Available online at www.sciencedirect.com
SCIENCE DIRECT®
Advances in Applied Mathematics 32 (2004) 439–453
ADVANCES IN Applied Mathematics
www.elsevier.com/locate/yaama

Estimating the expected reversal distance after a fixed number of reversals

Niklas Eriksen [*,1] and Axel Hultman

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

Received 11 February 2003; accepted 20 February 2003

Abstract

We address the problem of computing the expected reversal distance of a genome with n genes obtained by applying t random reversals to the identity. A good approximation is the expected transposition distance of a product of t random transpositions in S_n. Computing the latter turns out to be equivalent to computing the coefficients of the length function (i.e., the class function returning the number of parts in an integer partition) when written as a linear combination of the irreducible characters of S_n. Using symmetric functions theory, we compute these coefficients, thus obtaining a formula for the expected transposition distance. We also briefly sketch how to compute the variance.

Keywords: Sorting by reversals; Genome rearrangements; Permutations; Transpositions; Expected distances

1. Introduction

Over the last decade, the computational biology community has been looking at the problem of estimating evolutionary distances between species from their gene order. The probably most common, and by far most studied, evolutionary operation in this context is the *reversal*: a segment (that is a sequence of consecutive genes) of the genome is taken out and inserted at the same place, but in reversed order. In 1999, Hannenhalli and Pevzner [6] presented a formula for the minimal number of reversals needed to transform one sequence of distinct genes into a given permutation of them.

* Corresponding author.
E-mail address: niklas@math.kth.se (N. Eriksen).
[1] Niklas Eriksen was supported by a grant from the Swedish Research Council.

0196-8858/$ – see front matter
doi:10.1016/S0196-8858(03)00054-X

For distant genomes, the true evolutionary distance is in general much longer than the shortest distance. In order to find a better estimate of the true distance, we may instead look at the expected distance. Such attempts have been made by Wang and Warnow [11] and Eriksen [4], providing bounds for and an approximation of the expected reversal distance given the number of *breakpoints* between two genomes π and τ, respectively. (There is a breakpoint between two genes in π if they are adjacent in π but not in τ.)

The reversal distance "contains more information" than the breakpoint distance, so if we could find the expected reversal distance, we would probably obtain a biologically more relevant formula. For the same reasons, we would also expect this problem to be harder. The inverse problem seems to be of more reasonable difficulty:

Problem 1.1. Compute the expected reversal distance after t random reversals, taken independently from the uniform distribution.

In this paper (Section 3), we find an analogy between certain cycles used by Hannenhalli and Pevzner, and the ordinary cycles in the symmetric group. We reach the conclusion that one can obtain a good estimate to the expected reversal distance by solving the following analogous problem:

Problem 1.2. Compute the expected *transposition* distance in the symmetric group S_n after t random transpositions, taken independently from the uniform distribution.

In Sections 4 and 5 we find the solution to Problem 1.2 to be the following formula:

$$\mathbb{E}_{\mathrm{trp}}(n,t) = n - \sum_{k=1}^{n} \frac{1}{k} + \sum_{p=1}^{n-1} \sum_{q=1}^{\min(p,n-p)} a_{pq} \left(\frac{\binom{p}{2} + \binom{q-1}{2} - \binom{n-p-q+2}{2}}{\binom{n}{2}} \right)^t, \qquad (1)$$

where

$$a_{pq} = (-1)^{n-p-q+1} \frac{(p-q+1)^2}{(n-q+1)^2(n-p)} \binom{n-p-1}{q-1} \binom{n}{p}.$$

Finally, in Section 6 we show how the inverse of (1) can be used as an estimate for the expected evolutionary reversal distance and investigate numerically how well this formula behaves compared to previous methods when it comes to predicting the true evolutionary distance.

It should be noted that "expected reversal distances" have been studied as early as 1996 [1]. That paper, however, dealt with the expected reversal distance of a linear, unsigned genome taken from the uniform distribution. We will give the full answer to the problem of computing the expected transposition distance of a permutation taken from the uniform distribution, which gives an approximation of the expected reversal distances of circular and signed genomes, taken from the uniform distribution.

Remark 1.3. In [5], we use analogous, but somewhat more involved, methods to solve Problem 1.2 for the complex reflection groups $G(r,1,n) \cong (\mathbb{Z}/r\mathbb{Z}) \wr S_n$. The symmetric

group is then the special case $r = 1$. For $r > 1$, there are no immediate applications to computational biology.

2. Preliminaries

Let S_n be the symmetric group on n elements and let $d_{\mathrm{trp}}(\pi)$ be the transposition distance from a permutation π to the identity permutation, i.e., the minimal i such that π is a product of i transpositions. It is well known since Cayley that $d_{\mathrm{trp}}(\pi) = n - c_{\mathrm{trp}}(\pi)$, where $c_{\mathrm{trp}}(\pi)$ is the number of cycles in π.

A *genome with n genes* is a signed, circular permutation on n elements. All genomes are assumed to be read counterclockwise. Two genomes are equivalent if you can obtain one from the other by reading it backwards and changing all signs. Disregarding the signs yields an *unsigned genome*. We will denote the set of all genomes with n genes by G_n. The identity genome is denoted $\mathrm{id} = 1\,2\,\ldots\,n$. We take the liberty of writing a genome $\pi \in G_n$ in a linear fashion. It is then understood that the leftmost gene should be attached to the rightmost gene.

Example 2.1. The genome in Fig. 1 can be written as, for instance, $1\,3\,{-2}$ or $3\,{-2}\,1$ or even $-3\,{-1}\,2$ (reading in the opposite direction). Usually, we let 1 be the first element in the linear order.

In this paper, we will consider an evolutionary event called *reversal* (or inversion). A reversal between π_i and π_j, where $i \neq j$, is an operation that takes the segment $\pi_{i+1}\pi_{i+2}\ldots\pi_j$ out of the genome and inserts it at the same place backwards, changing the signs of all elements in the segment. This is depicted in Fig. 2.

2.1. The breakpoint graph

The breakpoint graph of a genome π was used by Hannenhalli and Pevzner in 1999 [6] to find the *reversal distance* $d_{\mathrm{rev}}(\pi)$ between π and id. One should note that since we can always rename the genes in two genomes such that one of them becomes the identity, this gives the reversal distance between any pair of genomes.

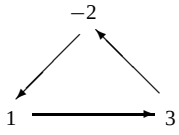

Fig. 1. An example genome.

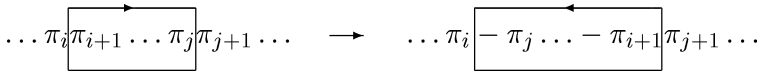

Fig. 2. The reversal.

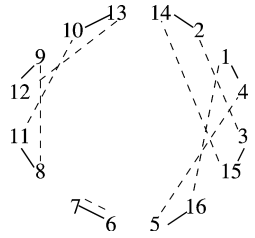

Fig. 3. The breakpoint graph of $\pi = 1\,{-7}\,{-5}\,{-6}\,{-4}\,{-3}\,{-8}\,2$.

Two genes a and b in a genome π are said to be *consecutive* if b follows directly after a or $-a$ follows directly after $-b$ in π. Observe that a and b is an ordered pair, so if a and b are consecutive in π, then b and a are in general not. There is a *breakpoint* between a and b in π (relative to id) if a and b are consecutive in id but not in π. We denote the number of breakpoints by $b(\pi)$.

Let U_{2n} denote the set of unsigned genomes with $2n$ genes. Following [6], we define the *genome transformation map* $\mathrm{gtm}: G_n \to U_{2n}$ as follows: each gene a in $\pi \in G_n$ is mapped to the pair of genes $(2a-1, 2a)$ if $a > 0$, and mapped to $(-2a, -2a-1)$ if $a < 0$. In the pair of genes obtained from a, we will denote the left element by a_L and the right by a_R. We then take these pairs in the same order as the corresponding genes appear in π. For instance, the genome $\pi = 1\,{-5}\,3\,2\,{-4}$ is mapped to the unsigned genome $\mathrm{gtm}(\pi) = 1\,2\,10\,9\,5\,6\,3\,4\,8\,7$. Note that the number of breakpoints relative to the identity is preserved by this transformation, that is $b(\pi) = b(\mathrm{gtm}(\pi))$.

The *breakpoint graph* $G(\pi)$ of $\pi \in G_n$ has the genes in $\mathrm{gtm}(\pi)$ as vertices. There is a solid edge between a_R and b_L if a and b are consecutive in π and there is a dashed edge between $2k$ and $2k+1$ and between $2n$ and 1. An example of a breakpoint graph can be viewed in Fig. 3.

It is fairly easy to see that each vertex in $G(\pi)$ has valency two, and that no vertex has two edges of the same colour. Hence, the edges form alternating cycles. We will call the number of solid edges in such a cycle the *length* of the cycle.

From now on, we assume the breakpoint graph of π to be drawn with its vertices on a circle, counterclockwise in the order given by $\mathrm{gtm}(\pi)$. A cycle is *oriented* if it has length 1 or if, when we traverse it, we do not traverse all the solid edges in the same direction (clockwise or counterclockwise). Otherwise, the cycle is *unoriented*.

We now present an equivalence relation on the cycles. An *interval* on a genome is a segment of consecutive genes. We say that two cycles are equivalent if, when we take one interval containing all the vertices of the first cycle and another interval containing all the vertices of the second cycle, the intervals are always intersecting. The equivalence classes are called *components*. A component is *oriented* if it contains at least one oriented cycle and *unoriented* otherwise.

If there is an interval that contains (the vertices of) an unoriented component τ, but no other unoriented components, then τ is known as a *hurdle*. If there is an interval which contains exactly two unoriented components and possibly some oriented ones, and exactly one of these unoriented components is a hurdle, then this hurdle is a *super hurdle*. Finally,

if a breakpoint graph contains an odd number of hurdles, all of which are super hurdles, then this graph is known as a *fortress*.

For $\pi \in G_n$, we define $c_{\text{rev}}(\pi)$ to be the number of cycles in $G(\pi)$. Similarly, $h(\pi)$ is its number of hurdles. Finally, $f(\pi)$ is one if $G(\pi)$ is a fortress, zero otherwise. Using these functions, we can formulate the theorem of Hannenhalli and Pevzner.

Theorem 1 [6]. *The reversal distance is given by*

$$d_{\text{rev}}(\pi) = n - c_{\text{rev}}(\pi) + h(\pi) + f(\pi).$$

It follows from Caprara [3] that genomes containing hurdles are very rare. For instance, for genomes of length 8, less than one percent of these contain hurdles, and for genomes of length 100, only one in 10^5 contains a hurdle. Thus, there is little harm in using the approximation $d_{\text{rev}}(\pi) \approx n - c_{\text{rev}}(\pi)$, since $h(\pi) + f(\pi) = 0$ if π does not contain any hurdle. Observe the similarity between this formula and the one governing the transposition distance in S_n.

3. The analogy

We shall now explore the analogy between unsigned transpositions and signed reversals. If we apply a transposition $\tau = (a\ b)$ to a permutation $\pi \in S_n$, one of the following things will happen.

- If a and b belong to different cycles in π, the number of cycles will decrease by one.
- If a and b belong to the same cycle in π, the number of cycles will increase by one.

Thus, applying a transposition to π will change the transposition distance by one.

On the other hand, if we apply the reversal $a \ldots b$ to $\pi \in G_n$, one of the following things will happen.

- If a_L and b_R belong to different cycles in $\text{gtm}(\pi)$, the number of cycles will decrease by one.
- If a_L and b_R belong to the same cycle and the solid edges connected to a_L and b_R are traversed in different directions when we traverse this cycle, the number of cycles will increase by one.
- If a_L and b_R belong to the same cycle and the solid edges connected to a_L and b_R are traversed in the same direction when we traverse this cycle, the number of cycles will stay the same.

Applying a reversal to a genome π will thus change $d_{\text{rev}}(\pi)$ by one, unless the reversal cuts two equally directed solid edges in the same cycle, while not creating or destroying a hurdle or altering the value of $f(\pi)$.

From this analysis, we find that if we apply a random transposition to a permutation π and the corresponding reversal to a genome σ with the same cycle structure (that is

there exists a length preserving bijection between the cycles of π and the cycles of $G(\sigma)$), then the distances to the identities will in most cases change by an equal amount. This approximation holds particularly well for permutations and genomes close to the identity. It seems reasonable that the expected distances after t operations will be approximately equal, at least for $t \leqslant n$, say.

We must not carry this analogy too far; there are major dissimilarities between S_n and G_n. Still, as we will see in the paper, the similarities described above are sufficient to draw conclusions on the behaviour of genomes from the behaviour of permutations, when subject to reversals and transpositions, respectively.

4. The Markov chain approach

We wish to compute $\mathbb{E}_{\mathrm{trp}}(n,t)$, the expected transposition distance in S_n given that t random transpositions have been applied to the identity permutation. One possible approach to calculating $\mathbb{E}_{\mathrm{trp}}(n,t)$ would be to let each one of the $n!$ permutations in S_n correspond to a state in a Markov chain, where at each step we apply a transposition, chosen randomly from a uniform distribution. A more economical approach, however, is obtained from the observation that all permutations in some conjugacy class are equally probable. We thus let the conjugacy classes, each one corresponding to an integer partition of n, constitute the states in our Markov chain.

We adopt the convention of sorting the integer partitions $\lambda = (\lambda_1 \geqslant \lambda_2 \geqslant \cdots)$ in reverse lexicographical order.

Calculating the transition matrix is not too hard. Say that we wish to compute the probability that we go between states λ and μ. Such a transition is possible if λ, say, has two parts a and b which sum up to one part c of μ, all other parts in λ equalling the other parts in μ. Then the probability that we go from λ to μ, given that λ has p parts equal to a and q parts equal to b, is $paqb/\binom{n}{2}$ if $a \neq b$ and $\binom{p}{2}a^2/\binom{n}{2}$ otherwise. The probability that we go from μ to λ, given that μ has r parts equal to c, is $cr/\binom{n}{2}$ if $a \neq b$ and $cr/2\binom{n}{2}$ otherwise. In order to obtain integer matrices, we multiply the transition matrices by $\binom{n}{2}$.

Example 4.1. For $n = 4$, the transition matrix multiplied by $\binom{n}{2}$ is given by

$$M_4 = \begin{pmatrix} 0 & 6 & 0 & 0 & 0 \\ 1 & 0 & 1 & 4 & 0 \\ 0 & 2 & 0 & 0 & 4 \\ 0 & 3 & 0 & 0 & 3 \\ 0 & 0 & 2 & 4 & 0 \end{pmatrix}.$$

What we wish to calculate is $\mathbb{E}_{\mathrm{trp}}(n,t) = e_1 M_n^t w_n^T / \binom{n}{2}^t$, where $e_1 = (1,0,\ldots,0)$ and $w_n = (n - \ell(\lambda))_{\lambda \vdash n}$, where $\ell(\lambda)$ is the number of parts in λ. In other words, w_n contains the transposition distances from the corresponding conjugacy classes to the identity class. In order to do this, we can diagonalise $M_n = V_n D_n V_n^{-1}$. It follows from Ito [7] that each irreducible character χ^λ contributes to the spectrum an eigenvalue $\binom{n}{2}\chi^\lambda(2,1^{n-2})/\chi^\lambda(1^n)$. These eigenvalues are easy to compute, as was noted already by Frobenius. In the Ferrers

diagram of a partition λ, fill each square (i, j) with its *content* $j - i$. Taking the sum c_λ of the contents of all squares in the Ferrers diagram of λ, we get the eigenvalues of M_n. Computing this leads to

$$c_\lambda = \sum_{i=1}^{\ell(\lambda)} \binom{\lambda_i}{2} - (i-1)\lambda_i$$

Moreover, the eigenvectors are given by the irreducible characters indexed by the corresponding partitions. Since the irreducible characters are orthonormal in the usual inner product

$$\langle \chi^\lambda, \chi^\nu \rangle = \sum_{\mu \vdash n} \frac{\chi^\lambda(\mu)\chi^\nu(\mu)}{z_\mu},$$

we obtain the inverse of V_n from V_n^T by dividing each column by the appropriate $z_\lambda = 1^{m_1} m_1! 2^{m_2} m_2! \ldots n^{m_n} m_n!$ for $\lambda = (1^{m_1}, 2^{m_2}, \ldots, n^{m_n})$.

Example 4.2. For $n = 4$, we have the eigenvalues 6, 2, 0, -2, and -6. The matrix V_4 is given by

$$V_4 = \begin{pmatrix} 1 & 3 & 2 & 3 & 1 \\ -1 & -1 & 0 & 1 & 1 \\ 1 & -1 & 2 & -1 & 1 \\ 1 & 0 & -1 & 0 & 1 \\ -1 & 1 & 0 & -1 & 1 \end{pmatrix}$$

and its inverse by

$$V_4^{-1} = \frac{1}{4!} \begin{pmatrix} 1 & -6 & 3 & 8 & -6 \\ 3 & -6 & -3 & 0 & 6 \\ 2 & 0 & 6 & -8 & 0 \\ 3 & 6 & -3 & 0 & -6 \\ 1 & 6 & 3 & 8 & 6 \end{pmatrix}.$$

With this information, we find that

$$e_1 M_n^t w_n^T = \sum_{\lambda \vdash n} \chi^\lambda(1^n) \left(\frac{c_\lambda}{\binom{n}{2}} \right)^t \sum_{\mu \vdash n} \frac{\chi^\lambda(\mu) w_\mu}{z_\mu}.$$

The information we need to compute this is gathered in the next theorem.

Theorem 2. *If* $\lambda_3 \geqslant 2$, *then*

$$\sum_{\mu \vdash n} \frac{\chi^\lambda(\mu) w_\mu}{z_\mu} = 0.$$

Otherwise, let $\lambda = (p, q, 1^{n-p-q})$. *Then, for* $p, q \geqslant 1$,

$$\sum_{\mu \vdash n} \frac{\chi^{\lambda}(\mu) w_{\mu}}{z_{\mu}} = (-1)^{n-p-q+1} \frac{p-q+1}{(n-q+1)(n-p)},$$

and, for $p = n, q = 0$,

$$\sum_{\mu \vdash n} \frac{\chi^{(n)}(\mu) w_{\mu}}{z_{\mu}} = n - \sum_{k=1}^{n} \frac{1}{k}.$$

We postpone the proof of this theorem to Section 5.1. Using this theorem, we can give a closed formula for the expected transposition distance after t random transpositions in S_n.

Corollary 4.3. *The expected transposition distance after* t *random transpositions in* S_n *is given by*

$$n - \sum_{k=1}^{n} \frac{1}{k} + \sum_{p=1}^{n-1} \sum_{q=1}^{\min(p,n-p)} a_{pq} \left(\frac{\binom{p}{2} + \binom{q-1}{2} - \binom{n-p-q+2}{2}}{\binom{n}{2}} \right)^{t},$$

where

$$a_{pq} = (-1)^{n-p-q+1} \frac{(p-q+1)^2}{(n-q+1)^2(n-p)} \binom{n-p-1}{q-1} \binom{n}{p}.$$

Proof. The character $\chi^{\lambda}(1^n)$ is given (see [8] or [10]) by the hook-length formula

$$\chi^{\lambda}(1^n) = \frac{n!}{\prod_{c \in \lambda} h_c}.$$

For $\lambda = (p, q, 1^{n-p-q})$, this yields

$$\chi^{\lambda}(1^n) = \frac{n!(p-q+1)}{(q-1)!(n-p-q)!(n-p)(n-q+1)p!} = \frac{(p-q+1)}{(n-q+1)} \binom{n-p-1}{q-1} \binom{n}{p}.$$

Since the c_{λ} of such a partition is

$$\binom{p}{2} + \binom{q-1}{2} - \binom{n-p-q+2}{2},$$

the corollary follows. □

Of interest is the behaviour of the expected distance as t grows (keeping n fixed). Depending on the parity of t, one of two limits is approached. It is not surprising that what we obtain for even (odd) t is exactly the expected distance of a randomly chosen

even (odd) permutation of $[n]$ from a uniform distribution. We leave the verification of this statement to the reader.

Corollary 4.4. *We have*

$$\lim_{t\to\infty} \mathbb{E}_{\mathrm{trp}}(n, 2t) = n - \sum_{k=1}^{n} \frac{1}{k} + (-1)^{n-1} \frac{1}{n(n-1)}$$

and

$$\lim_{t\to\infty} \mathbb{E}_{\mathrm{trp}}(n, 2t+1) = n - \sum_{k=1}^{n} \frac{1}{k} + (-1)^{n} \frac{1}{n(n-1)}.$$

Proof. As t grows, all terms but one in the double sum of Corollary 4.3 tend to zero, the exception being given by $p = q = 1$. This term is $(-1)^{t+n-1}\frac{1}{n(n-1)}$. Substituting $2t$ and $2t+1$, respectively, for t yields the result. □

5. Decomposing the length function

Recall that the *length*, $\ell(\lambda)$, of a partition λ is its number of parts. In this section we will use elements of symmetric functions theory in order to prove our main technical result: a decomposition formula for ℓ. To this end, we briefly review the material we need. For terminology not explained here, we refer the reader, e.g., to Macdonald [8] or Stanley [10, Chapter 7].

Let R^n be the vector space (over $\mathbb{Q}$, say) of *class functions*, i.e., functions $f : P_n \to \mathbb{Q}$, where P_n is the set of integer partitions of n. The irreducible S_n-characters, $\{\chi^\lambda\}_{\lambda\vdash n}$ form an orthonormal basis of R^n with respect to the inner product defined by $\langle f, g\rangle = \frac{1}{n!}\sum_{\pi\in S_n} f(\mathrm{type}(\pi))g(\mathrm{type}(\pi))$. As a vector space, R^n is isomorphic to the space Λ^n of symmetric functions of degree n via the *characteristic map*, $\mathrm{ch}^n : R^n \to \Lambda^n$, defined by $f \mapsto \sum_{\lambda\vdash n} \frac{p_\lambda}{z_\lambda} f(\lambda)$. Here, $\frac{n!}{z_\lambda}$ is the number of permutations of cycle type λ in S_n and $\{p_\lambda\}_{\lambda\vdash n}$ is the Λ^n-basis of *power sums*. We will use one more basis of Λ^n. The *Schur function* s_λ is the image of χ^λ under ch^n, hence the Schur functions form a basis.

If λ and μ are two partitions such that the Ferrers diagram of λ is contained in that of μ, then μ/λ denotes the part of the μ-diagram not contained in λ. We call μ/λ a *border strip* if it is connected (meaning that we can walk from any square to any other without crossing corners) and contains no 2×2 subsquare. The *height*, $\mathrm{ht}(\mu/\lambda)$, of the border strip μ/λ is one less than the number of rows in its diagram.

Richard Stanley pointed out to us the usefulness of the following two equations to proving Theorem 3 below. Letting the first t y-variables be equal to one and the rest be zero in [10, 7.20], then differentiating with respect to t, putting $t = 1$ and considering only terms of degree n yields

$$\sum_{\lambda\vdash n} \frac{\ell(\lambda)}{z_\lambda} p_\lambda = \sum_{k=1}^{n} \frac{1}{k} s_{(n-k)}\, p_{(k)}. \tag{2}$$

The following is a special case of [10, 7.72]. The sum is over all $\lambda \vdash n$ such that $\lambda/(n-k)$ is a border strip

$$s_{(n-k)}\, p_{(k)} = \sum_{\lambda} (-1)^{\mathrm{ht}(\lambda/(n-k))} s_{\lambda}. \tag{3}$$

We are now ready to prove the theorem.

Theorem 3. *Let $\lambda \vdash n$. We have $\ell(\lambda) = \sum_{\mu \vdash n} c_\mu \chi^\mu(\lambda)$, where*

$$c_\mu = \begin{cases} \displaystyle\sum_{k=1}^{n} \frac{1}{k} & \text{if } \mu = (n), \\ \displaystyle(-1)^{n-p-q} \frac{p-q+1}{(n-q+1)(n-p)} & \text{if } \mu = (p, q, 1^{n-p-q}), \\ 0 & \text{otherwise.} \end{cases}$$

Proof. Living in R^n, ℓ can be written uniquely as a linear combination of the $\{\chi^\mu\}_{\mu \vdash n}$. Hence, $\ell = \sum c_\mu \chi^\mu$ for some coefficients c_μ. Passing to Λ^n yields

$$\sum_{\mu \vdash n} c_\mu s_\mu = \sum_{\mu \vdash n} \frac{p_\mu}{z_\mu} \ell(\mu).$$

Using (2), we get

$$\sum_{\mu \vdash n} c_\mu s_\mu = \sum_{k=1}^{n} \frac{1}{k} s_{(n-k)}\, p_{(k)},$$

which, with the aid of (3), turns into

$$\sum_{\mu \vdash n} c_\mu s_\mu = \sum_{k=1}^{n} \sum_{\mu} \frac{1}{k} (-1)^{\mathrm{ht}(\mu/(n-k))} s_\mu,$$

so that

$$c_\mu = \sum \frac{1}{k} (-1)^{\mathrm{ht}(\mu/(n-k))},$$

where the sum now is over all k such that $\mu/(n-k)$ is a border strip. This immediately shows that $c_\mu = 0$ unless $\mu = (n)$ or $\mu = (p, q, 1^{n-p-q})$ for some $p \geqslant q \geqslant 1$. Now, $(n)/(n-k)$ is always a border strip of height zero, so $c_{(n)} = \sum_{k=1}^{n} 1/k$. Finally, $(p, q, 1^{n-p-q})/(n-k)$ is a border strip if and only if $q = n-k+1$ or $p = n-k$. Thus,

$$c_{(p,q,1^{n-p-q})} = (-1)^{n-p-q+1}\frac{1}{n-q+1} + (-1)^{n-p-q}\frac{1}{n-p}$$
$$= (-1)^{n-p-q}\frac{p-q+1}{(n-q+1)(n-p)}. \qquad \square$$

5.1. Proof of Theorem 2

Now we show that Theorem 2 is a consequence of Theorem 3. Note that $\mu \mapsto w_\mu$ is a class function and that for fixed $\lambda \vdash n$ we have

$$\sum_{\mu\vdash n} \frac{\chi^\lambda(\mu) w_\mu}{z_\mu} = \langle \chi^\lambda, w_\bullet \rangle.$$

Hence, $\sum_{\mu\vdash n} \chi^\lambda(\mu) w_\mu / z_\mu$ is the coefficient of χ^λ when the class function $w_\bullet$ is written as a linear combination of the irreducible S_n-characters. Now, $w_\mu = n - \ell(\mu)$. Hence, with c_μ as in Theorem 3, the coefficient of the trivial character $\chi^{(n)}$ is $n - c_{(n)}$, whereas the coefficient of χ^μ, $\mu \neq (n)$, is $-c_\mu$. This concludes the proof of Theorem 2.

5.2. Computing the variance

The methods used above apply not only to computing the expected transposition distance given n and t, but also to computing the variance. The formulae in this case are messier and we confine ourselves to briefly sketching the computations.

Since variance and expectation are related according to $V(X) = E(X^2) - E(X)^2$, what we need to compute is the expected value of the square of the transposition distance. Applying our Markov chain machinery, this amounts to computing the coefficients when the class function $\mu \mapsto (n - \ell(\mu))^2 = n^2 - 2n\ell(\mu) + \ell(\mu)^2$ is written as a linear combination of the irreducible S_n-characters. Passing to the space of symmetric functions, what we need to compute is the coefficients d_μ in the expansion $\sum_{\mu\vdash n} \frac{p_\mu}{z_\mu}\ell(\mu)^2 = \sum_{\mu\vdash n} d_\mu s_\mu$.

Again, we need two equations. The first is obtained in the same way as (2) except that we differentiate twice instead of once with respect to t

$$\sum_{\lambda\vdash n} \frac{p_\lambda}{z_\lambda}\ell(\lambda)\big(\ell(\lambda)-1\big) = \sum_{j=1}^{n-1}\sum_{k=1}^{n-j} \frac{1}{jk} s_{(n-j-k)} p_{(j)} p_{(k)}. \tag{4}$$

The other equation we need is a special case of [10, Theorem 7.17.3]. The first sum is over all $\lambda \vdash n$ such that $\lambda/(n-j-k)$ is a border strip, and the second sum is over all *border strip tableaux* of shape $\lambda/(n-j-k)$ and *type* $(\max(j,k), \min(j,k))$ (see [10] for definitions)

$$s_{(n-j-k)} p_{(j)} p_{(k)} = \sum_\lambda \sum_T (-1)^{\mathrm{ht}(T)} s_\lambda. \tag{5}$$

Combining (4) and (5), we obtain

$$d_\mu = c_\mu + \sum_{j=1}^{n-1} \sum_{k=1}^{n-j} \frac{1}{jk} \sum_T (-1)^{\mathrm{ht}(T)},$$

where the third sum is over all border strip tableaux of shape $\mu/(n-j-k)$ and type $(\max(j,k), \min(j,k))$. In particular, $d_\mu = 0$ unless $\mu = (n)$ or $\mu = (p, q, 1^{n-p-q})$ for some $p \geqslant q$.

6. Experimental results

We have deduced a closed formula for the expected transposition distance after t transpositions. We shall now use it as an approximation for the expected reversal distance after t reversals. By taking the inverse, we obtain an estimate for the expected number of reversals applied in creating a genome with some given reversal distance.

6.1. Predicting the true reversal distance

By computing the inverse of $\mathbb{E}_{\mathrm{trp}}$ numerically, we may use it in experiments. We have performed 10000 simulations of evolutionary processes, in which genomes of length 400 have had between 200 and 600 reversals applied to them. We have then used three methods to estimate this evolutionary distance from the resulting genome:

Expected transposition distance: This is the method presented in this paper.
Expected reversal distance given breakpoints: The is the method presented in Eriksen [4]. It is fairly accurate, but considers breakpoints only.
Reversal distance: The by now classical method of Hannenhalli and Pevzner. This is exact, but really measures something different from what we want to measure.

Figure 4 shows that the estimated evolutionary distance depends approximately linearly on the true evolutionary distance if we use any of the first two estimation methods, but not the third. We also see that we should probably not use any of these methods for more distant genomes than those in our experiments, since the results are getting unreliable at the right end of the diagram. This is only natural, since the distribution of genomes after t reversals will approach the uniform distribution as t grows.

Turning to Table 1, we have gathered the mean absolute error and standard deviation obtained using these three methods. As expected, the reversal distance estimates the true evolutionary distance quite poorly. The other two methods are better and quite on a par with each other. Looking at the absolute error, the expected transposition distance seems a better choice than the expected reversal distance given breakpoints. Looking at standard deviations, the situation is the opposite. Also note that their arithmetical mean is a slight improvement over both these estimates taken separately. It is an interesting question whether a more sophisticated use of these two methods will give further improvements.

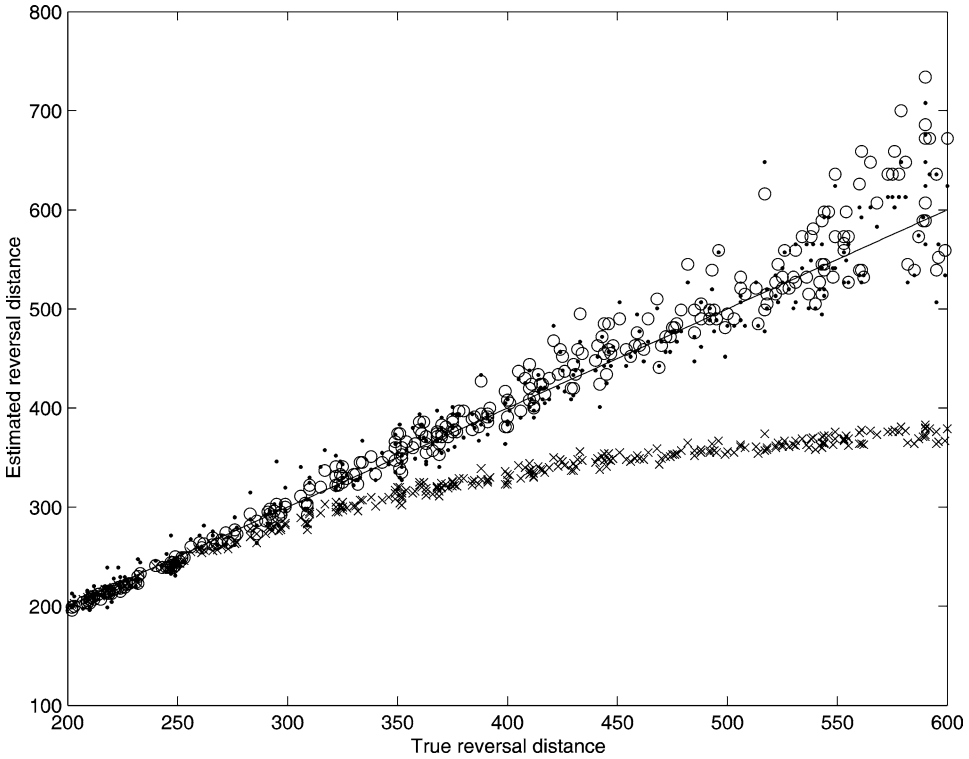

Fig. 4. Results from using three different methods of obtaining the evolutionary reversal distance in 300 of our simulations. The circles come from $\mathbb{E}_{\text{trp}}$, the dots from $\mathbb{E}_{\text{rev}}$ and the crosses from the reversal distance. The two former methods keep their linearity throughout this range, whereas the reversal distance estimate is far from linear.

Table 1
The mean absolute error and the standard deviation obtained from using four methods to estimate the evolutionary reversal distance in simulations

	$\mathbb{E}_{\text{trp}}$	$\mathbb{E}_{\text{rev}}$	$\frac{\mathbb{E}_{\text{trp}}+\mathbb{E}_{\text{rev}}}{2}$	d_{rev}
Mean abs	16.2	18.0	15.8	83.5
St. d.	25.8	24.2	23.0	108.4

The genomes had length 400, the evolutionary reversal distances were between 200 and 600 and the number of simulations was 10000.

7. Conclusions

In computational biology, one has to find the fine balance between models that are relevant and models that facilitate computation. For gene order rearrangements, the "reversals only" model has met both criteria as far as regarding minimal distances, but computations have proved harder for expected distances. With this in mind, it seems natural

to look for models with similar behaviour to the reversals model, but with properties better suited for computation.

One such model is ordinary permutations with transpositions. The symmetric group has been well studied over the years and its computational accessibility is undisputed. The interesting question is whether it is suited as a model in the biological context.

We have in this paper seen that as far as expected distances go, we get results that compare well to the best results obtained through other methods. This should encourage us to look for further areas where we could benefit from this model.

One related problem is the *reversal median problem*: compute the genome G (the median) such that the sum of the reversal distances from G to three given genomes is minimised. This problem is NP-hard, but attempts to use the reversal median for phylogenetic tree construction have still met with some success [2,9]. The use of transpositions in S_n is new to this area and we hope that it can be useful in the future, for instance by studying the transposition median.

We now turn to some computational issues. The double sum of Corollary 4.3 involves binomial coefficients with quite large parameters. Such calculations take some time and it would be useful to be able to discard some terms of minor importance. Are there any?

Using $n = 50$ and $t = 10$ as an example, we get $\mathbb{E}_{\mathrm{trp}}(n, t) = 9.91$. Still, most of the terms in our sum have an absolute value greater than one million (see Fig. 5)! There does not seem to be any terms of minor importance. If we are to exclude any terms, we need to know that the sum of these terms is small.

It turns out that if we sum over all q for fixed p, we do get small values for $p > n/2$ (see Fig. 6). For smaller p these sums have large absolute values. Summing over the last ten or twenty values of p seems to give a reasonable approximation of $\mathbb{E}_{\mathrm{trp}}(n, t)$. This reduces the computation quite a bit, depending on the size of the genomes.

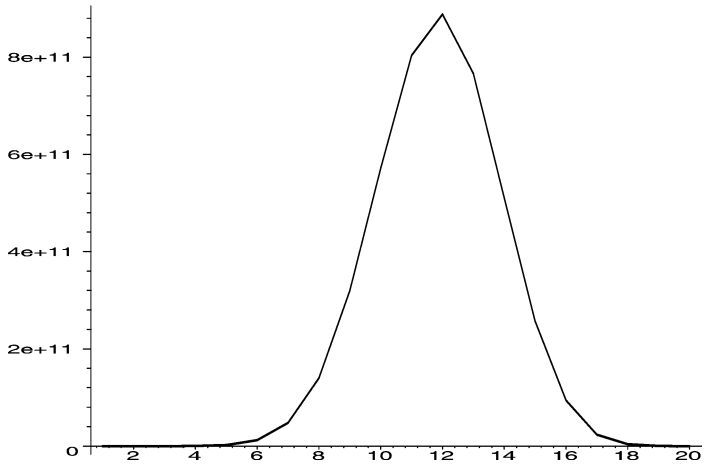

Fig. 5. Absolute values of the terms for different values of q (at the abscissa) for $n = 50$, $t = 10$, and $p = 30$. The terms are alternating and their absolute values form a bell-like shape. This appearance is typical for any $p \geqslant n/2$.

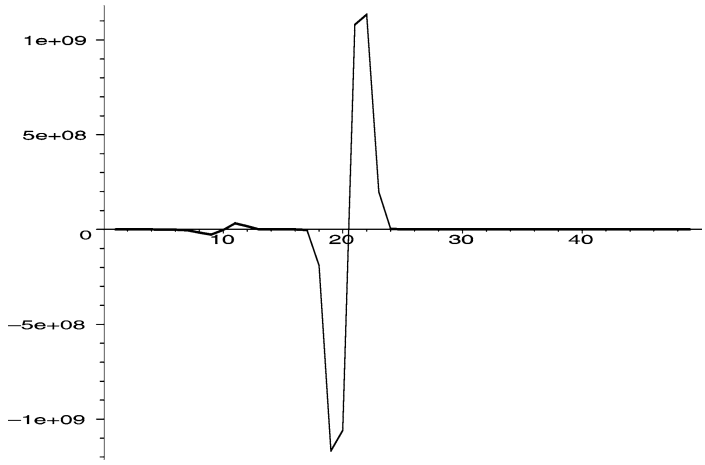

Fig. 6. Sums over q for different values of p (at the abscissa) for $n = 50$ and $t = 10$. For p slightly smaller than $n/2$, the terms are very large compared to the total sum.

Acknowledgments

The authors are most grateful to Richard Stanley for pointing out a way to prove Theorem 3. We also thank Kimmo Eriksson for his careful reading of this article.

References

[1] V. Bafna, P. Pevzner, Genome rearrangements and sorting by reversals, SIAM J. Comput. 25 (1996) 272–289.
[2] G. Bourque, P. Pevzner, Genome-scale evolution: Reconstructing gene orders in the ancestral species, Genome Res. 12 (2002) 26–36.
[3] A. Caprara, On the tightness of the alternating-cycle lower bound for sorting by reversals, J. Combin. Optim. 3 (1999) 149–182.
[4] N. Eriksen, Approximating the expected number of inversions given the number of breakpoints, in: Algorithms in Bioinformatics, Proceedings of WABI 2002, in: Lecture Notes in Comput. Sci., vol. 2452, 2002, pp. 316–330.
[5] N. Eriksen, A. Hultman, Expected reflection distance in $G(r, 1, n)$ after a fixed number of reflections, Preprint, 2002.
[6] S. Hannenhalli, P. Pevzner, Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations with reversals), J. ACM 46 (1999) 1–27.
[7] N. Ito, The spectrum of a conjugacy class graph of a finite group, Math. J. Okayama Univ. 26 (1984) 1–10.
[8] I.G. Macdonald, Symmetric Functions and Hall Polynomials, 2nd ed., Oxford Univ. Press, Oxford, 1995.
[9] B.M.E. Moret, A.C. Siepel, J. Tang, T. Liu, Inversion medians out-perform breakpoint medians in phylogeny reconstruction from gene-order data, in: Algorithms in Bioinformatics, Proceedings of WABI 2002, in: Lecture Notes in Comput. Sci., vol. 2452, 2002, pp. 521–536.
[10] R.P. Stanley, Enumerative Combinatorics, vol. 2, Cambridge Univ. Press, New York, 1999.
[11] L.-S. Wang, T. Warnow, Estimating true evolutionary distances between genomes, in: Proceedings of the ACM Symposium on the Theory of Computing (STOC 01), 2001, pp. 637–646.