

MS2a, Week 2

Rune Lyngsø

October 19, 2011

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

Let n be a node in a binary (internal nodes have two children) tree with a root, let n_L be the left child, n_R the right child. Let $d(,)$ be a distance function on nucleotides, with the distance between identical nucleotides being 0, the distance between nucleotides separated by a transition being 2, and the distance between nucleotides separated by a transversion being 5. $w(n, N)$ is the total cost of the evolution in the subtree hanging from n if the nucleotide N must be at node n .

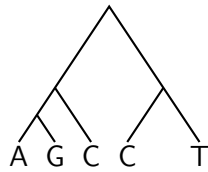
Basic recursion:

Initial condition $w(\text{leaf}, N) = 0$ 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 in $\{A, C, G, T\}$, the second min is taken over N_R element in $\{A, C, G, T\}$.

- Find the cheapest assignment of nucleotides to internal nodes in the tree below. What is the evolutionary cost of the tree with this assignment? Are the nucleotides assigned unambiguously?



- g. Why does this recursion work?
- h. 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)?
- i. Could this algorithm be modified so it could handle ambiguity in sequencing (say we only knew that the nucleotide at the first leaf was purine)?
- j. How would the recursion look if we were analyzing proteins?
- k. Could you make an algorithm that would minimize the number of amino acid changes if we had codons at the leaves?
- l. 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?