

Algorithms in Genome Research  
Winter 2013/2014

Exercises

Number 4, Discussion: 2013 November 29

1. Recently, Kováč [1, <http://online.liebertpub.com/doi/abs/10.1089/cmb.2013.0004>] 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 [2, 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; also, the matching does not need to be perfect).

Now, given genomes  $A = \{1_t, 1_h 2_t, 2_h 4_t, 4_h 3_t, 3_h\}$ ,  $B = \{1_t, 1_h 3_h, 3_t 2_h, 2_t 4_t, 4_h\}$  and  $C = \{1_t, 1_h 2_t, 2_h 3_h, 3_t 4_t, 4_h\}$ ,

- (a) Draw the weighted matching graph of Tannier et al., omitting zero-weight edges.
  - (b) Choose the “high weight” edges as discussed above, and redraw the graph.
  - (c) Find a maximum weight matching in this graph, and the corresponding genome that is a breakpoint median.
  - (d) Find the SCJ median of  $A$ ,  $B$  and  $C$ .
  - (e) Comparing with the input genomes, what median do you think is a better representation of an ancestral genome of  $A$ ,  $B$  and  $C$ ?
2. A well-known lower bound for the median of three genomes  $A$ ,  $B$  and  $C$  for **any** distance  $d$  is given by

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

- (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? How about the SCJ median?

## References

- [1] J. Kováč. On the complexity of rearrangement problems under the breakpoint distance. *Journal of Computational Biology*, 2013.
- [2] E. Tannier, C. Zheng, and D. Sankoff. Multichromosomal median and halving problems under different genomic distances. *BMC bioinformatics*, 10:120, 2009.