CMB 2019 Homework Set II

Due February 27th 2019

- Make the assignment by yourself.
- You may want to design a program to answer this assignment, then it should be written in C++, JAVA, or Python.
- Give your solutions in a *.zip* file with the following name **"<your student number><your name>_CMB_HW2.zip"**, e.g., **"012345janjansen_CMB_HW2.zip".**
- Send this .zip file as an attachment of an e-mail with subject **"CMB_HW2"** to *erwin@liacs.nl* .
- Grade will be PASS/NO PASS.

Given a set of DNA sequences S = $\{S_1,\ldots,S_N\}$. Note that the sequences are given in the *sequences.txt* file. Every line contains one sequence, line 1 sequence S_1, etc. Now consider the following procedure for determining a Neighbor Joining Tree (NJ-Tree) for the set S (For details refer also to [1-3]):

1. Calculate a distance matrix *D* for the given set of sequences S using the sequence alignment code from Homework Set I.
2. The initial network T consists of a star network, where for every $i, 1 \leq i \leq N$: S_i forms different taxa in T, all connected to a central node c.
3. Calculate/update the distance matrix *Q* defined by:

$$q_{ij} = d_{ij}(N-2) - \sum_{k=1}^{N} d_{ik} - \sum_{k=1}^{N} d_{jk}\,, where\ d_{ij} is\ the\ distance\ between\ taxa\ i\ and\ j$$

4. Find taxa *f* and *g* where q_{fg} is minimal. Create a new node *u* to which *f* and *g* are connected (instead of *c*), and connect the new *u* to *c*.
5. Update the distance matrix Δ (of T) for the taxa f, and g that are now paired in the new node u:

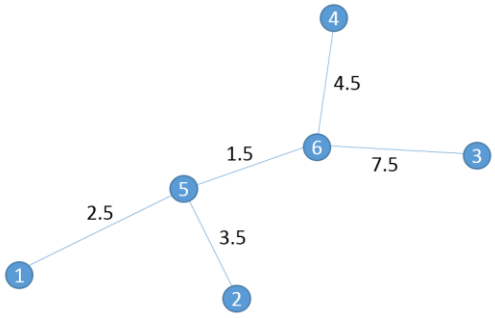
   $\partial_{fu} = 0.5 d_{fg} + \frac{1}{2(N-2)}\left[\sum_{k=1}^{N} d_{fk} - \sum_{k=1}^{N} d_{gk}\right]$ and $\partial_{gu} = d_{fg} - \partial_{fu}$
6. We update distance matrix *D* of step 1 by deleting the taxa *f and g* that were joined by *u*, and by adding the distances to the new *u* for all the other nodes *k* using the following calculation:

   $$d_{uk} = 0.5\left[d_{fk} + d_{gk} - d_{fg}\right]$$

   The updated *D* is now an *(N-1)x(N-1)* matrix.
7. N = N-1 and repeat from step 3) until T is fully resolved, i.e*., N=1*.

The answer to this assignment will be a list of triplets, each triplet consisting of a node with its direct 'ancestor' in *T*, and the length of the respective link as stored in distance matrix Δ.
Number/name nodes $S_1, S_2,\ldots,S_N$ as *1, 2, …, N* respectively, and give every new node a subsequent number, i.e., *N+1, N+2,* etc.

For example the tree T:



Would be listed as: (1, 5, 2.5) (2, 5, 3.5) (3, 6, 7.5) (4, 6, 4.5) (5, 6, 1.5)

Note: You may want to visualize a tree (without weights) by using the Newick format [4] and the web-site http://etetoolkit.org/treeview/ . The Newick format of the above tree would be: ((1,2)5,(4,3)6) .

References

1. N. Saitou, M. Nei, *The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution.* Vol. 4, No. 4, pp. 406-425, July 1987.
2. Olivier Gascuel, Mike Steel. Neighbor-Joining Revealed. Molecular Biology and Evolution, Oxford University Press (OUP), 2006, 23 (11), pp.1997-2000.
3. https://en.wikipedia.org/wiki/Neighbor_joining
4. https://en.wikipedia.org/wiki/Newick_format