

January 22, 2008

CS 395T: Computational Phylogenetics

Today

Review of last week

- representations of trees: distance-matrices, clades, splits (bipartitions), and quartets
- Computing trees from dissimilarity matrices: the “naïve” quartet method

Construction from data

- Connections to estimation of phylogenies from empirical data
- Identifiability of the tree under Markov models of evolution

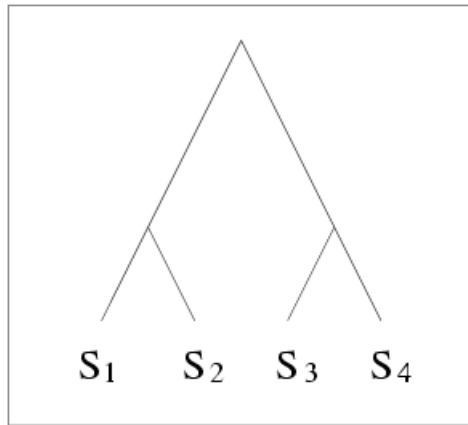
Computing trees

- Given $Q(T)$ (the quartet subtrees of T), can we determine T ?
- Given $C(T)$ (the bipartitions of S defined by the edges of T), can we determine T ?
- Given $\text{Clades}(T)$ (the sets of leaves defined by internal nodes in the rooted tree T), can we determine T ?

So?

- We can compute a tree from its set of clades, bipartitions, or quartets. But how do we get these sets?
- Primary data are generally characters (columns within alignments of biomolecular sequences, morphological features, or other such features). These don't directly produce these sets.
- But often evolutionary biologists have techniques for estimating “evolutionary distances” between taxa, and these are then useful for computing these sets!

Distance-based Phylogenetic Methods

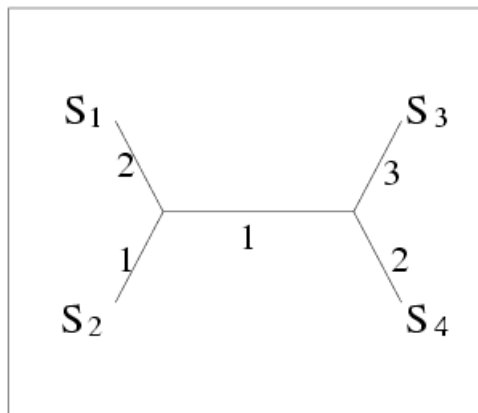


TRUE TREE

S₁ ACAATTAGAAC
S₂ ACCCTTAGAAC
S₃ ACCATTCCAAC
S₄ ACCAGACCAAC

DNA SEQUENCES

STATISTICAL
ESTIMATION
OF PAIRWISE
DISTANCES



INFERRED TREE

METHODS
SUCH AS
NEIGHBOR
JOINING



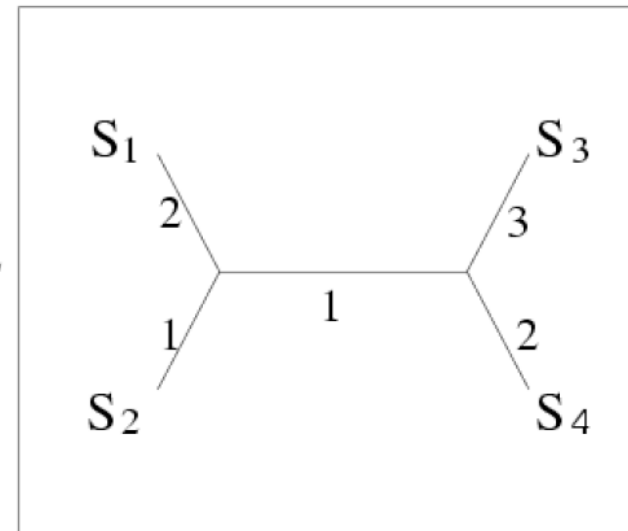
	S ₁	S ₂	S ₃	S ₄
S ₁	0	3	6	5
S ₂		0	5	4
S ₃			0	5
S ₄				0

DISTANCE MATRIX

Additive Distance Matrices

	S_1	S_2	S_3	S_4
S_1	0	3	6	5
S_2		0	5	4
S_3			0	5
S_4				0

POLYTIME
INVERTIBLE



Four-point condition

- A matrix D is additive if and only if for every four indices i, j, k, l , the maximum and median of the three pairwise sums are identical

$$D_{ij} + D_{kl} < D_{ik} + D_{jl} = D_{il} + D_{jk}$$

Four-point condition

- A matrix D is additive if and only if for every four indices i, j, k, l , the maximum and median of the three pairwise sums are identical

$$D_{ij} + D_{kl} < D_{ik} + D_{jl} = D_{il} + D_{jk}$$

The Four-Point Method computes trees on quartets using the Four-point condition

Four-point Method

- Given distance matrix $[D_{xy}]$ and four taxa, s_i, s_j, s_k, s_l , return tree $(s_i, s_j), (s_k, s_l)$ iff $D_{ij} + D_{kl}$ is the minimum of the three pairwise sums

Computing a tree from an additive distance matrix

- Compute the tree on each quartet using the four-point condition, thus producing $Q(T)$.
- Compute the tree T , using the quartet-based method given earlier:
 - Find a sibling pair A, B
 - Recurse on $S - \{A\}$, producing tree T'
 - Insert A into T' by making A sibling to B , and return the tree

But what about real data?

- Distances calculated for real data are rarely (basically never) additive. In fact, real data produce “dissimilarity matrices”, which may not satisfy the triangle inequality!
- What can we say about this quartet-based method, when the distance matrix is not additive?

Naïve Quartet Method

- Compute the tree on each quartet using the four-point condition
- Merge them into a tree on the entire set if they are compatible:
 - Find a sibling pair A,B
 - Recurse on $S-\{A\}$
 - If $S-\{A\}$ has a tree T, insert A into T by making A a sibling to B, and return the tree

Performance of the Naïve Quartet Method?

- How much error can it handle? (How far from the “true” additive distance matrix can the input dissimilarity matrix be?)
- Running time?