JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 13, Number 2, 2006

Pp. 554–566

Stability of Rearrangement Measures in the Comparison of Genome Sequences

MATTHEW MAZOWITA, LANI HAQUE, and DAVID SANKOFF

ABSTRACT

We present data-analytic and statistical tools for studying rates of rearrangement of whole genomes and to assess the stability of these methods with changes in the level of resolution of the genomic data. We construct datasets on the numbers of conserved syntenies and conserved segments shared by pairs of animal genomes at different levels of resolution. We fit these data to an evolutionary tree and find the rates of rearrangement on various evolutionary lineages. We document the lack of clocklike behavior of rearrangement processes, the independence of translocation and inversion rates, and the level of resolution beyond which translocations rates are lost in noise due to other processes.

Key words: comparative genomics, rearrangement rates, genome browsers, models of evolution, mammalian phylogeny.

1. INTRODUCTION

THE GOAL OF THIS PAPER IS TO PRESENT DATA-ANALYTIC and statistical tools for studying rates of rearrangement of whole genomes and to assess the stability of these methods with changes in the level of resolution of the genomic data. From secondary data provided by the UCSC Genome Browser, we construct datasets on the number of *conserved syntenies* (pairs of chromosomes, one from each species, containing at least one sufficiently long stretch of homologous sequence) and the number of *conserved segments* (i.e., the total number of such stretches of homologous sequence) shared by pairs of animal genomes at levels of resolution from 30 kb to 1 Mb. For each lineage, we calculate rates of interchromosomal and intrachromosomal rearrangement, with and without the assumption these are due to translocations and inversions of specific kinds.

The key to using whole genome sequences to study evolutionary rearrangements is being able to partition each genome into segments conserved by two genomes since their divergence. In the higher animals and many other eukaryotes, this can be extremely difficult, given the high levels of occurrence of transposable elements, retrotranspositions, paralogy and other repetitive and/or inserted sequences, deletions, conversion, and uneven sequence divergence. The inference of conserved segments becomes a multistage procedure parametrized by repeat-masking sensitivities, alignment scores, penalties and thresholds, and various numerical criteria for linking smaller segments into large ones. Successful protocols have been developed independently by two research groups (Kent *et al.*, 2003; Pevzner and Tesler, 2003a) using somewhat different strategies to combine short regions of elevated similarity to construct the conserved segments, bridging singly gapped or doubly gapped regions where similarity does not attain a threshold

Department of Mathematics and Statistics, University of Ottawa, Ottawa, Canada, K1N 6N5.

criterion and ignoring short inversions and transpositions that have rearranged one sequence or the other. We develop our datasets on syntenies and segments from the continuously updated output of the UCSC protocol made available on the genome browser.

Building on the ideas of Sankoff *et al.* (1997, 2000, 2005), we derive an estimator for the number of reciprocal translocations responsible for the number of conserved syntenies between two genomes and use simulations to show that the bias and standard deviation of this estimator are less than 5%, under two models of random translocation, with and without strict conservation of the centromere. By contrasting the number of conserved segments with the number of conserved syntenies, we can also estimate the number inversions or other intrachromosomal events.

The form of this *process-based* estimator suggests a derivation of yet another, *state-based*, estimator, which is even more accurate and is also directly applicable to the comparison of genomes with different numbers of chromosomes, which is the case for the genomes we shall examine here.

Our results include:

- A loss of stability of the data at resolutions starting at about 100 kb.
- A highly variable proportion of translocations relative to inversions across lineages.
- The relative stability of translocation rates compared to inversion rates as resolution is refined.
- The absence of correlation between accumulated rearrangements and chronological time elapsed, especially beyond 20 Myr.

2. THE DATA

We examined the UCSC browsers for five mammalian species and the chicken and constructed our sets of segments and syntenies for the pairs shown in Table 1. We did not use other browsers, or other nets on the four browsers in the table, because either the browser-net pairs are not posted, or because only a build older than the one used in our table was posted.

For each of the pairs in Table 1, four datasets were constructed, one at each of the 1 Mb, 300 kb, 100 kb, and 30 kb levels. These contain all segments larger than the resolution level as measured by the segment starting and ending points. We counted only segments in autosomes, except for the comparisons with chicken where the sex chromosomes were also included.

One complication in identifying the segment stems from the key technique of the net construction of Kent *et al.* (2003), which allows very long double gaps in alignments. These gaps are often so long that they contain nested large alignments with chromosomes other than that of the main alignment. Whenever a gap in an alignment contained a nested alignment larger than the level of resolution, we broke the main alignment in two and counted the segments before and after the gap separately, assuming they remained long enough, as well as the segment in the gap.

Table 2 shows the results of our data extraction procedure. Entries for $\hat{t}$ and $\hat{i}$ are calculated according to Equations (9) and (1), respectively.

TABLE 1. BROWSERS AND NETS PROVIDING SEGMENTS[a]

	Net species and build number					
Browser species ↓	*Human Hg17*	*Mouse Mm6*	*Chimp PanTro1*	*Rat Rn3*	*Dog CanFam1*	*Chicken GalGal2*
Human (22)		✓	✓	✓	✓	✓
Mouse (19)	✓		✓	✓	✓	✓
Chimp (23)		✓				
Rat (20)	✓	✓				
Dog (38)	✓	✓				
Chicken (30)	✓	✓				

[a]For each browser, parentheses contain number of autosomes (chromosomes in the case of chicken) being compared.

TABLE 2. DATA ON CONSERVED SYNTENIES c', CONSERVED SEGMENTS n, INFERRED TRANSLOCATIONS $\hat{t}$ AND INFERRED INVERSIONS $\hat{i}$[a]

	Human browser				*Mouse browser*				*Chimp browser*			
Net	c'	n	$\hat{t}$	$\hat{i}$	c'	n	$\hat{t}$	$\hat{i}$	c'	n	$\hat{t}$	$\hat{i}$
Human												
					197	1357	129.5 ± 1.1	538.7 ± 0.3				
					150	656	84.6 ± 0.9	233.1 ± 0.2				
					126	394	65.4 ± 0.9	121.3 ± 0.1				
					110	258	53.7 ± 0.9	65.1 ± 0.1				
Mouse												
30 kb	230	1076	169.1 ± 1.2	358.7 ± 0.5					296	2638	264.5 ± 2.0	1044.0 ± 1.0
100 kb	142	493	78.0 ± 0.9	158.3 ± 0.2					217	1048	146.9 ± 1.5	366.6 ± 0.5
300 kb	115	354	57.3 ± 0.9	109.5 ± 0.1					176	566	105.1 ± 1.3	167.4 ± 0.3
1 Mb	105	249	50.2 ± 0.9	64.1 ± 0.1					145	354	78.7 ± 1.2	87.8 ± 0.2
Chimp												
30 kb	92	2472	39.4 ± 0.3	1185.3 ± 0.1	191	2183	119.4 ± 1.4	961.6 ± 0.4				
100 kb	44	492	11.3 ± 0.3	223.4 ± 0.0	141	744	75.5 ± 1.2	286.0 ± 0.2				
300 kb	30	133	3.9 ± 0.3	51.4 ± 0.0	126	402	64.1 ± 1.2	126.4 ± 0.2				
1 Mb	25	64	1.3 ± 0.3	19.5 ± 0.0	109	268	52.1 ± 1.2	71.4 ± 0.2				
Rat												
30 kb	186	1107	119.6 ± 0.8	423.4 ± 0.3	140	2186	79.6 ± 0.3	1003.7 ± 0.0				
100 kb	124	566	64.2 ± 0.7	208.3 ± 0.2	80	735	34.8 ± 0.2	323.0 ± 0.0				
300 kb	104	325	49.6 ± 0.6	102.4 ± 0.1	58	235	21.2 ± 0.2	86.6 ± 0.0				
1 Mb	94	232	42.7 ± 0.6	62.8 ± 0.1	54	122	18.8 ± 0.2	32.4 ± 0.0				
Dog												
30 kb	221	715	115.9 ± 5.2	226.6 ± 1.2	280	1275	166.2 ± 6.4	457.1 ± 1.7				
100 kb	122	348	50.6 ± 4.6	108.4 ± 0.6	214	608	113.3 ± 5.9	176.4 ± 1.2				
300 kb	87	231	30.4 ± 4.4	70.1 ± 0.4	184	400	91.9 ± 5.7	93.8 ± 1.0				
1 Mb	80	182	26.5 ± 4.3	49.5 ± 0.3	157	278	73.8 ± 5.5	50.9 ± 0.8				
Chicken												
30 kb	204	1873	152.2 ± 2.8	770.8 ± 1.3	228	1677	205.4 ± 6.1	620.4 ± 3.9				
100 kb	126	1007	66.0 ± 1.8	424.0 ± 0.3	167	981	113.3 ± 4.1	364.5 ± 1.8				
300 kb	89	603	36.9 ± 1.6	251.1 ± 0.1	137	587	79.9 ± 3.4	200.9 ± 1.1				
1 Mb	76	334	28.0 ± 1.5	125.5 ± 0.0	108	322	53.4 ± 2.9	94.9 ± 0.6				

TABLE 2. (*Continued*)

	Rat browser				*Dog browser*				*Chicken browser*			
Net	c'	n	$\hat{t}$	$\hat{i}$	c'	n	$\hat{t}$	$\hat{i}$	c'	n	$\hat{t}$	$\hat{i}$
Human												
	179	1484	112.4 ± 0.8	619.1 ± 0.3	100	403	37.8 ± 4.4	148.7 ± 0.4	109	775	51.8 ± 1.7	322.2 ± 0.2
	128	711	67.3 ± 0.7	277.7 ± 0.2	87	275	30.4 ± 4.4	92.1 ± 0.4	93	549	39.7 ± 1.6	221.3 ± 0.1
	102	358	48.2 ± 0.6	120.3 ± 0.1	83	221	28.2 ± 4.3	67.3 ± 0.3	81	339	31.4 ± 1.5	124.6 ± 0.0
	94	241	42.7 ± 0.6	67.3 ± 0.1	79	169	26.0 ± 4.3	43.5 ± 0.3	62	176	19.2 ± 1.5	55.3 ± 0.0
Mouse												
30 kb	145	1867	84.0 ± 0.3	839.8 ± 0.0	198	719	101.7 ± 5.8	243.5 ± 1.1	170	1074	117.0 ± 4.2	407.3 ± 1.9
100 kb	97	748	46.2 ± 0.2	318.0 ± 0.0	175	434	85.8 ± 5.7	117.0 ± 0.9	146	709	89.2 ± 3.6	252.5 ± 1.3
300 kb	60	257	22.4 ± 0.2	96.4 ± 0.0	161	340	76.5 ± 5.6	79.3 ± 0.8	130	425	73.0 ± 3.2	126.7 ± 1.0
1 Mb	41	95	11.4 ± 0.2	26.3 ± 0.0	147	257	67.4 ± 5.5	45.3 ± 0.7	91	206	40.0 ± 2.6	50.2 ± 0.4
Chimp												
30 kb												
100 kb												
300 kb												
1 Mb												
Rat												
30 kb												
100 kb												
300 kb												
1 Mb												
Dog												
30 kb												
100 kb												
300 kb												
1 Mb												
Chicken												
30 kb												
100 kb												
300 kb												
1 Mb												

[a] Values of $\hat{t}$ and $\hat{i}$ represent average of two estimates, one where the browser chromosome sizes are used to estimate the p and the net ones the q, and the other where the roles are reversed. The ± indicates the difference beween either of the two individual estimates and the average.

3. MODELS OF TRANSLOCATION

In order to derive and validate our estimator of translocation rates, we model the autosomes of a genome as c linear segments with lengths $p(1), \ldots, p(c)$, proportional to the number of base pairs they contain, where $\sum_{i=1}^{c} p(i) = 1$. We assume the two breakpoints of a translocation are chosen independently according to a uniform distribution over all autosomes, conditioned on their not being on the same chromosome. There is no statistical evidence (Sankoff *et al.*, 2002) that translocational breakpoints cluster in a nonrandom way on chromosomes, except in a small region immediately proximal—within 50–300 kb—to the telomere in a wide spectrum of eukaryote lineages (Mefford and Trask, 2002).

A reciprocal translocation between two chromosomes h and k consists of breaking each one, at some interior point, into two segments and rejoining the four resulting segments such that two new chromosomes are produced.

In one version of our model, we impose a left–right orientation on each chromosome, such that a left-hand fragment must always rejoin a right-hand fragment. This ensures that each chromosome always retains a segment, however small it may become, containing its original left-hand extremity. This restriction models the conservation of the centromere without introducing complications such as trends towards or away from acrocentricity. With further translocations, if a breakpoint falls into a previously created segment on chromosome i, it divides that segment into two new segments, the left-hand one remaining in chromosome i, while the right-hand one, and all the other segments to the right of the breakpoint, are transferred to the other chromosome involved in the translocation. It is for this version of the model that we will derive an estimator of the number of translocations, and that we will simulate to test the estimator.

In another version of the model, an inverted left-hand fragment may rejoin another left-hand fragment and similarly for right-hand fragments. This models a high level of neocentromeric activity. We will also simulate this model to see how our estimator (derived from the previous model) fares.

We do not consider chromosome fusion and fission, so that the number of chromosomes is constant throughout the time period governed by the model. Later, in our analysis of animal genomes, we simply assume that the case where fusions or fissions occur will be well approximated by interpolating two models (with fixed chromosome number) corresponding to the two genomes being compared.

Moreover, in our simulations, we do not consider the effects of inversions on the accuracy of estimator. In previous work (Sankoff and Mazowita, 2005), we showed that high rates of long inversions would severely bias the estimator upwards, but that the rates and distribution of inversion lengths documented for mammalian genomes (Kent *et al.*, 2003; Pevzner and Tesler, 2003a) had no perceptible biasing effect.

In our simulations, we impose a threshold and a cap on chromosome size, rejecting any translocation that results in a chromosome too small or too large. Theories about meosis (e.g., Schubert and Oud [1997]) can be adduced for these constraints, though there are clear exceptions, such as the "dot" chromosomes of avian and some reptilian and other vertebrate genomes (Bed'Hom, 2000; Burt, 2002).

The total number of segments on a human chromosome i is

$$n^{(i)} = t^{(i)} + 2u^{(i)} + 1, \tag{1}$$

where $t^{(i)}$ is the number of translocational breakpoints on the chromosome and $2u^{(i)}$ is the number of inversion breakpoints.

4. PREDICTION AND ESTIMATION

We assume that our random translocation process is temporally reversible, and to this effect we show in Fig. 1 and Section 5.1 that the equilibrium state of our process well approximates the observed distribution of chromosome lengths in the human genome. In comparing two genomes, this assumption allows us to treat either one as ancestral and the other as derived, instead of having to consider them as diverging independently from a common ancestor.

At the outset, assume the first translocation on the lineage from genome A to genome B involves chromosome i. The assumption of a uniform density of breakpoints across the genome implies that the "partner" of i in the translocation will be chromosome j with probability $p_i(j) = \frac{p(j)}{1-p(i)}$. Thus, the

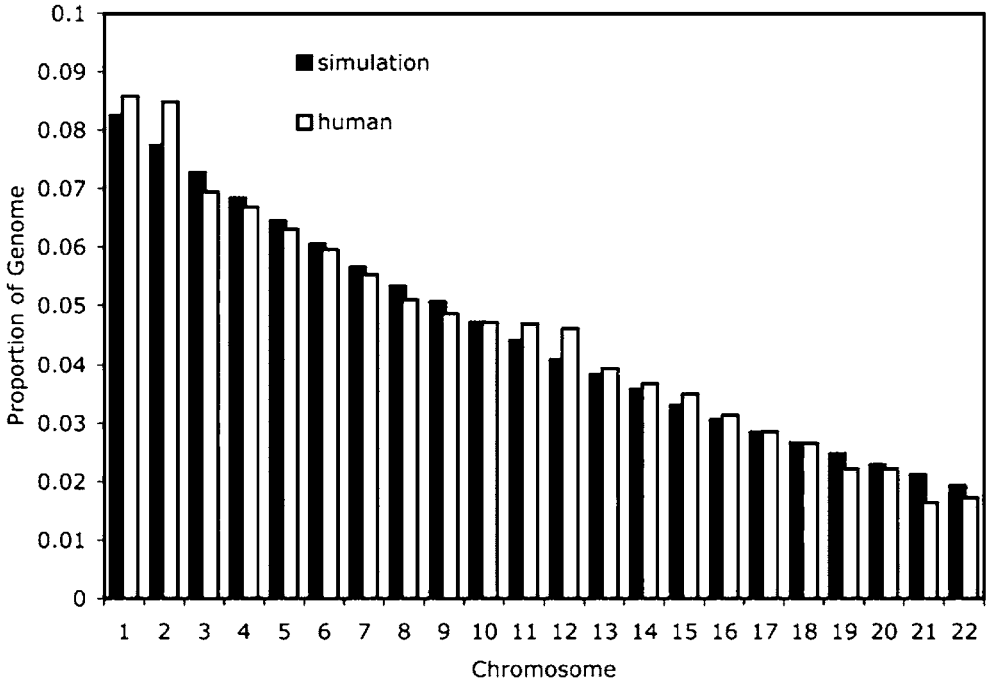

FIG. 1. Comparison of equilibrium distribution of simulated chromosome sizes with human autosome sizes.

probability that the new chromosome labeled i contains no fragment of genome A chromosome j, where $j \neq i$, is $1 - p_i(j)$. For small $t^{(i)}$, after chromosome i has undergone $t^{(i)}$ translocations, the probability that it contains no fragment of the genome A chromosome j is approximately $(1 - p_i(j))^{t^{(i)}}$, neglecting second-order events, for example, the event that j previously translocated with one or more of the $t^{(i)}$ chromosomes that then translocated with i, and that a secondary transfer to i of material originally from j thereby occurred.

Though even with moderate sized $t^{(i)}$, we can expect these secondary transfers to be quite significant, we will see that neglecting them does not detract from the accuracy of the estimator.

Then the probability that the genome B chromosome i contains at least one fragment from j is approximately $1-(1-p_i(j))^{t^{(i)}}$, and the expected number of genome A chromosomes with at least one fragment showing up on genome B chromosome i is

$$E(c^{(i)}) \approx 1 + \sum_{j \neq i} [1 - (1 - p_i(j))^{t^{(i)}}] \tag{2}$$

so that

$$c - E(c^{(i)}) \approx \sum_{j \neq i} (1 - p_i(j))^{t^{(i)}} \tag{3}$$

where the leading 1 in (2) counts the fragment containing the left-hand endpoint of the genome A chromosome i itself. We term $c^{(i)}$ the number of conserved syntenies on chromosome i.

Suppose there have been a total of t translocations in the evolutionary history. Then

$$\sum_i t^{(i)} = 2t. \tag{4}$$

We can expect these to have been distributed among the chromosomes approximately as

$$t^{(i)} = 2tp(i), \tag{5}$$

so that

$$c^2 - \sum_i E(c^{(i)}) \approx \sum_i \sum_{j \neq i} (1 - p_i(j))^{2tp(i)}. \tag{6}$$

Substituting the $c^{(i)}$ for the $E(c^{(i)})$ in Equation (6) suggests solving

$$c^2 - \sum_i c^{(i)} = \sum_i \sum_{j \neq i} (1 - p_i(j))^{2\hat{t}p(i)} \tag{7}$$

for t to provide an estimator of t. Newton's method converges rapidly for the range of parameters used in our studies, as long as not all $c^{(i)} = c$, the point at which every chromosome is completely scrambled. We call this estimator the *process-based* estimator because its derivation makes use of the translocational model in Section 3.

5. AN ESTIMATOR FOR GENOMES WITH DIFFERENT NUMBERS OF CHROMOSOMES

The preceding analysis assumed that there has been no change in chromosome number over the time interval in which genome B evolved. The genomes of the animals we will be studying, however, have different numbers of chromosomes. One solution is simply to assume that all chromosome fusions and fissions occurred as a first step in the evolutionary process, so that the translocational model could then apply to a fixed chromosomal complement, i.e., with constant c corresponding to the number in genome B. By then reversing the analysis, considering genome A as derived from genome B, so that c corresponds to genome A, we model the situation where all fissions and fusions occurred at the time of genome B. The hope is that averaging the two sets of results will produce figures consistent with the process of fissions and fusions occurring randomly over the evolutionary period. Experience with this approach, however, shows that the two analyses often result in rather different estimates, inspiring little confidence in the accuracy of their averages. This problem motivates us to construct an estimator that takes into account the numbers of chromosomes in both of the genomes.

That the process-based estimator derived from Equation (7) provides accurate results for the case where c is constant, as we shall see in Section 6, despite the breakdown in the assumptions in the derivation as t increases, suggests that this estimator, or something close to it, can be derived more directly based on the state of the two genomes, without recourse to the evolutionary model.

The process-based estimator is based on terms of form $1 - (1 - p_i(j))^{2tp(i)}$, which is the probability of at least one success in $2tp(i)$ binomial trials with parameter $p_i(j)$. This suggests that we view the segments on each of the d chromosomes of genome B as being produced by a set of binomial variables, one for each chromosome of genome A. For each of the $2tp(i)$ translocations affecting chromosome i in genome A, the probability that it produces a segment in chromosome j of genome B is just $q(j)$, the proportion of genome B represented by chromosome j. Then $1 - (1 - q(j))^{2tp(i)}$ is the probability that i in A and j in B share at least one segment, i.e., have a conserved synteny. The expected number of conserved syntenies would then be $cd - \sum_{i=1}^{c} \sum_{j=1}^{d} [1 - (1 - q(j))^{2tp(i)}]$.

This analysis, however, does not take into account that each chromosome i in A has one conserved synteny with some chromosome j in B, dating from before any translocation occurred, so that the expected number of conserved syntenies $c^{(i)}$ is really $1 + \sum_{k \neq j} [1 - (1 - q(k))^{2tp(i)}]$. Since j is unknown, we assume an average $q(j)$ in the sense that the expected number of conserved syntenies $c^{(i)}$ is

$$1 + \frac{d-1}{d} \sum_{k=1}^{d} [1 - (1 - q(k))^{2tp(i)}] = d - \frac{d-1}{d} \sum_{k=1}^{d} (1 - q(k))^{2tp(i)}. \tag{8}$$

A "state-based" estimator based on the total expectation over all chromosomes in A then satisfies

$$cd - \sum_i c^{(i)} = \frac{d-1}{d} \sum_i \sum_j (1 - q(j))^{2p(i)\hat{t}} \tag{9}$$

which can be solved as before.

TABLE 3. SHORTEST AND LONGEST CHROMOSOME, IN Mb

Genome	*Shortest*	*Longest*
Mouse	61	199
Human	47	246
Rat	47	268
Chimp	47	230
Dog	26	125
Chicken	0.24	188

6. SIMULATIONS

6.1. Equilibrium distribution of chromosome size

Models of accumulated reciprocal translocations for explaining the observed range of chromosome sizes in a genome date from Sankoff and Ferretti (1996). They proposed a lower threshold on chromosome size in order to reproduce the appropriate size range in plant and animal genomes containing from two to 22 autosomes. A cap on largest chromosome size has also been proposed (Schubert and Oud, 1997) and shown to be effective (De *et al.*, 2001). Economy and elegance in explaining chromosome size being less important in the present context than simulating a realistic equilibrium distribution of these sizes, we imposed both a threshold of 50 Mb and a cap of 250 Mb on the process described in Section 3, simply rejecting any translocations that produced chromosomes out of the range. These values were inspired by the relative stability across primates and rodents evident in the data in Table 3, though they are less pertinent for the dog, with a larger number of correspondingly smaller chromosomes, and chicken, which has several very small chromosomes.

Simulating the translocation process 100 times up to 10,000 translocations each produced the equilibrium distribution of chromosome sizes in Fig. 1. The superimposed distribution of human autosome sizes is very close to the equilibrium distribution.

6.2. Performance of the estimators

Figure 2 depicts the number of translocations inferred by the process-based estimator as a function of the true number t in the simulation. The estimator $\hat{t}$ appears very accurate, only lightly biased (less than 5% for $t < 200$), with small error rates (s.d. less than 5% for $t < 200$).

To compare the performance of the "state-based" estimator with the process-based one, we consider the case where $d = c$ in Equation (9), and use the corresponding estimator for the same simulations as in Fig. 2. The results, in Fig. 3, show that the state-based estimator is even more accurate than the process-based one.

Thus, we will use the estimator based on solving Equation (9) for our data analysis in the remainder of this paper.

7. FITTING THE DATA TO ANIMAL PHYLOGENY

To infer the rates of rearrangement on evolutionary lineages, we assumed the phylogenetic tree in Fig. 4. Because of the limited number of genome pairs for which we have data, we could not estimate the translocational divergence during the mammalian radiation, i.e., between the divergence of the dog lineage and the common primate/rodent lineage.

We fit the data in Table 2 to the tree by solving the system of linear equations between the additive path lengths in the tree and the inferred rearrangement distances, namely, $c' - c$ the number of new syntenies created on a path, $\hat{t}$ the number of translocations inferred to have occurred, $n - c$ the number of segments created, and $\hat{i}$, the inferred number of inversions. With one exception, data from all genome pairs (A,B) were available both with A as browser and B as net, and with B as browser and A as net, so we averaged the corresponding values in Table 2 to produce a single term. The one exception involves the lack of a human net on the chimp browser. Because the mouse–chimp comparison suggests that the chimp browser

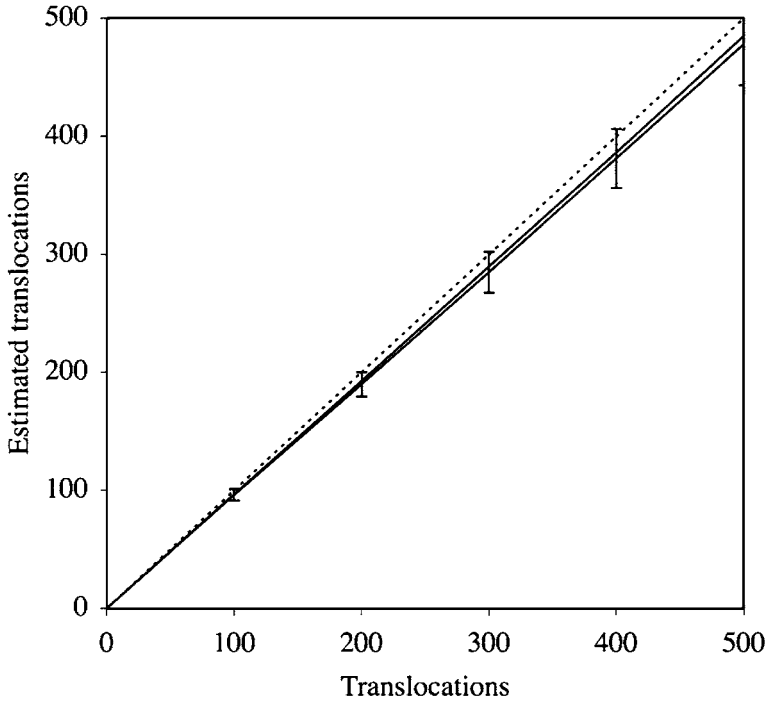

FIG. 2. Mean value, over 100 runs, of $\hat{t}$ as a function of t. Dotted line: $\hat{t} = t$. Lower line with ± 1 s.d. error bars: model with centromere. Upper line: model without centromere.

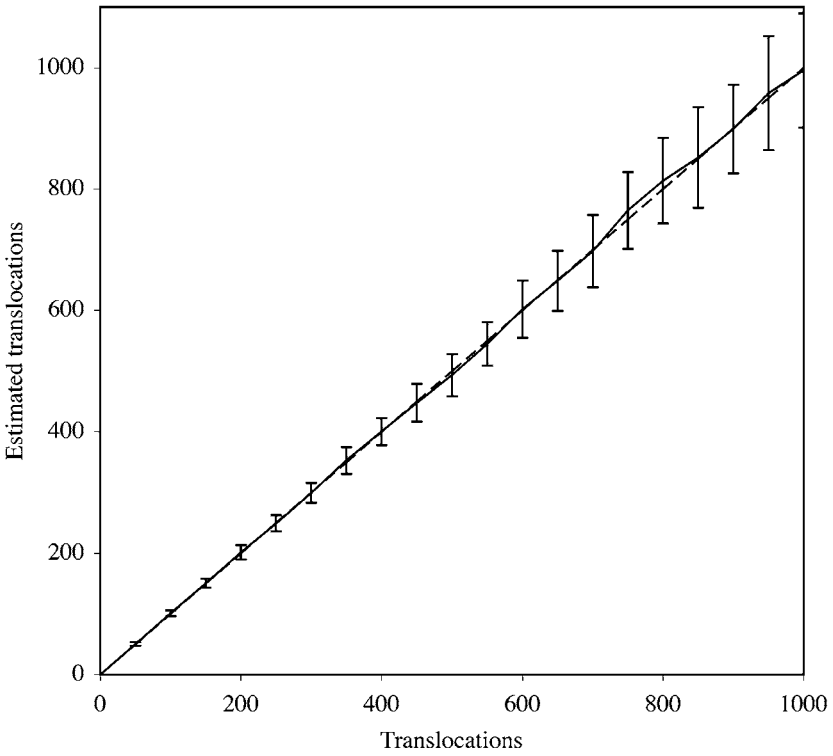

FIG. 3. Mean value, over 100 runs, of state-based $\hat{t}$ as a function of t. Dotted line: $\hat{t} = t$.

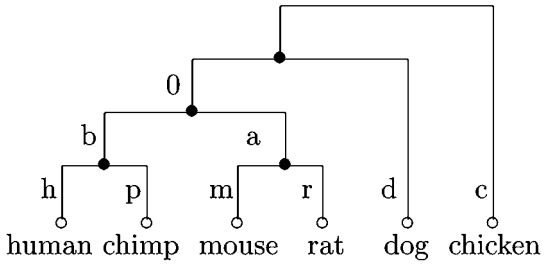

FIG. 4. Unrooted phylogeny for fitting rearrangement measures. Edge labeled "0" indicates the two endpoints are collapsed; none of the pairwise measures bear on the location of the vertex between chicken and dog. The length of this edge (representing the period of rapid mammalian radiation) can be assumed to be short in comparison to c, d, b, and a.

produces an exaggerated number of segments and syntenies, we simply discarded the results from this browser and used only the two comparable chimp nets, on the human and on the mouse browsers.

Since results from the more finished genomes are likely to be more reliable, we adopted a three-stage procedure to estimate the edge lengths:

- Use the three (best-established) comparisons among human, mouse, and rat to estimate m, r, and h+b+a.
- Combine the mouse–chicken and human–chicken comparisons to obtain c, making use of the previous estimates of m and h+b+a. Combine the mouse–dog and human–dog comparisons to obtain d. Similarly, estimate p and h from the two chimp comparisons.
- The estimates for m, h, c, and d provide two estimates for each of a and b when substituted in the chicken comparisons and the dog comparisons in the previous step. We then average the two values for each length.

The procedure produces the results in Table 4.

The very approximate temporal edge lengths given in Table 4 were based, for p and h, on the usual estimate (6 Myr) of chimp–human divergence; for m and r, the date for the rat–mouse divergence ($\approx$ 20 Myr); for d, the date for mammalian radiation; for a, the same date less the 20 Myr of murid evolution; for b, the same date less the 6 Myr of primate evolution; and for c, twice the mammalian–reptile divergence time (310 Myr) less the 85 Myr since the mammalian radiation.

8. OBSERVATIONS

The most striking trend in Table 4 is the dramatic increase in both conserved syntenies and conserved segments in almost every lineage (except a and b) as the level of resolution is refined, starting at 100 kb but accelerating rapidly at 30 kb. It seems likely that the increased level of translocations inferred is artifactual, the apparent level of conserved syntenies reflecting retrotransposition and other interchromosomal process and not reciprocal translocation (Trinh *et al.*, 2004).

That this increase does not reflect translocational distance is further evidenced by the loss at 30 kb of clear trends among the lineages visible at less refined resolutions, such as the very low values for human, mouse, and rat compared to the other lineages. Thus, we can conclude that below the 100 kb level, the study of translocational rearrangement by our statistical approach is no longer feasible.

Even at the less refined levels of resolution, any correlation between chronological time and translocational distance breaks down somewhere between 20 and 65 Myr. As has been remarked previously (Bourque *et al.*, 2005), the chicken evidences a low rate of translocation. The dog on the other hand, shows a high rate.

Turning to the results on inversions, the rapid increase in segments and inferred inversions at refined resolutions may, in contrast to translocations, be a real effect. It is known that the inversions of small size are very frequent in these genomes (Kent *et al.*, 2003), with a mean size less than 1 kb, so that it can be expected that the number of inversions inferred will continue to accelerate with increased resolution.

TABLE 4. TREE EDGE-LENGTHS ESTIMATED FROM PAIRWISE MEASURES IN TABLE 2. NEGATIVE ENTRIES INDICATE POOR FIT OF THE DATA TO TREE

	Time							
	h	*p*	*b*	*m*	*r*	*a*	*d*	*c*
	6	6	79	20	20	65	85	535
	New syntenies created							
	h	*p*	*b*	*m*	*r*	*a*	*d*	*c*
30 kb	57.3	34.8	26.9	86.8	55.8	42.6	93.0	55.8
100 kb	24.5	19.5	20.9	54.3	34.3	46.4	76.5	46.8
300 kb	12.3	17.8	20.1	38.3	20.8	49.9	68.5	36.8
1 Mb	11.8	13.3	19.9	30.5	17.0	45.4	62.0	23.3
	Translocations							
	h	*p*	*b*	*m*	*r*	*a*	*d*	*c*
30 kb	34.7	4.8	19.1	57.5	24.3	38.1	30.8	40.7
100 kb	8.5	2.8	11.4	28.0	12.5	33.3	29.4	24.0
300 kb	0.5	3.3	11.0	17.1	4.7	32.7	26.1	14.2
1 Mb	0.6	0.7	11.1	12.2	2.9	28.0	22.5	4.0
	New segments created							
	h	*p*	*b*	*m*	*r*	*a*	*d*	*c*
30 kb	752.8	1719.3	−229.5	973.8	1052.8	−280.5	169.8	666.8
100 kb	161.3	330.8	76.9	338.8	402.8	−2.4	129.0	484.3
300 kb	52.5	80.5	100.5	139.3	106.8	81.8	111.0	280.0
1 Mb	24.8	39.3	80.9	62.8	45.8	85.1	94.0	125.3
	Inversions							
	h	*p*	*b*	*m*	*r*	*a*	*d*	*c*
30 kb	336.2	849.1	−133.7	424.6	497.1	−178.4	44.6	284.5
100 kb	66.6	156.8	27.1	136.6	183.9	−34.6	25.6	209.9
300 kb	20.2	31.2	39.3	47.8	43.7	8.1	19.9	117.6
1 Mb	6.3	13.2	29.4	14.4	14.9	14.4	15.0	50.4

Indeed, even between 1 Mb and 300 kb, while translocation rates are relatively stable, inversion rates increase substantially.

The pattern of inversion rates among the lineages is very different from that of translocations. Here, chicken has a very high rate while dog has a moderate one, the opposite of what was seen with translocations. Perhaps most startling is the high rate recovered for the chimp lineage compared to the human lineage.

Again with inversions, as with translocations, there is little correlation of lineage-specific rates and the chronological span of the lineage.

Finally, when inversion and translocation rates are compared at a fixed level of resolution, no systematic association can be seen.

9. DISCUSSION

We have proposed an estimator of the number of translocations intervening between two rearranged genomes, based only on the numbers of conserved syntenies on each chromosome, the lengths of the

chromosomes, and a simplified random model of interchromosomal exchange. This estimator proves to be very accurate in simulations, which is remarkable given that it explicitly takes into account only the first-order effects of interchromosomal exchange.

In this paper, we applied our estimator to animal genome comparisons at various levels of resolution. This showed that translocation estimates are stable at coarse resolutions, while inversions increased markedly. This reflects the discovery of high numbers of smaller-scale local arrangements recognizable from genomic sequence (Kent *et al.*, 2003). At very detailed levels of resolution, inferred translocations numbers probably reflect processes other than translocation, though increased inversion inferences are more likely to reflect the inversion process.

Our estimates of the number of translocations and inversions in the evolutionary divergence of animals are only about half of what has been published by Bourque *et al.* (2005) for corresponding levels of resolution. Their estimates are based on an algorithmic reconstruction of the details of evolutionary history to account for how the segments are ordered on the chromosomes. Our Equation (1) assumes each translocation creates two new segments, but the algorithms used by Bourque *et al.* (2005), following Pevzner and Tesler (2003a), require a number of rearrangements almost equal to the number of segments, either because some segments are below threshold size, hidden to observation within small "hotspot" regions (Pevzner and Tesler, 2003b), or because other sources of noise interfere with the assumptions underlying the evolutionary reconstruction (Sankoff and Trinh, 2005). The general proportionality between the two sets of results assures that our observations in Section 8 do not depend on the differences in methodology.

ACKNOWLEDGMENTS

This work supported in part by a Discovery grant 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 Canadian Institute for Advanced Research.

REFERENCES

Bed'Hom, B. 2000. Evolution of karyotype organization in *Accipitridae*: A translocation model, *in* Sankoff, D., and Nadeau, J.H., eds., *Comparative Genomics*, 347–356, Kluwer, Dordrecht, NL.

Bourque, G., Zdobnov, E., Bork, P., Pevzner, P.A., and Tesler, G. 2005. Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome Res.* 15, 98–110.

Burt, D.W. 2002. Origin and evolution of avian microchromosomes. *Cytogenetic and Genome Research* 96, 97–112.

De, A., Ferguson, M., Sindi, S., and Durrett, R. 2001. The equilibrium distribution for a generalized Sankoff–Ferretti model accurately predicts chromosome size distributions in a wide variety of species. *J. Appl. Probab.* 38, 324–334.

Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D. 2003. Evolutions cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. USA* 100, 11484–11489.

Mefford, H.C., and Trask, B.J. 2002. The complex structure and dynamic evolution of human subtelomeres. *Nature Rev. Genet.* 3(229), 91–102.

Pevzner, P.A., and Tesler, G. 2003a. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc. Natl. Acad. Sci. USA* 100, 7672–7679.

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., Deneault, M., Turbis, P., and Allen, C.P. 2002. Chromosomal distributions of breakpoints in cancer, infertility and evolution. *Theoret. Population Biol.* 61, 497–501.

Sankoff, D., and Ferretti, V. 1996. Karotype distributions in a stochastic model of reciprocal translocation. *Genome Res.* 6, 1–9.

Sankoff, D., Ferretti, V., and Nadeau, J.H. 1997. Conserved segment identification. *J. Comp. Biol.* 4, 559–565.

Sankoff, D., and Mazowita, M. 2005. Estimators of translocations and inversions in comparative maps. *Proc. RECOMB 2004 Satellite Workshop on Comparative Genomics (RCG 2004), Lecture Notes in Bioinformatics* 3388, 109–122.

Sankoff, D., Parent, M.-N., and Bryant, D. 2000. Accuracy and robustness of analyses based on numbers of genes in observed segments, *in* Sankoff, D., and Nadeau, J.H., eds., *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and Evolution of Gene Families*, 299–306, Kluwer, Dordrecht, NL.

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

Schubert, I., and Oud, J.L. 1997. There is an upper limit of chromosome size for normal development of an organism. *Cell* 88, 515–520.

Trinh, P., McLysaght, A., and Sankoff, D. 2004. Genomic features in the breakpoint regions between syntenic blocks. *Bioinformatics* 20, I318–I325.

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

E-mail: sankoff@uottawa.ca