

Expected number of inversions after a sequence of random adjacent transpositions — an exact expression

Niklas Eriksen

*Department of Mathematics
Royal Institute of Technology
S-100 44 Stockholm, Sweden*

Abstract

A formula for calculating the expected number of inversions after t random adjacent transpositions has been presented by Eriksson et al. We have improved their result by determining a formula for the unknown integer sequence d_r that was used in their formula and also made the formula valid for large t .

Key words: Inversions, expectation, permutations, adjacent transpositions

1 Introduction

In a recent article [3], the Eriksson-Sjöstrand family calculated the expected number of inversions in a permutation, given the number of adjacent transpositions applied to it. Problems of this type have applications in computational biology, where the genome may be regarded as a permutation of genes. Consider two such genomes π and ρ , in which we have named the genes such that $\rho = \text{id}$. The evolutionary distance between π and ρ is assumed to be proportional to the number of evolutionary operations that have changed the gene order since the two genomes diverged. To calculate this number of operations, we can either calculate the least number of operations needed to transform π into $\rho = \text{id}$ (this corresponds to sorting π), which gives a lower bound of the true number of operations, or we can calculate the expected number of operations, given some measure on the difference between the two genomes.

¹ Supported by a grant from the Swedish Research Council

One such common measure is the number of breakpoints, that is the number of adjacent pairs in π that are not consecutive.

In the paper by Eriksson et al., they calculated the inverse of the second alternative: they found the expected measure of difference given a certain number of operations. With this information, we may determine this measure of difference between two given genomes and then extract the number of operations that is expected to produce this difference. The same approach has been taken by Wang and Warnow [7], for breakpoints and the reversals and block transpositions usually considered in computational biology.

As mentioned, Eriksson et al. considered inversions and adjacent transpositions. Their result is the following

Theorem 1 (Eriksson et al. [3]) *The expected number of inversions in a permutation in S_{n+1} after t random adjacent transpositions is, for $n \geq t$,*

$$\mathbb{E}_{\text{inv}}(n, t) = \sum_{r=0}^t \frac{(-1)^r}{n^r} \left[\binom{t}{r+1} 2^r C_r + 4d_r \binom{t}{r} \right],$$

where d_r is an integer sequence that begins with 0, 0, 0, 1, 9, 69, 510 and $C_r = \frac{1}{r+1} \binom{2r}{r}$ are the Catalan numbers.

There are a couple of things that can be improved in the result of Eriksson et al. First, their formula includes some numbers d_r that they have no expression formula for. Second, the formula is only valid for $n \geq t$.

In this paper, we will present an improved formula, where both these flaws have been eliminated. The theorem is given directly below, and the proof will appear in the following sections.

Theorem 2 *The expected number of inversions in a permutation in S_{n+1} after t random adjacent transpositions is*

$$\mathbb{E}_{\text{inv}}(n, t) = \sum_{r=1}^t \frac{1}{n^r} \binom{t}{r} \sum_{s=1}^r \binom{r-1}{s-1} (-1)^{r-s} 4^{r-s} g_{s,n}.$$

The integer sequence $g_{s,n}$ is given by

$$g_{s,n} = \sum_{l=0}^n \sum_{k \in \mathbb{N}} (-1)^k (n-2l) \binom{2\lceil \frac{s}{2} \rceil - 1}{\lceil \frac{s}{2} \rceil + l + k(n+1)} \sum_{j \in \mathbb{Z}} (-1)^j \binom{2\lfloor \frac{s}{2} \rfloor}{\lfloor \frac{s}{2} \rfloor + j(n+1)}.$$

Corollary 3 *For $n \geq t$, we get*

$$\begin{aligned} \mathbb{E}_{\text{inv}}(n, t) &= \sum_{r=0}^t \frac{(-1)^r}{n^r} \left[2^r C_r \binom{t}{r+1} \right. \\ &\quad \left. + 2 \binom{t}{r} \sum_{s=3}^r \binom{r-1}{s-1} (-1)^{s-1} 4^{r-s} \binom{2 \lfloor \frac{s}{2} \rfloor}{\lfloor \frac{s}{2} \rfloor} \sum_{l=0}^{\lfloor \frac{s-1}{2} \rfloor} l \binom{2 \lceil \frac{s}{2} \rceil - 1}{\lceil \frac{s}{2} \rceil + l} \right], \end{aligned}$$

where C_r are the Catalan numbers. Thus, the sequence d_r in Theorem 1 is given by

$$d_r = \frac{1}{2} \sum_{s=3}^r \binom{r-1}{s-1} (-1)^{s-1} 4^{r-s} \binom{2 \lfloor \frac{s}{2} \rfloor}{\lfloor \frac{s}{2} \rfloor} \sum_{l=0}^{\lfloor \frac{s-1}{2} \rfloor} l \binom{2 \lceil \frac{s}{2} \rceil - 1}{\lceil \frac{s}{2} \rceil + l}.$$

Having proved this theorem, we provide in short an alternative formula, which may work better in some situations, and take a look at another approach to this problem. Even though this second approach did not prove too successful here, it has led to remarkable results on similar problems taken directly from computational biology (see Eriksen [1] and Eriksen and Hultman [2]). In closing, we review some results for other Coxeter groups.

2 The heat flow model

To prove Theorem 2, we have used the heat flow model proposed by Eriksson et al. Before we state this model, we need a few definitions.

We look at the symmetric group S_{n+1} . The transposition that changes the elements π_i and π_{i+1} is denoted s_i . We let

$$\mathcal{P}_{nt} = \{s_{i_1} s_{i_2} \dots s_{i_t} : 1 \leq i_1, i_2, \dots, i_t \leq n\},$$

that is the set of sequences of exactly t adjacent transpositions.

Fix n . We define the “matrix” $(p_{ij})(t)$, where

$$p_{ij}(t) = \text{Prob}(\pi_i < \pi_j)$$

for a permutation $\pi \in \mathcal{P}_{nt}$, where the adjacent transpositions $s_k, 1 \leq k \leq t$ have been chosen randomly from the uniform distribution. Observe that the main diagonal has not been assigned any values. From this definition, it follows that

$$\mathbb{E}_{\text{inv}}(n, t) = \sum_{i>j} p_{ij}(t).$$

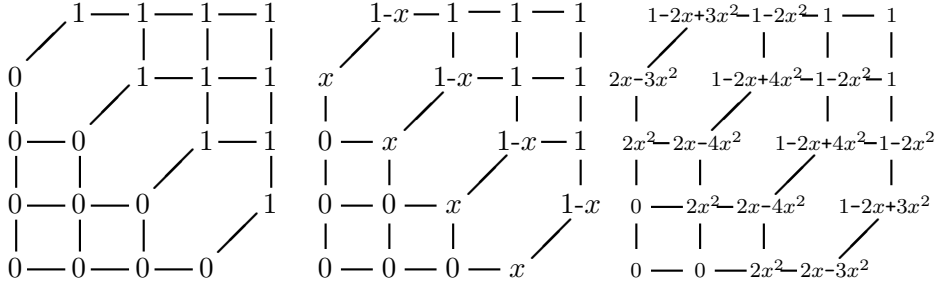


Fig. 1. The matrices $(p_{ij})(0)$, $(p_{ij})(1)$ and $(p_{ij})(2)$ for $n = 4$.

We now define a discrete heat flow process as follows. On a (finite or infinite) graph, every vertex has at time zero some heat associated to itself. In each time step, all vertices sends a fraction x of its heat to each of its neighbours. At the same time, it will receive the same fraction of each neighbours' heat. The following proposition is proven in [3].

Proposition 4 (Eriksson et al. [3]) *The sequence of (p_{ij}) -matrices for $t = 0, 1, 2, \dots$ describes a discrete heat flow process with conductivity $x = 1/n$ on the grid graph depicted in Figure 1 (left). The heat equation becomes*

$$p_{ij}(t) = p_{ij}(t-1) + \frac{1}{n} \sum (p_{\text{neighbour}}(t-1) - p_{ij}(t-1)),$$

where the sum is taken over all neighbours of vertex (i, j) .

In the same paper, it is also shown that we can replace the grid in Figure 1 by the grid in Figure 2. The sequence of (p_{ij}) -matrices for $t = 0, 1, 2, \dots$ describes a heat flow process on this grid graph. In this process, the heat on the diagonal will never change. Furthermore, we are only interested in the part below the diagonal, since this is where we record the probabilities of inversions. We thus get a model with two insulated boundaries (below and to the left) and one hot boundary (the diagonal). This is depicted in Figure 3.

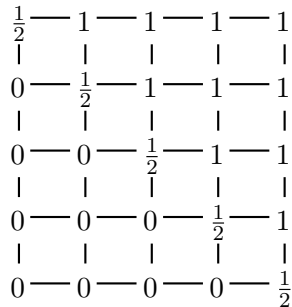


Fig. 2. Grid graph with initial values

By reflection, we can extend this graph to a graph with no insulated boundaries (as in Figure 3). We will now calculate the amount of heat that flows from one of the borders (say the northeast one) onto this grid. This will equal the

amount of heat in the upper right quarter of the grid, which is what we are trying to calculate. Remember that this heat equals $\mathbb{E}_{\text{inv}}(n, t)$.

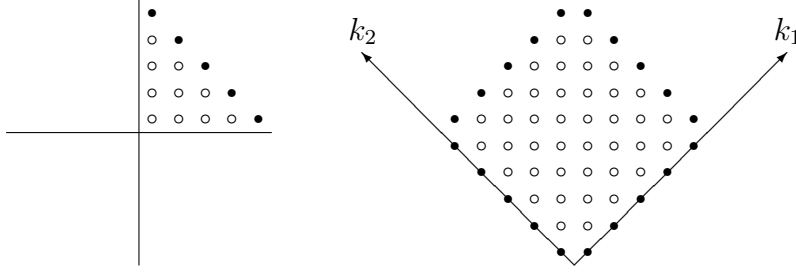


Fig. 3. By reflection, the graph with one hot and two insulated boundaries is extended to a diamond shaped graph with no insulated boundaries. The new set of coordinates (k_1, k_2) is introduced.

We can view the heat equation as describing how small packets of heat are sent back and forth on the grid. The amazing thing about the heat flow model is that we can calculate the contribution from every heat packet separately, and then add them all together. The vertices at the hot boundary send out heat packets with value $\frac{1}{2n}$ to their neighbours at each time step. These packets are then sent back and forth between the inner vertices. From our heat equation, there are three possible travel steps for a packet [3]:

- It stays on the vertex unchanged.
- It travels to a neighbouring vertex, getting multiplied by $\frac{1}{n}$.
- It travels halfway to a neighbouring vertex, gets multiplied by $\frac{-1}{n}$ and returns to the vertex it came from.

Now, in order to calculate the total heat at a vertex, we sum, over all travel routes from the boundary, the heat packets that have travelled these routes. We define new coordinates k_1 and k_2 on this grid as in Figure 3 (the origin is at the bottom of the graph). If a packet has travelled from the northeast border to (i, j) in t days, we know the following.

- Out of the t days, there are r travel days. They can be chosen in $\binom{t}{r}$ ways.
- From these travel days, we must choose s true travel days, in which the packet changes vertex. This can be done in $\binom{r-1}{s-1}$ ways, since the packet must change vertex the first travel day.
- If the packet does not change vertex on a travel day, it has four direction to choose from. This gives the factor 4^{r-s} .
- The heat that reaches the destination is $\frac{(-1)^{r-s}}{2} \frac{1}{n^r}$.
- For each of the true travel days, both coordinates k_1 and k_2 change. Only paths that do not touch the boundary are valid. We will enumerate these paths, which we call **two-sided Dyck paths**. Their proper definition is given below.

It should be noted that Eriksson et al. used a similar approach, but only on a

semi-infinite model, which gave a lower bound for $\mathbb{E}_{\text{inv}}(n, t)$.

We are now able to prove the first part of Theorem 2. We will sum over all vertices in the diamond graph, and for each vertex over all paths from the northeast border. These paths will display two-sided Dyck paths from $(0, a)$ to (s, b) (with a odd), where the y -coordinate corresponds to k_2 , and two-sided Dyck paths from $(0, 1)$ to $(s-1, b)$ where the y -coordinate corresponds to $2n+2-k_1$. Let $b_{s,n}$ and $c_{s-1,n}$ be the number of such two-sided Dyck paths, respectively. This yields, with $x = 1/n$,

$$\mathbb{E}_{\text{inv}}(n, t) = \frac{1}{2} \sum_{r=1}^t \frac{1}{n^r} \binom{t}{r} \sum_{s=1}^r \binom{r-1}{s-1} (-1)^{r-s} 4^{r-s} b_{s,n} c_{s-1,n}.$$

Thus, the first part of the theorem is proven (we have, of course, $g_{s,n} = b_{s,n} c_{s-1,n} / 2$).

3 Two-sided Dyck paths

We start by formally defining two-sided Dyck paths and then proceed to enumerate them.

Definition 5 *A two-sided Dyck path of height n is a path on the integer grid from $(0, a)$ to (s, b) , where $a, b \in \{1, 2, \dots, n-1\}$ and $s \geq 0$, allowing only the steps $(1, 1)$ and $(1, -1)$, such that $0 < y < n$ at all positions along the way.*

We see that the number of two-sided Dyck paths from $(0, 1)$ to $(2k, 1)$ is C_k (ordinary Catalan numbers) if the height is larger than $k+1$ (we can never hit the ceiling then).

Proposition 6 *The number of two-sided Dyck paths of height n from $(0, a)$ to (s, b) is given by*

$$\sum_{k \in \mathbb{Z}} \left(\binom{s}{\frac{s+b-a+2kn}{2}} - \binom{s}{\frac{s-b-a+2kn}{2}} \right)$$

or 0, if $s+b-a$ is an odd number.

This proposition can be proven using the standard reflection argument, in combination with the principle of inclusion-exclusion. It can also be found in Mohanty [4].

With this proposition, we are able to determine $b_{s,n}$ and $c_{s,n}$. We start with the latter.

Lemma 7 *The number of two-sided Dyck paths of height $2n + 2$ from $(0, 1)$ to (s, b) for all $0 < b < 2n + 2$ is given by*

$$c_{s,n} = \sum_{k \in \mathbb{Z}} (-1)^k \binom{s}{\frac{s+2k(n+1)}{2}},$$

if s is an even number, and

$$c_{s,n} = \frac{1}{2} c_{s+1,n},$$

if s is an odd number.

PROOF. We get, for even s ,

$$\begin{aligned} c_{s,n} &= \sum_{m=0}^n \sum_{k \in \mathbb{Z}} \left(\binom{s}{\frac{s}{2} + m + 2k(n+1)} - \binom{s}{\frac{s}{2} - m - 1 + 2k(n+1)} \right) \\ &= \sum_{k \in \mathbb{Z}} (-1)^k \binom{s}{\frac{s+2k(n+1)}{2}}. \end{aligned}$$

Most terms cancel by symmetry of the binomial coefficients. For odd s , we see that for each two-sided Dyck path to $x = s$ we get two such paths to $x = s + 1$.

□

Lemma 8 *The number of two-sided Dyck paths of height $2n + 2$ from $(0, a)$ to (s, b) for all $0 < a, b < 2n + 2$, a odd, is given by*

$$\begin{aligned} b_{s,n} &= 2 \sum_{l=0}^n \sum_{k \in \mathbb{N}} (-1)^k (n - 2l) \binom{s}{\frac{s+1}{2} + l + k(n+1)} \\ &= n2^s - 2 \sum_{l=0}^n 2l \sum_{k \in \mathbb{N}} (-1)^k \binom{s}{\frac{s+1}{2} + l + k(n+1)} \\ &= n2^s - 4 \beta_{s,n}, \end{aligned}$$

if s is an odd number, and

$$b_{s,n} = 2b_{s-1,n} = n2^s - 8 \beta_{s-1,n},$$

if s is an even number.

PROOF. Assume s is an odd number. For all odd a but $n + 1$, we get a term $\binom{s}{\frac{s+1}{2}}$. Hence, there are n such terms. Similarly, we get $n - 2$ ($n - 1$ positive and 1 negative) $\binom{s}{\frac{s+1}{2}+1}$ and $(n - 4)\binom{s}{\frac{s+1}{2}+2}$, etc. This continues similarly to $(n - 2n)\binom{s}{\frac{s+1}{2}+n}$. We then turn to get $(n - 2n)\binom{s}{\frac{s+1}{2}+n+1}$, $(n - 2(n - 1))\binom{s}{\frac{s+1}{2}+n+2}$, etc. Continuing in this fashion gives the first equality in the lemma. The leading 2 comes from symmetry, adding all paths going downwards.

For the second equality, we use that the row sums in Pascal's triangle are 2^n .

For even s , there are b_{s-1} paths to $x = s - 1$. For each of these paths, there are two valid options (up or down) for the last step.

□

We have now proved the second part of our main theorem. What remains to prove the corollary is the simplifications for $t \leq n$. Assuming this inequality, we can simplify our formula using the following lemma.

Lemma 9 *We have that*

$$\sum_{s=0}^r (-1)^s 2^{r-s} \binom{r}{s} \binom{s}{\lceil \frac{s}{2} \rceil} = C_r,$$

where C_r is the r :th Catalan number.

PROOF. Consider vectors v of length $2r + 1$, containing $r + 1$ zeroes and r ones. The number $T(r, s)$ of such vectors that contain exactly $2s + 1$ palindrome positions, i.e. positions i such that $v_i = v_{2r+2-i}$, can be found as follows. We concentrate on the first r positions. First choose which of these should be palindrome positions. Fill in the others arbitrarily. We then fill in the palindrome positions using $\lceil \frac{s}{2} \rceil$ zeroes and $\lfloor \frac{s}{2} \rfloor$ ones. All other positions can then be filled in so that the chosen palindrome positions really are palindrome positions and the other positions are not. It is easy to check that we get a valid palindrome vector, and that we do not miss any valid vectors. From this analysis, we find that

$$T(r, s) = 2^{r-s} \binom{r}{s} \binom{s}{\lceil \frac{s}{2} \rceil}.$$

It turns out that the element at position $r + 1$ is 0 if s is even and 1 otherwise. If we remove this position, we get vectors of length $2r$ with r zeroes and r ones, for even s , and $r + 1$ zeroes and $r - 1$ ones for odd s . The number of such vectors are $\binom{2r}{r}$ and $\binom{2r}{r+1}$, respectively. We thus get

$$\sum_{s=0}^r (-1)^s T(r, s) = \binom{2r}{r} - \binom{2r}{r+1} = C_r.$$

□

Now, for $n \geq t \geq r \geq s$, we get

$$g_{s,n} = b_{s,n} c_{s-1,n} = n 2^{s-1} \binom{s-1}{\lceil \frac{s-1}{2} \rceil} - 2 \binom{2 \lfloor \frac{s}{2} \rfloor}{\lfloor \frac{s}{2} \rfloor} \sum_{l=0}^{\lfloor \frac{s-1}{2} \rfloor} l \binom{2 \lceil \frac{s}{2} \rceil - 1}{\lceil \frac{s}{2} \rceil + l}.$$

Lemma 9 and Theorem 2 now prove Corollary 3.

4 An alternative formula

There is another way of writing $\mathbb{E}_{\text{inv}}(n, t)$ that can be obtained using a similar model. We start with the same heat flow model, but instead of the three possible travel steps previously described, we merge two of them, giving these options:

- The packet changes vertex. It will then get multiplied with $x = \frac{1}{n}$.
- The packet does not change vertex. If it has not changed vertex before, nothing happens. Otherwise, it gets multiplied with $(1 - 4x)$.

We no longer need to keep track of the true travel days, since there will be no other travel days. We must, however, keep track of the first day (q) of travel. With this in mind, we easily find this expression valid:

$$\mathbb{E}_{\text{inv}}(n, t) = \frac{1}{2} \sum_{q=1}^t \sum_{r=0}^{t-q} \binom{t-q}{r} \left(1 - \frac{4}{n}\right)^{t-q-r} \frac{1}{n^{r+1}} b_{r+1,n} c_{r,n}.$$

This gives the following theorem.

Theorem 10 *The expected number of inversions in a permutation in S_{n+1} after t random permutations is given by*

$$\mathbb{E}_{\text{inv}}(n, t) = \sum_{u=0}^{t-1} \left(\frac{n-4}{n}\right)^u \sum_{r=0}^u \binom{u}{r} \frac{1}{(n-4)^r} \left(2^r + \frac{2\beta_{r+1,n}}{n}\right) c_{r,n}.$$

PROOF. Trivial calculations give

$$\begin{aligned}
\mathbb{E}_{\text{inv}}(n, t) &= \frac{1}{2} \sum_{q=1}^t \sum_{r=0}^{t-q} \binom{t-q}{r} \left(1 - \frac{4}{n}\right)^{t-q-r} \frac{1}{n^{r+1}} b_{r+1, n} c_{r, n} \\
&= \frac{1}{2} \sum_{u=0}^{t-1} \sum_{r=0}^u \binom{u}{r} \left(1 - \frac{4}{n}\right)^{u-r} \frac{1}{n^{r+1}} b_{r+1, n} c_{r, n} \\
&= \sum_{u=0}^{t-1} \left(\frac{n-4}{n}\right)^u \sum_{r=0}^u \binom{u}{r} \frac{1}{(n-4)^r} \left(2^r + \frac{2\beta_{r+1, n}}{n}\right) c_{r, n}.
\end{aligned}$$

□

This expression seems particularly useful for fixed n (try for instance $n = 4$), since t only appears as the number of terms in the sum. This would be the standard case in applications. While our formulae are somewhat complicated, this indicates that they will be useful in practise.

Furthermore, it is easy to find out how much $\mathbb{E}_{\text{inv}}(n, t)$ increases when we increase t one step. This is given by

$$\begin{aligned}
\Delta_t \mathbb{E}_{\text{inv}}(n, t) &= \mathbb{E}_{\text{inv}}(n, t+1) - \mathbb{E}_{\text{inv}}(n, t) \\
&= \sum_{r=0}^t \binom{t}{r} \left(1 - \frac{4}{n}\right)^{t-r} \frac{1}{n^{r+1}} b_{r+1, n} c_{r, n}.
\end{aligned}$$

In a glance, we see that $\Delta_t \mathbb{E}_{\text{inv}}(n, t)$ is always positive for $n \geq 4$. This means that $\mathbb{E}_{\text{inv}}(n, t)$ is monotonically increasing for almost all n . It should be pointed out that although this may seem trivial, for $n = 1$ (permutations of length 2), $\mathbb{E}_{\text{inv}}(1, t)$ takes the values $0, 1, 0, 1, 0, 1, \dots$. While this sequence may be regarded as quite monotone, it is not monotonically increasing.

To be able to apply this in a biological context, where we wish to estimate the number of adjacent transpositions given the inversion number of a permutation, we need this monotonicity property. The reason is that when we have found an expectation $\mathbb{E}_{\text{inv}}(n, t)$ which is close to our number of inversions, we must be sure that we will not find a better expectation for a much larger t . If the sequence is monotone, this can never happen.

5 A Markov chain approach

We will now briefly discuss another method of obtaining an approximative formula for $\mathbb{E}_{\text{inv}}(n, t)$. It is built on the theory of Markov chains and depends

on our ability of calculating eigenvalues of the transition matrices.

We will use the Cayley graph of S_{n+1} with the adjacent transpositions as generators: each permutation in S_{n+1} corresponds to a vertex and there is an edge between two vertices if and only if the corresponding permutations differ by an adjacent transposition. We then form the adjacency matrix $A_n = (a_{ij})$ of this graph. The vertices will be sorted in increasing lexicographic order.

Since the Cayley graph of S_{n+1} is regular, $M_n = \frac{1}{n}A_n$ constitutes the transition matrix of the Markov chain of walks on the Cayley graph of S_n , giving equal probability to all edges. Thus, the entry $m_{ij}^{(t)}$ in M_n^t gives the probability that a walk of length t starting at permutation i ends at permutation j .

The expected number of inversions after t random transpositions can then be written as

$$\mathbb{E}_{\text{inv}}(n, t) = \sum_{\pi \in S_{n+1}} m_{\text{id}, \pi} w_{\pi},$$

where w_{π} is the number of inversions of permutation π . The w_{π} are easy to describe as follows:

Lemma 11 *Arrange all permutations in S_{n+1} in lexicographic order. Let w_{i-1} be the number of inversions in the i th permutation of this list. Then,*

$$w_i = \sum_{k \geq 1} \left\lfloor \frac{i \bmod (k+1)!}{k!} \right\rfloor.$$

In this order, permutation 1 is the identity. With $\bar{w}_n = (w_0, \dots, w_{(n+1)!-1})$ and $\bar{e}_1 = (1, 0, \dots, 0)$ we have

$$\mathbb{E}_{\text{inv}}(n, t) = \bar{e}_1 M_n^t \bar{w}_n.$$

Since M_n is real and symmetric, we can diagonalise it: $M_n = V_n D_n V_n^T$, where D_n is a diagonal matrix with the eigenvalues of M_n on the diagonal and V_n has the eigenvectors of M_n as columns. Letting (\cdot, \cdot) be the usual inner product, we then get

$$\mathbb{E}_{\text{inv}}(n, t) = \bar{e}_1 V_n D_n^t V_n^T \bar{w}_n = \sum_i v_i (\bar{v}_i, \bar{w}_n) \lambda_i^t,$$

where λ_i is the i th eigenvalue of M_n , \bar{v}_i the i th eigenvector of M_n and v_i is the first element in \bar{v}_i . Thus, we can write

$$\mathbb{E}_{\text{inv}}(n, t) = \sum_i a_i \lambda_i^t,$$

for some coefficients $a_i = v_i(\bar{v}_i, \bar{w}_n)$.

Example 12 *Let us take a look at S_3 . It contains six elements, which we sort in lexicographic order: $\{123, 132, 213, 231, 312, 321\}$. The adjacency matrix of the Cayley graph is*

$$A_2 = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}.$$

We see that from 123, we can get to 132 and 213 by an adjacent transposition, but the other permutations can not be reached. If we take the cube of A_2 , we get

$$A_2^3 = \begin{pmatrix} 0 & 3 & 3 & 0 & 0 & 2 \\ 3 & 0 & 0 & 2 & 3 & 0 \\ 3 & 0 & 0 & 3 & 2 & 0 \\ 0 & 2 & 3 & 0 & 0 & 3 \\ 0 & 3 & 2 & 0 & 0 & 3 \\ 2 & 0 & 0 & 3 & 3 & 0 \end{pmatrix}.$$

We see that there are three possible ways of reaching 213 from 123, but only two ways to reach 321. Thus, the probability of ending up at 213 after three moves is higher than for 321.

If we compute $\bar{e}_1 A_2^3 \bar{w}_2^T$, where $\bar{w}_2 = (0, 1, 1, 2, 2, 3)$, we get $0+3+3+0+0+6 = 12$, so the expected number of breakpoints after 3 transpositions is $\frac{12}{2^3} = \frac{3}{2}$.

To obtain the complete formula for $\mathbb{E}_{\text{inv}}(2, t)$, we diagonalise A_2 . The eigenvalues of A_2 , which become the eigenvalues of M_2 if we divide by $n = 2$, are given by $\{2, 1, 1, -1, -1, -2\}$. We observe that the largest eigenvalue equals n and that for each eigenvalue λ there is an eigenvalue $-\lambda$. In total, we get

$$\mathbb{E}_{\text{inv}}(2t) = \frac{\frac{3}{2}2^t - \frac{4}{3}1^t - \frac{1}{6}(-2)^t}{2^t} = \frac{3}{2} - \frac{4}{3} \frac{1}{2^t} - \frac{1}{6}(-1)^t.$$

Again, it is easy to see that the expected number of breakpoints after 3 transpositions is $\frac{3}{2}$.

The example is misleading; the eigenvalues of M_n are in general not integers, or even rational numbers. As a consequence, we have not been able to find out too much about these eigenvalues. We have, however, made some more or less trivial observations.

Lemma 13 *The greatest eigenvalue of M_n is n .*

PROOF. Otherwise, $\mathbb{E}_{\text{inv}}(n, t)$ would tend either to zero or to infinity. We could also use the Perron-Frobenius theorem on non-negative matrices.

□

Lemma 14 *The eigenvalues of M_n are symmetrical, i. e. the eigenvalues of M_n can be permuted such that each eigenvalue λ is mapped to $-\lambda$.*

PROOF. It is enough to show that the characteristic polynomial contains only even powers of λ . To do this, the key observation is that adjacent transposition transform odd permutations to even and vice versa. If we rearrange rows and columns of M_n , it can be written

$$M'_n = \left(\begin{array}{c|c} 0 & A \\ \hline B & 0 \end{array} \right).$$

We know that

$$|(M'_n - \lambda I)| = \sum_{\pi \in S_{(n+1)!}} \prod_{i=1}^{(n+1)!} m_{i, \pi(i)}.$$

If, for any $i \leq (n+1)!/2$, we have $i \neq \pi(i) \leq (n+1)!/2$, then the corresponding term is zero. The same goes for $i > (n+1)!/2$. Thus, the non-zero terms has a number of fixpoints and, in addition, $\pi(i) > (n+1)!/2$ for $i \leq (n+1)!/2, i \neq \pi(i)$ and vice versa. But then the number of $i \leq (n+1)!/2$ such that $i \neq \pi(i)$ must equal the number of $i > (n+1)!/2$ such that $i \neq \pi(i)$, which means that the number of fixpoints in π is even (since $(n+1)!$ is even). Thus we only get even powers of λ in the characteristic polynomial.

□

Lemma 15 *In the expression*

$$\mathbb{E}_{\text{inv}}(n, t) = \sum_i a_i \lambda_i^t,$$

the coefficient of the smallest eigenvalue $-n$ is zero for $n \geq 3$.

PROOF. It is clear that the eigenvalue $-n$ has multiplicity one and that its eigenvector has 1 at the components that correspond to even permutations and -1 otherwise. The coefficient is given by $v_i(\bar{v}_i, \bar{w}_n)$. We will now show that for $n \geq 3$, (\bar{v}_i, \bar{w}_n) is always zero.

Equivalently, we need to show that the sum of the even inversion numbers equals the sum of the odd inversion numbers, that is

$$\sum_{i=0}^{(n+1)!} (-1)^{w_i} w_i = 0$$

for $n \geq 3$. Now, we know (see for instance Stanley [5]) that

$$\sum_{\pi \in S_{n+1}} x^{\text{inv}(\pi)} = \prod_{k=1}^n (1 + x + x^2 + \dots + x^k).$$

Taking the derivative, we get

$$\begin{aligned} \sum_{\pi \in S_{n+1}} \text{inv}(\pi) x^{\text{inv}(\pi)-1} &= (1+x) \frac{d}{dx} \prod_{k=2}^n (1+x+\dots+x^k) \\ &\quad + \prod_{k=2}^n (1+x+\dots+x^k). \end{aligned}$$

If we let $x = -1$, we get the desired result for $n \geq 3$.

□

This lemma casts some light on the behaviour of $\mathbb{E}_{\text{inv}}(n, t)$ for large t .

Corollary 16 *As t goes to ∞ , $\mathbb{E}_{\text{inv}}(n, t)$ approaches the same limit $\frac{\binom{n}{2}}{2}$ for even and odd t , if $n \geq 3$.*

PROOF. As t goes to infinity, $\frac{\lambda^t}{n^t}$ goes to zero for all eigenvalues λ except for $\lambda = n$ and $\lambda = -n$. The coefficient of $\lambda = -n$ is zero, from the previous

lemma, and the coefficient of $\lambda = n$ is consequently given by $\binom{n}{2}/2$, since the probability that any two elements in a random permutation are in reversed order is exactly one half.

□

6 Solutions for other Coxeter groups

It is interesting to note that this problem, which can be solved by simple combinatorial arguments, does not lend itself to the Markov chain approach, which in general seems to be the most useful technique in this context. We have not been able to understand why this is the case, nor to give a characterisation of the problems that can be solved with one technique or the other. Solving more problems may assist in this search.

If we view S_{n+1} as the Coxeter group A_n , the adjacent transposition take the role of generators, or simple reflections. It is natural to consider the more general problem of finding the expected word length of a word made of t random generators in any Coxeter group, or at least some of them.

This has been done by Emma Troili in her Master's thesis [6]. She considered the Coxeter groups B_n , $I(m)$ and \tilde{A}_n . For B_n , the best approach seemed to be the Markov chain approach, which did not give a full solution. For $I(m)$, combinatorial reasoning gave the formula

$$\mathbb{E}(n, t) = 1 + \sum_{j=1}^{\lfloor \frac{t-1}{2} \rfloor} \frac{1}{4^j} \sum_{k=0}^{\frac{j}{m}} \binom{2j}{j - km} - \begin{cases} \sum_{j=1}^{\lfloor \frac{t-1}{2} \rfloor} \left(\frac{2}{4^j} \sum_{k=1}^{\frac{j}{m} + \frac{1}{2}} \binom{2j}{2j - (2k-1)m} \right) \\ \sum_{j=1}^{\lfloor t/2 \rfloor} \left(\frac{1}{4^{j-1}} \sum_{k=1}^{\frac{2j-1}{2m} + \frac{1}{2}} \binom{2j-1}{2j-1 - (2k-1)m} \right), \end{cases}$$

for even and odd m , respectively. For \tilde{A}_n , using a semi-infinite grid for the heat process was the appropriate line of thinking. This gave

$$\mathbb{E}(n, t) = \sum_{r=1}^t \binom{t}{r} \left(\frac{2}{n} \right)^{r-1} (-1)^r C_{r-1}.$$

Acknowledgments

For the proof of Lemma 9, the author is indebted to Axel Hultman and Sloane's On-Line Encyclopedia of Integer Sequences.

References

- [1] Niklas Eriksen, Approximating the expected number of inversions given the number of breakpoints. *Algorithms in Bioinformatics*, Proceedings of WABI 2002, LNCS 2452, 316–330.
- [2] Niklas Eriksen and Axel Hultman, Estimating the expected reversal distance after a fixed number of reversals, *Advances in Applied Mathematics*, **32** (2004), 439–453.
- [3] Henrik Eriksson, Kimmo Eriksson and Jonas Sjöstrand, Expected inversion number after k adjacent transpositions, in D. Krob, A.A. Mikhalev, A.V. Mikhalev, eds., *Proceedings of Formal Power Series and Algebraic Combinatorics* (Springer Verlag, 2000) 677–685.
- [4] Sri Gopal Mohanty, *Lattice path counting and applications* (Academic Press, London, 1979).
- [5] Richard Stanley, *Enumerative combinatorics*, vol. 1 (Cambridge University Press, New York/Cambridge, 1997).
- [6] Emma Troili, *Förväntade avstånd i Coxetergrupper (Expected distances in Coxeter groups)*, Master’s thesis (in Swedish), Department of Mathematics, KTH (2002).
- [7] Li-San Wang and Tandy Warnow, Estimating true evolutionary distances between genomes, *Proceedings of the Thirty-Third Annual ACM Symposium on the Theory of Computing (STOC’01)* (2001).