JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 21, Number 8, 2014

Pp. 622–631
DOI: 10.1089/cmb.2013.0066

Random Matrix Approach to the Distribution of Genomic Distance

NIKITA ALEXEEV[1] and PETER ZOGRAF[1,2]

ABSTRACT

The cycle graph introduced by Bafna and Pevzner is an important tool for evaluating the distance between two genomes, that is, the minimal number of rearrangements needed to transform one genome into another. We interpret this distance in topological terms and relate it to the random matrix theory. Namely, the number of genomes at a given 2-break distance from a fixed one (the Hultman number) is represented by a coefficient in the genus expansion of a matrix integral over the space of complex matrices with the Gaussian measure. We study generating functions for the Hultman numbers and prove that the two-break distance distribution is asymptotically normal.

Key words: combinatorics, graph theory, probability.

1. INTRODUCTION

In the Bafna and Pevzner (1998) paper on genome comparison and genome rearrangements, the authors raised the problem of decomposing a permutation into the minimal number of ''transpositions'' (here a transposition is understood as an exchange of two contiguous intervals of the permutation). An important tool they introduced to deal with this problem is the *cycle graph* of a permutation. We recall that the cycle graph of a permutation $\pi \in S_n$, denoted by $G(\pi)$, is the directed edge-colored graph with vertices $\{0, 1, \ldots, n\}$ and edges of two colors: gray edges going from $i - 1$ to i and black edges going from π_i to $\{0, 1, \ldots, n\}$ (throughout this note we assume that $\pi_0 = 0$ and consider i modulo $n + 1$). An *alternating cycle* in $G(\pi)$ is a directed cycle with edges of alternate colors. Notice that at every vertex of $G(\pi)$ there is one incoming edge and one outgoing edge of each color. This means that there is a unique disjoint decomposition of the edge set of $G(\pi)$ into alternating cycles (see Fig. 1).

In a bit more detail, suppose we have two circular genomes A and B built from the same set of genes. We may assume that the genes in A are numbered $\{0, 1, , n\}$, and the order of genes in B is $\{0, \pi_1, , \pi_n\}$, where $\pi \in S_n$. For each pair of genomes A and B one can associate the *break-point graph*—a graph on the set of $2n + 2$ vertices $\{0, 0', 1, 1', \ldots, n, n'\}$ with edges of two types (gray and black): gray edges connect the consecutive genes in A (i.e. i' to $i + 1$) and black edges connect the consecutive genes in B (i.e. π'_{i-1} to π_i). Such a graph splits into a disjoint union of cycles (if $A = B$, then the breakpoint graph consists of $n + 1$ cycles of length 2). In terms of the number of cycles in a breakpoint graph, one can evaluate the number of rearrangements necessary to transform one genome into another. Two genomes B and B' are related by an

[1]Chebyshev Laboratory, St. Petersburg State University, St. Petersburg, Russia.
[2]Steklov Mathematical Institute, Russian Academy of Sciences, St. Petersburg, Russia.

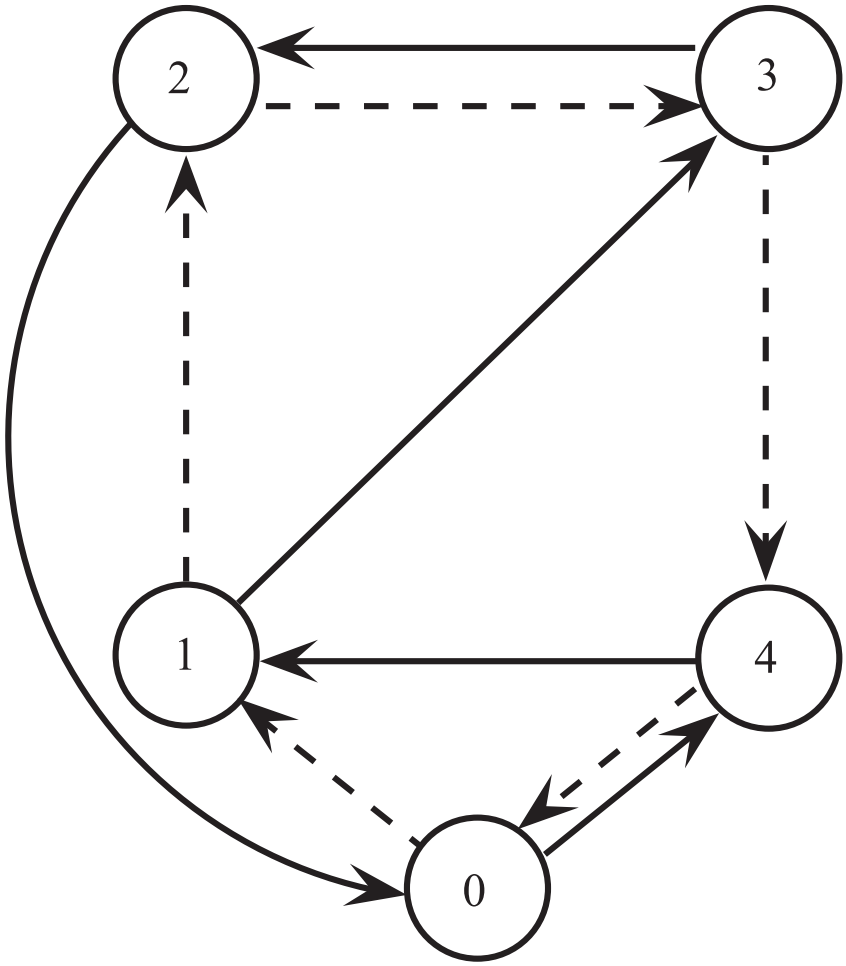

FIG. 1. The cycle graph $G(\pi)$ of the permutation $\pi = \binom{1234}{2314}$, where the gray edges are drawn by dashed arrows and the black edges are drawn by solid arrows. There are three alternating cycles: 0-1-3-4-1-2-0, 2-3-2, and 4-0-4.

m-break if after cutting each of them at m places we get two identical sets of chains of genes. Then the *m-break distance* between A and B is the minimal number of m-breaks needed to transform A to B. The most evolutionary relevant rearrangements of genomes are the 2-breaks (reversals, fusions, fissions, and translocations) and 3-breaks (in particular, transpositions) (cf. Alekseyev, 2008).

Remark 1. We emphasize here that the 2-break distance between two genomes is equal to $n + 1 - k$, where k is the number of cycles in the corresponding breakpoint graph (cf. Bafna and Pevzner, 1998). To see this, we notice that any cycle of length larger than 2 can be split into two cycles by a 2-break; we can continue doing that as long as the number of cycles is less than $n + 1$. At the same time, with a 2-break we cannot increase the number of cycles by more than one. Therefore, the 2-break distance is precisely $n + 1 - k$. On the other hand, the 3-break distance between two genomes cannot exceed $(n + 1 - k)/2$.

Remark 2. The cycle graph and the breakpoint graph are closely related to each other. Indeed, identifying the vertices i and i' in the breakpoint graph and choosing proper orientation of the edges we get the cycle graph.

In Hultman (1999), the author attempted to characterize the number $H(n, k)$ of permutations in S_n whose cycle graph has exactly k alternating cycles. These numbers, now carrying his name have later been studied by several authors (Bóna and Flynn, 2009; Doignon and Labarre, 2007, to name just few). As shown in (Bóna and Flynn, 2009), the Hultman numbers are closely related to the (unsigned) Stirling numbers of the first kind $S(n, k)$ that count permutations in S_n whose disjoint cycle decomposition consists of k cycles:

$$H(n,k)=\begin{cases}\frac{2S(n+2,k)}{(n+1)(n+2)} & \text{if } n-k \text{ is odd,}\\ 0 & \text{otherwise.}\end{cases} \tag{1}$$

A closed formula for $H(n, k)$ was obtained in (Doignon and Labarre, 2007).

In this note, we give two new interpretations of the Hultman numbers in the spirit of Harer and Zagier (1986): as numbers of certain polygon gluings and as integrals over the space of complex matrices. We also give a recursion relation for the Hultman numbers and derive some properties of their generating functions.

2. POLYGON GLUINGS

Consider a $2n$-sided polygon, whose boundary consists of n black sides followed by n gray sides; the black sides are oriented in the counterclockwise direction and the gray sides are oriented in the clockwise direction (see Fig. 2).

Pairwise gluing of black sides with gray sides (respecting orientation) gives an orientable topological surface without boundary of topological genus $g \geq 0$ (the genus g depends on the gluing). At the same time,

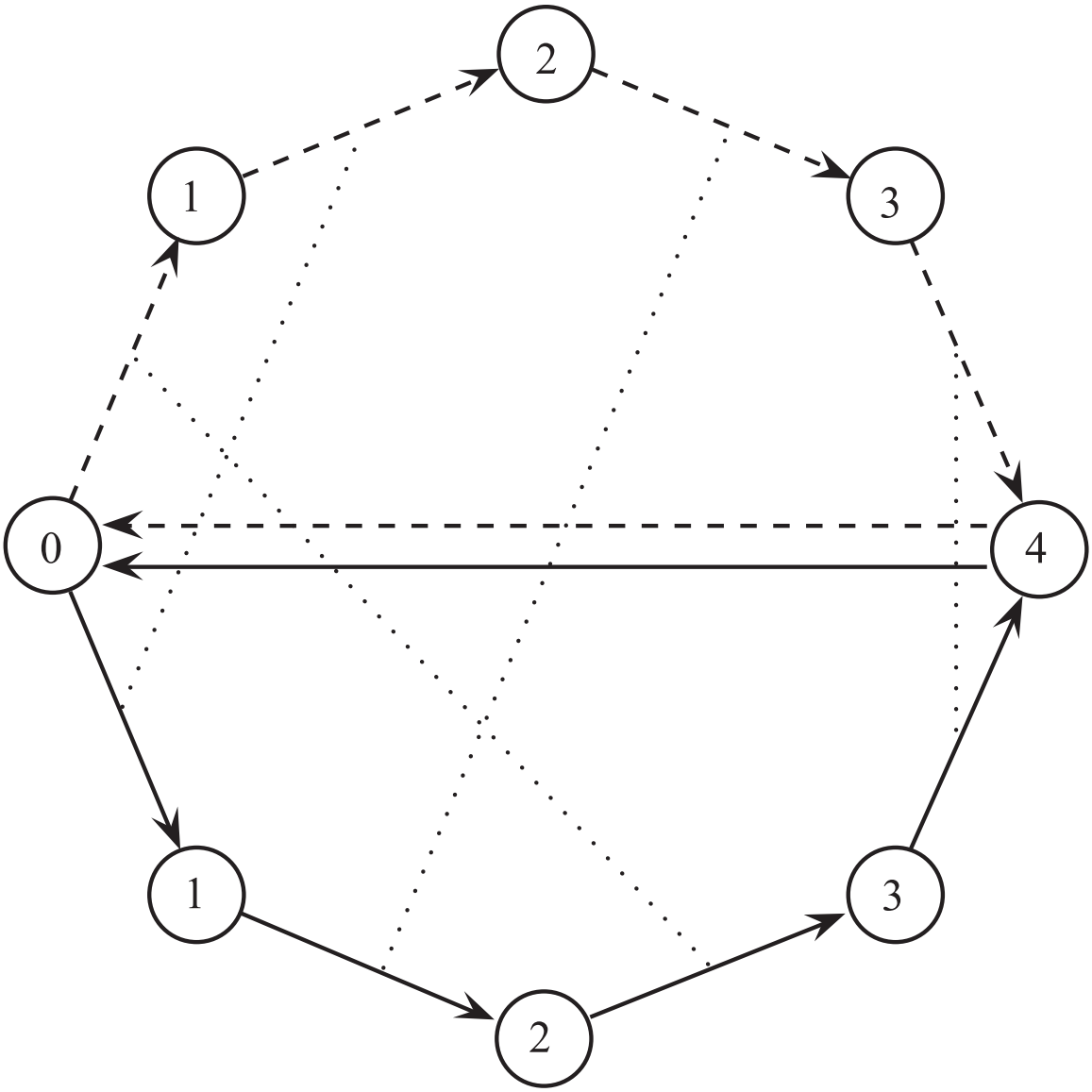

FIG. 2. A $2n$-gon ($n = 4$) with n black sides (solid arrows) and n gray sides (dashed arrows). The pairs of sides that are glued together by $\pi=\binom{1234}{2314}$ are connected with dotted lines.

the boundary of the polygon turns into an oriented graph with $k \geq 1$ vertices and n edges. The numbers g and k are related by the Euler characteristic formula $2 - 2g = k - n + 1$, so that $k = n + 1 - 2g$. We denote by $h_g(n)$ the number of genus g such gluings of a $2n$-gon.

Remark 3. In terms of the polygon gluings, the 2-break distance between genomes related by the permutation π is $2g$, that is, twice the genus of the surface glued according to π, and the 3-break distance is not larger than g.

Theorem 1. *The Hultman numbers $H(n, k)$ and the numbers $h_g(n)$ of genus g gluings of a $2n$-gon described above are related by the following formula:*

$$H(n, n+1-2g) = h_g(n). \tag{2}$$

Proof. We start with a slightly different interpretation of the cycle graph $G(\pi)$. Consider two oriented cycles (that is, two regular oriented graphs) of length $n + 1$, one colored in gray and the other colored in black. The vertex set in both cycles is $\{0, \ldots, n\}$, but in the gray cycle the vertices follow in the clockwise order, and in the black cycle they follow in the counterclockwise order. We identify the vertex π_i of the gray cycle with the vertex i of the black cycle (we assume $\pi_0 = 0$). Obviously, the obtained graph coincides with the cycle graph $G(\pi)$ (see Fig. 1).

We label the black sides of the polygon by numbers from 1 to n in the counterclockwise order, and the gray sides by numbers from 1 to n in the clockwise order, both times starting from the initial vertex 0. Clearly, a gluing of a $2n$-gon of the type considered above is uniquely described by a permutation $\pi \in S_n$, where π_i is the number of the gray side identified with the ith black side. Let us cut the polygon along the diagonal $(n, 0)$, that is, we add one black edge and one gray edge connecting the vertex n to the vertex 0 (see Fig. 2). Now we have two n-gons, one with black boundary and the other with gray boundary, whose sides are pairwise identified by means of the permutation π ($\pi_0 = 0$). These two boundaries glued together give a graph that we denote by $\Gamma(\pi)$. The construction is quite similar to that of the cycle graph $G(\pi)$, but instead of gluing vetices we now glue edges according to the same rule. The graphs $G(\pi)$ and $\Gamma(\pi)$ are closely related to each other: it is straightforward to verify that there is a one-to-one correspondence between the alternating cycles in the cycle graph $G(\pi)$ and the vertices in the polygon gluing graph $\Gamma(\pi)$. To complete the proof, we recall that $k = n + 1 - 2g$, where k is the number of vertices of $\Gamma(\pi)$ and g is the genus of the glued surface. ■

3. MATRIX INTEGRAL

Denote by $M(N) = \mathrm{Mat}_{\mathbb{C}}(N \times N)$ the linear space of complex $N \times N$ matrices; the (complex) dimension of $M(N)$ is N^2. The space $M(N)$ has a natural Gaussian probabilistic measure

$$d\mu_N = \left(\frac{1}{2\pi\sqrt{-1}}\right)^{N^2} e^{-\mathrm{Tr}(XX^*)} \bigwedge_{i,j=1}^{N} dx_{ij} \wedge d\bar{x}_{ij}, \tag{3}$$

where $X = \{x_{ij}\}_{i,j=1}^{N} \in M(N)$, the asterisk denotes the Hermitian conjugation and Tr is the trace. Note that the space $M(N)$ equipped with the measure μ_N is also called the complex Ginibre ensemble.

Theorem 2. *Put*

$$p_n(N) = \sum_{g=0}^{[n/2]} H(n, n+1-2g)\ , N^{n-2g+1}, \tag{4}$$

where $H(n, k)$ are the Hultman numbers. Then

$$p_n(N) = \int_{M(N)} \mathrm{Tr}(X^n X^{*n})\ , d\mu_N. \tag{5}$$

Remark 4. More general matrix integrals over the space $M(N)$ are considered in Alexeev et al. (2010).

Remark 5. Below is a list of the first several polynomials $p_n(N)$:

$$\begin{aligned}
p_0(N) &= N,\\
p_1(N) &= N^2,\\
p_2(N) &= N^3 + N,\\
p_3(N) &= N^4 + 5N^2,\\
p_4(N) &= N^5 + 15N^3 + 8N,\\
p_5(N) &= N^6 + 35N^4 + 84N^2,\\
p_6(N) &= N^7 + 70N^5 + 469N^3 + 180N,\\
p_7(N) &= N^8 + 126N^6 + 1869N^4 + 3044N^2,\\
p_8(N) &= N^9 + 210N^7 + 5985N^5 + 26060N^3 + 8064N,\\
p_9(N) &= N^{10} + 330N^8 + 16401N^6 + 152900N^4 + 193248N^2.
\end{aligned}$$

Proof. It is a fairly standard exercise in t'Hooft graphic calculus to reduce the matrix integral in Equation (5) to a sum over Feynman diagrams (polygon gluings) (cf., e.g., Mulase, 1998; Zvonkin, 1997). We will briefly explain how it works. By definition, we have

$$\mathrm{Tr}(X^n X^{*n}) = \sum_{i_1=1}^{N} \dots \sum_{i_{2n}=1}^{N} x_{i_1 i_2} \dots x_{i_n i_{n+1}} \bar{x}_{i_1 i_{2n}} \dots \bar{x}_{i_{n+2} i_{n+1}},$$

and a simple computation shows that

$$\begin{aligned}
&\int_{M(N)} x_{ij} \bar{x}_{kl} d\mu_N = \delta_{ik}\delta_{jl},\\
&\int_{M(N)} x_{ij} x_{kl} d\mu_N = \int_{M(N)} \bar{x}_{ij} \bar{x}_{kl} d\mu_N = 0.
\end{aligned}$$

Applying Wick's formula (cf. Mulase, 1998; Zvonkin, 1997), we get

$$\begin{aligned}
&\int_{M(N)} x_{i_1 i_2} \dots x_{i_n i_{n+1}} \bar{x}_{i_1 i_{2n}} \dots \bar{x}_{i_{n+2} i_{n+1}} d\mu_N\\
&\quad = \sum_{\pi \in S_n} \int_{M(N)} x_{i_1 i_2} \bar{x}_{i_{\alpha_1 + 1} i_{\alpha_1}} d\mu_N \times \dots\\
&\qquad \dots \times \int_{M(N)} x_{i_n i_{n+1}} \bar{x}_{i_{\alpha_n + 1} i_{\alpha_n}} d\mu_N\\
&\quad = \sum_{\pi \in S_n} \delta_{i_1 i_{\alpha_1 + 1}} \delta_{i_2 i_{\alpha_1}} \cdots \delta_{i_n i_{\alpha_n + 1}} \delta_{i_{n+1} i_{\alpha_n}},
\end{aligned}$$

where $\alpha_j = 2n + 1 - \pi_j$ (we assume that $i_{2n+1} = i_1$). Therefore,

$$\int_{M(N)} \mathrm{Tr}(X^n X^{*n}) \, , d\mu_N = \sum_{\pi \in S_n} \sum_{i_1=1}^{N} \dots \sum_{i_{2n}=1}^{N} \delta_{i_1 i_{\alpha_1 + 1}} \delta_{i_2 i_{\alpha_1}} \cdots \delta_{i_n i_{\alpha_n + 1}} \delta_{i_{n+1} i_{\alpha_n}}.$$

We note that the pairs of indices $\{i_k i_{k+1}\}$ correspond to the black edges of the polygon on Figure 2, and the pairs of indices $\{i_{\alpha_k + 1} i_{\alpha_k}\}$ correspond to the gray edges, so there is a one-to-one correspondence between the pairings of indices and polygon gluings. Moreover, it is not hard to see that for a given $\pi \in S_N$

$$\sum_{i_1=1}^{N} \dots \sum_{i_{2n}=1}^{N} \delta_{i_1 i_{\alpha_1 + 1}} \delta_{i_2 i_{\alpha_1}} \cdots \delta_{i_n i_{\alpha_n + 1}} \delta_{i_{n+1} i_{\alpha_n}} = N^{n - 2g + 1},$$

where g denotes the genus of the surface glued from the $2n$-gon by means of π. This yields

$$\int_{M(N)} \mathrm{Tr}(X^n X^{*n})\ , d\mu_N = \sum_{g=0}^{[n/2]} h_g(n) N^{n-2g+1},$$

and Equation (5) now follows from Theorem 1. ■

4. GENERATING FUNCTIONS AND RECURSIONS

Here we collect some simple facts about the recursive relations and generating functions for the Hultman numbers that we did not find in the literature.

Consider the generating functions

$$F(x, N) = \sum_{g=0}^{\infty} \sum_{n=2g}^{\infty} H(n, n+1-2g) N^{n-2g+1} \frac{x^n}{n!} \tag{6}$$

and

$$H_g(x) = \sum_{n=2g}^{\infty} H(n, n+1-2g) x^n. \tag{7}$$

Theorem 3. *We have*

(i)

$$F(x, N) = \frac{1}{x^2}\left(\frac{1}{(1-x)^N} - (1+x)^N\right);$$

(ii) $H(n, n+1-2g) = h_g(n)$ *satisfy the recursion*

$$(n+2)h_g(n) = (2n+1)h_g(n-1) - (n-1)h_g(n-2) + n^2(n-1)h_{g-1}(n-2);$$

(iii) the polynomials $p_n(N)$ *defined by Equation (4) satisfy the recursion*

$$(n+2)p_n(N) = (2n+1)Np_{n-1}(N) + (n-1)(n^2 - N^2)p_{n-2}(N)$$

with $p_0 = N$, $p_1 = N^2$;

(iv)

$$H_0(x) = \frac{1}{1-x}\ , \qquad H_g(x) = \frac{P_g(x)}{(1-x)^{1+4g}}\ , \quad g \geq 1,$$

where $P_g(x) = \sum_{i=2g}^{4g-2} a_{g,i} x^i$ *is a polynomial with integer coefficients,* $a_{g,2g} = \frac{(2g)!}{g+1}$, $a_{g,4g-2} = 1$, *and* $P_g(1) = \frac{(4g-1)!!}{2g+1}$.

Remark 6. Several first polynomials $P_g(x)$ are listed below:

$$\begin{aligned}
P_0(x) &= 1,\\
P_1(x) &= x^2,\\
P_2(x) &= x^4(8 + 12x + x^2),\\
P_3(x) &= x^6(180 + 704x + 528x^2 + 72x^3 + x^4),\\
P_4(x) &= x^8(8064 + 56160x + 98124x^2 + 53792x^3 + 8760x^4 + 324x^5 + x^6),\\
P_5(x) &= x^{10}(604800 + 6356160x + 19083456x^2\\
&\quad + 21676144x^3 + 9936360x^4 + 1759520x^5\\
&\quad + 103040x^6 + 1344x^7 + x^8).
\end{aligned}$$

Remarkably, all polynomials $P_g(x)$ have positive integer coefficients. Moreover, the integers $P_g(1)$ are well known—they enumerate genus g orientable gluings of a $4g$-gon (cf. Harer and Zagier, 1986), or the permutations in S_{4g-1} whose cycle graph alternating cycles are all of length 4 (cf. Doignon and Labarre, 2007).

Proof. Part (*i*) follows from Equation (1) and the fact that

$$(1+x)^N = \sum_{n=0}^{\infty} \sum_{k=0}^{n} (-1)^{n+k} S(n,k) N^k \frac{x^n}{n!},$$

where $S(n, k)$ are the Stirling numbers of the first kind. Similarly, the recursion $S(n + 1, k) = S(n, k - 1) + nS(n, k)$ for the Stirling numbers immediately implies (*ii*). Part (*iii*) is a direct consequence of (*ii*). The proof of (*iv*) is by induction on g and follows the proof of Theorem 1 in Andersen et al. (2013). The cases $g = 0, 1$ being easy, assume that the statements of part (*iv*) of the theorem hold for $g - 1$, $g \geq 2$. Put $\tilde{H}_g(x) = x^2 H_g(x)$, then the recursion (*ii*) is equivalent to the ordinary differential equation (ODE)

$$(1-x)^2 \tilde{H}_g'(x) + (1-x)\tilde{H}_g(x) = x^4 \tilde{H}_{g-1}'''(x) + 2x^3 \tilde{H}_{g-1}''(x)$$

with initial condition $\tilde{H}_g(0) = 0$. Therefore, we have

$$\tilde{H}_g(x) = (1-x) \int_0^x \frac{t^4 \tilde{H}_{g-1}'''(t) + 2t^3 \tilde{H}_{g-1}''(t)}{(1-t)^3} dt. \tag{8}$$

The elementary formula

$$\left(\frac{x^\alpha}{(1-x)^\beta}\right)' = \frac{\alpha x^{\alpha-1} + (\beta-\alpha)x^\alpha}{(1-x)^{\beta+1}} \tag{9}$$

immediately yields

$$x^4 \left(\frac{x^\alpha}{(1-x)^\beta}\right)''' + 2x^3 \left(\frac{x^\alpha}{(1-x)^\beta}\right)'' = \frac{\alpha^2(\alpha-1)x^{\alpha+1} + \ldots + (\beta-\alpha)^2(\beta-\alpha+1)x^{\alpha+4}}{(1-x)^{\beta+3}}. \tag{10}$$

Since, by assumption,

$$\tilde{H}_{g-1}(x) = \frac{x^2 P_{g-1}(x)}{(1-x)^{4g-3}} = \frac{\sum_{i=2g-2}^{4g-6} a_{g-1,i} x^{i+2}}{(1-x)^{4g-3}},$$

applying Equation (10) we get that

$$\frac{x^4 \tilde{H}_{g-1}'''(x) + 2x^3 \tilde{H}_{g-1}''(x)}{(1-x)^3} = \frac{Q_g(x)}{(1-x)^{4g+3}}, \tag{11}$$

where $Q_g(x) = \sum_{i=2g+1}^{4g} q_{g,i} x^i$ is a polynomial with integer coefficients,

$$q_{g,2g+1} = (2g)^2(2g-1)a_{g-1,2g-2} = 2(2g)!,$$
$$q_{g,4g} = 2a_{g-1,4g-6} = 2.$$

Consider the Laurent expansion

$$\frac{Q_g(x)}{(1-x)^{4g+3}} = \sum_{i=3}^{4g+3} \frac{r_{g,i}}{(1-x)^i}, \tag{12}$$

then we have

$$\frac{\tilde{H}_g(x)}{1-x} = \sum_{i=2}^{4g+2} \frac{r_{g,i+1}}{i(1-x)^i} + C,$$

where the initial condition $\tilde{H}_g(0)=0$ implies that

$$C=-\sum_{i=2}^{4g+2}\frac{r_{g,\,i+1}}{i}.$$

Now put

$$\tilde{P}_g(z)=\sum_{i=2}^{4g+2}\frac{r_{g,\,i+1}}{i}\left((1-x)^{4g+2-i}-(1-x)^{4g+2}\right)=\sum_{i=0}^{4g+2}p_{g,\,i}x^i. \tag{13}$$

By construction, we have $p_{g,0}=0$, therefore $\tilde{H}_g(x)=\tilde{P}_g(x)/(1-x)^{4g+1}$ since they both satisfy the same first order ordinary differential equation with the same initial condition. Moreover, since $h_g(1)=\ldots=h_g(2g-1)=0$, we also have $p_{g,\,1}=\ldots=p_{g,\,2g+1}=0$. Inverting (9), we see that

$$\begin{aligned}a_{g,\,2g}&=p_{g,\,2g+2}=q_{g,\,2g+1}/(2g+2)=(2g)!/(g+1),\\ a_{g,\,4g-2}&=p_{g,\,4g}=q_{g,\,4g}/2=1\end{aligned}$$

as claimed. Clearly, $P_g(x)=\tilde{P}_g(x)/x^2=(1-x)^{4g+1}H_g(x)$ must have integral coefficients because $H_g(x)$ does.

To complete the proof it is sufficient to show that

$$P_g(1)=\frac{(4g-1)(4g-3)(2g-1)}{2g+1}\;P_{g-1}(1)$$

[note that $P_0(1)=P_1(1)=1$]. We have

$$\begin{aligned}\tilde{H}'_{g-1}(x)&=\frac{(1-x)\tilde{P}'_{g-1}(x)+(4g-3)\tilde{P}_{g-1}(x)}{(1-x)^{4g-2}}=\frac{P_{g,\,1}(x)}{(1-x)^{4g-2}},\\ \tilde{H}''_{g-1}(x)&=\frac{(1-x)P'_{g,\,1}(x)+(4g-2)P_{g,\,1}(x)}{(1-x)^{4g-1}}=\frac{P_{g,\,2}(x)}{(1-x)^{4g-1}},\\ \tilde{H}'''_{g-1}(x)&=\frac{(1-x)P'_{g,\,2}(x)+(4g-1)P_{g,\,2}(x)}{(1-x)^{4g}},\end{aligned}$$

and from Equation (11) it then follows that

$$Q_g(x)=(1-x)\left(x^4P'_{g,\,2}(x)+2x^3P_{g,\,2}(x)\right)+(4g-1)x^4P_{g,\,2}(x).$$

From here we easily get

$$\begin{aligned}P_{g,\,1}(1)&=(4g-3)P_{g-1}(1),\\ P_{g,\,2}(1)&=(4g-2)P_{g,\,1}(1)=(4g-2)(4g-3)P_{g-1}(1),\\ Q_{g,\,1}(1)&=(4g-1)P_{g,\,2}(1)=(4g-1)(4g-2)(4g-3)P_{g-1}(1).\end{aligned}$$

Clearly, $Q_{g,1}(1)=r_{g,4g+3}$ in the Laurent expansion (12), and from Equation (13) we obtain $P_g(1)=\frac{1}{4g+2}\;Q_{g,\,1}(1)=\frac{(4g-1)(4g-2)(4g-3)}{4g+2}P_{g-1}(1)$ as claimed. ■

5. ASYMPTOTIC DISTRIBUTION OF GENOMIC DISTANCE

Consider the symmetric group S_n equipped with the uniform measure. Then the number of alternating cycles in the cycle graph of a random permutation is a random variable that we denote by K_n. Here we study the asymptotic distribution of the random variable K_n as $n\to\infty$.

Theorem 4. *The number K_n of alternating cycles in the cycle graph of a random permutation of length n has the expectation and the variance of order* ln *n. The variable $\frac{K_n-\ln n}{\sqrt{\ln n}}$ weakly converges to the standard Gaussian random variable.*

Proof. The probability $P\{K_n=k\}$ is equal to $\frac{H(n,k)}{n!}$. Therefore, by Equation (6) and Theorem 3, (i) the coefficient of $F(x, N)$ at x^n is the expectation of N^{K_n}, and

$$\mathbb{E}N^{K_n} = \binom{N+n+1}{n+2} - \binom{N}{n+2}.$$

Clearly, we have

$$\mathbb{E}K_n = \frac{\partial(\mathbb{E}N^{K_n})}{\partial N}\bigg|_{N=1},$$
$$\mathbb{E}K_n(K_n-1) = \frac{\partial^2(\mathbb{E}N^{K_n})}{\partial N^2}\bigg|_{N=1}.$$

A standard computation shows that

$$\frac{\partial\mathbb{E}(N^{K_n})}{\partial N} = \sum_{j=1}^{n+2}\frac{1}{j}\prod_{l\neq j}\frac{N-1+l}{l} - \sum_{j=1}^{n+2}\frac{1}{j}\prod_{l\neq j}\frac{N+1-l}{l},$$
$$\frac{\partial^2\mathbb{E}(N^{K_n})}{\partial N^2} = \sum_{i=1}^{n+2}\sum_{\substack{j=1\\ j\neq i}}^{n+2}\frac{1}{ij}\prod_{\substack{l\neq j\\ l\neq i}}\frac{N-1+l}{l} - \sum_{i=1}^{n+2}\sum_{\substack{j=1\\ j\neq i}}^{n+2}\frac{1}{ij}\prod_{\substack{l\neq j\\ l\neq i}}\frac{N+1-l}{l}.$$

Hence,

$$\mathbb{E}K_n = \sum_{j=1}^{n+2}\frac{1}{j} - \frac{(-1)^n n!}{(n+2)!}$$

and

$$\mathbb{E}K_n(K_n-1) = \left(\sum_{j=1}^{n+2}\frac{1}{j}\right)^2 - \sum_{j=1}^{n+2}\frac{1}{j^2} - \frac{(-1)^n n!}{(n+2)!}\sum_{j=1}^{n}\frac{1}{j}.$$

From here it is easy to see that for the mean value and variance of K_n we have

$$\mathbb{E}K_n = \ln n + \gamma + o(1)$$

and

$$\mathbb{E}(K_n - \mathbb{E}K_n)^2 = \ln n + \gamma - \frac{\pi^2}{6} + o(1)$$

as $n \to \infty$, where γ is the Euler-Mascheroni constant. This proves the first statement of the theorem. To prove the second statement, consider the Laplace transform of the random variable $\frac{K_n - \ln n}{\sqrt{\ln n}}$

$$\begin{aligned}
\mathbb{E}N^{\frac{K_n-\ln(n)}{\sqrt{\ln(n)}}} &= N^{-\sqrt{\ln n}}\left(\binom{N^{1/\sqrt{\ln n}}+n+1}{n+2} - \binom{N^{1/\sqrt{\ln n}}}{n+2}\right)\\
&\sim N^{-\sqrt{\ln n}}\prod_{j=1}^{n+2}\frac{N^{1/\sqrt{\ln n}}-1+j}{j}\\
&\sim N^{-\sqrt{\ln n}}\exp\sum_{j=1}^{n+2}\ln\left(1+\frac{N^{1/\sqrt{\ln n}}-1}{j}\right)\\
&\sim \exp\left(\left(N^{1/\sqrt{\ln n}}-1\right)\sum_{j=1}^{n}\frac{1}{j} - \sqrt{\ln n}\ln N\right)\\
&\sim \exp\left(\ln n\left(\frac{\ln N}{\sqrt{\ln n}} + \frac{1}{2}\frac{\ln^2 N}{\ln n}\right) - \sqrt{\ln n}\ln N\right) \to e^{\ln^2 N/2}.
\end{aligned}$$

The function $e^{\ln^2 N/2}$ is the Laplace transform of the standard Gaussian random variable. ■

Remark 7. In terms of genome rearrangements Theorem 4 claims that the 2-break distance between two genomes randomly built from the same set of n genes has the mean value of order $n - \ln n$ and is asymptotically Gaussian as $n \to \infty$.

ACKNOWLEDGMENTS

We thank P. Pevzner and R. Penner for discussions regarding this study. This work has been supported by the Government of the Russian Federation Megagrant 11.G34.31.0026; JSC ''Gazprom Neft''; the RFBR grants 13-01-12422-OF1-M, 14-01-00373-a, and 13-01-00935-a; the SPbSU grant 6.38.672.2013; and the Centre for Quantum Geometry of Moduli Spaces (QGM) at Aarhus University.

AUTHOR DISCLOSURE STATEMENT

The authors declare that no competing financial interests exist.

REFERENCES

Alekseyev, M. 2008. Multi-break rearrangements and breakpoint re-uses: from circular to linear genomes. *J. Comp. Biol.* 15, 1117–1131.

Alexeev, N., Goetze, F., and Tikhomirov, A. 2010. Asymptotic distribution of singular values of powers of random matrices. *Lithuanian Math. J.* 50, 121–132.

Andersen, J.E., Penner, R.C., Reidys, C.M., and Waterman, M.S. 2013. Topological classification and enumeration of RNA structures by genus. *J. Math. Bio.* 67, 1261–1278.

Bafna, V., and Pevzner, P.A. 1998. Sorting by transpositions. *SIAM Journal on Discrete Mathematics* 11, 224–240.

Bóna, M., and Flynn, R. 2009. The average number of block interchanges needed to sort a permutation and a recent result of Stanley. *Information Processing Letters* 109, 927–931.

Doignon, J.-P., and Labarre, A. 2007. On Hultman numbers. *Journal of Integer Sequences* 10, 1–13.

Harer, J., and Zagier, D. 1986. The Euler characteristic of the moduli space of curves. *Invent. Math.* 85, 457–485.

Hultman, A. 1999. Toric permutations [Master's thesis]. Department of Mathematics, KTH Royal Institute of Technology, Stockholm, Sweden.

Mulase, M. 1998. Lectures on the asymptotic expansion of a Hermitian matrix integral. *Supersymmetry and Integrable Models* 502, 91–134.

Zvonkin, A. 1997. Matrix integrals and map enumeration: an accessible introduction. *Math. Comput. Modeling*, 26, 281–304.

Address correspondence to:
Dr. Peter Zograf
Steklov Mathematical Institute
Russian Academy of Sciences
Fontanka 27
St. Petersburg 191023
Russia

E-mail: zograf@pdmi.ras.ru