Universität Bielefeld
Technische Fakultät

AG Genominformatik
Prof. Dr. Jens Stoye

Algorithms in Comparative Genomics
Summer 2018

Exercises

Number 8, return 2018 July 06

1. Given an all-duplicates genome with three linear chromosomes

$$G = [\ 3\ \ 5\ \ -4\ \ 2\ \ -5\ ]\ \ [\ 2\ \ 1\ ]\ \ [\ 3\ \ 4\ \ 1\ ],$$

   solve the genome halving problem under the DCJ distance, i.e., find a perfectly duplicated genome H with smallest DCJ distance to G.

2. The *guided* genome halving problem under the DCJ distance is the following: Given an all-duplicates genome G and an all-singleton genome A over the same set of genes, find a perfectly duplicated genome H such that the overall DCJ distance $d_{DCJ}(G, H)+d_{DCJ}(H_1, A)$ is minimized, where H_1 is the non-duplicated version of H.

   Try to develop an efficient algorithm to solve the guided genome halving problem under the DCJ distance.

3. The following paragraph about the genome halving problem for the breakpoint distance is from the paper "Multichromosomal median and halving problems under different genomic distances" by Eric Tannier, Chunfang Zheng and David Sankoff, BMC Bioinformatics 10:120, 2009. (Notation: For a gene g, $g1$ and $g2$ denote the two paralogous copies of g; similarly, for a gene extremity x, the two copies are denoted $x1$ and $x2$.)

   > Let Δ be an all-duplicates genome on a gene set $\mathcal{G}$, and G be the graph on the vertex set containing (1) all the extremities of the genes in $\mathcal{G}$, and (2) one supplementary vertex t_x for every gene extremity x. For any pair of gene extremities x, y, draw an edge in G weighted by zero, one or two according to the number of adjacencies in Δ among $x1y1$, $x1y2$, $x2y1$, and $x2y2$. Now for any vertex x, draw an edge xt_x weighted by half the number of telomeres among $x1$ and $x2$ in Δ. Finally, put an edge of weight 0 between t_xt_y for all pairs of gene extremities x, y.
   >
   > For a genome M on $\mathcal{G}$, define a perfect matching, also called M, by including edges xy and t_xt_y for each adjacency xy, and an edge xt_x for each telomere x. Let $w(M)$ be the weight of the matching M.

   (a) Prove the following:

   > *Claim.* For a genome M on $\mathcal{G}$, the perfect matching M thus constructed satisfies $w(M) = 2n - d(\Delta, M)$.

   (b) How can this construction be used to solve the genome halving problem under the breakpoint distance?

Hint: Have a closer look at the all-duplicates genome $\Delta = [\ 1\ \ 2\ \ 3\ ]\ \ [\ 1\ \ -3\ \ -2\ ]$.