EXPECTED NUMBER OF INVERSIONS AFTER A SEQUENCE OF RANDOM ADJACENT TRANSPOSITIONS — AN EXACT EXPRESSION

NIKLAS ERIKSEN

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.

RÉSUMÉ. Une formule pour calculer le nombre attendu d'inversions après t transpositions adjacentes aléatoires a été présentée par Eriksson et al. Nous avons amélioré ce résultat en déterminant une formule pour la séquence inconnue d'entiers d_r, qui était utilisée dans leur formule et qui rendait la formule valide lorsque t prend une grande valeur.

1. INTRODUCTION

In a recent article [1], 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 $\rho = id$. 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 $\rho = id$ (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. One such common measure is the number of breakpoints, that is the number of adjacencent 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 [2], for breakpoints and the long range inversions and transpostions usually considered in computational biology.

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

Theorem 1.1. *The expected number of inversions in a permutation in S_{n+1} after t random adjacent transpositions is, for $n \geq t$,*

$$E_{nt} = \sum_{r=0}^{t} \frac{(-1)^r}{n^r} \left[\binom{t}{r+1} 2^r C_r + 4 d_r \binom{t}{r} \right],$$

where d_r is an integer sequence that begins with $0, 0, 0, 1, 9, 69, 510$ and C_r 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 1.2. *The expected number of inversions in a permutation in S_{n+1} after t random adjacent transpositions is*

$$E_{nt} = \sum_{r=1}^{t} \frac{1}{n^r} \binom{t}{r} \sum_{s=1}^{r} \binom{r-1}{s-1}(-1)^{r-s}4^{r-s}g_{s,n}.$$

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

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

For $n \geq t$, we get

$$E_{nt} = \sum_{r=0}^{t}\frac{(-1)^r}{n^r}\left[2^r C_r\binom{t}{r+1}+2\binom{t}{r}\sum_{s=3}^{r}\binom{r-1}{s-1}(-1)^{s-1}4^{r-s}\binom{2\lfloor\frac{s}{2}\rfloor}{\lfloor\frac{s}{2}\rfloor}\sum_{l=0}^{\lfloor\frac{s-1}{2}\rfloor}l\binom{2\lceil\frac{s}{2}\rceil-1}{\lceil\frac{s}{2}\rceil+l}\right]$$

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

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

2. THE HEAT FLOW MODEL

To prove Theorem 1.2, 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

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

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

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

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

for a permutation $\pi \in \mathcal{P}_{nt}$, where the adjacent transpositions $s_k, 1 \leq k \leq t$ have been chosen randomly from a uniform distribution. From this, it follows that

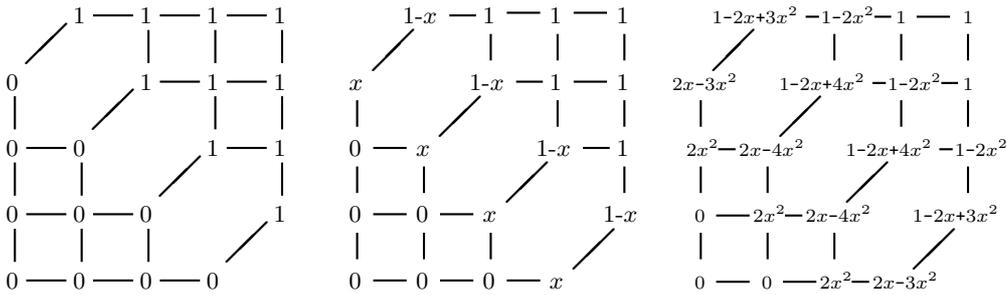$$E_{nt} = \sum_{i>j} p_{ij}(t).$$

FIGURE 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 [1].

Proposition 2.1. *(Eriksson et al. [1]) The sequence of (p_{ij})-matrices for $t = 0, 1, 2, \ldots$ describes a discrete heat flow process with conductivity $x = 1/n$ on the grid graph depicted in Figure 1 (left).*

In the same paper, they also show that we can replace the graph in Figure 1 by the grid in Figure 2. The sequence of (p_{ij})-matrices for $t = 0, 1, 2, \ldots$ 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. 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.
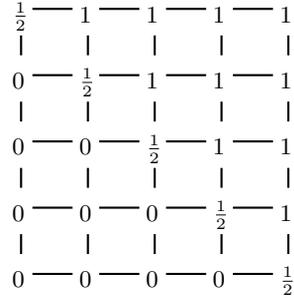

FIGURE 2. Grid graph with initial values

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 E_{nt}.
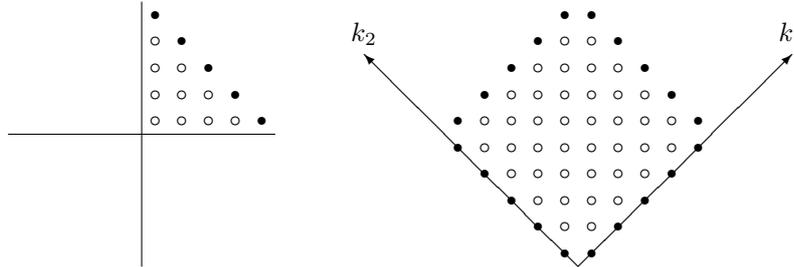

FIGURE 3. By reflection, the graph with one hot and two insulated boundaries is extended to a diamond shaped graph with no insulated boundaries. The new set of coordinates (k_1, k_2) is introduced.

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. In the model proposed by Eriksson et al., 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. There are three possible travel steps for a packet [1]:

- 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 traveled 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 traveled 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 walks, which we call **two-sided Catalan walks**.

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 E_{nt}.

We are now able to prove the first part of Theorem 1.2. 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 Catalan walks from $(0, a)$ to (s, b) (with a odd), where the y-coordinate corresponds to k_2, and two-sided Catalan walks 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 Catalan walks, respectively. This yields, with $x = 1/n$,

$$E_{nt} = \frac{1}{2}\sum_{r=1}^{t}\frac{1}{n^r}\binom{t}{r}\sum_{s=1}^{r}\binom{r-1}{s-1}(-1)^{r-s}4^{r-s}b_{s,n}c_{s-1,n}.$$

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

3. Two-sided Catalan walks

We start by formally defining two-sided Catalan walks and then proceed to enumerate them.

Definition 3.1. *A **two-sided Catalan walk of height** n is a walk on the integer grid from $(0, a)$ to (s, b), where $a, b \in \{1, 2, \ldots n - 1\}$ and $s > 0$, 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 Catalan walks 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 3.2. *The number of two-sided Catalan walks 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.

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

Lemma 3.3. *The number of two-sided Catalan walks 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,

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

Most terms cancel by symmetry of the binomial coefficients. For odd s, we see that for each two-sided Catalan walk to $x = s$ we get two such walks to $x = s + 1$. $\square$

Lemma 3.4. *The number of two-sided Catalan walks of height $2n + 2$ from $(0, a)$ to (s, b) for all $0 < a, b < 2n + 2$, a odd, is given by*

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

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. Similarily, 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. to $(n - 2n)\binom{s}{\frac{s+1}{2}+n}$. We then 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 down.

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. $\square$

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

Lemma 3.5.

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

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

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

$\square$

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

$$g_{s,n} = b_{s,n}c_{s-1,n} = n2^{s-1}\binom{s-1}{\lceil\frac{s-1}{2}\rceil} - 2\binom{2\lfloor\frac{s}{2}\rfloor}{\lfloor\frac{s}{2}\rfloor}\sum_{l=0}^{\lfloor\frac{s-1}{2}\rfloor}l\binom{2\lceil\frac{s}{2}\rceil-1}{\lceil\frac{s}{2}\rceil+l}.$$

This yields

$$E_{nt} = \sum_{r=0}^{t}\frac{(-1)^r}{n^r}\left[2^r C_r\binom{t}{r+1} + 2\binom{t}{r}\sum_{s=3}^{r}\binom{r-1}{s-1}(-1)^{s-1}4^{r-s}\binom{2\lfloor\frac{s}{2}\rfloor}{\lfloor\frac{s}{2}\rfloor}\sum_{l=0}^{\lfloor\frac{s-1}{2}\rfloor}l\binom{2\lceil\frac{s}{2}\rceil-1}{\lceil\frac{s}{2}\rceil+l}\right].$$

4. AN ALTERNATIVE FORMULA

There is another way of writing E_{nt} 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 need no longer keep track of the true travel days (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:

$$E_{nt} = \frac{1}{2}\sum_{q=1}^{t}\sum_{r=0}^{t-q}\binom{t-q}{r}\left(1-\frac{4}{n}\right)^{t-q-r}\frac{1}{n^{r+1}}b_{r+1,n}c_{r,n}.$$

This gives the following theorem.

Theorem 4.1. *The expected number of inversions in a permutation in S_{n+1} after t random permutations is given by*

$$E_{nt} = \sum_{u=0}^{t-1}\left(\frac{n-4}{n}\right)^u\sum_{r=0}^{u}\binom{u}{r}\frac{1}{(n-4)^r}(2^r + \frac{2\beta_{r+1,n}}{n})c_{r,n}.$$

Proof. Trivial calculations give

$$E_{nt} = \frac{1}{2}\sum_{q=1}^{t}\sum_{r=0}^{t-q}\binom{t-q}{r}\left(1-\frac{4}{n}\right)^{t-q-r}\frac{1}{n^{r+1}}b_{r+1,n}c_{r,n}$$

$$= \frac{1}{2}\sum_{u=0}^{t-1}\sum_{r=0}^{u}\binom{u}{r}\left(1-\frac{4}{n}\right)^{u-r}\frac{1}{n^{r+1}}b_{r+1,n}c_{r,n}$$

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

$\square$

This expression seems particularly useful for fixed n (try for instance $n = 4$). Also, it is easy to find out how much E_{nt} increases when we increase t one step. This is given by

$$\Delta_t E_{nt} = E_{n,t+1} - E_{nt} = \sum_{r=0}^{t}\binom{t}{r}\left(1-\frac{4}{n}\right)^{t-r}\frac{1}{n^{r+1}}b_{r+1,n}c_{r,n}.$$

It is easy to see that $\Delta_t E_{nt}$ is always positive for $n \geq 4$. This means that E_{nt} 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), $E_{1,t}$ takes the values $0, 1, 0, 1, 0, 1 \ldots$, which is not a monotone sequence.

To be able to apply this in the biological context, where we wish to estimate the number of transpositions given the inversion number of a permutation, we need this monotonicity property. The reason is that when we have found an expectation value E_{nt} which is close to our number of inversions, we must be sure that we will not find a better expectation value for a much larger t. If the sequence if monotone, this can never happen.

Acknowlegdments

References

[1] Henrik Eriksson, Kimmo Eriksson, Jonas Sjöstrand, Expected inversion number after k adjacent transpositions, in Formal Power Series and Algebraic Combinatorics, Krob, D., Mikhalev, A.A., Mikhalev, A.V. (Eds.), Springer Verlag (2000), 677–685.

[2] Li-San Wang, Exact-IEBP: A New Technique for Estimating Evolutionary Distances between Whole Genomes. Algorithms in Bioinformatics, Proceedings of WABI 2001, LNCS 2149, 175–188

E-mail address: niklas@math.kth.se

DEPARTMENT OF MATHEMATICS, ROYAL INSTITUTE OF TECHNOLOGY, S-100 44 STOCKHOLM, SWEDEN