Chromosomal Breakpoint Re-use in the Inference of Genome Sequence Rearrangement

David Sankoff
Department of Mathematics and Statistics
University of Ottawa
Ottawa, Canada, K1N 6N5
sankoff@uottawa.ca

Phil Trinh
Hillcrest High School
Ottawa, Canada, K1G 2L7
Phil_Trinh@Canada.com

ABSTRACT

In order to apply gene-order rearrangement algorithms to the comparison of genome sequences, Pevzner and Tesler [9] bypass gene finding and ortholog identification, and use the order of homologous blocks of unannotated sequence as input. The method excludes blocks shorter than a threshold length and ignores small block-internal rearrangements. Here we investigate possible biases introduced by eliminating and amalgamating short blocks, focusing on the notion of "breakpoint re-use" introduced by these authors. Analytic and simulation methods show that re-use is very sensitive to threshold size and to parameters of the rearrangement process. As is pertinent to the comparison of mammalian genomes, large thresholds in the context of high rates of small rearrangements risk randomizing the comparison completely. We suggest a number of mathematical, algorithmic and statistical lines for further developing the Pevzner-Tesler approach.

Categories and Subject Descriptors

F.2.2 [**Analysis of algorithms and problem complexity**]: Nonnumerical Algorithms and Problems

General Terms

algorithms

Keywords

comparative genomics, evolution, rearrangements, inversion, Hannenhalli-Pevzner algorithm, breakpoints, synteny blocks

1. INTRODUCTION.

Until recently algorithms for studying the evolution of gene order could only be applied to small genomes (mitochondria, chloroplasts, prokaryotes), the difficulty with mammalian and other larger eukaryotic nuclear genomes lying not so much in their much greater length but rather in the absence of comprehensive lists of genes and their orthologs. Pevzner and Tesler have suggested a way to bypass gene finding and ortholog identification by using the order of syntenic blocks constructed solely from sequence data as input to a genome rearrangement algorithm. The method focuses on major evolutionary events by glossing over small block-internal rearrangements, and neglecting intervening blocks smaller than a threshold length. This use of large "sanitized" blocks, and the neglect of short blocks may, however, blur important parts of the historical derivation of the genomes. We model the effects of eliminating and amalgamating short blocks, concentrating on the summary statistic of"breakpoint re-use" introduced by Pevzner and Tesler. They did not conceive of this as an evolutionary distance, but in the context of their protocol it effectively measures to what extent genomes have diverged in becoming random permutations of blocks with respect to each other. We use analytic and simulation methods to investigate breakpoint re-use as a function of threshold size and of rearrangement parameters. Throughout, we discuss the implication of our findings for the comparison of mammalian genomes and suggest a number of mathematical, algorithmic and statistical lines for further developing the Pevzner-Tesler approach.

RECOMB'04, March 27–31, 2004, San Diego, California, USA.

2. FROM GENOME SEQUENCE TO GENOME REARRANGEMENT

Algorithmic inference of genome rearrangement, as reviewed in [11], has been predicated on the representation of the genome as a signed permutation of $(12\cdots n)$, in the case of unichromosomal organisms, or as a fragmented signed permutation in the case of multichromosomal organisms. Each of the n terms in this representation of a genome corresponds to a unique term in the other, possibly in a different position or with a different sign. Each term represents a gene or other marker that has been mapped by genetic or molecular biological techniques or abstracted from the underlying sequence data. In large measure, these algorithms have been developed in the context of organellar or other small genomes, where gene finding and ortholog identification have not been major obstacles. Recent improvements in efficiency [1, 2, 12] enable this approach to handle many thousands of genes in reasonable computing time. Faced with large nuclear genome sequences, particularly from the higher eukaryotes, however, uncertainties in global alignments, lack of complete consensus inventories of

genes and the difficulties of distinguishing among paralogs widely distributed across the genome, constitute apparently insurmountable impediments to the direct application of the algorithms.

2.1 The Pevzner-Tesler protocol

In comparing drafts of the human and mouse genomes, Pevzner and Tesler [13, 8, 9, 10] adopted an ingenious strategem to leap-frog the global alignment, gene finding and ortholog identification steps. They analyzed almost 600,000 relatively short (average length 340 bp [8]) anchors of highly aligned sequence fragments as a starting point for building blocks of conserved synteny, and then amalgamated neighboring sub-blocks using a variety of criteria to avoid disruptions due to "microrearrangements" less than 1Mb. This procedure eventually succeeds in inferring a set of $b' = 281$ blocks larger than 1 Mb, which is comparable to the result in [7] and indeed to the most recent results (somewhat more than 200 in the current NCBI Human-Mouse Homology Map (http://www.ncbi.nlm.nih.gov/Homology/ComMapDoc.html) of the comparative mapping and other "pre-genome sequence" approaches. Pevzner and Tesler go a step further and use the order of these b' large blocks on the $c = 23$ chromosomes as input to an improved adaptation [12] of the gene order rearrangement algorithms originally devised by Hannenhalli and Pevzner [4, 3, 5], in order to reconstruct aspects of the actual sequence of d inversions and translocations that account for the divergent structures of the two genomes. Some familiarity with the basics of the Hannenhalli-Pevzner theory will be useful to the reader in the ensuing sections.

2.2 The re-use statistic r.

One of the key results reported by Pevzner and Tesler pertains to the "re-use" of the breakpoints between the b' syntenic blocks on the c chromosomes used as input to their rearrangement algorithms. Basic to the combinatorial optimization approach to inferring genome rearrangements are the bounds $\frac{b}{2} \leq d \leq b$, where $b = b' - c$. (This type of bound was first found in 1982 [14].) We define breakpoint re-use as $r = \frac{2d}{b}$. Then

$$1 \leq r \leq 2.$$

The lower value $r = 1$ is characteristic of an evolutionary trajectory where each inversion or translocation breaks the genome at two sites specific to that particular rearrangement; no other inversion or translocation breaks the genome at either of these sites. High values of r, near $r = 2$, are characteristic of evolutionary histories where each rearrangement after the first one breaks the genome at one new site and at one previously broken site. Pevzner and Tesler [10] found that $r = 1.9$ in their comparison of the human mouse genome, and argued that this was evidence that evolutionary breakpoints are concentrated in fragile regions covering a relatively small proportion of the genome.

Now it is also an observed property of random permutations of length n, where it can be shown that $b \approx n$, that the number of inversions needed to sort them (d) is very close to n, and thus breakpoint re-use is close to 2. Without disputing the substantive claim about fragile regions, for which there may be independent evidence [6], we may ask what breakpoint re-use in empirical genome comparison really measures: a bonafide tendency for repeated use of breakpoints or simply the degree of randomness of one genome with respect to the other at the level of synteny blocks. We will show here how this randomness may be an artifact of the Pevzner-Tesler protocol for constructing the synteny blocks.

3. SIMULATING INVERSION WITH A BLOCK-SIZE THRESHOLD

To see whether a high inferred rate of breakpoint re-use necessarily reflects a high rate when the genome was derived, we will generate a genome with NO breakpoint re-use ($r = 1$), then mimic the Pevzner-Tesler imposition of a block-size threshold and calculate r for the remaining configuration of blocks.

We generate a permutation of length $n = 1000$ or $n = 100$ by applying d "two-breakpoint" inversions to the identity permutation $(12 \cdots n)$. A two-breakpoint inversion is one that disrupts two hitherto intact adjacencies in the starting (i.e. identity) permutation. At each step, the two breakpoints are chosen at random among the remaining original adjacencies. This represents the extreme hypothesis of no breakpoint re-use at all during evolution, which is not unreasonable given the 3 x 10^9 distinct dinucleotide sites available in a mammalian genome.

Of course, our "blocks" are just elements in the permutation and have no associated size, and indeed the Hannenhalli-Pevzner procedures do not involve any concept of block size. Thus, to imitate the effect of imposing a block-size threshold we simply delete a fixed proportion of the terms at random, the same terms from both the starting and derived genomes, relabel the remaining terms according to their order in the starting (identity) genome, and apply the Hannenhalli-Pevzner algorithm.

It can be shown that before any deletions, the Hannenhalli-Pevzner algorithm will recover exactly d inversions. At each step it will find a configuration of form $\cdots gh \mid -(i-1), \cdots, -(h+1) \mid ij \cdots$ and will "undo" the inversion between h and i, removing two breakpoints. There being $b = 2d$ breakpoints, breakpoint re-use is 1.0.

What happens as terms are deleted? Suppose $j \neq i+1$ in the above example, and i is deleted. Then the two-breakpoint inversion from $-(i-1)$ to $-(h+1)$ is no longer available to undo. An inversion that erases the breakpoint between h and $-(i-1)$ will not eliminate a second breakpoint. So while the distance d drops by 1, the number of breakpoints b also only drops by 1, and r increases.

The probability that one, two, or more two-breakpoint inversions are "spoiled' in this way depends on the number of terms deleted.

Figure 1 shows how r increases with the proportion of terms deleted, for different values of d, for $n = 100$ and $n = 1000$.

We note

- r increases more rapidly for more highly rearranged genomes.
- the initial rate of increase of r depends only on d/n
- the increase in r levels off well below $r = 2$ and then descends sharply. The maximum level attained increases with n.

The first of these is readily explained. In more rearranged permutations, the deletion of term i is more likely to cause the configuration change described above, i.e. $\cdots gh \mid -(i-$

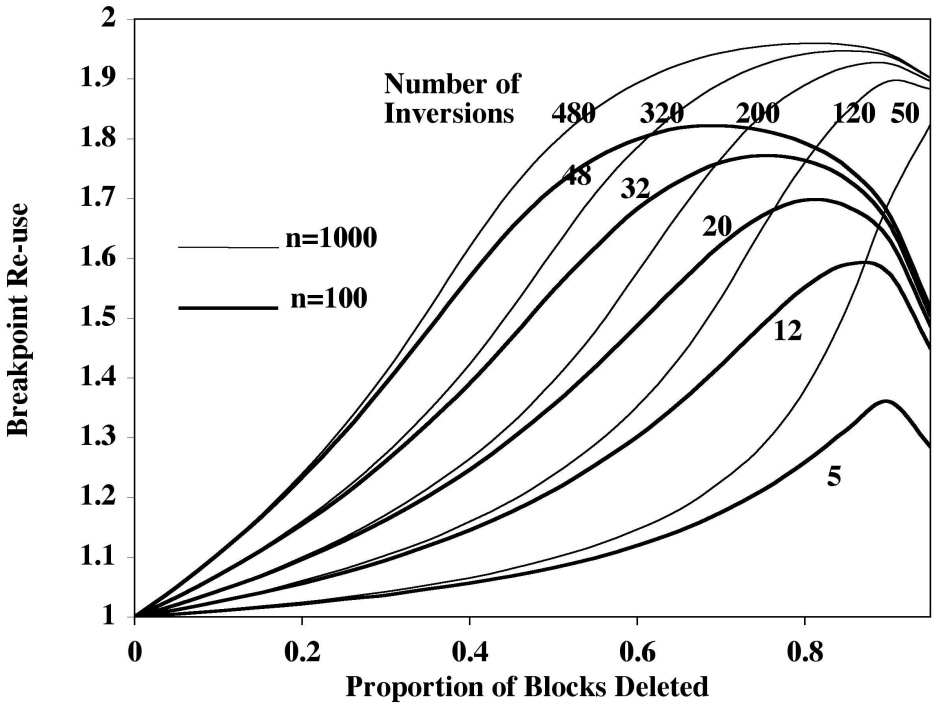

Figure 1: Effect of deleting random terms on breakpoint re-use, as a function of proportion of terms deleted, for various levels of rearrangement of the genome.

$1), \cdots, -(h+1) \mid ij \cdots$, simply because it is more likely that $j \neq i+1$. The third observation is also easily understood. For large n, the re-use rate r approaches 2 for random permutations. As n decreases, however, expected re-use drops as indicated in Table 1.

n	r
5	1.53
25	1.83
50	1.90
100	1.94
250	1.97

Table 1: Expected re-use as a function of n. Estimated from samples of size 500.

As more and more terms are dropped from a permutation, it loses its "structure", i.e., the pairs of breakpoints involved in the original inversions are wholly or partially deleted, and the remaining permutation becomes essentially random. We may consider that after a curve in Figure 1 attains its maximum, it is entering into the 'noisy" region where the historical signal becomes thoroughly hidden.

4. A MODEL FOR BREAKPOINT RE-USE

In this section, we explain the second observation above about the pertinence of d/n for the initial shape of the curves. Suppose a genome G has b breakpoints with respect to $12 \cdots n$ and the inversion distance is $d = d_2 + d_1$, involving no hurdles, where d_1 and d_2 represent the number of one-breakpoint inversions and two-breakpoint inversions required to sort G optimally. Then $2d_2 + d_1 = b$.

Suppose now that we delete one gene i at random and relabel genes $j = i+1, \cdots, n$ as $j = i, \cdots, n-1$, respectively. The number of breakpoints changes, and quantities b, d_1, d_2 and d can change only if the original gene i was flanked by two breakpoints. The probability of this event is $b(b-1)/n(n-1)$. The left and right-hand breakpoints may be involved in one or two inversions among the d_1 one-breakpoint inversions, with probabilities

$$p_{11} = \frac{d_1}{b(b-1)}$$

$$p_{12} = \frac{d_1(d_1-2)}{b(b-1)}$$

(Cases 11 and 12 in Table 2, respectively), respectively, or one or two inversions among the d_2 two-breakpoint inversions, with probabilities

$$p_{21} = \frac{1}{d_2-1}\frac{2d_2(2d_2-1)}{b(b-1)}$$

$$p_{22} = \frac{d_2-2}{d_2-1}\frac{2d_2(2d_2-1)}{b(b-1)},$$

(Cases 21 and 21), respectively, or one one-breakpoint inversion and one two-breakpoint inversion, with probability

$$p_3 = \frac{4d_1d_2}{b(b-1)},$$

(Case 3).

Case	configuration	probability	effect on d_1	effect on d_2
11	$-(i+1) \mid i \mid j$	$\frac{d_1}{b(b-1)}$	-1	0
12	$g \mid -(i-1), \cdots, h \mid i \mid j, \cdots, -(i+1) \mid k$	$\frac{d_1(d_1-2)}{4b(b-1)}$	-1	0
	$g \mid -(i-1), \cdots, h \mid i \mid -(k-1), \cdots, j \mid k$	$\frac{d_1(d_1-2)}{2b(b-1)}$	-1	0
	$g \mid h, \cdots, -(g+1) \mid i \mid -(k-1), \cdots, j \mid k$	$\frac{d_1(d_1-2)}{4b(b-1)}$	-1	0
21	$-(i+1) \mid i \mid -(i-1)$	$\frac{1}{d_2-1}\frac{2d_2(2d_2-1)}{b(b-1)}$	0	-1
22	$g \mid -(i-1), \cdots, -(g+1) \mid i \mid -(k-1), \cdots, -(i+1) \mid k$	$\frac{d_2-2}{d_2-1}\frac{2d_2(2d_2-1)}{b(b-1)}$	+1	-1
3	$g \mid -(i-1), \cdots, -(g+1) \mid i \mid -(k-1), \cdots, j \mid k$	$\frac{d_1 d_2}{b(b-1)}$	+1	-1
	$g \mid -(i-1), \cdots, -(g+1) \mid i \mid j, \cdots, -(i+1) \mid k$	$\frac{d_1 d_2}{b(b-1)}$	+1	-1

Table 2: Probabilities and usual effects of discarding gene i in various configurations, given it is flanked by two breakpoints. Probabilities include those of inverted or nested versions (not listed) of configurations shown. Special cases of configurations with order $O(1/n)$ probabilities not distinguished, e.g. $g \mid -(i-1), \cdots, h \mid i \mid h+1, \cdots, -(i+1) \mid k$.

Were these inversions available in G directly (in fact, some are set up by other inversions later, during the sorting of G), removing i would not only diminish b by one, in effect fusing two breakpoints (except in Case 21 where it eliminates two breakpoints), but it would also generally decrease the number of one-breakpoint inversions by one in Cases 11 and 12, the number of two-breakpoint inversions by one in Case 21 and effectively converts one two-breakpoint inversion to a one-breakpoint inversion in Cases 22 and 3, the two one-breakpoint inversions are replaced by one two-breakpoint inversion. (There are a few special cases of these events, occurring in $O(1/n)$ proportions, that differ in their effects on b, d_1 and d_2, but that we may ignore for present purposes.) These observations motivate the deterministic model:

$$d_2(t+1) = d_2(t) + \frac{b(t)(b(t)-1)}{(n-t)(n-t-1)}(-p_{21}(t) - p_{22}(t) - p_3(t))$$

$$= d_2(t) - \frac{2d_2(2d_2-1) + 4d_1 d_2}{(n-t)(n-t-1)},$$

$$d_1(t+1) = d_1(t) + \frac{b(t)(b(t)-1)}{(n-t)(n-t-1)}(p_{22}(t) + p_3(t) - \max[0, p_1(t)])$$

$$= d_1(t) + \frac{\frac{d_2-2}{d_2-1}2d_2(2d_2-1) + 4d_1 d_2 - \max[0, d_1(d_1-1)]}{(n-t)(n-t-1)}.$$

$$b(t+1) = 2d_2(t+1) + d_1(t+1),$$

where t ranges from 0 to n and with initial conditions $b(0) = 2d_2(0) = 2d(0)$ and $d_1(0) = 0$. (N.B. All the d terms on the RHS of the recurrence should be understood as indexed by t.)

Figure 2 shows how the recurrence models closely the average evolution of r as the number of terms randomly deleted increases, particularly at the outset, before there are large numbers of one-breakpoint inversions in the Hannenhalli-Pevzner reconstruction. As d_1 increases, the model renders less well the changing structure of optimal reconstructions. Finally, the loss of historical signal in the noisy zone for the reconstructions is not built into the model, which attains $r = 2$ as the last terms of the permutation are deleted.

Let $\theta = t/n$ represent the proportion of terms deleted. Formally, since $r = 2d/b$, and d is constant in a neighbourhood of of $t = 0$, while $db/dt \approx -(b/n)^2$, we can write that $dr/d\theta|_{\theta=0} = 2d/n$. This explains the coincidence between the curves for $n = 100$ and $n = 1000$ in Figure 1.

5. THE EFFECT OF MICROREARRANGEMENTS

In the previous sections we have investigated the effect of threshold size on r, albeit indirectly by varying the rate of random deletion of blocks. What is the effect of "repairing" small rearrangements that disrupt longer syntenic blocks? There are diverse considerations for how to detect and suppress the disruptions caused by such microrearrangements:

- how far does the disruption separate terms in one genome that are adjacent in the other? A threshold is necessary to decide when a this separation can no longer be categorized as a minor change.
- how large are the chromosomal segments created or eliminated by the disruption? Again, a threshold is necessary to decide when a segment is to large to be suppressed.
- does the disruption implicate more than one chromosome in either or both of the genomes being compared? We may wish to consider translocations between chromosomes as more significant evolutionary events than small inversions.
- how do we decide between two or more plausible reconstructions that infer conflicting syntenic blocks? Global evaluation criteria are necessary involving the lengths and quality of sets of competing alignments.
- do we wish to allow syntenic blocks to overlap? This may have the advantage of allowing the visualization of several competing solutions at once.

We do not survey the many protocols that reply to these questions in a variety of ways. Each of the recent human-mouse comparisons [8, 6] uses different procedures. In the context of our simulations on permutations rather than sequence rearrangements, we imitate the process of creating syntenic blocks as follows.

1. generate 150 random two-breakpoint inversions of a $n = 5000$ long genome. These set up 300 segments, in

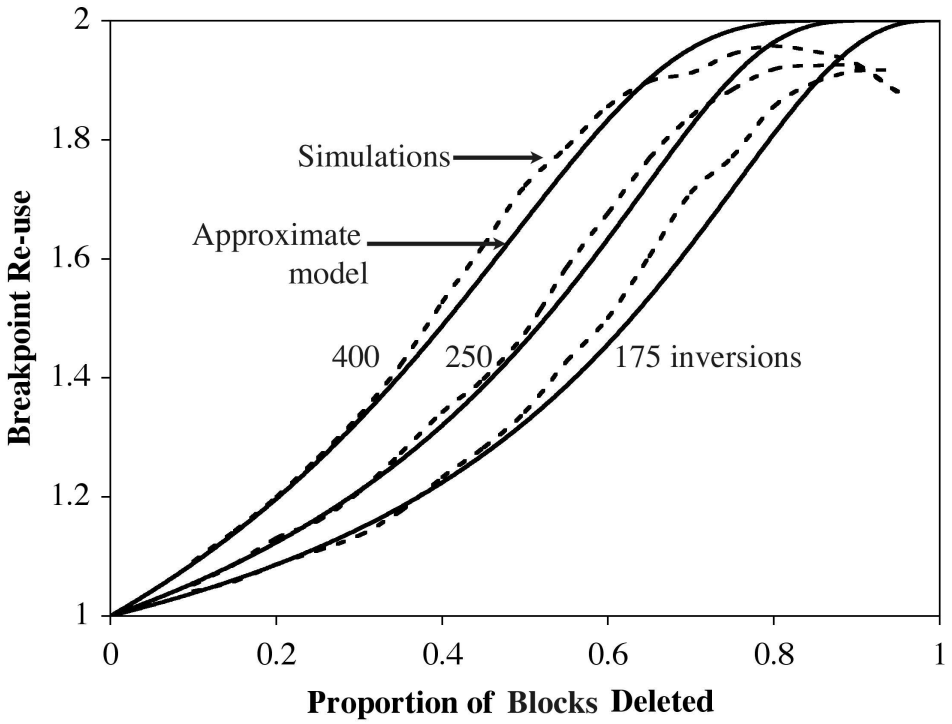

Figure 2: Plot of r predicted by the recurrence compared to true value estimated by simulation.

order to be comparable to the human-mouse comparison.

2. generate k "micro"-inversions of exactly w terms each randomly placed throughout the genome, without any restriction on whether breakpoints are re-used.

3. amalgamate all adjacent terms into blocks such that two adjacent blocks remain unamalgamated only if no two integers i and j, where term i or $-i$ is in one block and j or $-j$ is in the other, are within w of each other, i.e., if $|i - j| \leq w$, then the two blocks should be amalgamated.

4. delete any blocks containing only 1 or 2 terms. These represent blocks containing less than 1/2500 of the genome, comparable to the Pevzner-Tesler threshold of 1 Mb, approximately 1/3000 of the human genome.

5. apply the Hannenhalli-Pevzner algorithm to the remaining blocks and calculate the re-use rate r.

Figure 3 shows the average re-use rate in simulations of this process with and without Step 4. It can be seen that within the parameters of these experiments, the amalgamation process by itself substantially increases r, but not above 1.4, and that the maximum is attained for an intermediate size for block-size threshold. Moreover, beyond a certain number of microrearrangements, re-use actually decreases. It is with the deletions of short blocks, however, that re-use increases dramatically.

6. DISCUSSION.

This work was motivated by Pevzner and Tesler's [10] introduction of the re-use statistic. Though they used it to infer relative susceptibility of genomic regions to rearrangement, in our analysis it serves rather to measure the loss of signal of evolutionary history, due to the imposition of thresholds for retaining syntenic blocks and for repairing microrearrangements. We have taken account of the human-mouse comparison in fixing our parameters: a few hundred large rearrangements and a few thousand microrearrangements. However, we take issue neither with the imposition of thresholds, which seem to be methodologically reasonable, nor with the conclusions about inhomogeneities of genomes in their susceptibility to breakpoints. Nevertheless, we have shown that breakpoint re-use of the same magnitude as found in [10] may very well be artifacts of the use of thresholds in a context where NO re-use actually occurred. Indeed, while this may not have been their goal, Pevzner and Tesler have invented a statistic that is a measure of the noise affecting a genomic rearrangement process at the sequence level. Given some information about the parameters of rearrangement, the number of blocks and the size of the thresholds, the re-use rate tells us whether we can have confidence in evolutionary signal reconstructed, whether it must be considered largely random, or whether we are in the "twilight" zone.

Acknowledgments

Research supported by grants from the Natural Sciences and Engineering Research Council (NSERC). DS holds the Canada Research Chair in Mathematical Genomics and is a

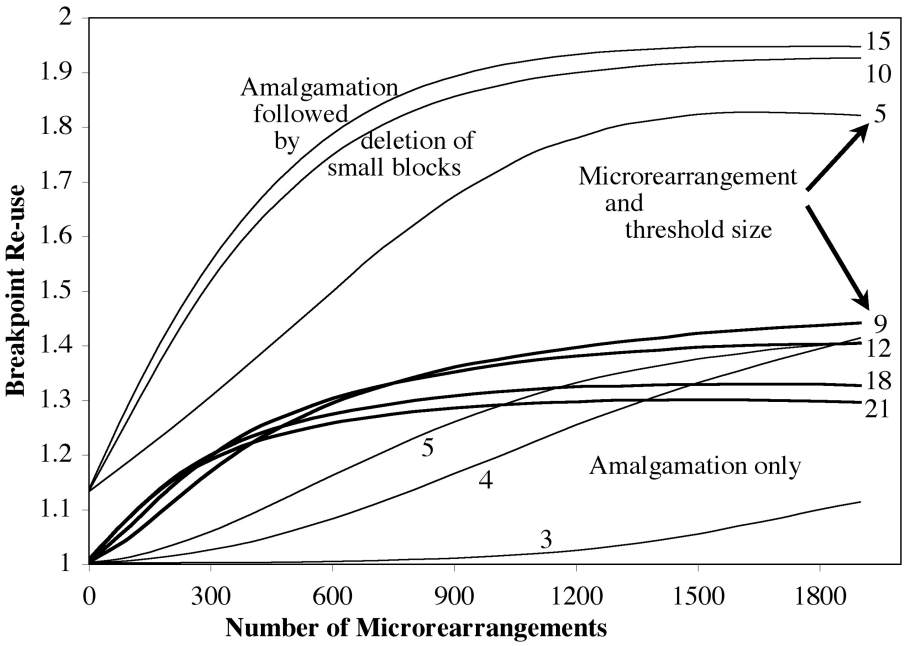

Figure 3: Effect of amalgamation, and of amalgamation followed by deletion of small blocks, on re-use rate r, as a function of threshold size w and number of microrearrangements.

Fellow in the Evolutionary Biology Program of the Canadian Institute for Advanced Research.

7. REFERENCES

[1] Bader DA, Moret BM, Yan M (2001) A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. *J Comput Biol* 8:483-91.

[2] Bergeron A. (in press) A very elementary presentation of the Hannenhalli-Pevzner theory. *Discrete Applied Mathematics*.

[3] Hannenhalli S (1996). Polynomial-time algorithm for computing translocation distance between genomes. *Discrete Applied Mathematics* 71:137–151.

[4] Hannenhalli S, Pevzner PA (1995) Transforming men into mice (polynomial algorithm for genomic distance problem). In *Proceedings of the IEEE 36th Annual Symposium on Foundations of Computer Science*: 581–592.

[5] Hannenhalli S, Pevzner PA (1999) Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). *Journal of the ACM* 48:1–27.

[6] Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D (2003) Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci USA* 100:11484-11489

[7] Nadeau JH, Taylor BA (1984) Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc Natl Acad Sci USA* 81:814-8.

[8] Pevzner PA, Tesler G (2003) Genome rearrangements in mammalian genomes: Lessons from human and mouse genomic sequences. *Genome Research* 13: 37-45

[9] Pevzner PA, Tesler G (2003) Transforming men into mice: The Nadeau-Taylor chromosomal breakage model revisited. In *Proceedings of RECOMB 03, Seventh International Conference on Computational Molecular Biology*. ACM Press: 247-256.

[10] Pevzner PA, Tesler G (2003) Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc Natl Acad Sci USA* 100:7672-7

[11] Sankoff D, El-Mabrouk N (2002) Genome rearrangement. In *Current Topics in Computational Biology*. Edited by Jiang T, Smith T, Xu Y, Zhang M. Cambridge, MA: MIT Press; 135-155.

[12] Tesler G (2002) GRIMM: Genome rearrangements web server. *Bioinformatics* 18:492-3.

[13] Waterston R *et al.* (2002) Initial sequencing and analysis of the mouse genome. *Nature* 420: 520-562.

[14] Watterson G, Ewens W, Hall T, Morgan A (1982) The chromosome inversion problem. *Journal of Theoretical Biology* 99:1–7.