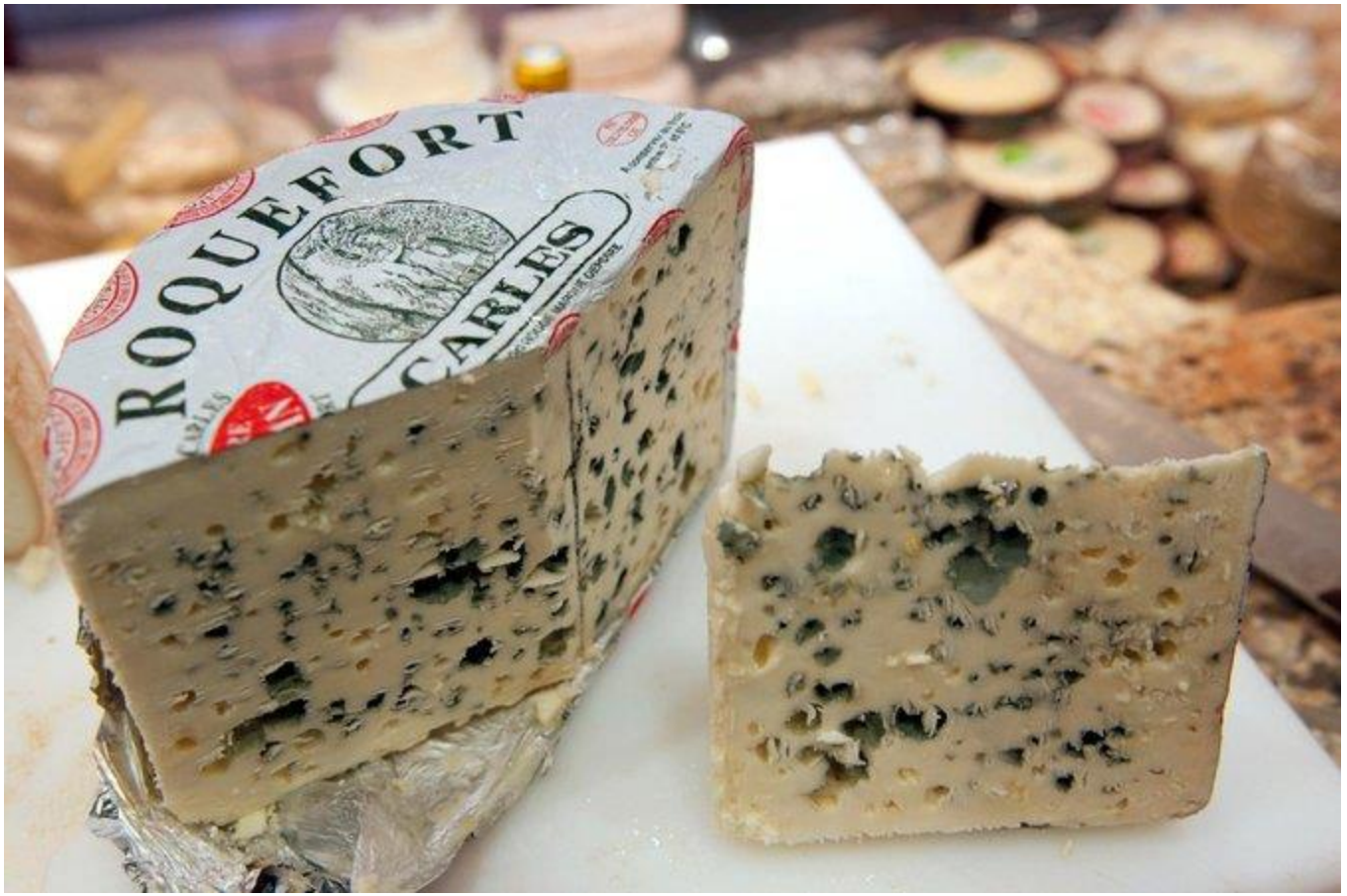


Eisen & Wu, *Genome Biology*, 2008

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



Fonte: NY Times, 24-set-2015

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*

Please cite this article in press as: Ropars et al., Adaptive Horizontal Gene Transfers between Multiple Cheese-Associated Fungi, Current Biology (2015), http://dx.doi.org/10.1016/j.cub.2015.08.025

Current Biology
Report

CellPress

Adaptive Horizontal Gene Transfers between Multiple Cheese-Associated Fungi

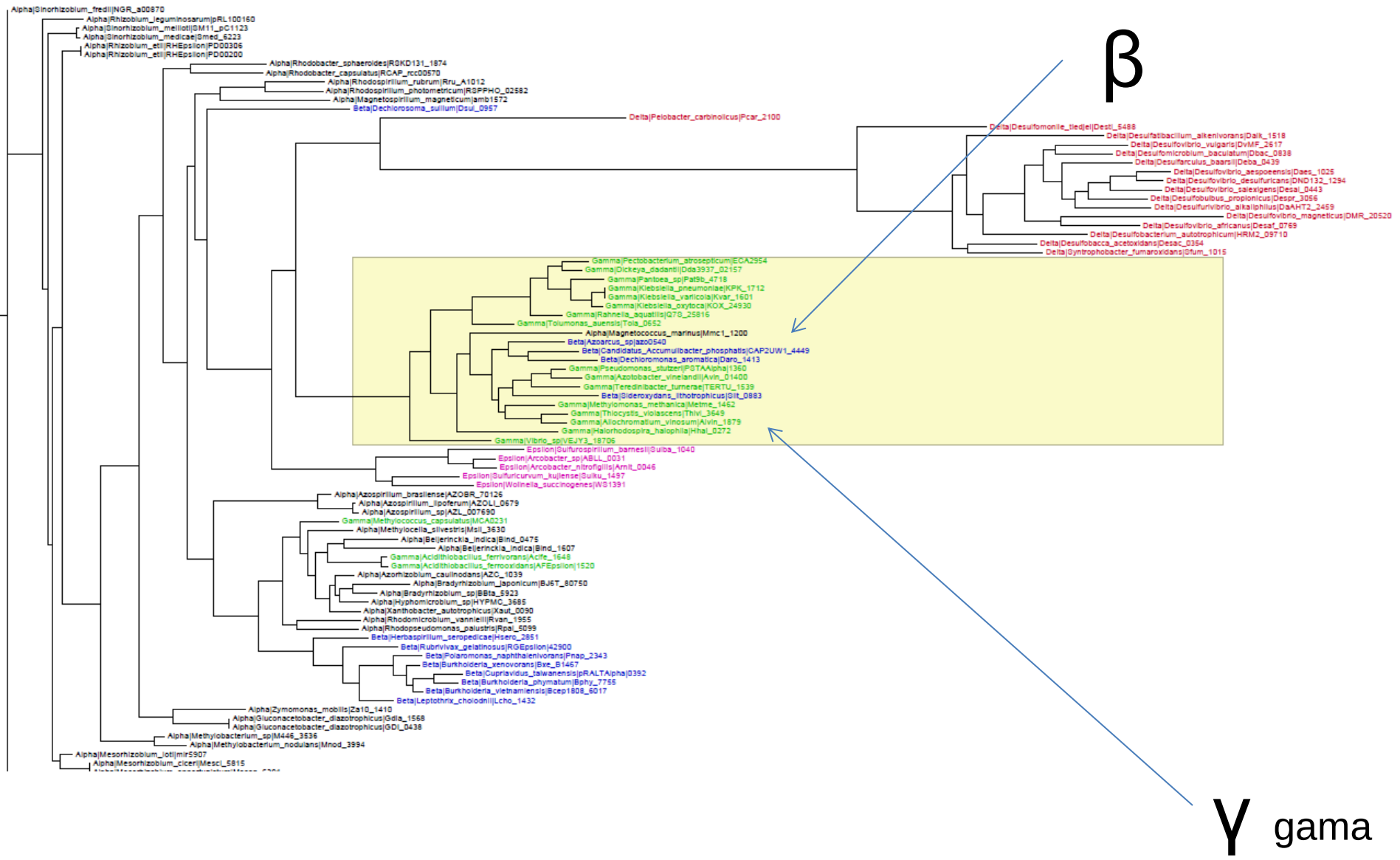
Jeanne Ropars,[1,2,8] Ricardo C. Rodríguez de la Vega,[1,2,8] Manuela López-Villavicencio,[3] Jérôme Gouzy,[4,5] Erika Sallet,[4,5] Émilie Dumas,[1,2] Sandrine Lacoste,[3] Robert Debuchy,[6,7] Joëlle Dupont,[3] Antoine Branca,[1,2,9,*] and Tatiana Giraud[1,2,9,*]
[1]Ecologie, Systématique et Evolution, UMR8079, Univ. Paris-Sud, 91405 Orsay, France
[2]Ecologie, Systématique et Evolution, UMR8079, CNRS, 91405 Orsay, France
[3]Institut 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
[4]Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR441, INRA, Castanet-Tolosan 31326, France
[5]Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR2594, CNRS, Castanet-Tolosan 31326, France
[6]Institut de Génétique et Microbiologie, UMR8621, Univ. Paris-Sud, 91405 Orsay, France
[7]Institut de Génétique et Microbiologie, UMR8621, CNRS, 91405 Orsay, France
[8]Co-first author
[9]Co-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

Transferência Horizontal de Genes

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

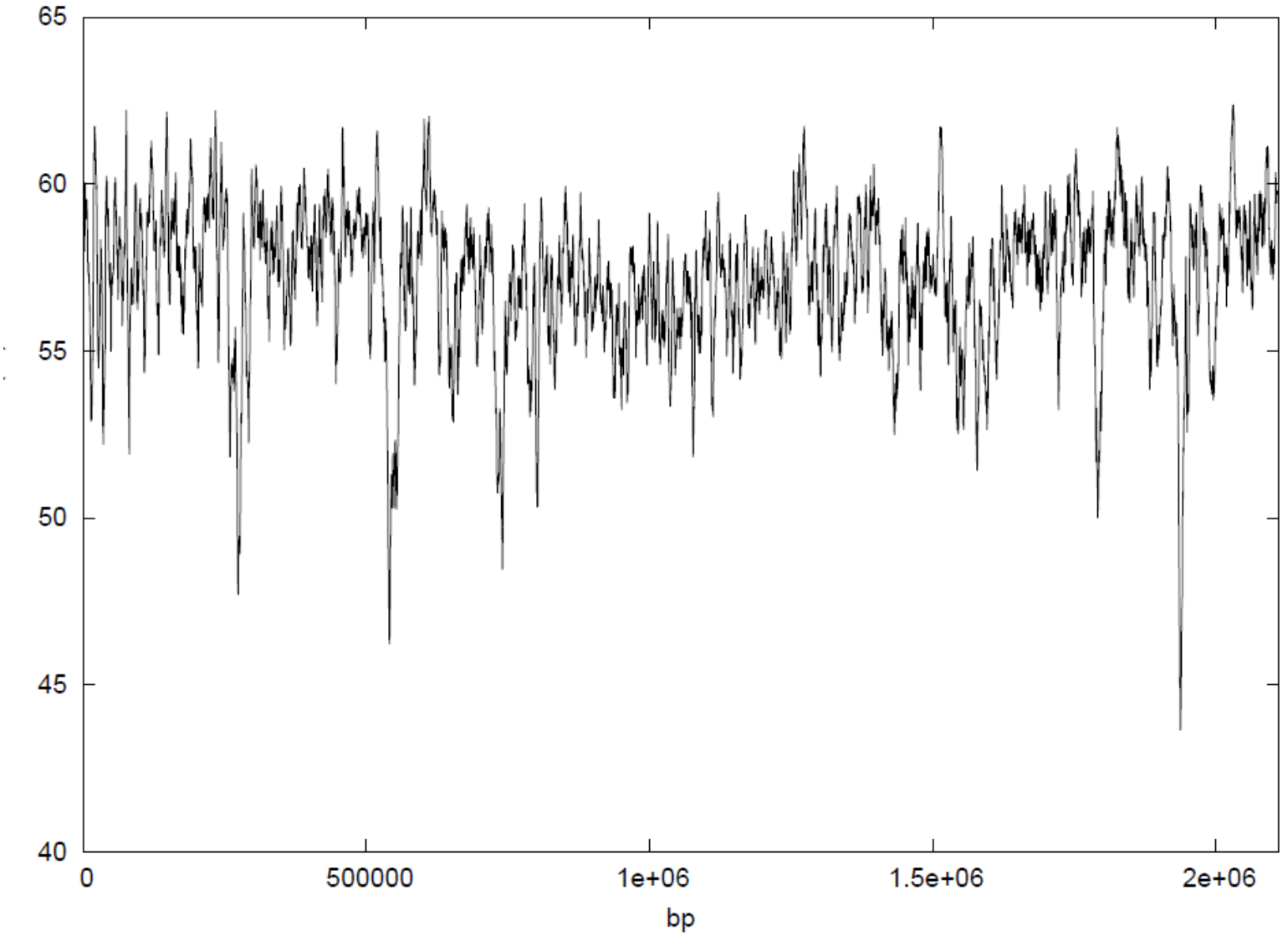
THG antiga

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

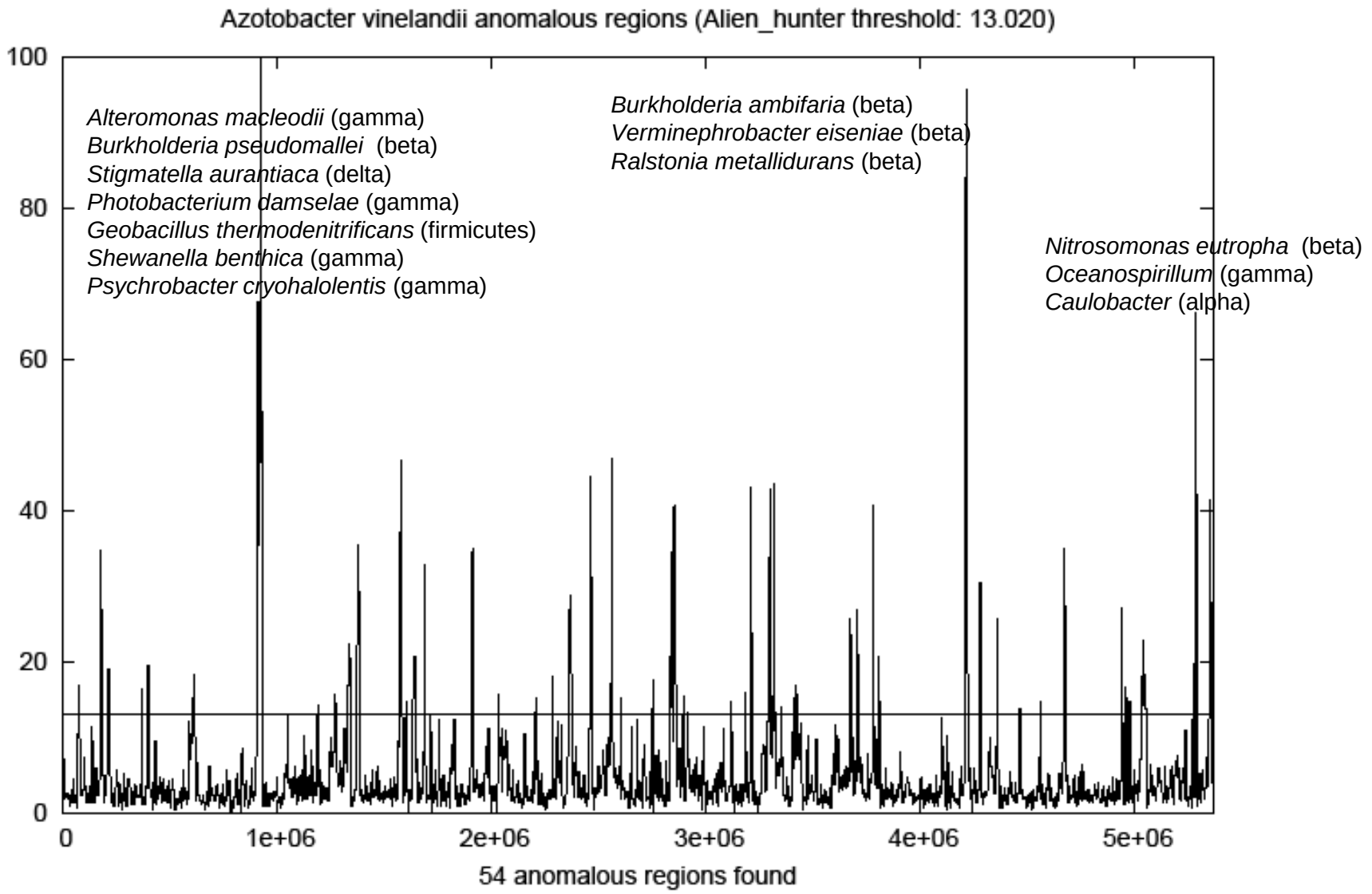

β
γ gama
Alpha|Sinorhizobium_fredii|NGR_a00870
Alpha|Rhizobium_leguminosarum|pRL100160
Alpha|Sinorhizobium_meliloti|SM11_pC1123
Alpha|Sinorhizobium_medicae|Smed_6223
Alpha|Rhizobium_etli|RHEpsilon|PD00306
Alpha|Rhizobium_etli|RHEpsilon|PD00200
Alpha|Rhodobacter_sphaeroides|RGKD131_1874
Alpha|Rhodobacter_capsulatus|RCAP_rcc00570
Alpha|Rhodospirillum_rubrum|Rru_A1012
Alpha|Rhodospirillum_photometricum|RGPPHO_02582
Alpha|Magnetospirillum_magneticum|amb1572
Beta|Dechlorosoma_suillum|Dsui_0957
Delta|Pelobacter_carbinolicus|Pcar_2100
Delta|Desulfomonile_tiedjei|Desti_5488
Delta|Desulfatibacillum_alkenivorans|Dalk_1518
Delta|Desulfovibrio_vulgaris|DvMF_2617
Delta|Desulfomicrobium_baculatum|Dbac_0838
Delta|Desulfarculus_baarsii|Deba_0439
Delta|Desulfovibrio_aespoeensis|Daes_1025
Delta|Desulfovibrio_desulfuricans|DND132_1294
Delta|Desulfovibrio_salexigens|Desal_0443
Delta|Desulfobulbus_propionicus|Despr_3056
Delta|Desulfurivibrio_alkaliphilus|DaAHT2_2459
Delta|Desulfovibrio_magneticus|DMR_20520
Delta|Desulfovibrio_africanus|Desaf_0769
Delta|Desulfotobacterium_autotrophicum|HRM2_09710
Delta|Desulfobacca_acetoxidans|Desac_0354
Delta|Syntrophobacter_fumaroxidans|Sfum_1015
Gamma|Pectobacterium_atrosepticum|ECA2954
Gamma|Dickeya_dadantii|Dda3937_02157
Gamma|Pantoea_sp|Pat9b_4718
Gamma|Klebsiella_pneumoniae|KPK_1712
Gamma|Klebsiella_variicola|Kvar_1601
Gamma|Klebsiella_oxytoca|KOX_24930
Gamma|Rahnella_aquatilis|Q79_25816
Gamma|Tolumonas_auensis|Tola_0652
Alpha|Magnetococcus_marinus|Mmc1_1200
Beta|Azoarcus_sp|azo0540
Beta|Candidatus_Accumulibacter_phosphatis|CAP2UW1_4449
Beta|Dechloromonas_aromatica|Daro_1413
Gamma|Pseudomonas_stutzeri|PSTAAlpha|1360
Gamma|Azotobacter_vinelandii|Avin_01400
Gamma|Teredinibacter_turnerae|TERTU_1539
Beta|Sideroxydans_lithotrophicus|Slit_0883
Gamma|Methylomonas_methanica|Metme_1462
Gamma|Thiocystis_violascens|Thivi_3649
Gamma|Allochromatium_vinosum|Alvin_1879
Gamma|Halorhodospira_halophila|Hhal_0272
Gamma|Vibrio_sp|VEJY3_18706
Epsilon|Sulfurospirillum_barnesii|Sulba_1040
Epsilon|Arcobacter_sp|ABLL_0031
Epsilon|Arcobacter_nitrofigilis|Arnit_0046
Epsilon|Sulfuricurvum_kujiense|Suku_1457
Epsilon|Wolinella_succinogenes|WS1391
Alpha|Azospirillum_brasilense|AZOBR_70126
Alpha|Azospirillum_lipoferum|AZOLI_0679
Alpha|Azospirillum_sp|AZL_007690
Gamma|Methylococcus_capsulatus|MCA0231
Alpha|Methylocella_silvestris|Msil_3630
Alpha|Beijerinckia_indica|Bind_0475
Alpha|Beijerinckia_indica|Bind_1607
Gamma|Acidithiobacillus_ferrivorans|Acife_1648
Gamma|Acidithiobacillus_ferrooxidans|AFEpsilon|1520
Alpha|Azorhizobium_caulinodans|AZC_1039
Alpha|Bradyrhizobium_japonicum|BJ6T_80750
Alpha|Bradyrhizobium_sp|BBta_5923
Alpha|Hyphomicrobium_sp|HYPMC_3685
Alpha|Xanthobacter_autotrophicus|Xaut_0090
Alpha|Rhodomicrobium_vannielii|Rvan_1955
Alpha|Rhodopseudomonas_palustris|Rpal_5099
Beta|Herbaspirillum_seropedicae|Hsero_2851
Beta|Rubrivivax_gelatinosus|RGEpsilon|42900
Beta|Polaromonas_naphthalenivorans|Pnap_2343
Beta|Burkholderia_xenovorans|Bxe_B1467
Beta|Cupriavidus_taiwanensis|pRALTAlpha|0392
Beta|Burkholderia_phymatum|Bphy_7755
Beta|Burkholderia_vietnamiensis|Bcep1808_6017
Beta|Leptothrix_cholodnii|Lcho_1432
Alpha|Zymomonas_mobilis|Za10_1410
Alpha|Gluconacetobacter_diazotrophicus|Gdia_1568
Alpha|Gluconacetobacter_diazotrophicus|GDI_0438
Alpha|Methylobacterium_sp|M446_3536
Alpha|Methylobacterium_nodulans|Mnod_3994
Alpha|Mesorhizobium_loti|mlr5907
Alpha|Mesorhizobium_ciceri|Mesci_5815

THG recente

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

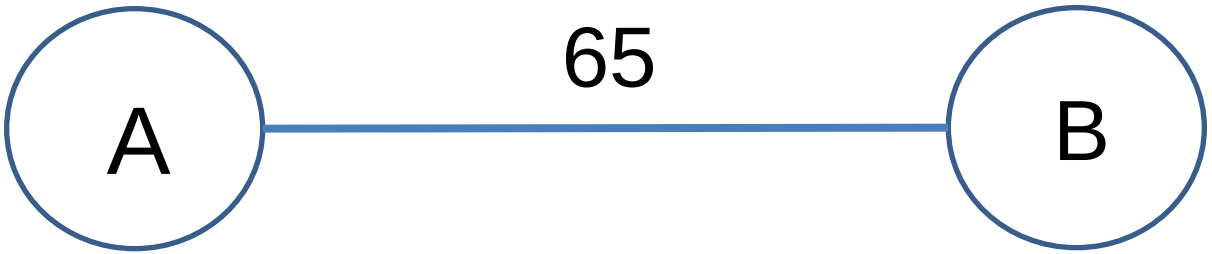


Variação do %GC no cromossomo principal de *Brucella ovis* ATCC25840


Azotobacter vinelandii anomalous regions (Alien_hunter threshold: 13.020)
100
80
60
40
20
0
0
1e+06
2e+06
3e+06
4e+06
5e+06
54 anomalous regions found
Alteromonas macleodii (gamma)
Burkholderia pseudomallei (beta)
Stigmatella aurantiaca (delta)
Photobacterium damselae (gamma)
Geobacillus thermodenitrificans (firmicutes)
Shewanella benthica (gamma)
Psychrobacter cryohalolentis (gamma)
Burkholderia ambifaria (beta)
Verminephrobacter eiseniae (beta)
Ralstonia metallidurans (beta)
Nitrosomonas eutropha (beta)
Oceanospirillum (gamma)
Caulobacter (alpha)

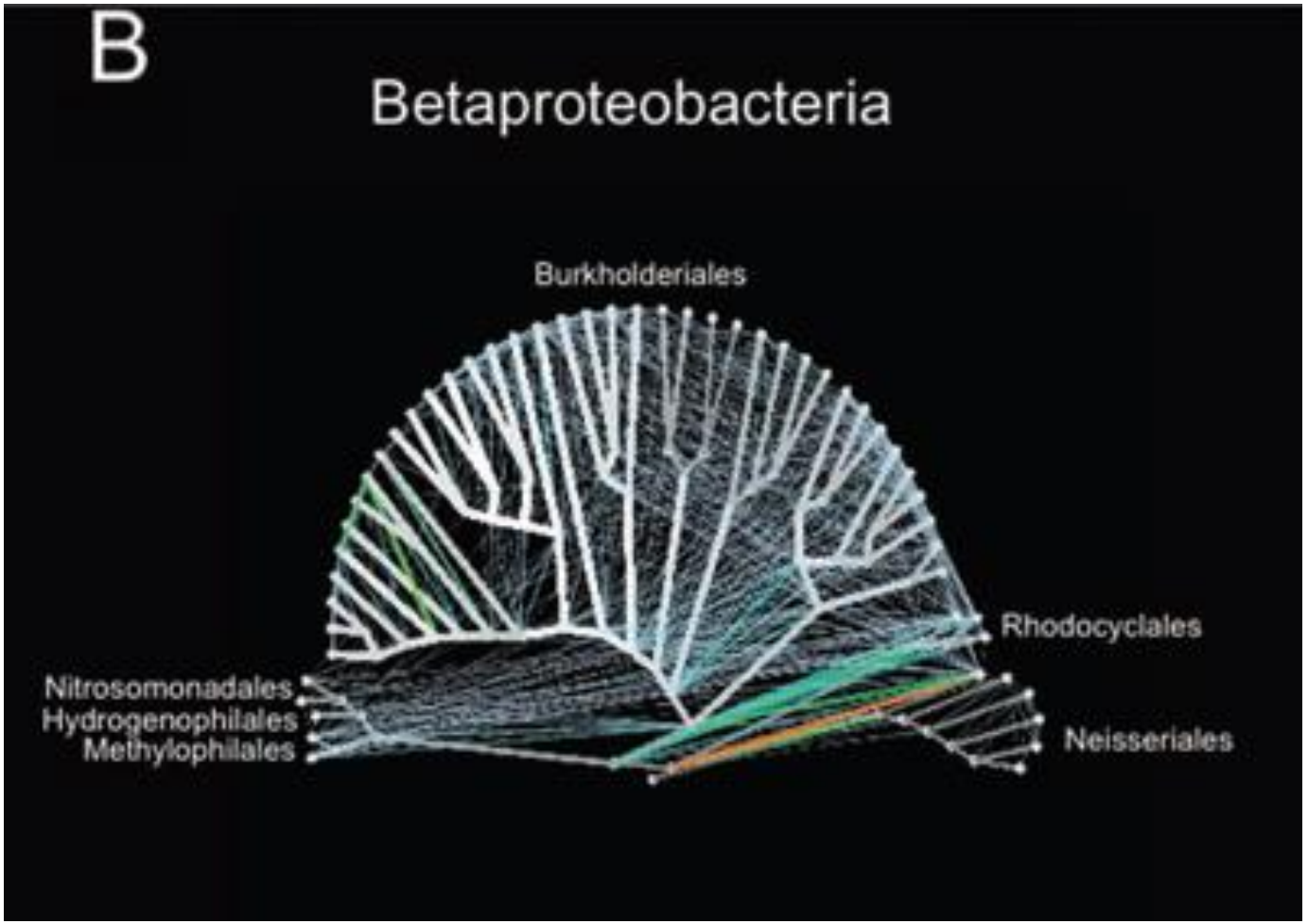
Redes filogenômicas

- Redes que mostram compartilhamento de genes

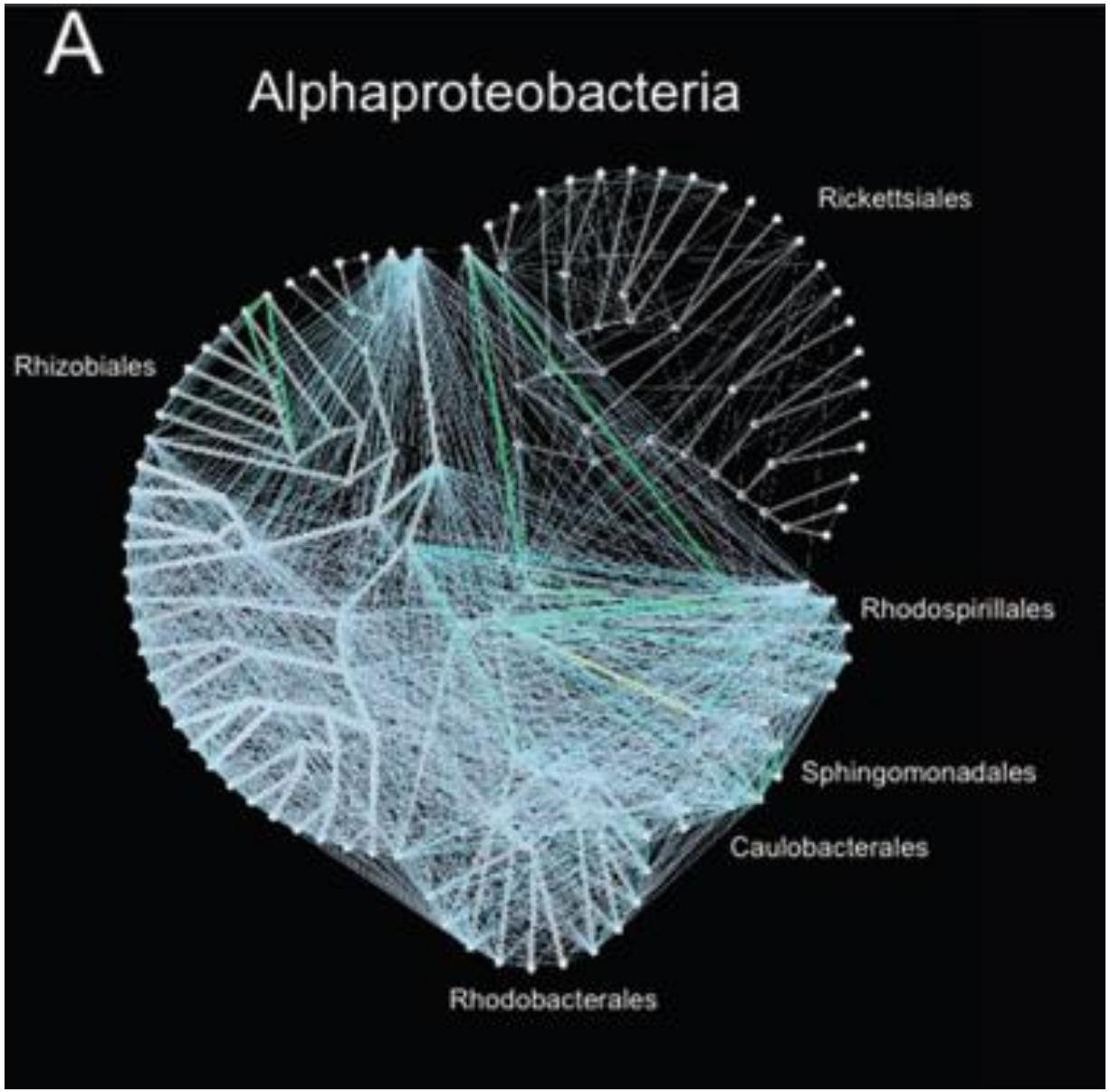


A superposição de uma árvore de espécies numa tal rede mostra possíveis eventos de transferência horizontal

Uma rede filogenômica

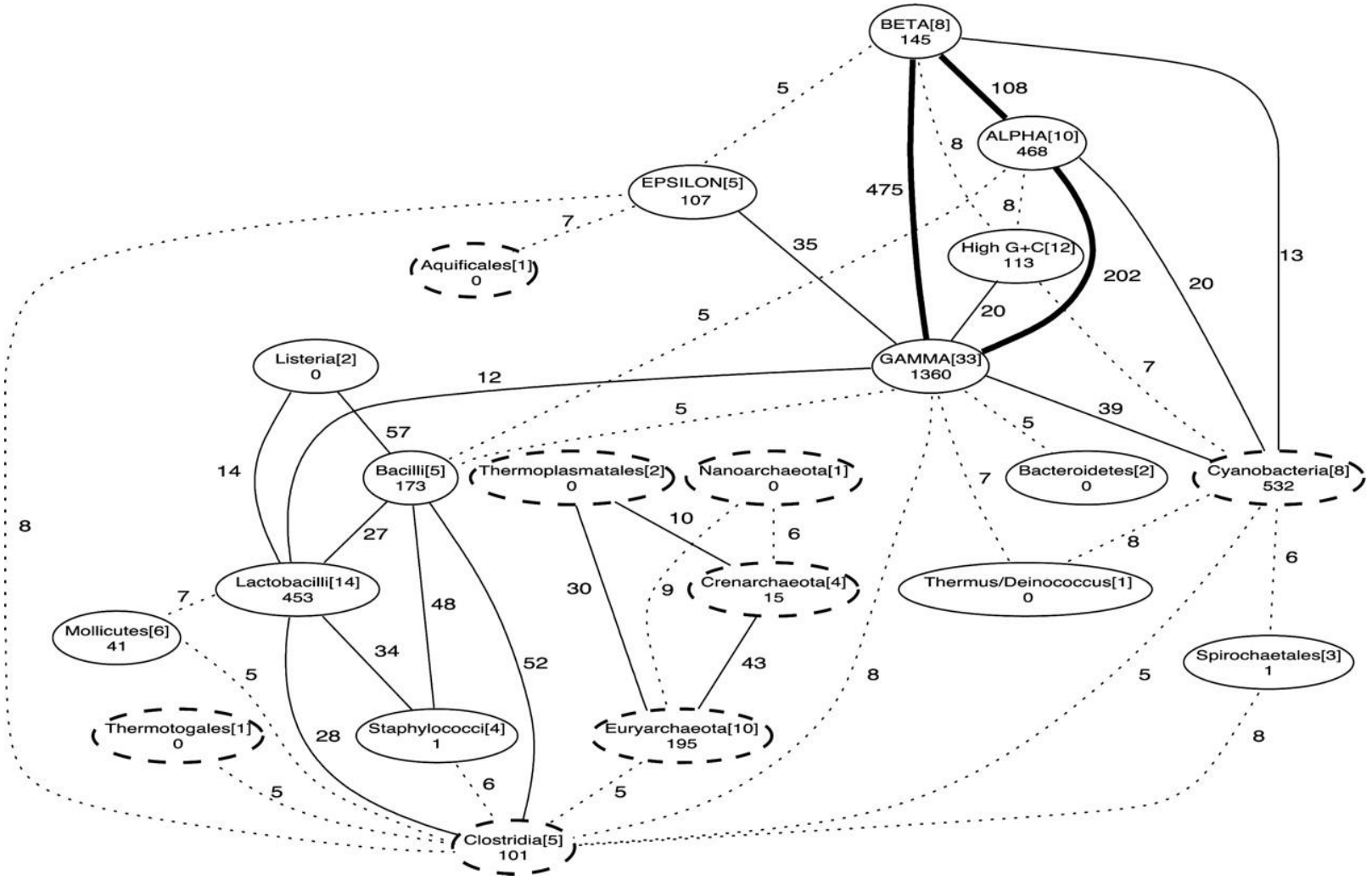


Kloesges et al, *Molecular Biology and Evolution*, 2011



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

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

NEWS FEATURE

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.

Citations, in which one paper refers to ear-

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 paper[2] 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 I really know it is not a

to other scientists what kind of wo
doing". Another common practice i
ensures that truly foundational d
ies — Einstein's special theory of r
for instance — get fewer citations t
might deserve: they are so important
quickly enter the textbooks or are incc
into the main text of papers as terms
so familiar that they do not need a cit

Citation counts are riddled with ot
founding factors. The volume of cita
increased, for example — yet older pa
had more time to accrue citations. E
tend to cite one another's work more fi
than, say, physicists. And not all fields
the same number of publications. Mo
liometricians therefore recoil from m
crude as simply counting citations w
want to measure a paper's value: inst
prefer to compare counts for papers c
age, and in comparable fields.

Nor is Thomson Reuters' list the o
ing system available. Google Scholar
its own top-100 list for *Nature*. It is
many more citations because the searc
culls references from a much greater (
poorly characterized) literature base, i
from a large range of books. In that l
able at www.nature.com/top100, ec
papers have more prominence. Goog
ar's list also features books, which T
Reuters did not analyse. But among th
papers, many of the same titles show

Yet even with all the caveats, the c
ioned hall of fame still has value. If
else, it serves as a reminder of the natu
entific knowledge. To make exciting a
researchers rely on relatively unsung
describe experimental methods, datal
software.

Here *Nature* tours some of the key
that tens of thousands of citations hav
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