6. The Genome Halving Problem

Literature:

Mixtacki, J. (2008). *Genome Halving under DCJ Revisited.* Proceedings of COCOON 2008, LNCS 5092(Chapter 28), pp. 276–286.

Motivation: reconstruct the original gene order within a genome after a whole genome duplication followed by genome rearrangement events.

Definition 22 (Duplicated Genome)**.** *In* a duplicated genome*, each gene appears twice, or (in adjacency notation) each head and tail appears twice. Further,*

- *a paralogous extremity of p is denoted by $\bar{p}$,*
- *a paralogous adjacency of $x = \{p, q\}$ is denoted by $\bar{x} = \{\bar{p}, \bar{q}\}$, and*
- *a paralogous chromosome of C is denoted by $\bar{C}$.*

Example 17. *A duplicated genome:* $(\circ\ -d_2\ a_2\ -d_1\ -c_2\ b_2\ \circ)\ (\circ\ -b_1\ c_1\ a_1\ \circ)$

Definition 23. *A genome is*

- linear-perfectly duplicated*, if for each linear chromosome C_i, there is also a chromosome $C_j = \bar{C}_i$ for some $j \neq i$,*
- circular-perfectly duplicated*, if for each circular chromosome C_i, either there is also a chromosome $C_j = \bar{C}_i$ for some $j \neq i$, or $C_i = C \cup \bar{C}$, where each adjacency of C_i occurs either in C or in $\bar{C}$, but not in both,*
- perfectly duplicated *if it is linear-perfectly duplicated and circular-perfectly duplicated.*

Example 18. *A linear-perfectly duplicated genome:* $(\circ\ a_1\ -d_2\ -c_1\ b_1\ \circ)\ (\circ\ a_2\ -d_1\ -c_2\ b_2\ \circ)$*; a perfectly duplicated genome containing a circular-perfectly duplicated chromosome:* $(a_1\ d_1\ a_2\ d_2)\ (\circ\ c_1\ b_1\ \circ)\ (\circ\ c_2\ b_2\ \circ)$.

Lemma 5. *A genome A is perfectly duplicated if and only if*

- *for each adjacency $\{u, v\}$ in A, also $\{\bar{u}, \bar{v}\}$ is in A and $u \neq \bar{v}$ and*
- *for each telomere $\{u\}$ in A, also $\{\bar{u}\}$ is in A.*

Problem 5 (Genome Halving Problem)**.** *Given a (rearranged) duplicated genome A, find a perfectly duplicated genome B such that the DCJ distance between A and B is minimal.*

Definition 24 (Natural Graph)**.** *The natural graph $NG(A)$ of genome A is a graph whose vertices are the adjacencies and telomeres of A and in which each vertex containing an extremity p is connected to the vertex containing the paralogous extremity $\bar{p}$.*

Definition 25. *The set of paths and cycles of a natural graph is divided into four sets:*

- EC = *set of even cycles,*
- EP = *set of even paths,*
- OC = *set of odd cycles,*
- OP = *set of odd paths*

Observation 14. *A genome is perfectly duplicated if and only if $n = |EC| + \frac{|OP|}{2}$ (all cycles are 2-cycles, all paths are 1-paths)*

Theorem 7. $d_{GH}(A) = \min_B dcj(A, B) = n - |EC| - \lfloor \frac{|OP|}{2} \rfloor$

Proof. (i) This is a lower bound (a DCJ can change the number of components only by 1) and (ii) there is an algorithm that achieves this lower bound, as described in Algorithm 3. □

Algorithm 3 Greedy sorting algorithm for duplicated genomes

1: construct the natural graph
2: maximize the number of even cycles and odd paths in the natural graph:
   - for each k-path with $k > 1$, create a 2-cycle (and $(k-2)$-path if $k > 2$) $\implies$ all paths have then length 1
   - for each k-cycle with $k > 2$, create a 2-cycle and $(k-2)$-cycle $\implies$ all cycles have then length 1 or 2
   - for each 1-cycle + 1-cycle, create a 2-cycle
   - for each 1-cycle, create a 1-path
3: reconstruct the perfectly duplicated genome from the resulting natural graph

Algorithm 3 runs in linear time and space.