

Week Plan 2

Reading material for lectures L1-L3 (substitution models) is Yang: Computational Molecular Evolution chapter 1

Reading material for lectures on phylogenies is Yang: Computational Molecular Evolution chapter 3

The Exercises that should be done in week 3 are:

A. Counting of trees

- No root inner nodes have 3 edges, only leaves are labelled. How many distinct trees with 10 leaves?
- No root and only leaves are labelled. How many distinct trees with 8 leaves.
- No root inner nodes have 3 edges all nodes are labelled. How many distinct trees with 4 leaves? Show 2 trees that would be identical if inner nodes were unlabelled.
- No root, inner nodes have 3 edges – no nodes are labelled. How many trees with 6 leaves?
- No root - no nodes are labelled. How many trees with 6 leaves?

B Ancestral nucleotides.

Find the cheapest assignment of nucleotides to internal nodes. What is the evolutionary cost of the tree then? Are the nucleotides assigned unambiguously?

Transitions 2, Transversions 5, Identity 0.

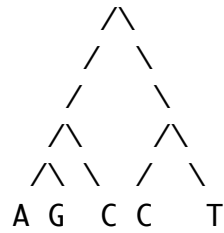
Let n be a node in a binary (internal nodes has two children) tree with a root, let n_L be the left child, n_R the right child. $d(,)$ is a distance function on nucleotides. $w(n, N)$ is the with of the evolution in the subtree hanging from n if the nucleotide N must be at node N .

Basic recursion:

Initialcondition $w(\text{leaf}, N) = 0$ if N is actually at this leaf, infinity if not.

$$w(n, N) = \min\{w(n_L, N_L) + d(N, N_L)\} + \min\{w(n_R, N_R) + d(N, N_R)\} .$$

First min is taken over N_L element $i \in \{A, C, G, T\}$, the second min N_R element $i \in \{A, C, G, T\}$.



1. Why does this recursion work?
2. Can you come up with a simple example where the method would fail if we used $w(n)$ instead of $w(n, N)$ (And also ignored N_L and N_R) ?
3. Could this algorithm be modified so it could handle ambiguity in sequencing (say we only knew that the nucleotide at first leaf was a purine)?
4. How would the recursion look if we were analyzing proteins?
5. Could you make an algorithm that would minimize the number of amino acid changes if we had codons at the leaves?
6. Given an alignment of 10 sequences, 100 nucleotides long how could the most parsimonious phylogeny be found? How much computation would be involved? Would it be slower if we had had proteins?