

Inferring phylogenetic trees

László Kozma
Lkozma@cc.hut.fi

Helsinki University of Technology
T-61.6070 Special course in bioinformatics: Modeling of proteomics data

2. 4. 2008

References

- ① Notes on phylogenetics, School of Computer Science, Tel-Aviv University
<http://www.cs.tau.ac.il/~rshamir/algmb/00/scribe00/html/lec08/lec08.html>
- ② Kim, Warnow: Tutorial on Phylogenetic Tree Estimation
<http://kim.bio.upenn.edu/~jkim/media/ISMBtutorial.pdf>
- ③ Lecture notes on phylogenetic tree construction, Frank Olken, Berkeley PGA
<http://pga.lbl.gov/Workshop/April2002/lectures/Olken.pdf>
- ④ L. A. Salter, Algorithms for Phylogenetic Tree Reconstruction (2000)
- ⑤ J. Felsenstein: Inferring Phylogenies (2003, Sinauer Associates)
- ⑥ Google, Wikipedia, ...

Phylogenetics

- Fundamental idea in biology:
All organisms living on Earth today share common ancestors (P.-L. de Maupertuis 1745; I. Kant 1790; Ch. Darwin, 1859)
- Evolution happened through:
 - **mutation**: DNA sequence changed due to single base changes, deletion/insertion
 - **selection bias**
 - **speciation**: physical separation into groups
- Phylogenetics: study the evolutionary relations between organisms
- Goal: recover order of speciation (evolution history)
- Evolution history can be (roughly) depicted as a tree

Phylogenetics

- **Taxon**: organism, species, gene, protein, etc.
- Infer tree from observed features of taxa:
 - morphological characters (used in classical taxonomies): number of legs, size, color, swims/flies, etc.
 - DNA sequences (alignment of selected sequences that are conserved in many organisms, example: Ribosomal RNA 16S)
 - protein sequences
 - metabolic pathways data, etc.
- Often different data produce similar trees (despite the fact that most genetic variation does not cause different external morphology) $\Rightarrow$ *indirect evidence for evolution*

Phylogenetic trees

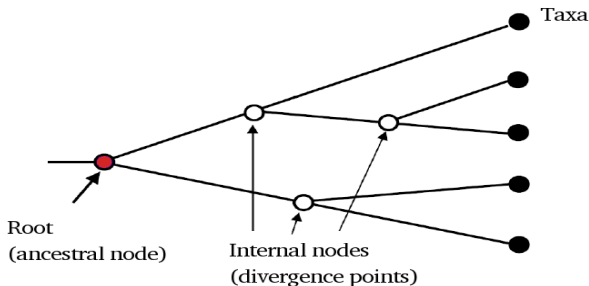


Figure: Phylogenetic tree.

- leaves: taxa, species observable today
- internal nodes: hypothetical common ancestors

Phylogenetic trees

- Evidence of extinct common ancestors:
 - fossil data:
 - rarely available
 - mostly recent
 - unreliable
 - Recapitulation theory (Ernst Haeckel, 19th century):
"ontogeny recapitulates phylogeny"
- Only data from current taxa used to infer tree
- Other data can still be used for verifying predictions

Phylogenetics

Why do phylogenetics:

- "Tree of Life"
- understand lineage of species
- understand how a function evolved
- design drugs, vaccines
- assist epidemiology
- study of relationships between parasite/host

theory applicable to other fields, for ex. natural languages

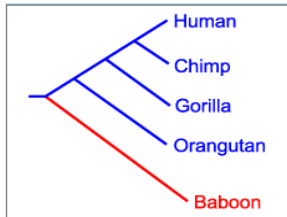
Phylogenetics

Example applications:

- Human, Chimpanzee, Gorilla tree
- analysis of mitochondrial DNA of 182 people, one common ancestor found who lived in Africa
- HIV origins
- Influenza virus phylogeny, matching frozen ancient samples

Phylogenetic trees

Rooted tree



Unrooted tree

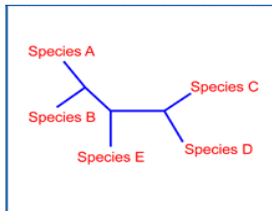


Figure: Rooted/Unrooted tree.

- In rooted tree one node given special significance (common ancestor)
- Can be easily picked if **molecular clock hypothesis** holds
- Edge length can indicate genetic distance (=? time)

Phylogenetic trees

Difficulties with data:

- Noise
- Hybridisation between species that were not nearest neighbors
- Recombination
- Convergent evolution
- Gene duplication, horizontal gene transfer
- Conserved sequences
- Extinct species

Phylogenetic trees

Methods for inferring the tree (binary/bifurcating):

- Distance-based methods
 - UPGMA (agglomerative clustering)
 - Neighbor Joining
- Character-based methods
 - Maximum Parsimony
 - Maximum Likelihood
 - Bayesian methods, etc.

Phylogenetic trees

Inferring phylogenetic trees is computationally expensive:

- Most of the criteria lead to NP-hard problems, therefore heuristics/approximations are used.
- Optimal sequence alignment itself NP-hard
- Exhaustive search:

$$\frac{(2n-3)!}{2^{n-2}(n-2)!} \text{ rooted}$$

$$\frac{(2n-5)!}{2^{n-3}(n-3)!} \text{ unrooted}$$

trees are possible with n given leaves.

Distance-based methods

Align relevant parts of DNA sequences:

Species A: ..ATGCTCAGT..

Species B: ..ATGCCCAGG..

Species C: ..ATGCTCAGC..

Species D: ..CGCCTCAGA..

Species E: ..GTGCTCAGT..

Similarity matrix:

	A	B	C	D	E
A	-	0.3	0.4	0.55	0.7
B		-	0.24	0.45	0.5
C			-	0.4	0.2
D				-	0.5
E					-

Distance-based methods

- Usual axioms for distance metrics must hold
- Possible distance measure: observed percent difference
- Extension: weighted using different transition probabilities
- Similar approaches for protein sequences
- Other distance measures: edit distance, more complex measures

Distance-based methods

Unweighted Pair Group Method with Arithmetic Mean (UPGMA):

- let d_{ij} be the distance between two nodes
- C_i a cluster of nodes, with n_i elements
- the distance between two clusters C_i, C_j is:

$$D_{ij} = \frac{1}{n_i n_j} \sum_{p \in C_i, q \in C_j} d_{pq} \quad (1)$$

UPGMA

Steps for building the tree:

- 1 Construct distance matrix between taxa.
- 2 Add each taxa to the tree as a leaf.
- 3 Assign each taxa to its own cluster.
- 4 Find the pair of clusters with the shortest distance D_{ij} .
- 5 In the matrix replace clusters i and j with new cluster k .
- 6 Calculate distances from k to all other clusters as
$$D_{kl} = \frac{n_i D_{il} + n_j D_{jl}}{n_i + n_j}$$
- 7 Add branches in the tree from k to i and from k to j such that distance to leaves is $D_{ij}/2$.
- 8 Repeat from 4. until only root node left.

UPGMA

Example

	A	B	C	D	E	F
A						
B	2					
C	4	4				
D	6	6	6			
E	6	6	6	4		
F	8	8	8	8	8	

UPGMA

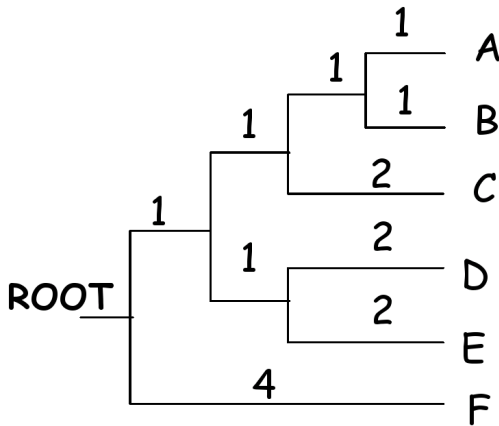


Figure: Resulting tree.

Distance-based methods

UPGMA

- Simple to implement, fast
- Not very accurate
- Estimates branch length
- Assumes that rate of evolution is constant
- Assumes distances are additive
- No statistical evolutionary model
- Does not give quality of tree

Character-based methods

- **Input:** n species, m characters/features (an $n \times m$ matrix)
- Can be the aligned DNA/protein sequences as seen before
- **Output:** A phylogenetic tree that maximizes some target function
- Possible other input: relative importance (weight) of characters, distribution of mutations, etc.
- We make two rather strong assumptions:
 - Characters mutually independent
 - After two species diverge, they continue to evolve independently

Character-based methods

Maximum Parsimony

- Predicts evolutionary tree with minimum number of evolutionary steps required to generate observed variation
- Parsimony, Occam's Razor, MDL, "most defensible theory, least number of assumptions"
- Not necessarily meaningful biologically
- Naive algorithm:
 - Construct all possible trees
 - Assign scores to all trees
 - Choose tree with best score

- Parsimony score:
$$S(T) = \sum_{(v,u) \in E(T)} |\{j : v_j \neq u_j\}|$$

Parsimony

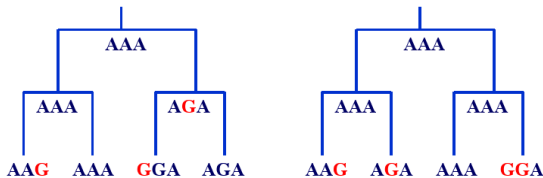


Figure: Two possible trees.

Parsimony suggests that the first tree is better, as it has fewer mutations.

Parsimony

Parsimony algorithms have two components:

- ① Search through the space of trees
 - ② Assign minimum number of changes to a given tree topology
- First problem NP-complete:
 - use heuristics, hill-climbing, local search
 - branch-and-bound
 - Second problem: Fitch's algorithm
 - any state can convert to any state
 - locations are uniform
 - each type of transition has same cost
 - Weighted variants exist

Parsimony

Fitch's algorithm:

- ➊ Traverse tree from leaves to root, determining possible states for each internal node
- ➋ Traverse tree from root to leaves, picking ancestral states for internal nodes

Run separately for each character location.

Parsimony

- 1 post-order traversal from leaves to root:

Possible states R_i of node i with children j and k :

$$R_i = R_j \cup R_k, \text{ if } R_j \cap R_k = \emptyset$$

$$R_i = R_j \cap R_k, \text{ otherwise}$$

- 2 pre-order traversal from root to leaves:

State r_j of node j with parent i :

$$r_j = r_i, \text{ if } r_i \in R_j$$

$$\text{arbitrary state} \in R_j, \text{ otherwise}$$

Parsimony

Example:

Aardvark	CAGGTA
Bison	CAGACA
Chimp	CGGGTA
Dog	TGCACT
Elephant	TGCGTA

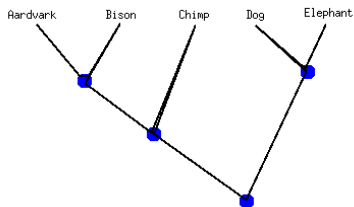


Figure: Resulting tree.

Parsimony

- Advantages:
 - Models evolutionary history
 - Gives score of tree
- Disadvantages:
 - Does not estimate branch length.
 - Computationally expensive
 - Biased tree under some conditions.

Maximum Likelihood

Algorithm for finding best tree:

- for each possible tree
 - for each character location in the alignment
 - compute the likelihood of the tree, given data and evolution model

Choose tree that has highest likelihood.

$$L(\text{tree}|\text{data}) \propto P(\text{data}|\text{tree})$$

explanation from:

<http://noble.gs.washington.edu/noble/genome373/lectures>

Maximum Likelihood

Maximum Likelihood

Questions left open:

- Heuristics to speed up the algorithm...
- How to find optimal branch lengths?

Maximum Likelihood

- Advantages:
 - Takes evolutionary model into account
 - More accurate than the other methods
 - All the sequence information is used
 - Evaluates all possible trees
- Disadvantages:
 - Very slow.
 - Impractical for analyzing large data sets.

Summary

- What is phylogenetics?
- What are phylogenetic trees?
- Information used for constructing phylogenetic trees
- Methods for building phylogenetic trees
 - Distance based: UPGMA
 - Character based: Parsimony, Maximum Likelihood