Bielefeld University
Faculty of Technology

AG Genominformatik
Prof. Dr. Jens Stoye
Dr. Pedro Feijão

Algorithms in Genome Research
Winter 2015/2016

Exercises

Number 11, Discussion: 2016 February 5

1. **BP Median Problem**. Recently, Kováč (2013) introduced a series of simplifications for the solution to the *breakpoint median problem*, improving the complexity of the matching algorithm needed to solve it. Given the same weighted graph used by Tannier et al (2009), as we saw in the lecture, he proved that edges with weight 2 or 3 (from adjacencies present in 2 or 3 of the input genomes) and also weight 3/2 (from telomeres in all 3 input genomes) *must* be present in the optimal matching, so they can be chosen and the corresponding vertices can be removed from the graph. Also, edges with weight 1, from 2 common telomeres, can also be chosen in this way, because they are present in at least one median (but *not* in all). In the resulting graph, only edges of weight 1 and 1/2 will be present, and edges with weight 0 can be ignored, making the problem easier to solve (one reason is that the number of edges is now linear in the number of vertices).

   Now, given genomes $A = \{1_t, 1_h2_t, 2_h4_t, 4_h3_t, 3_h\}$, $B = \{1_t, 1_h3_h, 3_t2_h, 2_t4_t, 4_h\}$ and $C = \{1_t, 1_h2_t, 2_h3_h, 3_t4_t, 4_h\}$,

   (a) Draw the weighted matching graph of Tannier et al., containing the adjacencies and telomeres of A, B and C.

   (b) Choose the "high weight" edges as discussed above, and redraw the graph removing these edges.

   (c) Find a maximum weight matching in this graph, and the corresponding genome that is a breakpoint median, having the edges from item (b) and the maximum weight matching edges.

2. **Median Problem Lower Bound**. A well-known lower bound for the median of three genomes A, B and C for **any** distance d is given by

$$d(M,A) + d(M,B) + d(M,C) \geq \frac{1}{2}\big(d(A,B) + d(A,C) + d(B,C)\big)$$

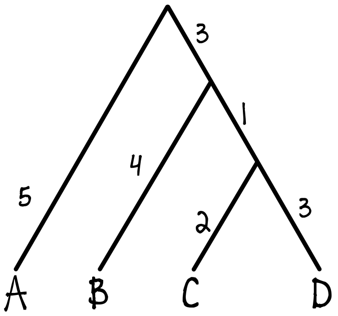
   (a) Prove this lower bound using the *triangle inequality* property:

$$d(X,Y) \leq d(X,W) + d(W,Y).$$

   (b) In exercise (1), does the breakpoint median reaches the lower bound?

3. **Adjacency Reconstruction**

Consider genomes $A = (1, -4, 3, 2, 5)$, $B = (1, -2, -3, 4, 5)$, $C = (1, 2, 3, 4, 5)$ and $D = (-4, -3, 2, -1, 5)$ (all *circular*, with 1 chromosome) and the phylogenetic tree below.



(a) Apply Fitch's algorithm on the adjacencies of the genomes to reconstruct adjacencies in the ancestral nodes of the tree. Choose always 1 (presence) on the root, whenever possible.

(b) In the case of conflicts, use the equation below to choose the adjacencies with higher weight,

$$w_\alpha(i,j) = \frac{D_L \cdot w_R(i,j) + D_R \cdot w_L(i,j)}{D_L + D_R}$$

where D_L and D_R are the branch lenghts from the current node to the left and right child, $w_R(i,j)$ and $w_L(i,j)$ are the weights of adjacency (i,j) in the left and right child. On leaf nodes, $w(i,j) = 1$ if the adjacency is present, 0 if it is not.