On the distribution of the number of cycles in the breakpoint graph of a random signed permutation

Simona Grusea

On the distribution of the number of cycles in the breakpoint graph of a random signed permutation

Simona Grusea, *LATP - UMR CNRS 6632, Équipe EBM, Université de Provence* and *Institut de Mathématiques de Toulouse, INSA de Toulouse, Université de Toulouse*

Abstract

We use the finite Markov chain embedding technique to obtain the distribution of the number of cycles in the breakpoint graph of a random uniform signed permutation. This further gives a very good approximation of the distribution of the reversal distance between two random genomes.

Index Terms

Markov processes, probabilistic algorithms, distribution functions, biology and genetics.

I. Introduction

A. The biological context

Comparing genomic organization between different species may help to decipher the evolutionary history of species and also to better understand the biology of contemporary species.

Orthologous genes are two genes, in two different species, that descend from the same gene at the ancestor of the two species, as the result of a speciation event. They tend, in general, to have similar functions. Therefore, finding a group of orthologous genes in close proximity in the genomes of two different species may represent a sign for evolutionary or functional relationships between these genes. For this to be the case, the observed orthologous gene clusters have to be *significant*, i.e. very improbable to have appeared by chance.

During the evolutionary time, the gene order in one genome can be affected by various genome rearrangement events, like inversions, translocations, transpositions, chromosomic fissions and fusions. Hence, in the absence of certain constraints due to functional selective pressures, the gene order is rapidly randomized. This is one reason why, in general, the null hypothesis taken in the significance tests for gene clusters is the hypothesis of random gene order.

In the "genomic comparison" literature various definitions for gene clusters exist, and also different statistical tests for detecting gene clusters which are significant from the point of view of the proximity of the orthologs (see [3], [4], [6], [7], [13], [15], [19], [20], [26], [35]). On the other hand, one might want to take into account also the order of the orthologs in these gene clusters, considering that the clusters in which the gene order is exceptionally conserved are even more biologically significant.

Sankoff and Haque [27] propose three adjacency disruption measures for comparing the order of the orthologs which are in common between two clusters in two genomes. They investigate

in more detail the "maximum adjacency disruption" criterion, giving analytic formulas for some values of its distribution under random gene order and also simulation results. Grusea [17] propose three measures based on the mathematical transposition distance between permutations, for assessing the exceptionality of the gene order in conserved genomic regions found by the reference region approach, and obtains analytic expressions for the distribution of these distances in the case of a uniform random permutation. In [12], Corteel *et al.* analyze the distribution of the number of common intervals in the case of a uniform random permutation and also study some generalized common intervals, in which gaps of a certain size are permitted.

In the "genome rearrangements" literature, several more biologically relevant distances have been studied, which take into account one or a combination of different types of genomic events: reversals, translocations, chromosomic fissions and fusions, biological transpositions, block-interchanges – see [23] for a review. The problem with using these distances as test statistics comes from the fact that their distributions for a random permutation are very difficult to obtain and there are very few results on this subject.

Recently, Doignon and Labarre [14] and Bona and Flynn [8] have found, in two different ways, the distribution of the number of (edge-disjoint) alternating cycles in the bicolored breakpoint graph of a random *unsigned* permutation, which can be used to deduce the exact distribution of the "block-interchange" distance of Christie [11]. However, for signed permutations, corresponding to the case when gene orientation is also known, the exact distribution of the number of cycles in the breakpoint graph is still unknown.

Sankoff and Haque [28] use a constructive approach to obtain asymptotic estimates for the distribution of the number of cycles in the breakpoint graph of two random *signed* permutations. Xu *et al.* [33] and Xu [32] use a similar approach to study the case of multichromosomal genomes.

The comparison of two genomes induces a decomposition of the genomes into *synteny blocks* (or *conserved segments*), chromosomic segments containing orthologous genes in the same or reverse order in the two genomes (see [10], [24]). The genomes could then be seen as permutations (unsigned or signed) of the set of synteny blocks. Some authors extend the notion of synteny block, allowing for some micro-rearrangements inside the synteny blocks (see [25]).

In the present work we are interested in finding the exact distribution of the number of alternating cycles in the breakpoint graph of a random *signed* permutation. The knowledge of

this distribution provides a very good approximation for the distribution of the reversal distance for a random signed permutation. This further allows us to use the reversal distance as a test statistic for assessing the exceptionality of the gene order in conserved genomic regions or of the order of the synteny blocks between two genomes.

We use the finite Markov chain embedding technique of Fu and Koutras [16] to obtain the distribution of the number of cycles in the breakpoint graph of a random signed permutation via a product of transition probability matrices of a certain finite Markov chain.

B. The breakpoint graph and the reversal distance

We let S_n denote the permutation group of order n. For a permutation $\pi \in S_n$ we will use the notation $\pi = [\pi(1), ..., \pi(n)]$. A *signed permutation* of n elements is a permutation $\pi = [\pi(1), ..., \pi(n)]$ in which the elements $\pi(i), i = 1, ..., n$ have a sign, either $+$ or $-$. In other words, $\pi(i) \in \{\pm 1, ..., \pm n\}$, for $i = 1, ..., n$ and $\{|\pi(1)|, ..., |\pi(n)|\} = \{1, ..., n\}$. We denote by B_n the set of all the signed permutations of n elements.

The *reversal* of the interval (i, j) in the signed permutation π reverses the subsequence $\pi(i), ..., \pi(j)$ while changing their signs, hence produces the signed permutation

$$\pi' = [\pi(1), ..., \pi(i-1), -\pi(j), -\pi(j-1), ..., -\pi(i+1), -\pi(i), \pi(j+1), ..., \pi(n)].$$

For $\pi \in B_n$, we let $d_{rev}(\pi, Id)$ denote its *reversal distance*, i.e. the minimum number of reversals needed to transform π into the identity permutation $Id = [+1, ..., +n]$.

Bafna and Pevzner [2] introduced the concept of *breakpoint graph* of a permutation and noticed important links between the cycle decomposition of this graph and the reversal distance. The breakpoint graph of a signed permutation is defined as follows. Given a signed permutation π in B_n, we first transform it into an unsigned permutation $\pi' \in S_{2n}$ by replacing the positive elements $+i$ by the pair $(2i-1, 2i)$ and the negative elements $-i$ by the pair $(2i, 2i-1)$. For instance, the signed permutation $\pi = [+3, -4, -2, +1, +5]$ is transformed into $\pi' = [5, 6, 8, 7, 4, 3, 1, 2, 9, 10]$. We then extend π' by adding two more elements, one at the beginning, which we will denote S (for *Start*) and one at the end, which we will denote T (for *Terminus*).

The breakpoint graph of the signed permutation $\pi \in B_n$ is the graph $G(\pi) = (V, B \cup C)$, having the set of vertices $V = \{S, 1, 2, ..., 2n, T\}$ and the edge set partitioned into two subsets:

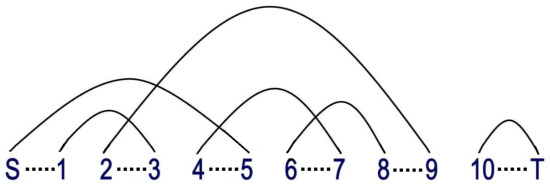

Fig. 1. The breakpoint graph of the permutation $\pi = [+3, -4, -2, +1, +5]$.

the set B of *solid* edges, corresponding to the adjacencies in the permutation π, and the set C of *dashed* edges, corresponding to the adjacencies in the identity permutation Id.

More precisely, if for every element a of the permutation π we denote respectively by a_L and a_R the left and right elements in the pair associated to a in π', then we will have a solid edge between a_R and b_L if a and b are consecutive in π. We have also solid edges between S and $(\pi_1)_L$ and between $(\pi_n)_R$ and T. We have dashed edges between the vertices $2i-1$ and $2i$, for every $i = 1, ..., n$, between S and 1 and between $2n$ and T.

Note that each vertex in $G(\pi)$ is of degree 2, having exactly one solid edge and one dashed edge incident to it. Consequently, the breakpoint graph decomposes uniquely into disjoint alternating cycles, i.e. cycles in which the solid edges and the dashed edges alternate. For a given cycle, we call its *length* the number of solid edges, or equivalently, the number of dashed edges it contains. In the example from Fig. 1, the breakpoint graph of the permutation $\pi = [+3, -4, -2, +1, +5]$ decomposes into two alternating cycles, one of length 1 and one of length 5.

For a signed permutation $\pi \in B_n$, we will denote $c(\pi)$ the number of alternating cycles in the breakpoint graph $G(\pi)$. The reason for introducing $c(\pi)$ in [2] was that it is easily seen to give the following lower bound, for every $\pi \in B_n$

$$d_{rev}(\pi, Id) \geq n + 1 - c(\pi). \tag{1}$$

Hannenhalli and Pevzner [18] proved that for every signed permutation $\pi \in B_n$ we have the exact formula

$$d_{rev}(\pi, Id) = n + 1 - c(\pi) + h(\pi) + f(\pi), \tag{2}$$

where $h(\pi)$ is the number of *hurdles* in $G(\pi)$ and $f(\pi)$ is 1 if π is a *fortress* and 0 otherwise (see [18] for the definitions of hurdle and fortress).

The problem of computing the reversal distance for signed permutations (without giving an optimal sequence of reversals) can be solved in linear time (see Bader *et al.* [1]). The problem becomes more complicated if one wants also an optimal sequence of reversals. The most efficient sorting algorithms at present are the one by Tannier *et al.* [31], than runs in $\mathcal{O}(n^{3/2}\sqrt{\log n})$, and the algorithm by Swenson *et al.* [30] that runs in $\mathcal{O}(n \log n)$ time for almost all permutations.

Caprara [9] showed that genomes containing hurdles are very rare. For example, less than one percent of the genomes with 8 genes contain hurdles and only one in 10^5 genomes with 100 genes. Swenson *et al.* [29] proved that the probability that a random signed permutation on n elements contains a hurdle is $\mathcal{O}(n^{-2})$ and the probability that it contains a fortress is $\mathcal{O}(n^{-15})$. It was also shown that the bound (1) approximates the reversal distance extremely well for both simulated (see Kececioglu and Sankoff [22]) and biological data (see Bafna and Pevzner [2]). Kececioglu and Sankoff [22] observed that the average difference between this bound and the exact distance is less than 1 for a random permutation.

One can therefore use the bound (1) as a very good approximation for the reversal distance.

Moreover, in the case of unichromosomal genomes, the bound (1) agrees exactly with the *double-cut-and-join* (DCJ) distance introduced by Yancopoulos *et al.* [34]. For more details on this distance see also [5].

The goal of the present work is to find the distribution of $c(\Pi)$ for a random (uniform) signed permutation Π.

II. The distribution of $c(\Pi)$

A. The finite Markov chain embedding technique

For obtaining the distribution of $c(\Pi)$ for a random signed permutation Π, we use the finite Markov chain embedding technique introduced by Fu and Koutras [16].

Let X_n (n a non-negative integer) be a non-negative integer random variable. As in Definition 2.1 of [16], we call X_n *finite Markov chain embeddable* if

- (i) there exists a (possibly inhomogeneous) finite Markov chain $\{Y_t : 1 \leq t \leq n\}$ with values in a finite state space $E = \{a_1, ..., a_m\}$,
- (ii) there exists a finite partition $\{C_x, x = 0, 1, ..., \ell\}$ of E, and
- (iii) for every $x = 0, 1, ..., \ell$ we have

$$\mathbb{P}(X_n = x) = \mathbb{P}(Y_n \in C_x).$$

The distribution of X_n can in this case be obtained via a product of transition matrices of the Markov chain $(Y_t)_{1\leq t\leq n}$. Indeed, if we define, for every $2 \leq t \leq n$, the transition matrix $P_t := (P_t(y,z))_{y,z\in E}$ by

$$P_t(y,z) = \mathbb{P}(Y_t = z | Y_{t-1} = y), \forall y, z \in E,$$

then, by Theorem 2.1 in [16], we have

$$\mathbb{P}(X_n = x) = \mu_1 P_2 \cdots P_n \sum_{i:a_i \in C_x} e_i, \tag{3}$$

where

$$\mu_1 = (\mathbb{P}(Y_1 = a_1), ..., \mathbb{P}(Y_1 = a_m))$$

is the row vector of the initial probability of the Markov chain and, for each $i = 1, ..., m$, e_i is the column vector having 1 at the i-th coordinate and 0 elsewhere.

B. The construction of the Markov chain

Let n be a fixed positive integer. We start with Π_1 being a random uniform signed permutation with one element, hence $\Pi_1 = [+1]$ with probability $1/2$ and $\Pi_1 = [-1]$ with probability $1/2$. For every $t = 2, ..., n$, we let Π_t represent the random signed permutation of t elements which is obtained from Π_{t-1} by inserting at random the element t uniformly into one of the t possible positions, with the "$+$" sign with probability $1/2$ and the "$-$" sign with probability $1/2$, the sign being independent of the position.

Note that $(\Pi_t)_{1\leq t\leq n}$ is an inhomogeneous Markov chain with initial distribution

$$\mathbb{P}(\Pi_1 = [1]) = \mathbb{P}(\Pi_1 = [-1]) = 1/2$$

and the following transition probability matrices: for every $2 \leq t \leq n$,

$$M_t(\sigma, \sigma^{+,i}) = \mathbb{P}(\Pi_t = \sigma^{+,i} | \Pi_{t-1} = \sigma) = \frac{1}{2t},$$
$$M_t(\sigma, \sigma^{-,i}) = \mathbb{P}(\Pi_t = \sigma^{-,i} | \Pi_{t-1} = \sigma) = \frac{1}{2t},$$

where

$$\sigma^{+,i} := [\sigma_1, ..., \sigma_{i-1}, t, \sigma_i, ..., \sigma_{t-1}],$$

$$\sigma^{-,i} := [\sigma_1, ..., \sigma_{i-1}, -t, \sigma_i, ..., \sigma_{t-1}]$$

and $M_t(\sigma,\sigma')=0$ for every other $\sigma'\in B_t$.

It is easy to see that for every $t=1,...,n$, Π_t is a random signed permutation of t elements, uniformly chosen among the $2^t t!$ elements of B_t. In our case, the random variable of interest is $X_n:=c(\Pi_n)$, which we will show to be finite Markov chain embeddable. We construct a finite Markov chain $(Y_t)_{1\leq t\leq n}$ verifying the conditions (i),(ii) and (iii), as follows.

For every $t=1,...,n$ we denote by $K_{j,t}, j=1,...,n+1$ the random variables representing the number of cycles of length j in the breakpoint graph of the permutation Π_t. We also denote by L_t the length of the cycle in $G(\Pi_t)$ which contains the terminal point T. For every $t=1,...,n$ we obviously have $K_{j,t}=0$ for $j=t+2,...,n+1$ and

$$\sum_{j=1}^{t+1} jK_{j,t}=t+1, \sum_{j=1}^{t+1} K_{j,t}=c(\Pi_t).$$

We let

$$Y_t:=(L_t,K_{1,t},...,K_{n+1,t}), t=1,...,n.$$

We call Y_t *the type* of the permutation Π_t. For example, the permutation $\pi=[+3,-4,-2,+1,+5]$ from Fig. 1 is of type $(1,1,0,0,0,1,0)$. Note that for every $1\leq t\leq n$, Y_t takes values in the finite set

$$E_t=\left\{(\ell,k_1,...,k_{t+1},\underbrace{0,...,0}_{n-t}) : \ell\in\{1,...,t+1\}, \sum_{j=1}^{t+1} jk_j=t+1, k_\ell\geq 1\right\}.$$

Let us denote $\overrightarrow{k}:=(k_1,...,k_{n+1})$. We have

$$\mathbb{P}(c(\Pi_n)=x)=\mathbb{P}(Y_n\in C_x),$$

where, for every $x=1,2,...,n+1$,

$$C_x=\left\{(\ell,\overrightarrow{k}) : \sum_{j=1}^{n+1} k_j=x, \sum_{j=1}^{n+1} jk_j=n+1, k_\ell\geq 1\right\}.$$

We will show that $(Y_t)_{1\leq t\leq n}$ is an inhomogeneous Markov chain. The initial distribution of Y_1 is

$$\mathbb{P}(Y_1=(1,2,0,0,...,0))=\mathbb{P}(Y_1=(2,0,1,0,...,0))=1/2,$$

the case $Y_1=(1,2,0,0,...,0)$ corresponding to $\Pi_1=[+1]$ and the case $Y_1=(2,0,1,0,...,0)$ to $\Pi_1=[-1]$.

For $2\leq t<n$, write $Y_{t-1}=(\ell,\overrightarrow{k})$. Note that necessarily $k_\ell\geq 1$.

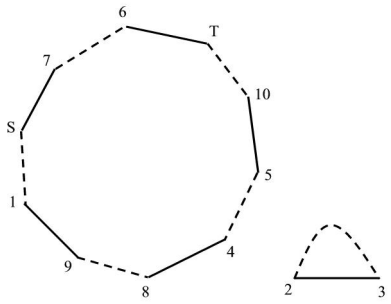

Fig. 2. The disjoint cycle decomposition of $G(\pi)$, for $\pi = [+4, -2, -1, +5, +3]$.

We have the following result.

Proposition 1: $(Y_t)_{1\leq t\leq n}$ is an inhomogeneous Markov chain of initial distribution

$$\mathbb{P}(Y_1 = (1, 2, 0, 0, ..., 0)) = \mathbb{P}(Y_1 = (2, 0, 1, 0, ..., 0)) = 1/2$$

and the following transition probabilities.

If $Y_{t-1} = (\ell, \overrightarrow{k})$, with $k_\ell \geq 1$, then the possible transitions are to $Y_t = (\ell', \overrightarrow{k'})$, where

(i) $\ell' = \ell + 1$ and $\overrightarrow{k'} = \overrightarrow{k} - e'_\ell + e'_{\ell+1}$, with probability $\ell/(2t)$;

(ii) $\ell' = j$, with $1 \leq j \leq \ell$, and $\overrightarrow{k'} = \overrightarrow{k} - e'_\ell + e'_j + e'_{\ell+1-j}$, with probability $1/(2t)$;

(iii) $\ell' = \ell + x + 1$, with $1 \leq x \leq t - \ell, x \neq \ell$ and $\overrightarrow{k'} = \overrightarrow{k} - e'_\ell - e'_x + e'_{\ell+x+1}$, with probability xk_x/t;

(iv) $\ell' = 2\ell + 1$ and $\overrightarrow{k'} = \overrightarrow{k} - 2e'_\ell + e'_{2\ell+1}$, with probability $\ell(k_\ell - 1)/t$,

where for each i, e'_i is the row vector having 1 at the i-th coordinate and 0 elsewhere.

Proof: For a permutation π of type $(\ell, \overrightarrow{k})$, we will show that $\mathbb{P}(Y_t = (\ell', \overrightarrow{k'})|\Pi_{t-1} = \pi)$ depends only on $\ell', \overrightarrow{k'}, \ell, \overrightarrow{k}$. This easily implies that $(Y_t)_{1\leq t\leq n}$ is a Markov chain.

Suppose now that $\Pi_{t-1} = \pi$, with π being of type $(\ell, \overrightarrow{k})$. In Fig. 2 we have the disjoint cycle decomposition of the breakpoint graph of the permutation $\Pi_5 = \pi = [+4, -2, -1, +5, +3]$. In this case we have $\ell = 5$ and $\overrightarrow{k} = (1, 0, 0, 0, 1, 0)$.

We will investigate the changes produced in the breakpoint graph when inserting the new element $\pm t$, at random, into one of the t possible positions of the permutation π, with the "$+$" sign with probability $1/2$ and the "$-$" sign with probability $1/2$.

The modifications concerning the dashed edges are simple. Disregarding the sign of $\pm t$, the dashed edge between $2(t-1)$ and T is deleted and replaced by a dashed edge between $2(t-1)$ and $2t-1$, and then another dashed edge is added between $2t$ and T (see for example Fig. 3).

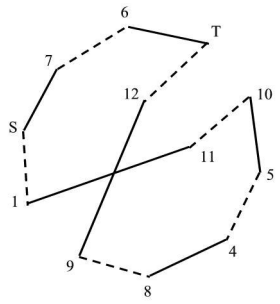

Fig. 3. For $\Pi_6 = [+4, -2, -1, +6, +5, +3]$ the cycle containing T grows to the length $\ell' = \ell + 1 = 6$. The element $+6$ is inserted into $\Pi_5 = [+4, -2, -1, +5, +3]$ in position $i = 4$, corresponding to the deletion of the solid edge 1–9.

Concerning the solid edges: we choose at random a solid edge among the t solid edges in the breakpoint graph of π, we delete it and then add two other solid edges to connect the two extremities of the deleted edge to $2t-1$ and $2t$ respectively, in one of the two possible ways. The choice of the solid edge to be deleted corresponds to the choice of the position in the permutation π where $\pm t$ is inserted. The way in which we connect the two extremities of the deleted edge to $2t-1$ and $2t$ respectively, corresponds to the sign of the element t.

More precisely, if we choose to insert the element $\pm t$ in the position i, where $2 \leq i \leq t-1$, then we will delete the solid edge between $(\pi(i-1))_R$ and $(\pi(i))_L$. If we insert $+t$, then we will add two solid edges between $(\pi(i-1))_R$ and $2t-1$ and between $2t$ and $(\pi(i))_L$. If we insert $-t$, then we will add two solid edges between $(\pi(i-1))_R$ and $2t$ and between $2t-1$ and $(\pi(i))_L$.

If we choose to insert the element $\pm t$ in the position 1, i.e. at the beginning of the permutation π, then we will delete the solid edge between S and $(\pi(1))_L$. If we insert $+t$ we add two solid edges between S and $2t-1$ and between $2t$ and $(\pi(1))_L$, and if we insert $-t$ we add two solid edges between S and $2t$, and between $2t-1$ and $(\pi(1))_L$.

If we choose to insert $\pm t$ in the position t, i.e. at the end of the permutation π, then we will delete the solid edge between $(\pi(t))_R$ and T. If we insert $+t$ we add two solid edges between $(\pi(t-1))_R$ and $2t-1$ and between $2t$ and T, and if we insert $-t$ we add two solid edges between $(\pi(t-1))_R$ and $2t$ and between $2t-1$ and T.

The cases (i) and (ii) in the statement correspond to the deletion of a solid edge from the cycle containing T, and the cases (iii) and (iv) correspond to the deletion of a solid edge belonging

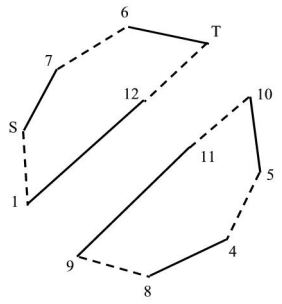

Fig. 4. For $\Pi_6 = [+4, -2, -1, -6, +5, +3]$ the cycle containing T splits into two and $\ell' = 3$. The element -6 is inserted into $\Pi_5 = [+4, -2, -1, +5, +3]$ in position $i = 4$, corresponding to the deletion of the solid edge 1–9.

to a cycle not containing T.

If we delete a solid edge belonging to the cycle of size ℓ which contains T, then we have two possible situations, depending on the deleted solid edge and on the permutation π. One situation is that, when we insert $+t$, the cycle containing T grows to the length $\ell + 1$ (see Fig. 3), and when we insert $-t$ it splits into two smaller cycles, of sizes which sum to $\ell+1$ (see Fig. 4). The other possible situation is the converse, i.e. when we insert $-t$ the cycle containing T becomes of size $\ell+1$ (see Fig. 5), and when we insert $+t$ it splits into two smaller cycles, of sizes which sum to $\ell + 1$ (see Fig. 6).

The event that the cycle containing T becomes of size $\ell + 1$ occurs with probability $\ell/(2t)$, because we have ℓ possible solid edges to choose in the cycle containing T. In the case when the cycle containing T splits into two cycles, the new size j of the cycle which will contain T is chosen at random, uniformly between 1 and ℓ. The size of the second cycle is then simply $\ell + 1 - j$. Each size j corresponds to a specific choice for the deleted solid edge, hence the event that the cycle containing T splits into two cycles and the size of the new cycle which will contain T becomes j, occurs with probability $1/(2t)$.

If we delete a solid edge from a cycle not containing T, then, disregarding the sign of t, this cycle will merge with the one containing T. If the cycle from which we have deleted a solid edge was of size x, then in the breakpoint graph of Π_t the cycle containing T will be of size $\ell + x + 1$.

In (iii), x represents the length of the cycle not containing T from which we choose a solid edge to be deleted. If $x \neq \ell$, the probability that this event occurs equals xk_x/t, because we

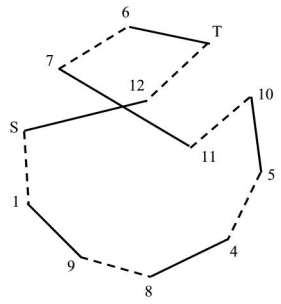

Fig. 5. For $\Pi_6 = [-6, +4, -2, -1, +5, +3]$ the cycle containing T grows to the length $\ell' = \ell + 1 = 6$. The element -6 is inserted into $\Pi_5 = [+4, -2, -1, +5, +3]$ in position $i = 1$, corresponding to the deletion of the solid edge S–7.

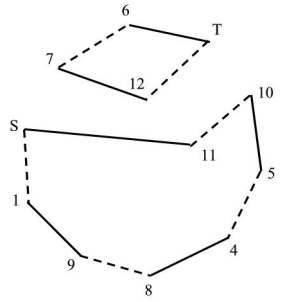

Fig. 6. For $\Pi_6 = [+6, +4, -2, -1, +5, +3]$ the cycle containing T splits into two and $\ell' = 2$. The element $+6$ is inserted into $\Pi_5 = [+4, -2, -1, +5, +3]$ in position $i = 1$, corresponding to the deletion of the solid edge S–7.

have k_x cycles of length x that we can choose, and each of them contains x solid edges.

The case (iv) corresponds to the case $x = \ell$, when we have only $k_\ell - 1$ possibilities to choose a cycle of size ℓ not containing T. ■

Proposition 1 describes the entries of the transition probability matrix P_t of the inhomogeneous Markov chain $(Y_t)_t$. As described in (3), we can therefore obtain the distribution of $c(\Pi_n)$ via the product of n transition matrices of this Markov chain.

C. Numerical results

We have implemented an iterative procedure which, for a given n, computes numerically the distribution of Y_n and then the distribution of $c(\Pi_n)$. At each step $t = 1, ..., n-1$, we compute the distribution of Y_{t+1} from the distribution of Y_t, using the transition probabilities described in Proposition 1. The complexity of our algorithm is $\mathcal{O}(n^2 \times p(n+1))$, where p is the partition function, i.e. for every positive integer m, $p(m)$ is the number of integer partitions of m. An

TABLE I

THE DISTRIBUTION OF $c(\Pi)$ FOR A RANDOM SIGNED PERMUTATION $\Pi \in B_{20}$.

k	1	2	3	4	5	6	7	8	9
p_k	0.19213	0.34805	0.27688	0.13047	0.04126	0.00938	0.00160	0.00021	0.00002

TABLE II

THE DISTRIBUTION OF $c(\Pi)$ FOR A RANDOM SIGNED PERMUTATION $\Pi \in B_{30}$.

k	1	2	3	4	5	6	7	8	9	10
p_k	0.15849	0.31791	0.28690	0.15704	0.05909	0.01639	0.0035	0.00059	0.00008	0.00001

asymptotic expression for $p(m)$ is given by

$$p(m) \sim \frac{\exp(\pi\sqrt{(2m)/3})}{4m\sqrt{3}}, \text{ as } m \to \infty.$$

In Table I and Table II we give the distribution of $c(\Pi)$ for a random uniform signed permutation Π of 20 and 30 elements, respectively. In the two tables p_k denotes the probability that $c(\Pi)$ takes the value k. For the values of k not appearing in the tables the corresponding probabilities are negligible. On a Pentium 4 processor, 3.1 Mhz, 512 Mb, the computation time was $13s$ for $n = 20$, $300s$ for $n = 30$, $4 \times 10^3 s$ for $n = 40$ and $4 \times 10^4 s$ for $n = 50$.

III. CONCLUDING REMARKS

In this article we have obtained the distribution of the number of alternating cycles in the breakpoint graph of a random signed permutation, in the form of a product of transition probability matrices of a certain finite Markov chain, using the finite Markov chain embedding technique. A drawback of our method is the fact that our Markov chain is inhomogeneous and of large dimension, which induces a high computational complexity.

A plan for a future work is to find a closed analytic formula for the exact distribution of the number of cycles in the breakpoint graph of a random signed permutation.

ACKNOWLEDGEMENTS

I would like to thank Etienne Pardoux, my thesis advisor, for all his support during this work, and Pierre Pontarotti, my second thesis advisor, for helpful biological discussions. I also wish to thank Anthony Labarre for explaining me his results on the number of cycles in the breakpoint graph of a random unsigned permutation.

This work was partially supported by the ANR MAEV under contract ANR-06-BLAN-0113.

REFERENCES

[1] D.A. Bader, B.M.E. Moret, and M. Yan, "A linear-time algorithm for computing inversion distance between signed permutations with an experimental study", Journal of Computational Biology, vol. 8, no. 5, pp. 483-491, Oct. 2001.

[2] V. Bafna and P. Pevzner, "Genome rearrangements and sorting by reversals", SIAM Journal of Computing, vol. 25, no. 2, pp. 272-289, Feb. 1996, doi:10.1137/S0097539793250627.

[3] A. Bergeron, C. Chauve, and Y. Gingras, "Formal models of gene clusters", in Bioinformatics Algorithms: Techniques and Applications, edited by I. Mandoiu and A. Zelikovsky, Wiley Series of Bioinformatics, 2008.

[4] A. Bergeron, S. Corteel, and M. Raffinot, "The algorithmic of gene teams", Lecture Notes in Computer Science, vol. 2452, pp. 464-476, Jan. 2002, doi:10.1007/3-540-45784-4_36.

[5] A. Bergeron, J. Mixtacki, and J. Stoye, "A unifying view of genome rearrangements", Proceedings of WABI 2006, Lecture Notes in Bioinformatics, vol. 4175, pp. 163-173, Sept. 2006, doi:10.1007/11851561_16.

[6] G. Blin and J. Stoye, "Finding nested common intervals efficiently", Lecture Notes in Bioinformatics, vol. 5817, pp. 5969, Sept. 2009, doi:10.1007/978-3-642-04744-2_6.

[7] S. Bocker, K. Jahn, J. Mixtacki, and J. Stoye, "Computation of median gene clusters", Journal of Computational Biology, vol. 16, no. 8, pp. 1085-1099, Aug. 2009, doi:10.1089/cmb.2009.0098.

[8] M. Bona and R. Flynn, "The average number of block interchanges needed to sort a permutation and a recent result of Stanley", Information Processing Letters, vol. 109, no. 16, pp. 927-931, July 2009, doi:10.1016/j.ipl.2009.04.019.

[9] A. Caprara, "On the tightness of the alternating-cycle lower bound for sorting by reversals", Journal of Combinatorial Optimization, vol. 3, no. 2-3, pp. 149-182, July 1999, doi:10.1023/A:1009838309166.

[10] V. Choi, C. Zheng, Q. Zhu, and D. Sankoff, "Algorithms for the extraction of synteny blocks from comparative maps", Lecture Notes in Bioinformatics, vol. 4645, pp. 277-288, Aug. 2007, doi:10.1007/978-3-540-74126-8_26.

[11] D.A. Christie, "Sorting permutations by block-interchanges", Information Processing Letters, vol. 60, no. 4, pp. 165-169, Nov. 1996, doi:10.1016/S0020-0190(96)00155-X.

[12] S. Corteel, G. Louchard, and R. Pemantle, "Common intervals in permutations", Discrete Mathematics and Theoretical Computer Science, vol. 8, no. 1, pp. 189216, 2006.

[13] E. Danchin and P. Pontarotti, "Statistical evidence for a more than 800-million-year-old evolutionarily conserved genomic region in our genome", Journal of Molecular Evolution, vol. 59, no. 5, pp. 587-597, Nov. 2004.

[14] J-P. Doignon and A. Labarre, "On Hultman numbers", Journal of Integer Sequences, vol. 10, Article 07.6.2, 2007.

[15] D. Durand and D. Sankoff, "Tests for gene clustering", Journal of Computational Biology, vol. 10, no. 3-4, pp. 453-482, June 2003, doi:10.1089/10665270360688129.

[16] J.C. Fu and M.V. Koutras, "Distribution theory of runs: a Markov chain approach", Journal of the American Statistical Association, vol. 89, no. 427, pp. 1050-1058, Sept. 1994.

[17] S. Grusea, "Measures for the exceptionality of gene order in conserved genomic regions", to appear in Advances in Applied Mathematics, doi:10.1016/j.aam.2010.02.002.

[18] S. Hannenhalli and P. Pevzner, "Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals", Journal of the ACM, vol. 46, no. 1, pp. 127, Jan. 1999.

[19] R. Hoberman and D. Durand, "The incompatible desiderata of gene cluster properties", Lecture Notes in Bioinformatics, Vol. 3678, pp. 73-87, Sept. 2005, doi:10.1007/11554714.

[20] R. Hoberman, D. Sankoff, and D. Durand, "The statistical analysis of spatially clustered genes under the maximum gap criterion", Journal of Computational Biology, vol. 12, no. 8, pp. 1083-1102, Oct. 2005, doi:10.1089/cmb.2005.12.1083.

[21] H. Kaplan, R. Shamir, and R.E. Tarjan, "Faster and simpler algorithm for sorting signed permutations by reversals", Proceedings of the 8th Annual ACM-SIAM Symposium on Discrete Algorithms, ACM Press, pp. 344-351, 1997.

[22] J. Kececioglu and D. Sankoff, "Exact and approximation algorithms for sorting by reversals, with application to genome rearrangement", Algorithmica, vol. 13, no. 1-2, pp. 180–210, Feb. 1995, doi:10.1007/BF01188586.

[23] Z. Li, L. Wang, and K. Zhang, "Algorithmic approaches for genome rearrangement: a review", IEEE Transactions on Systems, Man and Cybernetics, Part C, vol. 36, no. 5, pp. 636-648, Sept. 2006.

[24] J. Nadeau and B. Taylor, "Lengths of chromosomal segments conserved since divergence of man and mouse", Proceedings of the National Academy of Sciences USA, vol. 81, pp. 814818, 1984.

[25] P. Pevzner and G. Tesler, "Genome rearrangements in mammalian evolution: lessons from human and mouse genomes", Genome Research, vol. 13, no. 1, pp. 37-45, Jan. 2003.

[26] N. Raghupathy and D. Durand, "Gene cluster statistics with gene families", Molecular Biology and Evolution, vol. 26, no. 5, pp. 957968, Jan. 2009, doi:10.1093/molbev/msp002.

[27] D. Sankoff and L. Haque, "Power boosts for cluster tests", Lecture Notes in Bioinformatics, vol. 3678, pp. 121-130, Dec. 2005, doi:10.1007/11554714_11.

[28] D. Sankoff and L. Haque, "The distribution of genomic distance between random genomes", Journal of Computational Biology, vol. 13, no. 5, pp. 1005-1012, June 2006, doi:10.1089/cmb.2006.13.1005.

[29] K.M. Swenson, Y. Lin, V. Rajan, and B.M.E. Moret, "Hurdles hardly have to be heeded", In Proceedings of RECOMB-CG 2008, Lecture Notes in Bioinformatics, vol. 5267, pp. 239-249, 2008, doi:10.1007/978-3-540-87989-3_18.

[30] K.M. Swenson, V. Rajan, Y. Lin, and B.M.E. Moret, "Sorting signed permutations by inversions in $\mathcal{O}(n \log n)$ time", Lecture Notes in Computer Science, vol. 5541, pp. 386-399, May 2009, doi:10.1007/978-3-642-02008-7_28.

[31] E. Tannier, A. Bergeron, and M-F. Sagot, "Advances on sorting by reversals", Discrete Applied Mathematics, vol. 155, no. 6-7, pp. 881-888, April 2007, doi:10.1016/j.dam.2005.02.033.

[32] W. Xu, "The distance between randomly constructed genomes", Proceedings of the 5th Asia-Pacific Bioinformatics Conference, pp. 227-236, Oct. 2006, doi:10.1142/9781860947995_0025.

[33] W. Xu, C. Zheng, and D. Sankoff, "Paths and cycles in breakpoint graph of random multichromosomal genomes", Journal of Computational Biology, vol. 14, no. 4, pp. 423435, May 2007, doi:10.1089/cmb.2007.A004.

[34] S. Yancopoulos, O. Attie, and R. Friedberg, "Efficient sorting of genomic permutations by translocation, inversion and block interchange", Bioinformatics, vol. 21, no. 16, pp. 3340-3346, Aug. 2005, doi:10.1093/bioinformatics/bti535.

[35] Q. Zhu, Z. Adam, V. Choi, and D. Sankoff, "Generalized gene adjacencies, graph bandwidth, and clusters in Yeast evolution", Lecture Notes in Bioinformatics, vol. 4983, pp. 134-145, April 2008, doi:0.1007/978-3-540-79450-9_13.