Acta Univ. Sapientiae, Mathematica, **1**, 1 (2009) 5–20

On the expectation and variance of the reversal distance

László A. Székely
Department of Mathematics
University of South Carolina
Columbia, SC 29208, USA
email: szekely@math.sc.edu

Yiting Yang
Department of Mathematics
University of South Carolina
Columbia, SC 29208, USA
email: yang36@mailbox.sc.edu

Abstract. We give a pair of well-matched lower and upper bounds for the expectation of reversal distance under the hypothesis of random gene order by investigating the expected number of cycles in the breakpoint graph of linear signed permutations. Sankoff and Haque [9] proved similar results for circular signed permutations based on approximations based on a slightly different model; while our approach is discrete. We also provide an near-tight upper bound for the variance of reversal distance, which gives information on the distribution of reversal distance.

1 Introduction

In the late 1980s, Jeffrey Palmer[6] and his colleagues compared the mitochondrial genomes of cabbage and turnip, which are very closely related. To their surprise, these genomes, which are almost identical in gene sequences, differ dramatically in gene order. This discovery and many other studies in the last decade convincingly proved that genome rearrangements represent a common mode of molecular evolution.

A framework of possible models to study genome rearrangements is to represent genomes as *signed permutations* of genes and compute their distances based on the minimum number of certain operations (evolutionary events)

AMS 2000 subject classifications: 05C90, 05A16, 92B05
Key words and phrases: Reversal distance, breakpoint graph.

needed to transform one permutation into another. Under these models, the shorter the distance, the closer the genomes are.

In general, genes are represented as integers from 1 to n, and the genome is represented by a permutation π: $\{1, 2, \dots n\} \mapsto \{1, 2, \dots n\}$ by $(\pi_1\pi_2 \dots \pi_n)$, where π_i denotes $\pi(i)$.

Permutations (more precisely, at this point, the permuted elements) may get *signed*, reflecting whether a gene or its mirror image is present. In this case each entry π_i has a positive or negative sign to model the orientation of genes. We denote the set of all permutations of size n by S_n and the set of all signed permutations by $\overline{S_n}$, respectively. Clearly $|S_n| = n!$ and $|\overline{S_n}| = 2^n n!$. We call $\pi_i\pi_{i+1} \dots \pi_j$, where $(1 \leq i \leq j \leq n)$, a *segment* of the permutation $(\pi_1\pi_2 \dots \pi_n)$.

An extensively studied operation on genomes is *reversal*. A reversal is an operation that reverses the order of the genes on a certain segment of the permutation of S_n - this operation is usually called *unsigned reversal*. We avoid them in this paper. Reversals are also considered acting on $\overline{S_n}$, in this case they are called *signed reversals*, and they also change the sign of the genes in the segment which was reversed. (There are other operations considered in the literature that correspond to evolutionary events, such as transposition, block interchange, transversal and translocation, see [11].) In 1995, Hannenhalli and Pevzner [4] discovered an elegant polynomial time algorithm to compute the signed reversal distance of signed permutations. However, in 1999 Caprara [2] showed that computing the unsigned reversal distance is NP-hard. Even before that, in 1996, Bafna and Pevzner [1] gave a 1.5-approximation algorithm to compute the unsigned reversal distance.

1.1 Reversals

The formal definition for a signed reversal on a signed permutation follows. A *signed permutation* is a bijection of the set $[-n, n] \setminus \{0\}$ onto itself such that $\pi(-a) = -\pi(a)$ for all $a \in [-n, n] \setminus \{0\}$ holds. It is easy to see that these bijections make a group for the composition of bijections as the group operation. This group is usually known as the group of "signed permutations" on $[n]$, or as the *hyper-octahedral group* of rank n. A signed permutation can be (uniquely) represented by the sequence of values it assigns to $1, 2, ..., n$ — this is how we described them earlier. The signed permutation in the previous section is the array of images of $1, 2, ..., n$ under the bijection. We identify the group with $\overline{S_n}$. The *identity* signed permutation assigns values $1, 2, 3, \dots, n$ to $1, 2, 3, \dots, n$ in this order. We denote the identity of $\overline{S_n}$ by id. For any

$0 \leq i < j \leq n$, we define a signed reversal $\rho_{i,j}$ as a signed permutation, whose values on $1, 2, ..., n$ are $1, 2, ..., (i-2), (i-1), -j, ..., -i, (j+1), (j+2), ..., n$. Observe that its action is

$$\rho_{i,j}(\pi) = \pi \circ \rho_{i,j},$$

and the values of this signed permutation on the sequence $1, 2, ..., n$ are

$$\pi_1, \ldots, \pi_{i-1}, -\pi_j, \ldots, -\pi_i, \pi_{j+1} \ldots, \pi_n.$$

For $\pi_1, \pi_2 \in \overline{S_n}$, the *reversal distance* of π_1 from π_2 is the smallest k such that

$$\pi_1 = \pi_2 \circ \rho_{i_1,j_1} \circ \rho_{i_2,j_2} \circ \cdots \circ \rho_{i_k,j_k}, \tag{1.1}$$

where the ρ's are reversals. As reversals have order two, the reversal distance is symmetric. We define the reversal distance of a signed permutation π to be the reversal distance of π and the identity id, and denote it by $d(\pi)$.

We are interested in the expected reversal distance of two random signed permutations selected from the uniform distribution, π_1 and π_2. It follows from (1.1), that the reversal distance of π_1 and π_2 is the same as the reversal distance of $\pi_2^{-1} \circ \pi_1$ and id. Furthermore, $\pi_2^{-1} \circ \pi_1$ is equidistributed with π_1 and π_2. Therefore it is sufficient to compute or estimate the expected reversal distance of a random signed permutation selected from the uniform distribution.

1.2 Breakpoint graph

An efficient tool, widely applied in the research of genome rearrangement is the *breakpoint graph*.

We define the breakpoint graph $G(\pi)$, together with its layout, for a signed permutation $\pi = (\pi_1 \pi_2 \ldots \pi_n)$ as follows. If the entry π_i has positive sign, replace it by two vertices π_i^l, π_i^r in this order, and if the entry is negative, by π_i^r, π_i^l. Put these vertices in the order of π with two endpoints, 0^r on the extreme left and $(n+1)^l$ on the extreme right added—these vertices do not have a left (resp. right) companion. Connect any two vertices, which are consecutive in the layout (other than π_i^l and π_i^r from the same π_i) by a *black edge*, and connect i^r and $(i+1)^l$ by a *gray edge* for $0 \leq i \leq n$.

For convenience we call x the *value* of the vertex x^a, and a, which can be l or r, the *direction* of x^a. We call the vertices with the same value *conjugates*. For instance, the value of 3^r is 3, its direction is r, and it is the conjugate of 3^l, a relationship that we denote by $3^r = \overline{3^l}$. We define the *sign-function* s on $\{l, r\}$ such that $s(l) = 1$ and $s(r) = -1$. We will call the pair of vertices i^r and

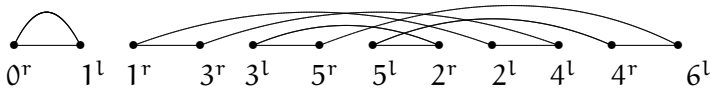

Figure 1: The breakpoint graph of π=(1,−3,−5,−2,4), straight lines are black edges and curved arcs are gray edges.

$(i+1)^l$, the *mates* of each other. For convenience, we denote the set of black edges of $G(\pi)$ by $B(\pi)$, and the vertex set $\{0^r, 1^l, 1^r, \ldots, n^l, n^r, (n+1)^l\}$ by V_n .

Each vertex is *adjacent* to exactly one black and one gray edge, so there is a *unique decomposition* of $G(\pi)$ into disjoint cycles of alternating edge colors. By the *length* of a cycle we mean the number of black edges it contains. We say that two gray edges g_1 and g_2 *cross*, if g_1 links vertices a and c, g_2 links vertices b and d, but these vertices are ordered a, b, c, d in $G(\pi)$. If g_1 and g_2 are crossing gray edges, and the cycle C_i contains the edge g_i for $i = 1, 2$, then we say that the two cycles, are *connected*. There is a finest equivalence relation on the set of cycles of $G(\pi)$, in which pairs of connected cycles fall in one class. A *component* of $G(\pi)$ is a class of this equivalence relation.

Using the breakpoint graph, Hannenhalli and Pevzner showed that the minimum number of reversals necessary to transform a signed permutation π to id is:

$$d(\pi) = n + 1 - c(\pi) + h(\pi) + fr(\pi), \tag{1.2}$$

where $c(\pi)$ is the number of cycles in the breakpoint graph, $h(\pi)$ is the number of hurdles, which are some special components, and $fr(\pi)$ takes value 1 or 0 based on whether G is a fortress. Caprara[3] showed that the probability of a random signed permutation of length n containing a hurdle is $\Theta(n^{-2})$ and the probability of a random signed permutation of length n including a fortress is $\Theta(n^{-15})$. Recently Swenson et al. [10] simplified Caprara's proof. Based on two facts above, in approximations for $d(\pi)$, the terms $h(\pi)$ and $fr(\pi)$ are often dropped. Hence we can easily find approximation for the distribution of reversal distance, if we find approximation to the distribution of the number of cycles in the breakpoint graph.

The permutations (unsigned or signed) that we discussed so far, are referred to as *linear permutations*, in order to distinguish them from the *circular permutations* which are arrangements of $\{1, 2, \ldots, n\}$ (or of $\{\pm 1, \pm 2, \ldots, \pm n\}$) along a cycle. Mathematically, the circular permutations are just the equivalence classes of linear permutations under rotation. The total number of circular unsigned permutations is $(n-1)!$ and the total number of circular signed permutations is $2^{n-1}(n-1)!$. As genomes or chromosomes can be linear or circular, the circular analogues of all concepts that we discussed so far are also relevant for bioinformatics. There exist concepts of reversals on circular permutations and breakpoint graphs of circular permutations, signed or unsigned. Hannenhalli and Pevzner [4] also computed that reversal distance of a circular permutation π of size n from the $\mathtt{id}$, an analogue of (1.2).

Sankoff and Haque [9] investigated the distribution of reversal distance among two randomly and uniformly selected circular signed permutations. They derived the expected number of cycles in the graph obtained by two random matchings on $2n$ vertices and claimed it approximately equals to the expected number of the cycles in the breakpoint graph of two random circular signed permutations. The precision of this approximation hinges on results of Kim and Wormald [5], which works with proof for sufficiently large n only (see the $\kappa < n/40$ condition in [9]). Depending this approximation, Sankoff and Haque [9] in their further calculations use continuous limit distributions to approximate the discrete probabilities in question, and put emphasis on the plotted simulation results as evidence for the result. Personal communication from Friedberg is cited in [9] as source for an asymptotic formula for the expected number of cycles in the breakpoint graph of circular (unsigned or signed) permutations.

It is expected that for linear signed permutations one would get similar results. Sankoff and Haque [9] claim that their approach extends to linear signed permutations, but do not give any specifics. The goal of this paper is to achieve such results, with rigorous and complete proofs, using a more discrete approach. We will give matching lower and upper bounds for the expected number of cycles in the breakpoint graph of a randomly and uniformly selected signed permutation.

We are not aware of any earlier results on the variance of the number of cycles in the breakpoint graph of a randomly and uniformly selected signed permutation, either linear or circular. The expectation result is analogous with the expected number of cycles in an unsigned linear permutation, which is also logarithmic, but the analogy fails for the variance, which is still logarithmic for

an unsigned linear permutation [8], but jumps to $\Theta(\log^2 n)$ for signed linear permutations.

1.3 B-cycle and test graph

Definition 1 *We call an alternating edge colored cycle in black and gray on a subset of the vertex set* V_n *a* B-cycle*, if there exists a signed permutation, whose breakpoint graph has it among its cycles.*

We will denote by [] the black edges and by () the gray edges. Notice that not every alternating colored cycle is a B-cycle. For example $[2^r, 5^l](5^r, 6^l)$ $[6^l, 1^r](1^r, 2^l)[2^l, 5^l](5^l, 4^r)[4^r, 3^l](3^l, 2^r)$ is a alternating colored cycle, but not a B-cycle, because if it is a B-cycle $[2^r, 5^r]$ determined 2 is followed by -5 or 5 is followed by -2 in the permutations whose breakpoint graphs contain this cycle, while $[2^l, 5^l]$ determined -2 is followed by 5 or -5 is followed by 2. Here we define another kind of auxiliary graph.

Definition 2 *Let* $\mathcal{E}$ *be a partial matching on the vertex set* V_n*. If for any* i*,* $(i^l, i^r) \notin \mathcal{E}, 0 < i < n+1$ *and* $(0^r, (n+1)^l) \notin \mathcal{E}$*, then we call* $\mathcal{E}$ *as a* standard partial matching*.*

Definition 3 *Let* $\mathcal{E}$ *be a standard partial matching on the vertex set* V_n*. The* test graph *of* $\mathcal{E}$*, denoted by* $T(\mathcal{E})$ *is defined as follows:*

- *The vertex set is* V_n*.*
- *The edge set consists of all the edges in* $\mathcal{E}$ *and the edges* $\langle i^l, i^r \rangle$ *for all* $0 < i < n+1$ *and* $\langle 0^r, (n+1)^l \rangle$*.*

In a test graph, we call the edges in $\mathcal{E}$ as *real edges* denoted by [] and the edges $\langle i^l, i^r \rangle, 0 < i < n+1$ and $\langle 0^r, (n+1)^l \rangle$ as *imaginary edges*. Notice that in a test graph each vertex is incident to one imaginary edge and at most one real edge; henceforth the test graph is composed of alternating cycles and paths. These pathes begin and end with imaginary edges. By the *length* of a cycle or a path we will understand the number of real edges in the cycle or path. We describe below a condition to tell whether a standard partial matching is a subset of the set of black edges of the breakpoint graph of some signed permutation.

Theorem 1.1 *Let* $\mathcal{E}$ *be a standard partial matching on the vertex set* V_n*. Then* $\mathcal{E}$ *is a subset of the black edges of the breakpoint graph of some permutation if and only if the test graph* $T(\mathcal{E})$ *is cycle-free (consists of paths only), or is a single cycle of length* $n+1$*.*

Proof. It is obvious that the test graph $T(B(\pi))$ for the set of all black edges of the breakpoint graph of some permutation π is an alternating cycle of length $n+1$. If $\mathcal{E}$ is a subset of $B(\pi)$, then $T(\mathcal{E})$ is a subgraph of $T(B(\pi))$. Since $T(B(\pi))$ is a cycle of length $n+1$, its subgraphs have to be itself or a set of pathes.

If $T(\mathcal{E})$ is a cycle of length $n+1$, we begin to read the numbers from 0^r to $(n+1)^l$ along the longer side of this cycle, and take the positive sign to the number if i^l proceed $i^r, 0 < i < n+1$, otherwise the negative sign. Thus we obtain a signed permutation, such that $\mathcal{E}$ is exactly the black edges set of this signed permutation. If $T(\mathcal{E})$ consists of pathes, we can connect them into an $n+1$- cycle, and from that, like above, we can get a signed permutation. Clearly, $\mathcal{E}$ is a subset of the black edge set of this signed permutation. ∎

2 Expected number of cycles

Selecting a signed permutation randomly and uniformly with probability $\frac{1}{2^n n!}$, the number of cycles in its breakpoint graph, c, will be a random variable and we are interested in the expectation and variance of this random variable c.

In order to get the expected number of cycles $E[c]$, it is enough to get the expected number of cycles of fixed lengths and sum them up, according to the linearity of expectation.

Lemma 1 *Let O_t be any B-cycle of length t, where $0 < t < n+1$. The probability that a randomly and uniformly selected signed permutation contains O_t is $p_t = \frac{(n-t)!}{2^t n!}$.*

Proof. We have to count how many signed permutations contain O_t as a cycle in their breakpoint graph. Since each cycle is determined by its black edges, we only have to count the signed permutations whose breakpoint graph contains the black edges of O_t. Let $B(O_t)$ be the set of black edges of O_t. Since O_t is a B-cycle of length less than $n+1$, from Theorem 1.1, $T(B(O_t))$ consists of alternating paths. The test graph of the whole black edge set of any breakpoint graph that contains $B(O_t)$ is a cycle of length $n+1$ that contains these paths as subgraphs. Just as in the proof of Theorem 1.1, we read the permutations from these cycles and we will observe that

(A) Let $P = \langle x_1^{a_1}, \overline{x_1^{a_1}}\rangle[\overline{x_1^{a_1}}, x_2^{a_2}]\dots[\overline{x_k^{a_k}}, x_{k+1}^{a_{k+1}}]\langle x_{k+1}^{a_{k+1}}, \overline{x_{k+1}^{a_{k+1}}}\rangle$ be a path of length k which does not contain 0^r and $(n+1)^l$ in $T(B(O_t))$. Then either the segment $(s(a_1)\cdot x_1, s(a_2)\cdot x_2 \dots\ s(a_{k+1})\cdot x_{k+1})$ or the segment $(-s(a_{k+1})\cdot x_{k+1}, -s(a_k)\cdot x_k, \dots - s(a_1)\cdot x_1)$ lies in the permutations whose breakpoint

graphs contain O_t depending on the direction of reading. Here notice the length of these segments is $k+1$.

For instance the path $\langle 3^l, 3^r\rangle[3^r, 5^l]\langle 5^l, 5^r\rangle[5^r, 7^r]\ \langle 7^r, 7^l\rangle[7^l, 4^r]\langle 4^r, 4^l\rangle$ is a path of length 3. A permutation whose breakpoint graph contains the black edges $[3^r, 5^l]$, $[5^r, 7^r]$ and $[7^l, 4^r]$ must contain the segment $(3, 5, -7, -4)$ or $(4, 7, -5, -3)$.

(B) Let $P = \langle x_1^{a_1}, \overline{x_1^{a_1}}\rangle[\overline{x_1^{a_1}}, x_2^{a_2}]\dots[\overline{x_i^{a_i}}, (n+1)^l]\langle(n+1)^l, 0^r\rangle[0^r, x_{i+1}^{a_{i+1}}]\langle x_{i+1}^{a_{i+1}}, \overline{x_{i+1}^{a_{i+1}}}\rangle\dots[\overline{x_{k-1}^{a_{k-1}}}, x_k^{a_k}]\langle x_k^{a_k}, \overline{x_k^{a_k}}\rangle$ be a path of length k which contains 0^r and $(n+1)^l$ in $T(B(O_t))$, then P determines that all the permutations whose breakpoint graphs contain O_t must begin with $s(a_{i+1})\cdot x_{i+1}, s(a_{i+2})\cdot x_{i+2}\dots s(a_k)\cdot x_k)$ and end with $(s(a_1)\cdot x_1, s(a_2)\cdot x_2, \dots s(a_i)\cdot x_i)$.

For example, the path $\langle 6^l, 6^r\rangle[6^r, 4^l]\langle 4^l, 4^r\rangle[4^r, 8^l]\langle 8^l, 0^r\rangle[0^r, 3^l]\langle 3^l, 3^r\rangle[3^r, 2^r]$ $\langle 2^r, 2^l\rangle$ implies that the permutation whose breakpoint graph contains these black edges in the path must begin with $(3, -2)$ and end with $(6, 4)$.

Let l_i be the number of paths of length i which do not contain vertices 0^r and $(n+1)^l$ in $T(B(O_t))$. Here $0 < i < n$ for a path of length n or $n+1$ must contain 0^r and $(n+1)^l$.

Case 1. The length of the path containing 0^r and $(n+1)^l$ is 0. In this case, we have $\sum_{i=1}^{n-1} i\cdot l_i = t$. From observation (A), each path of length i determine one segment of length $i+1$ in the permutation . So the number of permutations whose breakpoint graphs contain O_t is

$$2^{n-\sum_{i=1}^{n-1}(i+1)l_i+\sum_{i=1}^{n-1} l_i}(n-\sum_{i=1}^{n-1}(i+1)l_i+\sum_{i=1}^{n-1} l_i)!,$$

which turns out to be just $2^{n-t}(n-t)!$.

Case 2. The length of the path containing 0^r and $(n+1)^l$ is $s > 0$. In this Case, we have $\sum_{i=1}^{n-1} i\cdot l_i = t-s$. From observation (B), the start and end segments of the permutations of total length s are fixed. So we only need to consider the number of permutations on the remained $n-s$ numbers. According to case 1, it should be $2^{(n-s)-(t-s)}((n-s)-(t-s))!$ which still equals to $2^{n-t}(n-t)!$.

So from Cases 1 and 2, we conclude that for any B-cycle, there are $2^{n-t}(n-t)!$ signed permutations containing it. Since $|\overline{S_n}|$ is $2^n n!$, we have $p_t = \frac{2^{n-t}(n-t)!}{2^n n!} = \frac{(n-t)!}{2^t n!}$. ∎

The following lemma is a basic fact which will be applied in the coming computation.

Lemma 2

$$\log(n+1) \leq \sum_{i=1}^{n} \frac{1}{i} \leq \log n + 1.$$

2.1 Upper bound

Lemma 3 *Let $c_t(\pi)$ be the number of different B-cycles of length t which do not contain 0^r and 1^l in the breakpoint graphs of permutation. Then*

$$c_t < \frac{2^{t-1} n!}{t(n-t)!}.$$

Proof. A B-cycle of length t without 0^r and 1^l can be written as a circular sequence of edges as

$$\{[x'_t, x_1], (x_1, x'_1), [x'_1, x_2], (x_2, x'_2) \ldots, [x'_{t-1}, x_t], (x_t, x'_t)\}$$

or

$$\{[x_1, x'_t](x'_t, x_t)[x_t, x'_{t-1}] \ldots, (x'_2, x_2), [x_2, x'_1](x'_1, x_1)\}$$

for it is undirected, where $x_i \in \{1^r, 2^l, 2^r \ldots n^l, n^r, (n+1)^l\}$ and x'_i is the mate of x_i. Observe that the vertex set of a B-cycle of length t is just t pairs of mates and it corresponds to the circular sequence of the second vertex of each black edge. The first sequence corresponds to the circular t-permutation $x_1 x_2 \ldots, x_t$ of the vertices set $V_n \setminus \{0^r, 1^l\}$, and the other corresponds to $\{x'_t, x'_{t-1} \ldots, x'_1\}$. So each B-cycle corresponds to two circular t-permutations and each circular t-permutation corresponds to at most one B-cycle. Notice that if a t-permutation corresponds to a B-cycle, there is no pair of mates both in the t-permutation. Now lets count the number of such permutations. Let's select x_i's one by one to get a linear t-permutation. We have $2n$ choices for x_1, $2n-2$ choices x_2 since x'_1 can not be selected $\ldots 2n-2t+2$ choices for x_t. Thus we have totally $\frac{2^t n!}{(n-t)!}$ such linear t- permutations i.e $\frac{2^t n!}{t(n-t)!}$ such circular t-permutation. Hence there are at most $\frac{2^{t-1} n!}{t(n-t)!}$ cycles of length t for each cycle corresponding to a pair of circular t-permutations. ∎

Theorem 2.2 *Let $c(\pi)$ be the random variable counting the number of cycles in the breakpoint graph of a random signed permutation π. Then*

$$E[c] \leq \frac{1}{2} \log n + \frac{3}{2}.$$

Proof. Since 0^r and 1^l are mates, they are contained in one cycle. Thus we have

$$E[c] \leq \sum_{i=1}^{n} c_i p_i + 1 \leq \sum_{t=1}^{n} \frac{2^{t-1}n!}{t(n-t)!} \cdot \frac{(n-t)!}{2^t n!} + 1 = \sum_{t=1}^{n} \frac{1}{2t} + 1 \leq \frac{1}{2}\log n + \frac{3}{2},$$

where the last inequality is obtained by Lemma 2. ∎

2.2 Lower bound

Definition 4 *Let O_t be a B-cycle of length t. Let O_{t+1} be a B-cycle of length $t+1$ on V_n obtained by replacing a black edge of O_t with an alternating path of two black edges and a gray edge. Then we call O_t the shadow of O_{t+1} and O_{t+1} the shade of O_t.*

Lemma 4 *For any i, $1 < i < n$, let c_i be the number of different B-cycles of length i among the breakpoint graphs of all the signed permutations, then we have*

$$(i+1)c_{i+1} \geq 2i(n-i)c_i.$$

Proof. We prove it by counting the set of ordered pairs $\mathcal{P} = \{(O_i, O_{i+1}) | O_i \text{ is a shadow of } O_{i+1}\}$.

For a given B-cycle of length $i+1$, we can replace an alternating path of two black edges and a gray edge by a black edge to get a new cycle which could be a shadow depending on whether the new cycle is a B-cycle. Since we have $i+1$ ways to select the alternating path of two black edges and a gray edge, each B-cycle has at most $i+1$ shadows which implies that $|\mathcal{P}| \leq (i+1)c_{i+1}$.

For a given B-cycle O_i of length i, its shades must lie in the set of cycles which are obtained by replacing one black edge with an alternating path of two black edges and a gray edge. We denote that set by $\mathcal{S}(O_i)$. Next let's consider how many of cycles in $\mathcal{S}(O_i)$ are B-cycles.

From Theorem1.1, we only need to count how many of the cycles in $\mathcal{S}(O_i)$ with the test graph of their black edges are cycle-free.

Assume that O_i is written as

$$\{[x_i', x_1], (x_1, x_1'), [x_1', x_2], (x_2, x_2') \ldots, [x_{i-1}', x_i], (x_i, x_i')\}.$$

we replace the black edge $[x_{k-1}', x_k]$ by the path$[x_{k-1}, y](y, y')[y', x_k]$. Clearly y can not be in O_i. What happens to the test graph $T(B(O_i))$ is the following: we delete the black edge $[x_{k-1}', x_k]$ and add two black edges $[x_{k-1}, y]$ and $[y', x_k]$

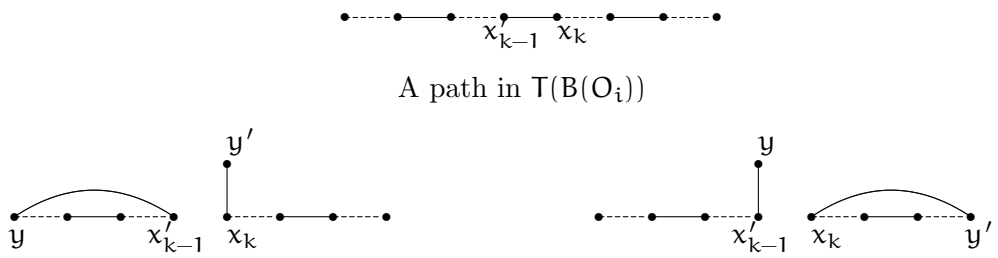

Figure 2: The path after deleting $[x'_{k-1}, x_k]$ and adding $[x'_{k-1}, y]$ and $[x_k, y']$

by selecting a vertex y. Let $[x'_{k-1}, x_k]$ be in a path in $T(B(O_i))$, only when y or y' is one of the two end points of this path could cause a cycle in the new test graph (see Fig. 2).

So here the "y" has at least $(2n+2) - (2i+2)$ choices. Hence $|\mathcal{P}| \geq i(2n-2i)c_i$. Thus we have

$$i(2n-2i)c_i \leq |\mathcal{P}| \leq (i+1)c_{i+1}.$$

which implies our lemma. ∎

We know that the number of B-cycles of length one is $n+1$. Recursively using Lemma 4 we get the following corollary:

Corollary 1 *Let c_t be the number of B-cycles of length t among the breakpoint graphs of all signed permutations of size n, where $1 < t < n+1$. Then*

$$c_t > \frac{2^{t-1}(n+1)(n-1)!}{t(n-t)!}.$$

Theorem 2.3 *Let $c(\pi)$ be the random variable by the number of cycles in the breakpoint graph of a random permutation π. Then*

$$E[c] \geq \frac{n+1}{2n}\log(n+1).$$

Proof.

$$E[c] \geq \sum_{i=1}^{n} c_i p_i \geq \sum_{i=1}^{n} \frac{2^{t-1}(n+1)(n-1)!}{t(n-t)!} \cdot \frac{(n-t)!}{2^t n!} \geq \frac{n+1}{2n}\log(n+1),$$

where the last inequality is obtained by Lemma 2. ∎

3 Variance of the number of cycles

Theorem 3.4 *Let $c(\pi)$ be the random variable counting the number of cycles in the breakpoint graph of of a randomly and uniformly selected signed permutation π, then*

$$E[c^2] \leq \frac{3}{4}\log^2 n + \frac{5}{2}\log n + \frac{7}{2}.$$

Proof. Let X_i be the random variable such that $X_i(\pi) = 1$ if $G(\pi)$ contains the B-cycle o_i and $X_i(\pi) = 0$ otherwise. Then we have:

- $c(\pi) = \sum_i X_i(\pi)$;
- $X_i(\pi)X_j(\pi) = 1$, if $G(\pi)$ contains the cycles o_i and o_j, $= 0$. Furthermore, if $o_i \cap o_j \neq \varnothing$, then $X_iX_j = 0$.
- $E[X_i^2(\pi)] = E[X_i(\pi)]$.

Hence

$$\begin{aligned} E[c^2] &= E[\sum_i X_i \sum_j X_j] = \sum_i E[X_i^2] + \sum_{\substack{i,j,\\ i\neq j}} E[X_iX_j] \\ &= E[c] + \sum_{\substack{i,j\\ i\neq j}} E[X_iX_j] = E[c] + \sum_{\substack{i,j\\ o_i\cap o_j=\varnothing}} E[X_iX_j] \end{aligned} \tag{3.1}$$

Let us be given two cycles A and B of length $a > 0$ and $b > 0$ with $A \cap B = \varnothing$. Then a permutation contains A and B only if it contains the $a+b$ black edges of them. From the proof of Lemma 1, we have there are $2^{n-a-b}(n-a-b)!$ permutations containing cycles A and B provided $a+b \leq n$. So the probability of a random permutation containing the cycles A and B is $\frac{2^{n-a-b}(n-a-b)!}{2^n n!}$ for $2 \leq a+b \leq n$. When $a+b = n+1$, the breakpoint graph of the permutation which contains A and B is uniquely determined by the black edges of A and B. Hence the probability of a permutation containing the cycles A and B is $\frac{1}{2^n n!}$ when $a+b = n+1$.

Now let us count how many ordered pairs of non-intersecting cycles have length $a > 0$ and $b > 0$. Since each cycle of length t is determined by a pair of circular t-permutations, the number of ordered pairs of non-intersecting cycles

which have length $a > 0$ and $b > 0$ is bounded by one fourth of the ways of getting a pair of circular a-permutation and b-permutation. Just as in the proof of Lemma 3, we could select a circular a-permutation with

$$\frac{(2n+2)(2n+2-2)\dots(2n+2-(2a-2))}{a} = \frac{2^a(n+1)!}{a(n+1-a)!}$$

ways and select a circular b-permutation which is not intersected with the a-permutation with

$$\frac{(2n+2-2a)(2n+2-2a-2)\dots(2n+2-2a-(2b-2))}{b} = \frac{2^b(n+1-a)!}{(n+1-a-b)!b}$$

ways. Thus totally we have at most

$$\frac{2^a(n+1)!}{a(n+1-a)!}\cdot\frac{2^b(n+1-a)!}{(n+1-a-b)!b} = \frac{2^{a+b}(n+1)!}{(n+1-a-b)!ab}$$

such pairs. Thus,

$$\begin{aligned}
\sum_{\substack{i,j\\ o_i\cap o_j=\varnothing}} E[X_iX_j] &\le \frac{1}{4}\sum_{a,b\ge 1}^{a+b\le n}\frac{2^{a+b}(n+1)!}{(n+1-a-b)!ab}\frac{2^{n-a-b}(n-a-b)!}{2^n n!}\\
&\quad+\frac{1}{4}\sum_{a,b\ge 1}^{a+b=n+1}\frac{2^{a+b}(n+1)!}{(n+1-a-b)!ab}\frac{1}{2^n n!}\\
&= \frac{1}{4}\sum_{a,b\ge 1}^{a+b\le n}\frac{n+1}{ab(n+1-a-b)}+\frac{1}{2}\sum_{a,b\ge 1}^{a+b=n+1}\frac{n+1}{ab}\\
&= \frac{1}{4}\sum_{t=2}^{n}\sum_{a=1}^{t-1}\frac{n+1}{a(t-a)(n+1-t)}+\frac{1}{2}\sum_{a=1}^{n}(\frac{1}{a}+\frac{1}{n+1-a})
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{4}\sum_{t=2}^{n}\sum_{a=1}^{t-1}\left(\frac{1}{a(t-a)}+\frac{1}{a(n+1-t)}+\frac{1}{(t-a)(n+1-t)}\right)+\sum_{a=1}^{n}\frac{1}{a}\\
&= \frac{1}{4}\sum_{t=2}^{n}\left((\frac{1}{t}+\frac{1}{n+1-t})\sum_{a=1}^{t-1}(\frac{1}{a}+\frac{1}{t-a})\right)+\sum_{a=1}^{n}\frac{1}{a}\\
&\le \sum_{t=2}^{n}\frac{1}{2}\left(\log(t-1)+1\right)\left(\frac{1}{t}+\frac{1}{n+1-t}\right)+(\log n+1)\\
&= \frac{1}{2}\sum_{t=2}^{n}\left(\frac{\log(t-1)}{t}+\frac{\log(t-1)}{n+1-t}\right)+\frac{1}{2}\sum_{t=2}^{n}\left(\frac{1}{t}+\frac{1}{n+1-t}\right)+(\log n+1)\\
&\le \frac{1}{2}\sum_{t=2}^{n}\left(\frac{\log(t-1)}{t-1}+\frac{\log(n-1)}{n+1-t}\right)+\frac{1}{2}(\log n+\log(n-1)+1)\\
&\quad +(\log n+1)\\
&\le \frac{1}{2}\left(\frac{\log^2 n}{2}+\log(n-1)(\log(n-1)+1)\right)\\
&\quad +\frac{1}{2}(\log n+\log(n-1)+1)+(\log n+1)\\
&\le \frac{3}{4}\log^2 n+\frac{5}{2}\log n+\frac{3}{2}.
\end{aligned}
$$

From (3.1) and Lemma 3, we have

$$E[c^2] \le \frac{1}{2}\log n+\frac{3}{2}+\frac{3}{4}\log^2 n+\frac{5}{2}\log n+\frac{3}{2}=\frac{3}{4}\log^2 n+3\log n+3.$$

∎

Theorem 3.5 *Let* $c(\pi)$ *be the random variable that counts the number of cycles in the breakpoint graph of* π*, then*

$$Var[c] \le \frac{1}{2}\log^2 n+3\log n+3.$$

Proof. For $Var[c]=E[c^2]-E[c]^2$, substitute the upper bound of $E[c^2]$ and the lower bound of $E[c]$ to obtain

$$\begin{aligned}\mathrm{Var}[c] &\le \frac{3}{4}\log^2 n + 3\log n + 3 - (\frac{n+1}{2n}\log(n+1))^2 \\ &\le \frac{3}{4}\log^2 n + 3\log n + 3 - (\frac{1}{2}\log(n))^2 \\ &= \frac{1}{2}\log^2 n + 3\log n + 3.\end{aligned}$$

We leave it to reader to verify that the calculations in Theorems 3.4 and 3.5 are asymptotically tight. ∎

4 Expectation and variance of the reversal distance

Recall that $h(\pi)$ is the number of hurdles and $fr(\pi)$ is the number of fortresses in the breakpoint graph of π. For a randomly and uniformly selected signed permutation π, we have $\mathrm{Prob}(h(\pi) \ge 1) = \Theta(\frac{1}{n^2})$ according to Swenson et al. [10]. There are at most n hurdles in the breakpoint graph of any permutation. So we have $E[h] = O(\frac{1}{n})$. Swenson et al. [10] also showed that $\mathrm{Prob}(fr(\pi) = 1) = \Theta(\frac{1}{n^{15}})$, which implies that $E[fr] = \Theta(\frac{1}{n^{15}})$ for $fr(\pi)$ only takes value 1 or 0. Hence from (1.2) we have

$$E[d(\pi)] = n + 1 - E[c(\pi)] + O(\frac{1}{n}),$$

which implyies the following theorem:

Theorem 4.6 $n+1-\frac{1}{2}\log n-\frac{3}{2}+O(\frac{1}{n}) \le E[d] \le n+1-\frac{n+1}{2n}\log(n+1)+O(\frac{1}{n})$.

Since $\mathrm{Prob}(h(\pi) \ge 1)$ and $\mathrm{Prob}(fr(\pi) = 1)$ are both very small, we can drop the terms $h(\pi)$ and $fr(\pi)$ in most of the cases when we compute $d(\pi)$. Let $\widetilde{d} = n + 1 - c(\pi)$. Then we have the following result as a consequence of Theorem 3.5:

Theorem 4.7

$$\mathrm{Var}[\widetilde{d}] \le \frac{1}{2}\log^2 n + 3\log n + 3.$$

Acknowledgment: *László A. Székely was supported in part by the NIH NIGMS contract 1 R01 GM078991-01, by the NSF DMS contract 0701111, by a Marie Curie Fellowship HUBI MTKD-CT-2006-042794, and by the 2007 Phylogeny program of the Isaac Newton Institute, Cambridge; Yiting Yang was supported in part by the NIH NIGMS contract 1 R01 GM078991-01 and by the NSF DMS contract 0701111.*

References

[1] V. Bafna and P.A. Pevzner, Genome rearrangements and sorting by reversals, *SIAM on Computing*, **25** (1996) 272–289.

[2] A. Caprara, Sorting Permutations by Reversals and Eulerian Cycle Decopostions, *SIAM J. Disc. Math.*, **12** (1999) 91–110.

[3] A. Caprara, On the tightness of the alternating-cycle lower bound for sorting by reversals, *J. Combin. Optimization*, **3** (1999) 149–182.

[4] S. Hannenhalli and P.A. Pevzner, Transforming cabbage into turnip: Polynomial algorithm for sorting signed permutations by reversals, *Journal of the ACM*, **46** (1999) 1–27.

[5] J. H. Kim, N. C. Wormald, Random matchings, which induce Hamiltonian cycles, and Hamiltonian decompositions of random graphs, *J. Combin. Theory* B **81** (2001) 20–44.

[6] J.D. Palmer, Intraspecific variation and multicircularity in Brassica mitochondrial DNAs, *Genetics*, **118** (1988) 341–351.

[7] P.A. Pevzner, *Computational Molecular Biology - An Algorithmic Approach*, The MIT Press, 2000.

[8] V. N. Sachkov, *Probabilistic methods in combinatorial analysis*, Encyclopedia of Mathematics and its Applications, vol. 56 Cambridge University Press, Cambridge, 1997.

[9] D.Sankoff and L. Haque, The Distribution of Genomic Distance between Random Genomes, *J. Comput. Biol.* **13** (2006) 1005–1012.

[10] K. M. Swenson, Y. Lin, V. Rajan and B. M. E. Moret, Hurdles hardly have to be heeded, *Proc. 6th RECOMB Workshop on Comparative Genomics* RECOMBCG'08, Springer, 241–251 (2008).

[11] Z. Li, L. Wang and K. Zhang, Algorithmic approaches for genome rearrangement: a review, *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews*, **36** (2006) 636–648.

Received: October 30, 2008