Paths and Cycles in Breakpoint Graphs of Random Multichromosomal Genomes

Wei Xu[1], Chunfang Zheng[2], and David Sankoff[1]

[1] Department of Mathematics and Statistics, University of Ottawa, Canada K1N 6N5
[2] Department of Biology, University of Ottawa, Canada K1N 6N5
{wxu060, czhen033, sankoff}@uottawa.ca

Abstract. We study the probability distribution of genomic distance d under the hypothesis of random gene order. We interpret the random order assumption in terms of a stochastic method for constructing the alternating colour cycles in the decomposition of the bicoloured breakpoint graph. For two random genomes of length n and χ chromosomes, we show that the expectation of $n + \chi - d$ is $O(\frac{1}{2}\log\frac{n+\chi}{2\chi} + \frac{3}{2}\chi)$. We then discuss how to extend these analyses to the case where intra- and interchromosomal operations have different probabilities.

1 Introduction

Though there is a large literature on chromosomal rearrangements in genome evolution and algorithms for inferring them from comparative maps, there is a need for ways to statistically validate the results. Are the characteristics of the evolutionary history of two related genomes as inferred from an algorithmic analysis different from the chance patterns obtained from two unrelated genomes? Implicit in this question is the notion that the null hypothesis for genome comparison is provided by two genomes, where the order of markers (genes, segments or other) in one is an appropriately randomized permutation of the order in the other. In a previous paper [5], we formalized this notion for the case of the comparison of two random circular genomes, such as are found in prokaryotes and in eukaryotic organelles. We found that the expected number of inversions necessary to convert one genome into the other is $n-O(\frac{1}{2}\log n)$, where n is the number of segments (or other markers). Related work has been done by R. Friedberg (personal communication) and by Eriksen and Hultman [1].

In another paper [4], we used simulations to throw doubt on whether the order of syntenic blocks on human and mouse retains enough evolutionary signal to distinguish it from the case where the blocks on each chromosome are randomly permuted.

In this paper, we begin to bridge the gap between mathematical analysis of simple genomes and simulation studies of advanced genomes. We extend the mathematical approach in [5] to the more difficult case of genomes with multiple linear chromosomes, such as those of eukaryotic nuclear genomes, which not only undergo inversion of chromosomal segments, but also interchromosomal translocation. The presence of chromosomal endpoints changes the problem

G. Bourque and N. El-Mabrouk (Eds.): RECOMB-CG 2006, LNBI 4205, pp. 51–62, 2006.

in a non-trivial way, requiring new mathematical developments. Key to our approach in this and previous papers is the introduction of randomness into the construction of the breakpoint graph rather than into the genomes themselves, which facilitates the analysis without materially affecting the results. One aspect of this is that the random genomes with multiple linear chromosomes may also include one or more small circular fragments, or *plasmids*.

Our main result is that the number of operations necessary to convert one genome into the other is $n - O(\frac{1}{2}\log\frac{n+\chi}{2\chi} + \frac{3}{2}\chi)$, where χ is the number of chromosomes in each genome. This result is validated by exact calculations of a recurrence up to large values of n and χ, by simulations, by analytic solution of a somewhat relaxed model, and by solving the limiting differential equation derived from the recurrence.

We also propose models where the the randomness is constrained to assure a realistic predominance of inversion over translocation. We use simulations of this model to demonstrate how key properties of the breakpoint graph depend on the proportion of intra- versus interchromosomal exchanges.

2 The Breakpoint Graph: Definitions and Constructions

Genomic distances can be efficiently computed using the bicoloured *breakpoint graph*. In this graph, the $2n$ vertices are the 5' and the 3' ends of each marker, be it a gene, a probe or a chromosomal segment, occurring orthologously in both genomes. The edges represent the *adjacencies* between the ends of successive markers on either DNA strand in the two genomes. We colour the edges from one genome (R) red, the other (B) black. With the addition of dummy vertices (*caps*) at the endpoints of the χ_R and χ_B linear chromosomes, and dummy edges connecting each cap to one marker end, the breakpoint graph decomposes automatically into alternating colour cycles and alternating colour paths, the latter starting and terminating at caps. The number b of breakpoints – marker adjacencies in one genome not occurring in the other genome – and the number κ of cycles and paths in the breakpoint graph are the dominant components in formulae for genomic distance. Indeed, with the slightly generalized notions of breakpoint and cycle in reference [7], the most inclusive formulation of genomic distance, 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 satisfies

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

We add red caps to both ends of each chromosome in R and black caps to both ends of each chromosome in B. The breakpoint graph has $2n + 2\chi_R + 2\chi_B$ vertices corresponding to the $2n$ marker ends and the $2\chi_R$ red caps and $2\chi_B$ black caps. The adjacencies in R determine $n - \chi_R$ red edges and the adjacencies in B determine $n - \chi_B$ black edges. The caps adjacent to chromosome ends determine a further $2\chi_R$ red edges and $2\chi_B$ black edges, for a total of $n + \chi_R$ red edges and

$n + \chi_B$ black edges. Because each marker vertex is incident to exactly one red and one black edge, the graph decomposes naturally into $\chi_R + \chi_B$ alternating colour *paths* and one or more disjoint alternating colour *cycles*.

3 The Randomness Hypothesis and the Relaxation of Linearity

The key to a mathematically tractable model of random genomes is to relax the constraint that genomes B and R are composed only of χ_R and χ_B linear chromosomes. The only structure we impose, in each genome separately, is that every cap is adjacent to a non-cap vertex and every vertex is adjacent to one other vertex or to a cap, 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 and $2\chi_R$ or $2\chi_B$ caps. Since every vertex is incident to exactly one black line and one red line as before, the breakpoint graph still decomposes into disjoint alternating colour paths and cycles. Studying the statistical structure of the set of paths and cycles is facilitated by relaxing the condition that genomes B and R are composed only of χ_R and χ_B linear chromosomes, but the consequence is that the random choice of vertices defines a genome that contains not only this number of linear chromosomes, but also in general several circular plasmids. There are partial mathematical results (Theorem 1.4 in [2]) which strongly suggest that this relaxation does no violence to the probabilistic structure of the breakpoint graph.

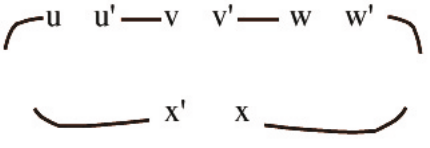

Fig. 1. Random matching gives rise to plasmids (when x and x' are two ends of the same marker) as well as linear chromosomes (when x and x' are two caps.)

For example, consider any vertex v, as in Figure 1. The chromosome containing v in genome R also contains v', the vertex at the other end of the same marker. It also contains u' and w, where u' and v are chosen by the random process to be adjacent in that genome, the vertex w adjacent to v', as well as w and u, and so on. Eventually, the two ends of construction will arrive at the two ends of a single marker, closing the circle, or two red caps, defining a linear chromosome.

Note that these considerations are independent of the properties of the alternating cycle containing v in the breakpoint graph, which would require information about *both* genomes R and B.

4 How Many Paths and Cycles?

In Section 3, we discussed the structure of the individual genomes. We now examine the structure of the breakpoint graph determined by the two genomes.

We will assume $\chi_R = \chi_B = \chi$ to simplify the presentation; the full version of this paper will give the details for the general case of different numbers of chromosomes in R and B.

4.1 The Case of No Caps – Circular Chromosomes

In [5], we pointed out that in the relaxed model for a random genome without caps, the the expected number of cycles approaches

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

where $\gamma = \lim_{n\to\infty}\left[\sum_{i=1}^{n}\frac{1}{i} - \log n\right] = 0.577\ldots$ is Euler's constant. We also cited partial mathematical results [2] and carried out simulations, both of which indicate that (2) also holds true without the relaxation, i.e. where each genome consists of a *single* DNA circle.

4.2 Linear Chromosomes

Where there are $\chi > 0$ linear chromosomes, so that each genome is assigned 2χ caps, we can take an initial view of the model as a random ordering of $2n$ "colourless" vertices and 2χ coloured vertices, where $2\chi - 1$ of the latter each represent the concatenation of two caps, one at the end of one path and one at the beginning of another path, while one vertex represents the end of the last (2χ-th) path. The vertices ordered after the latter must be on cycles rather than paths, and will be reordered in a later step.

What proportion of the vertices are on each path? As $n \to \infty$, the model becomes simply that of a random uniform distribution of 2χ points (the concatenated two-cap vertices and the final single cap) on the unit interval. The probability density of the distance between two order statistics x_k and x_{k+1}, representing the length of an alternating colour path, is the same for all $0 \le k \le 2\chi - 1$, where $x_0 = 0$:

$$q(x_{k+1} - x_k) = 2\chi(1 - (x_{k+1} - x_k))^{2\chi-1}, \tag{3}$$

with mean $1/(2\chi+1)$ and variance $\chi/[(2\chi+1)^2(\chi+1)]$. The probability density of the last order statistic, representing the sum of the lengths of all 2χ paths, is

$$q_\Sigma(x_{2\chi}) = 2\chi x_{2\chi}^{2\chi-1}, \tag{4}$$

with mean $2\chi/(2\chi+1)$ and variance $\chi/[(2\chi+1)^2(\chi+1)]$.

For the purposes of calculating genomic distance, it is convenient to introduce a certain asymmetry between the red and black genomes. Formula (1) is generally equivalent to

$$d = b + \chi - \kappa - \psi, \tag{5}$$

where ψ is the number of paths having at least one red cap. The explanation for this can be traced in [6] and [7]. Where the caps are distributed randomly, as in our model, the proportion of such paths is $\frac{3}{4}$, and the expected value of ψ is $\frac{3}{2}\chi$.

4.3 Cycles

The proportion of the genomes that is in cycles is just what is left over after the paths are calculated, in the limit $1 - x_{2\chi}$. We ignore the initial linear ordering of the these remaining vertices and instead choose two new random bipartite matchings among them, representing their configurations in genomes R and B. Then, in the limit, the number of cycles that will be constructed from this proportion of the genome is $\kappa_\chi(x_{2\chi}) = \log 2 + \frac{\gamma}{2} + \frac{1}{2}\log[(n+\chi)(1-x_{2\chi})]$, from (2). Thus, from (4), the expectation of of the random variable κ_χ is

$$\begin{aligned}
&\int_0^1 \kappa_\chi(x) q_\Sigma(x) dx \\
&= \int_0^1 \frac{2\chi}{2}\log(1-x)x^{2\chi-1}dx + \log 2 + \frac{\gamma}{2} + \frac{\log(n+\chi)}{2} \\
&= \log 2 + \frac{1}{2}\left[-\sum_{i=1}^{2\chi}\frac{1}{i} + \gamma + \log(n+\chi)\right] \\
&\sim \log 2 + \frac{1}{2}\log\frac{n+\chi}{2\chi}. \qquad (6)
\end{aligned}$$

While we will confirm the second term in (6) in subsequent sections, the $\log 2$ term is likely an extraneous result of our mathematical interpretation of the random linear chromosome model in the previous paragraph, Section 4.2, or the separate passages to the limit forms of κ_χ and q_Σ before integrating.

5 An Exact Recurrence for the Expected Number of Cycles

We build the random breakpoint graph as follows. First we construct R as a random match of the $2n$ labelled vertices and 2χ red caps, under the condition that no caps are matched to each other. Then we construct B as a random match of the $2n$ vertices, using black edges as in Figure 2[1] We will calculate the expected number of cycles completed during this construction by studying what happens as each edge in B is added,

Consider the connected components of the graph at any stage during its construction. They are either cycles, inner edges (paths incident to no cap), cap edges (paths incident to exactly one cap), or completed paths (with caps at either end). Let $N(\kappa, l, m)$ be the number of (equiprobable) ways graphs with κ

[1] We may ignore the black caps and cap edges in this process, since each step consists of joining two red vertices by a black edge.

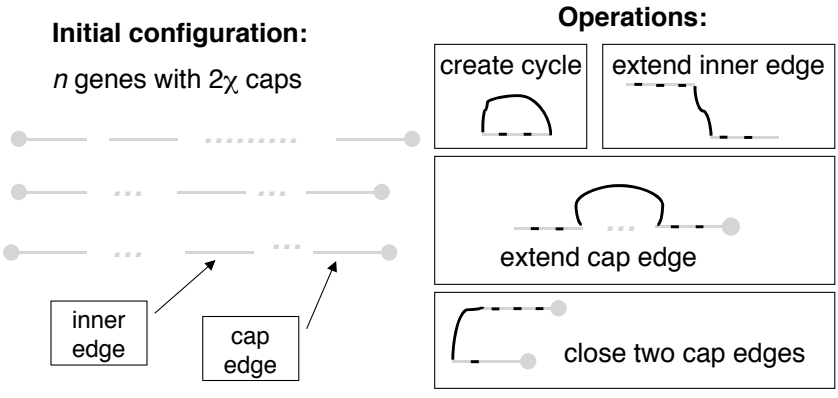

Fig. 2. Initial configuration of edges and caps. Operations of extending inner edges or cap edges and completing cycles or paths.

cycles are produced by this process starting with l inner edges and $2m$ cap edges. Initially $m = \chi$ and $l = n - \chi$, composed entirely of red edges.

There are $\binom{2m+2l}{2}$ ways of adding a black edge:

- The number of cycles can be increased by 1 if the two ends of an inner edge are connected. This decreases l by 1 and may happen in l ways.
- Two inner edges can be connected to form one inner edge. Again l decreases by 1. This can be done in $\binom{2l}{2} - l$ ways.
- One end of an inner edge is connected to a cap end. Again l decreases by 1. This can be done in $4lm$ ways
- Two cap ends are connected. Here $2m$ decreases by 2, and m decreases by 1. This can be done in $\binom{2m}{2}$ ways

Then

$$\begin{aligned} N(\kappa, l, m) = {} & lN(\kappa - 1, l - 1, m) \\ & + (\binom{2l}{2} - l + 4lm)N(\kappa, l - 1, m) \\ & + \binom{2m}{2} N(\kappa, l, m - 1), \end{aligned} \tag{7}$$

where $N(\kappa, 0, m) = 0$ for $\kappa > 0$ and $N(0, 0, m) = (2m)!/2^m$, for all positive m.

The expected number of cycles constructed during our procedure will be

$$\begin{aligned} E[\kappa(l, m)] &= \frac{\sum_\kappa \kappa N(\kappa, l, m)}{\sum_\kappa N(\kappa, l, m)} \\ &= \frac{\sum_\kappa \kappa N(\kappa, l, m)}{\prod_{i=1}^{l+m} \binom{2i}{2}} \end{aligned} \tag{8}$$

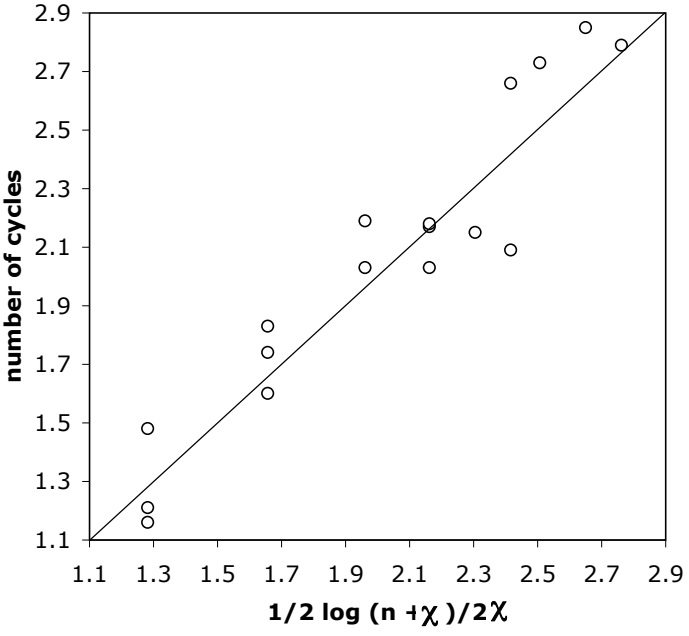

Fig. 3. Simulations for $\chi = 20$ and n ranging from 500 to 10,000. Each point represents the average of 100 pairs of random genomes.

since there are $\prod_{i=1}^{l+m} \binom{2i}{2}$ ways of adding black edges until the number of inner edges and the number of cap edges are both zero.

From (7) and (8), we find that

$$\begin{aligned} E[\kappa(l,m)] = &\frac{2m(2m-1)}{(2l+2m)(2l+2m-1)} E[\kappa(l,m-1)] \\ &+[1 - \frac{2m(2m-1)}{(2l+2m)(2l+2m-1)}] E[\kappa(l-1,m)] \\ &+\frac{2l}{(2l+2m)(2l+2m-1)}. \end{aligned} \tag{9}$$

6 Limiting Behavior of $E[\kappa(n,\chi)]$

Motivated by equation (6), if we calculate $E[\kappa(l,m)]$ for a large range of values of l and m, we find that to a very high degree of precision, the values fit

$$E[\kappa(n,\chi)] = \frac{1}{2} \log \frac{n+\chi}{2\chi}, \tag{10}$$

without the $\log 2$ term in equation (6).

Furthermore, when we simulate 100 pairs of random genomes with 20 chromosomes, for a large range of values of n, using a strictly ordered model rather than the relaxed models in Sections 4-5 above, and count the number of cycles in their breakpoint graphs, the average trend corresponds well to equation (10). This is seen in Figure 3.

Rewriting recurrence (9) for $t(l,m) = E[\kappa(l,m)]$ as

$$\begin{aligned} t(l,m) - t(l-1,m) = & \frac{2m(2m-1)}{(2l+2m)(2l+2m-1)} \\ & \times [t(l,m-1) - t(l-1,m-1) - t(l-1,m) + t(l-1,m-1)] \\ & + \frac{2l}{(2l+2m)(2l+2m-1)}, \end{aligned} \tag{11}$$

suggests that any limiting formulation for t should satisfy the equation:

$$\frac{dt}{dl} = \frac{2m(2m-1)}{(2l+2m)(2l+2m-1)}\left(\frac{dt}{dm} - \frac{dt}{dl}\right) + \frac{2l}{(2l+2m)(2l+2m-1)}. \tag{12}$$

The solutions of (12) are of form $t = \frac{1}{2}\log\frac{l+m}{2m} + C$, for any constant C. This confirms that $E[\kappa]$ is $O(\frac{1}{2}\log\frac{n+\chi}{2\chi})$. Comparison with the boundary condition $\chi = n$ in the discrete model, where each chromosome in our construction starts with two cap edges and no inner edges, so that $\kappa \equiv 0$, further confirms the computationally-established value of $C = 0$.

7 Differential Rates of Inversion and Translocation

The models we have been investigating assume that adjacencies between vertices are randomly established in one genome independently of the process in the other genome. For multichromosomal genomes, this means that the probability that any particular pair of adjacent vertices in the black genome are on the same chromosome in the red genome is of the order of χ^{-1}. This suggests that there are far fewer intrachromosomal exchanges during evolution than interchromosomal, in the approximate ratio of $\chi^{-1} : 1$, which, in the mammalian case, comes to about $0.05 : 1$, a tiny minority. In point of fact, intrachromosomal processes such as inversion represent not a minority, but a substantial majority of evolutionary events. Table 1 gives the estimated ratio of intrachromosomal events to interchromosomal events among six vertebrate species. This ratio depends on the resolution of the syntenic block evidence used to estimate the events; at finer resolutions than the 1 Mb used for the table, the ratio increases considerably. Even for the mouse-dog comparison the ratio is more than 1 at a 300 Kb resolution, while most of the other comparisons have a $2 : 1$ ratio or more.

What is the importance of this tendency for our theoretical analysis? First, there is no direct connection between the number of translocations among pairs of chromosomes and the number of adjacent vertex pairs in one genome that are on different chromosomes in the other, though they are roughly correlated; in

Table 1. Ratio of intrachromosomal events to interchromosomal ones, at a resolution of 1Mb. Calculated from estimates in [3]. Asymmetries due to construction of primary data sets in the UCSC Genome Browser and to asymmetry in the estimator used.

	human	mouse	chimp	rat	dog	chicken
human	\	1.2	-	1.6	1.7	2.9
mouse	1.3	\	1.1	2.3	0.7	1.3
chimp	15	1.4	\	-	-	-
rat	1.5	1.7	-	\	-	-
dog	1.9	0.7	-	-	\	-
chicken	4.5	1.8	-	-	-	\

some examples, a single translocation could "remove" many such edges. Nevertheless, these edges are the only obvious property of the breakpoint graph that we can model. For example, in our derivation of the recurrence (9) in Section 5, we could divide the end vertices of inner edges into χ classes corresponding to the χ chromosomes, as in Figure 4. Then by adjusting the relative probabilities of choosing intra-class edges versus inter-class edges, we can indirectly model differing proportions of inversions versus translocations. The removal of the simplifying assumption of equiprobable edge choice, however, would greatly complicate the analysis leading up to (9) and hence to (10).

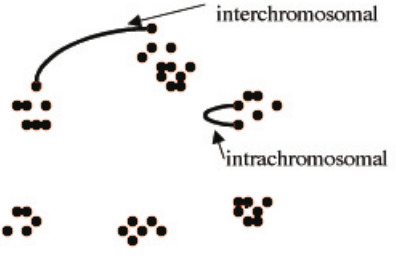

Fig. 4. Partitioning vertices into classes according to chromosomes in genome R. Two kinds of edges with differing probabilities, corresponding roughly to inversion versus translocation rates.

Leaving the theoretical aspects open, then, we propose a simulation approach to the question of the how the inversion-translocation ratio affects the breakpoint graph. For this simulation, our choice of parameters is inspired by the human-mouse comparison with 270 syntenic autosomal blocks at a resolution of 1 Mb. For simplicity we set $\chi = 20$ in both genomes. We know the genomic distance is about 240, but because of breakpoint reuse we need to use 405 operations for the algorithm to infer 240 (and there will obviously be little connection between the operations inferred by the algorithm and the operations actually producing the genomes).

We initialized the simulations with a genome having a distribution of chromosome sizes, in terms of numbers of blocks, patterned roughly after the human

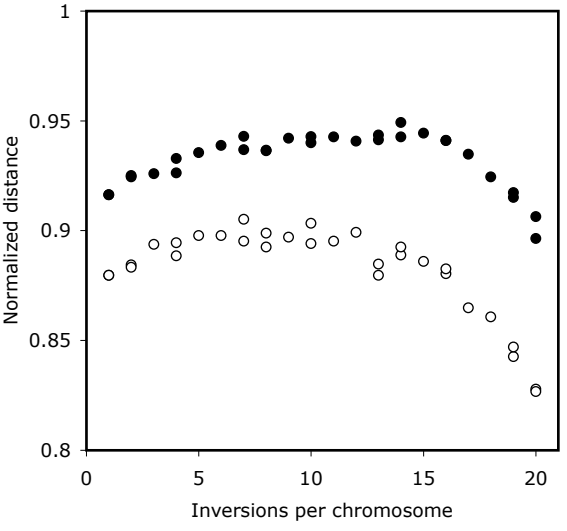

Fig. 5. Effect of changing inversion-translocation proportions. Open dots: before discarding 2-cycles. Filled dots: after discarding 2-cycles.

genome. We then used random inversions and random translocations to produce the second genome. The translocations were conditioned not to result in chromosomes smaller than a certain threshold or larger than a certain cap.

We sampled 10 runs with r inversions per chromosome and $405 - 20r$ translocations, for each $r = 1, \ldots, 20$. In Figure 5, we show that the average inferred distance (normalized by dividing by 270, the number of blocks) rises slowly with the increasing proportion of inversions, then falls precipitously as translocations became very rare. One artifact in this result is due to "two-cycles", representing genes that are adjacent in both genomes. In the breakpoint graphs of random genomes, 2-cycles occur rarely; the expected number of them has a limiting probability of $\frac{1}{2}$. And there are no 2-cycles in breakpoint graphs created from real genome sequence data. (If two syntenic blocks were adjacent and in the same orientation in both genomes, they would simply be amalgamated and treated as a single, larger, block.) Breakpoint graphs created from random inversions and translocations, however, will tend to retain some 2-cycles even after a large number of operations. It takes a very large number of operations before we can be sure that all adjacencies will be disrupted. For the sake of comparability, therefore, we should discard all two cycles and reduce n by a corresponding amount. This done in Figure 5 and it does reduce somewhat the variability of the normalized distance with respect to the inversion-translocation proportion, because the number of 2-cycles rises from about 10 per run when there are few inversions per chromosome, to more than 20 per run when there are 19 or 20 inversions per chromosome, and very few translocations.

Nevertheless, there remains two clear effects, an initial rise in the genomic distance, which will not discuss here, and a larger drop in the distance when

nearly all the operations are inversions. This drop is largely accounted for by an increase in the number of cycles from an average of 2 per run when there are less than 15 inversions per chromosome to 10 cycles per run, when there are 20 inversions per chromosome. To explain this, we observe that insofar as translocations do not interfere, the evolution of the genomes takes place as if each chromosome was evolving independently on is own. But from (10), we could then expect about $\frac{1}{2}\log(\frac{1}{2} + 270/40) \sim 1$ cycle per chromosome or 20 for the whole genome. Were the last five translocations removed from our simulation, we could expect almost that increase in the number of cycles, remembering of course, that with only 405 operations, we are still far from a complete randomization.

8 Discussion

We have continued the development of probabilistic models of random genomes, with a few to testing the statistical significance of genome rearrangement inferences. Here, we have focused on the breakpoint graphs of multichromosomal genomes and found that the expectation of the distance between two random genomes, based on equation (5), is $n + \chi - \frac{1}{2}\log\frac{n+\chi}{2\chi} - \frac{3}{2}\chi$.

A problem remains in the $\log 2$ discrepancy in equation (6). This may be dues to the approximate nature of the model or, more likely, to the premature conversion of the problem to a continuous analog.

We have suggested a new problem, how to construct breakpoint graphs reflecting differential rates of inversion and translocation. Our simulation show that the graphs are sensitive to this differential and so analytical work on this problem is important to the eventual utility of our approach in testing the significance of rearrangement inferences.

Acknowledgements

Research supported in part by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC). DS 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

1. Eriksen, N. and Hultman, A. 2004. Estimating the expected reversal distance after a fixed number of reversals. Advances of Applied Mathematics 32, 439–453.
2. Kim, J.H. and Wormald, N.C. 2001. Random matchings which induce Hamilton cycles, and Hamiltonian decompositions of random regular graphs, Journal of Combinatorial Theory, Series B 81, 20–44.
3. Mazowita, M., Haque, L. and Sankoff, D. 2006. Stability of rearrangement measures in the comparison of genome sequences. Journal of Computational Biology 13, 554–566.

4. Sankoff, D. 2006. The signal in the genomes. PLoS Computational Biology 2, e35.
5. Sankoff, D. and Haque, L. 2006. The distribution of genomic distance between random genomes. Journal of Computational Biology 13, 1005–1012.
6. Tesler, G. 2002. Efficient algorithms for multichromosomal genome rearrangements. Journal of Computer and System Sciences 65, 587–609.
7. Yancopoulos, S., Attie, O. and Friedberg, R. 2005. Efficient sorting of genomic permutations by translocation, inversion and block interchange. Bioinformatics 21, 3340 – 3346