Genomic Midpoints: Computation and Evolutionary Implications

Richard Durrett* and Yannet Interian†

Dept of Mathematics, Cornell University, Ithaca NY 14853*

Dept of Bioengineering, U. of California, Berkeley CA 94720†

Running Head: Genomic Midpoints

Keywords: genome rearrangement, breakpoint graph, parsimony method, local search

Corresponding Author:

Rick Durrett

523 Malott Hall

Cornell University

Ithaca NY 14853

Phone (607) 255-8282

FAX (607) 255-7149

Email: rtd1@cornell.edu

ABSTRACT

In this paper we describe a method for computing genomic midpoints, which minimizes the sum of the distances to three given genomes, and apply the method to three data sets which have been analyzed previously by other methods: a human-lemur-tree shrew comparison that estimates the ancestral primate karyotype, several versions at different resolution of a human-cow-cat comparison, and a human-mouse-rat map with 424 markers. Rather than report one particular midpoint as the "right answer," as previous analyses have done, we produce a large number of possible midpoints. As in the reconstruction of phylogenies, comparing these solutions helps us understand the confidence we can have in various features of our predicted midpoint, and in our estimates of the number of events that occurred on various lineages.

INTRODUCTION

Genomes evolve not only by nucleotide substitutions but also by inversions that rearrange the order of genes on a chromosome, by translocations that exchange chromosomal material between chromosomes, and by fissions and fusions that change chromosome number. Hannenhalli and Pevzner (1995a) developed a polynomial time algorithm for computing the inversion distance between two chromosomes, and later extended this to compute the distance between two genomes (1995b). While the parsimony distance gives an estimate of the number of events that happened, it does not give much insight into what events occurred. For example, the human X chromosome can be transformed into the mouse X chromosome in a minimum of 7 inversions but there are thousands of shortest paths.

In order to determine what rearrangement events took place and when they occurred in evolution, we need to consider multiple species. Here we are thinking of examples where the phylogeny is known. Simon and Larget (2001) and Larget *et al.* (2002) have used genome rearrangements to estimate phylogenetic relationships.

Sankoff and Blanchette (1997, 1998a, 1998b) have considered our problem for the "breakpoint" distance, which is 1/2 the number of markers adjacent in one genome that fail to be adjacent in the other, rounded up to the next integer. That is, given n genomes $G_1, \ldots, G_n$ having a known phylogeny, one seeks genomes $H_1, \ldots, H_{n-2}$ for the internal nodes of the tree, so that the sum of the breakpoint distance between end points of edges of the tree is minimized. Blanchette *et al.* (1999) used BPANALYSIS, an implementation of the breakpoint analysis of Blanchette *et al.* (1997) on a problem with

11 genomes and 35 markers. More recently, an improvement of the BPANALYSIS, call GRAPPA has been developed by Moret *et al.* (2001).

Here we consider the simplest problem of finding the midpoint of three genomes. Hannenhalli, *et al.* (1995) were the first to do this for three herpes viruses. However, there were only 7 total changes in the minimum solution, so they could find the midpoint by examining all of the arrangements within a fixed distance of the three genomes. Bourque and Pevzner (2002) have recently proposed a new approach, the Multiple Genome Rearrangment Median (MGR-MEDIAN) algorithm based on the genomic distance which applies to n species. When n=3 the algorithm works in two stages. In the first stage, rearrangement events in a genome that bring it closer to each of the other two of the three genomes are carried out "in a carefully selected order." In the second stage, moves are accepted if they bring two genomes closer together.

Here we introduce a new approach to the median problem. Our Metropolis-Hastings type algorithm combines greedy search (rearrangements that reduce the sum of the distances) with non-improving moves that allow the algorithm to escape from midpoints that are local minima and to perform a more wide-ranging investigation of the set of possibilities. Experiments with simulated data reported here show that our method performs better than GRAPPA and MGR-MEDIAN. Application of our method to three biological data sets finds better solutions and gives new insights in their evolutionary history.

MATERIALS

We will study three data sets. The first is a three way comparison of human, lemur (*Eulemer macaco macaco*) and tree shrew (*Taupaia belangeri*). Müller *et al.* (1997) did a reciprocal painting between human and lemur, and Müller *et al*. (1999) did a reciprocal painting between human and tree shrew, and a painting of lemur chromosomes with tree shrew paints. There are 37 segments in the comparison with EMA and 39 with TBE. Subdividing to obtain a common collection of segments, we arrive at the 41 homologous segments. See Table 1 for the data. Chromosome painting does not give information about orientation, so we have assigned signs to segments to minimize the distance.

Using chromosome painting data in an era when there are many completely sequences genomes may seem odd, but there are three advantages of doing this. (i) It produces a small data set where the proposed midpoint can be inspected visually. (ii) We can compare our midpoint with earlier analyses experts did by hand. (iii) We will use chromosome painting data that exists for a large number of primate species to obtain independent verification of predictions of our midpoint computation.

Table 3 lists the primate species we will use. The data we use can be found in the supplementary materials. A reciprocal painting is more informative than a one-way painting by human chromosomes since it tells us the order of segments in the human genome. However, with the exception of the lemur and tree shrew data mentioned above, there are only two reciprocal paints between humans and primate species: with African green monkey (*C. aethiops*), Finelli *et al.*(1999), and with woolly monkey (*Lagothrix*

lagotricha), Stanyon *et al.* (2001). In all of the other cases the primate is painted with human paints: macaque (*Macaca fuscata*), Weinberg *et al.* (1992); black and white colobine monkey (*Colobus guereza*), Bigoni *et al.* (1997) ; marmoset (*Callithrix jacchus*), Sherlock *et al.* (1996), capuchin monkey (*Cebus capucinus*), Richard *et al.* (1996); squirrel monkey (*Saimiri sciureus*) and dusky titi monkey (*Callicebus molloch*) by Stanyon *et al.* (2000); black handed spider monkey (*Ateles geoffroyi*), Moreslachi *et al.* (1997); and black-and-red howler monkey (*Alouatta belzebul*), Consigliere *et al.* (1998). There are also one-way paintings of siamang (*Hylobates syndactus*), and of concolor gibbon (*H. concolor*) by Koehler *et al.* (1995ab). However, there have been a large number of intra- and interchromosomal rearrangements in the Hylobates lineage, so this comparison is not informative.

Our second data set is a comparison of human, cat, and cattle constructed by Murphy *et al.* (2003a). They have 300 markers on autosomes. We have deleted 12 markers whose position is in conflict with chromosome painting experiments of Hayes (1995) and Chowdhary *et al.* (1996) that compared human and cattle or work of Weinberg *et al.* (1997), and Murphy *et al.* (1999, 2000) that used chromosome painting results and a radiation hybrid map to compare humans and cats. In addition we deleted 3 markers to make block boundaries coincide and to allow the creation of smaller data sets described below.

To have problems of varying degrees of difficulty, we use various techniques to simplify the human-cow-cattle comparison. Four groups of chromosomes do not tangle with the others and can be analyzed separately.

I. Human 17 Cat E1, Cow 19.

II. Human 14,15, Cat B3, Cow 10, 21.

III. Human 6, Cat B2, Cow 9, 23.

IV. Human: 11, Cat D1, Cow 15, 29.

If we take the remaining chromosomes and perform two inversions in cattle and two inversions in cat then the three genomes can be divided into 38 syntenic segments that contain the same markers, but not necessarily in the same order. See Table 2. The same genomic data can be divided into 79 conserved segments that contain the same markers in the same order. Finally, we have the entire data set that consists of 285 markers in 118 conserved segments. The larger data sets are given in the supplementary materials.

The third data set is a comparison of human, mouse, and rat constructed by Colin Dewey and Lior Pachter which appeared in the April 1, 2004 issue of Nature devoted to the sequencing of the rat genome, see page 498.

METHODS

Given three genomes A, B, and C, and a notion distance of d, we seek a genome M that minimizes the total number of events d(A,M) + d(B,M) + d(C,M). To do this, we initialize the search process with a proposed midpoint and then proceed by iteratively making small changes in the midpoint. The proposed change is always accepted if it reduces the total number of moves, and with a fixed probability if it does not. A precise

description of our algorithm which we call MEDBYLS (Median by Local Search) can be found in the supplementary materials. The computer code can be found at

http://www.cam.cornell.edu/~interian/MEDbyLS_code.html

Rather than using the genomic distance of Hannenhalli and Pevzner (1995a, 1995b), which is difficult to compute, we use the simpler *graph distance* based on the number of components in the breakpoint graph. The precise definition of the distance is not important for the discussion of our results, so again we refer the reader to the supplementary materials for details. In most biological examples the graph distance is equal to the genomic distance, see e.g., Bafna and Pevzner (1995) or Durrett, Nielsen, and York (2004). Caparara (1999) has shown that for a randomly chosen permutation of n markers, the probability the two distances differ is of order $1/n^5$, so discrepancies between the two distances are very unlikely when the number of markers n is large.

Yancopoulos *et al.* (2005) have recently shown that if you add a new operation which they call double cut and join, then the genomic distance for this enlarged set of events is the graph distance. A consequence of this observation is that the graph distance is a distance, and hence satisfies the triangle inequality: $d(A,B)+d(B,C)\geq d(A,C)$. Consider for concreteness the comparison of human (H), lemur (L), and tree shrew (TS). The distances between the three genomes are d(L,H)=21, d(L,TS)=19, and d(H,TS)=16. As Hannenhalli *et al.* (1995) observed, if M is any midpoint then the triangle inequality implies

$$d(L,M)+d(M,H)\geq d(L,H)$$
$$d(L,M)+d(M,TS)\geq d(L,TS)$$
$$d(H,M)+d(M,TS)\geq d(H,TS)$$

Add the three equations and letting $D = d(L,M) + d(H,M) + d(TS,M)$ be the total number of events then $2D \geq d(L,H) + d(L,TS) + d(H,TS) = 56$ so $D \geq 28$.

We can improve this bound if we look at the three color breakpoint graph. This graph has vertices the doubled markers and genome ends, and an edge of color A connecting two vertices if they are adjacent in genome A. This graph can be separated into its connected components. Some of the small components for the human-cow-cat comparison are shown in Figure 1. It is easy to see that a solution which achieves the graph distance will not combine two components of the graph, so we can apply the triangle inequality bound separately to each component. In the case of the human-lemur tree shrew comparison the graph has 18 components, and an improved lower bound of 31 results. For small components it is easy to find moves that achieve the lower bound. To see what makes these problems hard look at Figure 2, which shows part of the largest component of the breakpoint graph.

PERFORMANCE ON SIMULATED DATA

We compare our algorithm with GRAPPA, an implementation of the breakpoint analysis (Moret et al. 2001) and with MGR-MEDIAN (Bourque and Pevzner, 2002) using simulated data. For unichromosomal data we start with the identity permutation with *n* markers, and we perform *k* random reversals to get each genome G_i. For the case of multichromosomal genomes we start with identity permutation with *n* markers, we break it in five identical pieces, and then *k* rearrangements are applied at random (with

probability 0.2 the rearrangements are translocations and with probability 0.8 are inversions).

Part a of Figure 3 gives a comparison of GRAPPA and MEDBYLS. The score is

$$\sum_{i=1}^{3} d(M, G_i) - 3k$$

the sum of the distance between the genomes G_i and the midpoint M found by the algorithm, minus $3k$, the sum of the distances between the genomes G_i and the identity permutation, which is the true historical midpoint. Part a of Figure 3 shows that the quality of the solutions found by GRAPPA and MEDBYLS are very similar until r=0.86. But for the values r=1 and r=1.14 GRAPPA did not finish some of the instances in 24 hours.

Part b of Figure 3 shows the comparison between MGR-MEDIAN and MEDBYLS. Note that below the ratio r=0.75 the results of MGR-MEDIAN and MEDBYLS are similar but for r=0.9 the midpoints that MGR-MEDIAN finds are far from optimal. The values of the MGR-MEDIAN experiment were taken from the paper (Bourque and Pevzner 2002). The solver web interface, the only publicly available version of this program, only allows data sets with at most 30 markers. [We were not able to get the executable from the authors].

As we can see from Figure 3 there are some values of the ratio r of rearrangements to markers for which GRAPPA and MGR-MEDIAN stop working while MEDBYLS still finds good solutions. In the case of GRAPPA the time for finding solution blows up for r=1 and r=1.14 and for r=0.9 MGR-MEDIAN does not find good solutions.

Looking at Figure 1, the reader might be surprised to see negative values. However, these are to be expected. If the number of rearrangements on each lineage $k \geq n/4$, which corresponds to $r \geq 0.75$ the results in Berestycki and Durrett (2005) imply that

one can get from G_i to G_j in fewer that $2k$ steps. Thus, it is not surprising that the median we found is closer than the ancestral genome.

Figure 4 shows the results on simulated data for genomes with five chromosomes and n = 50, 100, 150, and 200 markers. For n = 50 and n = 100 we believe MEDBYLS finds midpoints very close to the best values, since we see curves similar to the ones from unichromosomal genomes. For n = 150 and 200 and for r greater than 0.6 MEDBYLS starts having difficulty finding the best midpoints. In particular, for n = 200 and large values of r the score is positive, while taking M to be the identity permutation results in a score of 0.

The Median problem is NP-hard in general Caprara (1999a). However, for many biological and random data seems relatively easy to find optimal or near-optimal solutions. One possible explanation for why algorithms are not finding optimal solutions in some situations comes from changes in the structure of the breakpoint graph. Consider our randomly generated data. Let $r = 3k/n$ where n is the number of markers and k the number of rearrangements along each lineage from the identity genome. The breakpoint graph for random data is a random graph. In most random graph models, as a certain parameter increases the size of the largest component abruptly changes from having size of order log n to a giant component with size of order n, see e.g., Durrett (2006). We think that this phenomenon is responsible for the increase in difficulty in this random data. The plots in Figure 5 show the fraction of the number of markers in the largest component of the three-genome breakpoint graph as a function of r. The appearance of the giant component for this model seems to take place around r = 0.6.

When *r* is small, all the components are small. Intuitively, one can then attack the problem by considering the components separately. This cannot always be done, since this might lead to a midpoint with a circular chromosome. However, in practice one can use this approach on many problems and it breaks even very large problems into a number of simple small ones.

RESULTS FOR COMPARATIVE MAPS

Human-lemur tree shrew. As mentioned in the methods section, applying the triangle inequality bound to the components of the breakpoint graph shows that at least 31 events are needed to produce the relationship between the three species. Table 1 gives our midpoint *M* which has *d(L,M) = 12, d(TS,M)=10, d(H,M)=10* for a total of 32 events. By examining the human-lemur-tree shrew breakpoint graph in more detail one can prove (results not shown) that the lower bound of 31 cannot be achieved and 32 is the minimum distance.

Müller *et al.* (1999) have proposed that the primitive primate karyotype consists of human autosomes 1a, 1b, 2a, 2b, 3/21, 4 to 11, 12a/22a, 12b/22b, 13, 14/15, 16a, 16b, 17, 18, 19a, 19b, and 20. Our interpretation of this midpoint *N* in terms of our segments is given in Table 1. We have performed two inversions in human chromosome 3 since this improves the performance of their solution: *d(L,N)=17, d(TS,N)=14, d(H,N)=7* for a total of 38 events. Note that the expert solution has many fewer events in the human genome, while ours distributes the changes almost equally over the three lineages.

Our computer generated solution has some features in common with the expert analysis of Müller *et al.* (1999). To check the accuracy of these predictions we will use the chromosome painting results mentioned earlier to make inferences about changes in the human genome. We are certainly not the first to have done this. For an extensive expert analysis of chromosome rearrangements in primates, see Haig (1999).

The first step is to observing that humans have 22 autosomes compared to 23 for chimp, gorilla, and orangutan. Human chromosome 2 is the result of a fusion of two chimpanzee chromosomes, 12 and 13. To locate other events within the primate lineage, we will use the results of chromosome painting between humans and the primate species described in the methods. Family names and three letter species abbreviations can be found in Table 3.

As expected human chromosome 2 corresponds to two chromosomes in each of the *Cercopithedae*. In all other cases human chromosomes paint only one MFU chromosome. Chromosomes 14 and 15 paint two parts of MFU7 and of CGU6, and in CAE we have 14 = 29.2+24 and 15=26+29.1, so it seems likely that HSA14 and 15 were created by a fission after the divergence from the *Cercopithedae*, and the ancestral 14/15 experienced two fissions in the CAE lineage to create its chromosomes 24, 29, and 26.

Moving to a comparison with New World Monkeys, there are twelve HSA chromosomes that are conserved in at least one of the species. This supports the presence of human chromosomes 7, 9, 11, 13, 17, and 20 in the ancestral genome. Chromosomes 4, 5, and 6 are all conserved in one of the species supporting Muller *et al.*'s solution. However these chromosomes are badly fragmented in lemur and tree shrew, which supports ours.

The fission of chromosome 1 (here and in what follows we refer to events as they would be seen moving backwards in time) is supported by the fact that it corresponds to three chromosomes in Cebidae and Pithicidae and to four chromosomes in Atelidae. Murphy *et al.* (2003b) have compared chromosome 1 with a number of other species to conclude that the fission occurred in 1q23.

Again chromosome 2 paints exactly two segments in all of the New World Monkeys except AGE where it paints three. Casting some doubt on our previous conclusion, HSA14/15 corresponds to two segments in all seven genomes. However if one looks at the q arm of SSC chromosome 2 on p.102 of Stanyon *et al.* (2001) one sees 14/15/14/15/14/15 suggesting that the fused chromosome underwent several inversions before being split by a translocation.

The association of 3 and 21 is supported by the fact that in 6 out of 7 cases one New World monkey chromosome contains parts of 3 and 21. However 3/21 corresponds to parts of three chromosomes in these six species, which is more consistent with our 3a, 3b/21 solution rather than a single ancestral 3/21. For a more detailed look at the evolution of 3/21 association in primates, see Figure 3 in Muller *et al.* (2000) which gives the result of painting with tree shrew chromosomes 6, 7, 24, and 28, which combine to make 3/21, see Table 1.

The translocation between 12 and 22, which is clearly present in lemurs and tree shrews is not evident in new world monkeys, so we assume this event preceded their divergence. Chromosomes 8 and 18 show a pattern that can be explained by a single translocation that for example changes 8a,8b and 16 into 8a,16 and 8b. However there is

no evidence of this in either solution, so perhaps this occurred in the New World Monkey lineage after divergence from humans.

Human-cow-cat. Table 4 gives the distances between each pair of genomes, the lower bound of Hannenhalli *et al.* (1995), and the number of events on the three lineages in one of our best solutions for each of our three data sets. Since cat and cow both begin with the letter c, our one letter abbreviations will be f for feline, and b for bovine. The first data set of syntenic segments primarily identifies translocations. Note that as the resolution of the data set increases, we not only see more events, but the relative rates of events on lineages changes dramatically. For example, as the number of segments increases from 38 to 118, the number of events on the human lineage only increases from 12 (34.3%) to 18 (16.4%) while the number on the cattle lineage goes from 14 (40%) to 56 (50.9%), and as the parenthetical numbers indicate, the percentage of the events on the different lineages changes dramatically.

Murphy *et al.* (2003a) have analyzed this data using the MGR-MEDIAN algorithm. This method has a parameter G that is a distance threshold used to filter out spurious markers that occur at isolated points. When G=4 singletons are deleted, while increasing G allows for more complex microrearrangments. The solution they present in their Figure 2 has G=6. This uses 276 of the 300 markers, compared to our selection of 285. They find distances d(h,m)=16, d(f,m)=21, d(b,m)=27 for a total of 64 events. In contrast to our result d(h,m)=18, d(f,m)=36, d(b,m)=56 that has a total of 110 events. This discrepancy is not due to a failure of our algorithm to find a good solution. The lower bound of Hannehalli *et al.* is 103 in our case and that results is not sharp.

Murphy *et al.*'s (2003a) human-cow-cat midpoint as well as their reconstruction of the boroeutherian ancestor by Murphy *et al.* (2005) present one solution as the right answer. However, in reality there are some aspects of the solution that are known with high confidence, but other predictions are less reliable. To quantify this, our approach is to find 100 good solutions and see how many times each adjacency occurs in the solutions. We present the results for the 38 marker data set in Table 5.

Note that all adjacencies that are present in two genomes are in 100% of the midpoints. To make contact with the approach of Bourque and Pevzner (2002), these adjacencies are associated with moves that will bring one genome closer to the other two. The adjacencies 24,25, 28,29, and 40,41 correspond to chromosomes 4, 5, and 9 being present in the ancestor. The isolated chromosomes described in the Materials section give us four more ancestral associations: 6, 11, 14/15, and 17. From the fact that 16,17 never occurs in the midpoint, we infer the fission of chromosome 2 in the human lineage. The adjacency 21,69 is the association of an inversion of the end of 3 with 21, i.e., -11,35 in terms of the original markers. The 53,71 and 52,73 associations which are present in 89 and 61 of the solutions suggest a 12 x 22 translocation, but in contrast to the human-lemur-tree shrew midpoint the result is 12a,22b and 22a,12b. Comparison with radiation hybrid maps of the pig and goat genomes suggest that 12a,22a and 22b,12b is the correct event. Perhaps rearrangements in cow 5 and 17 or cat B4 and D3 have confused the reconstruction.

To understand why some adjacencies in the solution are certain and others are not we look at some of the small connected components in the three species breakpoint graph in Figure 1. In part a of that figure, 2 and 3 are adjacent in human and cat, so we always

do a translocation of B-2 and 3-B in cow, i.e., a fusion that makes 2 and 3 adjacent. In part b, 19 and 31 are ends in both human and cow, so we always do a fission of 19-31 in cat. In part c, a translocation of 18-4 and 5-B in cow is always done. However in part d there are several ways of reducing this component and different solutions use different methods.

There are many other small components, most of which are similar to one of the examples drawn. The hard part of the computational problem comes from the "giant component" that contains 32 of the 76 markers. The appearance of a giant component is a well known phenomenon in random graphs. Following up on simulation work of Bourque and Pevzner (2002), Berestycki and Durrett (2005) have studied the situation of the reversal distance between two chromosomes. They have shown that if there are n markers and cn inversions then for $c < ½$ all the components are small and the computational problem is easy, but when $c > ½$ then there is a giant component, which makes the problem hard and causes the parsimony method to underestimate the actual distance.

We have drawn part of the giant component in Figure 2. Note that adjacencies which occur in two of the genomes occur in 100% of the solutions and this forces some nearby adjacencies to have frequencies 0 or 100%. However, many of the adjacencies have intermediate frequencies. At some of the nodes the numbers add up to less than 100 since these markers have an adjacency in the midpoint that does not occur in any of the three genomes.

Human-mouse-rat. In our final example we concentrate on the inference of the number of events rather than trying to reconstruct the changes. For reconstructions of the

midpoint, see Bourque et al. (2004, 2005), or Murphy et al. (2005). Figure 3 gives the distribution of the distances to the midpoint in 100 runs where the total distance was 347 in 98 cases and 346 in 2. The average distance to the midpoint are 43.01 for mouse, 62.75 for rat, and 241.22 for human, in contrast to the distances of 50, 56 and 247 and a total of 353 reported by the Rat Genome Sequencing Project Consortium (2004). Rather than reporting a single number, the range of values in 100 midpoints gives us some information about the uncertainty in the parsimony solution. It interesting to note that in all cases the number of events on the mouse lineage is larger than that on the rat lineage. However, we do not know how to assign a p value to assert our confidence in this conclusion.

If we use 15 million years for the divergence of mouse and rat, and 90 million years for their divergence from human, then the human to midpoint branch is 165 million years and our estimates of events per million years are 3.01 for mouse, 4.33 for rat and 1.48 for human. This contrasts with the estimates of 2.25 for mouse and 1.25 for rat from Figure 3 of Murphy *et al.* (2005) and it is interesting to note that the answer to the question: "do rats have more frequent rearrangements than mice?" is different

Murphy *et al.* (2005) give an estimate of 0.39 events per year for the 90 million years since divergence, and 2.11 per year on the branch from the divergence to the mouse-rat split. We cannot estimate rates separately for the two branches but weighting the rates by the interval lengths (0.39)(90/165) + (2.11)(75/165) = 1.17, gives a rate smaller than our 1.48.

CONCLUSIONS

Here we have developed a local search algorithm for the genomic midpoint problem, which performs better than previous approaches on simulated data, and we have applied it to three interesting biological examples. When applied to the human-lemur tree shrew data, it produces a more even distribution of events between lineages than the expert analysis, and the solution is consistent with rearrangements that can be inferred using chromosome painting data with primates. In contrast to methods that produce one solution and declare it to be the right answer, our approach is to produce 100 solutions and to place higher trust in features that are common to many of the solutions. In a similar way we can look at the distribution of the number of events on the various branches to understand the accuracy of those estimates. Our approach of producing many solutions is similar to phylogenetic methods that indicate support for branching in the reconstructed tree.

We would have liked to have taken a Bayesian approach to the midpoint problem, which produces a posterior distribution and a more rigorous analysis of the uncertainties in the answer, but we were discouraged by the convergence problems that Durrett *et al.* (2004) had in their two species comparisons. As a substitute for a Bayesian posterior, we will take advantage of the fact that local search will produce a large number of good solutions in order to gain insights into the reliability of the answer. Taking a cue from phylogenetic algorithms that indicate bootstrap support for various branches, we run the algorithm many times from random starting points, note the number of times each adjacency occurs in a family of solutions, and place a higher trust in the features that are

present in a large fraction of the solutions. In a similar way we can look at the distribution of the number of events on the various branches to understand the accuracy of those estimates. We believe that this is an important improvement compared to methods that give one midpoint which they declare to be the answer, see e.g., Murphy et al. (2003a, 2005).

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

Berestycki, N. and Durrett, R., (2005) A phase transition in the random transposition random walk. *Prob. Theory and Related. Fields*, to appear

Bigoni, F., Stanyon, R., Koehler, Morescalchi, A.M., and Weinberg, J. (1997) Mapping homology between human and black and white spider monkey chromosomes by fluorescent in situ hybridization. *Am. J. Primatology*. 42, 289-298

Blanchette, M., Bourque, G., and Sankoff, D. (1997) Breakpoint phylogenies. Pages 25-34.in *Genome Informatics* 1997, Miyano, S., and Takagi, T., eds., Universal Academy Press, Tokyo.

Blanchette, M., Kunisawa, T., and Sankoff, D. (1999) Gene order breakpoint evidence in animal mitochondrial phylogeny. *J. Mol. Evol*. 49, 193--203.

Blanchette, M., Green, E.D., Miller, W., and Haussler, D. (2004) Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Research.* 14, 2412-2423

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

Bourque, G., Pevzner, P.A. and Tesler, G. (2004) Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome Research.* 14, 507-516

Bourque, G., Zdobonk, E.M., 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 Research.* 15, 98-110

Caprara, A. (1999a) Formulations and hardness of multiple sorting by reversals. Pages 84-93 in Proceedings 3rd Conf. Computational Molecular Biology RECOMB99, ACM Press, New York

Caprara, A.(1999b) On the tightness of the alternating-cycle lower bound for sorting by reversals. *Journal of Combinatorial Optimization*. 3,149-18

Caprara, A. (2003) The Reversal Median Problem. INFORMS Journal on Computing. 15, 93-113

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

Consigliere, S., Stanyon, R., Koehler, U., Arnold, N., and Weinnberg, J. (1998) In situ hybridization (FISH) maps chromosomal homologies between Alouatta belzebul (Platyrrhini, Cebidae) and other primates reveals extensive interchromosomal rearrangments between howler monkey genomes. *Am. J. Primatology*. 46, 119-133

Doganlar, S., A. Frary, M. C. Daunay, R. N. Lester, and S. D. Tanksley, 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. (2006) *Random Graph Dynamics*. Cambridge University Press.

Durrett, R., Nielsen, R., and York, T.L. (2004) Bayesian estimation of genomic distance. *Genetics.* 166, 621—629

Finelli, P, Stanyon, R, Plesker, R, Ferguson-Smith, MA, O'Brien, PCM, Wienberg, J. (1999) Reciprocal chromosome painting shows that the great difference in diploid number between human and African green monkey is mostly due to non-Robertsonian fissions. *Mammalian Genome.* 10, 713-718

Haig, D. (1999) A brief history of human autosomes. *Phil. Trans. Roy. Soc., B*. 354, 1447-1470

Hannenhalli, S., Chappey, C., Koonin, E.V., and Pevzner, P.A. (1995) Genome sequence comparisons and scenarios for gene rearrangements: A test case. *Genomics.* 30, 299-311

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

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

Kececioglu, J.D. and Sankoff, D. (1995) Exact and Approximation Algorithms for Sorting by Reversals, with Application to Genome Rearrangement. *Algorithmica,* 13, 180-210

Koehler, U. Arnold, N., Weinberg, J., Tofanelli, S., and Stanyon, R. (1995a) Genomic rearrangement and disrupted chromosome synteny in siamang (*Hylobates syndactylus*) revealed by Flourescene In Situ Hybridization. *Am. J. Physical. Anthropology.* 97, 37-47

Koehler, U., Bigoni, F., Weinberg, J., and Stanyon, R. (1995b) Genomic reorganization in concolor gibbon (Hylobates concolor) revealed by chromosome painting. *Genomics* 30, 287-292

Larget, B., Simon, D.L., and Kadane, J.B. (2002) Bayesian phylogenetic inference from animal mitochondrial genome rearrangements. *J. Roy. Stat. Soc.* 64, 681-693

Morescalchi, M.A., Schempp, W., Consgliere, S., Biogni, F., Weinberg, J., and Stanyon, R. Mapping chromosomal homology between humans and the black-handed spider monkey by fluorescence in situ hybridization. *Chromosome Research* 5, 527-536

Moret, B., Wyman, S., Bader, D., Warnow, T., and Yan, M. (2001) A new implementation and detailed study of breakpoint analysis. Pages 583-594 in Proceedings of the 6th Pacific Symposium on Biocomputing, Hawaii

Müller, S., O'Brien, P.C.M., Ferguson-Smith, M.A., and Weinberg, J. (1997) Reciprocal chromosome painting between human and prosimians (*Elemer macaco macaco* and *E. fulvus mayottensis*). *Cytogenet. Cell Genet*. 78, 260-271

Müller, S., Stanyon, R., O'Brien, P.C.M., Ferguson-Smith, M.A., Plesker, R., and Weinberg, J. (1999) Defining the ancestral karyotype of all primates by multi-directional chromosome painting between tree shrews, lemurs, and humans. *Chromosoma.* 108, 393-400

Müller, S., Stanyon, R., Finelli, P., Archidianco, N., and Weinberg, J.(2000) Molecular cytogenetic dissection of human chromosomes 3 and 21 evolution. *Proc. Nat. Acad. Sci*. 97, 206-211

Murphy, W.J., Bourque, G., Tesler, G., Pevzner, P., and O'Brien, S.J. (2003a) Reconstructing the genomic architecture using multispecies comaprative maps. *Human Genomics* 1, 30-40

Murphy, W.J., Frönicke, L., O'Brien, S.J., and Stanyon, R. (2003b) The origin of human chromosome 1 and its homolog in placental mammals. *Genome Research*. 13, 1880-1888

Murphy, W.J., Larkin, D.M., Everts-van derWind *et al.* 25 coauthors. (2005) Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* 309, 613-617

Murphy, W.J., Sun, S., Chen, Z., Pecon-Slattery, J., and O'Brien, S.J. (1999) Extensive conservation of sex chromosome organization between cat and human revealed by parallel radiation hybrid mapping. *Genome Research.* 9, 1223-1230

Murphy, W.J., Sun, S., Chen, Z., Yuhki, N., Hirschman, D., Menotti-Raymond, M, O'Brien, S.J. (2000) A radiation-hybrid map of the cat genome: implications for comparative mapping *Genome Research.* 10, 691-702

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

Rat Genome Sequencing Project Consortium (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature.* 428, 493-521

Richard, F., Lombard, M., and Dutrillaux, B. Zoo-FISH suggests a complete homology between human and capuchin monkey (pltayrrhini) euchromatin. *Genomics.* 36 (1996), 417-423

Sankoff, D., and Blanchette, M. (1997) The median problem for breakpoints in comparative genomics. *Computing and Combinatorics, Proceedings of COCOON `97* Edited by T. Jiang and D.T. Lee. Springer Lecture Notes in Computer Science 1276, pages 251--263

Sankoff, D., and Blanchette, M. (1998a) Multiple genome rearrangement and breakpoint phylogeny. *J. Computational. Biology.* 5, 555-570

Sankoff, D., and Blanchette, M. (1998b) Multiple genome rearrangement. *RECOMB 98*, 243-247

Sherlock, J.K., Griffin, D.K., Delhanty, J.D.A., and Parrington, J.M. Homologies between human and marmoset (*Callithrix jacchus*) revealed by comparative chromosome painting. *Genomics* 33, 214-219

Simon, D.L. and Larget, B. (2001) Phylogenetic inference from mitochondrial genome rearrangement data. Springer Lecture Notes in Computer Science 2074, 1022-1028

Stanyon, R., Consigliere, S., Bigoni, F. Ferguson-Smith, M., O'Brien, P.C.M., and Weinberg, J. (2001) Reciprocal chromosome painting between a New World primate, the wooly monkey, and humans. *Chromosome Research.* 9, 97-106

Stanyon, R., Consigliere, S., Müller, S., Morescalchi, A., Neusser, and Weinberg, J. (2000) Flourescence in situ hybridization (FISH) maps chromosomal homologies between the dusky titi and squirrel monkey. *Am. J. Primatol.* 50, 95-107

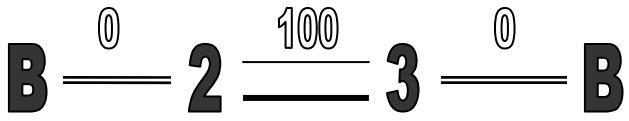
Tesler, G. (2002) GRIMM: genome rearrangement web server. *Bioinformatics* 18, 492-493

Weinberg, J., Stanyon, R., Jauch, A., and Cremer, T. (1992) Homologies in human and Macaca fuscata chromosomes revealed by in situ suppression hybridization with human chromosome specific DNA libraries. *Chromosoma.* 101, 265-270

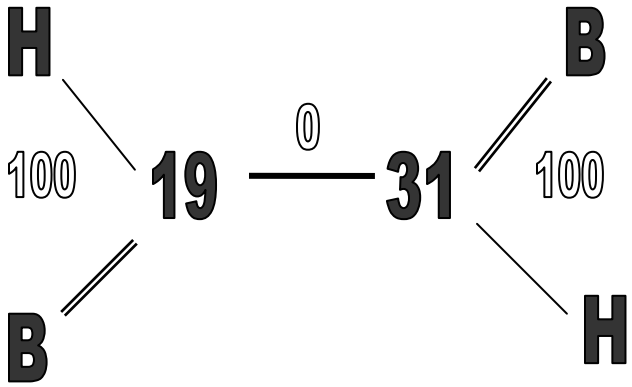
Wienberg, J, Stanyon, R, Nash, W.G., OBrien, P.C.M., Yang, F, O'Brien, S.J., Ferguson-Smith, M.A. (1997) Conservation of human vs. feline genome organization revealed by reciprocal chromosome painting. *Cytogenetics and Cell Genetics* 77, 211-217

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

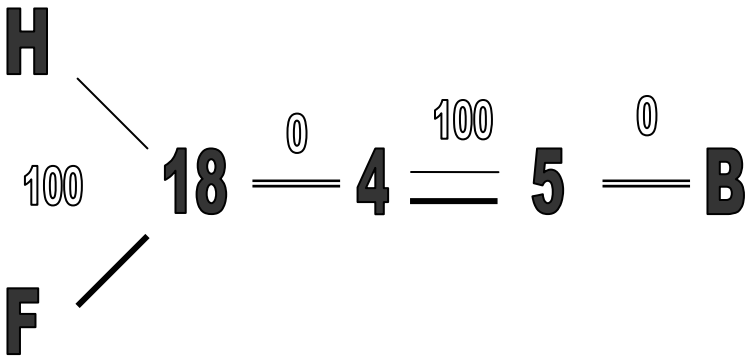
a.



b.



c.



d.

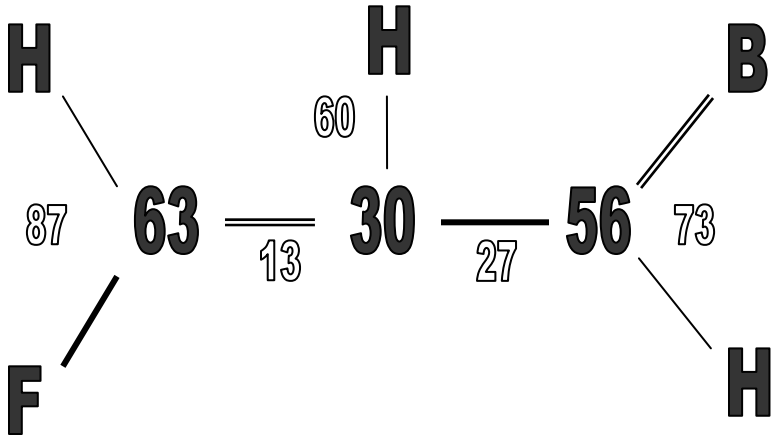


Figure 1. Small components of the three species breakpoint graph. Thin lines indicate adjacencies in human, thick lines in cat, and double lines in cow. Dark numbers are markers, light numbers are the number of times the adjacency was observed in the midpoint. In the first three cases there is one move that is always performed. In the last example there are several different ways of reducing the graph.

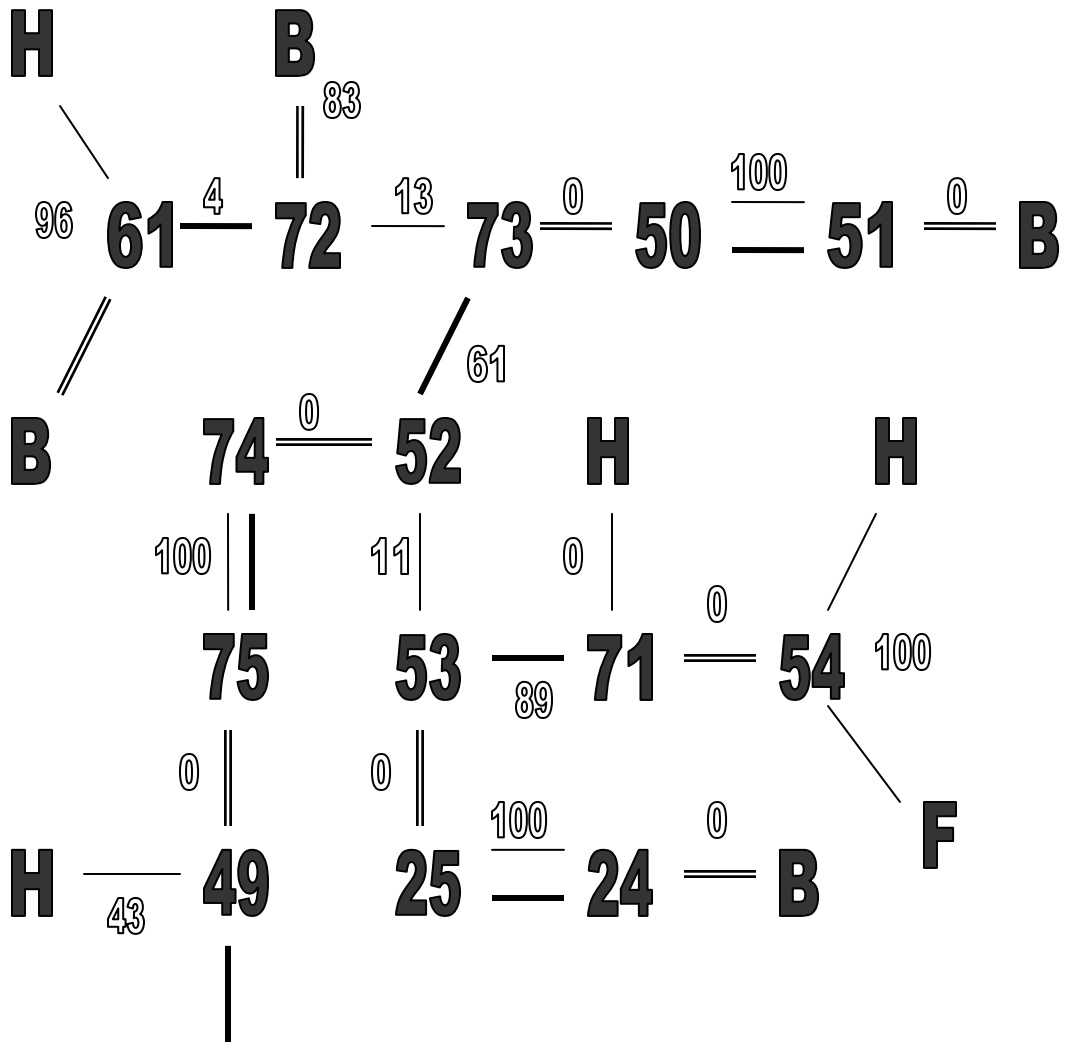
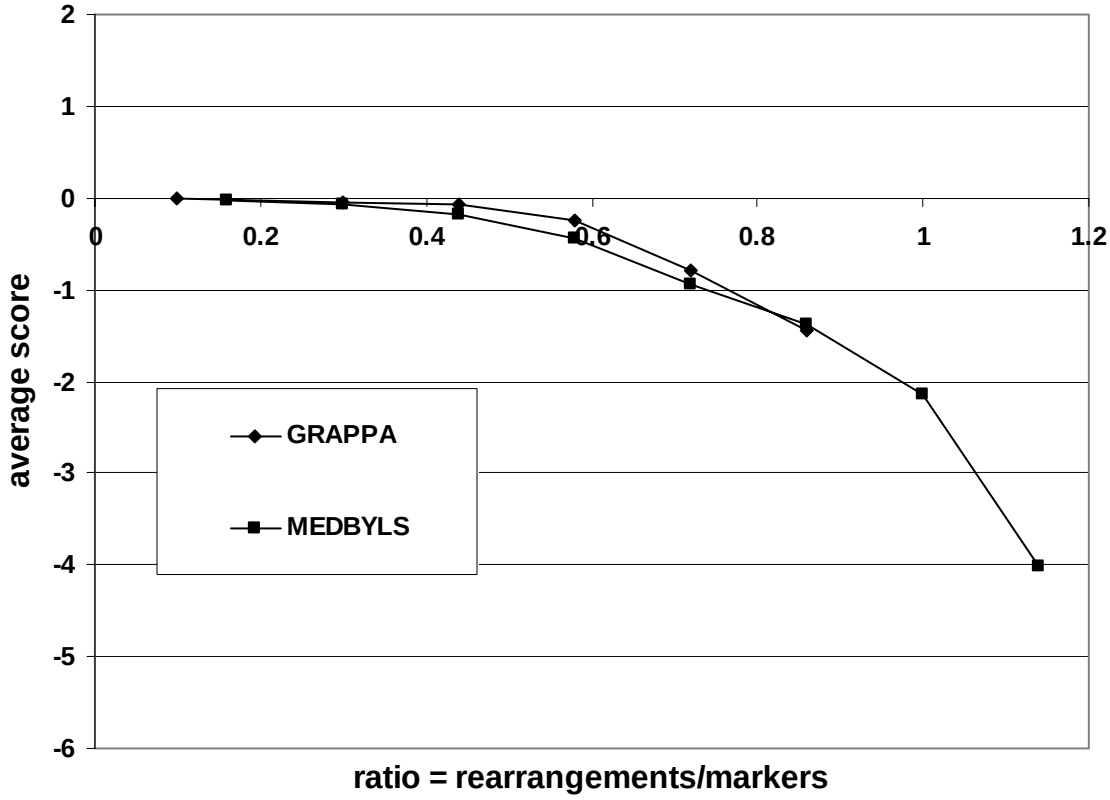


Figure 2. Frequency of midpoint adjacencies in one part of the giant component. Notation as in Figure 4.

a.


2
1
0
-1
-2
-3
-4
-5
-6
0
0.2
0.4
0.6
0.8
1
1.2
average score
GRAPPA
MEDBYLS
ratio = rearrangements/markers

b.

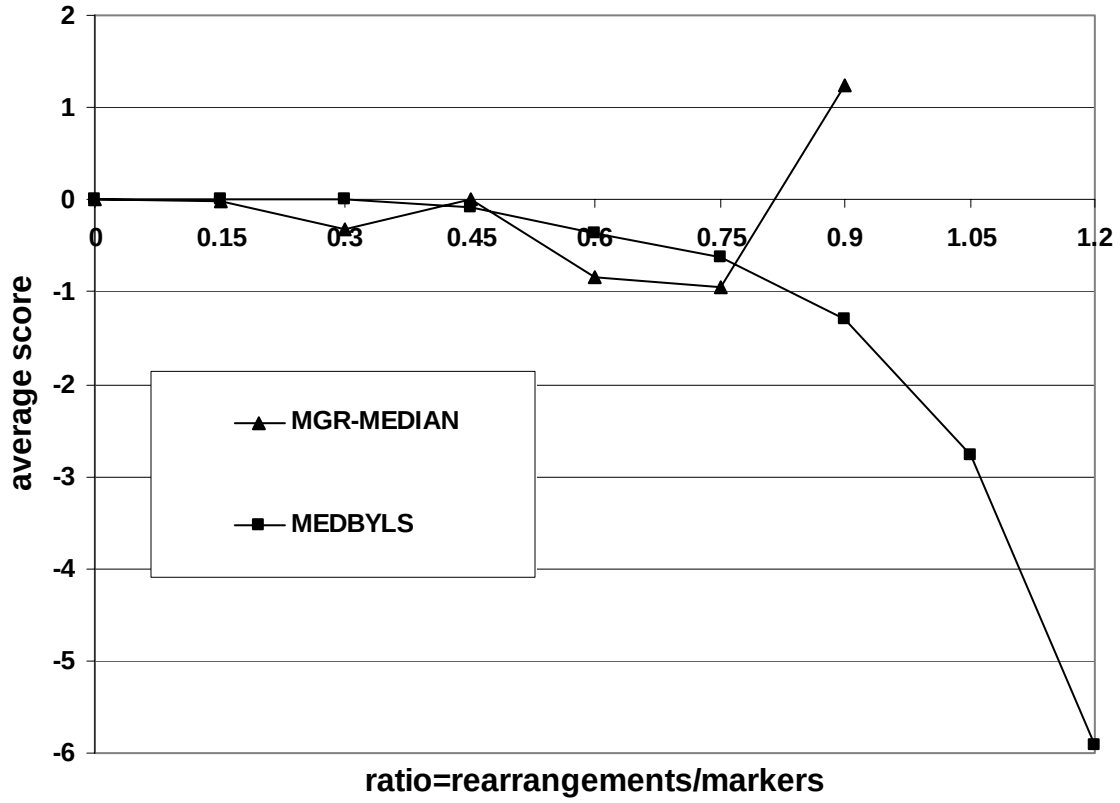


Figure 3. The score is the sum of the distances between the midpoint found by the algorithm and the three original genomes minus *3k*, where *k* is the number of inversions performed. Each point is the average value of the score for 30 simulations, as a function of the ratio k/n, where n is the number of markers. The first panel gives comparisons between GRAPPA and MEDBYLS for n = 150. The second the comparison between MGR-MEDIAN and MEDBYLS for n = 100.

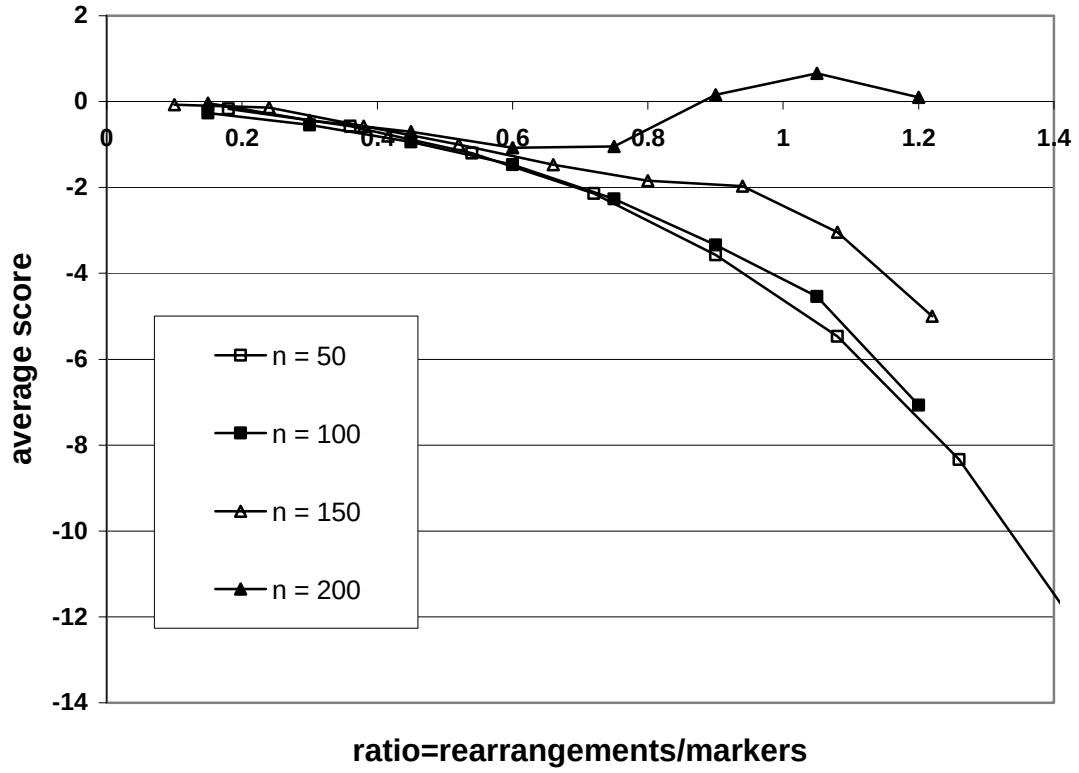


Figure 4. Results from MEDBYLS for genomes with five chromosomes and n = 50, 100, 150, 200 markers. As in Figure 3, each point is the average value of the score for 30 simulations as a function of $r = k/n$, where k is the number of rearrangements.

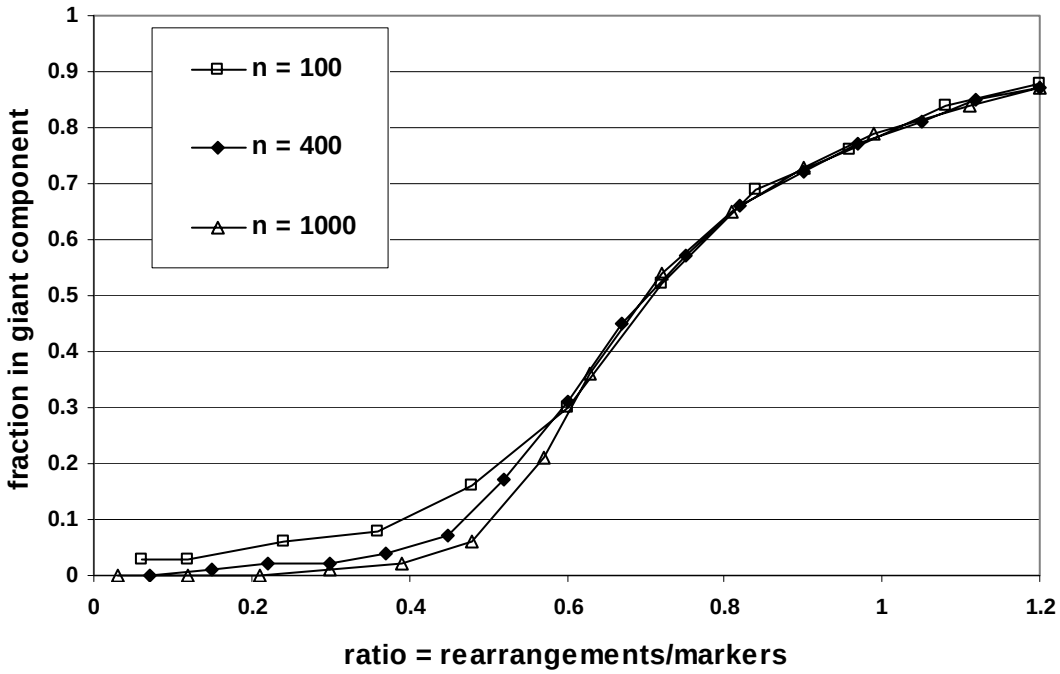


Figure 5. Fraction of the number of markers in the largest component of the breakpoint graph as a function of $r = k/n$, where n is the number of markers and k is the number of rearrangements.

Human	Lemur	Tree shrew	
1. 1,2,3	1. -22, -7,-6,-9,-8,39	1. 32,23,22	21. 15
2. 4,5	2. -5, -21, -31	2. -26, -38	22. 13
3. 6,7,8,9	3. 18, -14	3. 36	23. 11
4. 10,11,12,13	4. -29,-26,-25	4. 18,19	24. 7
5. 14,15	5. -24,27,-40	5. -37,-33	25. 10
6. 16,17	6. -38,10, 4	6. 9	26. 12
7. 18,19	7. 33,-36,30	7. -5,-39,8	27. 6
8. 20,21	8. -35,16,12	8. 17	28. -28,41
9. 22	9. 1	9. 27,-40	29. 16
10. 23,24	10. 32,-34	10. 1	30. 31
11. 25,26	11. 17	11. 30	
12. 27,28	12. -37,13	12. 4	
13. 29	13. -28,41	13. 20,21	
14/15. 30,31	14. 2	14. 14	
16. 32,33	15. 15	15. 35	
17. 34	16. 20	16. 24	
18. 35	17. 23	17. 29	
19. 36,37	18. 3	18. 34	
20. 38	19. 19	19. 2,3	
21. 39	20. -11	20. 25	
22. 40,41			

Our midpoint		Müller et al. (1999)	
1a. 1	7. 18, 19	1a. 1	12a/22a. 27,-40
1b. 2, 3	9. 22	1b. 2 3	12b/22b. -28,41
2a. 4	10. 23, 24	2a. 4	13. 29
-5, -21, -20	11. 25, 26	2b. 5	14/15. 30 31
3a. 6, 7	12a/22a. 27, -40	8. 20 21	16a. 32
3b/21. -9, -8, 39	12b/22b. -28, 41	3/21.-7 -6 -9 -8 39	16b. 33
10	13. 29	4.10 11 12 13	17. 34
11	14/15. 30, 31	5. 14 15	18. 35
-12, -16	16/19a. 32, 33, -36	6. 16 17	19a. 36
-13, 37	17. 34	7. 18 19	19b. 37
14	18. 35	9. 22	20. 38
15	20. 38	10. 23 24	
17		11. 25 26	

Table 1. Human genome compared to lemur and tree shrew.

Human	Cow	Cat
1: 1,2,3,4,5,6,7	1: -35,11	A1: 28,-15,-14
2: 8,9	2: 9,-2	A2: 32,-10,16
3: 10,11	3: -4,-3	A3: -34,-8
4: 12,13	4: 16	B1: 18,-13,-12
5: 14,15	5: 26,-37,-25,38	B4: 22,25,26,37,38
7: 16,17	6: 12	C1: 1,2,3,9
8: 18,19	7: -32,-15	C2: -35,11
9: 20,21	8: 20	D2: -7,-23,24
10: 22,23,24	11: -8,21	D3: -27,36,31
12: 25,26,27	12: 28	D4: 20,21
13: 28	13: -22,34	E2: -33,30
16: 29,30	14: -19	E3: 17,-29
18: 31	16: 6,5,1	F1: 5,-4,6
19: 32,33	17: -13,27,36	F2: 19
20: 34	18: -30,33	
21: 35	20: 14	
22: 36,37,38	22: 10	
	24: 31	
	25: 29,-17	
	26: 24	
	27: 18	
	28: 7,-23	

Table 2. 38 marker human-cat cow comparison.

Order Scadentia *Tupaia belangeri* (TBE) tree shrew

Order Primates

Suborder Lemuriformes

Family Lemuridae *Eulemer macaco* (EMA) black lemur

Suborder Anthropoidea

Infraorder Platyrrhini – New World Monkeys

Family Cebidae

Cebus capucinus (CCA) capuchin monkey
Samirini sciureues (SSC) squirrel monkey
Callithrix jacchus (CJA) marmoset

Family Pithecidae

Callicebus moloch (CMO) dusky titi monkey

Family Atelidae

Alouatta belzebul (ABE) howler monkey
Ateles goeffroyi (AGE) spider monkey
Lagotrix lagotrica (LLA) woolly monkey

Infraorder Catrrhini – Old World Monkeys

Family Cercopithecidae

Colobus guereza (CGU) colobine monkey
Cercopithecus aethiops (CAE) African green monkey
Macaca fuscata (MFU) macaques

Family Hominidae

Tribe Hyloblatini

Hylobates concolor (HCO) gibbon
Hylobates syndactus (HSY) siamang

Tribe Hominini

Pongo pygmaeus (PPY) orangutan
Gorilla gorilla (GGO) gorilla
Pan troglodytes (PTR) chimpanzee
Homo sapiens (HSA) humans

Table 3. Primate species in chromosome painting experiments, listed in decreasing order of distance from humans.

Markers	38	79	all
d(f,h)	18	35	51
d(f,b)	22	56	82
d(h,b)	23	51	72
lower bound	31	71	103
d(f,m)	9 (25.7%)	23 (30.7%)	36 (32.7%)
d(h,m)	12 (34.3%)	16 (21.3%)	18 (16.4%)
d(b,m)	14 (40%)	36 (48%)	56 (50.9%)
total	35	75	110

Table 4. Comparison of human (h), cow (b for bovine), cat (f for feline) based on our three data sets, and distances from the midpoint (m) in our solution.

always

2,3 hf; 4,5 hf; 6,7 hf; 21,69 bf; 24,25 hf; 28,29 hf; 34,58 bf; 40,41 hf; 50,51 hf; 59,65 bf; 74,75 hf; 16,67 f;

more than 50%

89:53,71 f; 68:10,11 h; 61:52,73 f; 56:44,49 f; 54: 8,9 h;

less than 50%

49:46,47 h; 37:12,13 h; 27:30,56 f; 26:36,37 h; 25:14,46 b; 23:13,46 f; 19:44,45 h; 19:26,36 f; 17,8,10 f; 15:1,10 b; 13:72,73 h; 13:30,63 b; 13:45,47 f; 11:52,53 h;

never

16,17 h; 20,21 h; 32,33 h; 58,59 h; 64,65 h; 4,18 b; 9,12 b; 15,41 b; 25,53 b; 43,67 b; 49,75 b; 50,73 b; 52,74 b; 54,71 b; 6,17 f; 7,11f; 19, 31f; 20,64 f;

Table 5. Adjacencies in 100 midpoints for the 38 marker human-cow-cat comparison. Markers have been doubled so the range of values is 1 to 76. Letters indicate the genome(s) in which the adjacency is found.

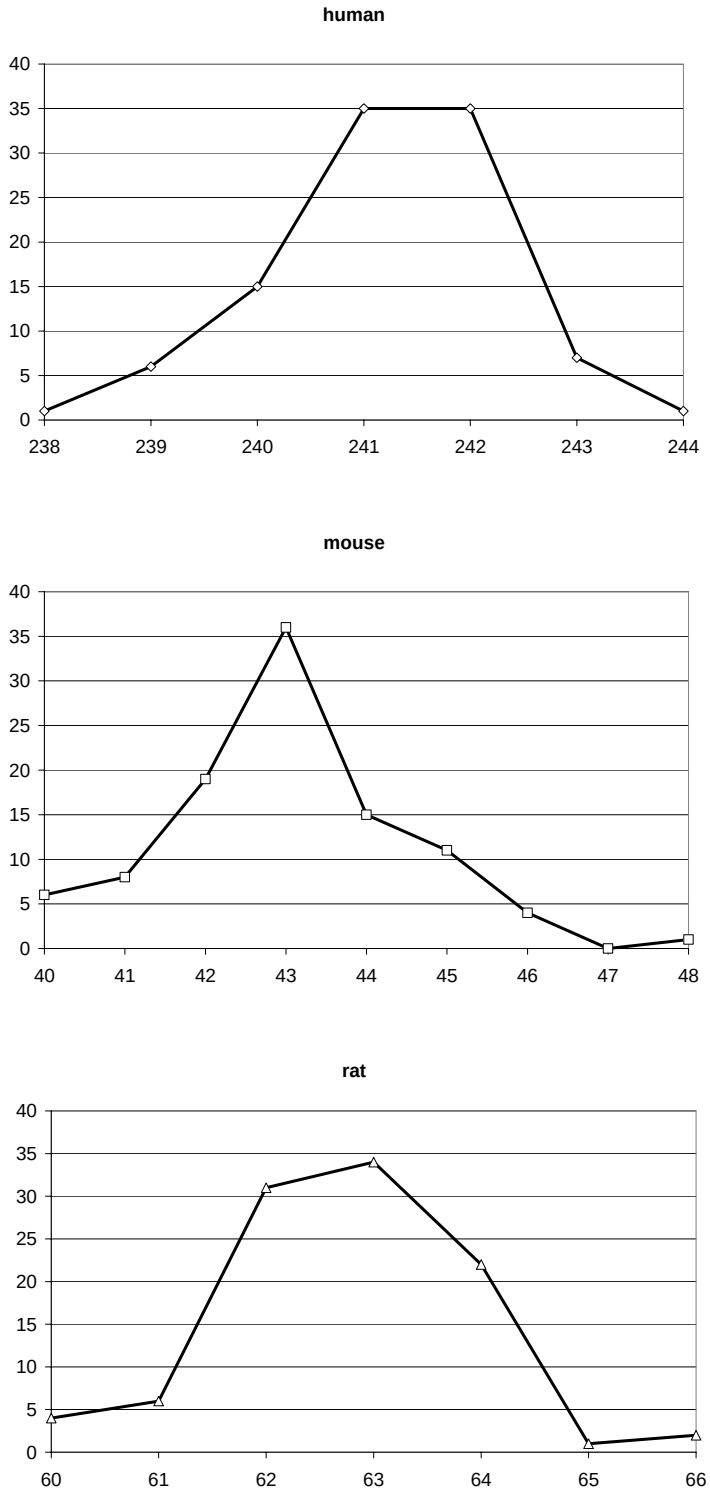


Figure 6. Distribution of the number of events on the human, mouse, and rat lineages in 100 midpoints.