

Approximating the expected number of inversions given the number of breakpoints

Niklas Eriksen

Dept. of Mathematics,
Royal Institute of Technology,
SE-100 44 Stockholm, Sweden,
niklas@math.kth.se,

WWW home page: <http://www.math.kth.se/~niklas>

Abstract We look at a problem with motivation from computational biology: Given the number of breakpoints in a permutation (representing a gene sequence), compute the expected number of inversions that have occurred. For this problem, we obtain an analytic approximation that is correct within a percent or two. For the inverse problem, computing the expected number of breakpoints after any number of inversions, we obtain an analytic approximation with an error of less than a hundredth of a breakpoint.

1 Introduction

For about a decade, mathematicians and computer scientists have been studying the problem of inferring evolutionary distances from gene order. We are given two permutations (of a gene sequence) and want to calculate the evolutionary distance between them. The most common distance studied is the shortest edit distance: the least number of operations needed to transform one of the permutations into the other, given a set of allowed operations. The operations primarily considered are inversions, transpositions and inverted transpositions. For an overview, see Pevzner's book [11].

The most prominent result in this field is the Hannenhalli-Pevzner inversion distance [10]. They provide a closed formula for the minimal edit distance of *signed* permutations, a formula which can be computed in linear time [1]. Other results worth mentioning are the NP-hardness of computing the inversion distance for unsigned permutations (Caprara, [3]) and the $(1 + \varepsilon)$ -approximation of the combined inversion and transposition distance, giving weight 2 to transpositions (Eriksen, [6]).

Although incorporating both inversions and transpositions makes the model more realistic, it seems that the corresponding distance functions usually do not differ much [2,8]. Thus, the inversion edit distance seems to correspond quite well to the evolutionary edit distance, even though this is not *a priori* obvious.

However, there is another problem with using the Hannenhalli-Pevzner formula for the shortest edit distance using inversions: For permutations that are

quite disparate, the shortest edit distance is much shorter than the expected edit distance. The reason is that as the distance between the permutations increases, so does the probability that the application of yet another inversion will not increase the distance. Obtaining the true evolutionary distance is of course impossible, but it will usually be closer to the expected value of the edit distance than to the the shortest edit distance. Therefore we want to study the following problem.

Problem 1. Given the number of breakpoints between two signed, circular, permutations, compute the expected number of inversions giving rise to this number of breakpoints.

For unsigned permutations, this problem has been solved by Caprara and Lancia [4]. One easily realises that if gene g_2 is not adjacent to gene g_1 , then the probability that a random inversion will place g_2 next to g_1 is the same for all such permutations. From this observation, the expected number of breakpoints after i random inversions is not hard to come by. This contrasts with the signed case, where the sign of g_2 will affect the probability.

For signed permutations, the problem has been attacked independently in two sequences of papers. One sequence includes papers by Sankoff and Blanchette [12] and Wang [13]. They have calculated transition matrices for a Markov chain, using which the expected number of breakpoints, given that i random inversions have been applied, can be computed. By inverting this, the expected number of inversions can be computed. This is a fair solution in practice, but it does not give an analytical expression and it is slow for large problems.

The other sequence contains papers by Eriksson et al. [9] and Eriksen [7]. They have looked at the similar problem of computing the expected number of pairs of elements in the wrong order (inversion number) given that t random **adjacent** transpositions have been applied to an ordinary (unsigned) permutation. Again, the purpose is to invert the result. This is an analogue of the above case and the study was initiated to raise ideas on how to solve the inversion problem. In the latter paper, Markov chain transition matrices for this problem are briefly investigated and the conclusion is drawn that an approximation of the expected inversion number can be found if we can compute the two largest eigenvalues of the transition matrix. This raises the question if it is possible to compute the largest eigenvalues of the matrices found by Sankoff, Blanchette and Wang.

We will show that the answer is yes. In fact, we can compute most of the eigenvalues of these transition matrices and sufficient information about the eigenvectors to get a very good approximation of the expected number of inversions, given b breakpoints:

$$i_{\text{appr}}(b) = \frac{\log\left(1 - \frac{b}{n(1 - \frac{1}{2n-2})}\right)}{\log\left(1 - \frac{2}{n}\right)}.$$

In this paper, we will derive this formula. We then take a look at some adjustments that can be made to improve it. The formula above is the inverse

of an approximation of the expected number of breakpoints after i random inversions. If we drop the demand that the formula for the expected number of breakpoints should be analytically invertible, then we can provide a significantly better formula, with an error that rarely exceeds 0.01 breakpoints.

2 Preliminaries

We are using signed, circular permutations as a model for bacterial genomes. Each gene corresponds to a unique element. The two possible orientations of a gene corresponds to the sign (+ or -) of the corresponding element in the permutation.

An **inversion** is an operation on a permutation which takes out a segment of consecutive elements and insert it at the same place in reverse order, altering the sign of all elements in the segment.

Example 1. Let $\pi = [g_1 -g_4 g_3 -g_6 g_5 g_2]$ be our permutation. It is understood that g_1 follows directly after g_2 , since the permutation is circular. If we invert the segment $[g_3 -g_6 g_5]$ in π , we get $\pi' = [g_1 -g_4 -g_5 g_6 -g_3 g_2]$. Had we inverted the segment $[g_2 g_1 -g_4]$ (the complement of the previous segment), the result would have been the same.

A measure of the difference between two permutations π_1 and π_2 is the number of **breakpoints** between them. There is a breakpoint between two adjacent genes $g_i g_j$ in π_1 if π_1 contains neither of the segments $[g_i g_j]$ or $[-g_j -g_i]$.

Example 2. The permutations $\pi_1 = [g_1 -g_4 g_3 -g_6 g_5 g_2]$ and $\pi_2 = [g_1 g_4 -g_5 g_6 -g_3 g_2]$ are separated by three breakpoints. In π_1 these are between g_1 and $-g_4$, between $-g_4$ and g_3 , and between g_5 and g_2 .

3 Obtaining the formula

We will in this paper consider Markov chains for circular genomes of length n . At each step in the process, an inversion is chosen at random from a uniform distribution, and the inversion is applied to the genome. The states in the process correspond to the position of the gene g_2 as follows. We fix the first element g_1 and consider the various places where the element g_2 can be located, relative to g_1 . Each such position (with orientation) is considered a state in the Markov process. This makes $2n - 2$ states, since there are $n - 1$ positions and two possible orientations at each position.

The transition matrix for this process was presented in 1999 by Sankoff and Blanchette [12] and it was generalised to include transpositions and inverted transpositions in a 2001 WABI paper [13] by Li-San Wang.

Theorem 1. (*Sankoff and Blanchette [12] and Wang [13]*) *Consider the Markov process where the states corresponds to the positions of g_2 , with a possible minus sign which signals that g_2 is reversed. The states are ordered as $\{2, -2, 3, -3, \dots,$*

$n, -n\}$. At each step an inversion is chosen at random from a uniform distribution. Then the transition matrix is $\frac{1}{\binom{n}{2}}M_n$, where n is the length of the genome, and $M_n = (m_{ij})$ is given by

$$m_{ij} = \begin{cases} \min\{|u| - 1, |v| - 1, n + 1 - |u|, n + 1 - |v|\}, & \text{if } uv < 0; \\ 0, & \text{if } u \neq v, uv > 0; \\ \binom{|u|-1}{2} + \binom{n+1-|u|}{2}, & \text{otherwise.} \end{cases}$$

Here $u = (-1)^{i+1} \left(\lceil \frac{i}{2} \rceil + 1\right)$ and $v = (-1)^{j+1} \left(\lceil \frac{j}{2} \rceil + 1\right)$, that is, u and v are the (signed) positions of the states that corresponds to row i and column j , respectively.

The proof is straightforward — just count the number of inversions that move g_2 from position u to position v .

Example 3. Let $n = 4$. At position $i = 3, j = 4$ in M_4 , we have $u = 3$ and $v = -3$. Thus, $m_{34} = \min\{3 - 1, 3 - 1, 5 - 3, 5 - 3\} = 2$. The entire matrix is given by

$$M_4 = \begin{pmatrix} 3 & 1 & 0 & 1 & 0 & 1 \\ 1 & 3 & 1 & 0 & 1 & 0 \\ 0 & 1 & 2 & 2 & 0 & 1 \\ 1 & 0 & 2 & 2 & 1 & 0 \\ 0 & 1 & 0 & 1 & 3 & 1 \\ 1 & 0 & 1 & 0 & 1 & 3 \end{pmatrix}$$

The transition matrix M_n can be used to calculate the estimated number of breakpoints in a genome, given that i inversions have been applied. The reason is that the entry at $(1, 1)$ of M_n^i gives the probability p that g_2 , after i inversions, is positioned just after g_1 , where it does not produce a breakpoint. The probability of a breakpoint is the same after any gene, so the expected number of breakpoints after i inversions is $n(1 - p)$. This is the same as

$$n \left(1 - \frac{\bar{e}_1 M_n^i \bar{e}_1^T}{\binom{n}{2}^i} \right),$$

where $\bar{e}_1 = (1, 0, 0, \dots, 0)$.

Now, M_n is a real symmetric matrix, so we can diagonalise it as $M_n = V_n D_n V_n^T$, where D_n is a diagonal matrix with the eigenvalues of M_n on the diagonal, and V_n is the orthogonal matrix of eigenvectors. We write $\bar{v}_n = \bar{e}_1 V_n$. The expected number of breakpoints after i inversions, $b(i)$, is then given by

$$b(i) = n \left(1 - \frac{\bar{e}_1 M_n^i \bar{e}_1^T}{\binom{n}{2}^i} \right) = n \left(1 - \frac{\bar{e}_1 V_n D_n^i V_n^T \bar{e}_1^T}{\binom{n}{2}^i} \right) = n \left(1 - \frac{\bar{v}_n D_n^i \bar{v}_n^T}{\binom{n}{2}^i} \right).$$

This analysis proves our first result.

Theorem 2. Let $\bar{v}_n = (v_1, v_2, \dots, v_{2n-2}) = \bar{e}_1 V_n$ and let $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{2n-2}$ be the eigenvalues of M_n . Then the expected number of breakpoints after i random inversions in a genome with n genes can be written as

$$b(i) = n \left(1 - \frac{\sum_{j=1}^{2n-2} v_j^2 \lambda_j^i}{\binom{n}{2}^i} \right),$$

where $\sum v_j^2 = 1$.

Calculating how fast the expected number of breakpoints approaches the obvious limit $n \left(1 - \frac{1}{2n-2} \right)$ primarily amounts to calculating the eigenvalues of M_n . We will prove the following result, which contains the most important information about the eigenvalues.

Theorem 3. Let M_n , $n \geq 2$, be the matrix described above and $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{2n-2}$ its eigenvalues. Then $\lambda_1 = \binom{n}{2}$ and $\lambda_2 = \lambda_3 = \lambda_4 = \binom{n-1}{2}$. The coefficient v_1^2 of λ_1 in the sum above will be $\frac{1}{2n-2}$ and the sum of the coefficients of $\lambda_2 = \lambda_3 = \lambda_4$ will be $\frac{3}{4} - \frac{1}{2n-2}$. The smallest eigenvalue λ_{2n-2} is greater than or equal to $-\lfloor \frac{n}{2} \rfloor$.

In the appendix, we have gathered information on some of the other eigenvalues and their coefficients.

If we set $v_2^2 = 1 - v_1^2$ and $v_j = 0$ for all $j \geq 3$ in Theorem 3, we get the following approximation.

Corollary 1. The expected number of breakpoints given i inversions, $b(i)$, can be approximated by

$$b_{appr}(i) = n \left(1 - \frac{1}{2n-2} \right) \left[1 - \left(1 - \frac{2}{n} \right)^i \right]$$

such that

$$\lim_{i \rightarrow \infty} (b(i) - b_{appr}(i)) = 0.$$

By taking the inverse of this map, we obtain an approximation of the expected number of inversions, given that we observe b breakpoints.

Corollary 2. The expected number of inversions given b breakpoints, $i(b)$, can be approximated by

$$i_{appr}(b) = \frac{\log \left(1 - \frac{b}{n(1 - \frac{1}{2n-2})} \right)}{\log \left(1 - \frac{2}{n} \right)}.$$

The quality of these approximations will be investigated in the next section. We now give a concrete example of what is going on, followed by the proof of Theorem 3.

Example 4. For $n = 4$, the matrix M_n looks like

$$M_4 = \begin{pmatrix} 3 & 1 & 0 & 1 & 0 & 1 \\ 1 & 3 & 1 & 0 & 1 & 0 \\ 0 & 1 & 2 & 2 & 0 & 1 \\ 1 & 0 & 2 & 2 & 1 & 0 \\ 0 & 1 & 0 & 1 & 3 & 1 \\ 1 & 0 & 1 & 0 & 1 & 3 \end{pmatrix}$$

and has eigenvalues $\{6, 3, 3, 3, 2, -1\}$. It is obvious that both 6 and 3 are eigenvalues, since the row sums are all 6 (hence $(1, 1, 1, 1, 1, 1)$ is an eigenvector with eigenvalue 6) and subtracting 3 from the diagonal would make the rows 1 and 5 equal, as well as rows 2 and 6. If we diagonalise $M_4 = V_4 D_4 V_4^T$, we get

$$V_4 = \begin{pmatrix} -0.6132 & -0.1687 & -0.4230 & 0.4082 & 0.2887 & 0.4082 \\ 0.2835 & 0.1893 & -0.6835 & -0.4082 & -0.2887 & 0.4082 \\ 0.2516 & -0.4719 & 0.2175 & -0.4082 & 0.5774 & 0.4082 \\ 0.2516 & -0.4719 & 0.2175 & 0.4082 & -0.5774 & 0.4082 \\ 0.3615 & 0.6406 & 0.2055 & 0.4082 & 0.2887 & 0.4082 \\ -0.5351 & 0.2826 & 0.4660 & -0.4082 & -0.2887 & 0.4082 \end{pmatrix}$$

and

$$D_4 = \begin{pmatrix} 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 6 \end{pmatrix}$$

From this, we can calculate

$$\bar{v}_n = \bar{e}_1 V_n = (-0.6132, -0.1687, -0.4230, 0.4082, 0.2887, 0.4082),$$

and thus

$$b(i) = 4 \left(1 - \frac{0.1667 \cdot 6^i + 0.5833 \cdot 3^i + 0.1667 \cdot 2^i + 0.0833 \cdot (-1)^i}{6^i} \right).$$

Our approximation (from Corollary 1) yields

$$b_{\text{appr}}(i) = \frac{10}{3} \left(1 - \frac{1}{2^i} \right),$$

the inverse of which is (from Corollary 2)

$$i_{\text{appr}}(b) = \frac{\log \left(1 - \frac{3b}{10} \right)}{\log \frac{1}{2}}.$$

We also have

$$M_5 = \begin{pmatrix} 6 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 6 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 4 & 2 & 0 & 2 & 0 & 1 \\ 1 & 0 & 2 & 4 & 2 & 0 & 1 & 0 \\ 0 & 1 & 0 & 2 & 4 & 2 & 0 & 1 \\ 1 & 0 & 2 & 0 & 2 & 4 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 6 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 6 \end{pmatrix}$$

with eigenvalues $\{10, 6, 6, 6, 4.8284, 4, 4, -0.8284\}$ and

$$M_6 = \begin{pmatrix} 10 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 10 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 7 & 2 & 0 & 2 & 0 & 2 & 0 & 1 \\ 1 & 0 & 2 & 7 & 2 & 0 & 2 & 0 & 1 & 0 \\ 0 & 1 & 0 & 2 & 6 & 3 & 0 & 2 & 0 & 1 \\ 1 & 0 & 2 & 0 & 3 & 6 & 2 & 0 & 1 & 0 \\ 0 & 1 & 0 & 2 & 0 & 2 & 7 & 2 & 0 & 1 \\ 1 & 0 & 2 & 0 & 2 & 0 & 2 & 7 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 10 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 10 \end{pmatrix}$$

with eigenvalues $\{15, 10, 10, 10, 8.7392, 7, 7, 7, 5.7759, -0.5151\}$. It is clear that in these examples, both $\binom{n}{2}$ and $\binom{n-1}{2}$ are eigenvalues of M_n . Also, it turns out that the eigenvalues of M_6 are related to the eigenvalues of M_4 . In fact, if we write

$$M_6 = \begin{pmatrix} 6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 3 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 & 2 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 2 & 2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 3 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6 \end{pmatrix} + \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{pmatrix} + 4I = B_6 + A_6 + 4I,$$

then A_6 has eigenvalues $\{5, 0, 0, 0, 0, 0, 0, 0, -5\}$, $4I$ has only 4 as eigenvalue and B_6 has eigenvalues $\{6, 6, 6, 6, 6, 3, 3, 3, 2, -1\}$. If we compare the latter with the eigenvalues of M_4 , we see that apart from the first four eigenvalues, we now have the same eigenvalues as for M_4 . This comes as no surprise, since if we remove the first and last two rows and columns, we get a matrix that is exactly M_4 . We will show below that we can always write $M_n = B_n + A_n + (n-2)I$, where B_n is a block matrix containing M_{n-2} in a similar fashion as B_6 contains M_4 .

Lemma 1. Let $A_n = (a_{ij})$ be a quadratic matrix of size $2n - 2$ with entries given by

$$a_{ij} = \begin{cases} 1, & \text{if } i + j \text{ is odd;} \\ 0, & \text{otherwise.} \end{cases}$$

Then the eigenvalues are $n - 1$ with multiplicity 1, 0 with multiplicity $2n - 4$ and $-(n - 1)$ with multiplicity 1.

Proof. Since A_n is real and symmetric, it has $2n - 2$ eigenvalues and an orthonormal set of equally many eigenvectors. It is easy to see that $(1, 1, \dots, 1)$ is an eigenvector with eigenvalue $n - 1$ and that $(1, -1, 1, -1, \dots, 1, -1)$ is an eigenvector with eigenvalue $-(n - 1)$. Clearly the rank of A_n is 2, and hence all remaining eigenvalues must be 0.

We are now able to prove Theorem 3.

Proof. (of Theorem 3) It is easy to see that $\binom{n}{2}$ and $\binom{n-1}{2}$ always are eigenvalues of M_n , since the common row sum equals the number of inversions on a genome with n genes, which is $\binom{n}{2}$, and since the second and last rows are equal, except for the term $\binom{n-1}{2}$ on the diagonal. However, we also need to show that there are no other eigenvalues as large as these.

We will use induction. Since our claim is true for M_4 and M_5 in the example above (and anyone can check that it also holds for M_2 and M_3), we can concentrate on the inductive step.

Consider M_n and define $B_n = M_n - A_n - (n - 2)I$, where A_n is given in the previous lemma and I is the identity matrix. Let $C_n = (c_{ij})$ be the matrix B_n with the first and last two rows and columns removed. C_n will have the same size as $M_{n-2} = (m_{ij})$ and we shall see that these matrices are in fact identical. For $i + j$ odd, we have

$$\begin{aligned} c_{ij} &= \min\{\lceil \frac{i+2}{2} \rceil - 1, \lceil \frac{j+2}{2} \rceil - 1, n + 1 - \lceil \frac{i+2}{2} \rceil, n + 1 - \lceil \frac{j+2}{2} \rceil\} - 1 \\ &= \min\{\lceil \frac{i}{2} \rceil - 1, \lceil \frac{j}{2} \rceil - 1, (n - 2) + 1 - \lceil \frac{i}{2} \rceil, (n - 2) + 1 - \lceil \frac{j}{2} \rceil\} = m_{ij}, \end{aligned}$$

and for $i + j$ even, with $i \neq j$, we have

$$c_{ij} = 0 = m_{ij}.$$

Finally, on the main diagonal, we have

$$\begin{aligned} c_{ii} &= \binom{\lceil \frac{i+2}{2} \rceil - 1}{2} + \binom{n + 1 - \lceil \frac{i+2}{2} \rceil}{2} - (n - 2) \\ &= \binom{\lceil \frac{i}{2} \rceil - 1}{1} + \binom{\lceil \frac{i}{2} \rceil - 1}{2} + \binom{n - 1 - \lceil \frac{i}{2} \rceil}{1} + \binom{n - 1 - \lceil \frac{i}{2} \rceil}{2} - (n - 2) \\ &= n - 2 + m_{ii} - (n - 2) = m_{ii}. \end{aligned}$$

Thus, with the only entries on the first and last two rows of B_n being four copies of $\binom{n-1}{2} - (n - 2) = \binom{n-2}{2}$ on the diagonal, the eigenvalues of B_n consists of

these four copies of $\binom{n-2}{2}$ and all the eigenvalues of M_{n-2} . Since the greatest eigenvalue of M_{n-2} is $\binom{n-2}{2}$, this is also the greatest eigenvalue of B_n .

We are now in the position to estimate the eigenvalues of M_n . If we let $\lambda_i(A)$ be the i th greatest eigenvalue of any matrix A , then it is known (see for instance [5], p. 52) that

$$\lambda_{1+i+j}(A+B) \leq \lambda_{1+i}(A) + \lambda_{1+j}(B).$$

We will apply this inequality to find the eigenvalues of $M_n = A_n + B_n + (n-2)I_n$. We know the eigenvalues of A_n from Lemma 1 and B_n by the induction hypothesis. Hence, we get

$$\begin{aligned} \lambda_1(M_n) &\leq \lambda_1(A_n) + \lambda_1(B_n) + \lambda_1((n-2)I) \\ &= (n-1) + \binom{n-2}{2} + (n-2) = \binom{n}{2} \end{aligned}$$

and

$$\begin{aligned} \lambda_2(M_n) &\leq \lambda_2(A_n) + \lambda_1(B_n) + \lambda_1((n-2)I) \\ &= 0 + \binom{n-2}{2} + (n-2) = \binom{n-1}{2}. \end{aligned}$$

Since we know that these are in fact eigenvalues, the inequalities are actually equalities.

So far, we have shown that $\lambda_1 = \binom{n}{2}$ and $\lambda_2 = \binom{n-1}{2}$. In order to see that $\binom{n-1}{2}$ is an eigenvalue with multiplicity at least 3, we need to check that the three linearly independent vectors $\bar{w}_1 = (1, 0, 0, \dots, 0, -1, 0)$, $\bar{w}_2 = (0, 1, 0, \dots, 0, 0, -1)$ and $\bar{w}_3 = (\frac{n-3}{2}, \frac{n-3}{2}, -1, -1, \dots, -1, \frac{n-3}{2}, \frac{n-3}{2})$ all are eigenvectors of M_n with eigenvalue $\binom{n-1}{2}$. It is clear for the two first, and for \bar{w}_3 , multiplying it with M_n give the entries

$$\frac{n-3}{2} \binom{n-1}{2} + 2 \frac{n-3}{2} - \frac{2n-2-4}{2} = \frac{n-3}{2} \binom{n-1}{2}$$

for the first and last two positions, and

$$2 \frac{n-3}{2} - \left(\binom{n}{2} - 2 \right) = \binom{n-1}{1} - \binom{n}{2} = -\binom{n-1}{2}$$

for the other entries. Thus, $M_n \bar{w}_3^T = \binom{n-1}{2} \bar{w}_3^T$.

We now turn to the coefficients of these eigenvalues in the sum giving $b(i)$. The coefficients are given by the first element in the normalised eigenvectors. For $\lambda_1 = \binom{n}{2}$, the first element is $v_1 = \frac{1}{\sqrt{2n-2}}$ and thus gives the coefficient $v_1^2 = \frac{1}{2n-2}$. For the eigenvalue $\lambda_2 = \lambda_3 = \lambda_4 = \binom{n-1}{2}$, we get

$$v_2^2 + v_3^2 + v_4^2 = \frac{1}{2} + 0 + \frac{\left(\frac{n-3}{2}\right)^2}{4 \left(\frac{n-3}{2}\right)^2 + 2n-6} = \frac{1}{2} + \frac{(n-3)}{4((n-3)+2)} = \frac{3}{4} - \frac{1}{2n-2}.$$

Finally, turning to the smallest eigenvalue, we can use

$$\lambda_{m-i-j}(A+B) \geq \lambda_{m-i}(A) + \lambda_{m-j}(B),$$

where $m \times m$ is the common size of A and B , to show that the smallest eigenvalue is greater than or equal to $-\lfloor \frac{n}{2} \rfloor$. This holds for M_2 and M_3 , and by the same procedure as above, we find that

$$\begin{aligned} \lambda_{2n-2}(A+B) &\geq \lambda_{2n-2}(A) + \lambda_{2n-2}(B) + \lambda_{2n-2}((n-2)I) \\ &\geq -(n-1) - \lfloor \frac{n-2}{2} \rfloor + (n-2) = -\lfloor \frac{n}{2} \rfloor. \end{aligned}$$

This ends the proof of Theorem 3.

4 Analysing and improving the formula

We will now leave the analytical trail and look at how well these approximations behave in practice. Based on abundant observations (every $n \leq 100$), we believe that the largest eigenvalues are distributed as follows.

Conjecture 1. Let M_n be the scaled transition matrix studied in the previous section and $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{2n-2}$ its eigenvalues. Then $\lambda_6 = \lambda_7 = \lambda_8 = \binom{n-2}{2} - 1$ for $n \geq 6$.

Since we know the four greatest eigenvalues, this conjecture implies that there is only one unknown eigenvalue λ_5 larger than $\binom{n-2}{2} - 1$. Knowledge of this eigenvalue and its coefficient v_5^2 would give a better approximation than the one found above. We now take a closer look at these parameters.

First we look at the unknown coefficients. We know that the sum of all coefficients is 1 and that, according to Theorem 3, the coefficients of the eigenvalues $\binom{n}{2}$ and $\binom{n-1}{2}$ sum to $\frac{3}{4}$. For the remaining $\frac{1}{4}$, numerical calculations for $n \leq 100$ indicate that almost everything has been given to λ_5 . We see in Figure 1, where the coefficients has been plotted for $n \leq 40$ that this coefficient fast approaches $\frac{1}{4}$. The sum of the remaining coefficients has been plotted in Figure 2. We see that for $n = 40$, their sum is less than 0.001. We can neglect this without worries.

Supported by this, we now propose an improved approximation of the expected number of breakpoints, given i inversion. By setting $v_5^2 = \frac{1}{4}$ and writing $\lambda_5 = \binom{n-1}{2} - \varepsilon(n)$, we get

$$\begin{aligned} b_{\text{appr2}}(i) &= n \left[1 - \frac{1}{2n-2} - \left(\frac{3}{4} - \frac{1}{2n-2} \right) \left(1 - \frac{2}{n} \right)^i - \frac{1}{4} \left(\frac{\binom{n-1}{2} - \varepsilon(n)}{\binom{n}{2}} \right)^i \right] \\ &\approx n \left(1 - \frac{1}{2n-2} \right) \left(1 - \left(1 - \frac{2}{n} \right)^i \right) + \frac{i\varepsilon(n)}{2n-2} \left(1 - \frac{2}{n} \right)^{i-1}. \end{aligned}$$

The final approximation was obtained by including only the first two terms in the binomial expansion of the last term. From this approximation of $b_{\text{appr2}}(i)$, we find that the error of $b_{\text{appr}}(i)$ is approximately $\frac{i\varepsilon(n)}{2n-2} \left(1 - \frac{2}{n} \right)^{i-1}$.

Figure 1. How the coefficients evolve with increasing n . The stars correspond to the eigenvalue $\binom{n}{2}$, the squares to $\binom{n-1}{2}$, the circles to the eigenvalue just below $\binom{n-1}{2}$ and the diamonds to the sum of all other coefficients.

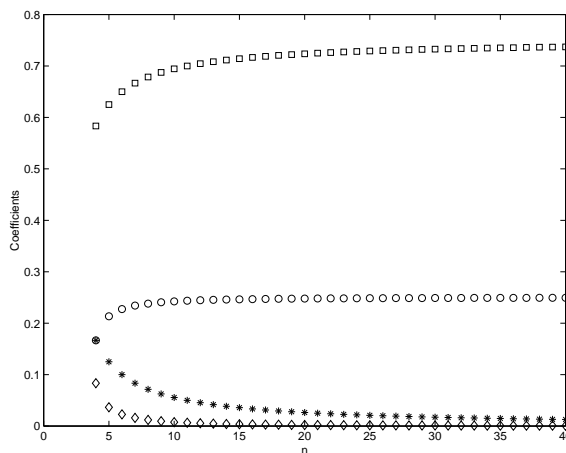
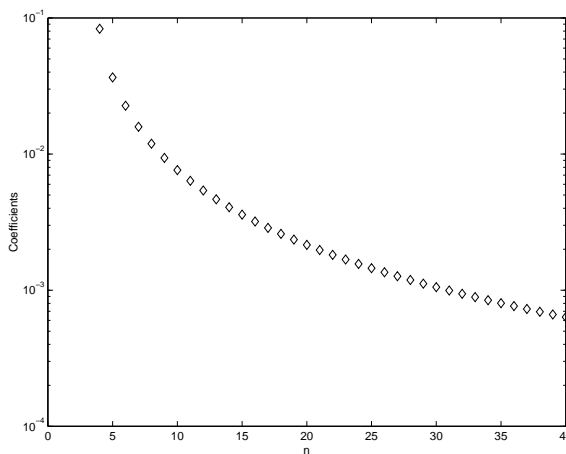


Figure 2. The coefficients of the eigenvalues we neglect. These tend to zero quite rapidly as n increases.



The usefulness of this improved approximation, which equals our first if we put $\varepsilon(n) = 0$, depends on our ability to calculate $\varepsilon(n)$. We have plotted $\varepsilon(n)$ as a function of n in Figure 3. We find (left graph) that for $40 \leq n \leq 100$, we can approximate $\varepsilon(n)$ with, for instance, $\varepsilon(n)_{\text{appr}} = 1.7 + 0.0016n$, and for larger n , $\varepsilon(n) = 2$ seems to be an as good approximation as one could hope for.

Figure 3. Two graphs of the function $\varepsilon(n)$. We have plotted this function for different ranges of n in order to make its behaviour clear for small n as well as for large n . For small n , it increases from less than $\frac{3}{2}$ to about 2, but for large n it stays fairly constant, just below 2.

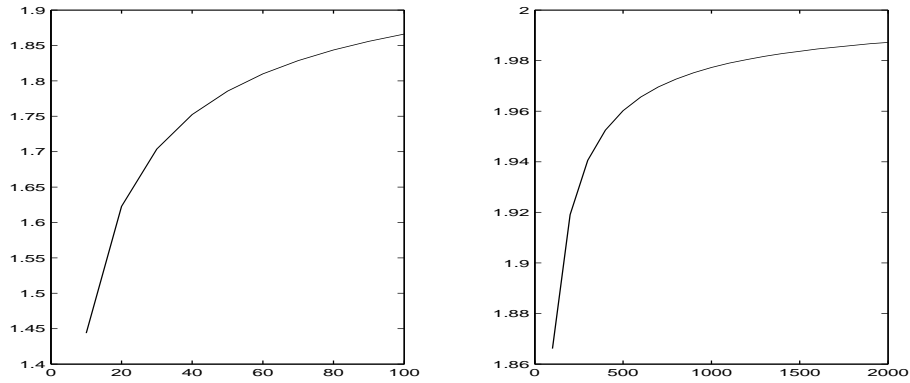
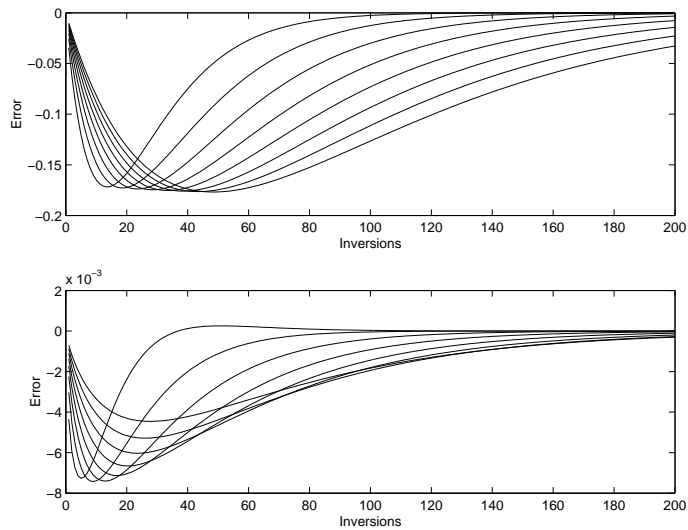
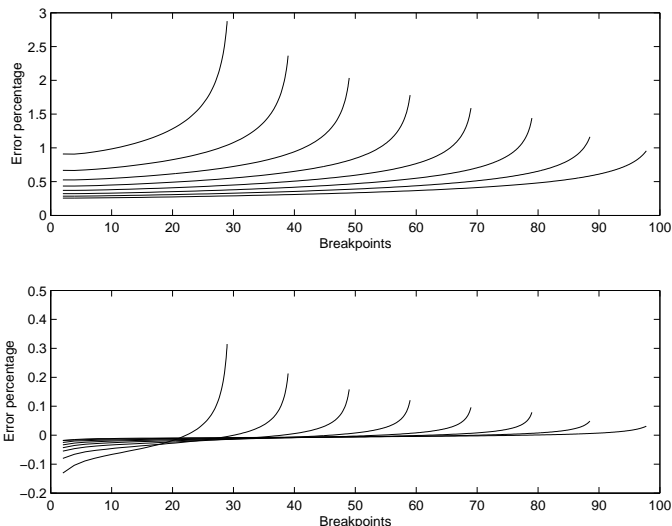


Figure 4. The error of $b_{\text{appr}}(i)$ (top) and $b_{\text{appr}2}(i)$ for these n : 30, 40, 50, 60, 70, 80, 90 and 100. The latter is one or two orders of magnitude lower.



A comparison between the quality of our approximations can be found in Figures 4 and 5. The true values have, of course, been taken from the Markov chain. We have plotted the error depending on i for the following values of n : 30, 40, 50, 60, 70, 80, 90 and 100. We see that for the approximations of $b(i)$, the

Figure 5. The error percentage of $i_{\text{appr}}(b)$ (top) and $i_{\text{appr}2}(b)$. Again, we have used the following values of n : 30, 40, 50, 60, 70, 80, 90 and 100. Note that there is no analytical expression for the second approximation, but it can be computed numerically from its inverse.



first approximation $b_{\text{appr}}(i)$ has an error that stays below 0.2 breakpoints, which is fairly good. With the second approximation, $b_{\text{appr}2}(i)$, the error is well below 0.01 breakpoints.

For the approximations of $i(b)$, we see that when b approaches its upper limit, the error increases. This is bound to happen, since the slope of $b(i)$ decreases towards zero for large i . Still, the error percentage is not too high, even for the analytical expression $i_{\text{appr}}(b)$, and using the numerical inverse of $b_{\text{appr}2}(i)$, the error vanishes in practice for most b .

5 Conclusions

Caprara and Lancia considered the problem of calculating the expected number of inversions leading to b breakpoints for *unsigned* permutations. By taking the sign of the genes into account, we use more information. Thus, using signed permutations, we should hope for more reliable results when applied to biological problems. In this case, contrary to calculating the minimal inversion distance, we gain this reliability at the cost of complexity. Where solving the unsigned case amounts to calculating the expected number of breakpoints in a random permutation ($n - 2$ by linearity of expectation values) and the probability that a random inversion creates or destroys the adjacency of two genes, the signed case requires the calculation of many permutation-specific probabilities and the

calculation of the eigenvalues of a matrix containing these probabilities. In other words: for signed permutations we calculate the eigenvalues of a $(2n-2) \times (2n-2)$ -matrix, for unsigned permutations the corresponding matrix has size 1×1 .

It is a truly remarkable property of the transition matrices M_n that most of their eigenvalues can be calculated without much effort. This insight provides us with the means to compute the expected number of inversions giving rise to b breakpoints. The error in this calculation is most certainly negligible, compared to the standard deviation. For the future, calculating this standard deviation seems to be an important problem. Also, it would be interesting to see if the information not used for the problem considered in this paper, for example the cycle structure of the permutations, can be used to obtain an even more realistic distance measure.

6 Acknowledgement

I wish to thank my advisor Kimmo Eriksson for valuable comments during the preparation of this paper.

Niklas Eriksen was supported by the Swedish Research Council and the Swedish Foundation for Strategic Research.

A Further information on the spectrum of M_n

There is some information known about the eigenvalues (and their coefficients) of M_n , which do not affect the analysis above. We will present this information here.

Theorem 4. *In addition to the eigenvalues found in Theorem 3, M_n has eigenvalues $\binom{n-2}{2} + \binom{2}{2}$, $\binom{n-3}{2} + \binom{3}{2}$, \dots , $\binom{n-\lceil \frac{n-2}{2} \rceil}{2} + \binom{\lceil \frac{n-2}{2} \rceil}{2}$ with multiplicity 3 (if n is odd, $\binom{n-\lceil \frac{n-2}{2} \rceil}{2} + \binom{\lceil \frac{n-2}{2} \rceil}{2}$ has multiplicity 2). The coefficients of these eigenvalues are 0.*

Proof. We saw in the proof for Theorem 3 that $\bar{w}_1 = (1, 0, 0, \dots, 0, -1, 0)$, $\bar{w}_2 = (0, 1, 0, \dots, 0, 0, -1)$ and $\bar{w}_3 = (\frac{n-3}{2}, \frac{n-3}{2}, -1, -1, \dots, -1, \frac{n-3}{2}, \frac{n-3}{2})$ are eigenvectors of M_n with eigenvalue $\binom{n-1}{2}$. In fact, if we add two zeros at the front and at the back of the eigenvector \bar{w}_i of M_{n-2} , we get a corresponding eigenvector for the eigenvalue $\binom{n-2}{2} + \binom{2}{2}$ in M_n . Using this inductively, it is easy to show that the eigenvalues $\binom{n-k}{2} + \binom{k}{2}$ for $2 \leq k \leq \lceil \frac{n-2}{2} \rceil$ have three-dimensional eigenspaces. The exception is, as mentioned, $\binom{n-\lceil \frac{n-2}{2} \rceil}{2} + \binom{\lceil \frac{n-2}{2} \rceil}{2}$ for odd n , which only has a two-dimensional eigenspace, since the eigenvector that would correspond to \bar{w}_3 then equals the sum of the other two eigenvectors.

Turning to their coefficients, we just found that all eigenvectors of $\binom{n-k}{2} + \binom{k}{2}$, $2 \leq k \leq \lceil \frac{n-2}{2} \rceil$ has a zero in the first position. Hence, the coefficients are also zero.

References

1. Bader, D. A., Moret, B. M. E., Yan, M.: A Linear-Time Algorithm for Computing Inversion Distance Between Signed Permutations with an Experimental Study. *Journal of Computational Biology*, **8**, 5 (2001), 483–491
2. Blanchette, M., Kunisawa, T., Sankoff, D.: Parametric genome rearrangement. *Gene* **172** (1996), GC 11–17
3. Caprara, A.: Sorting permutations by reversals and Eulerian cycle decompositions. *SIAM Journal of Discrete Mathematics* **12** (1999), 91–110
4. Caprara, A., Lancia, G.: Experimental and statistical analysis of sorting by reversals. Sankoff and Nadeau (eds.), *Comparative Genomics* (2000), 171–183
5. Cvetković, D. M., Doob, M., Sachs, H.: *Spectra of Graphs*. Johann Ambrosius Barth Verlag, Heidelberg, 1995
6. Eriksen, N.: $(1 + \epsilon)$ -Approximation of Sorting by Reversals and Transpositions. *Algorithms in Bioinformatics, Proceedings of WABI 2001, LNCS 2149*, 227–237
7. Eriksen, N.: Expected number of inversions after a sequence of random adjacent transpositions — an exact expression. Preprint
8. Eriksen, N., Dalevi, D., Andersson, S. G. E., Eriksson, K.: Gene order rearrangements with Derange: weights and reliability. Preprint
9. Eriksson, H., Eriksson, K., Sjöstrand, J.: Expected inversion number after k adjacent transpositions *Proceedings of Formal Power Series and Algebraic Combinatorics 2000*, Springer Verlag, 677–685
10. Hannenhalli, S., Pevzner, P.: Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations with reversals). *Proceedings of the 27th Annual ACM Symposium on the Theory of Computing* (1995), 178–189
11. Pevzner, P.: *Computational Molecular Biology: An Algorithmic Approach*. The MIT Press, Cambridge, MA 2000
12. Sankoff, D., Blanchette, M.: Probability models for genome rearrangements and linear invariants for phylogenetic inference. *Proceedings of RECOMB 1999*, 302–309
13. Wang, L.-S.: Exact-IEBP: A New Technique for Estimating Evolutionary Distances between Whole Genomes. *Algorithms in Bioinformatics, Proceedings of WABI 2001, LNCS 2149*, 175–188