JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 13, Number 5, 2006

Pp. 1005–1012

The Distribution of Genomic Distance between Random Genomes

DAVID SANKOFF and LANI HAQUE

ABSTRACT

We study the probability distribution of genomic distance d under the hypothesis of random gene order. We translate the random order assumption into a stochastic method for constructing the alternating color cycles in the decomposition of the bicolored breakpoint graph. For two random genomes of length n, we show that the expectation of $n - d$ is $O(\frac{1}{2}\log n)$.

Key words: breakpoint graph, cycle size, probability model, genomic rearrangements, random graph.

1. INTRODUCTION

GENOMIC DISTANCES based on the number of rearrangement steps—including inversions, transpositions, and reciprocal translocations—necessary to convert one genome to another are potentially meaningful measures of the evolutionary divergence among genomes. The significance of a particular value for such a measure depends on how much it differs from the value when the gene order of one genome is randomized with respect to the other. In this paper, we study the probability distribution of genomic distance under the null hypothesis of random gene order, and establish its asymptotic behavior for large genomes.

Genomic distances can be efficiently computed using the bicolored "breakpoint graph." In this graph, the vertices are the $5'$ and the $3'$ ends of each gene, and the edges represent the adjacencies between the ends of successive genes on either DNA strand in the two genomes. We color the edges from one genome red, the other black. With the addition of dummy vertices at the endpoints of linear chromosomes (and possibly some dummy edges in the case of multichromosomal genomes), the breakpoint graph decomposes automatically into alternating colored cycles. The number b of breakpoints—gene adjacencies in one genome not occurring in the other genome—and the number c of cycles in the breakpoint graph are the dominant components in formulae for genomic distance. Indeed, in the most recent and most inclusive formulation of genomic distance (Yancopoulos *et al.*, 2005), encompassing inversions, reciprocal translocations, chromosome fission and fusion, and block exchange (including transposition), where the latter operation is counted as two steps, the genome distance d is, remarkably, simply

$$d = b - c. \tag{1}$$

Some time ago, it was observed (Kececioglu and Sankoff, 1994) in simulations of random genomes with hundreds of genes that the distance d seldom differed from n by more than a few rearrangements, even though it is easy to construct examples where d is as low as $\frac{b}{2}$. Our approach to explaining this

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

will be to investigate the consequences of random gene order on the distribution of b and c, and on their limiting behavior for large genomes. Our initial goal here is to characterize the case of a single circular chromosome; our methods carry over to the treatment of genomes consisting of one or more linear chromosomes, with analogous results, but the treatment of chromosomal endpoints and dummy vertices introduces some complications, which we will detail elsewhere.

In Sections 2–4, we define the breakpoint graph, discuss the distribution of b, and derive the distribution of the size of a cycle containing a specified vertex. Our main result in Sections 5–7 is that, in expectation, $n-d = O(\log n)$. We first derive this in Sections 5 and 6 for a relaxed model where a genome need not be a single circular chromosome—it may consist of several circles, e.g., one large circle and a number of smaller "plasmids"—since the combinatorics of this case are very simple. We then show through simulations in Section 7 that the case where both genomes are single circles (no plasmids) behaves virtually identically to the relaxed model for large n and indeed for genomes as small as $n = 10$. Moreover, these simulations indicate that $n - d = C + \frac{1}{2}\log n$, where the constant $C \approx 1$.

2. DEFINITIONS AND CONSTRUCTIONS

Consider two genomes R and B, each with a circularly ordered set of n numbered or named chromosomal segments (e.g., genes), and a 1-1 "orthology" relation between the two sets. For each segment, its two ends are distinguished, e.g., as the $5'$ and $3'$ ends in the case where the segments are genes. The ends of any two consecutive segments in a genome that abut, or are closest to each other in a genome, define an "adjacency."

If an adjacency occurs in both genomes, whether or not the two segments involved have reversed order, this adjacency is "conserved"; if it occurs only in one genome, it is a "breakpoint."

The breakpoint graph has $2n$ vertices corresponding to the $2n$ segment ends. The adjacencies in R determine n red edges and the adjacencies in B determine n black edges. Because each vertex is incident to exactly one red and one black edge, the graph decomposes naturally into one or more disjoint alternating-color cycles. If such a cycle has h red edges and h black edges, we say it is an h-cycle.

3. THE RANDOMNESS HYPOTHESIS AND THE NUMBER OF BREAKPOINTS

If the segments in B are randomly ordered and randomly signed in comparison with their orthologous counterparts in R, for each of the n adjacencies in R, the probability that it will also occur in B is $\frac{1}{2n}$. The expected number of conserved adjacencies, then, is $\frac{1}{2}$ and, for large n, the number of conserved adjacencies $n - b$ will be distributed approximately as a Poisson distribution with parameter $\frac{1}{2}$.

4. THE RANDOMNESS HYPOTHESIS AND THE SIZE OF A CYCLE

In this section, we relax the single-circle condition on genomes B and R. The only structure we impose, in each genome separately, is that every vertex is adjacent to one other vertex, and that these pairings, which define the black lines from genome B and the red lines from genome R in the breakpoint graph, are constructed at random from the $2n$ vertices. Since every vertex is incident to exactly one black line and one red line as before, the breakpoint graph still decomposes into disjoint alternating color cycles. Studying the statistical structure of the set of cycles is facilitated by relaxing the single-circle condition, but the consequence is that the random choice of vertices defines a genome that may be a single circle, but which in general consists of several circular "plasmids."

For example, consider a gene whose two vertices in the breakpoint graph are v and v'. If the random pairing in one of the genomes involves v adjacent to v', the gene constitutes a circular plasmid by itself in that genome. This does not however tell us anything about the size of the alternating cycle containing the two vertices in the breakpoint graph, which would require information about *both* genomes.

Consider any vertex v. The circle containing v in a genome contains v and v', the two vertices representing the two ends of the same gene. It also contains u' and w, where u' and v are chosen by the random

process to be adjacent in that genome, the adjacent pair v' and w, and u, and w', and so on. Eventually, the two ends of construction will arrive at the two ends of a single gene, closing the circle. Note that one possibility is that all the genes are in a single circle.

The two preceding paragraphs discussed the structure of the individual genomes. We now examine the structure of the breakpoint graph determined by the two genomes. For any vertex v, consider the red edge vu incident to v and the black edge vw incident to v. By randomness,

$$\text{Prob}(w = u) = \frac{1}{2n-1}, \tag{2}$$

since there are $2n-1$ equally probable values of w, once v is removed from consideration.Then if $P_n(h)$ is the probability that any particular vertex is in a h-cycle,

$$P_n(1) = \frac{1}{2n-1}. \tag{3}$$

Similarly, we can show that

$$P_n(2) = \frac{2n-2}{2n-1}\frac{1}{2n-3}. \tag{4}$$

and in general

$$P_n(h) = 2^{2(h-1)}\frac{(2n-2h)!}{(2n-1)!}\left(\frac{(n-1)!}{(n-h))!}\right)^2, \qquad 1 \le h \le n. \tag{5}$$

Using Stirling's approximation,

$$P_n(h) \sim \frac{1}{\sqrt{2(2n-1)(n-h)}} \tag{6}$$

Let $x = \frac{h}{n}$. As $n \to \infty$, $P_n(h)$ converges to $p(x)\frac{1}{n}$, where p is a probability density on the interval $[0, 1)$ defined by

$$p(x) = \frac{1}{2\sqrt{(1-x)}}, \qquad 0 \le x < 1, \tag{7}$$

with expectation $\frac{2}{3}$.

5. ADDING CYCLES

Once a cycle of size h, using up a proportion $x = \frac{h}{n}$ of the vertices, has been constructed, we can build a second cycle by starting with any vertex not in the first cycle and proceeding as before, but confined to the remaining vertices. In the limiting model, the probability density $p_2(x)$ of the proportion of vertices contained in the two cycles is then found by the convolution:

$$p_2(x) = \int_0^x \frac{1}{2\sqrt{(1-y)}}\frac{1}{2\sqrt{\left(1-\frac{x-y}{1-y}\right)}}(1-y)^{-1}dy \tag{8}$$

$$= \frac{-\log(1-x)}{4\sqrt{1-x}}, 0 \le x < 1. \tag{9}$$

In the same way, we can show that the probability density $p_\kappa(x)$ of the proportion of vertices contained in κ cycles is:

$$p_\kappa(x) = \frac{[-\log(1-x)]^{\kappa-1}}{A(\kappa)\sqrt{1-x}}, \qquad 0 \le x < 1, \tag{10}$$

where

$$A(\kappa) = 2^{\kappa}(\kappa - 1)!. \tag{11}$$

The mean of this density is

$$E_{\kappa}(x) = \frac{3^{\kappa} - 1}{3^{\kappa}} \tag{12}$$

and its variance is

$$\mathrm{Var}_{\kappa}(\mathrm{x}) = \frac{1}{5^{\kappa}} - \frac{1}{9^{\kappa}}. \tag{13}$$

6. HOW MANY CYCLES?

We will discuss the number of cycles in the breakpoint graph of two random genomes, under the relaxed model, from two points of view. The first is based on an algorithm for constructing the cycles, where we focus the number of cycles necessary to use up all the vertices in the graph. The second is based on the expected number of cycles in the breakpoint graph. While the latter approach gives a more unequivocal result, the former approach will prove useful later, in comparing the more difficult single-circle case with the more tractable relaxed model, using simulations.

6.1. The constructive approach

In a breakpoint graph, all vertices are incorporated into cycles. To estimate when this will happen in the course of our procedure for accumulating cycles one at a time, we ask, motivated by Equation (12), for what κ is $E_{\kappa}(x) = \frac{3^{\kappa}-1}{3^{\kappa}} \geq 1 - \frac{1}{2n}$, i.e., for what κ can we expect the last remaining vertex to be incorporated into a cycle? This condition is simply

$$\kappa \geq C_1 + C_2 \log n, \tag{14}$$

where $C_1 = \frac{\log 2}{\log 3}$ and $C_2 = \frac{1}{\log 3}$. We can specify a stronger criterion for all vertices to be included in cycles by requiring $E_{\kappa}(x) - C_3 \mathrm{Var}_{\kappa}(\mathrm{x}) \geq 1 - \frac{1}{2\mathrm{n}}$, using Equation (13), but this will just produce Condition (14) again, with different values of C_1 and C_2.

A "stopping time" approach to the same question is based on

$$Q(\kappa) = \mathrm{Prob}\left(x > 1 - \frac{1}{n} \,\middle|\, \kappa\right), \tag{15}$$

representing the probability that the first κ cycles contain a proportion greater than $1 - \frac{1}{n}$ of the vertices. From Equation (10), we calculate

$$Q(\kappa) = \int_{1-\frac{1}{n}}^{1} p_{\kappa}(x)dx \tag{16}$$

$$= Q(\kappa - 1) + \frac{(\log \sqrt{n})^{\kappa - 1}}{\sqrt{n}(\kappa - 1)!}, \tag{17}$$

where $Q(0) = 0$. Then $Q(\kappa)$ is the $(\kappa - 1)$-st partial sum of a Poisson distribution with parameter $\log \sqrt{n}$. The value of $Q(\kappa) - Q(\kappa - 1)$ represents the probability of a proportion $x > 1 - \frac{1}{n}$ of the vertices being used up after κ but not $\kappa - 1$ cycles, in other words that $1 - \frac{1}{n}$ will be exceeded exactly when the κ-th cycle is constructed. The mean of this distribution is $\frac{1}{2} \log n$.

6.2. The expected number of cycles

Richard Friedberg (personal communication) has proved, based on Equation (7), that the expected number of cycles approaches

$$\kappa = \log 2 + \frac{\gamma}{2} + \frac{1}{2}\log n, \tag{18}$$

where $\gamma \sim 0.577\ldots$ is Euler's constant.

7. SINGLE-CIRCLE GENOMES

Though the analysis in the previous sections carries through without any change when either one of the two genomes being compared is required to be a single circle, imposing this condition on *both* genomes results in complex restrictions on which adjacencies may co-occur in a genome, so there is no easy way to salvage the analysis in Sections 4–6. On the other hand, it is not clear how these restrictions would affect the expected number of cycles $E(\kappa)$. Indeed, results of Kim and Wormald (2001) ensure that, restricted to the case where $\kappa < \frac{n}{40}$, the probabilistic structure of the breakpoint graph is asymptotically independent of whether the genomes are single circles or not. We would conjecture that the restriction to $\kappa < \frac{n}{40}$ could be removed, but this would not be a mathematically easy task.

Finally, we carried out simulations on pairs of random circular genomes. As seen in Fig. 1, these simulations are completely consistent with Friedberg's result in Equation (18) for the case without the restriction to single-circle genomes. Note that $\frac{\gamma}{2} + \log 2 \sim 0.98\ldots$ would be indistinguishable from 1 in the trend line, given the sample size of our simulations.

To explore the similarity between the unrestricted case and the single-circle case in more detail, we examine 1,000 simulations of random genomes of length $n = 20{,}000$. For each genome, as each cycle is constructed as described in Section 4, we calculate the proportion of the available vertices incorporated in

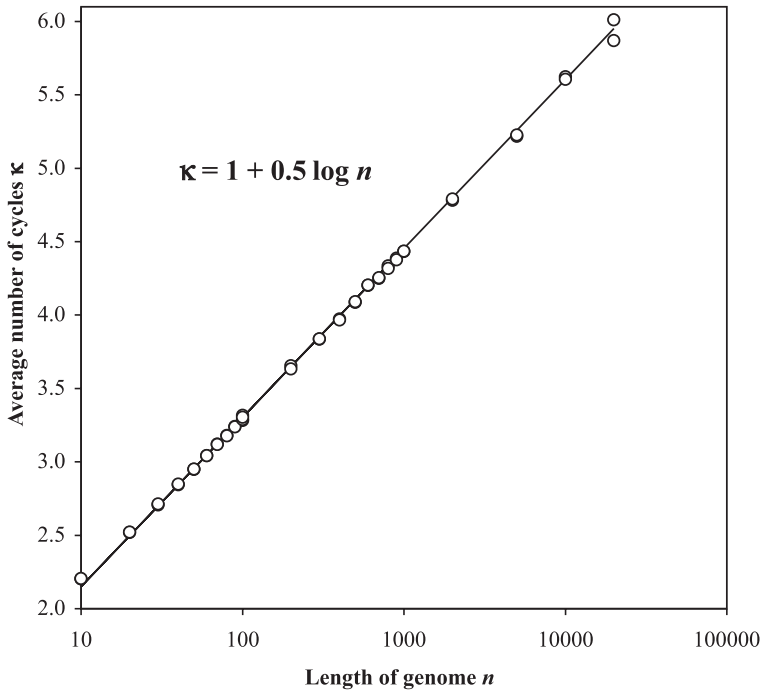

FIG. 1. Average number of cycles induced by random permutation of circular genomes of length n. Sample size 100,000 for $n \leq 100$, 10,000 for $100 < n \leq 1{,}000$, and 1,000 for $1{,}000 < n$. Two sets of simulations depicted for each n.

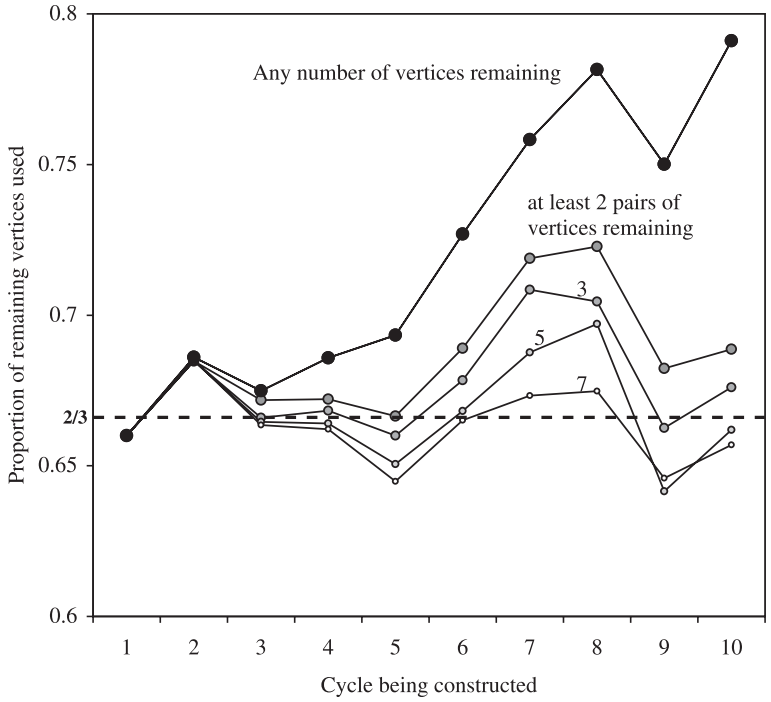

FIG. 2. Average proportion of the available vertices incorporated in the h-th cycle, restricted to cases where there are a minimum number of available vertices.

the cycle. We know from Sections 4 and 5 that without the single-circle restriction, in the limiting case $\frac{2}{3}$ of the available vertices are used at each step. In plotting the average proportion of the 1,000 simulations, we see (top curve in Fig. 2) that, while $\frac{2}{3}$ of the vertices are used in the first cycle, an increasing proportion is used in each successive cycle.

It must be remembered, however, that starting with the second cycle, there may be very few vertices left to construct the cycle, and that the proportions used in this context will be higher than $\frac{2}{3}$. For example, if only two vertices are left, 100% of them must be used in the cycle. Figure 2 also shows how, when simulations with only a few vertices left to build a cycle are excluded from the average calculation, the increase in proportion used is virtually eliminated. The curves rapidly approach a constant function: Proportion used in h-th cycle $= \frac{2}{3}$.

We can conclude that the single-circle restriction, while having a clear effect on the structure of genomes, does not have discernible effect on the statistics of cycle size. While it remains an open mathematical question, the coincidence between Equation (18) and Figure 1 is unmistakable.

8. DISCUSSION

Richard Friedberg (personal communication) has ponited out that, in the comparison of *unsigned* random genomes of length n, the average number of cycles approaches $\log n + \gamma$. The analogy of this to Condition (14), Equation (18), and Fig. 1 is yet another indication that the latter equation also holds for the single-circle case.

One reason this may be so is that, as n increases, the number of pieces of chromosome in the model of Section 4 grows very slowly, and the number of vertices per cycle grows relatively quickly, so that the effects of circularity are attenuated. The expected number of pieces in a genome with random adjacencies can be obtained in exactly the same way as the expected number of cycles in the breakpoint graph of two random genomes, and is indeed given by the same formula in Equation (18). Wei Xu (personal communication) showed this by pointing out that we can add a set of edges between the endpoints of each gene in the breakpoint graph and use a third color, say blue, to distinguish them. Then we can analyze the

bicolored graph with blue and red edges in the same way as we analyzed that with black and red edges. The idea that a third color might be a useful way of understanding the breakpoint graph is hinted at in the figures in Pevzner and Tesler (2003b), and was also mentioned to us recently by Pavel Pevzner (personal communication), though not in connection with this work.

Given that the distribution of $n - b$ approaches a Poisson distribution with fixed parameter 0.5 while c is basically logarithmic in n, we can now understand from Equation (1) why observed values of d for simulated random genomes were so close to n (Kececioglu and Sankoff, 1994).

When order data on genomes are constructed in terms of homologous segments of genome sequences, all the n segment adjacencies are breakpoints; two segments adjacent in both genomes would be considered a single segment. Our cycle construction under this constraint is slightly different from that in Section 4, but the Limiting Distribution (7) is the same. Whether $n - b$ is Poisson distributed or $n - b \equiv 0$ is of little consequence here.

Consider the "reuse" statistic (Pevzner and Tesler, 2003a; Sankoff and Trinh, 2005), $r = \frac{2d}{b}$. Two genomes that have diverged through d rearrangements may generate as many as $b = 2d$ breakpoints, in which case their breakpoint graph contains d cycles, each containing four edges, and $r = 1$. In other cases, however, a single breakpoint can be an endpoint of several segments involved in different rearrangements. To the extent this breakpoint reuse occurs, $b - d$ diminishes towards zero, cycles become larger and less numerous, and r approaches 2.

If we consider random genomes,

$$\frac{2d}{b} \to \frac{2\left(n - \frac{1}{2}\log n\right)}{n} \to 2, \tag{19}$$

as well.

Inference about breakpoint reuse as an evolutionary phenomenon, then, is compromised by the fact that the pattern of larger and fewer cycles occurs both when comparing genomes that have actually experienced such reuse and in genomes that are purely randomly ordered with respect to each other. This randomness, or apparent randomness, can be the result of noise, such as that produced by not including all the segments in the analysis (Sankoff and Trinh, 2005), or the result of processes that cannot be inferred from the breakpoint graph in the same way as, for example, inversion, transposition, and translocation.

Thus, there is no way of determining, without additional information, whether large cycles and apparent breakpoint reuse are genuine reflections of evolutionary process or simply the consequence of noise and randomness. However large the shortfall of b compared to $2d$, this should raise commensurate suspicions about both the data and the assumptions about the rearrangement process.

ACKNOWLEDGMENTS

In commenting on a first draft of this paper, Richard Friedberg pointed out the essential fact that the analysis in Sections 4 and 5 requires relaxing the "no plasmids," i.e., single-circle restriction. We also thank him for communicating his results on unsigned genomes as well as signed genomes without the single-circle restriction, and tracking down and explaining the Kim and Wormald (2001) reference to us. We also thank Wei Xu for his contribution cited in the Discussion and Nick Wormald for helpful comments. Research was supported in part by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC). D.S. holds the Canada Research Chair in Mathematical Genomics and is a Fellow of the Evolutionary Biology Program of the Canadian Institute for Advanced Research.

REFERENCES

Kececioglu, J., and Sankoff, D. 1994. Efficient bounds for oriented chromosome inversion distance. *Lect. Notes Comput. Sci.* 807, 307–325.

Kim, J.H., and Wormald, N.C. 2001. Random matchings which induce Hamilton cycles, and Hamiltonian decompositions of random regular graphs. *J. Combin. Theor. B* 81, 20–44.

Pevzner, P., and Tesler, G. 2003a. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc. Natl. Acad. Sci. USA* 100, 7672–7677.

Pevzner, P.A., and Tesler, G. 2003b. Genome rearrangements in mammalian genomes: Lessons from human and mouse genomic sequences. *Genome Res.* 13, 37–45.

Sankoff, D., and Trinh, P. 2005. Chromosomal breakpoint re-use in genome sequence rearrangement. *J. Comput. Biol.* 12, 812–821.

Yancopoulos, S., Attie, O., and Friedberg, R. 2005. Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics* 21, 3340—3346.

Address correspondence to:
David Sankoff
Department of Mathematics and Statistics
University of Ottawa
Ottawa, ON, Canada K1N 6N5

E-mail: sankoff@uottawa.ca