The solution space of genome rearrangement problems: a graph theoretical and linear algebraic approach

Research proposal, 2022 Summer

May 6, 2022

1 Problem description

This research proposal offers a research on genome rearrangement problems and linear algebraic problems related to genome rearrangements. Basic genome rearrangement models assume that a genome is a sequence of unequivocally identifiable segments, and during genome rearrangement only the order (and in some models, also the direction) of these segments are changed. Therefore, genomes can naturally be described in the language of permutations (or signed permutations). The genome rearrangement models also define what kind of elementary operations are possible, and then the goal is to determine the minimum number of elementary operations necessary to transform one genome (permutation) into another one (another permutation). Without loss of generality, the target permutation can be the identical permutation, and in this way, the genome rearrangement problem boils down to sorting permutations with a prescribed set of elementary operations. That is, we are looking for the shortest series of elementary operations sorting a permutation.

Such a series of elementary operations is called a *genome rearrangement scenario* or *sorting scenario*. There might be multiple sorting scenarios of a permutation. The set of all sorting scenarios is called the *solution space*. Although efficient algorithms exist to compute the minimum number of elementary operations to sort a permutation and to obtain one sorting scenario, quite little is known about the solution space. It is known that the solution space might grow exponentially with the length of the permutation, and it is also conjectured that calculating the size of the solution space of a given permutation is a hard computational task[1]. Therefore, instead of listing all sorting scenarios and/or computing the size of the solution space, we would like to explore the solution space using a random walk. Such a random walk should perturb the current sorting scenario into another sorting scenario. The central question is what kind of perturbations are necessary to transform any sorting scenario into any another one.

[1] For those who are familiar with computational complexity classes: it is conjectured to be #P-hard, that can be considered as “at least as hard as NP”.

1.1 Sorting by block interchanges

We are interested in two genome rearrangement models. Here we give the description of one of them on which we already had some results in a previous research class.

A *block interchange* swaps two, not necessarily consecutive blocks in a permutation of the numbers between 1 and n. For example, if we swap $2, 3, 7$ and $4, 1$ in the permutation

$$5, 2, 3, 7, 6, 4, 1$$

we get the permutation

$$5, 4, 1, 6, 2, 3, 7.$$

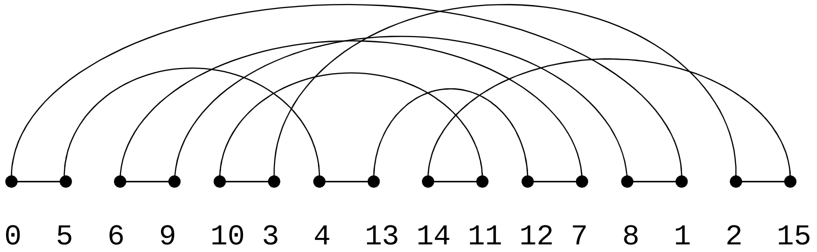
The *Sorting by block interchanges* problem asks for the minimum number of block interchange operations to transform a permutation into the identity permutation. This number is called the *block interchange distance*. To obtain this number, we have to consider the following discrete mathematical object, called the *graph of desire and reality*. This is a drawn multigraph. Multigraph means that two vertices might be connected with multiple (at most 2) edges. The "drawn" adjective emphasises that the drawing is also considered, that is, two graphs of desire and reality might not be identical even if they are identical as (multi)graphs. The construction is the following: replace each number k by $2k-1, 2k$ in the permutation π. Furthermore, frame these numbers between 0 and $2n+1$ where n is the length of π. For example, if π is

$$3, 5, 2, 7, 6, 4, 1$$

then we get

$$0, 5, 6, 9, 10, 3, 4, 13, 14, 11, 12, 7, 8, 1, 2, 15.$$

These numbers will be the vertices of the graph. The vertices are drawn along a line. We connect every second vertex with a straight line, that is, 0 is connected to 5, 6 to 9, etc. These edges are called the *reality edges*. Also every even number is connected to the next odd number with an arc, that is, 0 with 1, 2 with 3, etc. These edges are called the *desire edges*. In our example, we will get:



It is known that the minimum number of block interchanges necessary to sort the permutation is

$$\frac{n+1-c(\pi)}{2}$$

where n is the length of the permutation and $c(\pi)$ is the number of cycles in the graph of desire and reality. In our example, $n = 7$, and the graph of desire and reality contains 2 cycles. Therefore, 3 block interchange operations are sufficient to transform the permutation into the identity. Indeed, first swap $3, 5, 2, 7$ and 1 to get

$$1, 6, 4, 3, 5, 2, 7.$$

Then swap 2 and 6 to get

$$1, 2, 4, 3, 5, 6, 7.$$

Then finally, swapping 4 and 3 sorts the permutation. There might be multiple solutions, that is, there might be many sorting scenarios for a single permutation. In a previous reseach class, we proved the following theorem:

Theorem 1. *Let π be an arbitrary permutation. Consider the following graph $G(\pi) = (V, E)$. V is the solution space of π. For any v and $w \in V$, $(v, w) \in E$ if and only if v and w (as sorting scenarios) differ in two consecutive steps. Then $G(\pi)$ is connected.*

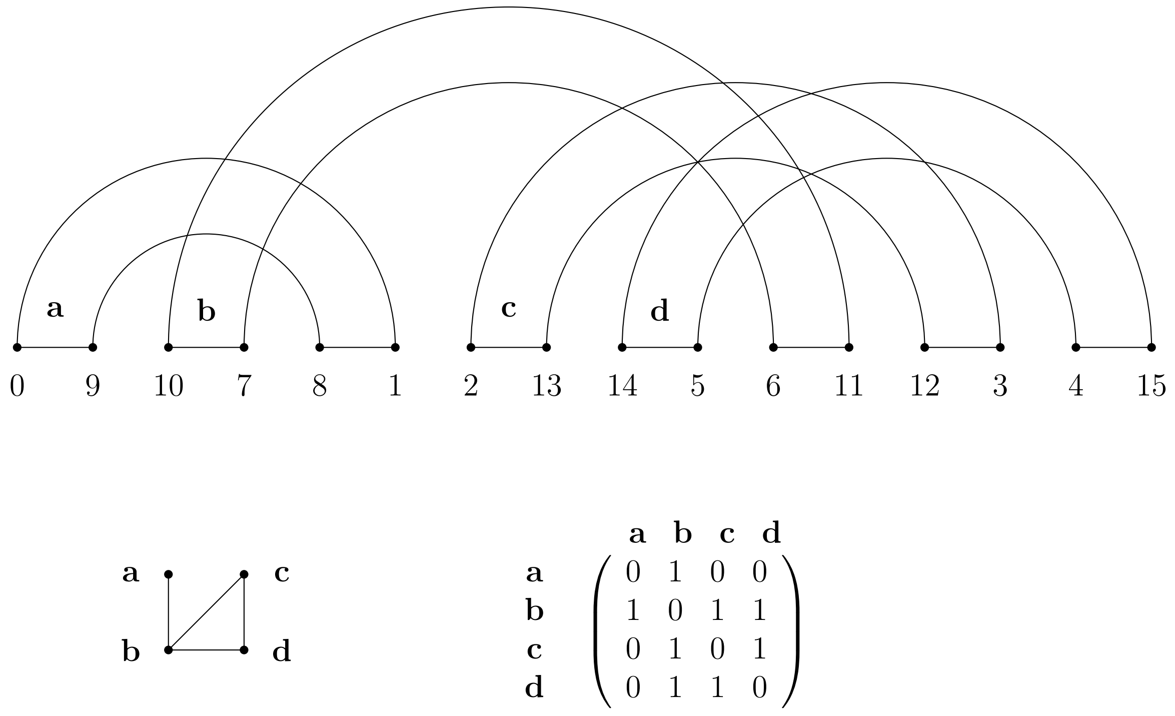
With other words: the solution space can be explored by small perturbations on the current sorting scenarios.

1.2 The linear algebraic approach

One particular set of permutations are those whose graph of desire and reality contains only cycles of length 4 (2 desire edges and 2 reality edges). Two such cycles *overlap* if they cannot be drawn without crossing edges. We can define the *overlap graph* of these permutations. The vertices of the overlap graph are the cycles in the graph of desire and reality, and two vertices in the overlap graph are adjacent if the cycles represented by them overlap. See an example in Figure 1. Each block interchange that decreases the block interchange distance can be considered as two Gaussian elimination steps in the adjacency matrix of the overlap graph over the field $\mathbb{F}_2$ followed by the elimination of the two rows used in the Gaussian elimination (see also the qualifying exercises where you can learn more about this Gaussian elimination). A corollary is that the rank of the adjacency matrix of the overlap graph over the field $\mathbb{F}_2$ is maximal.

1.3 Open problems

There are symmetric matrices with all 0 diagonal over the field $\mathbb{F}_2$ that are not adjacency matrices of overlap graphs. However, it can be shown that they all have even rank, and also, the minimum number of paired Gaussian elimination steps needed to transform them into the all 0 matrix is half their rank. Therefore, we can naturally define the "solution space" of symmetric all 0 matrices as the set of possible scenarios of paired Gaussian elimination steps transforming them into the all 0 matrix. What can we say about this solution space?



$$\begin{array}{c|cccc} & \mathbf{a} & \mathbf{b} & \mathbf{c} & \mathbf{d} \\ \mathbf{a} & 0 & 1 & 0 & 0 \\ \mathbf{b} & 1 & 0 & 1 & 1 \\ \mathbf{c} & 0 & 1 & 0 & 1 \\ \mathbf{d} & 0 & 1 & 1 & 0 \end{array}$$

Figure 1: A permutation whose graph of desire and reality contains only cycles of length 4. The cycles are labeled by **a**, **b**, **c** and **d**. Also shown its overlap graph and the adjacency matrix of the overlap graph.

This problem is related to the solution space of sorting by reversals of specific signed permutations, where we also have partial results, see `https://arxiv.org/pdf/1303.6799.pdf`[2]. Its relation to linear algebra over the field $\mathbb{F}_2$ is also known, see `https://link.springer.com/chapter/10.1007/11880561_23`. We are interested in how these seemingly different problems (sorting by block interchanges, sorting by reversals) can be related to each other, and if there is a unified linear algebraic approach that could help solve a further, long-standing open problem. This open problem is called the "four-reversal conjecture"[3]. It conjectures that the solution space of sorting by reversals is always connected (in terms as described in Theorem 1) if any neighboring sorting scenarios differ in at most four, not necessarily consecutive steps. The rational behind the idea that there might be a common linear algebraic approach describing these two problems is based on the following observations:

[2] also published as: Bixby, E, Flint, T, Miklós, I., (2016) Proving the Pressing Game Conjecture on Linear Graphs Involve, 9(1):41-56.

[3] This conjecture was first publised in Miklós, I., Mélykúti, B., Swenson, K. (2010) The Metropolized Partial Importance Sampling MCMC mixes slowly on minimum reversal rearrangement paths ACM/IEEE Transactions on Computational Biology and Bioinformatics, 4(7):763-767., however, it was already set up in 2006.

1. A reversal can be described as one Gaussian elimination step in the adjacency matrix of the overlap graph
2. A block interchange can be described as two Gaussian elimination steps in the adjacency matrix of the overlap graph
3. The solution space of sorting by block interchanges is connected if the neighbor scenarios differ in two block interchange steps
4. $2 \times 2 = 4$:)

1.4 Further reading

Some chapters from a textbook on genome rearrangement is available at `https://users.renyi.hu/~miklosi/2022SummerRES/GenomeRearrangementTextbook.pdf`. The most important is Chapter 10, which is about sorting by block interchanges. It proves the theorem on block interchange distance and should be understandable without reading the previous chapters. Chapter 7 provides a brief history of discovering genome rearrangement. It is for readers interested in the possible application of this quite theoretical research project. Chapter 9 gives an introduction to sorting by reversals. This will be taught during the research class up to Theorem 9.1. It is unlikely that we will consider other genome rearrangement models in this research class, however, for the sake of completeness a further genome rearrangement model is presented in Chapter 8.

2 Qualifying problems

Please, solve the following exercises.

1. (a) Draw the graph of desire and reality of the permutation

   $$\pi = 9 \quad 2 \quad 7 \quad 4 \quad 11 \quad 6 \quad 3 \quad 8 \quad 1 \quad 10 \quad 5,$$

   and compute its block interchange distance.

   (b) Observe that the graph of desire and reality of π contains only cycles of length 4 (2 reality edges and 2 desire edges in each cycle). Draw its overlap graph, and also obtain the adjacency matrix of the overlap graph.

   (c) Find a sorting scenario of π. (Hint: in each step, a block interchange has to transform 2 overlapping cycles of length 4 into 4 cycles of length 2, that is, with 1 reality edge and 1 desire edge). Since there are 6 cycles of length 4 in π, there will be 4 cycles of length 4 after one block interchange step, and 2 cycles of length 4 after 2 steps. Obtain the overlap graph of the intermediate permutations considering only the remaining cycles of length 4, and also obtain the corresponding 4×4 and 2×2 adjacency matrices.

(d) Let the adjacency matrix of π be denoted by A_6 and let the intermediate adjacency matrices be denoted by A_4 and A_2. "Blow up" A_4 and A_2 to 6×6 matrices by adding all 0 rows and columns to those positions where the cycles of length 4 were before the block interchange(s). Let these matrices be denoted by A'_4 and A'_2. Show the following: if r_i and r_j are the two rows corresponding to the cycles of length 4 on which the first block interchange acts, then A'_4 can be obtained from A_6 by adding r_i and r_j to some of the rows of A_6 (considering these operations over the field $\mathbb{F}_2$). Prove a similar statement on A'_4 and A'_2. (You may read the next exercise for a hint.)

(e) Prove that in each step, the rank of the adjacency matrix over the field $\mathbb{F}_2$ is decreased by 2.

2. We call the addition of r_i and r_j to certain rows of the current adjacency matrix *Gaussian elimination steps*. The rational behind this naming is the fact that it makes the corresponding c_i and c_j all 0 columns. In this exercise, we slightly abuse the notation that both a cycle and its corresponding row in the adjacency matrix will be denoted in the same way.

   Prove the following. Let π be a permutation whose graph of desire and reality contains only cycles of length 4, and let A be the adjacency matrix of its overlap graph. Let a block interchange transform 2 cycles of length 4, r_i an r_j, into 4 cycles of length 2. Furthermore, let A' be the "blown up" adjacency matrix of the so-obtained permutation. Show that A' can be obtained from A in the following way:

   (a) For each cycle r_k overlapping with r_i (including r_j), add r_j to r_k.

   (b) For each cycle r_k overlapping with r_j (including r_i), add r_i to r_k.

   Also show that the rank of A' is the rank of A minus 2 (over the field $\mathbb{F}_2$).

   Remark: The first part of this exercise has a straightforward solution, however, it needs quite a case study. Please make sure that you cover all the possible cases.

3. Find a symmetric $0-1$ matrix with all 0 diagonal, with dimension of $2k \times 2k$ for some positive integer k and full rank over the field $\mathbb{F}_2$ that cannot be the adjacency matrix of the overlap graph of a permutation whose graph of desire and reality contains only cycles of length 4.