Distribution of the genomic distance

Nikita Alexeev

Chebyshev Laboratory
Department of Mathematics and Mechanics
Saint-Petersburg State University

Saint-Petersburg
13 June
Problem

Let genomes of two species have $n$ common genes. What is the evolutionary scenario for transforming one genome into the other?
The 2-break distance $d_2(A, B)$ between genomes A and B is the minimal number of 2-breaks needed to transform A to B.
Figure: Let genomes $A$ and $B$ contain $n$ common genes.
Figure: Draw both genomes $A$ and $B$ on the same graph.
Figure: Erase all “genes” edges. The breakpoint graph $G(A, B)$ consists of several alternating cycles.
Theorem (Bafna and Pevzner ’98)

Let $A$ and $B$ be two genomes with $n$ common genes, and the breakpoint graph $G(A, B)$ consists of $k$ alternating cycles. Then the 2-break distance $d_2(A, B)$ between $A$ and $B$ is equal to

$$d_2(A, B) = n + 1 - k.$$
Figure: The 2-break distance between (1234) and (4123) is 2.
Topological interpretation

The metric $d_2$ on the permutation group $S_n$ is naturally induced by the metric $d_2$ on the set of genomes.

Figure: An octagon is glued according to the permutation $\pi = (4123)$.

Remark

The distance $d_2(\pi, \text{id})$ between some permutation $\pi$ and the identity is twice the genus of the surface glued according to $\pi$. 

Nikita Alexeev
Distribution of the genomic distance
**Definition**

_Hultman numbers_ $H(n, k) = \#\{\pi \in S_n : d_2(\pi, id) = k\}$.  

$$p_n(N) = \sum_{k=1}^{n+1} H(n, k)N^k$$

**Theorem (Alexeev and Zograf ’11)**

Let $X$ be an $N \times N$ random matrix, whose entries $x_{ij}, 1 \leq i, j \leq N$ are independent standard Gaussian complex random variables. Then  

$$\mathbb{E} \text{Tr } X^n X^{n*} = p_n(N).$$
\[ p_0(N) = N, \]
\[ p_1(N) = N^2, \]
\[ p_2(N) = N^3 + N, \]
\[ p_3(N) = N^4 + 5N^2, \]
\[ p_4(N) = N^5 + 15N^3 + 8N, \]
\[ p_5(N) = N^6 + 35N^4 + 84N^2, \]
\[ p_6(N) = N^7 + 70N^5 + 469N^3 + 180N, \]
\[ p_7(N) = N^8 + 126N^6 + 1869N^4 + 3044N^2, \]
\[ p_8(N) = N^9 + 210N^7 + 5985N^5 + 26060N^3 + 8064N, \]
\[ p_9(N) = N^{10} + 330N^8 + 16401N^6 + 152900N^4 + 193248N^2. \]
The Hultman numbers satisfy the recurrence relation

\[(n + 2)H(n, k) = (2n + 1)H(n - 1, k - 1) - (n - 1)H(n - 2, k - 2) + n^2(n - 1)H(n - 2, k);\]

The polynomials \(p_n(N) = \sum_{k=1}^{n+1} H(n, k)N^k\) 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 \(p_0 = N, p_1 = N^2;\)
Theorem

Let $F(x, t)$ be the generating function of the sequence $H(n, k)$

$$F(x, t) = \sum_{n=0}^{\infty} \sum_{k=1}^{n+1} H(n, k) t^k \frac{x^n}{n!}$$  \hspace{1cm} (1)

Then

$$F(x, t) = \frac{1}{x^2} \left( \frac{1}{(1-x)^t} - (1+x)^t \right);$$
Consider the symmetric group $S_n$ equipped with the uniform measure. Then the number of alternating cycles in the breakpoint 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 $n \to \infty$.

The probability $P\{K_n = k\}$ is equal to $\frac{H(n,k)}{n!}$. Therefore,

$$F(x, t) = \sum_{n=0}^{\infty} \sum_{k=1}^{n+1} x^n t^k P\{K_n = k\},$$

and the coefficient of $F(x, t)$ at $x^n$ is the expectation of $t^{K_n}$:

$$\mathbb{E} t^{K_n} = \binom{t + n + 1}{n + 2} - \binom{t}{n + 2}.$$
Theorem (Alexeev and Zograf ’13)

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

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

The variable $\frac{K_n - \ln n}{\sqrt{\ln n}}$ weakly converges to the standard Gaussian random variable.

Remark

*In terms of genome rearrangements this theorem 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 $n \to \infty$.  

Nikita Alexeev

Distribution of the genomic distance
Thank you!