Exercises – Phylogenetics

Universität Bielefeld, WS 2015/2016, Dipl.-Inform. Damianos Melidis, B. Sc. Kevin Lamkiewicz
http://wiki.techfak.uni-bielefeld.de/gi/Teaching/2015winter/Phylogenetik

Exercise List 4 — 10.11.2015

Due to: 17.11.2015

Exercise 1 Small Parsimony – Fitch-algorithm. **(3 Points)**

In this task we'll look at the original work of Walter M. Fitch: "Towards Defining the Course of Evolution: Minimum Change for a Specific Tree Topology", publiziert in dem Journal "Systematic Zoology". You can find this article as an PDF online: http://www.jstor.org/stable/2412116.

(a) Apply the Fitch-Algorithm on the tree on the right. Write down a solution and their *parsimony*-cost that can be found with the algorithm. Specify for each internal node the set S (see figure on page 28).

(b) We talked about the original work of Walter M. Fitch (1971) in the lecture. You can apply an extra step after the bottom-up phase such that the top-down phase will find *all* optimal labels.
Use this algorithm to enrich the set S for the tree above (the corresponding page is printed on the next page). Are there new solutions that weren't found in task (a)? Indicate such a solution and its *parsimony*-cost.

Exercise 2 Small Parsimony – Sankoff-Algorithm. **(3 Points)**

Apply the Sankoff-Algorithm (**with unit costs**) on the tree from exercise 1, in order to determine a *most parimonious* label for the internal nodes. Specify the values for $C(u, a)$ for each internal node (figure at page 30).

Write down all solutions, that weren't found in task 1(a).

1 Bonus Point: Repeat the exercise with the following cost function:

cost	A	C	G	T
A	0	2	1	2
C	2	0	2	1
G	1	2	0	2
T	2	1	2	0

Exercise 3 Number of binary trees. **(3 Points)**

Conni Count, the most unsuccessful bioinformatician in his time, wants to find a *most parsimonious* phylogenetic tree by specfying *all possible* unrooted trees and calculating the parsimony cost for each tree.

(a) He is able to calculate 1.000.000 trees in one second with his implementation of the Fitch-Algorithm. Conni is 30 years old. How old does Conni have to become, if he wants to get the result for a dataset with 17 species?

(b) Conni received a computing cluster for christmas. He is now able to calculate $10^{12} = 1.000.000.000.000$ trees per second with one terahertz and one calculation per clock. Our universe is roughly 15.000.000.000 years old. If Conni would have started his program on this computing cluster at the big bang, how many leaves could have been processed at most until today?

Hint: Do **not** try to transform the formula $U_n = \prod_{i=3}^{n}(2i-5)$ to n. Just calculated U_n for growing n instead. (A spreadsheet works wonders.)

Turn around!

the data. But this additional information about the upper ancestral node also makes it clear that the first node can not then be a C. The only formulation that will permit the descendent positions to be accounted for in a single replacement requires that replacement to be from A to C in the descent from the first node as shown in Figure 2b (upper right). The elimination of the C from the first node is determined by what may be called the *rule of diminished ambiguity*. Its precise formulation is encompassed in steps I and II of the algorithm, to be presented further on, that contains the complete set of rules for the final phase of reconstructing the nodal sets.

In Figure 2c (middle left) is shown another preliminary phase reconstruction which accounts, using two replacements, for the descent of the characters of the three taxonomic units given. Figure 3d (middle right), however, shows an equally adequate solution which is not encompassed by the possible alternatives available in Figure 3c. Clearly G is a valid alternative for the first node. This case is encompassed by the *rule of expanded ambiguity* which is precisely described in steps III and IV of the forthcoming algorithm.

In Figure 2e (lower left) is shown a third preliminary phase reconstruction that accounts for four descendants using two replacements. In Figure 2f (lower right) is an equally valid solution. Indeed, the C at the lowest node in the preliminary reconstruction is a valid alternative to the A if and only if a C is allowed at the penultimate node above. It is characteristic of this type of case that two nodes, separated by a single node, both contain a nucleotide not present in the intervening node because of the intersection process. Hence, this is called the *rule of encompassing ambiguity* which is formulated as step V of the forthcoming algorithm.

In the preliminary phase, the nodes in Figure 1 were formulated in the order of increasing ancestral remoteness (1→5, with the order for formulating nodes 1 and 2 being arbitrary). In the final phase, the order for correcting the nodal sets must be reversed (5→1).

The preliminary set for the ultimate node is made the final set for that node. We then go to the penultimate node (4 in this case) and proceed according to the following six step algorithm.

I. If the preliminary nodal set contains all of the nucleotides present in the final nodal set of its immediate ancestor, go to II, otherwise go to III.
II. Eliminate all nucleotides from the preliminary nodal set that are not present in the final nodal set of its immediate ancestor and go to VI.
III. If the preliminary nodal set was formed by a union of its descendent sets, go to IV, otherwise go to V.
IV. Add to the preliminary nodal set any nucleotides in the final set of its immediate ancestor that are not present in the preliminary nodal set and go to VI.
V. Add to the preliminary nodal set any nucleotides not already present provided that they are present in both the final set of the immediate ancestor and in at least one of the two immediately descendent preliminary sets and go to VI.
VI. The preliminary nodal set being examined is now final. Descend one node as long as any preliminary nodal sets remain and return to I above.

Figure 1 illustrates the operation of the algorithm. The left hand side (Figure 1a) depicts the preliminary nodal sets. The ultimate ancestral nodal set 5 (AU) is considered the final set and we turn our attention to preliminary nodal set 4. This nodal set does not contain an A and therefore, according to step I, we proceed to step III. Nodal set 4 was not formed by a union and therefore we are directed by step III to go to step V. Following the directions of step V we discover that A is present in both nodal sets 3 and 5 (the rule of encompassing ambiguity) and must therefore be added to nodal set 4. (Mathematically, $((1 \cap 5) \cup (3 \cap 5)) = AU$.

1

[1]Fitch uses some different terms compared to the lecture notes: A *nodal set* corresponds to a *set S*, an *immediate ancestor* is a *parent node* and an *immediate descendant* is a *child node*. The *ultimate (ancestral) node* is simply the *root*. Instead of *character states* Fitch uses *nucleotides*.