Genome Rearrangement

Rick Durrett[1]

Dept. of Math., Cornell U., Ithaca NY, 14853 rtd1@cornell.edu

Genomes evolve by chromosomal fissions and fusions, reciprocal translocations between chromosomes, and inversions that change gene order within chromosomes. For more than a decade biologists and computer scientists have studied these processes by parsimony methods, i.e., what is the minimum number of events needed to turn one genome into another? We have recently begun to develop a stochastic approach to this and related questions, which has the advantage of producing confidence intervals for estimates and allowing tests of hypotheses concerning mechanisms.

1 Inversions

We begin with the simplest problem of the comparison of two chromosomes where the genetic material differs only due to a number of inversions that have reversed the order of chromosomal segments. This occurs for mitochondrial DNA, mammalian X chromosomes and chromosome arms in some insect species (e.g., *Drosophila* and *Anopheles*). To explain the problem, we begin with an example. The relationship between the human and mouse X chromosomes may be given by a signed permutation (see Figure 2 in Pevzner and Tesler 2003)

$$1 \quad -7 \quad 6 \quad -10 \quad 9 \quad -8 \quad 2 \quad -11 \quad -3 \quad 5 \quad 4$$

In words if we look at the positions of genes then in the first segment of each chromosome the genes appear in the same order. The genes in the second segment of the mouse X chromosome are the same as those in the 7th segment of the human X chromosome but the order is reversed, etc.

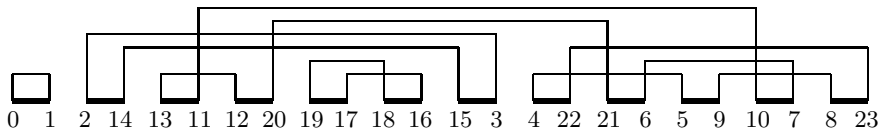
Hannenhalli and Pevzner (1995a) developed a polynomial algorithm for computing the inversion distance between chromosomes, i.e., what the smallest number of inversions needed to transform one chromosome into another? The first step in preparing to use the HP algorithm is to double the markers.

When segment i is doubled we replace it by two consecutive numbers $2i-1$ and $2i$, e.g., 6 becomes 11 and 12. A reversed segment $-i$ is replaced by $2i$ and $2i-1$, e.g., -7 is replaced by 14 and 13. The doubled markers use up the integers 1 to 22. To these we add a 0 at the front and a 23 at the end. Using commas to separate the ends of the markers we can write the two genomes as follows:

$$\begin{aligned} \text{mouse} \quad & 0, 1\,2, 14\,13, 11\,12, 20\,19, 17\,18, 16\,15, \\ & 3\,4, 22\,21, 6\,5, 9\,10, 7\,8, 23 \\ \text{human} \quad & 0, 1\,2, 3\,4, 5\,6, 7\,8, 9\,10, 11\,12, 13\,14, \\ & 15\,16, 17\,18, 19\,20, 21\,22, 23 \end{aligned}$$

The next step is to construct the breakpoint graph which results when the commas are replaced by edges that connect vertices with the corresponding numbers. In the picture we write the vertices in their order in the mouse genome. Commas in the mouse order become thick lines (black edges), while those in the human genome are thin lines (gray edges).

Fig. 1. Breakpoint graph for human-mouse X chromosome comparison



Each vertex has one black and one gray edge so its connected components are easy to find: start with a vertex and follow the connections in either direction until you come back to where you start. In this example there are five cycles:

$$0-1-0 \qquad 2-14-15-3-2 \qquad 4-22-23-8-9-5-4$$
$$19-17-16-18-19 \qquad 13-11-10-7-6-21-20-12-13$$

To compute a lower bound for the distance now we first count the number of commas seen when we write out one genome. In this example that is 1 plus the number of segments ($n = 11$). We then subtract the number of connected components, $c(n)$, in the breakpoint graph. This is a lower bound on the distance since any inversion can at most reduce this quantity by 1, and it is 0 when the two genomes are the same. In symbols,

$$d(\pi) \geq n + 1 - c(\pi) = 12 - 5 = 7$$

In general the distance between genomes can be larger than the lower bound from the breakpoint graph. There can be obstructions called *hurdles* that can prevent us from decreasing the distance and hurdles can be intertwined in a *fortress of hurdles* that takes an extra move to break. (See Hannenhalli and Pevzner 1995a.) If π is the signed permutation that represents the relative order and orientation of segments in the two genomes then

$$d(\pi) = n + 1 - c(\pi) + h(\pi) + f(\pi)$$

where $h(\pi)$ is the number of hurdles and $f(\pi)$ is the indicator of the event π is a fortress of hurdles.

Fortunately the complexities associated with hurdles rarely arise in biological data sets. Bafna and Pevzner (1995) considered the inversion distance problem for 11 chloroplast and mitochondrial data sets and in all cases they found that the distance was equal to the lower bound. We can verify that 7 is the minimum distance for the human-mouse comparison by constructing a sequence of 7 moves that transforms the mouse X chromosome into the human order. There are thousands of solutions, so we leave this as an exercise for the reader. Here are some hints: (i) To do this it suffices to at each step choose an inversion that increases the number of cycles by 1. (ii) This never occurs if the two chosen black edges are in different cycles. (iii) If the two black edges are in the same cycle and are (a, b) and (c, d) as we read from left to right, this will occur unless in the cycle minus these two edges a is connected to d and b to c, in which case the number of cycles will not change. For example in the graph above an inversion that breaks black edges 19-17 and 18-16 will increase the number of cycles but the one that breaks 2-14 and 15-3 will not. See Section 5.2 of Durrett (2002) or Chapter 10 of Pevzner (2000) for more details.

Ranz, Segarra, and Ruiz (1997) did a comparative study of chromosome 2 of *Drosophila repleta* and chromosome arm 3R of *D. melanogaster.* If we number the 26 genes that they studied according to their order on the *D. repleta* chromosome then their order on *D. melanogaster* is given by

12 7 4 *2 3 21 20* 18 1 13 9 16 6 14 *26 25 24* 15 *10 11* 8 5 *23 22* 19 17

where we have used italics to indicate adjacencies that have been preserved. Since the divergence of these two species, this chromosome region has been subjected to many inversions. Our first questions is: How many inversions have occurred? To answer this question we need to formulate and analyze a model. Before we do this, the reader should note that in contrast to the human-mouse comparison, here we do not have enough markers to determine the relative orientation of the segments, so we have an unsigned permutation.

n-inversion chain. Consider n markers on a chromosome, which we label with $1, 2, \ldots n$, and can be in any of the $n!$ possible orders. To these markers we add two others: one called 0 at the beginning and one called $n + 1$ at the

end. Finally for convenience of description we connect adjacent markers by edges. For example when $n = 7$ the state of the chromosome might be

$$0 - 5 - 3 - 4 - 1 - 7 - 2 - 6 - 8$$

In biological applications the probability of an inversion in a given generation is small so we will formulate the dynamics in continuous time. The labels 0 and $n + 1$ never move. To shuffle the others, at times of a rate one Poisson process we pick two of the $n + 1$ edges at random and invert the order of the markers in between. For example, if we pick the edges $5 - 3$ and $7 - 2$ the result is

$$0 - 5 - 7 - 1 - 4 - 3 - 2 - 6 - 8$$

If we pick $3 - 4$ and $4 - 1$ in the first arrangement there is no visible change. However, allowing this move will simplify the mathematical analysis and only amounts to a small time change of the dynamics in which one picks two markers $1 \le i < j \le n$ at random and reverses the segment with those endpoints.

It is clear that if the chromosome is shuffled repeatedly then in the limit all of the $n!$ orders for the interior markers will have equal probability. The first question is how long does it take for the marker order to be randomized. To explain the answer, we recall that the total variation distance between two distributions μ and ν is $\sup_A |\mu(A) - \nu(A)|$.

Theorem 1. *Consider the state of the system at time $t = cn \ln n$ starting with all markers in order. If $c < 1/2$ then the total variation distance to the uniform distribution ν goes to 1 as $n \to \infty$. If $c > 2$ then the total variation distance goes to 0.*

For a proof see Durrett (2003). There is a gap between the upper bound and the lower bound, but on the basis of other results it is natural to guess that the lower bound is right, i.e., convergence to equilibrium takes about $(n \ln n)/2$ shuffles. When $n = 26$, this is 42.3. Consequently, when the number of inversions is large (in the example more than 40) the final arrangement is almost independent of the initial one and we do not expect to be able to accurately estimate the actual number of invrsions.

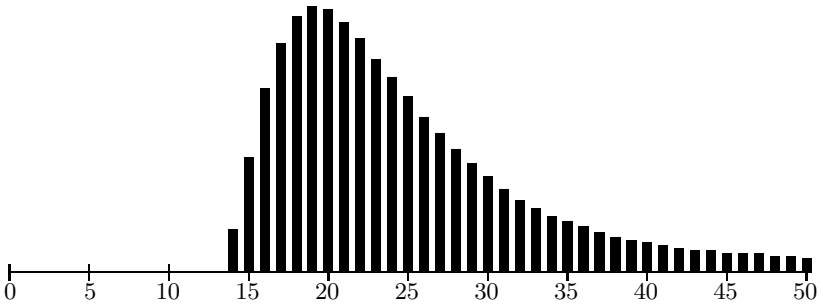
While Theorem 1 may be interesting for card shuffling algorithms, its conclusion does not tell us much about the number of inversions that occurred in our data set. To begin to investigate this question, we note that there are 6 conserved adjacencies. This means that at least $27 - 6 = 21$ edges have been disturbed, so at least 11 inversions have occurred. Biologists often use this easy to compute estimate, which is called the *breakpoint distance*. However, this lower bound is usually not sharp. In this example it can be shown that at least 14 inversions are needed to put the markers in order.

The maximum parsimony solution is 14 but there is no guarantee that nature took the shortest path between the two genomes. York, Durrett, and

Nielsen (2002) have introduced a Bayesian approach to the problem of inferring the history of inversions separating two chromosomes. They assume that the differences between the gene arrangements in two species come from running the n-inversion chain for some unknown time λ. Given a number of inversions ℓ, let $\pi_0, \pi_1, \ldots, \pi_\ell$ be the proposed evolutionary sequence that connects the two genomes, with each π_k differing from the previous one by one inversion. Let Ω be the set of all such sequences (of any length) and X be a generic member of Ω.

Let D (for data) be the marker order in the two sampled genomes. The Markov chain Monte Carlo method of York, Durrett, and Nielsen (2002) consists of defining a Markov chain on $\Omega \times [0, \infty)$ with stationary density $P(X, \lambda|D)$. They alternate updating λ and X. First a new λ is chosen according to $P(\lambda|X, D)$, then a new path is produced by choosing a segment to cut out of the current path and then reconnecting the two endpoints. In generating the new path they use the graph distance $n + 1 - c(\pi)$ as a guide and prefer steps that reduce the distance. We refer the reader to the cited paper for more details. Figure 2 show a picture of the posterior distribution of the number of inversions for the Ranz, Segarra, and Ruiz (1997) data set. Note that this density assigns a small probability to the shortest path (with length 14) and has a mode at 19.

Fig. 2. Posterior distribution of inversions for Drosophila data.



An alternative and simpler approach to our question comes from considering $\phi(\eta) =$ the number of conserved edges minus 2. Subtracting 2 makes ϕ is orthogonal to the constant eigenfunction. A simple calculation shows that ϕ is an eigenfunction of the chain with eigenvalue $(n-1)/(n+1)$. In our case $n = 26$ and $\phi = 4$ so solving

$$27\left(\frac{25}{27}\right)^m = 4 \quad \text{gives} \quad m = \frac{\ln(4/27)}{\ln(25/27)} = 24.8$$

gives a moment estimate of the number of inversions which seems consistent with the distribution in Figure 2.

Ranz, Ruiz, and Casals (2001) enriched the comparative map so that 79 markers can be located in both species. Again numbering the markers on the *D. repleta* chromosome by their order on *D. melanogaster* we have:

36	*37*	17	40	*16*	*15*	*14*	63	*10*	*9*	55	28
13	51	22	79	39	70	66	*5*	*6*	*7*	35	64
33	*32*	*60*	*61*	18	65	62	12	1	11	23	20
4	52	68	29	48	3	21	53	8	43	72	*58*
57	*56*	19	49	34	59	30	77	31	67	44	2
27	38	50	*26*	*25*	76	69	41	24	75	71	78
73	47	54	45	74	42	46					

The number of conserved adjacencies (again indicated with italics) is 11 so our moment estimate is

$$m = \frac{\ln(9/80)}{\ln(78/80)} = 86.3$$

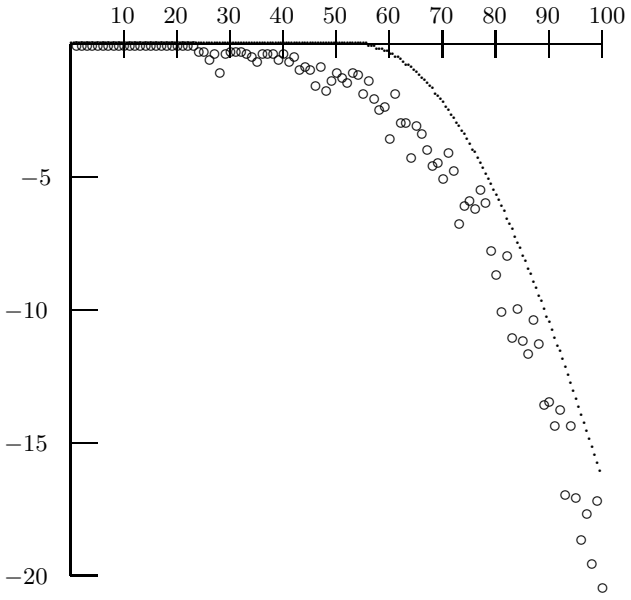
This agrees with the Bayesian analysis in York, Durrett, and Nielsen (2002) where the mode of the posterior distribution is 87. However these two numbers differ drastically from the parsimony analyses. The breakpoint distance is $(80-11)/2 = 35$, while the parsimony distance is 54. This lies outside the 95% credible interval of $[65, 120]$ that comes from the Bayesian estimate. Indeed the posterior probability of 54 is so small that this value that it was never seen in the 258 million MCMC updates in the simulation run.

2 Distances

In the last two examples we saw that the breakpoint distance was likely to be an underestimate of the true distance. This brings up the question: when is the parsimony estimate reliable? Bourque and Pevzner (2002) have approached this question by taking 100 markers in order performing k randomly chosen inversions, computing D_k the minimum number of inversions needed to return to the identity and then plotting the average value of $D_k - k \le 0$ (the circles in Figure 3). They concluded based on this and other simulations that the parsimony distance based on n markers was a good as long as the number of inversions was at most $0.4n$. The smooth curve, which we will describe in Theorem 2.1 below, gives the limiting behavior of $(D_{cn} - cn)/n$.

The first step is to consider the analogous but simpler problem for random transpositions. In that case the distance from the identity can be easily computed: it is the number of markers n minus the number of cycles in the permutation. For an example, consider the following permutation of 14 objects written in its cyclic decomposition:

Fig. 3. Bourque-Pevzner simulation results vs. Theorem 2.1



(1 7 4) (2) (3 12) (5 13 9 11 6) (8 10 14)

which indicates that $1 \to 7$, $7 \to 4$, $4 \to 1$, $2 \to 2$, $3 \to 12$, $12 \to 3$, etc. There are 5 cycles so the distance from the identity is 9. If the we perform a transposition that includes markers from two different cycles (e.g., 7 and 9) the two cycles merge into ine, while if we pick two in the same cycle (e.g., 13 and 11) it splits into two.

The situation is similar but slightly more complicated for inversions. There if we ignore the complexity of hurdles, the distance is $n+1$ minus the number of components in the breakpoint graph. An inversion that involves edges in two different components merges them into one but an inversion that involves two edges of the same cycle may or may not increase the number of cycles. To have a cleaner mathematical problem, we will consider the biologically less relevant case of random transpositions, and ask a question that in terms of the rate 1 continuous time random walk on the permutation group is: how far from the identity are we at time cn?

The first step in attacking this problem is to notice that by our description the cycle structure evolves according to a *coagulation-fragmentation process*. Suppose that for the moment we ignore fragmentation and draw an edge from i to j whenever we transpose i and j. In this case the cycles are the components of the resulting random graph. There are $n(n-1)/2$ potential

edges, so results of Erdös and Renyi imply that when $c < 1/2$ there are no very large components and we can ignore fragmentations. In this phase the distance will typically increase by 1 on each step or in the notation of Bourque and Pevzner, $D_k - k \approx 0$. When $n = 100$ this phase lasts until there have been about 50 inversions.

When $c > 1/2$ a giant component emerges in the percolation model and its behavior is much different from the large cycles in the permutation which experience a number of fragmentations and coagulations. The dynamics of the large components are quite complicated but (i) there can never be more than $\sqrt{n}$ of size $\sqrt{n}$ or larger and (ii) an easy argument shows that the number of fragmentations occurring to clusters of size $\le \sqrt{n}$ is $O(\sqrt{n})$. These two observations plus results from the theory of random graphs (see Theorem 12 in Section V.2 of Bollobás 1985) imply

Theorem 2.1. *The number of cycles at time $cn/2$ is $g(c)n + O(\sqrt{n})$ where*

$$g(c) = \sum_{k=1}^{\infty} \frac{1}{k} p_k(c) \quad \textit{and} \quad p_k(c) = \frac{1}{c} \frac{k^{k-1}}{k!} (ce^{-c})^k$$

Using Stirling's formula $k! \sim k^k e^{-k} \sqrt{2\pi k}$ it is easy to see that g' is continuous but $g''(1)$ does not exist. It is somewhat remarkable that $g(c) = 1 - c/2$ for $c < 1$. Thus there is a phase transition in the behavior of the distance of the random transposition random walk from the identity at time $n/2$.

As stated the result only applies to transpositions. However, the same exact conclusion applies to inversions. To show this, we note that the only difference between the two systems is that picking the same cycle twice may or may not increase the number of cycles in the breakpoint graph, and our proof has shown that fragmentations can be ignored.

To explain the strange function $g(c)$ that appears in the answer, we begin with Cayley's result that there are k^{k-2} trees with k labeled vertices. At time cn each edge is present with probability $\approx (cn/2)/\binom{n}{2} \approx c/n$ so the expected number of trees of size k is present is

$$\binom{n}{k} k^{k-2} \left(\frac{c}{n}\right)^{k-1} \left(1 - \frac{c}{n}\right)^{k(n-k)+\binom{k}{2}-(k-1)}$$

since each of the $k-1$ edges need to be present and there can be no edges connecting the k point set to its complement ($k(n-k)$ edges) or any other edges connecting the k points ($\binom{k}{2} - (k-1)$ edges). For fixed k, $\binom{n}{k} \approx n^k/k!$ so the above is

$$\approx n \frac{k^{k-2}}{k!} (2c)^{k-1} \left(1 - \frac{2c}{n}\right)^{kn}$$

from which the result follows easily. We have written the conclusion in the form given above so that $p_k(c)$ is the probability in an Erdös-Renyi graph with edge occupancy probability c/n that 1 belongs to a component of size k.

Having found laws of large numbers for the distance, it is natural to ask about fluctuations. This project is being carried out as part of the Ph.D. thesis of Nathaniel Berestycki. Since these results are only exact for transpositions, and are merely a lower bound for inversions, we will only state the first two results. The subcritical regime ($cn/2$ with $c < 1$) is easy. Let F_t be the number of fragmentations at time t in a system in which transpositions occur at rate one. The continuous time setting is more convenient since it leads to a random graph with independent edges. If N_t is the number of transpositions at time t then $D_t - N_t = -2F_t$ so we study the latter quantity.

Theorem 2.2. *Suppose $0 \le c < 1$. As $n \to \infty$, $F_{cn/2}$ converges in distribution to a Poisson random variable with mean $(-\ln(1-c) - c)/2$.*

Since a Poisson with large mean rescales to approximate a normal, it should not be surprising that if we change time to make the variance linear, the result is a Brownian motion.

Theorem 2.3. *Let $c_n(r) = 1 - n^{-r/3}$ for $0 \le r \le 1$. As $n \to \infty$*

$$X_n(r) = (F_{c_n(r)n/2} - (r/6)\log n)/((1/6)\log n)^{1/2}$$

converges to a standard Brownian motion.

Expected value estimates (see Luczak, Pittel, and Wierman 1994) imply that the number of fragmentations in $[1-n^{-1/3}, 1]$ is $O(1)$ and hence can be ignored. It follows from this that

$$(F_{n/2} - (1/6)\log n)/((1/6)\log n)^{1/2}$$

has approximately a normal distribution. To connect with the simulations of Boruque and Pevnzer, we note that this implies $EF_{50} \approx (1/6)\log 50 = 0.767$ which seems consistent with the data in Figure 3, even though all we know from the comparison is that this is an upper bound on the difference between N_t and the distance.

3 Genomic Distance

In general genomes evolve not only by inversions within chromosomes but also due to translocations between chromosomes, and fissions and fusions that change the number of chromosomes. To reduce the number of events considered from four to two, we note that a translocation splits two chromosomes (into say $a - b$ and $c - d$) and then recombines the pieces (to make $a - d$ and $b - c$ say). A fission is the special case in which the segments c and d are empty, a fusion when b and c are. To illustrate the problem we will consider

part of the data of Doganlar et al. (2002) who constructed a comparative genetic linkage map of eggplant (*Solanum melongena*) with 233 markers based on tomato cDNA, genomic DNA and ESTs. Using the first letter of the common name to denote the species they found that the marker order on T1 and E1 and on T8 and E8 were identical, while in four other cases (T2 vs. E2, T6 vs. E6, T7 vs. E7, T9 vs. E9) the collections of markers were the same and the order became the same after a small number of inversions was performed (3, 1, 2, and 1 respectively).

In our example we will compare of the remaining six chromosomes from the two species. The first step is to divide the chromosomes into *conserved segments* where the adjacency of markers has been preserved between the two species, allowing for the possibility of the overall order being reversed. When such segments have two or more markers we can determine the relative orientation. However as the HP algorithm assumes one knows the relative orientation of segments we will have to assign orientations to conserved segments consisting of single markers in order to minimize the distance. In the case of the tomato-eggplant comparison there are only five singleton segments, so one can easily consider all $2^5 = 32$ possibilities. The next table shows the two genomes with an assignment of signs to the singleton markers that minimizes the distance.

Eggplant	Tomato
1 2 3 4 5 6	1 -5 2 6
7 8	21 -22 -20 8
9 10	-4 14 11 -15 3 9
11 12 13 14 15 16 17 18	7 16 -18 17
19 20 21 22	-19 24 -26 27 25
23 24 25 26 27	-12 23 13 10

As in the inversion distance problem, our first step is to double the markers. The second step is to add ends to the chromosomes and enough empty chromosomes to make the number of chromosomes equal. In this example, no empty chromosomes are needed. We have labeled the ends in the first genome by 1000 to 1011 and in the second genome by 2000 to 2011. The next table shows the result of the first two preparatory steps. Commas indicate separations between two segments or between a segment and an end.

Eggplant
1000, 1 2 , 3 4 , 5 6 , 7 8 , 9 10 , 11 12 , 1001
1002, 13 14 , 15 16 , 1003
1004, 17 18 , 19 20 , 1005
1006, 21 22 , 23 24 , 25 26 , 27 28 , 29 30 , 31 32 , 33 34 , 35 36 , 1007
1008, 37 38 , 39 40 , 41 42 , 43 44 , 1009
1010, 45 46 , 47 48 , 49 50 , 51 52 , 53 54 , 1011

Tomato

2000, 1 2 , 10 9 , 3 4 , 11 12 , 2001
2002, 41 42 , 44 43 , 40 39 , 15 16 , 2003
2004, 8 7 , 27 28 , 21 22 , 30 29 , 5 6 , 17 18 , 2005
2006, 13 14 , 31 32 , 36 35 , 33 34 , 2007
2008, 38 37 , 47 48 , 52 51 , 53 54 , 49 50 , 2009
2010, 24 23 , 45 46 , 25 26 , 19 20 , 2011

As before, the next step is to construct the breakpoint graph which results when the commas are replaced by edges that connect vertices with the corresponding numbers. We did not draw the graph since to compute the distance we only need to know the connected components of the graph. Since each vertex has degree two, these are easy to find: start with a vertex and follow the connections. The resulting component will either be an path that connects two ends or a cycle that consists of markers and no ends. In our example there are five paths of length three: $1000 - 1 - 2000$, $1001 - 12 - 2001$, $1002 - 13 - 2006$, $1003 - 16 - 2003$, and $1005 - 20 - 2011$. These paths tell us that end 1000 in genome 1 corresponds to end 2000 in genome 2, etc. The other correspondences between ends will be determined after we compute the distance. The remaining components in the breakpoint graph are listed below.

1004 17 6 7 27 26 19 18 2005
1006 21 28 29 5 4 11 10 2 3 9 8 2004
1007 36 32 33 35 34 2007
1008 37 47 46 25 24 2010
1009 44 42 43 40 41 2002
1010 45 23 22 30 31 14 15 39 38 2008
1011 54 49 48 52 53 51 50 2009

To compute a lower bound for the distance now we start with the number of commas seen when we write out one genome. In this example that is 33. We subtract the number of connected components in the breakpoint graph. In this example that is $5 + 7 = 12$, and then add the number of paths that begin and end in the same genome, which in this case is 0. The result which is 21 in this case is a lower bound on the distance since any inversion or translocation can at most reduce this quantity by 1, and it is 0 when the two genomes are the same. As before this is only a lower bound. For the genomic distance problem the full answer is quite complicated and involves 7 quantities associated with genome. (For more details see Hannenhalli and Pevzner 1995b or Pevzner 2000.)

At least in this example, nature is simpler than the mathematically worst possible case. It is easy to produce path of length 21 to show that the lower bound is achieved. For a solution see Durrett, Nielsen, and York (2003). That paper extends the methods of York, Durrett, and Nielsen (2002) to develop a Bayesian estimate the number of inversions and translocations separating the

two genomes. As we have just calculated, the parsimony solution for the comparison of all 12 chromosomes is $21 + 7 = 28$. The Bayesian analysis produces 95% credible intervals of [5,7], [21,31], and [28,37] for the number of translocations, inversions, and the total number of events (respectively) separating tomato and eggplant. The mode of the posterior joint distribution of the number of translocations and inversions occurs at (6.6,25.9). Thus even in the case of these two closely related genomes, the most likely number of inversions and translocations are somewhat higher than their parsimony estimates.

When distances between the markers are known in one genome, there is another method due to Nadeau and Taylor (1984) that can be used to estimate the number of inversions and translocations that have occurred. The basic data for the process is the set of lengths of conserved segments, i.e., two or more consecutive markers in one genome that are adjacent (possibly in the reverse order) in the other. The actual conserved segment in the genome is larger than the distance r between the two markers at the end of the conserved interval. Thinking about what happens when we put n points at random in the unit interval, which produces $n + 1$ segments with $n - 1$ between the left-most and the right-most points, we estimate the length of the conserved segment containing these markers by $\hat{r} = r(n + 1)/(n - 1)$ where n is the number of makers in the segment.

Let D be the density of markers, i.e., the total number divided by the size of the genome. If the average length of conserved segments is L and we assume that their lengths are exponentially distributed then since we only detect segments with two markers the distribution of their lengths is

$$(1 - e^{-Dx} - Dxe^{-Dx})\frac{1}{L}e^{-x/L}$$

normalized to be a probability density. A little calculus shows that the mean of this distribution is $(L^2D + 3L)/(LD + 1)$.

Historically the first application of this technique was to a human-mouse comparative map with a total of 56 markers. Based on this limited amount of data they estimated that there were 178 ± 39 conserved segments. For more than fifteen years, this estimate held up remarkably well as the density of the comparative map increased. See Nadeau and Sankoff (1998). However the completion of the sequencing of the mouse genome (Mouse Genome Sequencing Consortium, 2002, see Figure 3) has revealed 342 conserved segments of size $>$ 300Kb.

To illustrate the Nadeau and Taylor computation we will use a comparative map of the human and cattle autosomes (non-sex chromosomes) constructed by Band et al. (2000). Using resources on the NCBI home page we were able to determine the location in the human genome of 422 genes in the map. These defined 125 conserved segments of actual average length 7.188 Mb (megabases) giving rise to an adjusted average length of 14.501 Mb. Assuming 3200 Mb for the size of the human genome the marker density was $D = 1.32 \times 10^{-4}$ or one every 7.582 Mb. Setting $14.501 = (L^2D + 3L)/(LD + 1)$ and solving the

quadratic equation for L gives an estimate $\hat{L} = 7.144$ Mb, which translates into approximately 448 segments. Subtracting 22 chromosome ends we infer there were 424 breakpoints, which leads to an estimate of 212 inversions and translocations. As a check on the assumptions of the Nadeau and Taylor computation, we note that if markers and segment endpoints are distributed randomly then the number of markers in a conserved segment would have a geometric distribution. The next table compares the observed counts with what was expected

markers	observed	expected
0	–	222.9
1	85	108.1
2	76	52.5
3	29	25.4
4	10	12.3
5	5	6.0
6	3	2.9
7	1	1.4
8	1	0.7

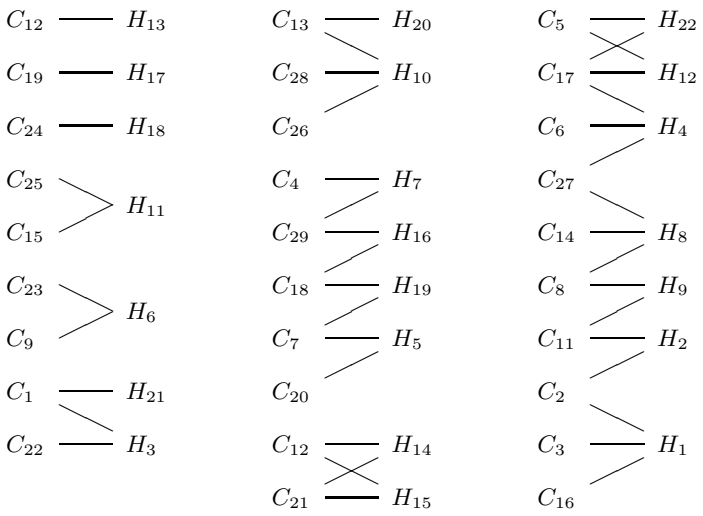
To get an idea of the number of translocations that have occurred we will look at the human-cattle correspondence through the eyes of FISH (fluorescent in situ hybridization) data of Hayes (1995) and Chowdary et al. (1996). In this technique one takes individual human chromosomes labels them with fluorescent chemicals and the determines where they hybridize to cattle chromosomes. To visualize the relationship between the genomes it is useful to draw the bipartite graph with vertices the chromosome numbers in the two genomes and an edge from C_i to H_j if there is part of cattle chromosome i is homologous part of human chromosome j. We call this the *Oxford graph* since the adjacency matrix of this graph is what biologists would call an *Oxford grid*.

Parsimony analysis reveals that a minimum of 155 moves (20 translocations and 135 inversions) is needed to rearrange the cattle genome to match the chromosomes of the human genome. Durrett, Nielsen, and York (2003) have applied their Bayesian methods to this example but experienced convergence problems. Figures 5 and 6 of their paper give posterior distributions from four runs. In the case of inversions, the modes are 20, 21, 21, and 25 with the overall shape of the fourth posterior distribution being considerably different. The modes for translocations are all in the range 185-191 but the variance differs considerably from run to run.

4 Nonuniformity of inversions

Define a *syntenic segment* to be a segment of chromosome where all of the markers come from the same chromosome in the other species but not nec-

Fig. 4. Comparison of cattle and human autosomes.



essarily in the same order. A remarkable aspect of the cattle data is that although our estimates suggest that there have been roughly 20 translocations and 190 inversions, each chromosome consists of only a few syntenic segments. If inversions were uniformly distributed on the chromosome we would expect inversions that occur after a translocations would mingle the two segments.

A second piece of evidence that not all inversions are equally likely comes from the 79 marker *Drosophila* data. The estimated number of inversions is large but there is still a strong correlation between the marker order in the two genomes. Spearman's rank correlation $\rho = 0.326$ which is significant at the $p = 0.001$ level. From the point of view of Theorem 1 this is not surprising: our lower bound on the mixing time predicts that $39.5 \ln 75 = 173$ inversions are needed to completely randomize the data. However, simulations in Durrett (2003) show that the rank correlation is randomized well before that time. In 10,000 runs the average rank correlation is only 0.0423 after 40 inversions and only 4.3% of the runs had a rank correlation larger than 0.325.

To seek a biological explanation of the non-uniformity we note that the gene-to-gene pairing of homologous chromosomes implies that if one chromosome of the pair contains an inversion that the other does not, a loop will form in the region in which the gene order is inverted. (See e.g, page 367 of Hartl and Jones 2000.) If a recombination occurs in the inverted region then

the recombined chromosomes will contain two copies of some regions and zero of others, which can have unpleasant consequences. A simple way to take this into account is

θ-inversion model. Inversions that reverse markers i to $i+j$ occur at rate $\theta^{j-1}/n(1-\theta)$.

The reasoning here is that the probability of no recombination decreases exponentially with the length of the segment reversed.

We expect that the likelihood methods of Durrett, Nielsen, and York can be extended to the θ-inversion model in order to estimate inversion tract lengths. A second way to approach the problem is to see how estimates of the number of inversions depend on the density of markers in the map. If n markers (blue balls) are randomly distributed and we pick two inversion end points (red balls) at random then the relative positions of the $n+2$ balls are all equally likely. The inversion will not be detected by the set of markers if there are 0 or 1 blue balls between the two red ones an event of probability

$$\frac{n+1+n}{\binom{n+2}{2}} = \frac{4n+2}{(n+2)(n+1)} \approx \frac{4}{n+2}$$

This means that the 26 markers in the first *Drosophila* data set should have missed only 1/7 of the inversions in sharp contrast to the fact that our estimate jumped from 24.8 with 26 markers to 86.3 with 79.

Suppose now that markers are distributed according to a Poisson process with mean spacing M while inversion tract lengths have an exponential distribution with mean L. If we place one inversion end point at random on the chromosome and then move to the right to locate the second one then the probability a marker comes before the other inversion endpoint is

$$\frac{1/M}{1/M+1/L} = \frac{L}{L+M}$$

so the fraction detected is $L^2/(L+M)^2$. If we take 30Mb as an estimate for the size of the chromosome arm studied, we see that the marker spacings in the two studies are: $M_1 = 30/27 = 1.11$ Mb and $M_2 = 30/80 = .375$ Mb respectively. Taking ratios we can estimate L by

$$\frac{86.3}{24.8} = \frac{(L+1.1)^2}{(L+0.375)^2}$$

Taking square roots of each side and solving we have $1.865L + 0.375 = L + 1.1$ or $L = 0.725/0.765 = 0.948$ Mb. If this is accurate then the larger data set only detects

$$\left(\frac{0.948}{1.273}\right)^2 = 0.554$$

or 55.4% of the inversions that have occurred. That is our best guess is that the chromosome arm has experienced $86.3/0.554 = 157$ inversions. This simple calculation is only meant to illustrate the possibilities of the method, which needs to be developed further and tested on other examples.

REFERENCES

Bafna, V. and Pevzner, P. (1995) Sorting by reversals: Genome rearrangement in plant organelles and evolutionary history of X chromosome. *Mol. Biol. Evol.* **12**, 239-246

Band, M.J. et al. (2000) An ordered comparative map of the cattle and human genomes. *Genome Research.* **10**, 1359–1368

Bollobás, B. (1985) *Random Graphs.* Academic Press, New York

Bourque, G., and Pevzner, P.A. (2002) Genome-scale evolution: Reconstructing gene orders in ancestral species. *Genome Research.* **12**, 26–36

Chowdhary, B.P., Fronicke, L., Gustavsson, I., Scherthan, H. (1996) Comparative analysis of cattle and human genomes: detection of ZOO-FISH and gene mapping-based chromosomal homologies. *Mammalian Genome.* **7**, 297–302

Doganlar, S., Frary, A., Daunay, M.C., Lester, R.N., and Tanksley, S.D. (2002) A comparative genetic linkage map of eggplant (*Solanum melongea*) and its implications for genome evolution in the Solanaceae. *Genetics.* **161**, 1697–1711

Durrett, R. (2002) *Probability Models for DNA Sequence Evolution.* Springer-Verlag, New York

Durrett, R. (2003) Shuffling chromosomes. *J. Theoretical Prob.* **16**, 725–750

Durrett, R., Nielsen, R., and York, T.L., (2003) Bayesian estimation of genomic distance. *Genetics.*, to appear

Hannenhalli, S., and Pevzner, P.A. (1995a) Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). Pages 178–189 in *Proceedings of the 27th Annual ACM Symposium on the Theory of Computing.* Full version in the *Journal of the ACM.* **46**, 1–27

Hannenhalli, S., and Pevzner, P. (1995b) Transforming men into mice (polynomial algorithm for the genomic distance problem. Pages 581-592 in *Proceedings of the 36th Annual IEEE Symposium on Foundations of Computer Science.* IEEE, New York

Hartl, D.L., and Jones, E.W. (2000) *Genetics: The Analysis of Genes and Genomes.* Jones and Bartlett, Sudbury, MA

Hayes, H. (1995) Chromosome painting with human chromosome-specific DNA libraries reveals the extent and distribution of conserved segments in bovine chromosomes. *Cytogenet. Cell. Genetics.* **71**, 168–174

Luczak, T., Pittel, B., and Weirman, J.C. (1994) The structure of the random graph at the point of phase transition. *Trans. Amer. Math. Soc.* **341**, 721–748

Mouse Genome Sequencing Consortium. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature.* **420**, 520–562

Nadeau, J.H., and Taylor, B.A. (1984) Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Nat. Acad. Sci.* **81**, 814–818

Nadeau, J.H., and Sankoff, D. (1998) The lengths of undiscovered conserved segments in comparative maps. *Mammalian Genome.* **9**, 491–495

Pevzner, P.A. (2000) *Computational Molecular Biology: An Algorithmic Approach.* MIT Press, Cambridge.

Pevzner, P.A. and Tesler, G. (2003) Genome rearrangment in mammalian evolution: Lessons from human and mouse genomes. *Genome Research.* **13**, 37–45

Ranz, J.M., Casals, F., and Ruiz, A. (2001) How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus *Drosophila*. *Genome Research.* **11**, 230–239

Ranz, J.M., Segarra, S., and Ruiz, A. (1997) Chromosomal homology and molecular organization of Muller's element D and E in the *Drosophila repleta* species group. *Genetics.* **145**, 281–295

York, T.L., Durrett, R., and Nielsen, R. (2002) Bayesian estimation of the number of inversions in the history of two chromosomes. *J. Comp. Bio.* **9**, 805–818