

Combinatorics problems in genome rearrangement

Istvan Miklos 10.3.08

The parsimony approach to the genome rearrangement can be considered as finding a minimum length transition path consisting of reversals between two signed permutations. An n long signed permutation is permutation of the first n positive integers together with a + or – sign before each number. For example, this is a 5 long signed permutation:

$$+3 -4 +1 +2 -5$$

A reversal reverts a consecutive part of a signed permutation. It changes both the order of the numbers and their sign. A consecutive part might contain a single number, in that case, the reversal changes its sign.

It is easy to show that the signed permutations form a group for the usual concatenation of permutations and the multiplications of signs (this group is isomorphic to $S_n \times (Z_2^+)^n$), and reversals have a group action on it. Therefore any transformation path that transforms a signed permutation g_1 into g_2 will transform $g_2^{-1}g_1$ into the identical permutation (assuming that reversals act from the right). Due to this, we are talking about *sorting* signed permutations instead of transforming one into another. Finding the minimal sorting reversal scenario is easy: the best algorithm has sub-quadratic running time (the problem size is n , the length of the signed permutation to sort) (Tanier & Sagot, 2004), and the length of the minimal sorting scenario can be calculated in linear time (without giving a path) (Bader et al., 2001). However, several related problems are hard or have unknown complexity. Below we give a list of them:

- Sorting unsigned permutations by reversals is proven to be NP-complete.
- An optimal reversal median of three signed permutations g_1, g_2, g_3 is a signed permutation g that minimizes the sum of distances $d(g, g_1) + d(g, g_2) + d(g, g_3)$. It is proven that the optimal reversal median problem is NP-complete (Caprara, 1999).
- There might be several minimal sorting reversal paths. It is an open question if the number of minimal sorting reversal paths can be calculated easily or it is a #P problem.

Further problems

We are going to introduce further problems related to genome rearrangement that needs more definitions.

Problem 1. A signed permutation can be represented as a *graph of desire and reality*. In this representation, the signed permutation is transformed into a double-length non-signed permutation replacing $+i$ by $2i-1, 2i$ and replacing $-i$ by $2i, 2i-1$. This unsigned permutation is framed by 0 and $2n+1$, where n is the length of the signed permutation. Vertices of the graph of desire and reality are the numbers of the unsigned permutation together with 0 and $2n+1$. Unlikely in the general graph theory, the spatial order of the vertices is fixed (see Fig.1.). Starting with 0, every other pair of vertices are connected in the unsigned permutation with a black line, and they are called *reality edges*, since they show the reality, i.e. what the neighbor of 0 is, etc. Also starting with 0, every node $2i$ and $2i+1$ are connected with a grey arc above the row of vertices, and these grey arcs are called *desire edges*, since they show which nodes should be neighbors to get the identity permutation. An edge is *oriented* if it spans an odd number of vertices, otherwise it is unoriented.

The vertices of the associated *overlap* graph are the edges of the graph of desire and reality, and two vertices are connected if the edges they represent overlap in the graph of desire and reality. The vertices of the overlap graph are colored, a vertex gets black color if it represents an oriented edge otherwise it is unoriented. The overlap graph might contain several components, a component is called oriented if it contains at least one black edge.

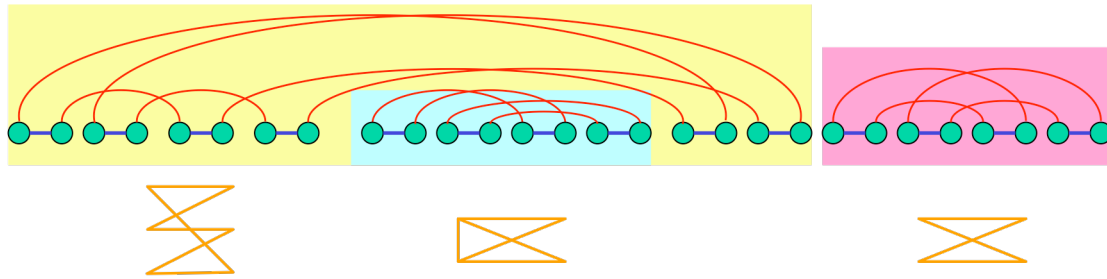


Fig. 1. A graph of desire and reality and its corresponding overlap graph

We can play the following game on overlap graphs: we can press a black vertex and it becomes a white one, loses all of its connection and their old neighbors change colors and any pair of neighbors changes connection. Namely, if both x and y are connected to the black vertex we would like to press, then after the transition they become connected if they were not, and will lose the edge they connect them otherwise. It is easy to show that any overlap graph in which all connected components contain at least one black edge can be transformed to an empty, all-white graph, and any kind of transformation path that transforms a fixed overlap graph have the same length (Bergeron, 2001).

Not all black-white graphs are an overlap graph for some signed permutation. However, it can be shown that any black-white graph in which each connected component contains at least one black edge can be transformed into an empty, all-white graph using the above-described transformations. The open question is that all transformation paths for a fixed black-white graph have the same length.

Problem 2. Let a reversal be described as a double cut-and-join (DCJ) mutation (Bergeron et al. 2006). The DCJ representation of a reversal tells which adjacencies are changed in the signed permutation. Let sorting paths be described by their series of reversals in DCJ representation. For example, the sorting path:

$$+3, +4, -1, -2 \rightarrow +1, -4, -3, -2 \rightarrow +1, +2, +3, +4$$

is represented by $(0, b_3|b_1, e_2), (e_1, e_4|b_2, 5)$. This means that before the first reversal, the beginning of number 3 was at the beginning of the permutation (represented as 0), the beginning of gene 1 was in adjacency with the end of gene 2, and the first reversal swapped the positions b_3 and b_1 . Similarly, the second reversal breaks the adjacencies between e_1 and e_4 and between b_2 and the end of the permutation by swapping e_4 and b_2 . Note that $(a, b|c, d)$ means the same reversal than $(d, c|b, a)$, but differs from, for example, $(b, a|c, d)$.

Let the vertices of a graph be the minimal reversal paths of a signed permutation. Let two points of this graph be connected iff at most four, not necessarily consecutive reversals can be removed from each of their DCJ representations such that the remaining patterns will be the same (note that the remaining representations of DCJ mutations might not represent valid DCJ operations). Our conjecture is that the graph will always be connected if the signed permutation contains only oriented components (more precisely, its overlap graph contains only oriented components).

Problem 3. It is easy to see that two reversals in a sorting path can be swapped if the intervals they revert do not overlap. Let two sorting paths be equivalent if one of them can be transformed into another by swapping consecutive swappable reversals. Given a representation of an equivalence class, is it easy to calculate the cardinality of the class?

References

- Bader, D.A., Moret, B.M.E., Yan, M.: A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. *J. Comp. Biol.*, vol. 8, num. 5, pp 483–491, 2001.
 Bergeron, A.: A very elementary presentation of the Hannenhalli-Pevzner theory. *Proceedings of CPM2001*, pp. 106–117, 2001.
 Bergeron, A., Mixtacki, J., Stoye, J.: A unifying view of genome rearrangements. *Proceedings of WABI2006*, pp 163–173, 2006.
 Caprara, A.: Formulations and hardness of multiple sorting by reversals. *Proc. 3rd Annual International Conference on Research in Computational Molecular Biology*, pp. 84–94, 1999.
 Tannier, E., Sagot, M.-F.: Sorting by reversals in subquadratic time. *Proceedings of the 15th CPM, Lecture Notes in Computer Science*, pp. 1–13, 2004.