Pattern avoiding permutations in genome rearrangement problems: the transposition model

Giulio Cerbai, Luca Ferrari (Dipartimento di Matematica e Informatica, University of Firenze, Italy)

One of the major trends in bioinformatics and biomathematics is the study of the genome rearrangement problem. Roughly speaking, given a genome, one is interested in understanding how the genome can evolve into another genome. To give a proper formalization, several models for rearranging a genome have been introduced, each of which defines a series of allowed elementary operations to be performed on a genome in order to obtain an adjacent one. For several models, it is possible to define a *distance* between two genomes, by counting the minimum number of elementary operations needed to transform one genome into the other. The investigation of the main properties of such a distance becomes then a key point in understanding the main features of the model under consideration.

A common formalization of any such models consists of encoding a genome using a *permutation* (in linear notation) and describing an elementary operation as a *combinatorial* operation on the entries of such a permutation. Many genome rearrangement models have been studied under this general framework. For instance, the *reversal* model consists of a single operation, defined as follows: a new permutation is obtained from a given one by selecting a cluster of consecutive elements and reversing it. More formally, given $\pi = \pi_1\pi_2\cdots\pi_n$, a reversal is performed by choosing $i < j < n$ and then forming the permutation $\sigma = \pi_1\cdots\pi_{i-1}\boxed{\pi_j\pi_{j-1}\cdots\pi_{i+1}\pi_i}\pi_{j+1}\cdots\pi_n$. This model was introduced in [9], then studied for instance in [1, 8]. Another interesting model, proposed in [6], is the *tandem duplication-random loss model*, in which a cluster of consecutive elements of a permutation is replicated (next to the original one), then one copy of each duplicated element is deleted at random. As a final example, a very popular and studied model is the *transposition model*, see [2]. Given a permutatation $\pi = \pi_1\cdots\pi_n$, a *transposition operation* consists of taking two adjacent clusters of consecutive elements and interchanging their positions. Formally, one has to choose indices $i < j < k < n$, then form the permutation $\sigma = \pi_1\cdots\pi_{i-1}\boxed{\pi_{j+1}\pi_{j+2}\cdots\pi_k}\boxed{\pi_i\pi_{i+1}\cdots\pi_j}\pi_{k+1}\cdots\pi_n$.

Independently from the chosen model, there are some general questions that can be asked in order to gain a better understanding of its combinatorial properties. First of all, the operations of the model often (but not always) allow to define a *distance* d between two permutations ρ and σ, as the minimum number of elementary operations needed to transform ρ into σ. Moreover, when the operations are nice enough, the above distance d could even be *left-invariant*, meaning that, given permutations π, ρ, σ (of the same length), $d(\pi, \rho) = d(\sigma\pi, \sigma\rho)$. As a consequence, choosing for instance $\sigma = \rho^{-1}$, the problem of evaluating the distance $d(\pi, \rho)$ reduces to that of sorting π with the minimum number of elementary allowed operations. Now, if d is a left-invariant distance on the set S_n of all permutations of the same length, define the *k-ball* of S_n to be the set $B_k^{(d)}(n) = \{\rho \in S_n \mid d(\rho, id_n) \leq k\}$, where id_n is the identity permutation of length n. The following questions appear quite natural to ask:

- compute the diameter of $B_k^{(d)}(n)$, i.e. the maximum distance between two permutations of $B_k^{(d)}(n)$;

- compute the diameter of S_n, i.e. the maximum distance between two permutations of S_n;
- characterize the permutations of $\partial B_k^{(d)}(n)$, i.e. the permutations of $B_k^{(d)}(n)$ having maximum distance from the identity;
- characterize the permutations of ∂S_n, i.e. the permutations of S_n having maximum distance from the identity;
- characterize and enumerate the permutations of $B_k^{(d)}(n)$;
- design sorting algorithms and study the related complexity issues.

In the literature there are several results, concerning several evolution models, which give some insight into the above problems. Our work starts from the observation that, in many cases, the balls $B_k^{(d)}(n)$ can be characterized in terms of *pattern avoidance*. This idea is not new; as far as we know, the first model which has been investigated from this point of view is the (whole) tandem duplication-random loss model: Bouvel andd Rossin [5] have in fact shown that, in such a model, the ball $B_k^{(d)} = \bigcup_{n\geq 0} B_k^{(d)}(n)$ is a class of pattern avoiding permutations. Subsequent works [4, 3, 7] have been done concerning the characterization and enumeration of the basis permutations of such classes. However, this appears to be an isolated case, and no further models have been analyzed using this approach. This is in sharp contrast with the fact that, in many interesting cases, the same observation is true. What we propose here is then a systematic investigation of the evolution models of the genome rearrangement problems using the permutation pattern paradigm. Specifically, for any given left-invariance distance on S_n, it is interesting to understand if the balls $B_k^{(d)}$ are classes of pattern avoiding permutations and, in the affirmative case, to investigate the property of such a class (starting of course from its basis).

In this talk we just scratch the surface of a single case, namely the *transposition model*. Given permutations σ and π of length n, denoting by dt the minimum number of block transpositions needed to transform σ into π, it is not difficult to show that dt is in fact a left-invariant distance on S_n. From now on, we will denote with $dt(\pi)$ the distance of π from the identity permutation (of the correct length). Our first result, very easy to show, is the observation that this model can in fact be studied with the tools of pattern avoidance.

Proposition 1. *Given $\pi \in S_n$ and $\sigma \in S_m$, if $\sigma \leq \pi$ then $dt(\sigma) \leq dt(\pi)$. As a consequence, if $B_k = \{\pi \mid dt(\pi) \leq k\}$ is the ball of radius k, then B_k is a class of pattern avoiding permutations, for all k.*

So the main issue is now to characterize B_k as a pattern avoiding class; in particular, we aim at investigating the structure of the permutations of B_k and enumerating its basis. We have only some partial results, which we are going to illustrate.

Before stating what we have obtained, we need a few notations and definitions.

A *strip* of $\pi = \pi_1\pi_2\cdots\pi_n \in S_n$ is a maximal consecutive substring $\pi_i\cdots\pi_{i+k-1}$ such that, for all $j = 1, \ldots i+k-2$, $\pi_{j+1} = \pi_j + 1$.

A permutation π is said to be *reduced* when, for all $i = 1, \ldots, n-1$, $\pi_{i+1} \neq \pi_i + 1$. In other words, π is a reduced permutation when it does not have points that are adjacent both in positions and values. Equivalently, a permutation is reduced if and only if all of its strips have length 1.

Any permutation π can be associated with a reduced permutation, denoted $red(\pi)$, which is obtained by replacing each string of π with its minimum element, then suitably rescaling the resulting word. It is easy to observe that $red(\pi) \leq \pi$. Moreover, for every permutation π, we have that $dt(\pi) = dt(red(\pi))$.

Given $\pi \in S_n$, let $v_1, \dots v_n$ be nonnegative integers. The *monotone expansion* of π through $v = (v_1, \dots v_n)$ is the permutation $\pi[v] = \pi[id_{v_1}, \dots, id_{v_n}]$ obtained from π by replacing each element π_i of π with the identity permutation id_{v_i} of length i suitably rescaled, so to mantain the relative order of the elements of π. So, for instance, if $\pi = 41352$ and $v = (0, 2, 1, 3, 2)$, we have $\pi[v] = \underbrace{\dots}_{4}\underbrace{12}_{1}\underbrace{5}_{3}\underbrace{678}_{5}\underbrace{34}_{2}$. The notion of monotone expansion is clearly related to those of inflation and of geometric grid class.

For a given permutation π, we denote with $EM(\pi)$ the set of all monotone expansions of π. More generally, if C is a set of permutations, we set $EM(C) = \bigcup_{\pi \in C} EM(\pi)$.

Lemma 2. *Given a $\{-1, 0, 1\}$-matrix M, denote with $Geom(M)$ the geometric grid class of permutations determined by M. Given a permutation π, let M_π be its permutation matrix. Then:*

1. $Geom(M_\pi) = Geom(M_{red(\pi)})$;
2. $EM(\pi) = Geom(M_\pi)$;
3. $EM(\pi) = EM(red(\pi))$.

From general facts of the theory of geometric grid classes, we are lead to the following result.

Corollary 3. *If C is a set of reduced permutations, then $EM(C)$ is a class of pattern avoiding permutations. Moreover, $EM(C)$ is strongly rational and finitely based.*

Coming back to our starting problem, we have complete results concerning the set B_1 of permutations having distance 1 from the identity.

Theorem 4. $B_1 = EM(1324)$.

Theorem 5. *$\pi \in B_1$ if and only if π avoids the patterns $321, 2143, 2413, 3142$.*

We are also able to enumerate the class B_1.

Theorem 6. *For every $n \geq 1$, let $f_n = B_1 \cap S_n$ be the number of permutations of length n having distance 1 from the identity. Then*

$$f_n = \binom{n+3}{n} - 2\binom{n+2}{2} + n + 2,$$

and its generating function is

$$F(x) = \sum_{n \geq 0} f_n x^n = \frac{1 - 3x + 4x^2 - x^3}{(1-x)^4}.$$

The associated sequence is sequence A050407 in the OEIS.

Concerning larger values of k, we have been able to prove the following general result.

Theorem 7. *Let $k \geq 1$.*

1. *There exist $N = N(k)$ reduced permutations $\alpha^{(1)}, \ldots, \alpha^{(N)}$ of length $3k+1$, each at distance k from the identity, such that*
$$B_k = \bigcup_{j=1}^{N} EM(\alpha^{(j)}.$$

2. *B_k is a strongly rational and finitely-based permutation class; moreover, each permutation of its basis has length at most $3k+1$.*

References

[1] V. Bafna, P. A. Pevzner. Genome rearrangements and sorting by reversals. *34th Annual Symposium on Foundations of Computer Science (Palo Alto, CA, 1993), IEEE Comput. Soc. Press, Los Alamitos, CA*, pp. 148–157, 1993.

[2] V. Bafna, P. A. Pevzner. Sorting by transpositions. *SIAM J. Discrete Math.*, 11:224–240, 1998.

[3] M. Bouvel, L. Ferrari. On the enumeration of d-minimal permutations. *Discrete Math. Theor. Comput. Sci.*, 15:33–48, 2013.

[4] M. Bouvel, E. Pergola. Posets and permutations in the duplication-loss model: minimal permutations with d descents. *Theoret. Comput. Sci.*, 411:2487–2501, 2010.

[5] M. Bouvel, D. Rossin. A variant of the tandem duplication-random loss model of genome rearrangement. *Theoret. Comput. Sci.*, 410:847–858, 2009.

[6] K. Chaudhuri, K. Chen, R. Mihaescu, S. Rao. On the tandem duplication-random loss model of genome rearrangement. *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms, ACM, New York*, pp. 564-570, 2006.

[7] W. Y. C. Chen, C. C. Y. Gu, K. J. Ma. Minimal permutations and 2-regular skew tableaux. *Adv. in Appl. Math.*, 47:795–812, 2011.

[8] S. Hannenhalli, P. A. Pevzner. Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. *J. ACM*, 46:1–27, 1999.

[9] G. A. Watterson, W. J. Ewens, T. Hall, A. Morgan. The chromosome inversion problem. *J. Theor. Biol.*, 99:1–7, 1982.