Power Boosts for Cluster Tests

David Sankoff and Lani Haque

Department of Mathematics and Statistics, University of Ottawa,
585 King Edward Street, Ottawa, Canada, K1S 4T8
{sankoff, lhaque}@uottawa.ca

Abstract. Gene cluster significance tests that are based on the number of genes in a cluster in two genomes, and how compactly they are distributed, but not their order, may be made more powerful by the addition of a test component that focuses solely on the similarity of the ordering of the common genes in the clusters in the two genomes. Here we suggest four such tests, compare them, and investigate one of them, the maximum adjacency disruption criterion, in some detail, analytically and through simulation.

1 Introduction

The detection of a number of genes in close proximity in two genomes may suggest an evolutionary association of these genes or indicate a functional relationship among them. Recently studied tests of significance of such gene clusters [2–4] have focused on the number of genes in common in relatively short chromosomal intervals in the two genomes, and not on the order of these genes. For different gene clusters having the same number of genes, spatially distributed in the same way, we might want to consider those clusters where the order is more similar, though not necessarily identical, in the two genomes as being more significant, and more indicative of a historical or functional relationship, than clusters whose gene orders in the two genomes bear little relationship. For tests such as those in the above-cited studies, where the significance level is independent of gene order within the cluster, this level can be enhanced by taking into account the significance of the gene-order similarity within the cluster in the two genomes. This combined test would have greater power against the null hypothesis of random gene order at the genomic scale in favour of alternative hypotheses derived from evolutionary or functional models.

In this this paper, we first suggest four different ways of defining an order-based "boost" to cluster significance, and investigate one of them in enough detail so that it can be used for all cluster sizes and all values of the similarity criterion. In Section 2 we define four measures of gene-order divergence and link them to previous work on genome rearrangement and gene clusters. In Section 3, we show some properties of the maximum adjacency disruption measure of order similarity. In the rest of the paper we develop tests based on this measure; in Section 4, an exact test for certain values of the maximum difference measure, and in Section 5 exact tests for all small clusters. For large clusters,

A. McLysaght et al. (Eds.): RECOMB 2005 Ws on Comparative Genomics, LNBI 3678, pp. 121–130, 2005.

we tabulate significance levels based on large scale simulations in Section 6, and on a simplified probabilistic model in Section 7. In Section 8, we compare the critical regions three of the four proposed tests, on moderate-sized clusters. In Section 9, we discuss the applicability of our method, and directions for further research.

2 Four Measures of Gene-Order Similarity and Their Motivations

Suppose we have identified, using some existing criterion, e.g., one of those in [2–4], k genes that form a cluster in both genome A and genome B. Number the clustered genes in genome A in order from 1 to k (ignoring any intervening genes that are not in the scope of the cluster in genome B) and let $g_1, \cdots, g_k$ be the order of these same genes in genome B. Similarly, re-number the genes from 1 to k according to their order on genome B, and let $h_1, \cdots, h_k$ be the order of these same genes in genome A.

1. The maximum adjacency disruption criterion (MAD):

$$\text{MAD} = \max_{i=1,\cdots,k-1} \{\max\{|g_i - g_{i+1}|, |h_i - h_{i+1}|\}\},$$

   the maximum, over all pairs of adjacent genes in the cluster in either genome, of the difference in their positions in the gene order in the cluster in the other genome. A low value of MAD means that no gene in the cluster has drifted far from its position in the ancestral genome. MAD is symmetric with respect to A and B. An asymmetric criterion somewhat similar to MAD was used in [1].
2. The summed adjacency disruption criterion (SAD):

$$\text{SAD} = \sum_{i=1,\cdots,k-1} \{|g_i - g_{i+1}| + |h_i - h_{i+1}|\},$$

   the sum, over all pairs of adjacent genes in the cluster in both genomes, of the difference in their positions in the gene order in the cluster in the other genome. This measures the overall movement of genes within the cluster from their positions in the ancestral genome.
3. The breakpoint metric (BAD):

$$\text{BAD} = \#_{(i=1,\cdots,k-1)}\{|g_i - g_{i+1}| > 1\},$$

   the number of times a pair of genes adjacent in the cluster in one genome is not adjacent in the other. This is in effect a simple count of the adjacency disruptions and has been used in comparisons of entire gene orders of genomes [5], whereas here we are focusing on the order of genes in the cluster only.
4. The rearrangement distance (RAD). The number of rearrangement operations (e.g., inversions, transpositions, block interchanges) required to transform the order of the genes in one cluster into the order in the other one [6]. This is the only measure we do not analyze here, for reasons detailed in Section 9.

To test for a significant level of gene-order correspondence between a cluster's realizations in two genomes, we need to know the distribution of the test statistic under a suitable null hypothesis, normally that gene order is purely random. This may be done by counting the number of permutations of the integers from 1 to k having a given value of the statistic, either exhaustively, or by means of a computing formula if this is available, or estimated through simulation or approximated by a continuous model. We will calculate the distribution of the MAD statistic, using all these approaches, depending on the cluster size k, and we will compare it to SAD and BAD on all permutations of size 11 and 12.

3 Maximum Adjacency Disruption; One-Sided and Two-Sided

Defining the one-sided versions of MAD,

$$\mathrm{MAD}_{AB} = \max_{i=1,\cdots,k-1}\{|g_i - g_{i+1}|\}, \mathrm{MAD}_{BA} = \max_{i=1,\cdots,k-1}\{|h_i - h_{i+1}|\},$$

it is not the case that MAD_{AB} always equals MAD_{BA}.

Example 1. : Consider B = 3 1 2 4. Then $\mathrm{MAD}_{AB} = 2$. Renumbering the genes in order in B as 1 2 3 4 translates into A = 2 3 1 4, so that $\mathrm{MAD}_{BA} = 3$.

Limiting the MAD value of a cluster not only constrains pairs of adjacent genes in one genome from being far apart in the other, it also keeps genes in the larger neighbourhood of a given gene from dispersing too widely. For a given k,

Proposition 1. *If*

$$|g_i - g_{i+1}| \leq a, \text{ for } 1 \leq i < k$$

then

$$|g_i - g_j| \leq a|i - j|, \text{ for } 1 \leq i, j \leq k.$$

Proof:

$$|g_i - g_j| = |\sum_{m=i}^{j-1} g_m - g_{m+1}| \tag{1}$$

$$\leq \sum_{m=i}^{j-1} |g_m - g_{m+1}| \tag{2}$$

$$\leq a|i - j| \tag{3}$$

Corollary 1. *Let $f(x)$ be an increasing concave function on $[1, \cdots, k]$. Then*

$$f(|i - j|) \leq f(1)|i - j|.$$

4 Enumerating Clusters Where $a = 1, 2$ and 3, and Where $a = k - 2$ and $a = k - 1$

Being able to calculate the number of clusters of size k with a specific value of a criterion is helpful for the construction of tests based on this criterion. Computing formulae are currently known only for BAD (entry A001100 in Sloane's On-Line Encyclopedia of Integer Sequences [7].) Here we give partial results for MAD. Let $n(a, k)$ be the number of clusters of length k where MAD $\leq a$.

Proposition 2. *The following statements hold:*

1. $n(1, k) = 2$, *for all* k
2. *For* $a = 2$, $k > 5$

$$n(2, k) = n(2, k - 1) + n(2, k - 3).$$

3. *For* $a = 3$, $k > 12$

$$\begin{aligned} n(3, k) = n(3, k - 1) + n(3, k - 2) + 3n(3, k - 4) \\ +3n(3, k - 5) - 3n(3, k - 6) - 3n(3, k - 7) \end{aligned}$$

4. $n(k - 1, k) = k!$, *for* $k > 1$.
5. $n(k - 2, k) = (k - 2)!(k^2 - 5k + 8)$, *for* $k > 2$.

Proof:

1) The only clusters that have MAD $= 1$ are $1 \cdots k$ and $k \cdots 1$.

2) One type of "valid" cluster, i.e., one that has MAD ≤ 2, is of form $k\ \gamma$, where γ is any valid cluster on the first $k-1$ integers starting with $k-1$ or $k-2$, or of form $\kappa\ k$, where κ is any valid cluster on the first $k - 1$ integers ending with $k-1$ or $k-2$. There are $n(2, k-1)$ of these. The other type of cluster that has MAD ≤ 2 is of form $k-1\ k\ k-2\ \gamma$, where γ is any valid cluster on the first $k - 3$ integers starting with $k - 3$ or $k - 4$, or of form $\kappa\ k-2\ k\ k-1$, where κ is any valid cluster on the first $k - 3$ integers ending with $k - 3$ or $k - 4$. There are $n(2, k - 3)$ of these. Thus, for large enough k (starting with $k = 6$), we have $n(2, k) = n(2, k - 1) + n(2, k - 3)$.

3) An argument similar to that for $a = 2$, but with many more cases, shows that for k even,

$$\begin{aligned} n(3, k-2) = n(3, k-3) + n(3, k-5) + 3n(3, k-6) + 3n(3, k-7) \\ + \sum_{i=2}^{\frac{k-4}{2}} n(3, k - (2i + 3)) \end{aligned}$$

So that

$$\begin{aligned} & n(3, k) - n(3, k - 2) \\ &= n(3, k - 1) + 3n(3, k - 4) + 3n(3, k - 5) - 3n(3, k - 6) - 3n(3, k - 7) \end{aligned}$$

and the analogous calculation for k odd yields the same result. Therefore,

$$n(3,k) = n(3,k-1) + n(3,k-2) \\ +3[n(3,k-4) + n(3,k-5) - n(3,k-6) - n(3,k-7)].$$

4) The criterion $a = k-1$ holds for any of the $k!$ clusters.

5) This follows from $n(k-2,k) = n(k-1,k) - n'(k-1,k)$, where $n'(k-1,k) = 4(k-2)(k-2)!$ is the number of clusters with a maximum disruption score of exactly $k-1$. The latter is the number of clusters such that $\{g_1, g_k\} = \{1,k\}$ plus the number such that $\{h_1, h_k\} = \{1,k\}$ less the number where both conditions hold. All of these may be evaluated in a straightforward manner, yielding the above expression for $n'(k-1,k)$.

The first few values of $n(2,k)$ are $0,0,6,8,12,18,26,\cdots$. The recurrence $n(2,k) = n(2,k-1)+n(2,k-3)$ only "kicks in" at $k=6$. The particular series of numbers starting with $0,0,6,8,12,18$ has not been mentioned before, but the recurrence, with different initial values, has cropped up in many different contexts, listed as sequence A000930 in Sloane's On-Line Encyclopedia of Integer Sequences [7].

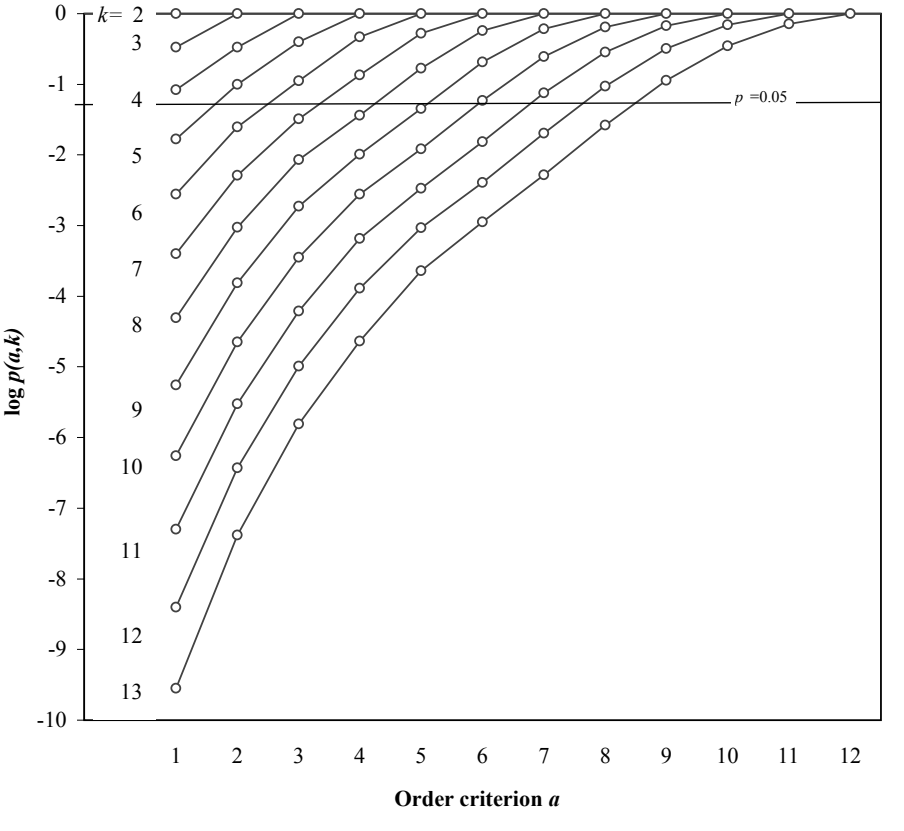

Fig. 1. Proportion of clusters of size k with maximum adjacency disruption $\leq a$; exact counts for small k

5 Exact Enumeration for Moderate Values of *k*

As a first step in constructing a usable test based on MAD, for $k \leq 13$, we generated all permutations on the integers from 1 to k and calculated the MAD value for each when compared to the identity permutation. We then calculated the $n(a, k)$ as defined in Section 4, normalized by $k!$ to compute the p-value $p(a, k)$, and plotted the results on a logarithmic scale in Figure 1. Given a cluster with MAD $= a$, then, its statistical significance can be assessed from the curve for clusters of size k.

6 Simulations for Large *k*

Though it would be feasible using high-performance methods to calculate the test exactly for k somewhat larger than 13, this would exceed 16 or 17 with great difficulty. Instead, we simply constructed 100,000 random clusters with

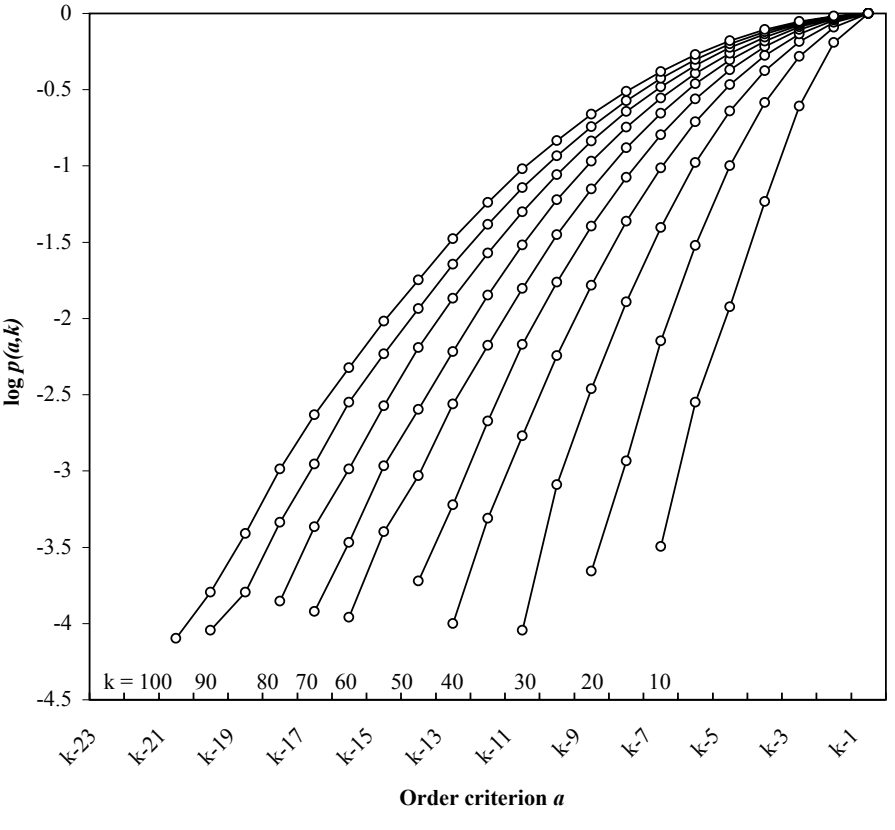

Fig. 2. Prop1ortion of clusters of size k with maximum adjacency disruption $\leq a$; estimated values for larger k based on 100,000 randomly generated clusters

k terms, for k running from 10 to 100, in steps of 10. The curves in Figure 2 were constructed in the same way as the previous figure, except that the normalization was by 100,000 instead of by $k!$, and that the curves are plotted against $a = k-1, k-2, \cdots$ rather than against $a = 1, 2, \cdots$.

7 A Model for Large k

Let $\boldsymbol{\alpha} = \frac{a}{k}$. To model the MAD criterion $|g_i - g_{i+1}| \leq a$ for large k under the null hypothesis, i.e., for genes randomly ordered within clusters, we consider two points at random on the real unit interval, and ask what is the probability they are within $\boldsymbol{\alpha}$ of each other, where $0 < \boldsymbol{\alpha} < 1$. This is just $1-(1-\boldsymbol{\alpha})^2$, representing the probability of adjacency disruption of less than $\boldsymbol{\alpha}$. Since k is large, we explore the assumption that these disruptions are independent across all $2k$ adjacencies in the two genomes. Then the probability that all the disruptions are less than $\boldsymbol{\alpha}$ is $[1-(1-\boldsymbol{\alpha})^2]^{2k} = [\boldsymbol{\alpha}(2-\boldsymbol{\alpha})]^{2k}$.

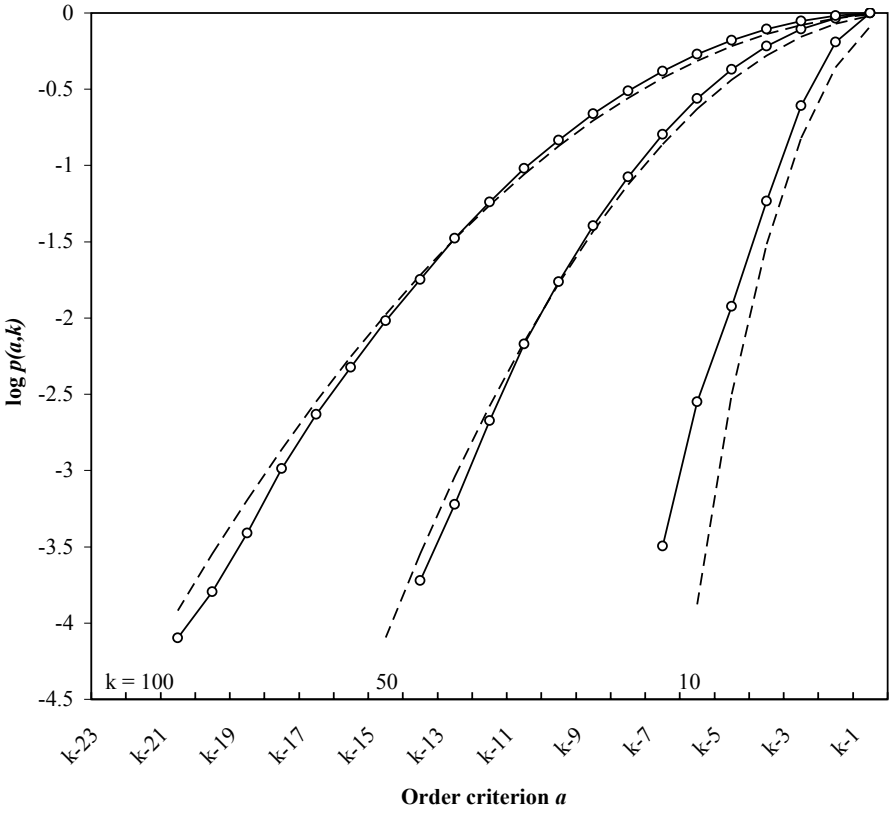

Fig. 3. Fit of simplified model to exact or simulated data for $k = 10, 50$ and 100. Dashed lines represent model, lines with data points drawn from exact values ($k = 10$) or simulations ($k = 50, 100$).

Plotting $p = [\frac{a}{k}(2 - \frac{a}{k})]^{2k}$ in Figure 3 demonstrates an improved fit of this simplified model for large values of k (50-100), compared with the case $k = 10$.

8 A Comparison of Adjacency Disruption Measures

Tests based on the four measures listed in Section 2 will not, of course, all have the same critical region. To compare the MAD, BAD and SAD criteria, we calculated them on all 4×10^7 clusters of size 11, and counted how many clusters fell into the critical regions of both members of a pair of tests. We aimed for equal critical regions of size close to 5%, but because we were restricted to discrete choices of a, the closest we could come was defined by an a for each test (detailed in the table legend) that resulted in approximately 8% of the criteria values falling on or below this threshold. We then repeated this for the 4.79×10^8 clusters of size 12; this time the critical regions closest to 5% all had size approximately 2%.

Table 1. Differences among critical regions of three tests. $k = 11$, total clusters 4×10^7, criteria for MAD: $a \leq 7$, BAD: $a \leq 6$, SAD: $a \leq 63$. $k = 12$, total clusters 4.79×10^8, MAD: $a \leq 7$, BAD: $a \leq 6$, SAD: $a \leq 67$.

Clusters in critical regions ($\alpha = 7.5 - 9.0\%$)	MAD	SAD	BAD
$(k = 11) R_i (\times 10^6)$	3.01	3.06	3.55
	Compared with SAD	BAD	MAD
Intersection $R_i \cap R_j$	1.09	1.7	0.51
Union $R_i \cup R_j$	4.98	4.91	6.05
Symmetric difference $R_i \Delta R_j$	3.89	3.2	5.54
Normalized difference $\frac{R_i \Delta R_j}{R_i \cup R_j}$	**0.782**	**0.653**	**0.916**
Clusters in critical regions ($\alpha = 2.0 - 2.3\%$)	MAD	SAD	BAD
$(k = 12) R_i (\times 10^6)$	9.65	10.1	11
	Compared with SAD	BAD	MAD
Intersection $R_i \cap R_j$	3.28	3.91	0.81
Union $R_i \cup R_j$	16.46	17.18	19.84
Symmetric difference $R_i \Delta R_j$	13.18	13.27	19.03
Normalized difference $\frac{R_i \Delta R_j}{R_i \cup R_j}$	**0.801**	**0.772**	**0.959**

Table 1 shows that SAD is closer to both MAD and BAD than they are to each other. This is understandable in that the sum of the adjacency disruptions should reflect not only the number of disruptions but also the size of the largest one. The number of disruptions (BAD), however, and the size of the largest one (MAD), would seem only very indirectly related. Indeed, almost all the clusters satisfying both the MAD and BAD criteria (0.51×10^6 for the $k = 11$ experiment, and 0.81×10^6 for the $k = 12$ experiment) also satisfy the SAD criterion (0.48×10^6 and 0.75×10^6, respectively).

9 Discussion

The study of gene clusters is increasingly focused on genomes where the genes have been located by virtue of genomic sequencing. Such clusters include not only gene position and hence gene order, but also gene orientation, or strandedness. In comparative genomics, such data are represented not by ordinary permutations, but by signed permutations, where the sign on a term in genome B indicates the DNA strand, or reading direction, relative to the direction of the same gene in genome A. In this context, we can simply ignore the sign, or else devise some way of taking it into account. For BAD, and particularly for RAD, signed permutations are the natural domain of application. This is one reason why we did not analyze RAD in the present study. The main adaptation is that in signed permutations, the configuration $i+1, i$ in genome B is considered a disruption of adjacency, but $-(i+1), -i$ is not, when the two genes are ordered as $i, i+1$ in genome A. For MAD and SAD, there is no natural way of taking sign into account and it seems most appropriate to ignore it.

Note that both MAD and SAD are asymmetrical, in that MAD_{AB} is not always equal to MAD_{BA} and SAD_{AB} is not always equal to SAD_{BA}. MAD is symmetrical by virtue of taking the maximum over both directions, and SAD by summing over both directions. Both BAD and RAD are symmetrical, on the other hand, and that is why it suffices to define them asymmetrically as we did in Section 2.

How should our tests for gene order be combined with tests for gene clustering such as r-windows and max-gap in [2–4]? The intuitive notion of a cluster involves spatial proximity of a group of genes similarly ordered in both genomes. But the most straightforward way of combining the two kinds of test, simply multiplying the two significance levels, is not an acceptable strategy, since it then only requires that one of the two tests be significant for the combined test to be significant. For example, if a putative cluster with k genes is evenly spaced across the entire genome, so that it is really the antithesis of the intuitive notion a cluster, but the order is $1 \cdots k$, then it will be still be highly significant (for large k) when the two significance levels (clustering and order) are multiplied together. The critical region of the combined test thus includes groups of genes which cannot be considered clusters. This problem does not seem to have been studied in the literature, where it is sometimes assumed that the significance level of a cluster of k genes may be enhanced by a factor of $(k!)^{-1}$ if the gene order is identical in both genomes.

For clusters with borderline (or better) p-values, however, the multiplicative strategy effectively boosts the power of the cluster test against evolutionarily or functionally meaningful alternatives.

Acknowledgements

Research supported in part by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC). DS holds the Canada Research Chair in

Mathematical Genomics and is a Fellow of the Evolutionary Biology Program of the Canadian Institute for Advanced Research. The authors thank the referees for numerous corrections and helpful comments.

References

1. Calabrese, P.P., Chakravarty, S. and Vision, T.J. 2003. Fast identification and statistical evaluation of segmental homologies in comparative maps. Bioinformatics 19, i74–i80.
2. Durand, D. and Sankoff, D., 2003. Tests for gene clustering. Journal of Computational Biology 10, 453–482.
3. Hoberman, R., Sankoff, D. and Durand, D. 2005. The statistical significance of max-gap clusters. in Lagergren, J. (ed) RECOMB Satellite Workshop on Comparative Genomics. LNBI 3388, 55–71. Berlin, Heidelberg: Springer Verlag.
4. Hoberman, R., Sankoff, D. and Durand, D. 2005. The statistical analysis of spatially clustered genes under the maximum gap criterion. Journal of Computational Biology, in press.
5. Sankoff, D. and .Blanchette, M. 1998. Multiple genome rearrangement and breakpoint phylogeny. Journal of Computational Biology 5, 555-570
6. Sankoff, D. and El-Mabrouk, N. 2002. Genome rearrangement. in Jiang, T., Smith, T., Xu, Y. and Zhang, M (eds) Current Topics in Computational Biology, 135–155. Cambridge, MA: MIT Press.
7. N. J. A. Sloane. 2005. The On-Line Encyclopedia of Integer Sequences, http://www.research.att.com/∼njas/sequences/.