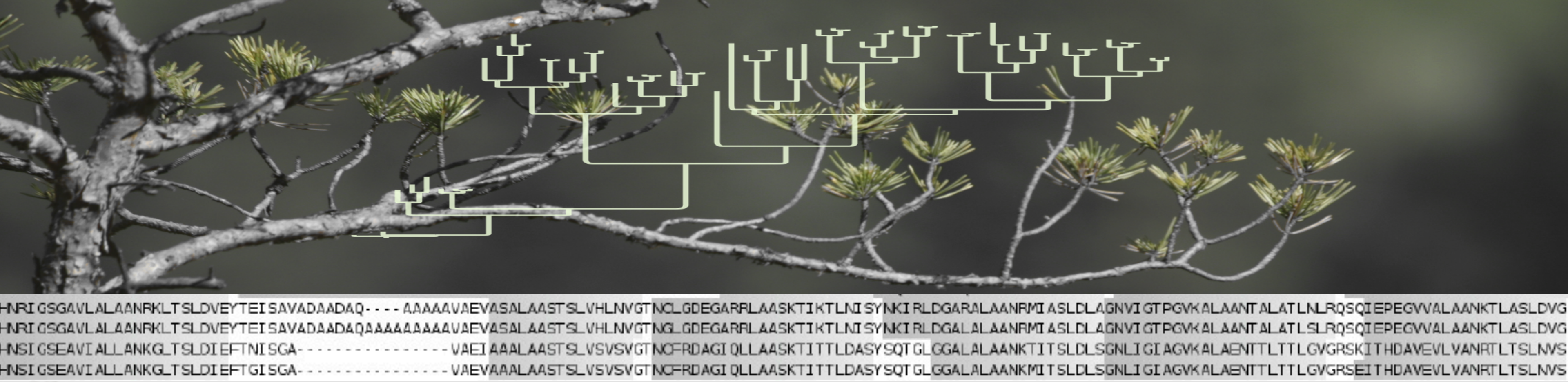


Inferring phylogenies

Maria Anisimova

Applied Computational Genomics Team – ACGT

Zürich University of Applied sciences

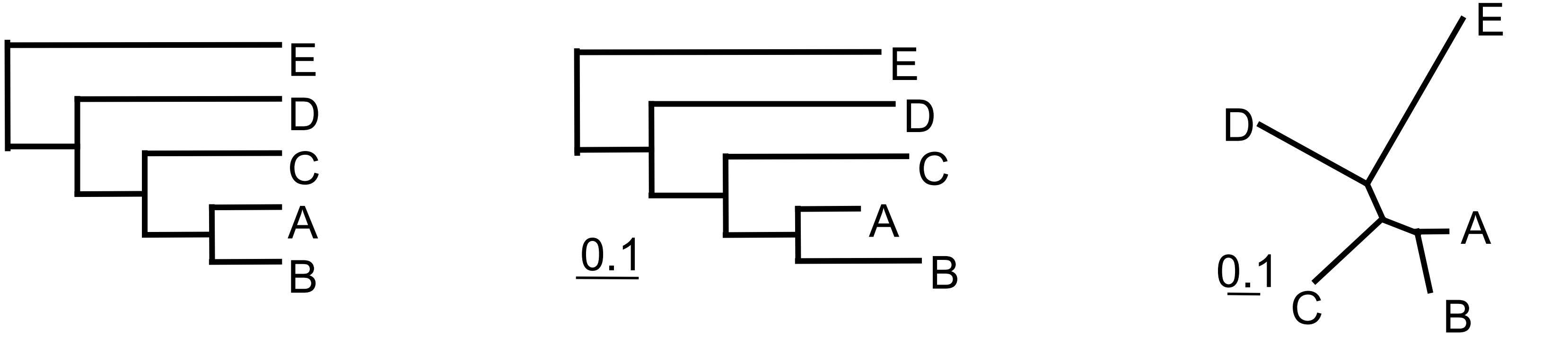
Swiss Institute of Bioinformatics

Zürcher Hochschule
für Angewandte Wissenschaften

SIB Swiss Institute of Bioinformatics

2017

Tree representations



(a) cladogram (b) phylogram (c) unrooted tree

(a) : ((((A,B),C),D),E)

(b) : ((((A: 0.1,B:0.2):0.12,C:0.3):0.123,D:0.4):0.1234,E:0.5)

(c) : (((A: 0.1,B:0.2):0.12,C:0.3):0.123,D:0.4,E:0.6234)

Visualization software:

TreeViewX, Forester ATV, FigTree, ITOL (itol.embl.de), Dendroscope

Figures from Yang (2006) Computational Molecular Evolution, OUP

Rearrangements that leave tree intact

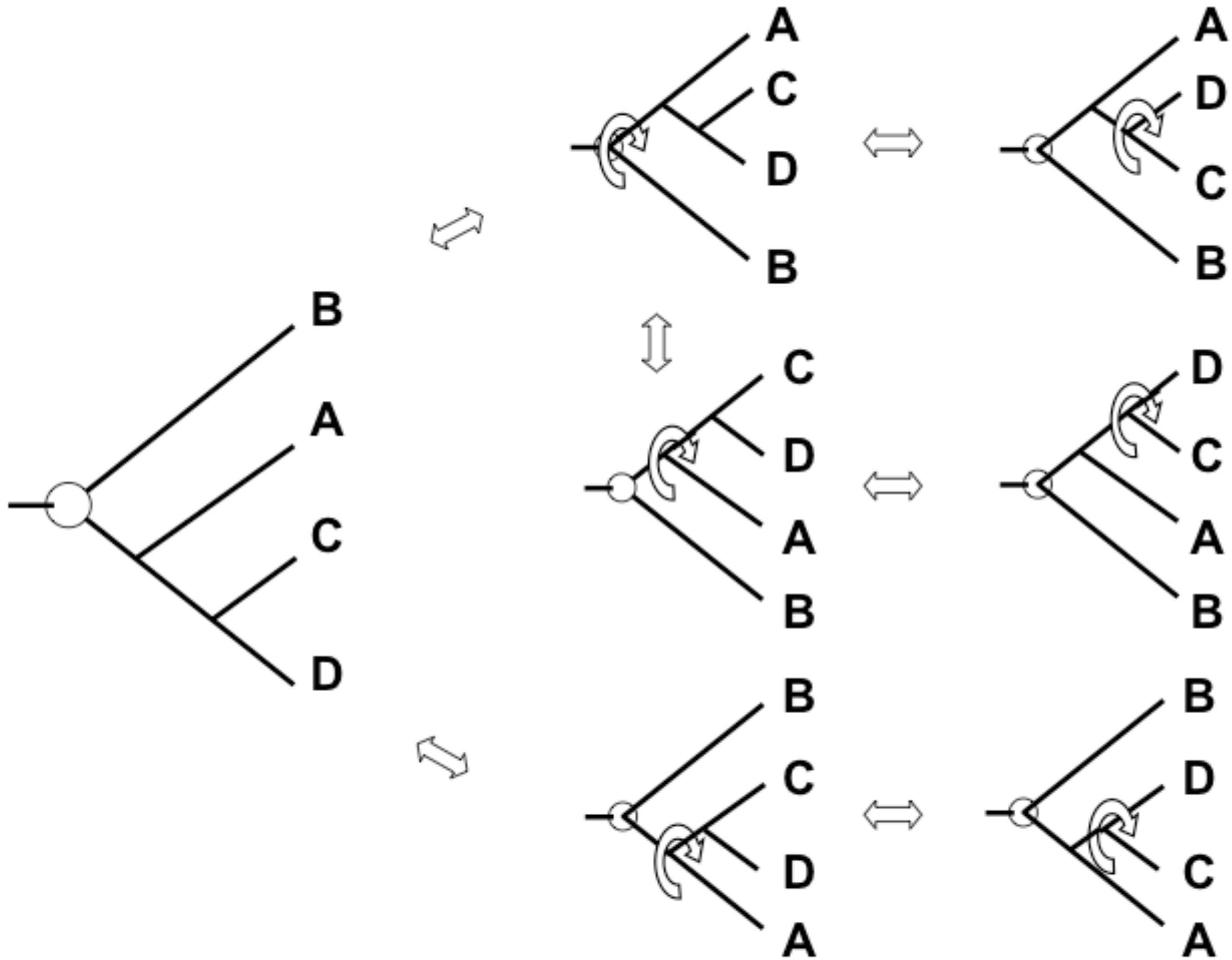
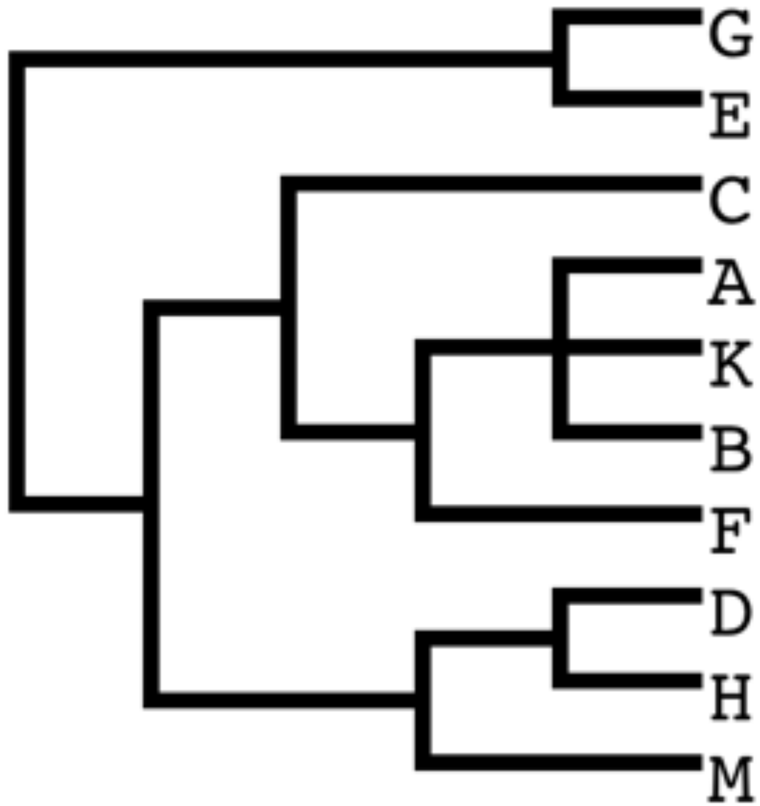


figure by Caro-Beth Stewart

Tree representations: exercise



Write down this tree as a NEWICK string

Spot the difference

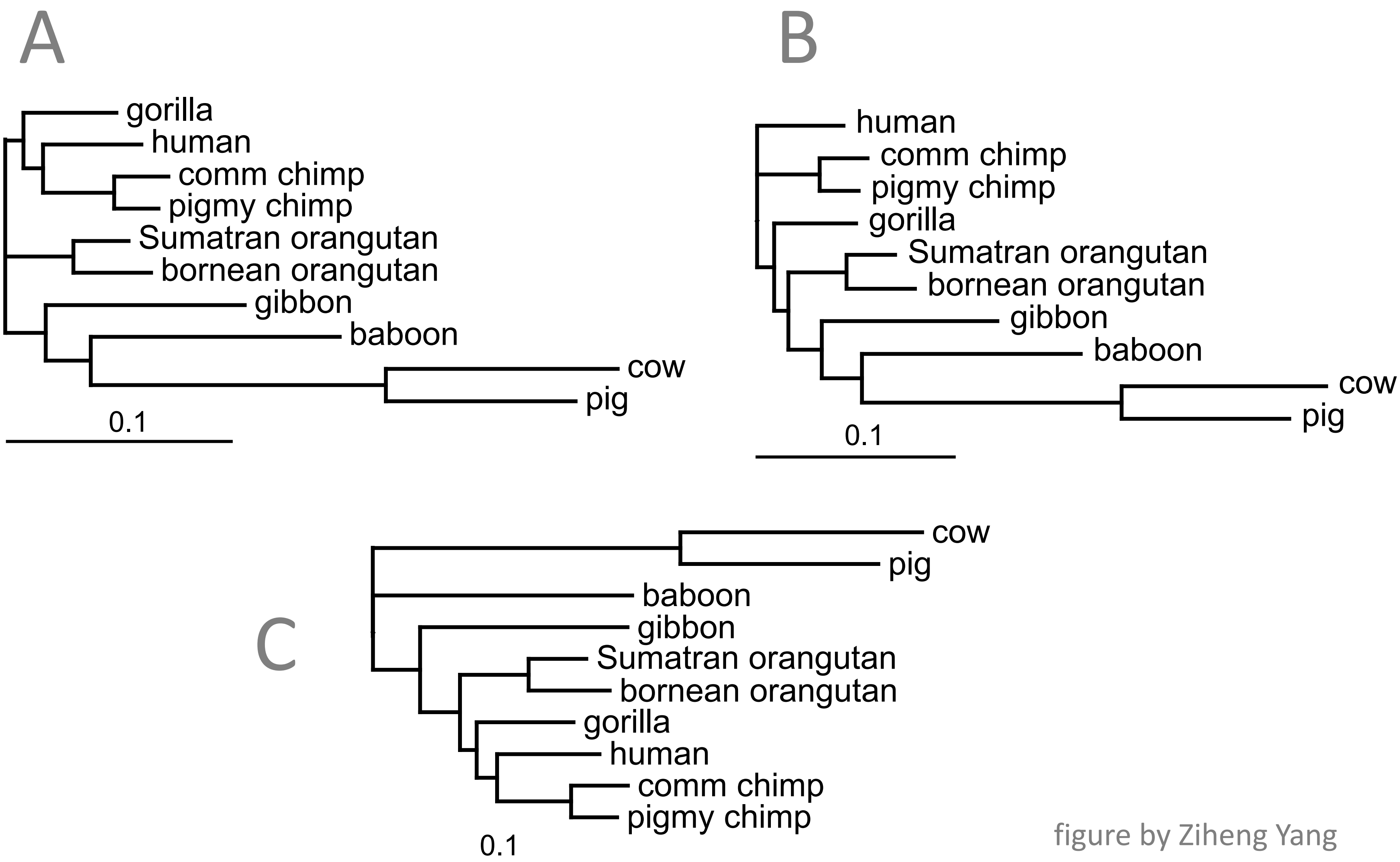
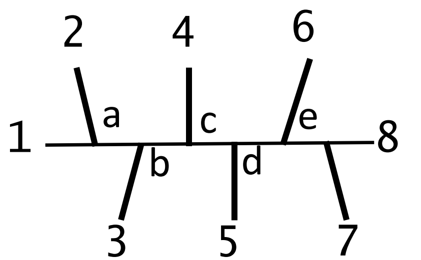


figure by Ziheng Yang

How different are two trees?

The partition distance is the total number of bipartitions that are in one tree but not in the other (Robinson & Foulds 1981)

Each internal branch defines a bipartition (split) on a tree



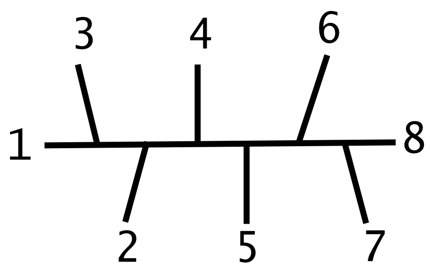
a: 1, 2 | 3,4,5,6,7,8

b: 1,2,3 | 4,5,6,7,8

c: 1,2,3,4 | 5,6,7,8

d: 1,2,3,4,5 | 6,7,8

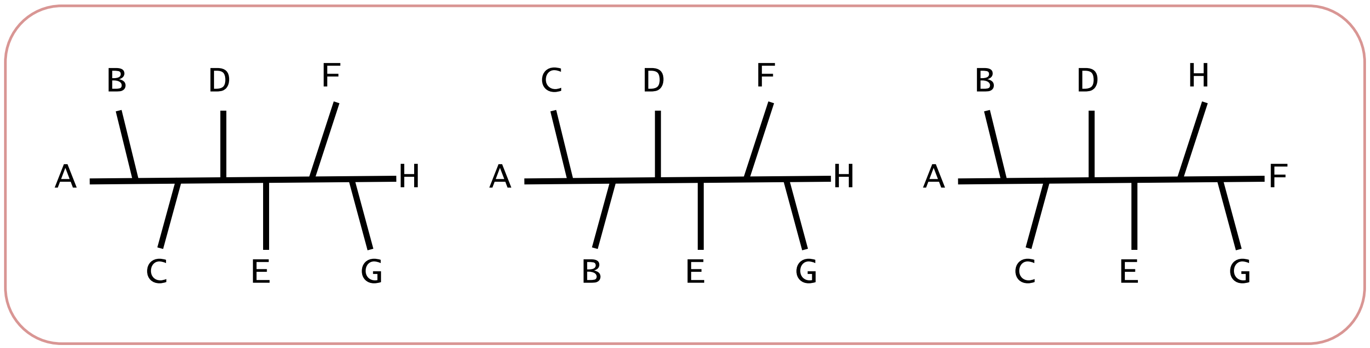
e: 1,2,3,4,5,6 | 7,8

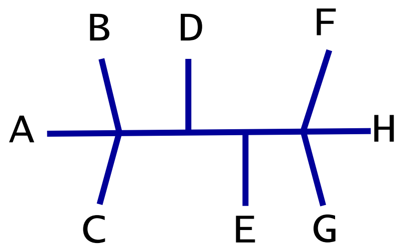


What is the partition distance between these two trees?

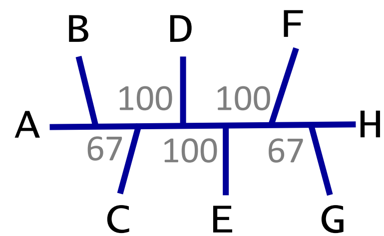
The partition distance ranges from 0 to $2(n – 3)$ for n sequences

Consensus trees





strict consensus



majority-rule consensus

A consensus tree shows clades that are shared by a set of trees

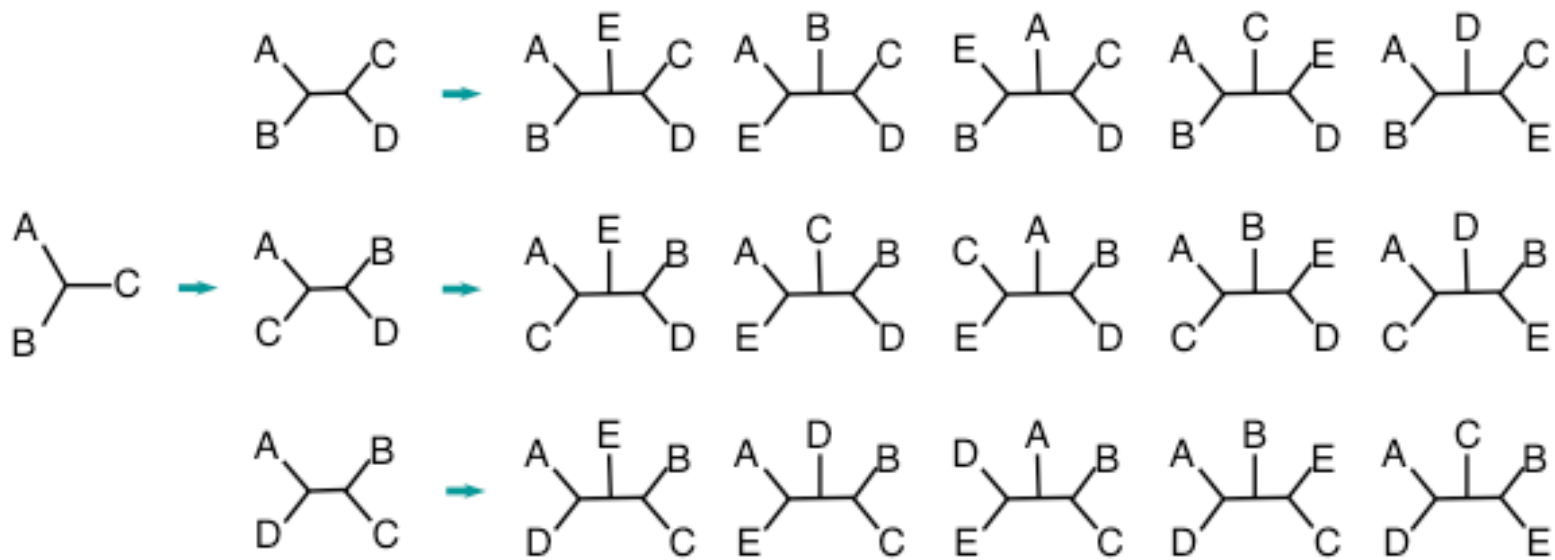
The *strict consensus tree* shows a clade only if it is in every tree of a set

The *majority-rule consensus tree* shows a clade if it is in >50% of a set


Quizz?

How many trees?

Step-wise addition algorithm (Cavalli-Sforza & Edwards 1967):



$T_3 = 1$ $\quad$ $T_4 = 1 \times 3$ $\quad$ $T_5 = 1 \times 3 \times 5$

unrooted trees for n+1 taxa: $T_{n+1} = T_n \times (2n\text{-}3)$

Figure from Yang (2006) Computational Molecular Evolution, OUP

How many trees?

n	Unrooted	Rooted
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10,395
8	10,395	135,135
9	135,135	2,027,025
10	2,027,025	34,459,425
20	$\sim 2.22 \times 10^{20}$	$\sim 8.20 \times 10^{21}$
50	$\sim 2.84 \times 10^{74}$	$\sim 2.75 \times 10^{76}$

Table from Yang (2006) Computational Molecular Evolution, OUP

unrooted trees for n+1 taxa: $T_{n+1} = T_n \times (2n-3)$

Classification of tree inference methods

	Distance-based	Character-based
Cluster methods	UPGMA Neighbour-joining (NJ)	
Optimality criterion	Minimum evolution (ME)	Maximum parsimony (MP) Maximum likelihood (ML) Bayesian

Optimality criteria

- **Maximum parsimony:** The parsimony score is the minimum number of required changes or steps. Given two trees, the one minimizing the parsimony score is the better.
- **Maximum likelihood:** The log likelihood value measures the fit of the tree to data. Given two trees, the one with the higher log likelihood is the better.
- **Minimum evolution:** The sum of branch lengths measures the fit of the tree to data. Shorter trees are preferred. This is a distance-based method.
- **Bayesian methods:** The posterior probability of a tree (clade) is the probability that the tree (clade) is correct, given the data and model. The MAP tree has the maximum posterior probability.

Heuristics

Tree search under optimality criterion:

- *Exhaustive tree search* evaluates all possible trees (only possible with very few taxa)

- *Heuristic tree search* does not guarantee finding the optimal tree
  - stepwise addition
  - star decomposition
  - branch swapping
    - nearest neighbor interchange (NNI)
    - subtree-pruning and regrafting (SPR)
    - tree bisection and reconnection (TBR)
    - …

> *"They are, of their very nature, are a bit ad hoc.."*
> Felsenstein (2004)
> *Inferring Phylogenies*

Stepwise addition

Illustrated under maximum parsimony criterion

Number of trees evaluated for *n* taxa:

$3+5+7+\ldots+(2n-5) = (n-1)(n-3)$

Often performed several times with a different starting tree

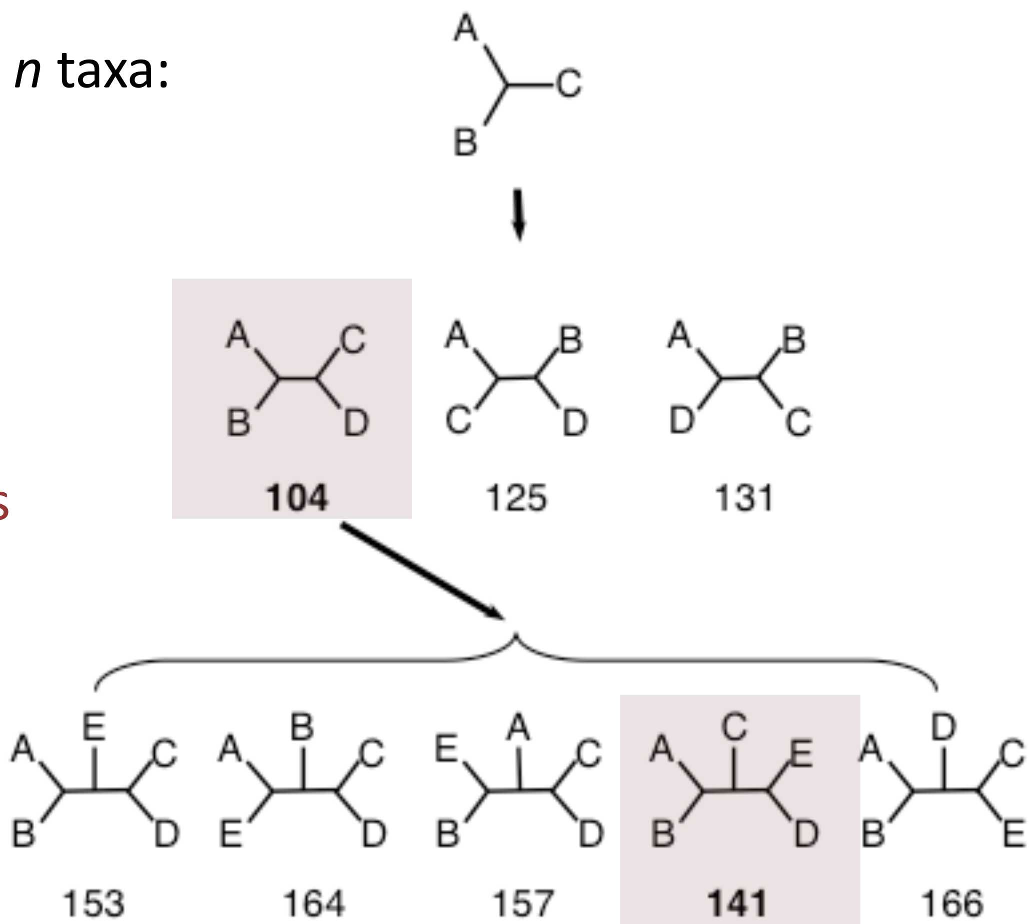


Figure from Yang (2006) Computational Molecular Evolution, OUP

Star decomposition

Illustrated under maximum likelihood criterion

Number of trees evaluated for *n* taxa:

$n(n\text{-}1)/2 + (n\text{-}1)(n\text{-}2)/2 + \ldots + 3$
$= n(n^2\text{-}1)/6 \text{ -}7$

Evaluates more trees:
slower than stepwise addition

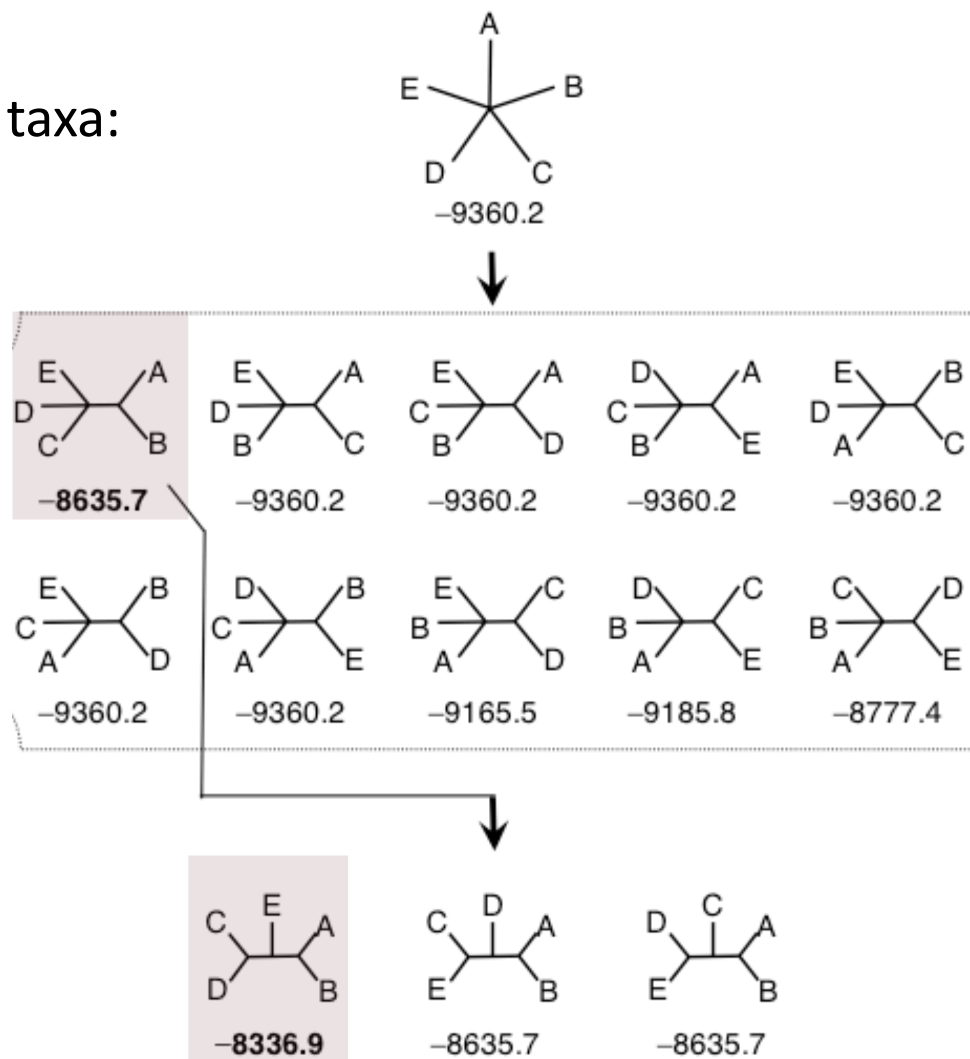
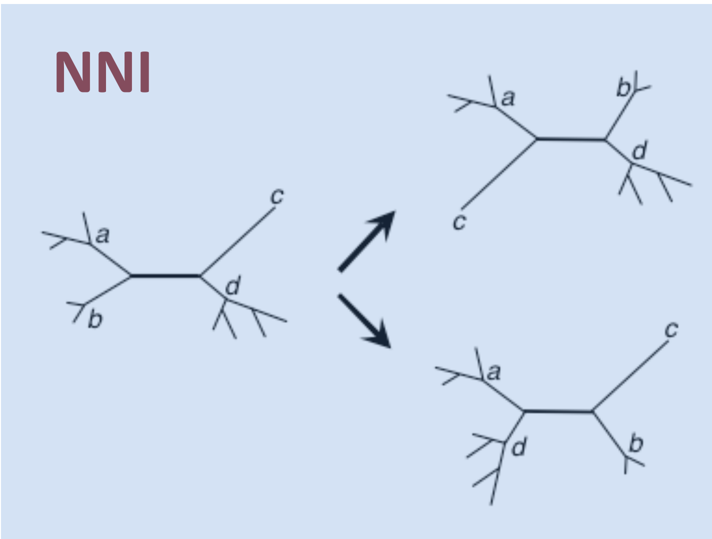
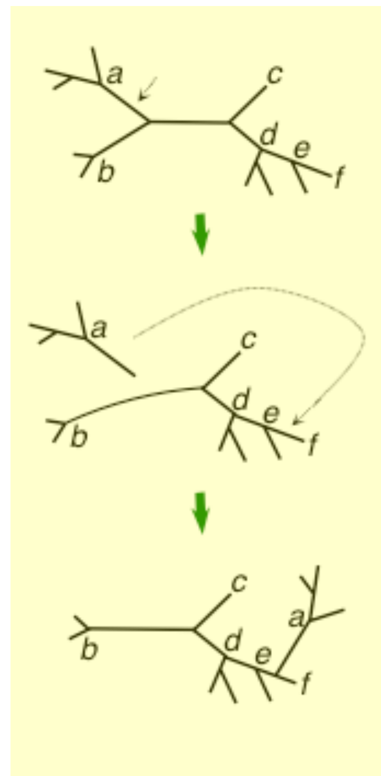


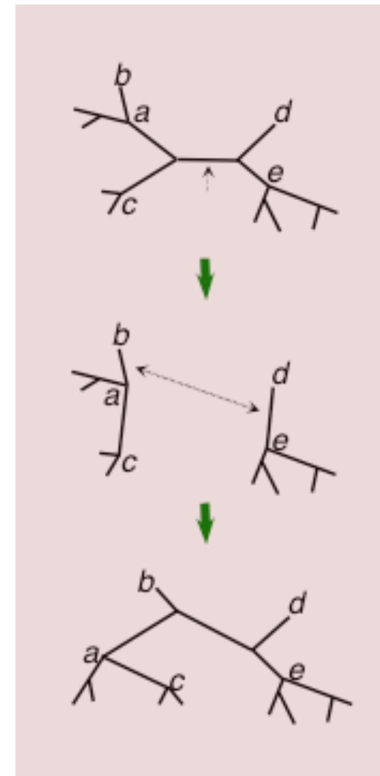
Figure from Yang (2006) Computational Molecular Evolution, OUP

Branch-swapping heuristics



SPR





TBR

The heuristic algorithm affects the chance of finding the best fitting tree

Figures from Yang (2006) Computational Molecular Evolution, OUP

III. Time complexity of tree search

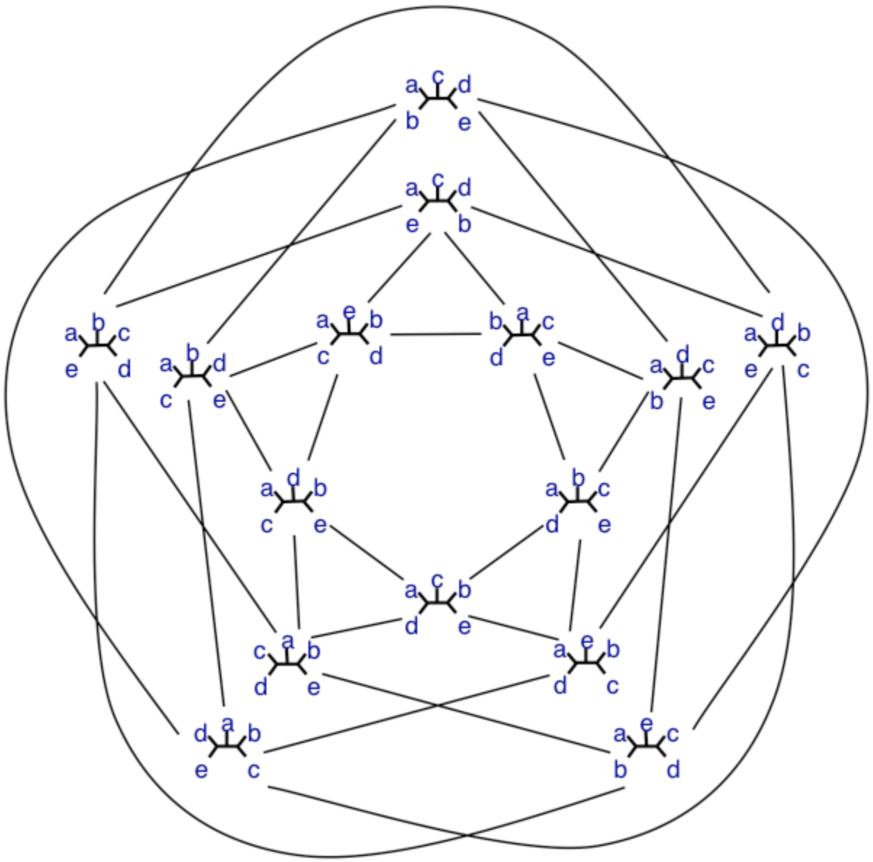
• Running times depend on the size of the data: number of taxa (n), sites, alphabet size, number of rates categories…

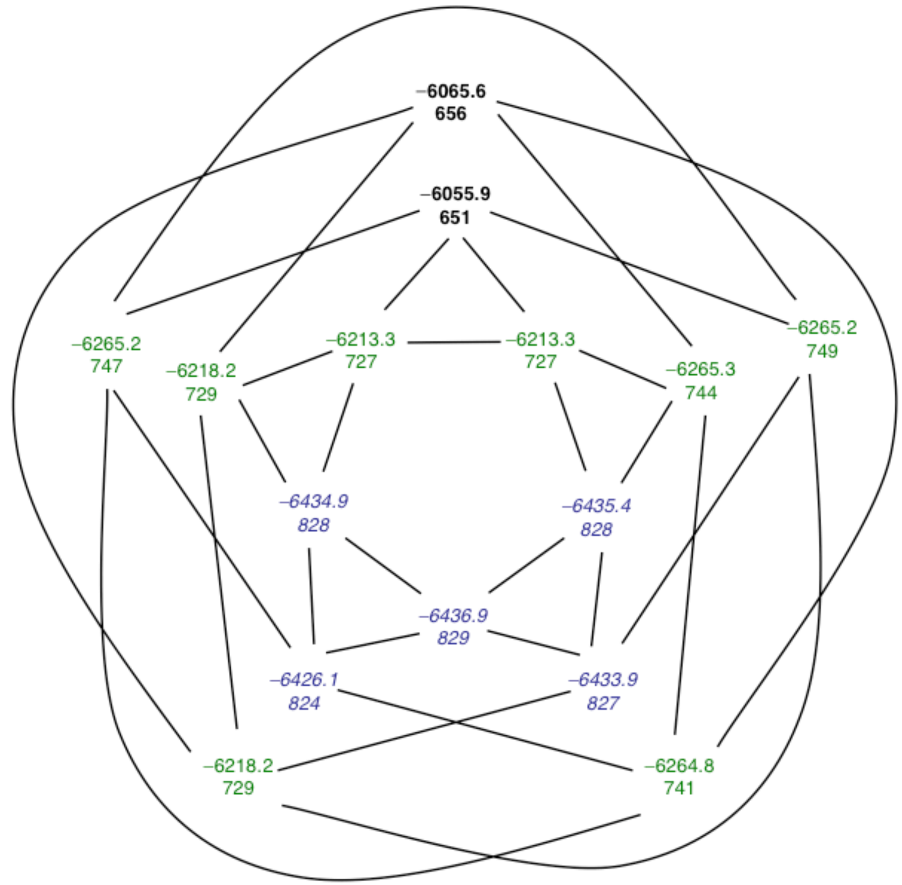
• O(f(parameters)) notation means that the running time is proportional to f(parameters)

• For example, for NNI, SPR and TBR the time complexity is *$O(n)$, $O(n^2)$ and $O(n^3)$ respectively*

• Exhaustive searches (with MP or ML) are NP-hard: The best tree(s) has worst case running times in *$O(e^n)$*

Local & global optima in tree space





15 trees for 5 species with neighbor relationships

modified by Yang (2006) after Felsenstein (2004)

Methods of phylogenetic inference

- Maximum parsimony (MP)
- Distance methods
- Maximum likelihood (ML)
- Bayesian inference

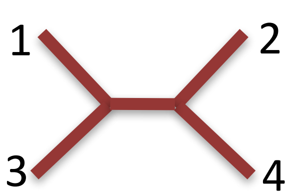
Maximum parsimony

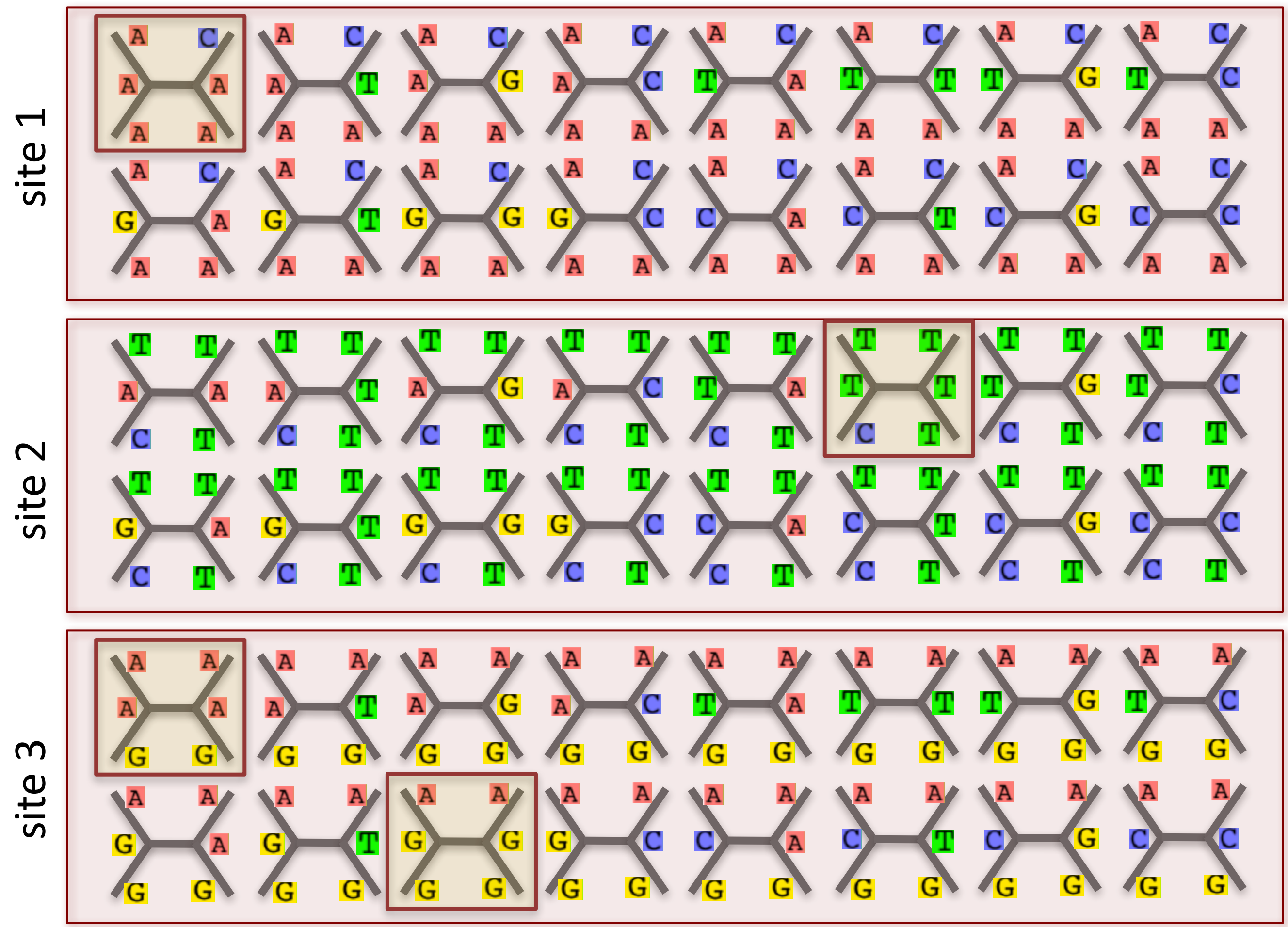
MP selects a tree with a min. number of changes

1	A	T	A
2	C	T	A
3	A	C	G
4	A	T	G

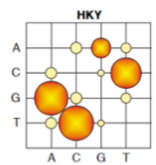
To score a tree min numbers of changes are summed for sites:

MP score: 1+1+2 =4

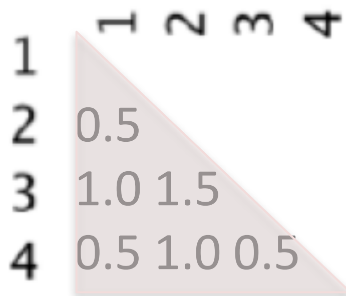




Distance-based inference



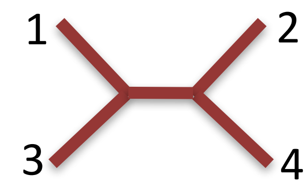
construct distance matrix



Tree inference from distance matrix

eg, minimize:

$$\sum_{i,j}(t_{ij}-d_{ij})^2/v_{ij}$$



Methods:

Least squares (LS), minimum evolution (ME), neighbor-joining (NJ)

Disadvantage:

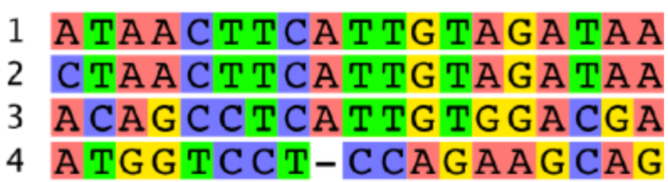
Pairwise distance estimation is not reliable for large divergences

Advantage: algorithmic approaches are fast

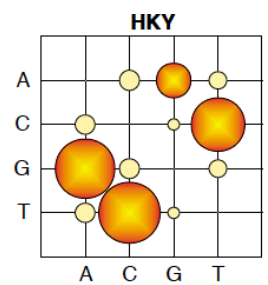
Maximum Likelihood (ML)

Estimate tree and model parameters by maximizing the probability of observing data:

Pr(



|



)

Data

Model

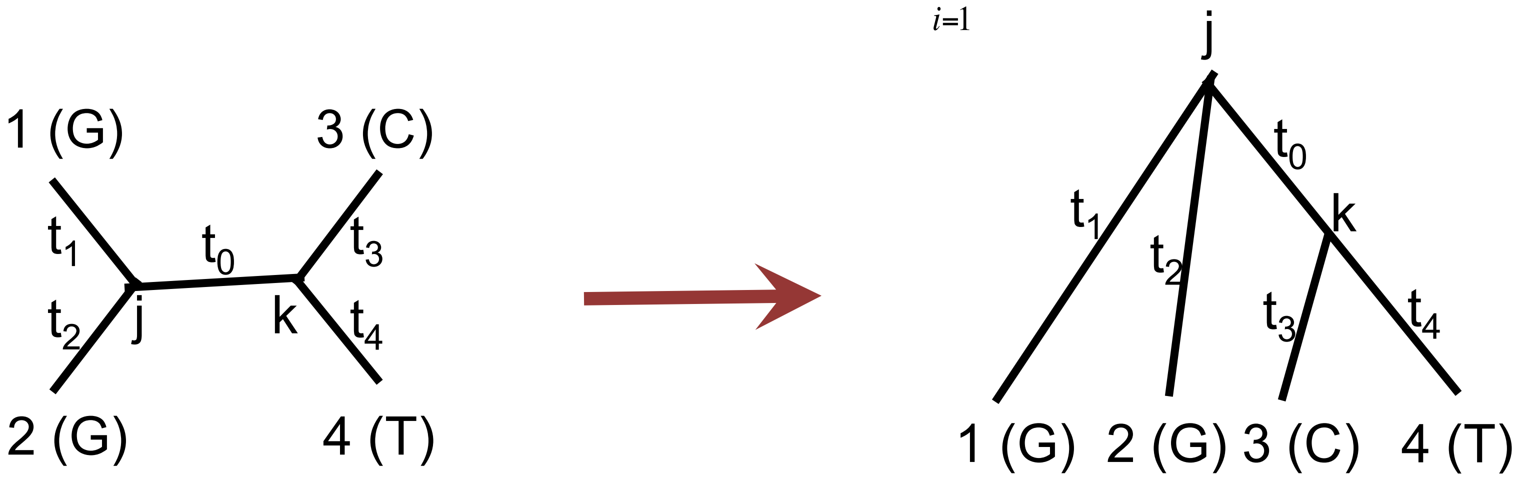
Tree,
br. lengths

Maximum Likelihood (ML)

Site	1	2	3	4	5	...	**i**	...					n
Seq 1	C	T	C	A	T	...	**G**	...	G	T	A	A	T
Seq 2	C	T	A	G	T	...	**G**	...	C	T	A	G	T
Seq 3	C	T	A	G	T	...	**C**	...	G	T	A	G	T
Seq 4	C	C	A	A	C	...	**T**	...	C	C	A	A	T
Probability	p_1	p_2				...	$\mathbf{p_i}$	...					p_n

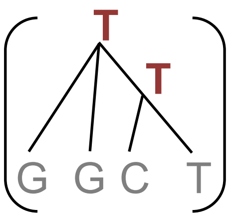
$$L = p_1 \times p_2 \times \ldots p_i \times \ldots \times p_n = \prod_{i=1}^{n} p_i$$

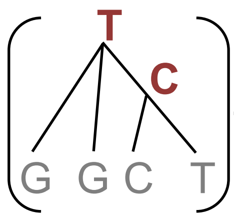
$$\ell = \log L = \log p_1 + \log p_2 + \ldots \log p_n = \sum_{i=1}^{n} \log p_i$$

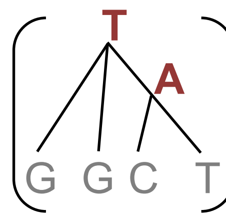


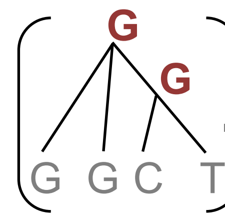
Maximum Likelihood (ML)

The probability of each site is a sum over all possible ancestral states

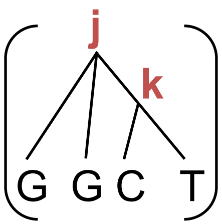
$p_i = \Pr$ 

$+ \Pr$ 

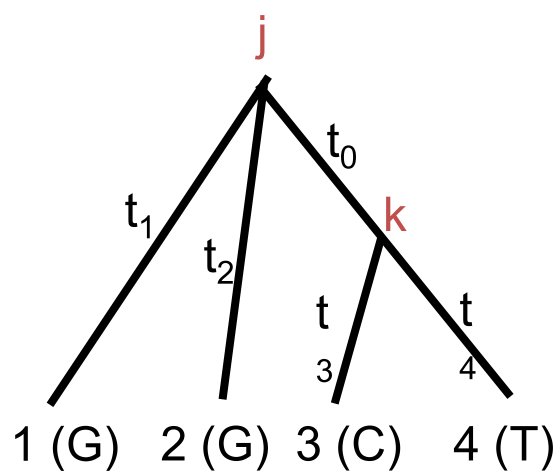
$+ \Pr$ 

$+ \ldots + \Pr$ 

.

$\Pr$ 

$= \pi_j p_{j\mathrm{G}}(t_1) p_{j\mathrm{G}}(t_2) p_{jk}(t_0) p_{k\mathrm{C}}(t_3) p_{k\mathrm{T}}(t_4)$



Use Felsenstein's pruning algorithm

ML summary

The log likelihood ℓ is a sum of the log probabilities over all sites. For each ancestral reconstruction, the probability is a product of the transition probabilities over branches.

$$\ell(t_0, t_1, t_2, t_3, t_4 \mid X) = \sum_{i=1}^{n} \log(p_i)$$

ℓ is a function of the branch lengths t_0, t_1, t_2, t_3, t_4 (and substitution parameters, if any), which are estimated by maximizing ℓ. The optimum ℓ corresponding to the MLEs of parameters is the score for the tree. We repeat this process for all possible trees (or during heuristic search). The ML tree is the one with the highest score.

ML summary

Advantages

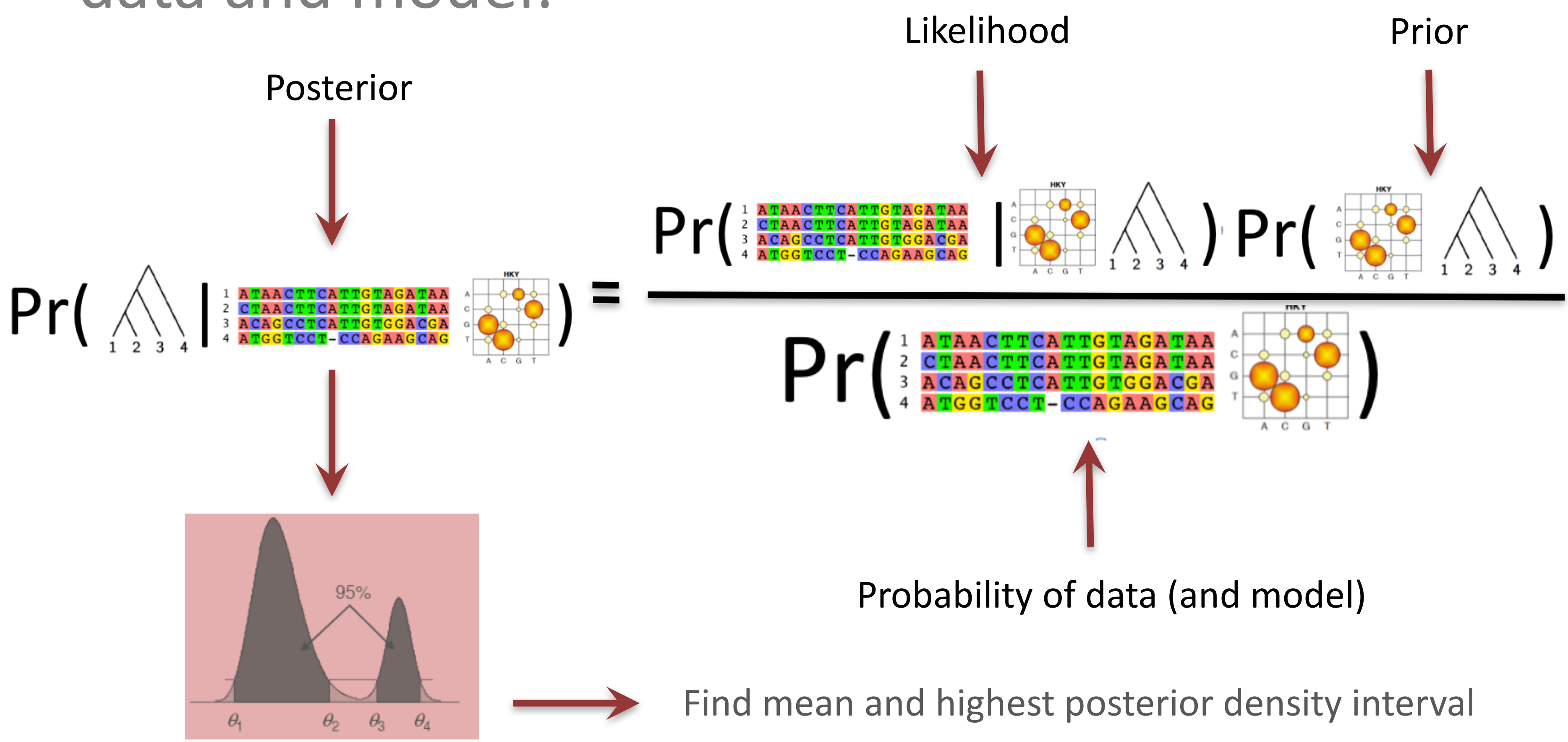
- Flexible statistical framework for testing evolutionary hypotheses
- Models can be tested and improved to fit data

Disadvantages

- Slow, but fast programs now exist (PhyML, RAxML, Garli)
- Difficulties in applying standard theory to tree comparison

Bayesian phylogenetic inference

Estimate the posterior distribution of trees given data and model:



Bayesian phylogenetic inference

$$P(\tau_i \mid X) = \frac{\iint f(\theta) f(\tau_i) f(\mathbf{b}_i \mid \theta, \tau_i) f(X \mid \theta, \tau_i, \mathbf{b}_i)\, \mathrm{d}\mathbf{b}_i \mathrm{d}\theta}{f(X)}$$

Parameters that need priors:

- tree topology τ_i (uniform)
- branch lengths b_i (uniform or exponential)
- parameters in the substitution model θ

Markov chain Monte Carlo

MCMC: used for **sampling from probability distributions** by constructing a Markov chain with the desired stationary distribution.
The state of the chain after a large number of steps is used as a sample from the desired distribution (after discarding burn-in).
The quality of the sample improves as a function of the number of steps.

In Bayesian inference:

Target distribution is the posterior distribution of interest
Proposal distribution is used to generate a candidate for the next sampled point, which is accepted or rejected with some probability

General idea: MCMC robot

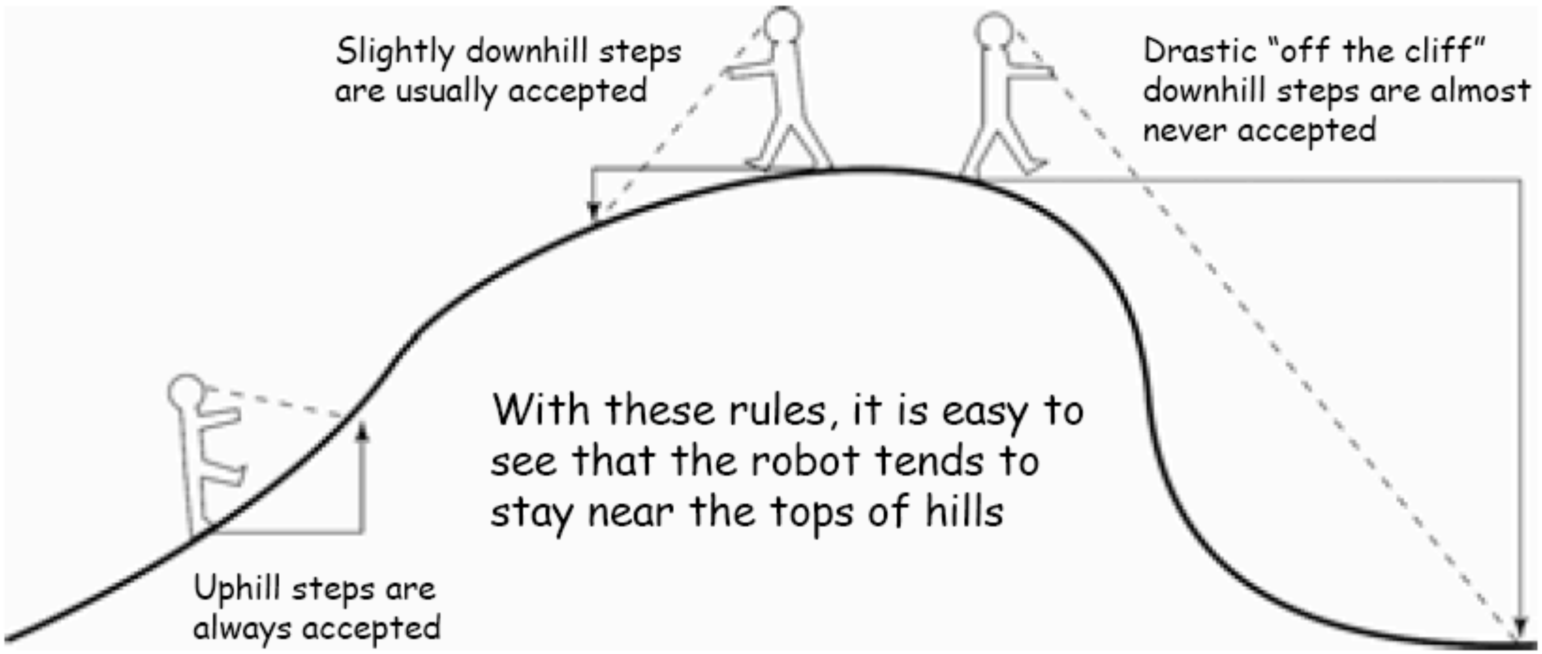


figure © Paul O. Lewis 2007

Markov chain Monte Carlo

The ratio of posteriors is easier to calculate than the posterior itself:

$$f(\theta \mid D) = \frac{f(D \mid \theta) f(\theta)}{f(D)}$$

$$\frac{f(\theta^* \mid D)}{f(\theta \mid D)} = \frac{\dfrac{f(D \mid \theta^*) f(\theta^*)}{\cancel{f(D)}}}{\dfrac{f(D \mid \theta) f(\theta)}{\cancel{f(D)}}} = \frac{f(D \mid \theta^*) f(\theta^*)}{f(D \mid \theta) f(\theta)}$$

Bayesian inference: summaries

- MAP tree: tree topology with the maximum posterior probability
- 95% credibility set of trees: add trees with the highest posterior probabilities until the total probability ≥ 95%
- Posterior clade probability: proportion of sampled trees that contain the clade, shown on the majority-rule consensus tree

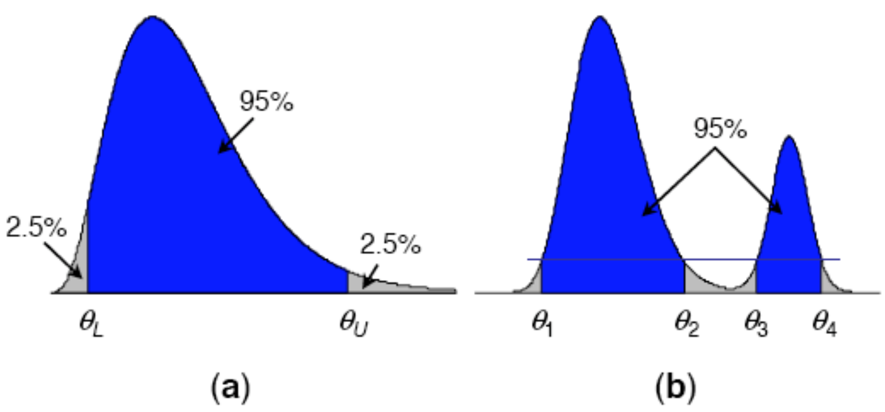


figure from Ziheng Yang

More generally:

Mean, median, mode as point estimate
95% equal tail credibility interval (a)
95% highest posterior density interval (b)

Sketch of MCMC for tree inference

- Start with a random tree τ, with random branch lengths b, and random substitution parameters θ.
- In each iteration do the following:
  - Propose a change to the tree, by using tree rearrangement algorithms (such as nearest neighbour interchange or subtree pruning and regrafting). The step may change b as well.
  - Propose changes to branch lengths b.
  - Propose changes to parameters θ.
  - Decide: accept or not?
- Every *k* iterations, sample the chain: save τ, b, θ to disk.
- At the end of the run, summarize the results.

Bayesian phylogenetic inference

- Posterior probability distribution for each branch may be estimated from MCMC samples of trees (convergence?)
- Theoretically, these posteriors may be interpreted as probabilities (under the true model!)
- Dependency on prior for trees and model parameters (unlike likelihood)

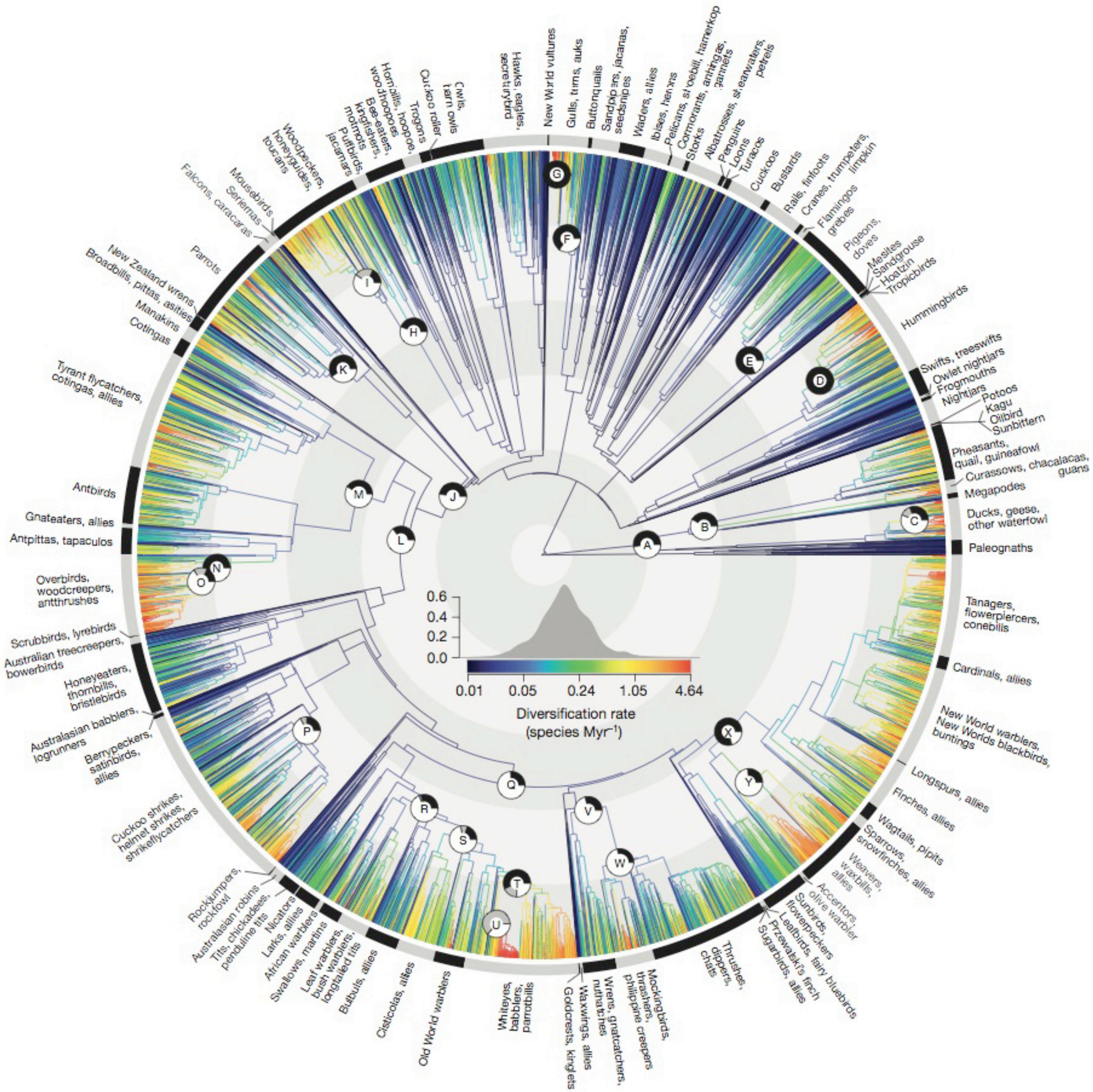
Some known trends

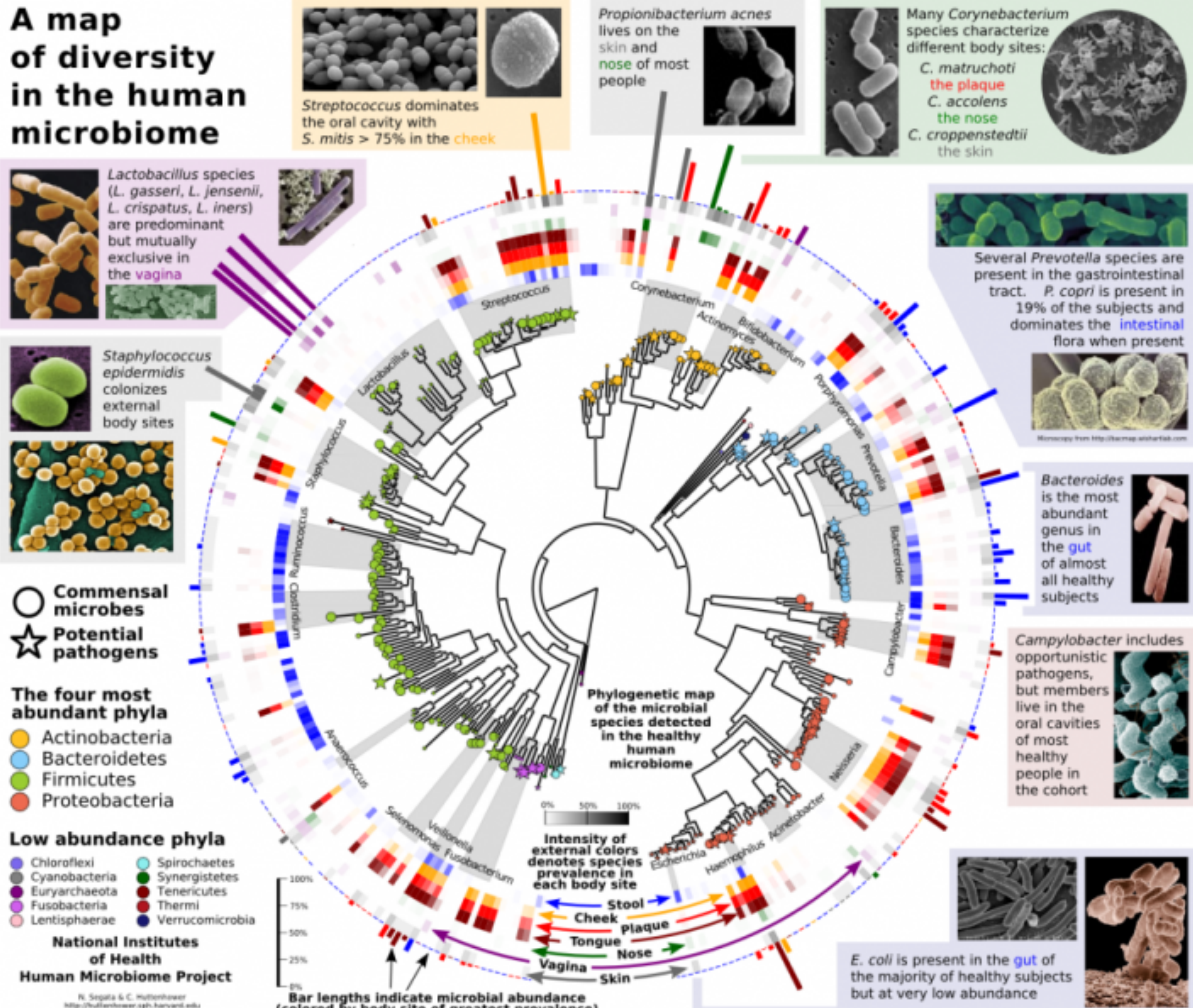
- LBA-like artefacts affect parsimony, as well other methods under over-simplistic models
- Bayesian and ML tree inference is generally more accurate than parsimony and distance, but model is important
- Distance methods perform poorly for highly divergent or “gappy” sequences
- Lack/loss of information for too similar/divergent data: no method can recover the true tree with confidence
- Success of reconstruction also depends on the tree shape: “easy” trees have long internal branches relative to external, “hard” trees have short internal branches relative to external

Applications of phylogenies

- Reconstruct molecular history
- Study ancient proteins (ancestral reconstruction)
- Molecular dating of speciation events
- Study change of gene function
- Find molecular changes that cause disease
- Study host pathogen dynamics
- Choose model organism for drug design
- Distribution and cohabitation in metagenomics

Diversity of birds (9993 species)




A map of diversity in the human microbiome
Streptococcus dominates the oral cavity with S. mitis > 75% in the cheek
Propionibacterium acnes lives on the skin and nose of most people
Many Corynebacterium species characterize different body sites:
C. matruchoti the plaque
C. accolens the nose
C. croppenstedtii the skin
Lactobacillus species (L. gasseri, L. jensenii, L. crispatus, L. iners) are predominant but mutually exclusive in the vagina
Several Prevotella species are present in the gastrointestinal tract. P. copri is present in 19% of the subjects and dominates the intestinal flora when present
Staphylococcus epidermidis colonizes external body sites
Bacteroides is the most abundant genus in the gut of almost all healthy subjects
Campylobacter includes opportunistic pathogens, but members live in the oral cavities of most healthy people in the cohort
E. coli is present in the gut of the majority of healthy subjects but at very low abundance
Commensal microbes
Potential pathogens
The four most abundant phyla
Actinobacteria
Bacteroidetes
Firmicutes
Proteobacteria
Low abundance phyla
Chloroflexi
Cyanobacteria
Euryarchaeota
Fusobacteria
Lentisphaerae
Spirochaetes
Synergistetes
Tenericutes
Thermi
Verrucomicrobia
National Institutes of Health Human Microbiome Project
Phylogenetic map of the microbial species detected in the healthy human microbiome
0%
50%
100%
Intensity of external colors denotes species prevalence in each body site
Stool
Cheek
Plaque
Tongue
Nose
Vagina
Skin
100%
75%
50%
25%
0%
Bar lengths indicate microbial abundance (colored by body site of greatest prevalence)
Streptococcus
Lactobacillus
Staphylococcus
Ruminococcus
Clostridium
Anaerococcus
Selenomonas
Veillonella
Fusobacterium
Corynebacterium
Actinomyces
Bifidobacterium
Porphyromonas
Prevotella
Bacteroides
Campylobacter
Neisseria
Acinetobacter
Haemophilus
Escherichia
N. Segata & C. Huttenhower

MOLECULAR EPIDEMIOLOGY

HIV-1 and HCV sequences from Libyan outbreak

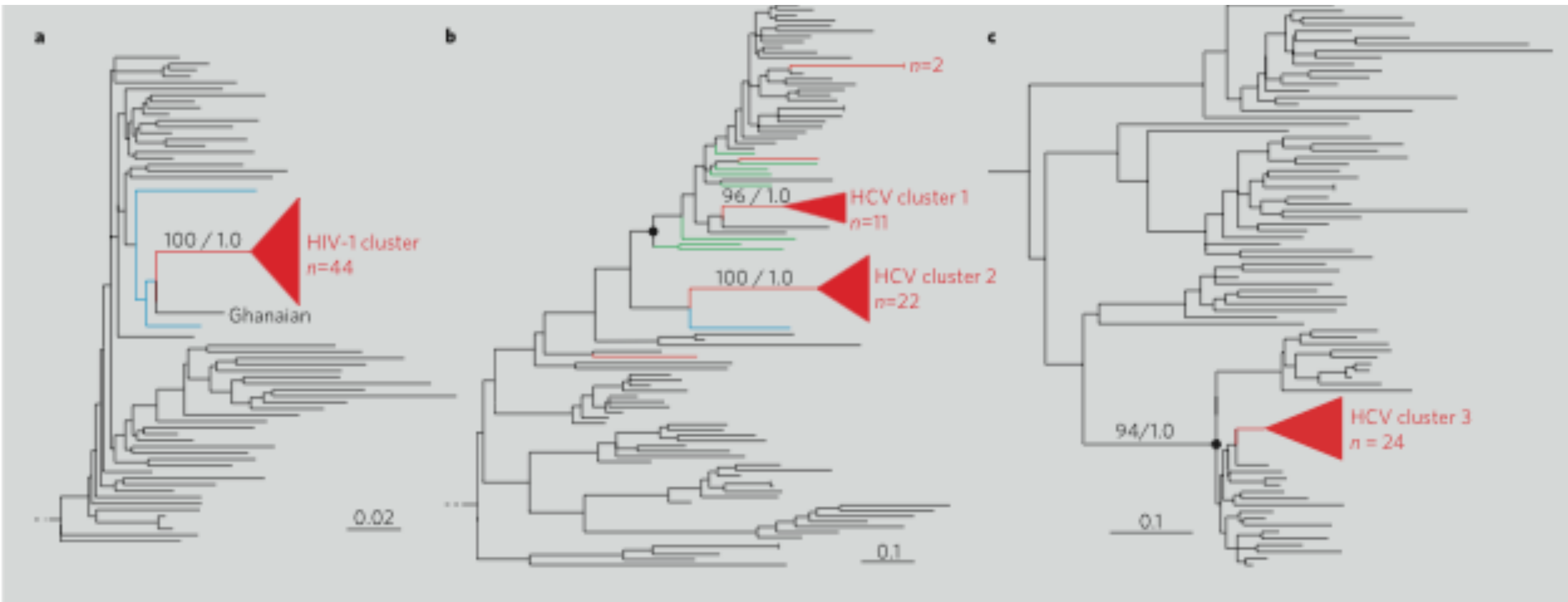


Figure 1 | HIV-1 and HCV sequences from 1998 Al-Fateh Hospital (AFH) outbreak. a–c, Estimated maximum-likelihood phylogenies for HIV-1 CRF02_AG (**a**), HCV genotype 4 (**b**) and HCV genotype 1 (**c**). Source of sequences used for analysis: AFH, red; Egypt, green; Cameroon, blue. Black circles mark the common ancestor of HCV subtype 4a and 1a; numbers above AFH lineages give clade support values using bootstrap and bayesian methods, respectively. Scale bar units are nucleotide substitutions per site. For visual clarity, AFH clusters are represented by triangles and some non-informative reference strains are excluded.

CORRESPONDENCE

Libya should stop denying scientific evidence on HIV

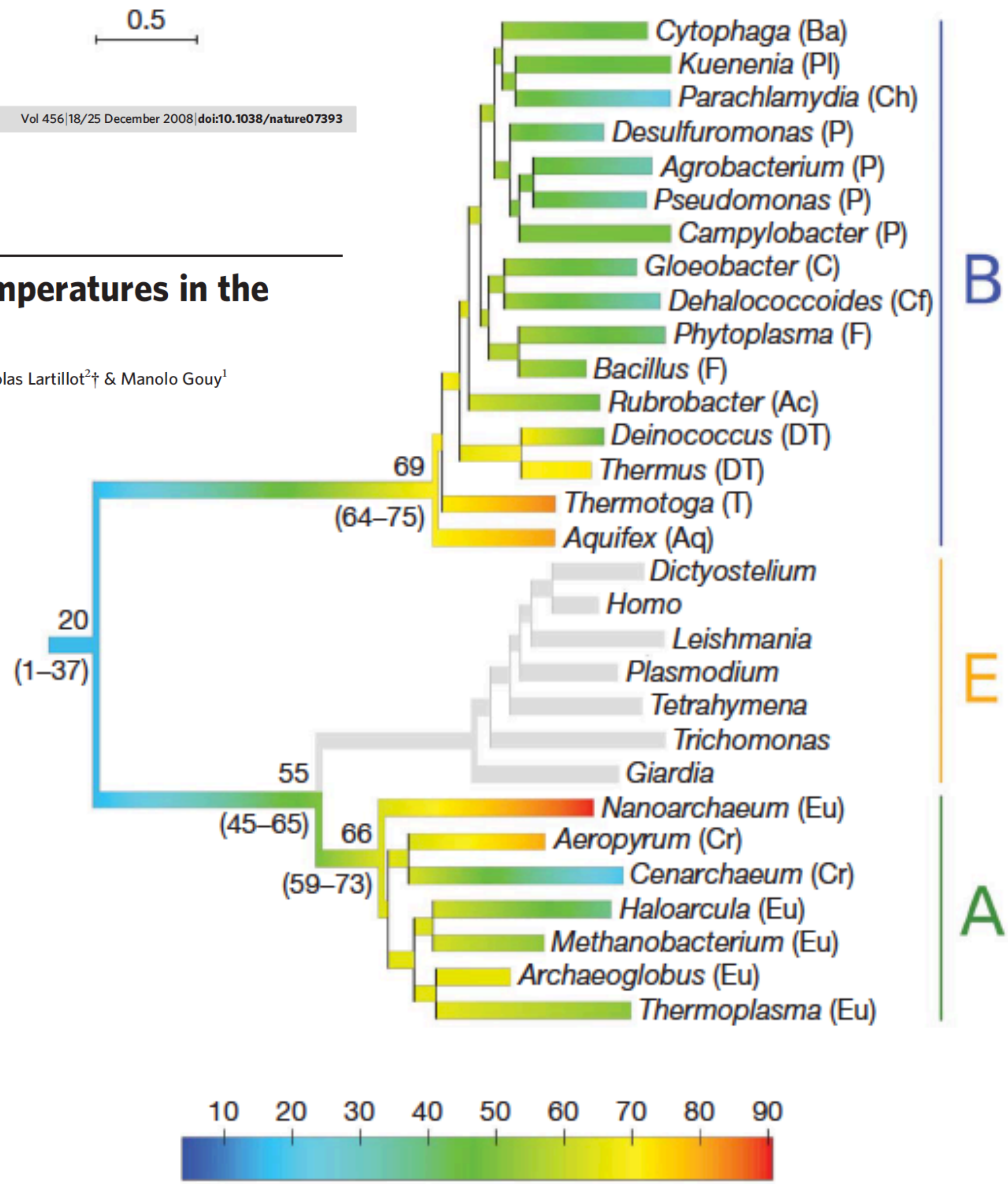
Vittorio Colizzi*, Tulio de Oliveira†, Richard J. Roberts‡

nature Vol 456|18/25 December 2008|doi:10.1038/nature07393

LETTERS

Parallel adaptations to high temperatures in the Archaean eon

Bastien Boussau[1]*, Samuel Blanquart[2]*, Anamaria Necsulea[1], Nicolas Lartillot[2]† & Manolo Gouy[1]



Anisimova et al. BMC Evolutionary Biology 2013, 13:161
http://www.biomedcentral.com/1471-2148/13/161

BMC Evolutionary Biology

EDITORIAL Open Access

State-of the art methodologies dictate new standards for phylogenetic analysis

Maria Anisimova[1,2*†], David A Liberles[3†], Hervé Philippe[4†], Jim Provan[5†], Tal Pupko[6†] and Arndt von Haeseler[7†]

Abstract

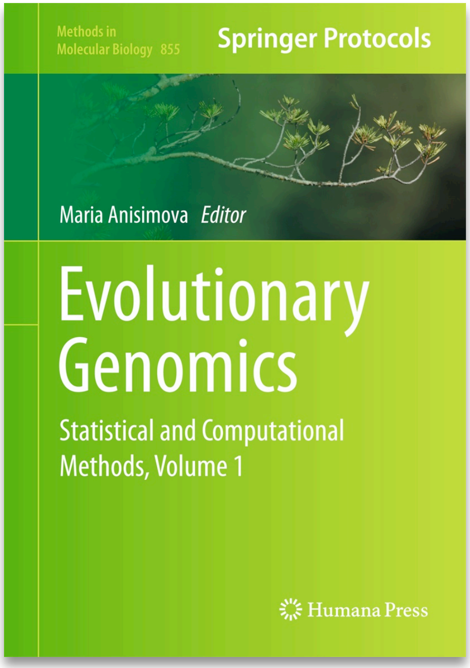
The intention of this editorial is to steer researchers through methodological choices in molecular evolution, drawing on the combined expertise of the authors. Our aim is not to review the most advanced methods for a specific task. Rather, we define several general guidelines to help with methodology choices at different stages of a typical phylogenetic 'pipeline'. We are not able to provide exhaustive citation of a literature that is vast and plentiful, but we point the reader to a set of classical textbooks that reflect the state-of-the-art. We do not wish to appear overly critical of outdated methodology but rather provide some practical guidance on the sort of issues which should be considered. We stress that a reported study should be well-motivated and evaluate a specific hypothesis or scientific question. However, a publishable study should not be merely a compilation of available sequences for a protein family of interest followed by some standard analyses, unless it specifically addresses a scientific hypothesis or question. The rapid pace at which sequence data accumulate quickly outdates such publications. Although clearly, discoveries stemming from data mining, reports of new tools and databases and review papers are also desirable.

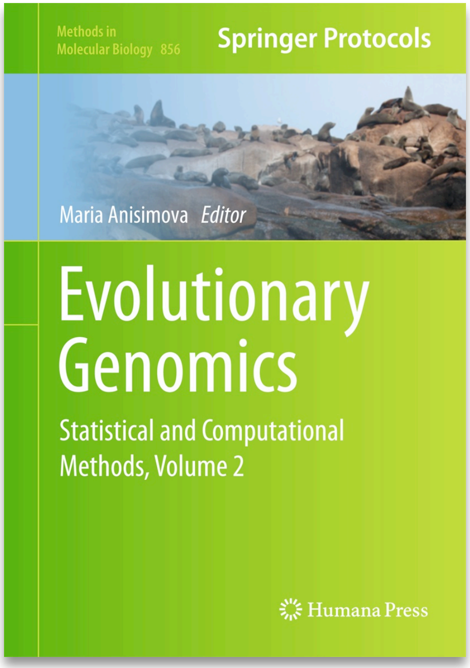
Criteria for a publishable phylogenomic study

- Strong biological motivation
- Justification for methods choice
- Use alternative methodologies
- Account for uncertainty and data filtering
- Reproducibility and data/code sharing

Reviews of the state-of the art

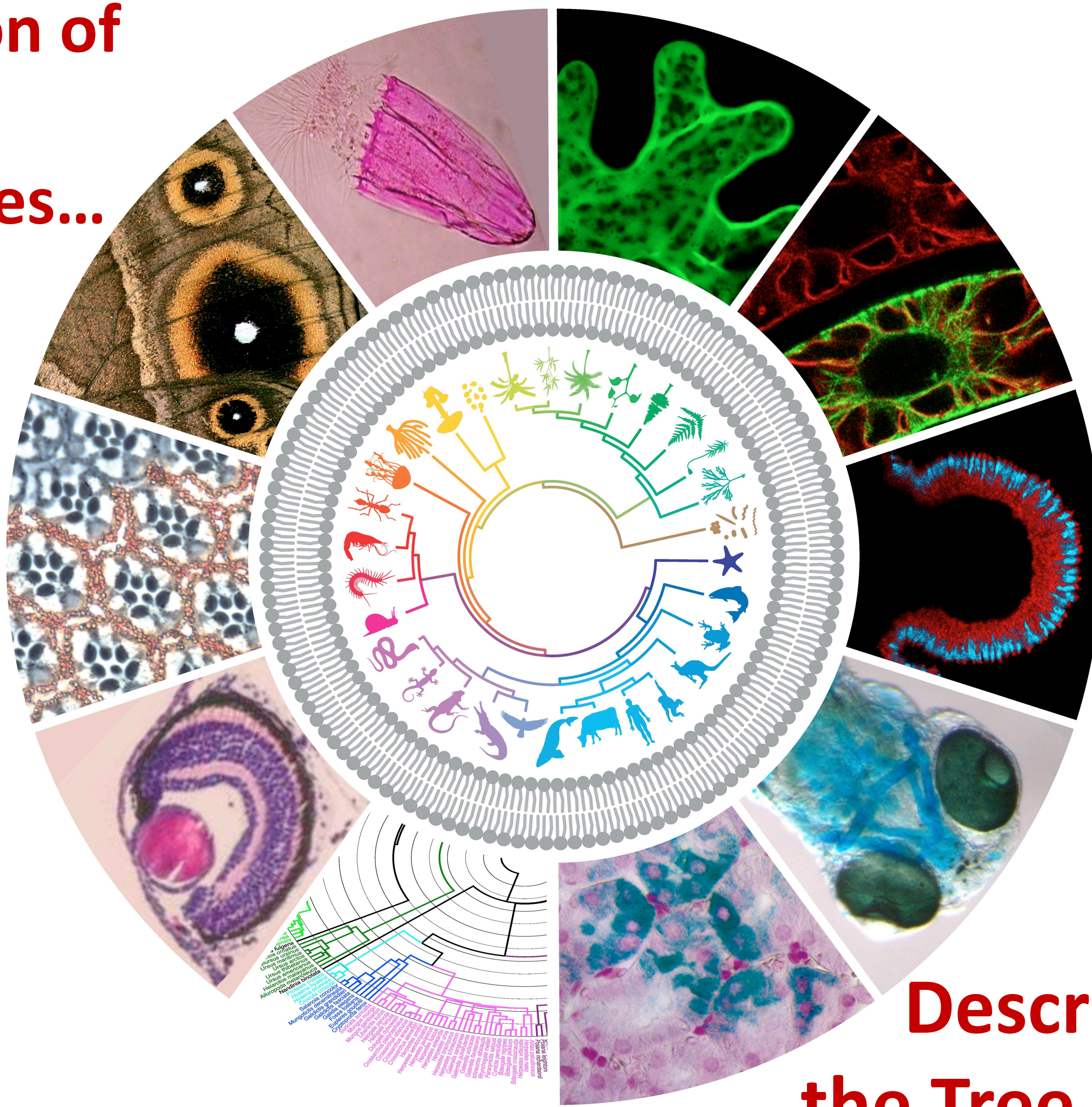
From genome assembly
and gene prediction …





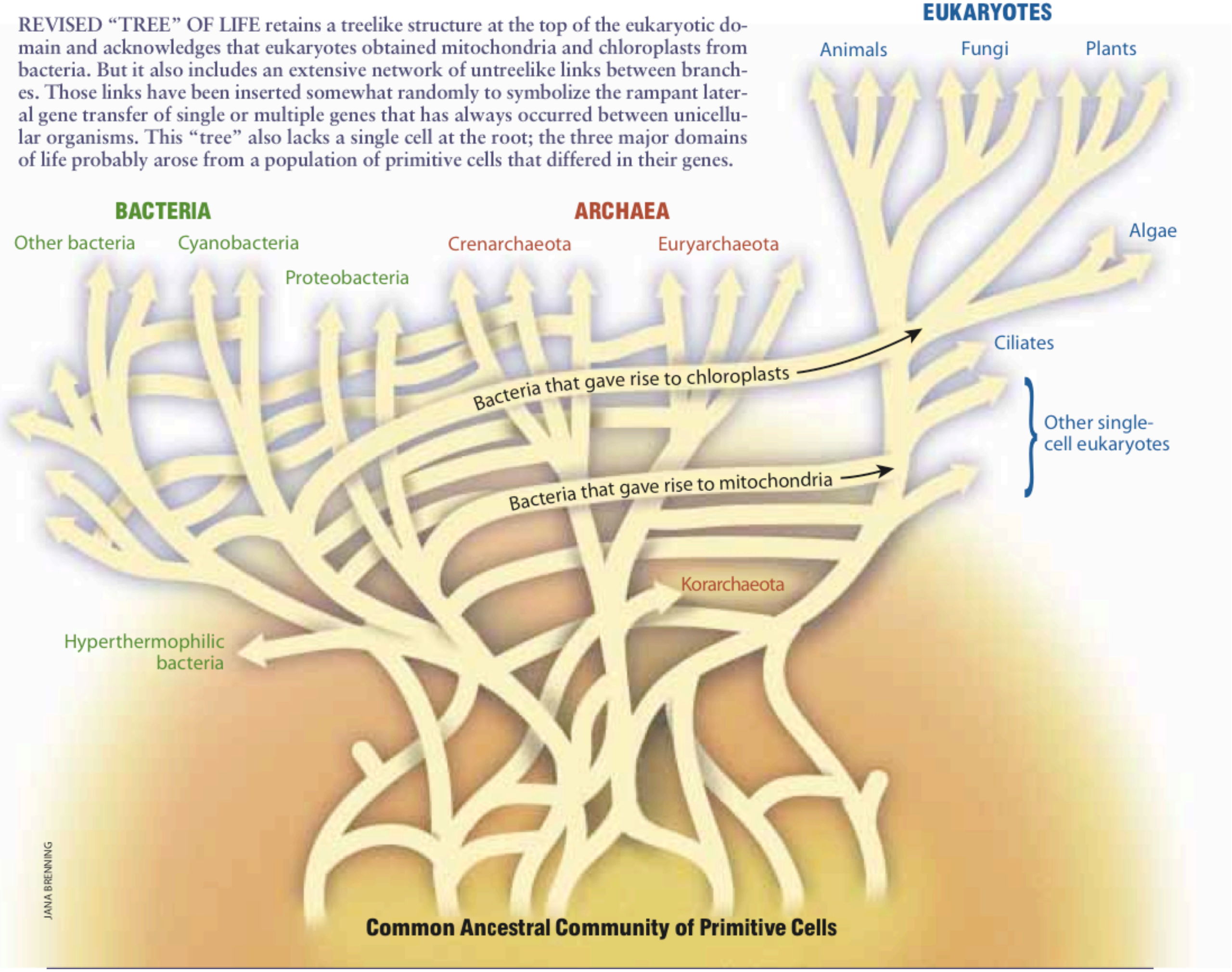
…to population genomics, omics and
aspects of data sharing and representation

Evolution of millions of species…



Described by the Tree of Life?

Problems with the Tree of Life

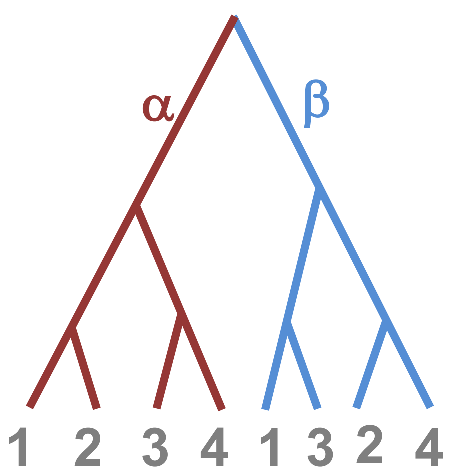


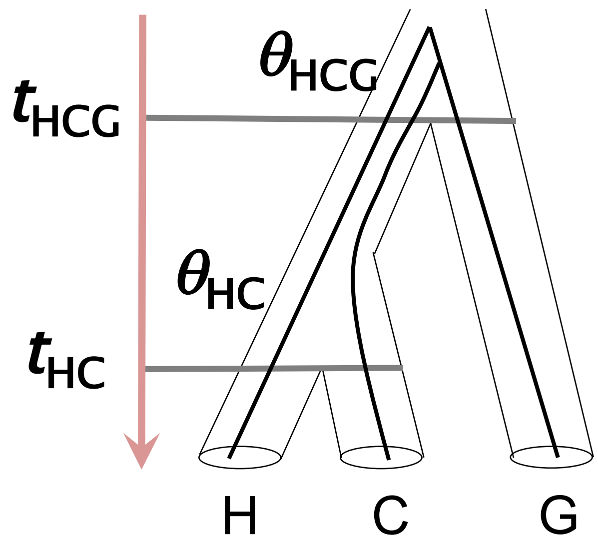
Doolittle (2000) "Uprooting the tree of Life", *Scientific American*

Gene trees vs species trees

Trees estimated from individual genes may differ from the species tree due to estimation errors, horizontal gene transfers, or use of paralogous sequences.

In closely related species, ancestral polymorphism (or lineage sorting) can also cause such conflicts. Sequences from multiple neutral loci can be used to estimate ancestral population sizes.





Slide from Ziheng Yang

Takahata, et al. 1995. *Theor. Popul. Biol.* 48:198-221
Yang 2002. *Genetics* 162:1811-1823
Rannala & Yang 2003. *Genetics* 164:1645-1656
Burgess, R. and Z. Yang. 2008 *Mol. Biol. Evol.* 25: 1979-1994

How about Forest of Life?

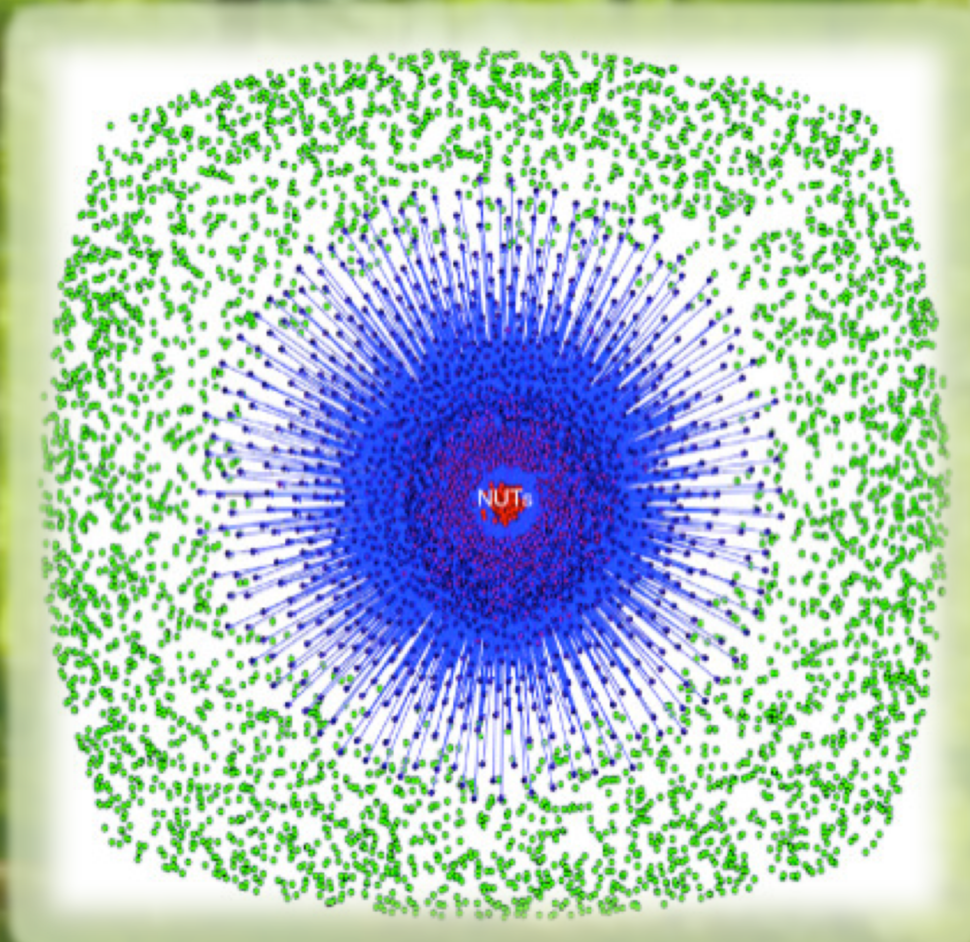


Figure from Puigbo, Wolf, Koonin (2009) J. Biol