Bayes estimators for phylogenetic reconstruction

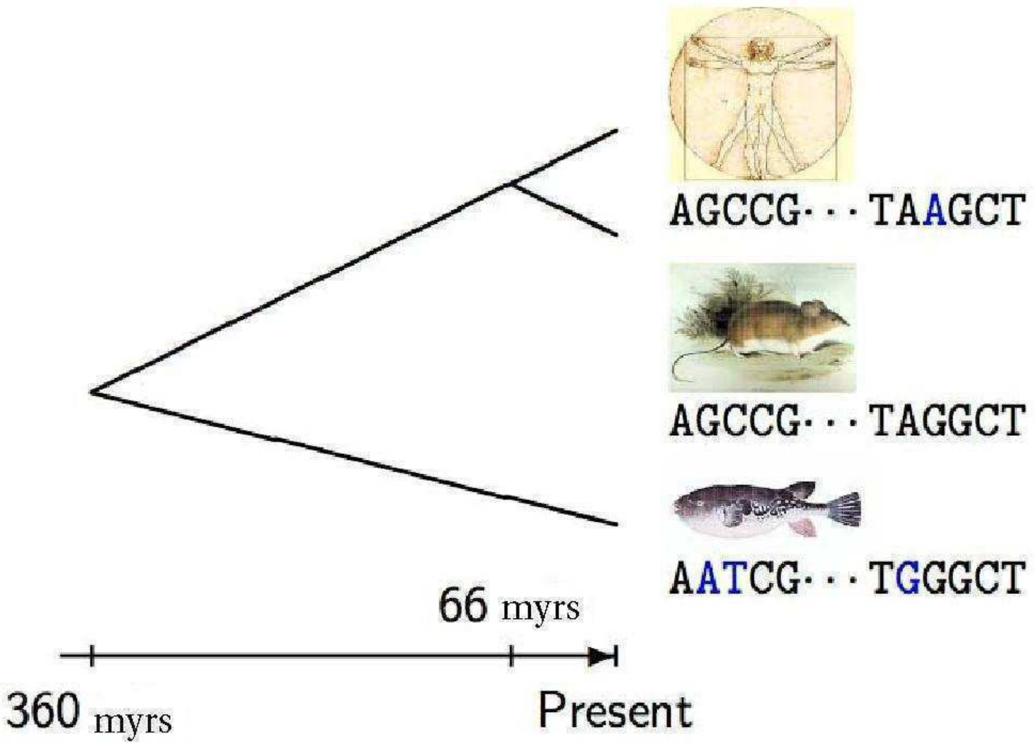
Joint work with W. Li, P. Huggins, D. Haws, and T. Friedrich

January 29, 2010

Ruriko Yoshida
Dept. of Statistics University of Kentucky

Phylogeny

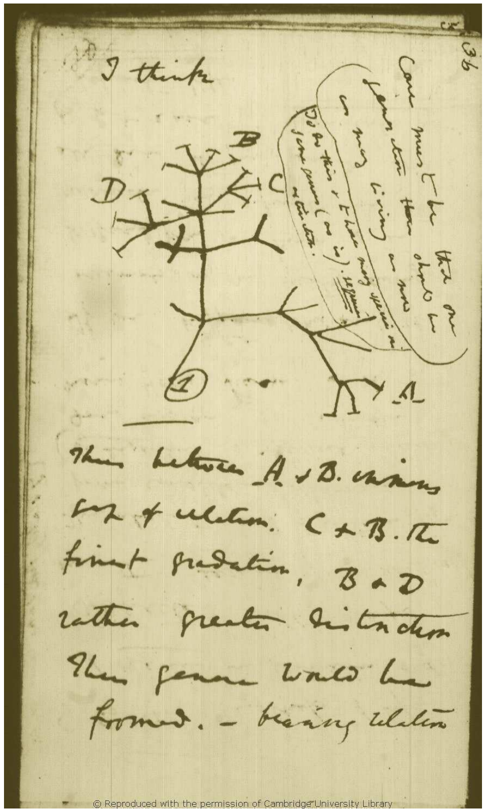
Phylogenetic trees describe the evolutionary relations among groups of organisms.



Why we care?

- We can analyze changes that have occurred in evolution of different species.
- Phylogenetic relations among different species help predict which species might have similar functions.
- We can predict changes occurring in rapidly changing species, such as HIV virus.
- Analyze cospeciation between hosts and their parasites.
- etc....

150 years since the introduction of the theory of evolution



I think

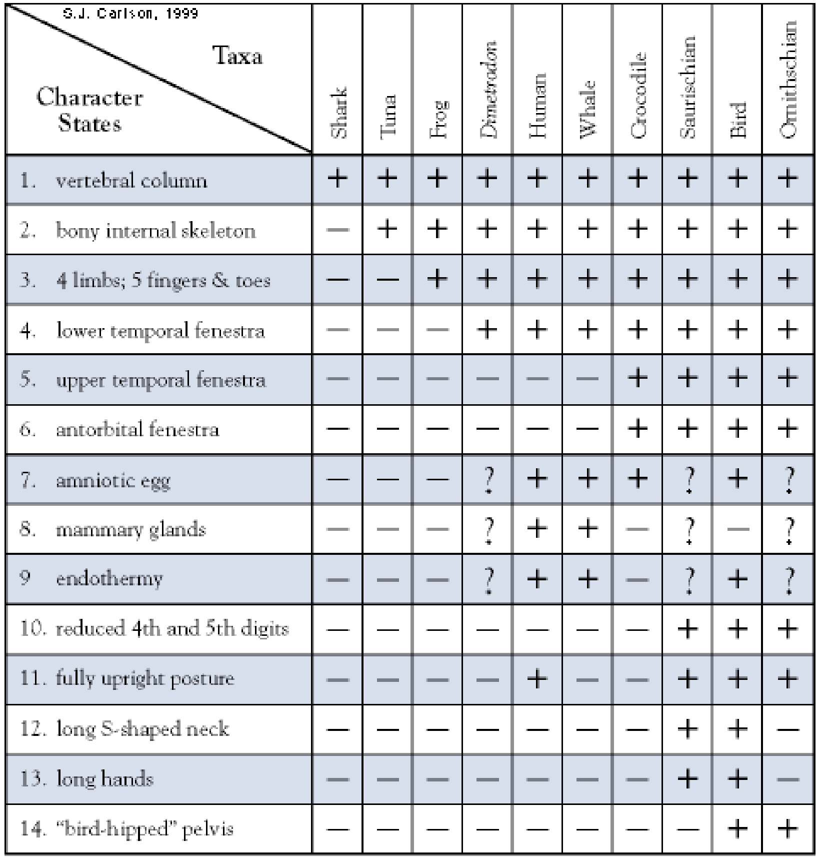
Case must be that one generation then should be as many living as now. To do this & to have many species in same genus (as is). requires extinction.

Thus between A & B. immens gap of relation. C & B. the finest gradation, B & D rather greater distinction Thus genera would be formed. — bearing relation

© Reproduced with the permission of Cambridge University Library

Rise of cladistics

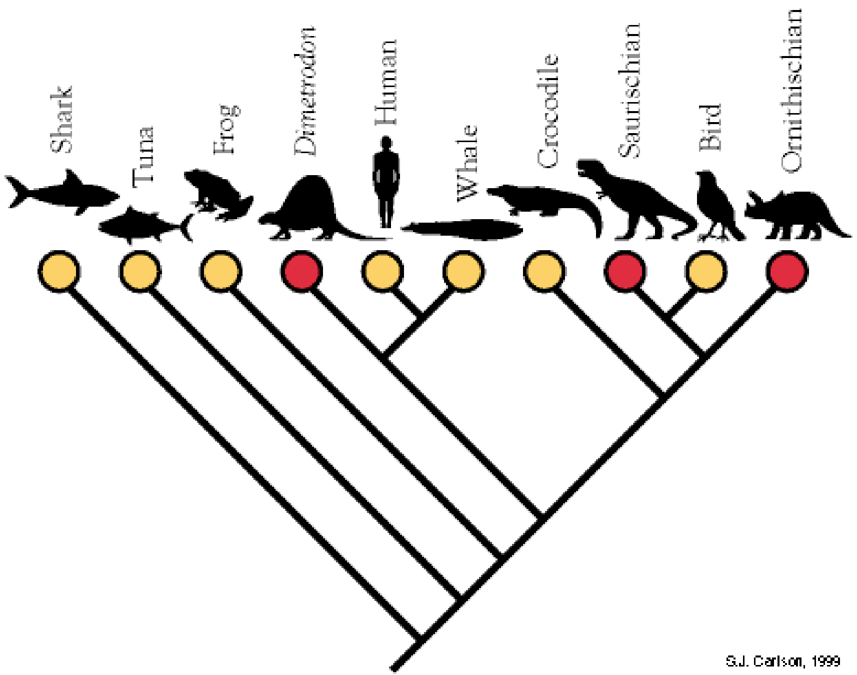
E.C. Zimmerman (30s) and W. Hennig (50s) began to define objective measures for reconstructing phylogenies (cladograms) based on the analysis of shared morphological ancestral characteristics of fossils and living organisms



S.J. Carlson, 1999

Character States / Taxa	Shark	Tuna	Frog	*Dimetrodon*	Human	Whale	Crocodile	Saurischian	Bird	Ornithschian
1. vertebral column	+	+	+	+	+	+	+	+	+	+
2. bony internal skeleton	—	+	+	+	+	+	+	+	+	+
3. 4 limbs; 5 fingers & toes	—	—	+	+	+	+	+	+	+	+
4. lower temporal fenestra	—	—	—	+	+	+	+	+	+	+
5. upper temporal fenestra	—	—	—	—	—	—	+	+	+	+
6. antorbital fenestra	—	—	—	—	—	—	+	+	+	+
7. amniotic egg	—	—	—	?	+	+	+	?	+	?
8. mammary glands	—	—	—	?	+	+	—	?	—	?
9 endothermy	—	—	—	?	+	+	—	?	+	?
10. reduced 4th and 5th digits	—	—	—	—	—	—	—	+	+	+
11. fully upright posture	—	—	—	—	+	—	—	+	+	+
12. long S-shaped neck	—	—	—	—	—	—	—	+	+	—
13. long hands	—	—	—	—	—	—	—	+	+	—
14. “bird-hipped” pelvis	—	—	—	—	—	—	—	—	+	+

Table 1. Distribution of selected character states among some living and extinct vertebrates.



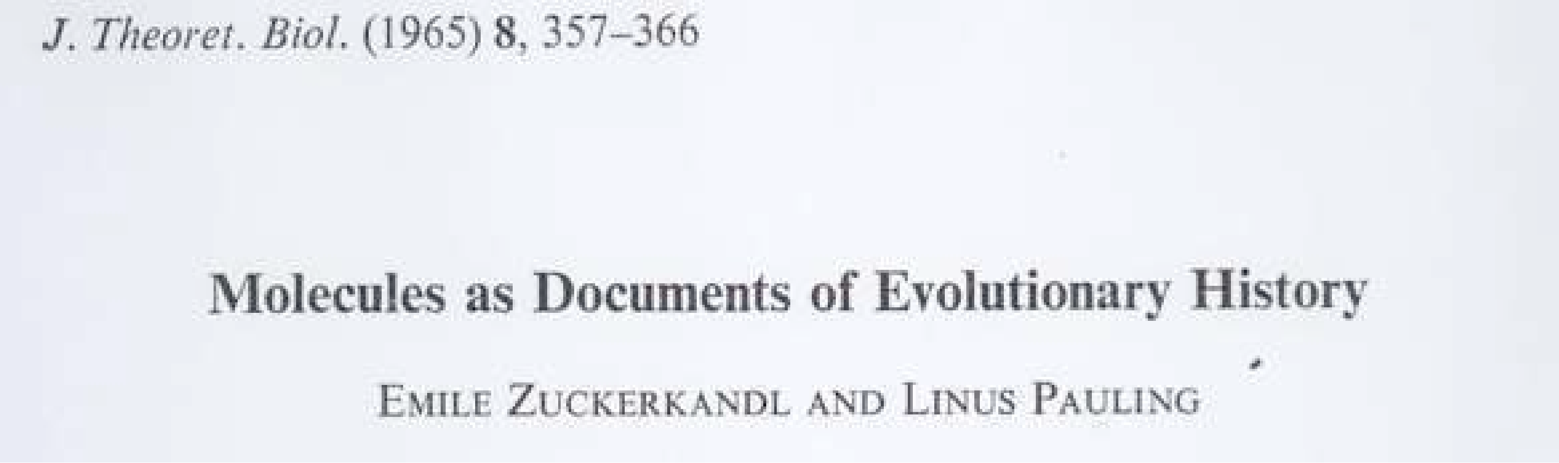
S.J. Carlson, 1999

Figure 4. Cladogram illustrating phylogenetic relationships among ten familiar kinds of animals, interrested in nine clades denoted by the nine nodes. Open circles denote living taxa; filled circles denote extinct taxa. The common names and Linnean binomials of the terminal taxa are: shark (*Carcharodon carcharias*); tuna (*Thunnus albacares*); frog (*Rana pipiens*); fin-backed pelycosaur (*Dimetrodon grandis*); human (*Homo sapiens*); whale (*Balaenoptera musculus*); crocodile (*Crocodylus acutus*); saurischian dinosaur (*Tyrannosaurus rex*); bird (*Melospiza melodea*); ornithischian dinosaur (*Triceratops horridus*). Redrawn from Carlson, 1995.

Phylogeny reconstruction based on molecular data

Zuckerkandl and Pauling (60s) first invoked the idea of using molecular data for reconstructing phylogenetic history.

They published a comparison of several species' hemoglobin fingerprints, observing that the level of dissimilarity of protein fingerprints corresponded roughly to the phylogenetic distance between source species

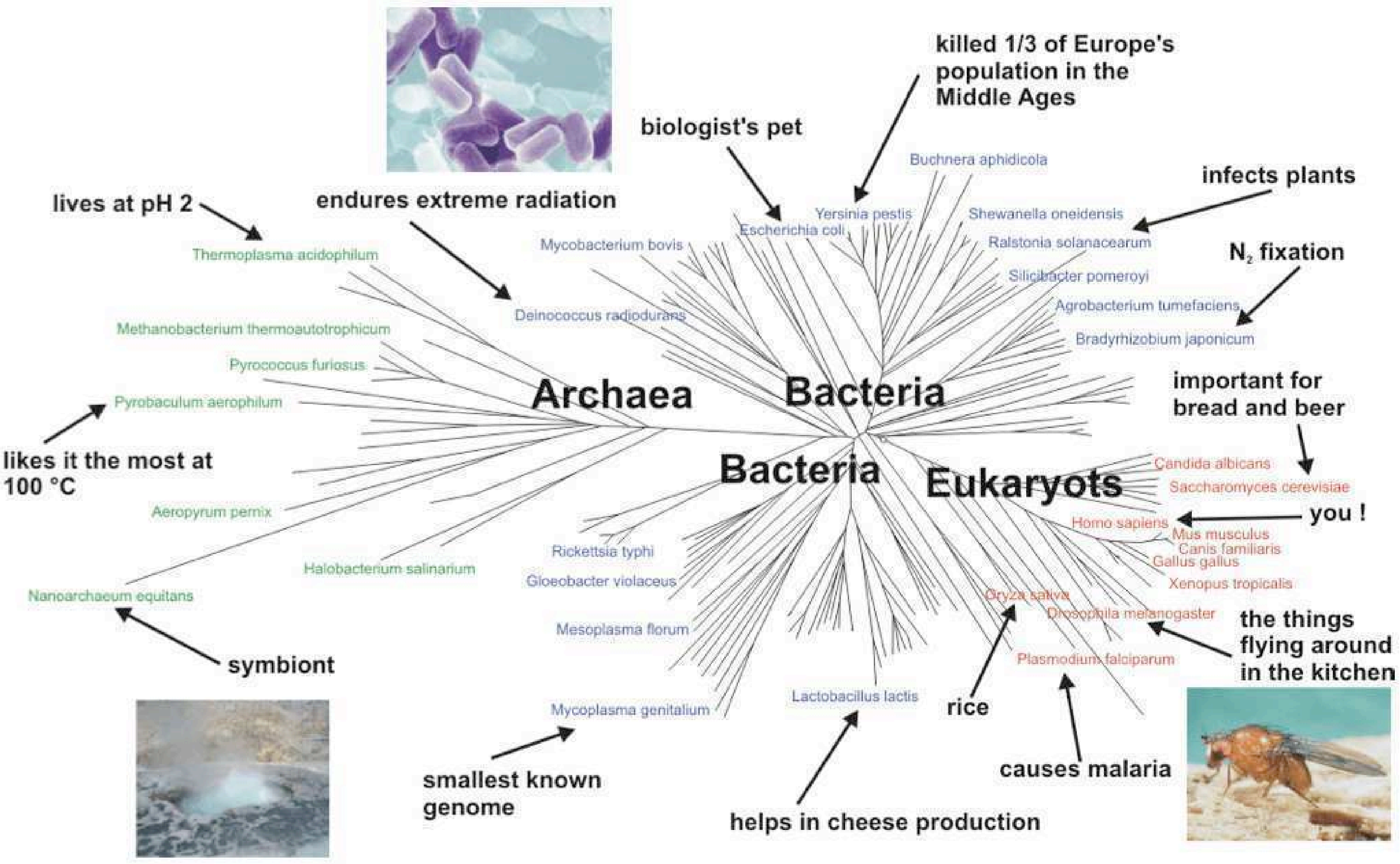


J. Theoret. Biol. (1965) 8, 357–366

Molecules as Documents of Evolutionary History

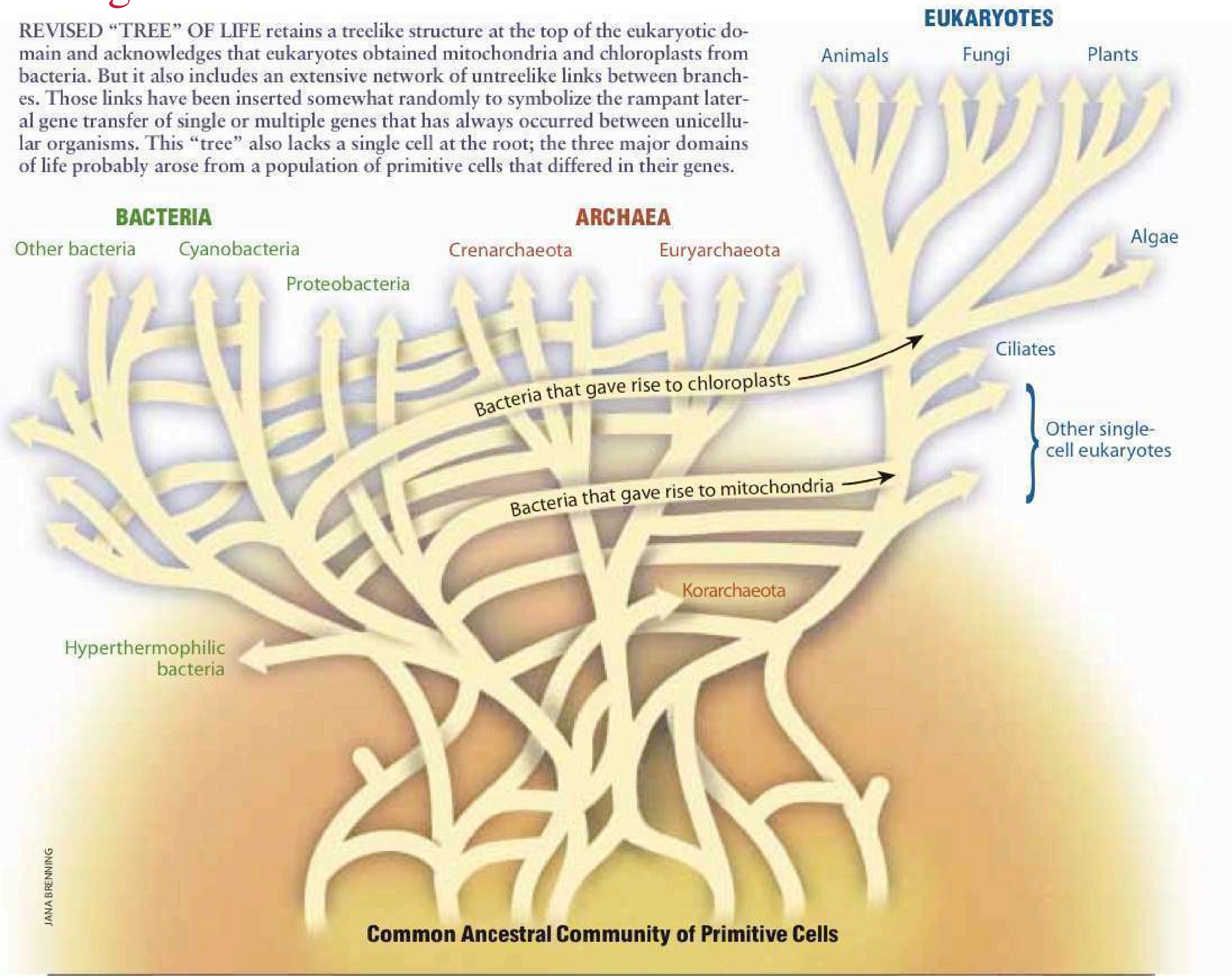
EMILE ZUCKERKANDL AND LINUS PAULING

Major methodological developments



All genomes, Wed Aug 3 11:11:38 2005, 197 species, 132924 orthologs, 11 used here, distance tree, ToCompl=36, Fit=0.4633; modified for educational purposes (drm)

But see Doolittle 2000 (*in Scientific American*) "Uprooting the tree of Life"



REVISED "TREE" OF LIFE retains a treelike structure at the top of the eukaryotic domain and acknowledges that eukaryotes obtained mitochondria and chloroplasts from bacteria. But it also includes an extensive network of untreelike links between branches. Those links have been inserted somewhat randomly to symbolize the rampant lateral gene transfer of single or multiple genes that has always occurred between unicellular organisms. This "tree" also lacks a single cell at the root; the three major domains of life probably arose from a population of primitive cells that differed in their genes.

EUKARYOTES
Animals
Fungi
Plants
Algae
Ciliates
Other single-cell eukaryotes
BACTERIA
Other bacteria
Cyanobacteria
Proteobacteria
ARCHAEA
Crenarchaeota
Euryarchaeota
Bacteria that gave rise to chloroplasts
Bacteria that gave rise to mitochondria
Korarchaeota
Hyperthermophilic bacteria
Common Ancestral Community of Primitive Cells

JANA BRENNING

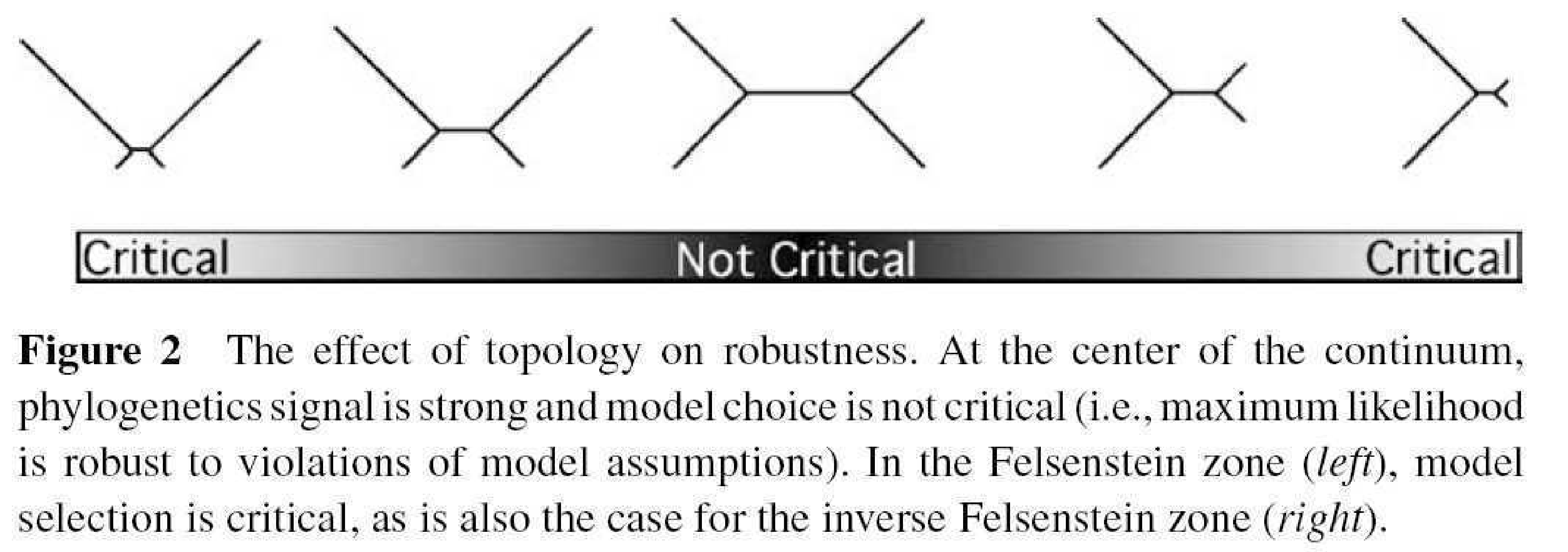


Figure 2 The effect of topology on robustness. At the center of the continuum, phylogenetics signal is strong and model choice is not critical (i.e., maximum likelihood is robust to violations of model assumptions). In the Felsenstein zone (*left*), model selection is critical, as is also the case for the inverse Felsenstein zone (*right*).

From Sullivan & Joyce (2005)

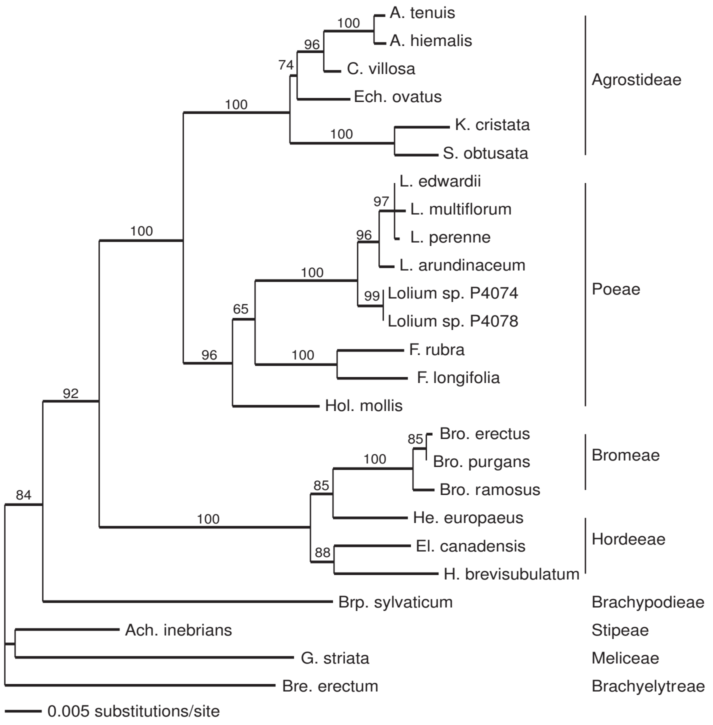


Figure 1: Parametric ML tree estimated from cpDNA intron and intergenic sequences. Numbers above branches indicate bootstrap support percentages (over 50%) obtained by 1000 maximum parsimony searches with branch swapping.

Applications of phylogeny

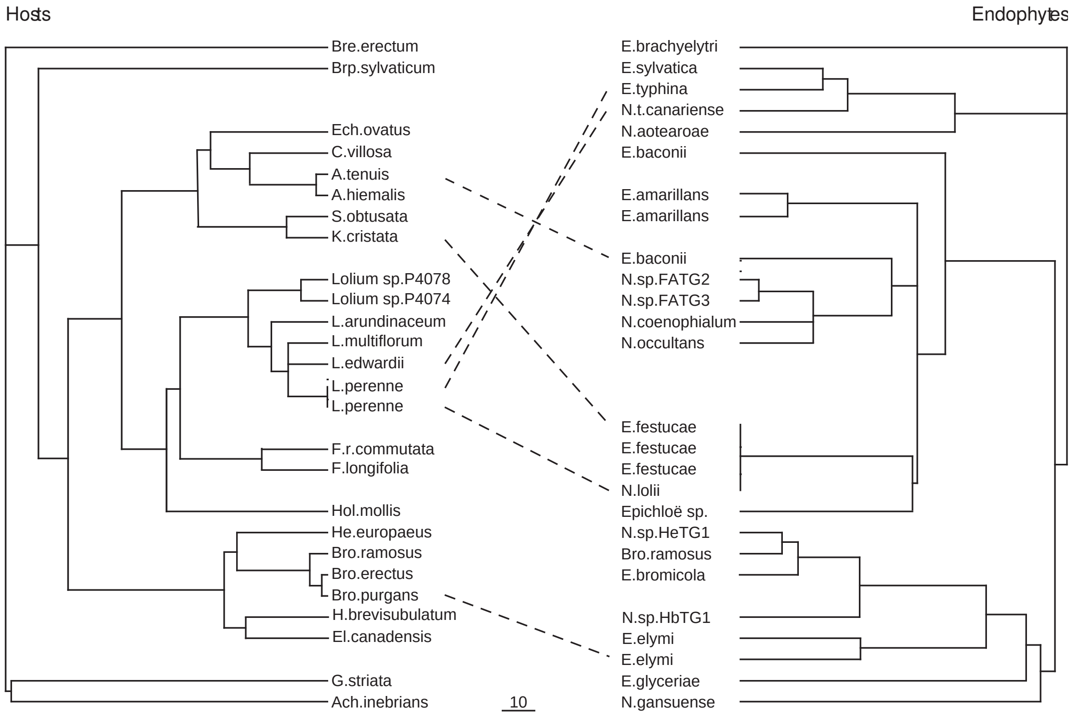


Figure 2: Ultrametric ML trees for host grasses and their endophytes. Hosts and their endophytes are indicated opposite each other or by connecting dashed lines.

We use statistical methods to reconstruct and to analyze a phylogenetic tree from DNA sequences. For example, to reconstruct a phylogenetic tree we use:

The maximum likelihood estimation (MLE) methods: These describe evolution in terms of a discrete-state continuous-time Markov process.

The Balanced Minimum Evolution (BME) method: This is a **distance based method** and **weighted Least Square method**. **The Neighbor-Joining (NJ) method** is a greedy algorithm of the BME method (Steel and Gasquel, 2008).

Bayesian inference for trees: Use Bayes Theorem and MCMC to estimate the posterior distribution rather than obtaining the point estimation.

Closeness

Tree uncertainty is a pervasive issue in phylogenetics.

To help cope with tree uncertainty, bootstrapping and Bayesian sampling methods provide a collection of possible trees instead of a single tree estimate.

Using bootstrapping or Bayesian sampling, one common practice is to identify highly supported tree features (e.g. splits) which occur in almost all the tree samples. **Highly supported features are regarded as likely features of the true tree.**

Similarly, in simulation studies it is common to judge reconstruction methods based on how **close** they get to the true tree.

This leads us to ask whether reconstruction accuracy (i.e. closeness to the true tree) can be improved, by attempting to directly optimize accuracy.

Bayes Estimator

Even though the true tree is unknown, we can still optimize reconstruction accuracy using a Bayesian approach.

In the Bayesian view, the true tree is a random variable T distributed according to the posterior distribution $P(T \mid D)$, where D is input data such as sequence data.

If $d()$ measures distance between trees, and T' is a tree estimate, then the expected distance between T' and the true tree is $\mathbb{E}_{T \sim P(T \mid D)} d(T, T')$.

Thus, to maximize reconstruction accuracy, we should choose our tree estimate to be $T^* = \text{argmin}_{T'} \mathbb{E}_{T \sim P(T \mid D)} d(T, T')$ where T^* is known as a **Bayes estimator**.

Tree Distances

Many popular distances between trees can be easily expressed as a squared euclidean distance, after embedding trees in an appropriately chosen vector space. Important examples include Robinson–Foulds distance (symmetric difference), quartet distance, and the squared path difference.

With Robinson–Foulds distance, Holder showed the Bayes estimator with the RF distance is the majority-rule consensus tree (Holder, Systematics Biology 2008).

We focus on squared euclidean distances, specifically the squared path difference.

Path Difference Metrics

The path difference metric is a **topological** distance, i.e. it only depends on the topologies of the tree.

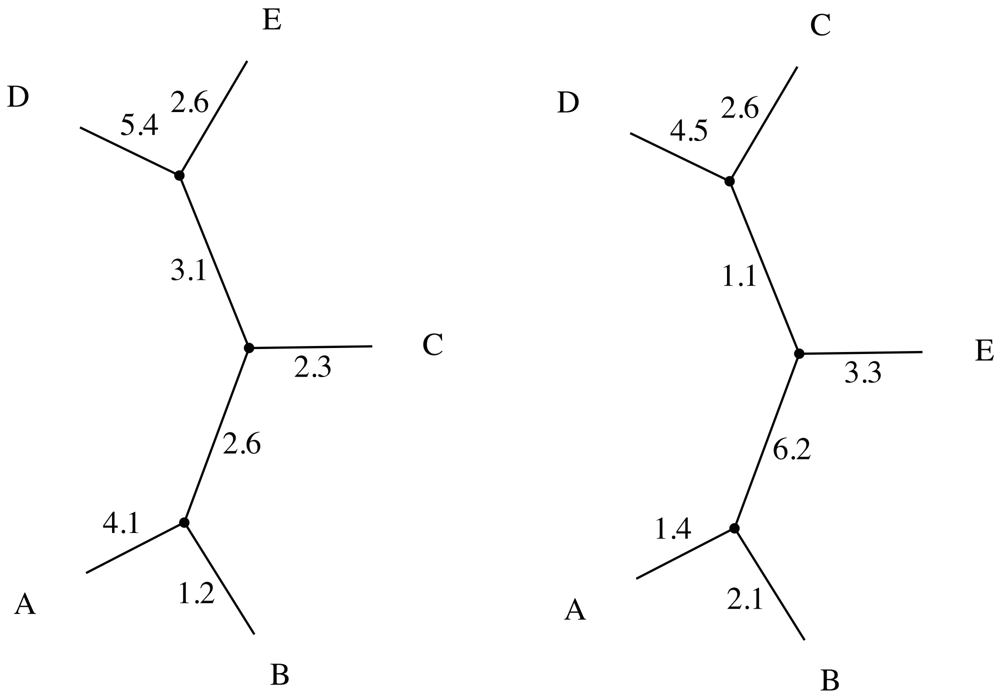
The path difference metric $v_p(T) \in \mathbb{R}^{\binom{n}{2}}$, where n is the number of taxa, is the integer vector whose ijth entry counts the number of edges between leaves i and j in the tree T.

The path difference metric was studied in (Steel and Penny, 1993).

The squared path difference is

$$d_p(T', T) = ||v_p(T) - v_p(T')||^2.$$

Example



Example

For the trees T_1 and T_2 in the previous figure, we have

$$v_p(T_1) = (2, 3, 4, 4, 3, 4, 4, 3, 3, 2),$$

$$v_p(T_2) = (2, 4, 4, 3, 4, 4, 3, 2, 3, 3),$$

and

$$d_p(T_1, T_2) = ||v_p(T_1) - v_p(T_2)||^2 = 6.$$

Here the coordinates of $v_p(T_1)$ and $v_p(T_2)$ are given by

$$\Big(v_{1,2}, v_{1,3,}, v_{1,4}, v_{2,3}, v_{2,4}, \ldots, v_{4,5}\Big),$$

where $v_{i,j}$ is the number of edges between leaf i and j.

Optimization

We want to find the optimal solution T^* such that

$$T^* = \text{argmin}_{T'} \mathbb{E}_{T \sim P(T \mid D)} d_p(T, T').$$

Since the number of tree topologies on n taxa grows exponentially in n, computing the Bayes estimator T^* under a general distance function can be computationally hard (NP hard).

Hill Climbing Techniques

However, hill climbing techniques such as those used in ML methods often work quite well in practice for tree reconstruction. Hill climbing techniques can similarly be used to find local minima of the empirical expected loss.

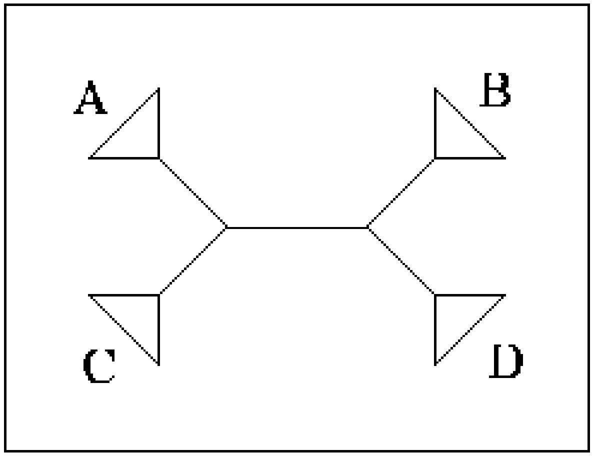
Hill climbing requires a way to move from one tree topology to another. Three types of combinatorial tree moves are often used for this purpose; *Nearest Neighbor Interchange (NNI)*, *Subtree-Prune-and-Regraft (SPR)*, and *Tree-Bisection-Reconnect (TBR)*.

Here we use NNI moves.

Nearest Neighbor Interchange

Here we assume unrooted trees (if we want to have a rooted tree we add an outgroup and make it as unrooted tree with $n+1$ taxa).

Suppose we have

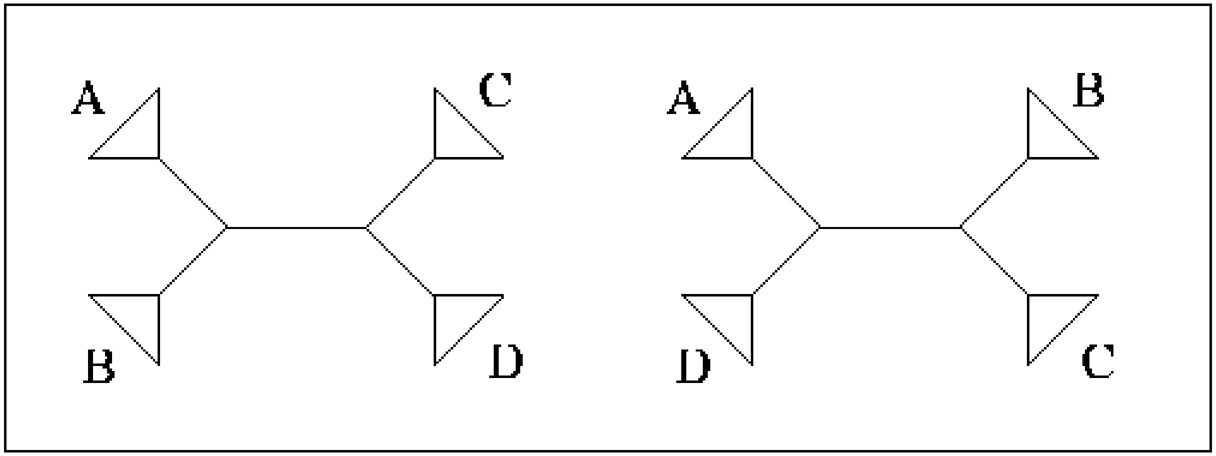


where A, B, C, D are clades

Nearest Neighbor Interchange

So we can think of this as a tree with 4 leaves and since there are only 3 tree topologies we swap this tree with one of the following two different tree topologies.

Suppose we have



Strict Hill Climbing algorithm

1. Choose a starting tree T'

2. Calculate $\mathbb{E}_{T \sim P(T \mid D)} d(T, T')$

3. Compute all possible NNI neighbor trees $T_1, \cdots T_{2\cdot(n-3)}$

4. Then do:

   (a) For $i = 1 \cdots 2 \cdot (n-3)$ compute $\mathbb{E}_{T \sim P(T \mid D)} d(T, T_i)$
   (b) Move to the tree with the lowest expectation.

5. Record the tree at each step

Simulations

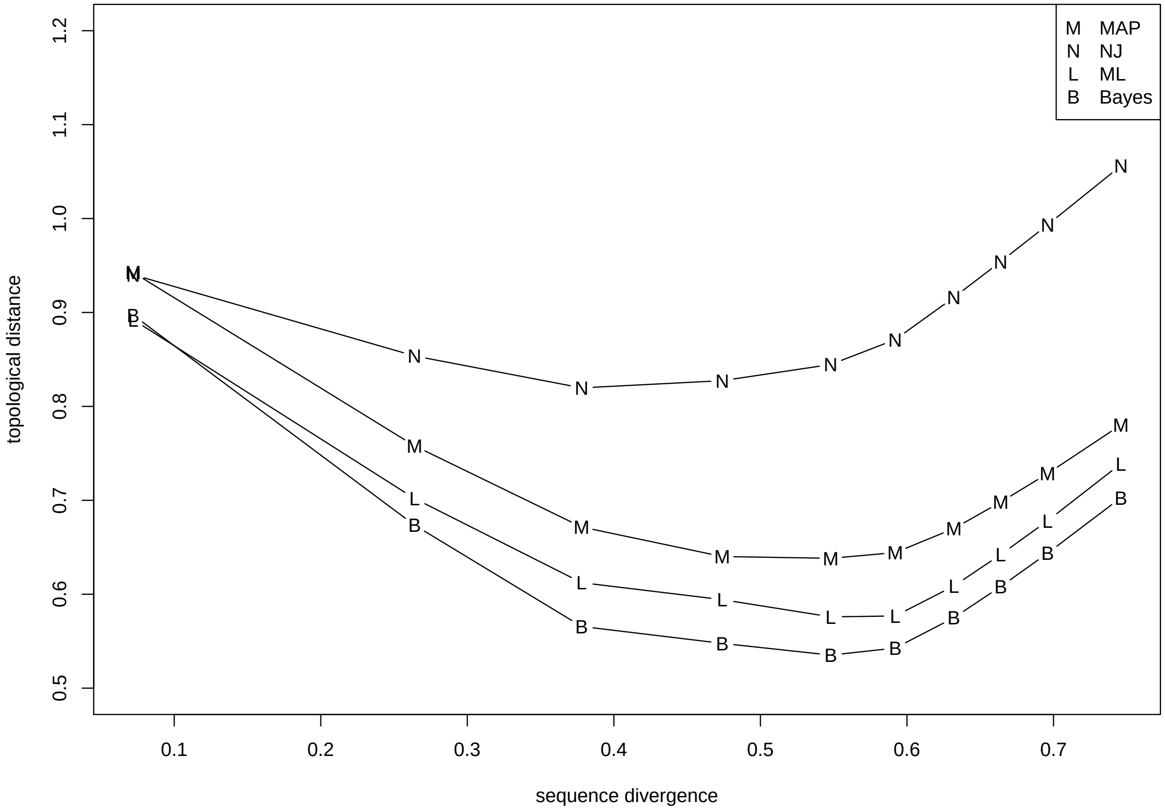
For simulated data, we used the first 1000 examples from the data set presented in (Guindon and Gasquel, 2003).

Trees on 40 taxa were generated according to a Markov process. For each generated tree, 40 homologous sequences (no indels) of length 500 were generated, under the Kumura two-parameter (K2P) model, with a transition/transversion ratio of 2.0. Specifically the Seq-Gen program was used to generate the sequences. The data is available from the website `http://www.atgc-montpellier.fr/phyml/datasets.php`.

For each set of homologous sequences D in the simulated data, we used the software `MrBayes` to obtain 15000 samples from the posterior distribution $P(T|D)$. Specifically, we ran `MrBayes` under the K2P model, discarded the initial 25% of samples as a burn-in, used a 50 generation sample rate, and ran for $1,000,000$ generations in total.

We computed a ML tree estimate for each data set, using the hill climbing software `PHYML` as described in the paper. We also computed a NJ tree using the software `PHYLIP`, using pairwise distances computed by `PHYLIP`.

Here NJ (N) is the neighbor joining tree constructed via `neighbor` in `PHYLIP` package, ML (L) is the `PHYML` tree, MAP (M) is the `MrBayes` sample with the highest posterior probability, and Bayes (B) is the Bayes Estimator tree, estimated from `MrBayes` samples.



Here NJ (N) is the neighbor joining tree constructed via `neighbor` in `PHYLIP` package, ML (L) is the `PHYML` tree, MAP (M) is the `MrBayes` sample with the highest posterior probability, and Bayes (B) is the Bayes Estimator tree, estimated from `MrBayes` samples.

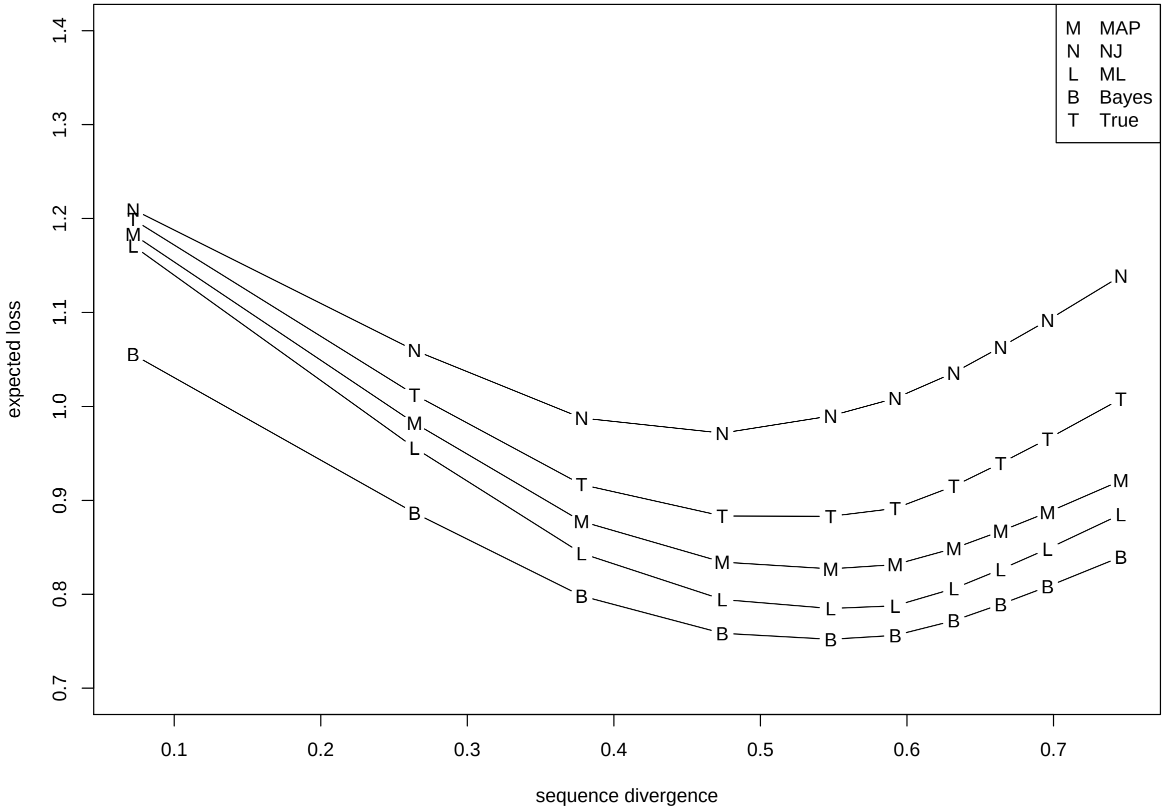


Table 1: We give the performance of hill climbing, applied to several different initial trees. The first two columns summarize how the local minimum compared to the initial tree, on the 1000 simulated data sets. The third column gives the average percentage by which hill climbing decreases the path difference distance to the true tree. This is computed as $1-$ mean$(d_{initial}/d_{final})$, where mean() denotes denotes geometric mean. If either the initial or final distance to the true tree is zero, we add 1 to both distances.

Initial tree	Hill climbing improves distance to T^{true}?	Hill climbing worsens distance to T^{true}?	Avg drop in distance to T^{true}
ML tree	380	253	5.9%
Empirical MAP tree	508	185	17.9%
NJ tree	693	229	39.6%

Table 2: The first two columns summarize how the local minimum compared to the initial tree, on the 1000 simulated data sets. The third column gives the average percentage by which hill climbing decreases $\hat{\rho}_p$. This is computed as $1-$ mean($\sqrt{\hat{\rho}_p^{initial}/\hat{\rho}_p^{final}}$), where mean() denotes denotes geometric mean. If either $\hat{\rho}_p^{initial}$ or $\hat{\rho}_p^{final}$ is zero, we add 1 to both.

Initial tree	Hill climbing improves	Hill climbing worsens	Avg drop in
	$\hat{\rho}_p$?	$\hat{\rho}_p$?	$\hat{\rho}_p$
ML tree	690	0	5.9%
Empirical MAP tree	870	0	8.6%
NJ tree	961	0	20.3%

UK Statistical Phylogenetics Group

The group of biologists, computer scientists, and statisticians at UK:

`http://www.cophylogeny.net/`

Why interdisciplinary and we have to collaborate?

UK Statistical Phylogenetics Group: The group of biologists, computer scientists, and statisticians at UK:

Problems in Molecular Evolution arise from the area of Biology.

To analyze huge/messy data sets we need to use statistical methods.

To compute huge/messy data sets using statistics we need algorithms to compute them efficiently.

Therefore we need to know how to collaborate with each other and we have to know a little bit from each area.

GOAL: Learn how to communicate and collaborate with people from other area.

Thank you....

The ms is available at `http://arxiv.org/abs/0911.0645` and the source code, written in java, is available at `http://cophylogeny.net/research.php`

Supported by NIH 1R01GM086888-01