Pacific Symposium on Biocomputing 13:37-48(2008)

SENSITIVITY ANALYSIS FOR REVERSAL DISTANCE AND BREAKPOINT REUSE IN GENOME REARRANGEMENTS

AMIT U SINHA

Department of Computer Science, University of Cincinnati, Cincinnati, OH 45221, USA (sinhaam@ececs.uc.edu)

JAROSLAW MELLER

Department of Environmental Health, University of Cincinnati College of Medicine, Cincinnati, OH 45267, USA; Department of Informatics, Nicholas Copernicus University, 87-100 Torun, Poland (jmeller@cchmc.org)

Identifying syntenic regions and quantifying evolutionary relatedness between genomes by interrogating genome rearrangement events is one of the central goals of comparative genomics. However, identification of synteny blocks and the resulting assessment of genome rearrangements are dependent on the choice of conserved markers, the definition of conserved segments, and the choice of various parameters that are used to construct such segments for two genomes. In this work, we performed an extended sensitivity analysis of synteny block generation using alternative sets of markers in multiple genomes. A simple approach to synteny block aggregation is used, which depends on two principle parameters: the maximum gap (*max_gap*) between adjacent blocks to be merged, and the minimum length (*min_len*) of synteny blocks. In particular, the dependence on the choice of conserved markers and *max_gap*/*min_len* aggregation parameters is assessed for two important quantities that can be used to characterize evolutionary relationships between genomes, namely the reversal distance and breakpoint reuse. We observe that the number of synteny blocks depends on both parameters, while the reversal distance depends mostly on *min_len*. On the other hand, we observe that relative reversal distances between mammalian genomes, which are defined as ratios of distances between different pairs of genomes, are nearly constant for both parameters. Similarly, the breakpoint reuse rate was found to be almost constant for different data sets and a wide range of parameters. Breakpoint reuse is also strongly correlated with evolutionary distances, increasing for pairs of more divergent genomes. Finally, we demonstrate that the role of parameters may be further reduced by using a multi-way analysis that involves markers conserved in multiple genomes, which opens a way to guide the choice of a correct parameterization.

Supplementary Materials (SM) at http://cinteny.cchmc.org/doc/sensitivity.php

1. Introduction

An increasing number of newly sequenced genomes greatly enhance our ability to construct evolutionary models from their comparative analysis. One problem of central importance is the identification of blocks of genes (or other discrete

markers) with evolutionary conserved order. These synteny blocks help in tracing back the evolution of genomes in terms of rearrangement events, such as inversion, translocation, fusion, fission, etc. Consequently, genome evolution and phylogenetic (phylogenomic) trees may be reconstructed from the analysis of synteny [1], [2], [3].

Nadeau and Taylor [4] argued that translocation and inversion (reversal) are the main evolutionary events that affect gene (and other markers) order. They concluded that the effect of transposition is not very significant. In fact, for the sake of computational efficiency, most of the algorithms for finding the evolutionary distance mimic translocation, fission and fusion in terms of inversions, while neglecting the effect of transpositions [5]. In particular, once two genomes are represented in terms of blocks of markers with conserved order, each genome may be transformed into a signed permutation (sign representing the strand of genes/markers). As a result, one genome may be transformed into the other by applying reversal operations, providing a model of genome rearrangements. Consequently, analyses of genome rearrangements within this model typically involve calculating the reversal distance between two genomes, which is defined as the minimum number of reversals required to sort one (signed) permutation to the other [5]. Thanks to recent algorithmic advances, the reversal distance can be computed in linear time [6].

Another quantity that we consider here is the breakpoint reuse rate (BRR), which is defined as $2d/b$, where d is the reversal distance and b is the number of breakpoints, as estimated from the observed synteny blocks. BRR can be interpreted as a simple measure of the extent to which breakpoints are used on average during rearrangement events [7]. However, this interpretation is contested by some groups [8], partly because of the divergence between alternative estimates of the numerical value of BRR, as obtained using different parameterizations of the problem, and partly because it largely disregards the mechanistic nature of rearrangement events that tend to occur within repetitive DNA fragments of certain (potentially large) length [9]. These debates clearly underscore the need for further assessment of current models of genome evolution, and methods for synteny block identification, in particular.

A set of discrete markers that represents the genome of interest in a simple model considered here, consists either of orthologous genes or conserved sequence tags (anchors). Obviously, the choice of a set of markers affects the results and attempts have been made to assess the impact of such choices [7], [10]. Another problem in identifying synteny blocks is that large potential blocks may be interrupted by local disruptions in the order of markers. However, there is no precise definition of such local disruptions. Consequently, many different methods have been devised to filter out these micro-rearrangements,

using heuristics or statistical models to assess the significance of associations (co-localization) between markers [7], [11], [12], [13], [14], [15], [16], [17].

As discussed in section 2.2, many of these algorithms for constructing synteny blocks can be cast using a general framework, in which there are two principal parameters. The first parameter defines blocks to be removed from consideration if their length (either in terms of the minimum number of markers, or in terms of their physical length) is too short. The second parameter defines how adjacent blocks will be merged (effectively disregarding the markers in between that locally disrupt the order), depending on the distance (gap) between these blocks. In what follows, these parameters are referred to as the minimum length (*min_len*) of individual blocks and the maximum gap (*max_gap*) between adjacent blocks to be merged, respectively.

Since the identification of synteny blocks is a crucial step in measuring reversal distance, breakpoint reuse and other related quantities, it is important to systematically assess its sensitivity with respect to the choice of the set of markers (including the use of markers conserved in multiple genomes), *min_len* and *max_gap* parameters, and other arbitrary choices. In fact, the impact of these parameters on the analysis of evolutionary relatedness within this model has recently been highlighted in attempts to estimate breakpoint reuse rate between human and mouse genomes, leading to debates about random vs. fragile breakage model of genome evolution [11], [10].

Here, we used an efficient computational framework [18] for a comprehensive analysis of the sensitivity of the reversal distance and breakpoint reuse in multiple genomes, using both homolog and sequence tag data sets. In particular, we performed a systematic assessment of the role of critical parameters in the model. Based on our result, we suggest that using a subset of genes common to more than two related species may provide more stable results and yield improved estimates of the evolutionary relatedness. Furthermore, we find that relative measures of divergence between two pairs of genomes are less dependent on the choice of arbitrary parameters. This observation provides an additional support for the construction of robust phylogenetic (phylogenomic) trees and other analyses relying on such relative (rather than absolute) distance measures.

2. Methods

The results presented in this contribution were generated using the Cinteny server for the analysis of synteny and genome rearrangements, which is available at http://cinteny.cchmc.org/ [18]. The server allows one to use alternative data sets, including both ortholog and sequence tags (anchors)-based sets of markers in multiple genomes. It also allows the user to set parameters

that affect the synteny blocks identification, as well as the computation of reversal distances and breakpoint reuse rates, enabling systematic analysis of sensitivity of the results with respect to these arbitrary choices.

2.1. *Data Sets*

While sequence tags in general provide greater coverage of the genome, the conservation of non-functional regions may not be of equal importance as gene conservation, or could simply result from spurious sequence matches, introducing noise in the model. On the other hand, the identification of orthologs is often marred by limited sensitivity of sequence searches and other annotation problems. Therefore, we used both orthologs and conserved sequence tags for a more comprehensive analysis. The orthologs from NCBI HomoloGene [19] and RoundUp Orthology Database [20] and a data set of conserved sequence tags from an earlier study by Bourque et al. [3], which will be referred to as GRIMM, were used. HomoloGene contains orthologs for human, mouse, rat, dog and chimp genomes, whereas RoundUp also contains rhesus macaque and cow. GRIMM data set has conserved markers in human, mouse and rat genomes.

2.2. *Forming Synteny Blocks*

Synteny blocks are identified as segments of the genomes in which the order of homologous markers is conserved. Typically, local rearrangement events that concern only few markers within a synteny block, and are referred to as micro-rearrangement, are being ignored. The rationale is that smaller conserved blocks do not represent a significant evolutionary signature, and might add noise to the model. This process takes the form of the aggregation of initial (entirely ordered) blocks to create larger synteny blocks, and effectively filter out these micro-rearrangements. While such an aggregation may be parameterized differently, two parameters are typically used in this context [10]:

- *max_gap*: maximum gap between blocks that are allowed to be merged;
- *min_len*: minimum length of a synteny block.

Specifically, if the gap between two adjacent synteny blocks is less than *max_gap* then they may be merged together to form a larger block. The relative order (orientation) of the two blocks has to be accounted for. This process of aggregation is continued until no more blocks may be merged. Subsequently, the blocks of length less than *min_len* are rejected.

Many algorithms for forming synteny blocks follow this paradigm, and we follow in their footsteps. For example, the GRIMM-Synteny [7] and MAUVE [15] algorithms define the parameter *min_len* as ‘minimum cluster size C’ and *w(cb)*, respectively. An alternative is to set a lower limit on the size of synteny

blocks in terms of the number of markers within the block. This corresponds to the parameter Δ in ST-Synteny algorithm [11] and *h* in [17]. Here, *min_len* is varied to test its effect on the results. In addition, unless stated otherwise, we reject synteny blocks with less than 3 markers.

There is more ambiguity in the notion of *max_gap* [17], which prescribes how blocks should be aggregated. We define it as the threshold on the maximum gap between the two synteny blocks in *each* species that are allowed to be merged. This corresponds to the parameter 'maximum gap size G' in the GRIMM-Synteny algorithm [7] and d_c, in FISH algorithm [12], which is defined as the *sum* of gap between adjacent blocks in the two species. The parameter *MaxDist* in [14] is similar to *max_gap* as well. In some cases, *max_gap* is defined in terms of the numbers of markers, i.e., by putting a threshold on the number of out-of-order markers while merging blocks [13], [17], [16]. Some methods avoid this gap constraint by coalescing blocks after removing smaller blocks [15]. However, this behavior may still be captured by some parameterization of *max_gap*. In general, while direct comparison between different definitions may be difficult, *max_gap* has relatively small impact on measures of genome rearrangement, as we show in the results section.

2.3. *Measuring Reversal Distance*

Once the synteny blocks are identified, the relative order of blocks in two multi-chromosomal genomes is represented as numeric signed permutation. The Hannenhalli-Pevzner [5] algorithm calculates the reversal distance in linear time when used with modifications proposed by [6] and [21], which we implemented in the Cinteny server, to enable comprehensive assessment for a large range of parameters considered here. It should be noted that we do not address block or genome duplications, and we use a heuristic choice of unique markers for paralogs (see Supplementary Materials).

2.4. *Using Multiple Genomes*

Working with a set of markers conserved across multiple species instead of those conserved in individual pairs of genomes may lead to more stable results. For example, at present HomoloGene includes 16,330 orthologs for human and mouse. When using a '5-way' approach, 10,574 genes having orthologs in human, mouse, rat, dog and chimp are identified. Pairwise synteny between human and mouse can now be identified using only these 10,574 genes. The advantage of using this approach is that aggregation of synteny blocks occurs naturally, as only highly conserved segments are used. The same logic may be extended for any multi-way approach, with the hope that a subset of markers

conserved and/or better annotated in multiple species may help filtering out micro-rearrangements and minimizing the effects of errors in homology prediction. A similar method was demonstrated for chromosome level comparison to yield more meaningful relationships between canine and other mammalian genomes [22].

3. Results

We used two independent ortholog data sets and a data set of conserved sequence tags, as described in the Methods section, in order to measure the variation in number of synteny blocks, reversal distance, breakpoint re-use rate, etc., by changing the parameters *max_gap* and *min_len*.

3.1. *Ortholog v/s Conserved Sequence Tags*

3.1.1. *Number of Synteny Blocks*

Figure 1 shows the variation in the Number of Synteny Blocks (NSB) due to parameters *max_gap* and *min_len* for human-mouse pair using HomoloGene data set. The parameters *max_gap* (y-axis) and *min_len* (x-axis) were increased from 0 to 1 Mb in steps of 20 Kb and NSB is plotted (z-axis). We observe that NSB decreases on increasing *max_gap* and *min_len*. When the latter is increased, more synteny blocks (of smaller size) are rejected leading to a decrease in NSB. As *max_gap* is increased, adjacent synteny blocks are aggregated and their total number decreases too, although to a smaller degree. In general, the results obtained with the RoundUp orthologs and GRIMM sequence tags were similar to those obtained with HomoloGene (SM Figure S1). However, when using sequence tags, the number of synteny blocks with small *max_gap* is large. This is because the number of sequence tags is much larger than the number of orthologs and little aggregation takes place when *max_gap* is low. As *max_gap* is increased, more aggregation takes place and there is a steep decline in the total number of synteny blocks, which becomes very close to the value observed for gene-based analysis. Similar pattern in sensitivity of NSB was observed for human-dog, human-rat, rat-mouse and other pairs of genomes (see SM Figure S2).

3.1.2. *Reversal Distance*

Once the synteny blocks are found, the disruption of the order of the blocks is measured as the Reversal Distance (RD). Figure 2 shows the variation of RD due to *min_len* for human-mouse genomes. For each value of *min_len*, the RD is calculated for different values of *max_gap* (between 60 Kb and 1 Mb) and the

variation is displayed as box plots. The low heights of the boxes indicate that the variation in RD due to *max_gap* for a given value of *min_len* is limited. This is because increasing *max_gap* preferentially aggregates blocks which have a similar order in both genomes, so the reversal distance does not change much. On the other hand, there is a steep and uneven decrease in RD as *min_len* is increased but the median values start to flatten at higher values of *min_len*. Some outliers are observed for high values of *min_len* and low values of *max_gap*. Orthologs and sequence tags based data sets give similar results qualitatively. Sequence tag based analysis gives a higher RD for low values of *min_len* because the number of synteny blocks is higher. At higher values of *min_len*, the value of RD for both types of data begins to converge. The results obtained with the RoundUp orthologs were similar (see SM Figure S3).

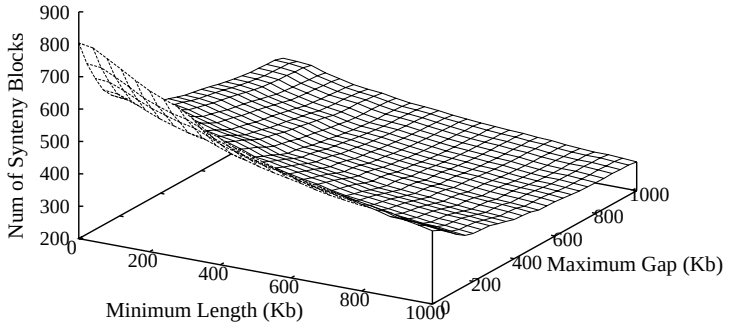

Figure 1: Variation in number of synteny blocks due to *max_gap* and *min_len* in human-mouse genomes for orthologs based analysis

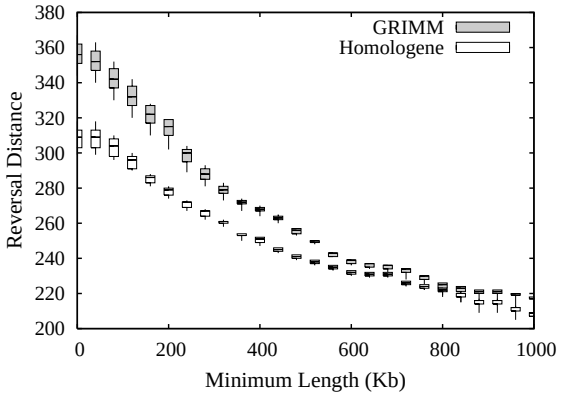

Figure 2: Variation of reversal distance due to *min_len* in human-mouse for ortholog (HomoloGene) and sequence tag (GRIMM) based analysis. The height of the boxes shows the variation in reversal distance due to *max_gap* for a given value of *min_len*.

3.2. *Breakpoint Reuse Rate*

Measurement of Breakpoint Reuse Rate (BRR) and its dependence on parameters has been debated a lot in the last few years [10], [8]. In particular, its numerical value was used as an argument in the dispute over fragile vs. random breakage model of genome evolution. We first assess the effects of the parameters on BRR for human and mouse genomes. The parameters *max_gap* and *min_len* were varied from 0 to 1 Mb in steps of 20 Kb and the BRR was calculated for different data sets. The mean and standard deviation as well as minimum and maximum values of BRR over the range of parameters are shown in Table 1. We observe that unlike RD, BRR (which is a relative quantity) shows very little variation due to the parameters or due to data sets. These results are consistent with previous findings by Peng and colleagues [10] for human-mouse genomes, for which they reported a BRR of 1.61 and 1.67 for ortholog-based and sequence-based analysis, respectively. To extend this analysis, we investigate BRR further in the next section by comparing it with other measures of evolutionary divergence.

Table 1: Breakpoint reuse rate for human-mouse genomes with varying *max_gap* and *min_len* from 0 to 1 Mb for all 3 data sets and parameterizations of the problem (see text for details)

	BRR			
Data set	**Mean**	**SD**	**Min**	**Max**
HomoloGene	1.64	0.02	1.58	1.66
RoundUp	1.63	0.02	1.59	1.67
Grimm	1.62	0.03	1.53	1.66

3.3. *Correlation of Reversal Distance and Breakpoint Reuse Rate*

One expects an increase in the number of genome rearrangement events as species evolve and diverge from their ancestral genomes. Additionally, when the number of rearrangement events is high, the chance of a breakpoint region being reused increases. Indeed, this is found to be the case for many genome pairs. Figure 3 shows BRR and RD of 5 genomes with human genome. The RD and BRR were calculated for both *min_len* and *max_gap* equal to 500 Kb. Pearson correlation coefficient was found to be 0.996 ($p < 0.001$). A correlation of 0.995 and 0.990 was found for *min_len* equal to 300 Kb and 1000 Kb, respectively, showing that the correlation stands for different values of these parameters.

There are, however, some intriguing exceptions from this general trend. For example, the human-dog and mouse-rat genomes have similar BRR (1.40 and 1.43, respectively) even though the RD is very different (150 and 71, respectively). Despite such outliers, it is evident that BRR increases, as the number of rearrangement events increases. Closely related genomes, such as human and chimp, show a BRR of 1.1, while human-mouse has a BRR of 1.64.

Similarly, mouse-rat genomes have a BRR of 1.42, while mouse-dog genomes have a BRR of 1.62. These data suggest that BRR may be used as an alternative measure of evolutionary distance, as it is largely independent of parameters.

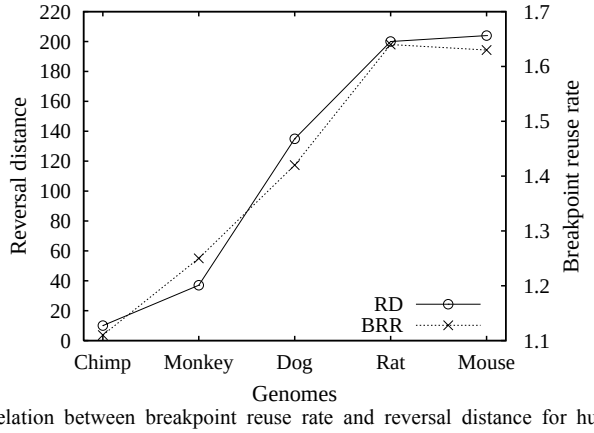

Figure 3: Correlation between breakpoint reuse rate and reversal distance for human and other genomes. The trend is independent of the parameterization of the synteny block identification.

Finally, in the context of the on-going discussion about numerical estimates of BRR and the validation of the proposed fragile breakage model [7], we would like to comment that BRR is an average quantity. In particular, it may be possible that some breakpoints are used more frequently than others, especially if they occur within large repetitive regions in the genome [9]. Since the evolutionary pathway can not be uniquely determined for a given reversal distance using Hannenhalli-Pevzner model, it is not possible to determine the actual number of breakpoints which are, in fact, reused (perhaps more than once) during the transformation of one genome to another. Consequently, the numerical value of BRR may be more informative as a relative (and weakly dependent on parameters) measure of evolutionary distances, rather than supporting (or not) one of the models of rearrangement events.

3.4. *Relative Divergence*

In light of the above conclusions regarding BRR, we investigated another relative measure of evolutionary relatedness. The absolute values of RD (reversal distance) are found to be very sensitive to the choice of *min_len*. Therefore, we define a relative divergence measure, as the ratio of RD of two different pairs of genomes. For this analysis, we measured RD and relative divergence as a function of *min_len*. The results in Table 2 show the absolute value of RD in human-mouse (H-M), rat-mouse (R-M), human-dog (H-D) and human-chimp (H-C) genomes for different choice of parameters. The table also shows the ratio of human-mouse RD with other pairs.

We observe that even though individual RD changes with the parameters, as shown earlier, the ratio of RD between pairs of genome shows negligible variation for *min_len* greater than 200 Kb. The mean relative divergence of human-mouse with respect to rat-mouse, human-dog and human-chimp genomes is almost constant at 3.29 (σ = 0.05), 1.63 (σ = 0.03) and 21.91 (σ = 0.79), respectively. This information (relative divergence) may be more useful than a simple RD as it shows very little variation due to the parameters. This perhaps bode well for attempts to use RD as a measurement of inter-genomic distances in relative terms, e.g., to construct phylogenomic trees. The ratio of NSB was also found to be constant for different choice of parameters between two pair of genomes.

Table 2: Variation of RD in different genome pairs and ratios of RD

min_len (kb)	**Reversal distances**				**Relative divergence between pairs**		
	H-M	**R-M**	**H-D**	**H-C**	**H-M/R-M**	**H-M/H-D**	**H-M/H-C**
1000	206	62	131	9	3.32	1.57	22.89
800	224	66	135	10	3.39	1.66	22.40
600	233	71	143	11	3.28	1.63	21.18
400	248	76	151	12	3.26	1.64	20.67
200	281	92	174	14	3.05	1.61	20.07
100	313	128	186	22	2.45	1.69	14.23
0	346	170	211	34	2.04	1.64	10.18

3.5. *Using Multiple Genomes*

In order to assess the behavior of more highly conserved elements, we compared the variation of reversal distance using 2-way and 5-way approaches. The former was done using the genes common to human and mouse and the latter was done using the genes common to human, mouse, dog, rat, chimp genomes. The number of orthologs for the 2-way and 5-way analysis was 16,330 and 10,574, respectively. Figure 4 shows the variation of RD due to *min_len* for human-mouse genomes. For each value of *min_len*, RD is calculated for different values of *max_gap* and the variation is displayed as a box plot.

Since fewer orthologs are used for a 5-way analysis, RD is smaller in absolute terms than in 2-way analysis. It is also evident from Figure 4 that the variation due to *max_gap* (height of the boxes) is almost negligible in the case of 5-way comparison. Furthermore, the overall variation due to *min_len* is less pronounced in 5-way comparison. This suggests that multi-way analysis reduces the role of the parameters, albeit to different degrees. We also performed an extended analysis of BRR using 5-way approach for five mammalian genomes. The results are similar to those obtained using 2-way approach, indicating again relatively low sensitivity of BRR with respect to the parameterization of the problem.

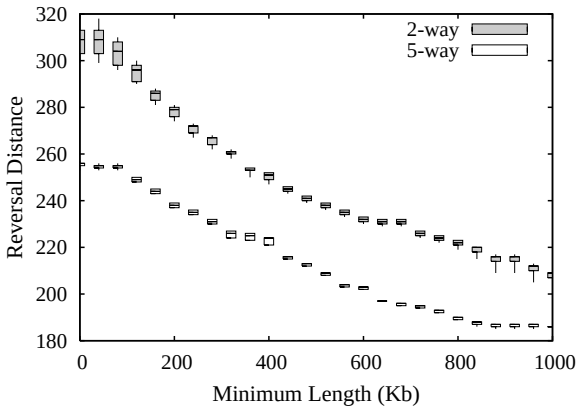

Figure 4: Variation in RD due to *min_len* in human and mouse genomes for 2-way and 5-way analysis. The height of the boxes shows the variation in reversal distance due to *max_gap* for a given value of *min_len*. The observed variation is smaller when using multiple genome approach.

4. Conclusions

Genome rearrangement analysis is often marred by the lack of a clear strategy for selecting critical parameters, choosing appropriate data sets, etc. We performed a systematic analysis of the sensitivity of genome rearrangement measures to the choice of critical parameters for several mammalian genomes. Both ortholog-based and sequence tag-based approaches are compared. Two specific parameters, i.e., the maximum allowable gap between adjacent blocks for aggregation and the minimum length of synteny blocks, are varied systematically to assess their effect. We found that the number of synteny blocks depends on both parameters, while the reversal distance depends mostly on the latter. Therefore, one needs to exercise caution while using (absolute values of) reversal distances as a measure of evolutionary relatedness. The breakpoint reuse rate was found, on the other hand, to have a negligible change due to variation in these two parameters. At the same time, it showed a strong correlation with reversal distances, indicating that high breakpoint reuse rates may simply reflect the expected higher number of inversions with increasing evolutionary divergence. This, however, opens a way to use BRR as an alternative measure of evolutionary distance, which may be more informative for inferring evolutionary relatedness, building phylogenetic trees and other applications. Another relative measure with similar properties that we consider is the relative divergence, which is defined as ratios of reversal distances between different pairs of genomes. In this context, the distance for a pair of well defined and annotated genomes, such as human and mouse, may be used to normalize all other pair-wise distances with the same parameterization. Using multiple-way comparisons decreases the dependence on parameters, when

compared with two-way analysis, suggesting rational strategies to choose parameters for the identification of synteny blocks.

Acknowledgements

We would like to thank the reviewers for their insightful comments and suggestions. This work has been partially supported by NIH grant R01 AR050688.

References

1. Sankoff D, Blanchette M, In *Proc. of COCOON*, 251-63, (1997).
2. Moret BME, Wyman S, Bader DA, et al., In *Proc. of Pac Symp on Biocomputing*, 583-94, (2001).
3. Bourque G, Pevzner PA, Tesler G, *Genome Res*, 14: 507-16, (2004).
4. Nadeau JH, Taylor BA, *Proc Natl Acad Sci USA*, 81: 814-18, (1984).
5. Hannenhalli S, Pevzner PA, In *Proc. of IEEE Symp on Found. of Comp Sci*, 581-92, (1995).
6. Bader DA, Moret BME, Yan M, *J. of Comp. Bio*, 8: 483-91, (2001).
7. Pevzner PA, Tesler G, In *Proc. of RECOMB*, 247-56, (2003).
8. Sankoff D, *PLoS Comput Biol*, 2: e35, (2006).
9. Ruiz-Herrera A, Castresana J, Robinson TJ, *Genome Biol*, 7:R115, (2006).
10. Peng Q, Pevzner PA, Tesler G, *PLoS Comput Biol*, 2: e14, (2006).
11. Sankoff D, Trinh P, In *Proc. of RECOMB*, 30-35, (2004).
12. Calabrese PP, Chakravarty S, Vision TJ, *Bioinformatics*, 19 Suppl. 1: i74-i80, (2003).
13. Hampson S, McLysaght A, Gaut B, et al., *Genome Res*, 13: 999-1010, (2003).
14. Haas BJ, Delcher AL, Wortman JR, et al., *Bioinformatics*, 20: 3643-46, (2004).
15. Darling ACE, Mau B, Blattner FR, et al., *Genome Res*, 14:1394-1403, (2004).
16. Mouse Genome Sequencing Consortium, *Nature*, 420: 520-62, (2002).
17. Hoberman R, Sankoff D, Durand D, In *Proc. of RECOMB Workshop on Comparative Genomics*, 55–71, (2005).
18. Sinha AU, Meller J, *BMC Bioinformatics*, 8: 82, (2007).
19. Wheeler DL, Barrett T, Benson DA, et al., *Nuc Acids Res*, 35:D5-12, (2007).
20. Deluca TF, Wu IH, Pu J, et al., *Bioinformatics*, 22: 2044-46, (2006).
21. Tesler G, *J. of Computer and System Sciences*, 65: 587-609, (2002).
22. Andelfinger G, Hitte C, Guyon R, et al., *Genomics*, 83: 1053-62, (2004).