

Additive distances

Let T be a tree on leaf set S and let $w : E \rightarrow R^+$ be an edge-weighting of T , and assume T has no nodes of degree two.

Let $D_{ij} = \sum_{e \in P_{ij}} w(e)$, where P_{ij} is the path in T from i to j . Then the matrix $[D_{ij}]$ is said to be *additive*.

Four Point Condition: If $[D_{ij}]$ is an additive matrix, then for all i, j, k, l , the median and maximum of the three pairwise sums are identical:

$$D_{ij} + D_{kl}$$

$$D_{ik} + D_{jl}$$

$$D_{il} + D_{jk}$$

Four Point Method: Given a dissimilarity matrix d on four indices i, j, k, l , return the tree $ij|kl$ if and only if $d_{ij} + d_{kl} \leq \min\{d_{ik} + d_{jl}, d_{il} + d_{jk}\}$.

Error Tolerance of NQM

Let $[D_{ij}]$ be an $n \times n$ additive matrix for a tree (T, w) and let $f = \min\{w(e)\}$.

Let $[d_{ij}]$ be an $n \times n$ dissimilarity matrix such that $L_\infty(d, D) < f/2$.

Then $NQM(d) = T$.

Proof (sketch)

The key is the observation that for every four indices i, j, k, l , if $L_\infty(d, D) < f/2$, then the Four Point Method on the dissimilarity matrix d will return the same tree on i, j, k, l as if it were applied to the additive matrix D .

Phylogeny reconstruction as a statistical estimation problem

Initially phylogeny reconstruction was based upon maximum parsimony analyses of morphology, or simple distance-based analyses of molecular sequences.

However, phylogeny reconstruction changed dramatically beginning in the 1960's with the introduction of stochastic models of evolution (Jukes-Cantor, Kimura 2-parameter, HKY, etc.).

Markov models of evolution

A Jukes-Cantor model tree is a pair (T, λ) where

- T is a rooted binary tree,
- λ is a function so that λ_e is the expected number of mutations of a random site on edge e

Assumptions:

1. The site at the root of T is drawn from the uniform distribution.
2. The number of times each site changes on each edge obeys a Poisson distribution.
3. The sites (i.e. positions) evolve identically and independently.

From these two assumptions one can prove that

$\lambda_e = -3/4 \log(1 - 4/3 p(e))$, where $p(e)$ is the probability that a random site has different states at the endpoints of e .

Other models of sequence evolution

- Most other models are still focused on single site evolution, and only differ from Jukes-Cantor in the constraints on the stochastic substitution matrix. They do not tend to relax the strong assumption of *i.i.d.* site evolution. (Note that *i.i.d.* does not mean the same rate of evolution, only that the rates are drawn from a distribution, but those rates are drawn identically and independently.) Amino-acid sequence evolution models tend to fall in this category.
- Some models are based upon codons (triplets of nucleotides that code for amino-acids).
- Some (fairly unused) models include rearrangement events within the sequence, duplications, insertions, and deletions of substrings.

Statistical Consistency

A phylogeny reconstruction method Φ is said to be **statistically consistent** under the model M if for all model trees $(T, params)$ in M , given random sequences S generated at the leaves of $(T, params)$, the probability that $M(S) = T$ goes to 1 as the sequence length increases.

Statistical estimation

- Is the model *identifiable*?
- Is a given phylogeny reconstruction method *statistically consistent* under the model?
- How much data does a given method need to reconstruct a given model tree correctly with high probability?

A brief history of mathematical phylogenetics

- 1960's and on: stochastic models of evolution, with *i.i.d.* evolution between sites
- 1978: Maximum Parsimony and Maximum Compatibility are not statistically consistent (Felsenstein)
- Mid-1990's and on:
 - Proofs of statistical consistency for basic methods (neighbor joining and maximum likelihood)
 - First mathematical analyses bounding the sequence length requirements of different methods
 - The Short Quartet Methods (the first “fast converging” methods)
 - The Disk-Covering Methods: turning exponentially converging methods into fast converging methods

Jukes-Cantor distance estimation

Let $\lambda_{ij} = \sum_{e \in P_{ij}} \lambda_e$ be the additive matrix of *model distances* corresponding to the Jukes-Cantor model tree (T, λ) .

Let $d_{ij} = -\frac{3}{4} \ln(1 - \frac{4}{3} h_{ij})$, where h_{ij} is the normalized Hamming distance between sequences i and j .

Then

$$d_{ij} \rightarrow \lambda_{ij}$$

in probability as the sequence length increases.

Topology-Invariant Neighborhoods

Many additive matrices correspond to the same unrooted tree, only with different edge lengths!

Let (T, w) and (T', w') be two binary trees leaf-labelled by set S , with w and w' positive edge-weightings of their respective trees, and let D and D' be the two additive matrices corresponding to T and T' , respectively. If $L_\infty(D, D') < f/2$ where $f = \min\{w(e) : e \in E(T)\}$, then $T = T'$.

Hence, most distance methods are *statistically consistent* for the JC model (and for most others).

(Proof is in the picture)

Sequence length requirements

Question: Let (T, λ) be a Jukes-Cantor model tree, and let $\epsilon > 0$ be given. For what sequence length k will $Pr[M(S) = T] > 1 - \epsilon$, for S a set of sequences of length k generated on T ?

Factors affecting this:

- ϵ
- $f = \min \lambda_e$
- $g = \max \lambda_e$
- n , the number of leaves in the tree,
- M , the method!

Theorem 1: Let $\{d_{ij}\}$ be an $n \times n$ dissimilarity matrix, $\{\lambda_{ij}\}$ the matrix of tree distances defined by JC tree (T, λ) , $f = \min_e \lambda_e$, and $\epsilon > 0$. Then there is a constant $C > 0$ such that, if the sequence length exceeds

$$C \log n e^{O(\max \lambda_{ij})}$$

then, with probability at least $1 - \epsilon$, the Naive Quartet Method recover the true tree.

Comments:

- Since $\max \lambda_{ij} = O(g \text{diam}(T))$, and $\text{diam}(T) \leq n - 1$, we say that NQM is *exponentially converging*.
- The proof follows from the condition that if $L_\infty(d, \lambda) < f/2$, then $NQM(d) = T$.

Other distance-based methods

- Neighbor Joining (and its variants BioNJ and Weighted NJ) and FastME are distance-based methods that are used in practice. These perform better in simulation (and empirically) than NQM or other theoretically obtained methods, but (as far as we can tell) have the same theoretical performance.
- UPGMA is another polynomial time method that is sometimes used, but it doesn't perform well due to its essential reliance on a strong molecular clock. (More on this later.)