BIOINFORMATICS

Vol. 24 ECCB 2008, pages i146–i152
doi:10.1093/bioinformatics/btn295

Poisson adjacency distributions in genome comparison: multichromosomal, circular, signed and unsigned cases

Wei Xu*, Benoît Alain and David Sankoff

Department of Mathematics and Statistics, University of Ottawa, Ottawa, Ontario, Canada

ABSTRACT

The number of common adjacencies of genetic markers, as a measure of the similarity of two genomes, has been widely used as indicator of evolutionary relatedness and as the basis for inferring phylogenetic relationships. Its probability distribution enables statistical tests in detecting whether significant evolutionary signal remains in the marker order. In this article, we derive the probability distributions of the number of adjacencies for a number of types of genome—signed or unsigned, circular or linear, single-chromosome or multichromosomal. Generating functions are found for single-chromosome cases, from which exact counts can be calculated. Probability approaches are adopted for multichromosomal cases, where we find the exact values for expectations and variances. In both cases, the limiting distributions are derived in term of numbers of adjacencies. For all unsigned cases, the limiting distribution is Poisson with parameter 2; for all signed cases, the limiting distribution is Poisson with parameter ½.

Contact: wxu060@uottawa.ca

1 INTRODUCTION

The linear order of markers, such as genes or other sites, on chromosomes is a characteristic structural feature of the genome shared by all individuals in a species. During evolution, various *rearrangement events* disrupt this order by moving or inverting segments of chromosomes. At the time of the event, the order of the markers within a segment is either conserved or inverted and the order of markers outside the segment is conserved. A *breakpoint* may be created between two markers that have hitherto been adjacent in the order if one is inside the segment while the other remains outside the segment affected.

The number of breakpoints b in the comparison of two genomes is the oldest and simplest metric representing the evolutionary divergence of species through chromosomal rearrangements (Nadeau and Taylor, 1984; Sankoff and Blanchette, 1998).

Chromosomes are generally *circular* in prokaryotes, mitochondria and chloroplasts, and a single chromosome contains the entire genome, although sometimes smaller circles, *plasmids*, contain some of this information. Eukaryotic nuclear genomes are partitioned among *linear* chromosomes, from a few to a few dozen in number.

Available data on chromosomes may or may not identify the DNA strand on which the markers are found (called *signed* or *unsigned* data, respectively).

*To whom correspondence should be addressed.

Let a be the number of pairs of markers adjacent in two genomes with the same markers $1,2,\ldots,n$ and the same number of linear chromosomes χ_l. Then

$$a+b=n-\chi_l.$$

As evolutionary rearrangements continue to disrupt adjacencies, b increases and a decreases. Now, even pairs of randomly constructed genomes may have some adjacencies, and pairs of genomes clearly related at the DNA sequence level may have highly scrambled marker order. Then, for any given pair of genomes the question arises of whether a is significantly larger than the random case. To answer this, i.e. to test a for statistical significance, we need its distribution under the null hypothesis of randomness.

In this article, we study the distribution of the number of common adjacencies under the null hypothesis that the n markers are ordered completely randomly on the genomes (N.B. it suffices to randomize just one of the genomes, since relabeling markers can convert one of the genome to a canonical order, e.g. $1,2,\ldots,n$, without changing a). For multichromosomal genomes, the number of markers on each chromosome is also random. We study the cases of single or multiple chromosomes, which can be linear or circular, and signed or unsigned. The unsigned single-chromosome case is related to the dinner table problem (Robbins, 1980) and non-attacking kings problem (Abramson and Moser, 1966). G. Tesler has previously derived results for linear, single-chromosome genomes (Tesler, 2005) .

In Section 2, we find the generating functions for the case of a single-chromosome genome and approximate the probability distributions for large n. In Section 3, we extend our discussion to multichromosomal genomes and directly derive the exact formulae for the expectation and variance of a as well as an approximation of the probability distribution for large n.

The remainder of this section introduces notation and recalls fundamental theorems used in this article.

1.1 Notation and terminology

In the single-chromosome unsigned case we call the genome R where the markers are ordered from 1 to n the *reference genome*, i.e. $R=1,2,\ldots,n$. The *random genome* $G=g_1,g_2,\ldots,g_n$ is sampled from a uniform distribution on the set of permutations of $1,\ldots,n$. For multichromosomal genomes, the beginning and ending marker of each chromosome in the reference genome are given, while in the random genome only the number of chromosomes is given. In the signed case, all the markers in the reference genome R have positive sign, while in the random genome G, a positive or negative sign is assigned at random to each marker independently.

If for some i we have $|g_i - g_{i+1}| = 1$ in the unsigned case, or $g_{i+1} - g_i = 1$ in the signed case, we say there is an adjacency in common between the two genomes, otherwise there is a breakpoint. In the multichromosomal case, if g_i is the last term on a chromosome and/or g_{i+1} is the first, we identify neither an adjacency nor a breakpoint.

1.2 Generating functions

The ordinary generating function (OGF) is defined as the formal power series:

$$F(z) = \sum_{n=0}^{\infty} A_n z^n, \tag{1}$$

where $[z^n]$ denotes A_n the coefficient of z^n in F.

If $A_{n,m}$ denotes the number of (random) genomes with n markers and m adjacencies in common with the reference genome, consider the bivariate generating function

$$F(z,u) = \sum_{n,m} A_{n,m} z^n u^m. \tag{2}$$

For a given n, if X is the random variable counting the number of adjacencies in a random genome and $A_n = \sum_m A_{n,m}$, then the probability

$$\mathbf{P}_n(X=k) = \frac{A_{n,k}}{A_n} = \frac{[z^n u^k]F(z,u)}{[z^n]F(z,1)}. \tag{3}$$

Let $X_{(k)} = X(X-1)\ldots(X-k+1)$. Then the kth factorial moment is

$$\mathbf{E}[X_{(k)}] = \mathbf{E}[X(X-1)\ldots(X-k+1)] \tag{4}$$

$$= \frac{[z^k]\left(\partial_u^k F(z,u)\right)|_{u=1}}{[z^k]F(z,1)}, \tag{5}$$

where ∂_u^k denotes the k-th partial derivative with respect to u.

The probability generating function

$$P_n(u) = \frac{[z^n]F(z,u)}{[z^n]F(z,1)} \tag{6}$$

$$= \sum_m \frac{A_{n,m}}{A_m} u^m \tag{7}$$

$$= \sum_m \mathbf{P}_n(X=m) u^m. \tag{8}$$

Substituting u by e^t gives the familiar moment generating function.

1.3 Convergence of probability distributions

A probability measure is *determined by its moments* if it has finite moments $\alpha_k = \mathbf{E}[x^k]$ of all orders and the power series $\sum_k \alpha_k \frac{r^k}{k!}$ has a positive radius of convergence for r.

THEOREM 1 [Theorem 30.2 in (Billingsley, 1995)]. *Suppose that the distribution of X is determined by its moments and the X_n have moments of all orders and that $\lim_n \mathbf{E}[X_n^r] = \mathbf{E}[X^r]$ for $r = 1, 2, \ldots$. Then the distribution of X_n converges to the distribution of X.*

THEOREM 2. *For probability distributions of X_n, if their kth factorial moment converges to μ^k, then their probability distributions converge to Poisson distribution with mean μ.*

PROOF. A distribution is determined by its moments if $\sum_k \mathbf{E}[x^k]\frac{r^k}{k!}$ converges for any value of r. For a Poisson$[\mu]$ distribution, $\sum_k \mathbf{E}[x^k]\frac{r^k}{k!} = e^{\mu(e^r-1)}$, which converges for any μ and r. Hence, Theorem 1 applies. Since regular moments $\mathbf{E}[X^r]$ are just the linear combinations of factorial moments $\mathbf{E}[X_{(r)}]$, the conclusion in Theorem 1 is also true for factorial moments. Finally for Poisson$[\mu]$, $\mathbf{E}[X_{(r)}] = \mu^r$. ■

2 THE GENERATING FUNCTION APPROACH TO SINGLE-CHROMOSOME GENOMES

In this section, we consider four different cases of genomes containing only one chromosome: unsigned linear, unsigned circular, signed linear and signed circular chromosomes. We first derive the generating functions for each case. For unsigned cases, the limiting probability distributions are derived via factorial moments while for signed cases the direct derivation of the exact distribution from generating functions is possible.

We first introduce an operation (Flajolet and Sedgewick, 2008) that we call the *star operation*. For any genome, from the existing adjacencies, this operation distinguishes an arbitrary set of adjacencies and labels them with stars. For a genome with m adjacencies, there are 2^m different ways of picking starred adjacencies.

By using starred genomes, we can avoid complications due to overcounting certain nested configurations. We can then make use of a straightforward relation (Lemma 1) between starred genomes and genomes without stars to derive the main result in Theorem 3.

LEMMA 1. *If $F(z,u)$ is the bivariate generating function counting the number of genomes with n elements and m adjacencies and $G(z,v)$ is the bivariate generating function counting the number of genomes with n elements and l starred adjacencies, then the star operation corresponds to the substitution $u \to 1+v$ and*

$$G(z,v) = F(z, 1+v) \tag{9}$$

$$F(z,u) = G(z, u-1). \tag{10}$$

PROOF. If $f_{n,m}$, the coefficient of $z^n u^m$ in F is the number of genomes with n elements and m adjacencies, the star operation on these genomes will produce $f_{n,m}\binom{m}{l}$ genomes with l starred adjacencies, where $l = 0, 1, \ldots, m$. We have $\sum_{l=0}^{m} f_{n,m}\binom{m}{l} z^n v^l = f_{n,m} z^n (1+v)^m$. Comparing $G(z,v) = \sum_{n,m} f_{n,m} z^n (1+v)^m$ with $F(z,u) = \sum_{n,m} f_{n,m} z^n u^m$, we have the desired result. ■

2.1 The four generating functions

THEOREM 3. *Denote by $F^{u,l}(z,u), F^{u,c}(z,u), F^{s,l}(z,u)$ and $F^{s,c}(z,u)$ the generating functions correspondingly to unsigned linear, unsigned circular, signed linear and signed circular single-chromosome genomes, where the coefficients of powers of z and u*

count markers and adjacencies, respectively. Then we have:

$$F^{u,l}(z,u)=\sum_{n=0}^{\infty}n!z^n\left(\frac{1+uz-z}{1-uz+z}\right)^n \tag{11}$$

$$F^{u,c}(z,u)=\sum_{n=0}^{\infty}(n-1)!z^n\left(\frac{1+uz-z}{1-uz+z}\right)^n \tag{12}$$

$$F^{s,l}(z,u)=\sum_{n=0}^{\infty}n!(2z)^n\left(\frac{1}{1-uz+z}\right)^n \tag{13}$$

$$F^{s,c}(z,u)=\sum_{n=0}^{\infty}(n-1)!(2z)^n\left(\frac{1}{1-uz+z}\right)^n. \tag{14}$$

PROOF. We count the numbers for the corresponding starred configurations first and derive the generating functions for the original questions via Equation (10). Define a *synteny block* as a block of markers numbered successively either in a increasing or in a decreasing order (for the signed case, there is a minus sign before decreasing ordered markers). For a synteny block of size s, there are $s-1$ adjacencies. The *starred synteny block* is just the synteny block where each adjacency is starred. The generating function $S(z,v)$ counting the number of starred synteny blocks is $2z^2v+2z^3v^2+2z^4v^3+\cdots=\frac{2z^2v}{1-zv}$. The starred genomes are the compositions of starred synteny blocks and free markers—markers that not involved in any starred adjacencies. Call these starred synteny blocks and free markers *free components*. Now we derive the expressions for $G(z,v)$ using the fact that starred genomes are permutations (signed/unsigned, linear/circular) of these free components.

1. Unsigned linear genomes. If there are n free components, there are $n!$ unsigned linear permutations of them. Each free component can be a free marker or a starred synteny block, so the generating function for free components is $z+S(z,v)=z+\frac{2z^2v}{1-zv}$. Then we have

$$G^{u,l}(z,v)=\sum_{n=0}^{\infty}n!\left(z+\frac{2z^2v}{1-zv}\right)^n.$$

So

$$F^{u,l}(z,u)=G^{u,l}(z,u-1)=\sum_{n=0}^{\infty}n!\left(z\frac{1+zu-z}{1-zu+z}\right)^n.$$

2. Unsigned circular genomes. Given n free components, there are $(n-1)!$ circular permutations. So that:

$$G^{u,c}(z,v)=\sum_{n=0}^{\infty}(n-1)!\left(z+\frac{2z^2v}{1-zv}\right)^n$$

$$F^{u,c}(z,u)=\sum_{n=0}^{\infty}(n-1)!\left(z\frac{1+zu-z}{1-zu+z}\right)^n.$$

3. Signed linear genomes. Given n free components there are $n!$ linear permutations. Each free component is either a free marker, which may take two different signs, or a starred synteny block. So the generating function for free components is $2z+S(z,v)=2z+\frac{2z^2v}{1-zv}$. We have:

$$G^{s,l}(z,v)=\sum_{n=0}^{\infty}n!\left(2z+\frac{2z^2v}{1-zv}\right)^n$$

$$F^{s,l}(z,u)=\sum_{n=0}^{\infty}n!2^n\left(z\frac{1}{1-zu+z}\right)^n.$$

4. Signed circular genomes. Similarly we have:

$$G^{s,c}(z,v)=\sum_{n=0}^{\infty}(n-1)!\left(2z+\frac{2z^2v}{1-zv}\right)^n$$

$$F^{s,c}(z,u)=\sum_{n=0}^{\infty}(n-1)!2^n\left(z\frac{1}{1-zu+z}\right)^n.$$

■

2.2 From factorial moments to limiting probability distributions for unsigned genomes

After expanding the generating functions $F(z,u)$, the coefficient of the term z^nu^m is just the number of permutations with n elements and m adjacencies. While this approach is easily followed for signed genomes, it leads to complicated multiple summations for the unsigned cases. Next we derive the probability distribution for unsigned cases by means of factorial moments.

THEOREM 4. *X is the random variable counting the number of adjacencies for unsigned linear or circular single-chromosome genomes. Its k-*th *factorial moment is*

$$\mathbf{E}[X_{(k)}]=2^k\left(1+o\left(\frac{1}{n}\right)\right), \tag{15}$$

while its probability distributions

$$\mathbf{P}[X=k]=e^{-2}\frac{2^k}{k!}\left(1+o\left(\frac{1}{n}\right)\right). \tag{16}$$

The limiting distribution is Poisson[2].

PROOF. Set $P(z,u)=1+uz-z, Q(z,u)=1-uz+z$. Then

$$F^{u,l}=\sum_{n=0}^{\infty}n!z^nP^nQ^{-n}$$

$$F^{u,c}=\sum_{n=0}^{\infty}(n-1)!z^nP^nQ^{-n}.$$

The kth derivative of P^nQ^{-n} can be expanded as

$$\partial_u^k[P^nQ^{-n}]=\sum_{i=0}^{k}\binom{k}{i}\partial_u^iP^n\cdot\partial_u^{k-i}Q^{-n}.$$

Then we have

$$\partial_u^iP^n=n_{(i)}z^iP^{n-i}$$

$$\partial_u^iQ^{-n}=(n+i-1)_{(i)}z^iQ^{-n-i},$$

where $n_{(i)}$ stands for $n(n-1)\ldots(n-i+1)$ and $n_{(0)}=1$.

Since $P(z,u)|_{u=1}=1$ and $Q(z,u)|_{u=1}=1$, we have

$$\partial_u^i P^n|_{u=1}=n_{(i)}z^i \quad \text{and} \quad \partial_u^i Q^{-n}|_{u=1}=(n+i-1)_{(i)}z^i$$

and

$$\partial_u^k\left[P^nQ^{-n}\right]|_{u=1}=z^k\sum_{i=0}^{k}\binom{k}{i}n_{(i)}\cdot(n+i-1)_{(i)}. \tag{17}$$

1. Unsigned linear case:

$$\partial_u^k F^{u,l}|_{u=1}=\sum_{l=0}^{\infty}l!z^{l+k}\sum_{i=0}^{k}\binom{k}{i}(l+k-i+1)_{(k-i)}\cdot l_{(i)}. \tag{18}$$

Then for the ith factorial moment, we can calculate:

$$\begin{aligned}\mathbf{E}[X_{(k)}^{u,l}]&=\frac{1}{n!}[z^n]\partial_u^k F^{u,l}|_{u=1}\\&=\frac{(n-k)!}{n!}\sum_i\binom{k}{i}(n-i-1)_{(k-i)}\cdot(n-k)_{(i)}\\&=2^k\left(1-\frac{k(k+1)}{2n}+O\left(\frac{k^4}{n^2}\right)\right).\end{aligned} \tag{19}$$

2. Unsigned circular case:

$$\partial_u^k F^{u,c}|_{u=1}=\sum_{l=0}(l-1)!z^{l+k}\sum_{i=0}^{k}(l+k-i-1)_{(k-i)}\cdot l_{(i)} \tag{20}$$

$$\begin{aligned}\mathbf{E}[X_{(k)}^{u,c}]&=\frac{1}{(n-1)!}[z^n]\partial_u^k F^{u,c}|_{u=1}\\&=\sum_i\binom{k}{i}\frac{n-k}{n-1}\frac{n-k-1}{n-2}\cdots\frac{n-k-i+1}{n-i}\\&=2^k\left(1-\frac{k(k-1)}{2n}+O\left(\frac{k^4}{n^2}\right)\right).\end{aligned} \tag{21}$$

Let X^u be the number of common adjacencies for unsigned single-chromosome genomes, either linear or circular.

$$\mathbf{E}[X_{(k)}^u]=2^k\left(1+O\left(\frac{1}{n}\right)\right)\overset{n\to\infty}{\longrightarrow}2^k.$$

From Theorem 2, we conclude that the limiting distribution for the number of adjacencies is Poisson[2].

The probability generating function:

$$\begin{aligned}P^u(u)&=\sum_{k=0}^{n}\mathbf{P}[X^u=k]u^k\\&=\sum_{k=0}^{n}\frac{\mathbf{E}[X_{(k)}^u]}{k!}(u-1)^k\\&=\sum_{k=0}^{\infty}\frac{2^k\left(1+O\left(\frac{1}{n}\right)\right)}{k!}(u-1)^k\\&=e^{2u-2}\left(1+O\left(\frac{1}{n}\right)\right).\end{aligned} \tag{22}$$

$$\mathbf{P}[X^u=k]=[u^k]P^u(u)=e^{-2}\frac{2^k}{k!}\left(1+O\left(\frac{1}{n}\right)\right). \tag{23}$$

■

2.3 Derivation of distributions for signed cases

The relatively simpler generating functions for signed genomes enable the direct derivation of probability distributions. We have

THEOREM 5. *For signed linear or circular single-chromosome genomes, the probability distributions of the number of adjacencies are:*

$$\mathbf{P}[X^{s,c}=k]\overset{n\to\infty}{\longrightarrow}e^{-\frac{1}{2}}\frac{1}{k!2^k}\left(1-\frac{2k-1}{2n}\right) \tag{24}$$

$$\mathbf{P}[X^{s,l}=k]\overset{n\to\infty}{\longrightarrow}e^{-\frac{1}{2}}\frac{1}{k!2^k} \tag{25}$$

and their limiting distributions are Poisson[$\frac{1}{2}$].

PROOF. From the generating functions $F(z,u)$ we get the corresponding probability generating functions $P_n(u)$, which give us the probability distribution immediately.

$$\begin{aligned}P_n^{s,l}(u)&=\frac{[z^n]F^{s,l}(z,u)}{[z^n]F^{s,l}(z,1)}\\&=\frac{\sum_{l=0}^{n}l!2^l\binom{n-1}{n-l}(u-1)^{n-l}}{n!2^n}\\&=\exp\left(\frac{u-1}{2}\right)\left(1-\frac{u-1}{2n}+O\left(\frac{\left(\frac{u-1}{2}\right)^n}{(n+1)!}\right)\right)\end{aligned} \tag{26}$$

$$\begin{aligned}P_n^{s,c}(u)&=\frac{[z^n]F^{s,u}(z,u)}{[z^n]F^{s,u}(z,1)}\\&=\frac{\sum_{l=0}^{n}(l-1)!2^l\binom{n-1}{n-l}(u-1)^{n-l}}{(n-1)!2^n}\\&=\sum_{i=0}^{n}\frac{2^{n-i}}{(n-1)!2^n}\frac{(n-1)!(n-i-1)!}{i!(n-i-1)!}(u-1)^i\\&=\sum_{i=0}^{n}\frac{\left(\frac{u-1}{2}\right)^i}{i!}\\&=\exp\left(\frac{u-1}{2}\right)\left(1+O\left(\frac{\left(\frac{u-1}{2}\right)^n}{(n+1)!}\right)\right).\end{aligned} \tag{27}$$

From $\mathbf{P}[X^s=k]=[u^k]P_n^s(u)$ we have

$$\mathbf{P}[X^{s,l}=k]=[u^k]P_n^{s,l}(u)$$
$$=e^{-\frac{1}{2}}\frac{1}{k!2^k}\left(1-\frac{2k-1}{2n}+O\left(\frac{1}{(n-k)!2^{n-k}}\right)\right) \quad (28)$$
$$\mathbf{P}[X^{s,c}=k]=[u^k]P_n^{s,c}(u)$$
$$=e^{-\frac{1}{2}}\frac{1}{k!2^k}\left(1+O\left(\frac{1}{(n-k)!2^{n-k}}\right)\right). \quad (29)$$

■

3 THE PROBABILITY APPROACH FOR MULTICHROMOSOMAL GENOMES

For multichromosomal genomes, the variation in the number of chromosomes, shape (linear or circular) and length (the number of markers) of each chromosome complicate the exact calculation. However, some dominant tendencies emerge when the number of markers is much larger than the number of linear chromosomes. We use a probabilistic approach to characterize these tendencies.

Since the methods for unsigned and signed genomes are essentially the same, we treat the two cases at the same time. For either case, suppose there are n markers, χ_l linear chromosomes and χ_c circular chromosomes in the reference genome and n genes, χ_l' linear chromosomes and χ_c' circular chromosomes in the random genome.

Let γ_i be the event that marker g_i and g_i+1 form an adjacency, in the form of either (g_i, g_i+1) or (g_i+1, g_i). (In the signed case, (g_i, g_i+1) or $(-g_i-1, -g_i)$.)

Denote Λ as the set of adjacencies in the reference genome, i.e. markers i and $i+1$ where i is not the end of a chromosome. Clearly $|\Lambda|=n-\chi_l$.

Let Γ_i^u (Γ_i^s for signed cases) be the indicator random variable for the event γ_i, i.e. Γ_i^u (or Γ_i^s) counts 1 when γ_i occurs, 0 otherwise. In the random genome, let p^u (or p^s) be the probability of event γ_i, where i takes any value from set Λ.

LEMMA 2. *In the random genome, the probability of the event γ_i, where $i\in\Lambda$, is $p^u=\frac{2(n-\chi_l')}{n(n-1)}$ for the unsigned case and $p^s=\frac{n-\chi_l'}{2n(n-1)}$ for the signed case.*

PROOF. In the random genome, marker g_i can be located at the end of some linear chromosome, with probability $\frac{2\chi_l'}{n}$. When this happens, for unsigned genomes, there is only one possible position for g_i+1 to form an adjacency with g_i, which gives the probability $\frac{1}{n-1}$. For signed genomes, γ_i happens when g_i, g_i+1 is located at the left end of the chromosome or $-g_i-1$, $-g_i$ is located at the right end of the chromosome. Either of the two cases gives the probability $\frac{1}{2}\frac{1}{2(n-1)}$.

Gene g_i can also be placed in the interior of chromosomes with probability $\frac{n-2\chi_l'}{n}$. For unsigned genomes, two possible positions are available for g_i+1 to form an adjacency with g_i, with total probability $\frac{2}{n-1}$. While for signed genomes, one possible position is available depending on the sign of g_i, with probability $\frac{1}{2(n-1)}$.

Summing up we have,

$$p^u=\frac{2\chi_l'}{n}\frac{1}{n-1}+\frac{n-2\chi_l'}{n}\frac{2}{n-1}=\frac{2(n-\chi_l')}{n(n-1)} \quad (30)$$
$$p^s=\frac{2\chi_l'}{n}\frac{1}{4(n-1)}+\frac{n-2\chi_l'}{n}\frac{1}{2(n-1)}=\frac{n-\chi_l'}{2n(n-1)}.$$

■

THEOREM 6. *The expected number of adjacencies is*

$$\mathbf{E}[X]=\begin{cases}2-\frac{2(\chi_l+\chi_l'-1)}{n}+O\left(\frac{1}{n^2}\right), & \text{unsigned genome}\\ \frac{1}{2}-\frac{\chi_l+\chi_l'-1}{2n}+O\left(\frac{1}{n^2}\right), & \text{signed genome.}\end{cases} \quad (31)$$

PROOF. Let X^u (X^s) be the number of adjacencies for unsigned genomes (signed genomes), which is just the summation of Γ_i^u's (Γ_i^s's) for all i in Λ.

$$X^u=\sum_{i\in\Lambda}\Gamma_i^u \qquad \text{and} \qquad X^s=\sum_{i\in\Lambda}\Gamma_i^s .$$

The expectations are easily derived:

$$\mathbf{E}[X^u]=\sum_{i\in\Lambda}\mathbf{E}[\Gamma_i^u]=\sum_{i\in\Lambda}p^u=\frac{2(n-\chi_l)(n-\chi_l')}{n(n-1)}$$
$$=2-\frac{2(\chi_l+\chi_l'-1)}{n}+O\left(\frac{1}{n^2}\right) \quad (32)$$
$$\mathbf{E}[X^s]=\sum_{i\in\Lambda}\mathbf{E}[\Gamma_i^s]=\sum_{i\in\Lambda}p^s=\frac{1}{2}\frac{(n-\chi_l)(n-\chi_l')}{n(n-1)}$$
$$=\frac{1}{2}-\frac{\chi_l+\chi_l'-1}{2n}+O\left(\frac{1}{n^2}\right). \quad (33)$$

■

THEOREM 7. *The variance of the number of adjacencies is*

$$\mathbf{V}[X]=\begin{cases}2-\frac{2(\chi_l+\chi_l'+1)}{n}+O\left(\frac{1}{n^2}\right), & \text{unsigned genome}\\ \frac{1}{2}-\frac{\chi_l+\chi_l'-1}{2n}+O\left(\frac{1}{n^2}\right), & \text{signed genome.}\end{cases} \quad (34)$$

PROOF. $\mathbf{V}[X]=\mathbf{E}[X^2]-\mathbf{E}^2[X]$, and we first calculate the non-centered second moment $\mathbf{E}[X^2]$, which can be expressed as the following summation.

$$\mathbf{E}[X^2]=\mathbf{E}\left[\left(\sum_{u\in\Lambda}\Gamma_i\right)^2\right] \quad (35)$$
$$=\sum_{|i-j|>1}\mathbf{E}[\Gamma_i\Gamma_j]+\sum_{|i-j|=1}\mathbf{E}[\Gamma_i\Gamma_j]+\sum_i\mathbf{E}[\Gamma_i^2].$$

In the last expression, there are $(n-\chi_l)(n-\chi_l-3)+2\chi_l$, $2(n-2\chi_l)$ and $n-\chi_l$ summands in the three summations correspondingly. For the present version of this article, we will not detail the case-by-case calculation of these summands, which results in the quantities in the statement of this theorem. ■

THEOREM 8. *The limiting probability distribution is* Poisson[2] *for unsigned genomes and* Poisson[$\frac{1}{2}$] *for signed genomes.*

PROOF. We first prove kth factorial moment converges to 2^k for unsigned genomes and 2^{-k} for signed genomes. Then by Theorem 2, we get the above conclusion.

Since Γ_i is the indicator random variable of value 0 or 1, the k-th factorial moment can be written as the following summation, where the k index runs over all k-tuples on the set Λ and no two indices take on the same value.

$$\mathbf{E}[X(X-1)\ldots(X-k+1)] = \sum_{i_1,i_2,\ldots,i_k\in\Lambda} \mathbf{E}[\Gamma_{i_1}\Gamma_{i_2}\ldots\Gamma_{i_k}]. \quad (36)$$

Since the value of conditional expection depends on indices $i_1, i_2, \ldots, i_k$, the summation on the right hand side of (36) is split into two summations:

$$\Sigma_1 = \sum_{\substack{i_1,i_2,\ldots,i_k\in\Lambda \\ |i_l-i_m|>1,\forall l,m}} \mathbf{E}[\Gamma_{i_1}\Gamma_{i_2}\ldots\Gamma_{i_k}] \quad (37)$$

$$\Sigma_2 = \sum_{\substack{i_1,i_2,\ldots,i_k\in\Lambda \\ |i_l-i_m|=1,\exists l,m}} \mathbf{E}[\Gamma_{i_1}\Gamma_{i_2}\ldots\Gamma_{i_k}]. \quad (38)$$

Denote $\mathbf{E}[\Gamma_{i_1}\Gamma_{i_2}\ldots\Gamma_{i_k}|g_{i_1}, g_{i_2}, \ldots, g_{i_k}]$ as the conditional expectation for given indices $\{i_j : 1 \le j \le k\}$, when the i_l-th element on the random genome is g_{i_l} for all $1 \le l \le k$. Then the unconditional expectation can be expressed as:

$$\mathbf{E}[\Gamma_{i_1}\Gamma_{i_2}\ldots\Gamma_{i_k}] = \frac{1}{\binom{n}{k}} \sum_{g_{i_1},g_{i_2},\ldots,g_{i_k}} \mathbf{E}[\Gamma_{i_1}\Gamma_{i_2}\ldots\Gamma_{i_k}|g_{i_1}, g_{i_2}, \ldots, g_{i_k}]. \quad (39)$$

The set $\{g_{i_1}, g_{i_2}, \ldots, g_{i_k} : i_1, i_2, \ldots, i_k \in \Lambda\}$ can be split into:

$$\mathcal{A}_1 = \{g_{i_1}, g_{i_2}, \ldots, g_{i_k} : |g_{i_l} - g_{i_m}| > 2, \forall l, m\}$$
$$\mathcal{A}_2 = \{g_{i_1}, g_{i_2}, \ldots, g_{i_k} : |g_{i_l} - g_{i_m}| \le 2, \exists l, m\}.$$

Then the expection $\mathbf{E}' = \mathbf{E}[\Gamma_{i_1}\Gamma_{i_2}\ldots\Gamma_{i_k}]$ on the right hand side of (37) becomes:

$$\mathbf{E}' = \mathbf{P}(\mathcal{A}_1)\mathbf{E}[\Gamma_{i_1}\Gamma_{i_2}\ldots\Gamma_{i_k}|\mathcal{A}_1] + \mathbf{P}(\mathcal{A}_1)\mathbf{E}[\Gamma_{i_1}\Gamma_{i_2}\ldots\Gamma_{i_k}|\mathcal{A}_2]$$

$$\mathbf{P}(\mathcal{A}_1) \ge \frac{(n-2\chi_l')(n-2\chi_l'-5)\ldots(n-2\chi_l'-5k+5)}{n(n-1)\ldots(n-k+1)} = 1 - \frac{2k(\chi_l'+k-1)}{n} + O\left(\frac{1}{n^2}\right) \quad (40)$$

$$\mathbf{P}(\mathcal{A}_2) = 1 - \mathbf{P}(\mathcal{A}_1) \le \frac{2k(\chi_l'+k-1)}{n} + O\left(\frac{1}{n^2}\right). \quad (41)$$

Since $\mathcal{A}_1$ asymptotically occurs with probability 1, as n goes to ∞, then $\mathbf{E}' \overset{n\to\infty}{\longrightarrow} \mathbf{E}[\Gamma_{i_1}\Gamma_{i_2}\ldots\Gamma_{i_k}|\mathcal{A}_1]$, which takes the maximum value among all conditional expectations as:

$$\mathbf{E}[\Gamma_{i_1}\Gamma_{i_2}\ldots\Gamma_{i_k}|\mathcal{A}_1] = \begin{cases} \frac{2^k}{(n-k)(n-k-1)\ldots(n-2k+1)}, & \text{unsigned genomes} \\ \frac{1}{2^k(n-k)(n-k-1)\ldots(n-2k+1)}, & \text{signed genomes.} \end{cases} \quad (42)$$

Since the number of summands in Σ_1 is at least $n(n-3)(n-6)\ldots(n-3k+3) = n^k + O(n^{k-1})$, the number of summands in Σ_2 is in the order $O(n^{k-1})$ and the unconditional expectation in Σ_2 is no larger than $\mathbf{E}[\Gamma_{i_1}\Gamma_{i_2}\ldots\Gamma_{i_k}|\mathcal{A}_1]$, then

$$\Sigma_1 = \left(\sum_{\substack{i_1,i_2,\ldots,i_k\in\Lambda \\ |i_l-i_m|>1,\forall l,m}} \mathbf{E}[\Gamma_{i_1}\Gamma_{i_2}\ldots\Gamma_{i_k}|\mathcal{A}_1]\right)\cdot\left(1+O\left(\frac{1}{n}\right)\right) = \left(n^k + O(n^{k-1})\right)\cdot O(n^{-k}) = O(1) \quad (43)$$

$$\Sigma_2 = O(n^{k-1})\cdot O(n^{-k}) = O\left(\frac{1}{n}\right). \quad (44)$$

So Σ_2 is at least one order of magnitude smaller than Σ_2.

$$\mathbf{E}[X(X-1)\ldots(X-k+1)] = \Sigma_1\cdot\left(1+O\left(\frac{1}{n}\right)\right) = \left(n^k + O(n^{k-1})\right)\cdot\mathbf{E}[\Gamma_{i_1}\Gamma_{i_2}\ldots\Gamma_{i_k}|\mathcal{A}_1] = \begin{cases} 2^k + O\left(\frac{1}{n}\right), & \text{unsigned genome} \\ 2^{-k} + O\left(\frac{1}{n}\right), & \text{signed genome.} \end{cases} \quad (45)$$

From the convergence of the factorial moments, we have the convergence of the probability distribution: the limiting probability distribution of number of adjacencies is Poisson[2] for unsigned genomes and Poisson[$\frac{1}{2}$] for signed genomes. ■

REMARK 1. *Using the methods of Theorem 6, we can calculate the covariance between Γ_i and Γ_j as $Cov(\Gamma_i, \Gamma_j) = \mathbf{E}[\Gamma_i\Gamma_j] - \mathbf{E}[\Gamma_i]\mathbf{E}[\Gamma_j]$:*

$$Cov(\Gamma_i^u, \Gamma_j^u) = \frac{4}{n^3} + O(\frac{1}{n^4}), \qquad |i-j| > 1$$
$$Cov(\Gamma_i^u, \Gamma_j^u) = -\frac{2}{n^2} + \frac{4\chi_l'-2}{n^3} + O\left(\frac{1}{n^4}\right), \qquad |i-j| = 1$$
$$Cov(\Gamma_i^s, \Gamma_j^s) = \frac{1}{4n^3} + O\left(\frac{1}{n^4}\right), \qquad \text{for all cases.} \quad (46)$$

Since $\mathbf{V}[\Gamma_i^u] = \frac{2}{n} + O(\frac{1}{n^3})$ and $\mathbf{V}[\Gamma_i^s] = \frac{1}{2n} + O(\frac{1}{n^3})$, the covariances are at least one order of magnitude smaller than the variance. The Γ_i's can be treated as independent identical random variables under a mild approximation, which leads to Poisson distributions.

4 CONCLUSION

In this article, we used a combinatorial approach to find the generating functions counting the number of genomes with given numbers of markers and adjacencies for genomes with only one chromosome. We used probabilistic methods to calculate the exact values for random expectations and variances of the number of adjacencies for genomes with any number of linear and circular chromosomes. The overall conclusion is that the limiting probability distribution is Poisson[2] for the unsigned case and Poisson[$\frac{1}{2}$] for the signed case.

Based on the limiting Poisson distribution, we can devise a statistical test for whether the two genomes contain a significant evolutionary signal, when the number of markers is not too small.

Table 1. *P*-values for given number of adjacencies when n is large

Number of adjacencies	*P*-value for unsigned case	*P*-value for signed case
0	1	1
1	0.8647	0.3935
2	0.5940	0.0902
3	0.3233	0.0144
4	0.1429	0.0018
5	0.0527	0.00017
6	0.0166	0.000014
7	0.0045	1.00×10^{-6}
8	0.0011	6.23×10^{-8}
9	2.37×10^{-4}	3.50×10^{-9}
10	4.64×10^{-5}	4.10×10^{-10}

For unsigned genomes with number of adjacencies a, the *P*-value is calculated by $p^u(a)=1-\sum_{i=0}^{a-1} e^{-2}\frac{2^i}{i!}$. For signed genomes with number of adjacencies a, the *P*-value is calculated by $p^s(a)=1-\sum_{i=0}^{a-1} e^{-\frac{1}{2}}\frac{2^{-i}}{i!}$. Based on Table 1, when the unsigned distance is >5 or the signed adjacency distance is >2, a statistical test with a critical region of 5% will reject the null hypothesis of randomness and accept that there is a significant evolutionary signal between the two genomes involved.

ACKNOWLEDGEMENTS

We thank Glenn Tesler for discussing his previous work on this topic with us, and Daniel Panario for his suggestions and encouragement. Research supported in part by a grant to D.S. from the Natural Sciences and Engineering Research Council of Canada (NSERC). D.S. holds the Canada Research Chair in Mathematical Genomics.

Conflict of Interest: none declared.

REFERENCES

Abramson,M. and Moser,W. (1966) Combinations, successions and the n-kings problem. *Math. Mag.*, **39**, 269–273.

Billingsley,P. (1995) *Probability and Measure*, 3rd edn. Wiley InterScience, New York.

Flajolet,P. and Sedgewick,R. (2008) *Analytic Combinatorics*. Cambridge University Press.

Nadeau,J.H. and Taylor,B.A. (1984) Lengths of chromosomal segments conserved since divergence of man and mouse. *Proceedings of the National Academy of Sciences (USA)*, **81**, 814–818.

Robbins,D.P. (1980) The probability that neighbors remain neighbors after random rearrangements. *Am. Math. Mon.*, **87**, 122–124.

Sankoff,D. and Blanchette,M. (1998) Multiple genome rearrangement and breakpoint phylogeny. *J. Comp. Biol.*, **5**, 555–570.

Tesler,G. (2005) Decomposition of permutations into rising and falling subsequences. unpublished manuscript, 2005.