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

Research proposal, 2020 Fall

July 3, 2020

Problem description

In this research proposal, we would like to work on genome rearrangement problems and linear algebraic problems related to genome rearrangements. Although genome rearrangement problems can naturally be described on the language of permutations, graph theoretical and linear algebraic approaches turned to be fruitful to study these problems.

We are interested in several open problems. 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

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

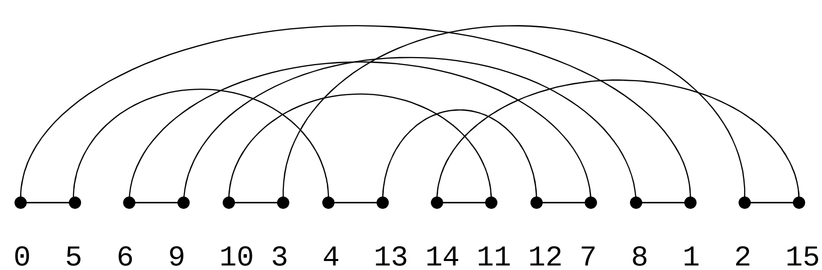
The *Sorting by block interchanges* problem asks for the minimum number of block interchange operations to transform a permutation into the identity permutation. To obtain this number, we have to consider the following graph, called the *graph of desire and reality*. This graph is a drawn multigraph, that is, the drawing is important and also, two vertices might be connected with multiple (at most 2) edges. The construction is the following: replace any number k with $2k-1, 2k$ 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. These 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. Also every even number is connected to the next odd number with an arc. That is, 0 with 1, 2 with 3, etc. 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 operqations 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. This series of block interchange operartions is called a *sorting scenario*. There might be multiple solutions, that is, there might be many sorting scenarios for a single permutation. The set of sorting scenarios are called the *solution space*. 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.

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. Each block interchange that decreases the block interchange distance can be considered as two Gaussian elimination step in the adjacency matrix of the overlap graph over the field $\mathbb{F}_2$. A corollary is that the block interchange distance is half the rank of the adjacency matrix of the overlap graph over the field $\mathbb{F}_2$.

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 al 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?

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`. Its relation to linear algebra over the field $\mathbb{F}_2$ is also knon, 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 further open problems.

Qualifying problems

Read Chapters 9 and 10 from this electronic note:
`https://users.renyi.hu/~miklosi/AlgorithmsOfBioinformatics.pdf`.

This chapter gives the detailed description of the theorem of sorting by reversals and sorting by block interchanges. Please, also solve the following exercises:

1. The block interchange distance is the minimum number of necessary block interchange operations to sort a permutation. What is the largest possible block interchange distance of a permutation of size n?

2. Prove that the solution space might grow faster than any exponential function of the length of the permutation.

3. Let s be a sorting scenario of π such that 1 is moved to the begining of the permutation in the second block interchange in s. Prove that there is a sorting scenario s' such that 1 is moved to the beginning of the permutation in the first block interchange in s', and s and s' differ only in the first two block interchanges.

4. Let s be a sorting scenario of π. Prove that there is a finite series of sorting scenarios $s = s_0, s_1, \ldots s_k$ such that for each $i = 0, 1, \ldots k-1$, s_i and s_{i+1} differ only in two consecutive steps, and in s_k, 1 is moved to the first position of the permutation in the first block interchange.

5. Let A be the adjacency matrix of an overlap graph of some permutation π. (We assume that the graph of desire and reality of π contains only cycles of length 4. Consider a block interchange that decreases the block interchange distance on π, let π' be the resulting permutation when this block interchange operation is applied on π, and let A' be the adjacency matrix of π'.

   Describe the pair of Gaussian elimination steps that transforms A to A'.

6. Find a symmetric $0-1$ matrix with all 0 diagonal that cannot be adjacency matrix of the overlap graph of a permutation whose graph of desire and reality contains only cycles of length 4.