THE DISTANCE BETWEEN RANDOMLY CONSTRUCTED GENOMES

WEI XU

Department of Mathematics and Statistics
University of Ottawa, Canada K1N 6N5
email: wxu060@uottawa.ca

In this paper, we study the exact probability distribution of the number of cycles c in the breakpoint graph of two random genomes with n genes or markers and χ_1 and χ_2 linear chromosomes, respectively. The genomic distance d between the two genomes is $d = n - c$. In the limit we find that the expectation of d is $n - \frac{2\chi_1\chi_2}{2\chi_1+2\chi_2-1} - \frac{1}{2}\ln\frac{n+\min(\chi_1,\chi_2)}{\chi_1+\chi_2}$.

1. Introduction

The study of genome rearrangements has developed a sophisticated technology for inferring a minimizing sequence of operations necessary to transform one genome into another, where the genomes are represented by signed permutations on $1, \cdots, n$ and the operations are modeled on the biological processes of inversion, reciprocal translocation, chromosome fusion and fission, transposition of chromosomal segments, excision and reintegration of circular chromosomal segments, among others. Once these inferences are made, however, there is a need for some way to statistically validate both the inferences and the assumptions of the evolutionary model.

Our approach has been to see to what extent there is an signal remaining in the comparative structure of the two genomes, or whether evolution has largely scrambled the order of each one with respect to the other, in terms of the evolutionary model assumed. This has led to the study of completely scrambled, i.e., randomized, genomes as a null baseline for the detection of a evolutionary signal. Insofar as a pair of genomes retain some evidence of evolutionary relationship, this should be detectible by contrast to randomized genomes. In previous papers, we have worked out the statistical properties of random genomes consisting of one or more circular chromosomes, [1] and those of two random genomes containing the same number c of linear chromosomes.[2] The latter paper concentrated on showing that the number of circular chromosomes inevitably associated with random linear chromosomes is very small with realistic numbers of chromosomes. It only included a rough estimation of the statistical properties of the linear chromosomes.

The present paper introduces a new way of representing the comparison of linear genomes, requiring only a single source/sink vertex in the breakpoint graph of the two genomes, instead of the numerous "chromosomal caps" used in other treatments. This facilitates a more rigorous treatment of the case of linear chromosomes, including the more realistic situation where the number of linear chromosomes may be different (χ_1 and χ_2)

2

in the two genomes being compared.

2. Genome rearrangement with linear chromosomes

In our framework, each genome consists of n markers (genes, chromosomal segments, etc.), divided among a number of disjoint chromosomes. We fix the number of linearly ordered chromosomes, but in our construction of random genomes we will permit some additional, circularly ordered, chromosomes as well. In graph-theoretical terms, we usually represent each marker by two distinct vertices, marking the beginning and end of the marker, respectively. We call all of these *inner vertices*. For each linear chromosome, two extra vertices, named *caps* are added to represent the ends of the chromosome. In comparing two genomes containing different numbers χ_1 and χ_2 of linear chromosomes, we equalize their numbers at $\chi = \max(\chi_1, \chi_2)$ by adding an appropriate number of null chromosomes, each of which consists only of two caps, to one of the genomes.

2.1. *The Breakpoint Graph*

When two genomes, say a red one and a black one, containing the same n markers, are compared, we use red edges to connect the the nearest vertices of two adjacent markers according to their order in the red genome – this may be the end of one mark and the beginning of the other, or two ends, or two beginnings, depending on the orientation or "strandedness" of the markers on the chromosome. The first and last inner vertices are connected to caps. Each cap may only be connected to one inner vertex. We also connect the two caps of any null chromosome in the red genome by a red edge. Similarly, we use black edges to connect the vertices and caps in the black genomes. There are thus $2n$ inner vertices, 2χ caps, $n + \chi$ red edges and $n + \chi$ black edges in the graph.

Since each vertex is connected to one red and one black edge (one adjacency in each genome), a 2-regular graph is formed. A 2-regular graph always can be decomposed into number of cycles c, and in our bicoloured graph, the edge colours alternate around each cycle. Yancopoulos, Attie and Friedberg[3] showed that the edit distance d is related to the number of cycles c by

$$d = n + \chi - \max c, \tag{1}$$

when block interchanges (each counting as two operations) are allowed besides inversions and reciprocal translocations. The number of cycles depends on which red chromosome and which black chromosome are incident to the same cap, a choice which is left free in the graph definition. The maximal number of cycles in equation (1) refers to the optimal choice of this cap assignment. We refer to this particular graph as the breakpoint graph of the two genomes.
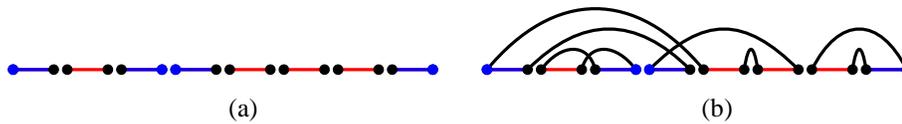
Figure 1.   The construction of a random breakpoint graph. We start with the red genome, represented by a set of cap edges (in blue) and a set of inner edges (in red), and add the black edges randomly, one by one, until every vertex is connected by one black edge. In (b) there are 3 cycles. Caps denoted by blue dots and inner vertices by black ones.

2.2. *Random Genomes*

Were we to construct genomes by successively adding markers or caps in random order, it would be very difficult to say anything precise about the breakpoint graph, because the linearity condition on chromosomes induces great complexity to the events whose probabilities we wish to calculate. Instead, we introduce the randomness directly in the construction of the breakpoint graph, leading to simple expressions for probabilities of the sizes and numbers of cycles. This simplicity comes at a cost, however, since the construction of a random genome at the level of the breakpoint graph does not exclude some circular chromosomes. As we shall mention later, there is good reason to believe that this feature does not affect our results on the limits of expectations.

To obtain two genomes randomized with respect to each other, it suffices to fix the gene order in one of them, say the red genome, and to introduce randomness into the black genome only. Because we are interested in calculations pertaining to the breakpoint graph, we simply postulate that at each step a black edge may be added to connect any two inner vertices that are not already incident to black edges. We do not at this stage really connect caps to inner vertices using black edges, because these edges are implicitly determined by the cycle optimization procedure applied to the rest of the graph. Thus we start with $2n + 2\chi$ vertices (inner vertices and caps), with red edges connected. We distinguish between two kinds of edges: 2χ *cap edges* incident to a cap and $n - \chi$ *inner edges* not incident to a caps. To construct the random breakpoint graph, we connect two inner vertices at random by a black edge until every vertex is incident to a black edge. Note that in randomly adding black edges we are not guaranteed to end up with linear chromosomes, since there is the possibility that the black genome so constructed will contain one or more circular chromosomes, with no caps. As χ becomes large, the number of such circles and the number of markers in them, will be small. Nevertheless this possibility is not part of the original problem involving two random genomes with linearly ordered chromosomes. Fortunately, partial mathematical results indicate that in the limit, the possible presence of circular chromosomes does not affect the probability structure of the breakpoint graph[4].

4

2.3. *Cap Optimization*

In the procedure of cap optimization, the breakpoint graph is decomposed into cycles and 2χ paths (whose two ends are caps or inner vertices incident to only one cap edge). The ψ_{HO} *homogenous paths* terminate with caps via two red edges (type 1) or with two inner vertices (type 2), with an equal number of the two types, and the ψ_{HE} *heterogenous paths* end with one cap and one inner vertex. The optimization principle developed by Hannenhalli and Pevzner[5] and Tesler[6], comes down to, in the reformulation by Yancopoulos *et al.*[3], to the addition of two black edges joining one homogenous path of type one to another homogeneous path of type 2 to form a cycle and the addition of a single black edge to each heterogeneous path to form a cycle. It can be seen that the maximized cap cycle number ψ is

$$\max \psi = \chi + \frac{1}{2}\psi_{HE} \qquad (2)$$

2.4. *The* Flower *Representation*

To facilitate the construction of the random breakpoint graph, including the cycle optimization, we abandon the regular graph representation and introduce a modified model as follows.
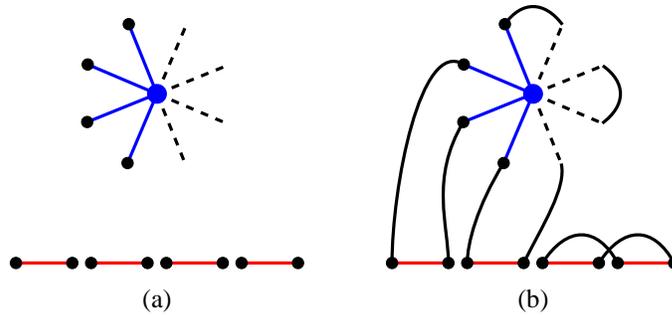

Figure 2.   The illustration of the modified model. At the initial state (a), all the caps have been merged into one source/sink vertex C. The dashed black edges are reserved for the 2χ black cap edges to be added later. At the end (b), all the cap edges should be connected via inner edges, except for some that are composed of two cap edges or a single edge with C at both ends. The rest of the inner edges form the inner cycles. In the figure, two homogenous paths, two heterogenous paths and one inner cycle are depicted.

We replace all the caps by a single source/sink vertex C. Then we may portray the cap edges as distributed around C as in Figure 2, while the inner edges are unaffected. In Figure 2(a), there are 2χ red cap edges and the same number of dashed edges incident to C indicating where the black cap edges will eventually connect. Some same-coloured pairs of these cap edges may represent null chromosomes. The construction proceeds by adding

black edges one by one at random as detailed in the next section, and terminates when a complete structure as in Figure 2(b) is achieved.

The cycles are of two sorts, those in the flower structure, named *cap cycles* and the rest, *inner cycles*. In the next section, recurrence equations will derived for both kinds. Note that each "petal" of the flower, connected to the source/sink vertex, represents a path, either a homogenous or heterogeneous. The cap cycles are not explicitly depicted in the graph. Their total number is determined by the capping optimization formula.

3. The Recurrence Equations

3.1. *The Number of Heterogenous Paths ψ_{HE}*

From the cap optimization principle, the number of cap cycles should be equal to $\chi + \frac{1}{2}\psi_{HE}$.

During the construction, at each step it suffices to keep track only of the number of extended red cap edges, where this includes paths with C connected to a red edge at one end and a red edge at the other, the number of extended black cap edges, where this includes either dashed edges of paths with C connected to a black edge at one end and a red edge at the other.

We start from a general situation where there are r extended red cap edges and s extended black cap edges. The problem is denoted as (r, s).

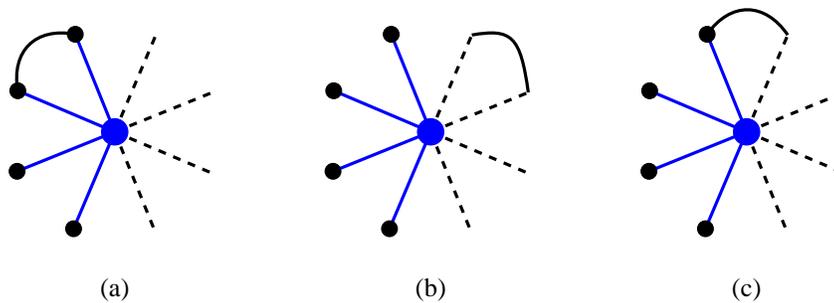| (a) | (b) | (c) |

Figure 3.   The three possible ways of completing a cap path. Homogenous paths are shown in (a) and (b) and a heterogenous path in (c).

At each step, one black edge is added, connecting two extended red cap edges, two dashed or extended black cap edges or one extended red cap edge and one dashed or extended black cap edge. Once a path forms, the total number of extended paths (i.e., the edges that remain to be connected) decreases by 2. The three possible ways of adding the black edge lead to the smaller problems $(r-2, s)$, $(r, s-2)$, $(r-1, s-1)$, respectively. The numbers of ways of doing each are $\binom{r}{2}$, $\binom{s}{2}$ and rs, respectively. Only in the last situation is a heterogenous path completed. Denote $\mathbf{n}(\psi_{HE}, r, s)$ as the total number of ways to get a

6

breakpoint graph with ψ_{HE} heterogenous paths for the (r, s) problem. Since each problem with size (r, s) can be constructed from three smaller problems of sizes $(r - 2, s)$, $(r, s - 2)$ and $(r - 1, s - 1)$, respectively, we have the recurrence :

$$\mathbf{n}(\psi_{HE}, r, s) = \binom{r}{2} \mathbf{n}(\psi_{HE}, r-2, s) + \binom{s}{2} \mathbf{n}(\psi_{HE}, r, s-2) + rs\, \mathbf{n}(\psi_{HE}-1, r-1, s-1) \tag{3}$$

Denote by $\bar{\psi}_{HE}(r, s)$ the average number of heterogenous paths in the breakpoint graph for (r, s), defined as:

$$\bar{\psi}_{HE}(r, s) = \frac{\sum_{\psi_{HE}=0}^{min(r,s)} \mathbf{n}(\psi_{HE}, r, s)\, \psi_{HE}}{\sum_{\psi_{HE}=0}^{min(r,s)} \mathbf{n}(\psi_{HE}, r, s)}$$

$$= \frac{\sum_{\psi_{HE}=0}^{min(r,s)} \mathbf{n}(\psi_{HE}, r, s)\, \psi_{HE}}{\prod_{i=0}^{\frac{r+s}{2}} \binom{r+s-2i}{2}}$$

where $\sum_{\psi_{HE}=0}^{min(r,s)} \mathbf{n}(\psi_{HE}, r, s) = \prod_{i=0}^{\frac{r+s}{2}} \binom{r+s-2i}{2}$ is the total number of ways to construct the breakpoint graph.

By summing over equation (3), we get the recurrence equation for the average number of heterogenous paths.

$$\bar{\psi}_{HE}(r, s) = \frac{\binom{r}{2}\bar{\psi}_{HE}(r-2, s) + \binom{s}{2}\bar{\psi}_{HE}(r, s-2) + rs\bar{\psi}_{HE}(r-1, s-1) + rs}{\binom{r+s}{2}} \tag{4}$$

Equation (4) has a probabilistic interpretation, since (r, s) can be decomposed into $(r - 2, s)$, $(r, s - 2)$ and $(r - 1, s - 1)$ with probabilities $\frac{r(r-1)}{(r+s)(r+s-1)}$, $\frac{s(s-1)}{(r+s)(r+s-1)}$ and $\frac{2rs}{(r+s)(r+s-1)}$, respectively.

3.2. *The Number of Inner Cycles*

The number of the inner cycles depends on the number of inner edges not used by the paths. Suppose we start with $2m$ extended cap edges (the extended red cap edges and the dashed or extended black edges) and l inner edges, which it will be convenient to denote (m, l).[a] In the random construction of the breakpoint graph, each addition of one black edge can lead to four different situations:

(1) two cap edges are connected – there are $\binom{2m}{2}$ ways of doing this – and the size of the problem becomes $(m - 1, l)$
(2) one cap edge and one inner edge are connected – there are $4ml$ ways of doing this – and the size of the problem becomes $(m, l - 1)$

---

[a]Rather than $(2m, l)$.

(3) two different inner edges are connected – there are $2l(l-1)$ ways of doing this – and the size of the problem becomes $(m, l-1)$

(4) the two ends of the same inner edges are connected – there are l ways of doing this. The size of the problem becomes $(m, l-1)$ and the number of inner cycles increases by one.
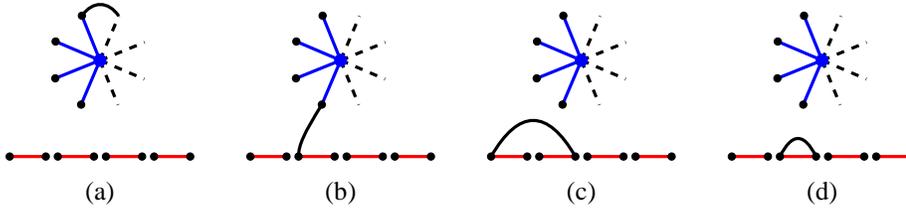

Figure 4.   The four possible ways to build a black edge in counting the inner cycles. Two cap edges are connected (a); one cap edge and one inner edge are connected (b); two different inner edges are connected (c); the two ends of the same inner edge are connected (d). And only in the last case, one inner cycle is formed.

Denote by $\mathbf{n}(\kappa, m, l)$ the number of ways to get a breakpoint graph with κ inner cycles for a (m, l) problem. Similarly define $\bar{\kappa}(m, l)$ as the average number of inner cycles for the problem (m, l)

$$\bar{\kappa}(m, l) = \frac{\sum_{\kappa=0}^{l} \mathbf{n}(\kappa, m, l)\,\kappa}{\sum_{\kappa=0}^{l} \mathbf{n}(\kappa, m, l)}$$

$$= \frac{\sum_{\kappa=0}^{l} \mathbf{n}(\kappa, m, l)\,\kappa}{\prod_{i=0}^{m+l} \binom{2m+2l-2i}{2}}$$

We then get the corresponding recurrences

$$\mathbf{n}(\kappa, m, l) = \binom{2m}{2} \mathbf{n}(\kappa, m-1, l) + [4ml + 2l(l-1)]\,\mathbf{n}(\kappa, m, l-1)$$
$$+ l\,\mathbf{n}(\kappa-1, m, l-1) \tag{5}$$

$$\bar{\kappa}(m, l) = \frac{\binom{2m}{2}\bar{\kappa}(m-1, l) + [4ml + 2l(l-1)]\,\bar{\kappa}(m, l-1) + l\,\bar{\kappa}(m, l-1) + l}{\binom{2m+2l}{2}} \tag{6}$$

Equation (6) also has a probabilistic interpretation, associating the probabilities $\frac{2m(2m-1)}{(2m+2l)(2m+2l-1)}$, $\frac{8ml}{(2m+2l)(2m+2l-1)}$, $\frac{4l(l-1)}{(2m+2l)(2m+2l-1)}$ and $\frac{2l}{(2m+2l)(2m+2l-1)}$ with the four possible smaller problems.

8

4. The solution to the problems

4.1. *The cap cycles*

The recurrence equations (4) and (6) enable rapid calculation of $\bar{\psi}_{HE}$ and $\bar{\kappa}$, but there is no easy way to convert them into a closed form solution.

We can, however, deduce these quantities through another combinatoric approach. The total number of ways to form any kind of flower structure is $\prod_{i=0}^{\frac{r+s}{2}} \binom{r+s-2i}{2}$. The ways to form a result with $2\psi_{HE}$ heterogenous paths (which should be always even) is

$$\mathbf{n}(2\psi_{HE}, r, s) = \binom{\frac{r+s}{2}}{2\psi_{HE}} \frac{(r)!}{(r - 2\psi_{HE})!} \frac{(s)!}{(s - 2\psi_{HE})!} \binom{\frac{r+s}{2} - 2\psi_{HE}}{\frac{r}{2} - \psi_{HE}}$$

$$\prod_{i=0}^{\frac{r+s}{2}} \binom{r - \psi_{HE} - 2i}{2} \prod_{i=0}^{\frac{r+s}{2}} \binom{s - \psi_{HE} - 2i}{2}$$

$$= \frac{(r)!(s)!(r+s)!}{(\frac{r+s}{2})!} \frac{4^{\psi_{HE}}}{(2\psi_{HE})!(\frac{r}{2} - \psi_{HE})!(\frac{s}{2} - \psi_{HE})!} \tag{7}$$

Averaging over $\mathbf{n}(2\psi_{HE}, r, s)$ and rewriting $r = 2\chi_1$ and $r = 2\chi_2$, we get

$$\bar{\psi}_{HE}(\chi_1, \chi_2) = \frac{4\chi_1\chi_2}{2\chi_1 + 2\chi_2 - 1} \tag{8}$$

So the average number of cap cycles is

$$\psi = \max(\chi_1, \chi_2) + \frac{2\chi_1\chi_2}{2\chi_1 + 2\chi_2 - 1} \tag{9}$$

When $\chi_1 = \chi_2 = \chi$, it becomes $\chi + \frac{2\chi^2}{4\chi - 1}$, approaching 1.5χ as χ becomes large, confirming a result which we have previously derived in another way.[2]

4.2. *The Inner Cycles*

In the flower structure where the numbers of chromosomes are equal, suppose we traverse all the edges, starting with a black cap edge, and each time we visit C, we choose an outgoing edge of colour different from the incoming edge. This will order the edges as in Figure 5(a). The last edge will be a red cap edge and C will be the last vertex. We then add the edges in the inner cycles to the right of the flower structure edges.

We define the position x of an edge, as the number of edges to the left of, and including, that edge. We assume there are χ linear chromosomes in each genome. So the smallest value possible for x is 1 and the largest one is $n - \chi$. The 2χ cap edges occupy random positions in the sequence. The constraints on the model are that the last cap edge should have C on its right, the ith cap edge can only distributed from $x_{i-1} + 1$ to $n + \chi - (2\chi - i) = n - \chi + i$. Only the inner edges to the right of the variable $x_{2\chi}$, the position of the last cap edge, are in inner cycles. Once we know the distribution of $x_{2\chi}$, we then use the formula[1] for the expected number of cycles in circular genomes to calculate the number of inner cycles.
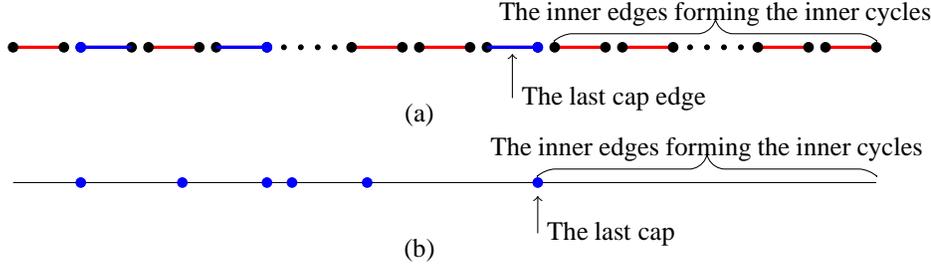
Figure 5.   The exact model (a) for counting the inner cycle number and the approximate model (b). Model (a) is discrete. The cap in the last cap edge should be in the right side in order to correspond to the flower structure. Model (b) is a continuous approximation of model (a) when the number of inner edges are large enough and has no constraint on the last cap edge.

When n becomes large, we may define a continuous approximation to this construction. The x_i become the order statistics of χ uniformly distributed points on $(0, n + \chi)$. Using the distribution for the position of the x_χ, we find the expected number of inner cycles is

$$c(m = 2\chi, l = n - \chi) = \frac{1}{2} \ln \frac{\chi + n}{2\chi} + B \qquad (10)$$

where B is some constant.[2]

Note that equation (10) is the asymptotic solution of equation (6), with $m = 2\chi, l = n - \chi$. B can be found from an initial condition: when $n = \chi$, then there are no inner edges and hence no inner cycles. So $\frac{1}{2} \ln \frac{\chi + \chi}{2\chi} + B = 0$, i.e., $B = 0$. In numerical comparison as well, the equation $c = \frac{1}{2} \ln \frac{\chi + n}{2\chi}$ confirms the recurrence equation (6).

4.3. *Two Genomes Having Different Numbers of Linear Chromosomes*

Suppose the two genomes being compared have χ_1 and χ_2 linear chromosomes, respectively. We have already found the formula for the cap cycles which is $\max(\chi_1, \chi_2) + \frac{2\chi_1\chi_2}{2\chi_1 + 2\chi_2 - 1}$. For the number of inner cycles, the approximate model only deals with the case where $\chi_1 = \chi_2 = \chi$. But that solution is also the asymptotic solution for the recurrence equation (6), which depends only on m and l. We can thus substitute the values for m and l in the case of unequal number of linear chromosomes. Note that in the case of equality, $m = 2\chi$ and $l = n - \chi$ and in the unequal case $m = \chi_1 + \chi_2$ and $l = n - \max(\chi_1, \chi_2)$.

$$\begin{aligned}
c &= \frac{1}{2} \ln \frac{\chi + n}{2\chi} = \frac{1}{2} \ln \frac{2\chi + n - \chi}{2\chi} = \frac{1}{2} \ln \frac{m + l}{m} \\
&= \frac{1}{2} \ln \frac{\chi_1 + \chi_2 + n - \max(\chi_1, \chi_2)}{\chi_1 + \chi_2} \\
&= \frac{1}{2} \ln \frac{n + \min(\chi_1, \chi_2)}{\chi_1 + \chi_2} \qquad (11)
\end{aligned}$$

10

Hence in the limit the total number of cycles is

$$c = \max(\chi_1, \chi_2) + \frac{2\chi_1\chi_2}{2\chi_1 + 2\chi_2 - 1} + \frac{1}{2}\ln\frac{n + \min(\chi_1, \chi_2)}{\chi_1 + \chi_2} \qquad (12)$$

And the genomic distance d is

$$d = n - \frac{2\chi_1\chi_2}{2\chi_1 + 2\chi_2 - 1} - \frac{1}{2}\ln\frac{n + \min(\chi_1, \chi_2)}{\chi_1 + \chi_2}. \qquad (13)$$

5. Conclusion

The mathematical essence of the question with two genomes with linear chromosomes, is the number of the cycles in the 2-regular breakpoint graph whose vertices consist of a set of labeled vertices and another set of interchangeable caps vertices. We have shown that collapsing all the caps to a single source/sink facilitates the optimal capping problem as well as the calculation of cycle expectations.

The final result equation (13) can be applied to the comparison of two genomes with the same or different number of linear chromosomes plus any number of circular chromosomes. This is true under the condition that inversions, translocations and block interchanges are the mechanism of genomic rearrangement, where the latter count as if they were each two operations.[3]

References

1. Sankoff, D. and Haque, L. 2006. The distribution of genomic distance between random genomes. Journal of Computational Biology 13, 1005–1012.
2. Xu, W., Zheng, C. and Sankoff, D. 2006. Paths and cycles in breakpoint graphs of random multichromosomal genomes. Proceedings of RECOMB Satellite Conference on Comparative Genomics 2006, G. Bourque and N. El-Mabrouk, eds., Lecture Notes in Computer Science 4205. Heidelberg: Springer, 51–62.
3. Yancopoulos, S., Attie, O. and Friedberg, R. 2005. Effi cient sorting of genomic permutations by translocation, inversion and block interchange. Bioinformatics 21, 3340 – 3346
4. 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.
5. Hannenhalli, S. and Pevzner, P.A. 1995. Transforming men into mice (polynomial algorithm for genomic distance problem. Proceedings of the IEEE 36th Annual Symposium on Foundations of Computer Science. 581–92.
6. Tesler, G. 2002. Effi cient algorithms for multichromosomal genome rearrangements. Journal of Computer and System Sciences 65, 587–609.