Drzewa filogenetyczne

Bartek Wilczyński

19. marca 2019

How sequences evolve?

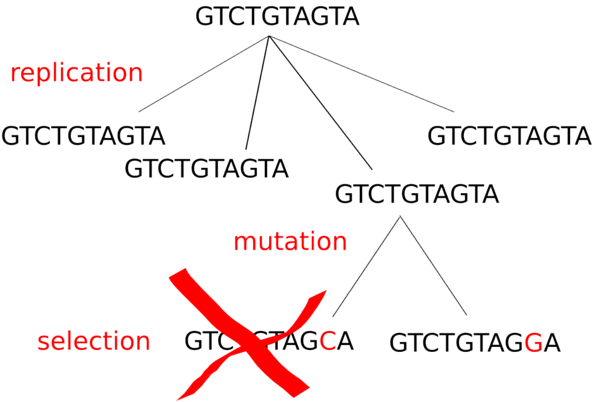


image (c) BW

Tree of life?

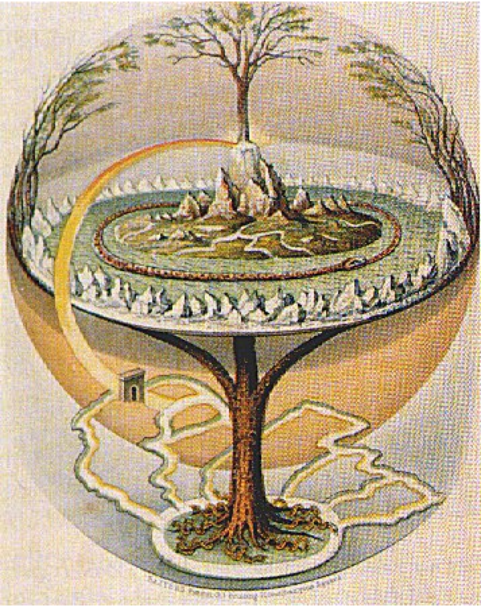


image (c) Wikimedia

Molecular tree of life

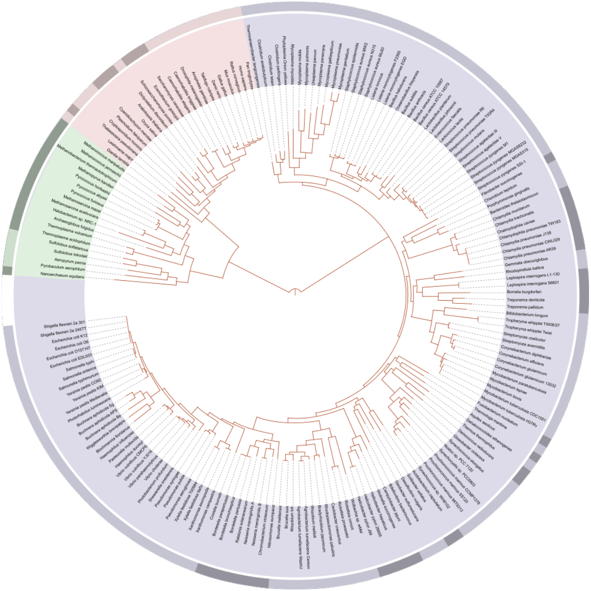


image (c) I. Letunic – itol.org

evolutionary distance vs. similarity

- We are interested in measuring evolutionary distances by looking at molecular sequences
- We expect *distances* to *grow* with **decreasing similarity**
- The sequence alignment problem allows us to find the optimal alignment, however the score of the alignment is a measure of *similarity*, rather than distance
- problem of *maximizing similarity* is similar to *minimizing distance*
- However:
  - We expect $d(x, x) = 0$ while for most a, b , $sim(a, a) \neq sim(b, b)$
  - Distances have triangle inequality, and similarities not

Bifurcating vs. multifurcating trees

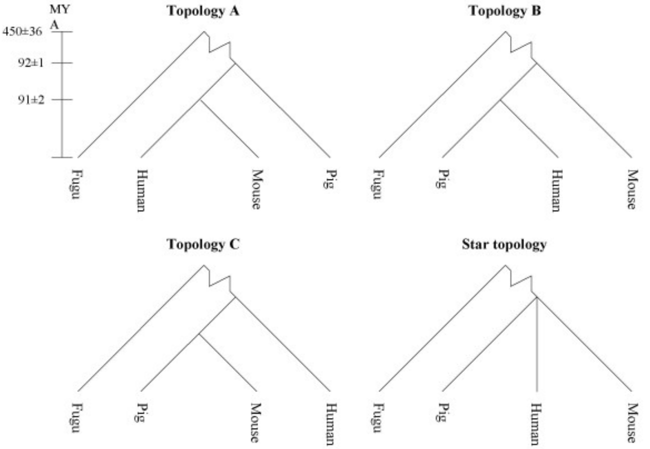
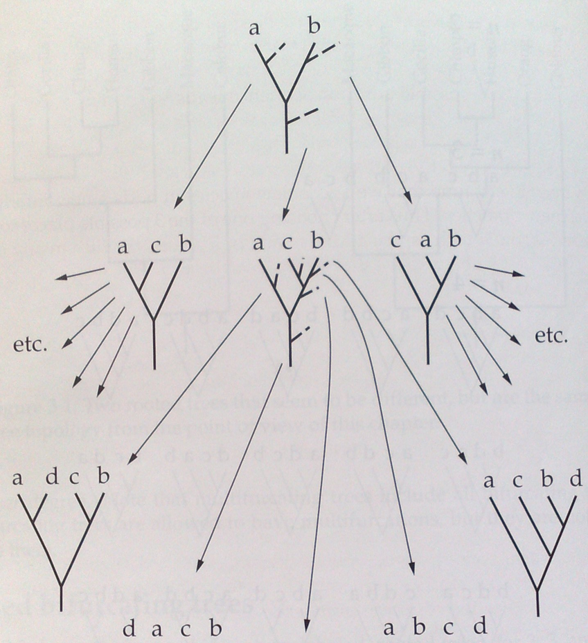
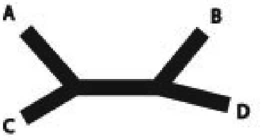


image (c) Jorgensen et al. 2005

How many binary trees are there



Rooted vs. unrooted trees



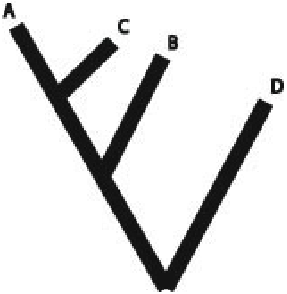
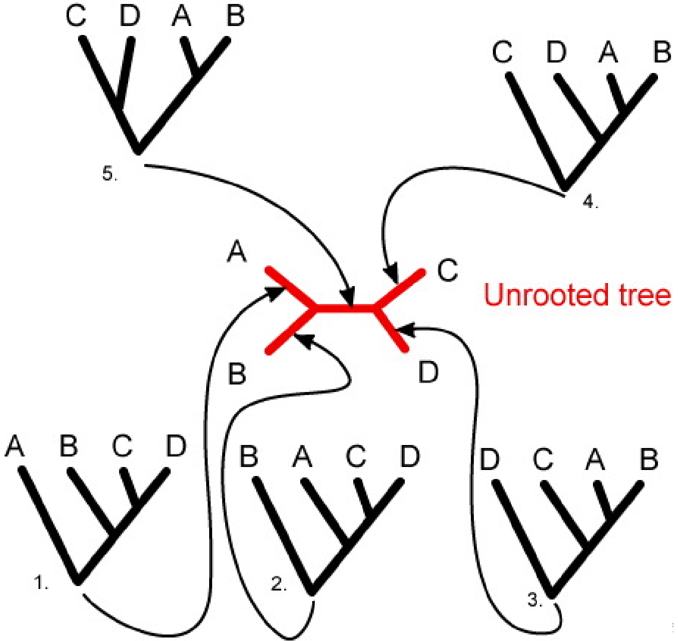


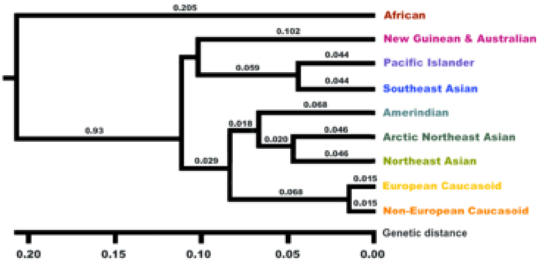
image (c) embl.org

Rooting an unrooted tree

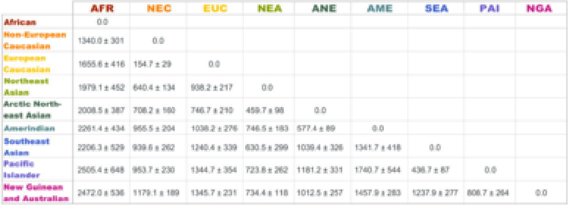


Trees vs. distance matrices

Linkage tree for 9 population clusters showing genetic distances (F_{ST})
(Cavalli-Sforza *et al.*, 1994:80)



F_{ST} distance matrix for the 9 clusters shown above
(x10,000 with standard errors obtained by bootstrap analysis)



	AFR	NEC	EUC	NEA	ANE	AME	SEA	PAI	NGA
African	0.0								
Non-European Caucasian	1340.0 ± 301	0.0							
European Caucasian	1655.6 ± 416	154.7 ± 29	0.0						
Northeast Asian	1979.1 ± 452	640.4 ± 134	938.2 ± 217	0.0					
Arctic Northeast Asian	2008.5 ± 387	708.2 ± 160	746.7 ± 210	459.7 ± 98	0.0				
Amerindian	2261.4 ± 434	955.5 ± 204	1038.2 ± 276	746.5 ± 183	577.4 ± 89	0.0			
Southeast Asian	2206.3 ± 529	939.6 ± 262	1240.4 ± 339	630.5 ± 299	1039.4 ± 326	1341.7 ± 418	0.0		
Pacific Islander	2505.4 ± 648	953.7 ± 230	1344.7 ± 354	723.8 ± 262	1181.2 ± 331	1740.7 ± 544	436.7 ± 87	0.0	
New Guinean and Australian	2472.0 ± 536	1179.1 ± 189	1345.7 ± 231	734.4 ± 118	1012.5 ± 257	1457.9 ± 283	1237.9 ± 277	808.7 ± 264	0.0

image (c) Cavalli-Sforza 1994

Finding an optimal tree

- Given a tree with branch lengths T, we can easily generate distance matrix d_{ij}
- Can we solve the reverse problem, and how does it relate to the original problem?
- Formally, for a given distance matrix D, we want to find a labelled tree T, optimizing the least squares criterion:

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (D_{ij} - d_{ij})^2$$

- In general, it is equivalent to solving the Steiner tree problem – one of the the original NP-complete problems
- Can we find any approximate or specialized solutions?

ultrametric trees: “Evolutionary clock”

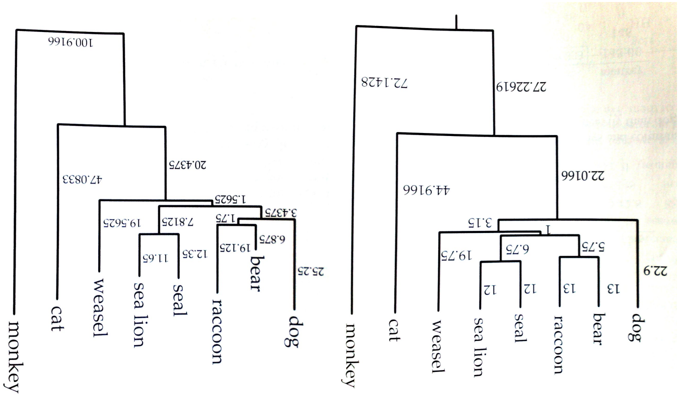


image (c) J. Felsenstein

ultrametric trees: properties

metric requirements

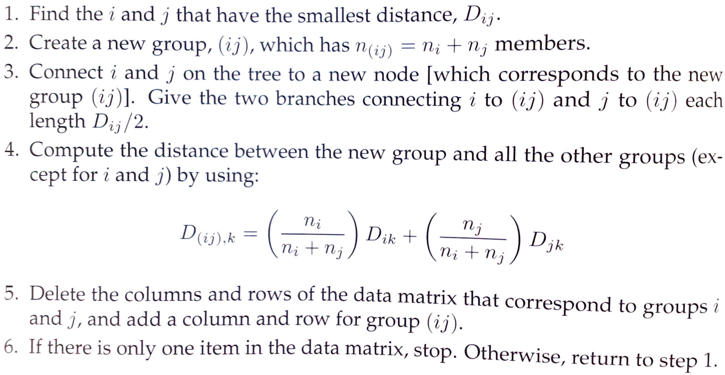
$$\begin{array}{rcll} d(x,y) & > & 0 & \mathit{for}\ x \neq y \\ d(x,y) & = & 0 & \mathit{for}\ x = y \\ d(x,y) & = & d(y,x) & \forall\ x,y \\ d(x,y) & \leq & d(x,z) + d(y,z) & \forall\ x,y,z \quad \text{(triangle inequality)} \end{array}$$

ultrametric – any three nodes can be relabelled so, that

$$d(x,y) \leq d(x,z) = d(y,z)$$

If you have a distance matrix induced from a tree, is it ultrametric?

Greedy approach 1 – hierarchical clustering



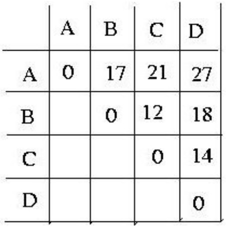
1. Find the i and j that have the smallest distance, D_{ij}.
2. Create a new group, (ij), which has $n_{(ij)} = n_i + n_j$ members.
3. Connect i and j on the tree to a new node [which corresponds to the new group (ij)]. Give the two branches connecting i to (ij) and j to (ij) each length $D_{ij}/2$.
4. Compute the distance between the new group and all the other groups (except for i and j) by using:

$$D_{(ij),k} = \left(\frac{n_i}{n_i + n_j}\right) D_{ik} + \left(\frac{n_j}{n_i + n_j}\right) D_{jk}$$

5. Delete the columns and rows of the data matrix that correspond to groups i and j, and add a column and row for group (ij).
6. If there is only one item in the data matrix, stop. Otherwise, return to step 1.

image (c) J. Felsenstein

Greedy approach 1 – example



	A	B	C	D
A	0	17	21	27
B		0	12	18
C			0	14
D				0

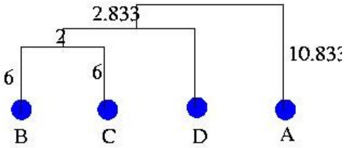
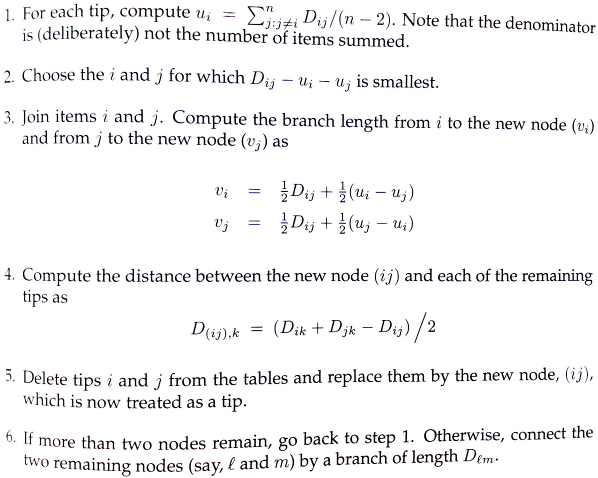


image (c) P. Winter

Greedy approach 2 – Neighbor-joining



1. For each tip, compute $u_i = \sum_{j:j \neq i}^{n} D_{ij}/(n-2)$. Note that the denominator is (deliberately) not the number of items summed.
2. Choose the i and j for which $D_{ij} - u_i - u_j$ is smallest.
3. Join items i and j. Compute the branch length from i to the new node (v_i) and from j to the new node (v_j) as

$$
\begin{aligned}
v_i &= \tfrac{1}{2}D_{ij} + \tfrac{1}{2}(u_i - u_j) \\
v_j &= \tfrac{1}{2}D_{ij} + \tfrac{1}{2}(u_j - u_i)
\end{aligned}
$$

4. Compute the distance between the new node (ij) and each of the remaining tips as

$$D_{(ij),k} = (D_{ik} + D_{jk} - D_{ij}) \Big/ 2$$

5. Delete tips i and j from the tables and replace them by the new node, (ij), which is now treated as a tip.
6. If more than two nodes remain, go back to step 1. Otherwise, connect the two remaining nodes (say, ℓ and m) by a branch of length $D_{\ell m}$.

image (c) J. Felsenstein

Properties of Neighbor-Joining

- The same complexity as average linkage hierarchical clustering $\mathcal{O}(n^3)$
- Guaranteed to return the correct answer if the distance matrix D originates from a tree
- Works also for non-ultrametric trees

Further links

- More information: *Inferring phylogenies* J. Felsenstein
- More advanced methods based on probabilistic approaches (Maximum likelihood, Bayesian approaches)
- Tree reconstruction might give different results for different genes, we will discuss this issue later
- Pairwise distances might lead to "unrealistic" phylogenies, We will discuss this problem next week.