JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 12, Number 8, 2005
© Mary Ann Liebert, Inc.
Pp. 1083–1102

The Statistical Analysis of Spatially Clustered Genes under the Maximum Gap Criterion

ROSE HOBERMAN,[1] DAVID SANKOFF,[2] and DANNIE DURAND[3]

ABSTRACT

Statistical validation of gene clusters is imperative for many important applications in comparative genomics which depend on the identification of genomic regions that are historically and/or functionally related. We develop the first rigorous statistical treatment of max-gap clusters, a cluster definition frequently used in empirical studies. We present exact expressions for the probability of observing an individual cluster of a set of marked genes in one genome, as well as upper and lower bounds on the probability of observing a cluster of *h* homologs in a pairwise whole-genome comparison. We demonstrate the utility of our approach by applying it to a whole-genome comparison of *E. coli* and *B. subtilis*. Code for statistical tests is available at *www.cs.cmu.edu/∼durand/Lab/software.html*.

Key words: comparative genomics, genome evolution, genome rearrangements, max-gap clusters, gene teams.

1. INTRODUCTION

There are many important applications in genomic comparison that require the identification of homologous regions. Researchers are interested in finding conserved groups of genes for identification of large-scale duplications (surveyed by Wolfe [2001]), reconstructing chromosomal rearrangements (surveyed by Sankoff [2003], and by Sankoff and Nadeau [2003]), and phylogenetic reconstruction (Blanchette *et al.*, 1999; Cosner *et al.*, 2000; Hannenhalli, *et al.*, 1995; Sankoff *et al.*, 2000a, b; Tamames *et al.*, 2001), as well as detecting operons, horizontal transfer, and functional selection in bacteria (surveyed by Chen *et al.* [2004], Lawrence and Roth [1996], and Tamames [2001]).

Our goal is to provide formal statistical models to test the hypothesis that two genomic regions in distantly related genomes share a common ancestor, against a null hypothesis of random gene order. In diverged genomes, the signature of such regions, or *gene clusters*, is similar gene content, but neither content nor order are strictly preserved. In order to construct formal statistical models, this intuitive notion of a gene cluster must be formalized into a precise mathematical definition.

While a number of definitions have been proposed, we focus in the current work on the *max-gap cluster*, a definition that has emerged as perhaps the most popular in empirical studies (Blanc *et al.*, 2003; Bourque *et al.*, 2005; Chen, *et al.*, 2004 Friedman and Hughes, 2001; Luc *et al.*, 2003; McLysaght *et al.*, 2002;

[1] Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, 15213, USA.
[2] Department of Mathematics and Statistics, University of Ottawa, Ottawa, ONT K1N 6N5, Canada.
[3] Departments of Biological Sciences and Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA.

Overbeek *et al.*, 1999; Simillion *et al.*, 2002; Tamames, 2001; Vandepoele *et al.*, 2002 Vision *et al.*, 2000). In a max-gap cluster, the distance between genes in the cluster is constrained to be no more than a constant g. Although max-gap clusters are widely used in practice, no analytical statistical tests have been developed for them. Studies based on max-gap criteria currently use randomization to estimate the significance of clusters (Blanc *et al.*, 2003; McLysaght *et al.*, 2002; Overbeek *et al.*, 1999; Simillion *et al.*, 2002; Vandepoele *et al.*, 2002; Vision *et al.*, 2000).

In this paper, we present the first formal, rigorous mathematical model of max-gap gene cluster probabilities. Models are developed for two basic cluster finding scenarios. In the first scenario, we wish to find clusters of a subset of genes that are prespecified, or *marked*. These genes may be of interest, for example, because their homologs are contiguous in another region or genome, or because they participate in the same functional pathway. In the second scenario, the set of genes in a cluster emerges from the comparison of two whole genomes; we are given two genomes and a mapping between their homologs, and we wish to find clusters of homologs found in close proximity in both genomes.

For the marked genes scenario (Section 3), we present an exact expression for the probability of observing a complete max-gap cluster containing all m marked genes within a randomly ordered genome of size n. Next we extend this analysis to evaluate the probability of observing a cluster containing only a subset of the marked genes. We present an algorithm that calculates the exact probability of observing an incomplete cluster containing $h < m$ marked genes, as well as an analytic solution for the case where $m/2 < h \leq m$. In Section 4, we discuss the statistical implications of choosing different algorithms for identifying max-gap clusters in a whole-genome comparison. In Section 5, we develop upper and lower bounds for the probability of observing a max-gap cluster of h homologs by a pairwise whole genome comparison scenario.

To investigate trends in the cluster probabilities for both models, we use the equations presented in Sections 3 and 5 to calculate the probability of clusters for common choices of parameter values and for a range of typical sizes of prokaryotic and eukaryotic genomes. We present these results graphically and discuss the influence of genome size, gap size, the number of marked genes, and the cluster size on cluster significance. We also determine the regions of the parameter space that yield clusters that are statistically significant. To demonstrate the use of our methods, a whole genome comparison of the *E. coli* and *B. subtilis* genomes was conducted and the results analyzed statistically.

2. PRELIMINARIES

We employ a commonly used genome model in which a genome is represented as an ordered set of n genes: $G = 1 \ldots n$. We assume a single unbroken chromosome and that genes do not overlap. The distance between two genes in this model is simply the number of genes between them.

A *max-gap cluster* is described by a single parameter g:

Definition 2.1. *A set of genes forms a chain if the distance from any gene in the chain to at least one other gene in the chain is never more than g. A set of genes form a max-gap cluster if they form a max-gap chain and are not included within any larger chain (i.e., the chain is* maximal*).*

For example, when the maximum gap allowed is $g = 3$, two max-gap clusters (indicated by boxes) are found in the example genome in Fig. 1. The remaining gene cannot be clustered with any other genes. Clusters are characterized by their size and length, where the *size* of a cluster is simply the number of marked genes it contains and the *length* is the total number of marked and unmarked genes. These concepts are illustrated graphically in Fig. 1.

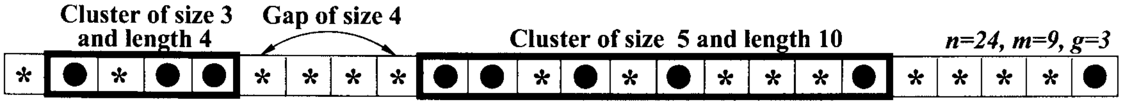


FIG. 1. A sample genome ($n = 24$), with $m = 9$ marked genes shown in black. Two clusters are found when the maximum gap allowed is $g = 3$. The rightmost marked gene is not part of any cluster.

Preliminary computations. In the following sections, it will be helpful to have an expression for the number of ways of arranging k marked genes to form a max-gap cluster within a window of length l. When both endpoints of the window contain marked genes, the cluster will be of length *exactly* l. Since such a cluster has $k-1$ gaps with a cumulative sum of $l-k$, the number of ways of creating a max-gap cluster of size k and length l is equivalent to the number of ways of rolling $k-1$ dice, each with faces numbered 0 to g, such that the sum of their faces is equal to $l-k$.[1] We can compute this quantity using the following expression:

$$d_e(k, g, l) = \sum_{i=0}^{\lfloor (l-k)/(g+1) \rfloor} (-1)^i \binom{k-1}{i} \binom{l-i(g+1)-e}{k-e}, \tag{1}$$

where the parameter e is set to 2, indicating that *both* endpoints of the window are constrained to contain marked genes. The number of ways of generating a cluster with length *no greater than* l is equivalent to requiring that a marked gene be placed in the leftmost endpoint of the window (the rightmost window may contain a marked gene but this is not required). This is simply $\sum_{r=k}^{l} d_2(k, g, r)$, which can be shown to be equivalent to $d_1(k, g, l)$ (see Hoberman *et al.* [2005] for the proof). For large values of l, however, there is a simpler expression for $d_1(k, g, l)$. Let w_{kg} be the maximum length of a chain of size k with maximum gap g, that is, $k + (k-1)g$. If $l \geq w_{kg}$, then $d_1(k, g, l) = (g+1)^{k-1}$. Finally, the number of ways of arranging m genes so that they form a max-gap cluster *anywhere* within a window of size l is $\sum_{r=k}^{l} d_1(m, g, r) = d_0(k, g, l)$.

3. STATISTICS FOR MARKED GENES

In this search scenario, m genes of interest are selected, or *marked*. These genes may be of interest, for example, because their homologs are contiguous in another region or genome, or because they participate in the same functional pathway. We are interested in the probability that all m genes, or a sizable subset, appear in close proximity within the genome of interest.

3.1. Exact probabilities for complete clusters

In this section, I consider the significance of a *complete* chain, containing all m marked genes. The null hypothesis is that the marked genes are randomly distributed in the genome, i.e., each permutation of the n genes is equally likely to occur. For this scenario, the test statistic is the maximum gap between the marked genes. We want to calculate a p-value, the probability of observing a gap between adjacent marked genes of at least g in a random genome.

Given a random permutation of n genes, the probability of observing all m marked genes (in any order) such that the gap between adjacent marked genes does not exceed g is

$$P(n, m, g) = \frac{1}{\binom{n}{m}} \begin{cases} (n - w_{mg} + 1)(g+1)^{m-1} + \dfrac{w_{mg} - m}{2}(g+1)^{m-1}, & \text{if } w_{mg} \leq n+1 \\ d_0(m, g, n) & \text{otherwise,} \end{cases} \tag{2}$$

where $n - w_{mg} + 1$ is the number of ways of placing the first marked gene, $(g+1)^{m-1}$ is the number of ways of placing the remaining marked genes (or equivalently, the number of ways of choosing $m-1$ gaps each between 0 and g), and the last term is the number of ways of constructing a max-gap cluster within the last $w-1$ genes in the genome. The derivation of the final term is given in Appendix 6.

[1] This in turn is equivalent to the number of ways of rolling a set of $k-1$ dice, each of which has faces numbered 1 to $g+1$, so that their cumulative sum in equal to $l-1$ (Uspensky, 1937).

3.2. Complete clusters with length restriction

Cluster definitions are often designed to control the density of genes in the cluster based on the intuition that clusters with very low density will often not be distinguishable from clusters that form by chance. For max-gap clusters, the maximum gap parameter prevents clusters from becoming too sparse in local regions. However, *global density*, i.e., the ratio of size to length, is only weakly constrained: it cannot fall below $1/(g+1)$. Additional control over the global density can be achieved by constraining the total length of the cluster. We simply add the restriction that all m marked genes must appear in a window of size at most r. Given a genome of size n, the probability of observing all m marked genes (in any order) in a max-gap cluster that is completely contained within a window of size r is

$$\frac{[(n-r+1)d_1(m,g,r)+d_0(m,g,r-1)]}{\binom{n}{m}}. \tag{3}$$

This length constraint may also be useful in situations in which only a local region of the genome is of interest.

3.3. Exact probabilities for incomplete clusters

In practice, researchers often look for clusters that contain only a subset of the m marked genes (Amores *et al.*, 1998; Coulier *et al.*, 1997; Endo *et al.*, 1997; Gibson and Spring, 2000; Hughes, 1998; Kasahara, 1997; Katsanis *et al.*, 1996; Lipovich *et al.*, 2001; Pebusque *et al.*, 1998; Ruvinsky and Silver, 1997; Trachtulec and Forejt, 2001). Thus, in this section, we provide a statistical test for *incomplete* max-gap clusters of size $h < m$, against a null hypothesis of random gene order. In this case, the maximum gap value g is fixed in advance, we search the genome for all max-gap clusters of marked genes, and we want to determine the significance of a cluster of size h. In this case, the test statistic is the size of the largest cluster, and the p-value is the probability under the null hypothesis that the largest cluster will be of size h or greater. This is simply the probability of observing *at least* one cluster of size h or greater in a random genome.

Unlike complete clusters, there can be more than one incomplete cluster in the same genome. A simple extension of Equation (2) to incomplete clusters would therefore lead to overcounting permutations containing more than one cluster. Instead, we use dynamic programming to count those permutations which *do not* contain a cluster of size h or larger and subtract to obtain the probability of observing at least one incomplete cluster.

The algorithm is defined by the following recursion relation

$$\eta[n,m,j,c] = \begin{cases} 0, & \text{if } c = h \text{ or } n < m \\ 1, & \text{else if } m = 0 \\ \eta[n-1,m,j+1,c] + \eta[n-1,m-1,0,c+1], & \text{else if } j \le g \\ \eta[n-1,m,j+1,c] + \eta[n-1,m-1,0,1], & \text{otherwise.} \end{cases}$$

This algorithm moves along the genome, adding a marked or unmarked gene at each step. It keeps track of runs of marked genes that satisfy the max-gap cluster criterion and avoids creating a cluster of size h or larger by judicious placement of unmarked genes. The quantity $\eta[n,m,j,c]$ represents the number of ways to place m marked genes in n slots without creating a max-gap cluster of size greater than or equal to h, where j is the distance to the previous marked gene and c is the size of any cluster created so far.

The probability of observing at least one incomplete cluster of size at least h is then just one minus the probability of observing no incomplete clusters:

$$Q(n,m,h,g) = 1 - \frac{\eta[n,m,g+1,0]}{\binom{n}{m}}. \tag{4}$$

The complexity of computing $Q(n, m, h, g)$ is $O(nmgh)$. Since $h < m$, this is bounded above by $O(nm^2g)$. However, in practice m will be significantly smaller than n. For example, the size of typical bacterial genomes ranges from 500 to 5,000 (Skovgaard *et al.*, 2001), whereas the average number of genes in an operon is predicted to be between two and four, and the large majority of operons contain fewer than fifteen genes (Zheng *et al.*, 2002). Vertebrate genomes can be much larger. For example, the estimated size of the human genome is around 25,000 genes (International Human Genome Sequencing Consortium, 2001), but duplicated or conserved regions reported in the literature tend to include only five to thirty genes in a window containing a hundred genes at most (surveyed in Durand and Sankoff, 2003; also Abi-Rached *et al.*, 2002, Danchin *et al.*, 2003). If we make the conservative assumption that $m \leq \sqrt{n}$ and that g is a small constant, then the running time will be bounded above by $O(n^2)$.

When $h > \frac{m}{2}$, each permutation can contain at most one cluster of size h or greater, so overcounting permutations with more than one cluster is not an issue, and the dynamic programming algorithm is not required. Instead, we directly count the number of permutations containing a cluster, enumerating them by the length l of the last h marked genes in the cluster. There are $n - l - g$ positions in which to place the last marked gene in the cluster. The number of ways to choose the gaps to obtain a cluster length of exactly l is $d_2(h, g, l)$. The remaining $m - h$ genes must be placed outside the window of size l. Note that they also must be placed at a distance of at least $g + 1$ from the end of the cluster to avoid extending the length of the cluster. Thus there are $n - l - g - 1$ ways of placing the remaining $m - h$ genes. When $h > \frac{m}{2}$, the probability of observing an incomplete cluster of size at least h is

$$\frac{1}{\binom{n}{m}} \sum_{l=h}^{h+g(h-1)} \left[\max(0, n - l - g) \cdot d_2(h, g, l) \cdot \binom{n - l - g - 1}{m - h} + \sum_{i=0}^{g} d_2(h, g, l) \cdot \binom{n - l - i}{m - h} \right], \quad (5)$$

where l ranges over all possible lengths of a cluster of size h and the final term addresses edge effects. The final term counts those clusters in which the last marked gene in the cluster is close to (within g of) the end of the genome. If the last marked gene is the i^{th} gene from the end of the genome, then there are $n - l - i$ ways to place the remaining $m - h$ genes. Using the upper summation identity (Graham *et al.*, 1989) the final term can be further simplified to

$$\sum_{i=0}^{g} d_2(h, g, l) \cdot \binom{n - l - i}{m - h} = d_2(h, g, l) \left[\binom{n - l + 1}{m - h + 1} - \binom{n - l - g}{m - h + 1} \right],$$

where the binomials are defined to be zero when the upper value is smaller than the lower value.

The complexity of computing Equation (5) depends on the extent to which subcomputations are reused, but empirically we observe that even a naive implementation has a substantially faster running time than Equation (4) (data not shown).

3.4. Experiments

The behavior of max-gap cluster statistics for a marked gene scenario was investigated by plotting the probabilities computed by Equations (2), (4), and (5) graphically. We selected parameter values corresponding to the range of values seen in real analyses. For example, we selected values of g ranging from 0 to 50, since typical values of this parameter used in genomic analyses range from 3 in bacteria (Tamames, 2001) to about 30 in human (McLysaght *et al.*, 2002). We calculated probabilities for genomes sizes of 500, 1000, 5000, 20,000, and 25,000, corresponding to typical gene sets for bacteria, yeast, worm, and higher eukaryotes like human and *Arabidopsis*.

3.4.1. Complete clusters. The probability of observing a complete cluster in a random genome was calculated from Equation (2), for varying values of n, m, and g. For complete clusters, we tested all values of m ranging from 2 to the genome size n.

Figure 2 shows the probability of observing a complete cluster containing all m marked genes in a genome of size $n = 1{,}000$, as m ranges from 2 to 1,000 and g increases from 5 to 50. The probability of

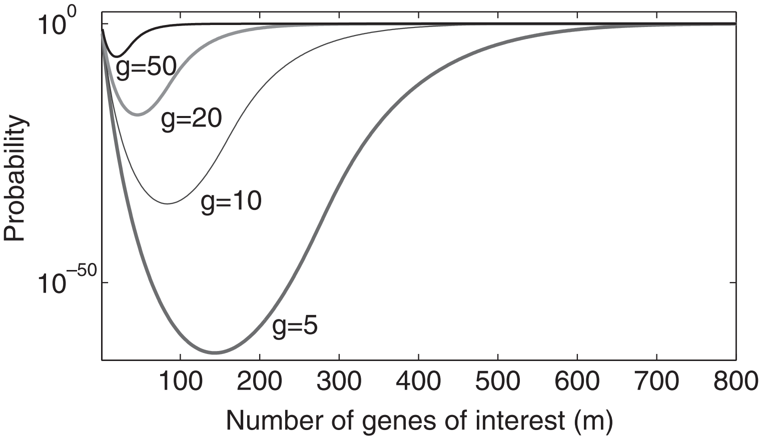


FIG. 2. Probability of a complete max-gap cluster of m marked genes in a genome of size $n = 1{,}000$ as a function of m, for $g = \{5, 10, 20, 50\}$.

observing a complete cluster increases monotonically with g. We might also expect that this probability will increase monotonically with m, or equivalently, that larger clusters will always be more significant, but this is not the case. As Fig. 2 shows, as m increases, the cluster probabilities first decrease and then increase. This makes sense intuitively if ones considers the extreme cases: when $m = 1$ or $m = n$ the probability of observing a complete cluster will clearly be one, and the values of m in between these two extremes will have probabilities of less than one. Calculations with larger genome sizes indicate that as n increases the probabilities decrease but the general trends seen in Fig. 2 remain the same (data not shown).

Another question of interest is the range of values of m and g for which it is possible to obtain a significant cluster. Figure 3 shows the parameter values for which the probability of observing a complete cluster in a genome of size 1,000 is no more than 0.001. The significant region of the parameter space is shown in black, indicating that as gap size increases, the range of values of m for which it is possible to obtain a significant cluster becomes more and more restricted.

3.4.2. Incomplete clusters. We calculated the probability of observing an incomplete cluster from Equations (4) and (5) for the values of n and g as stated above, and values of h ranging from 3 to m.

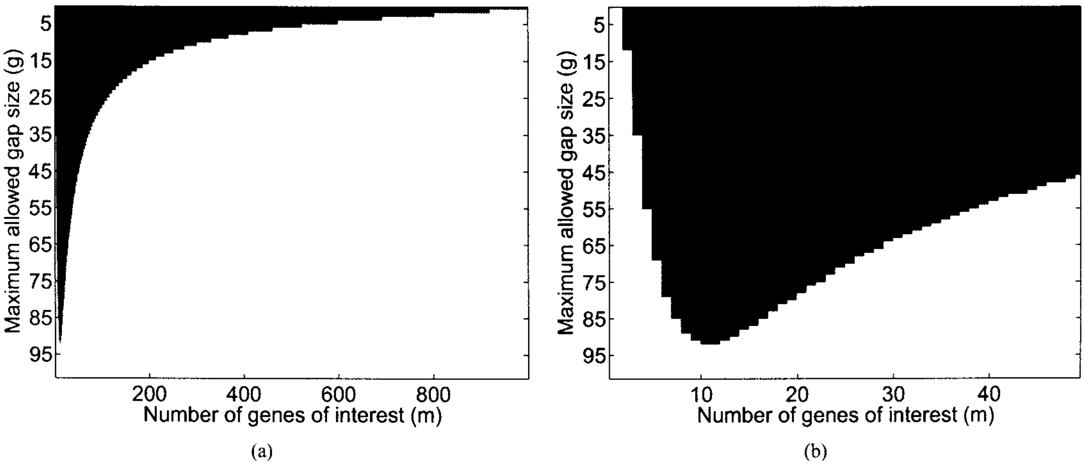


FIG. 3. Region of the parameter space that is statistically significant (shown in black) at the $\alpha = 0.001$ level for a complete cluster in a genome of size $n = 1{,}000$. (**a**) Complete parameter space where m ranges from 1 to 1,000. (**b**) Detail for $m \leq 50$.

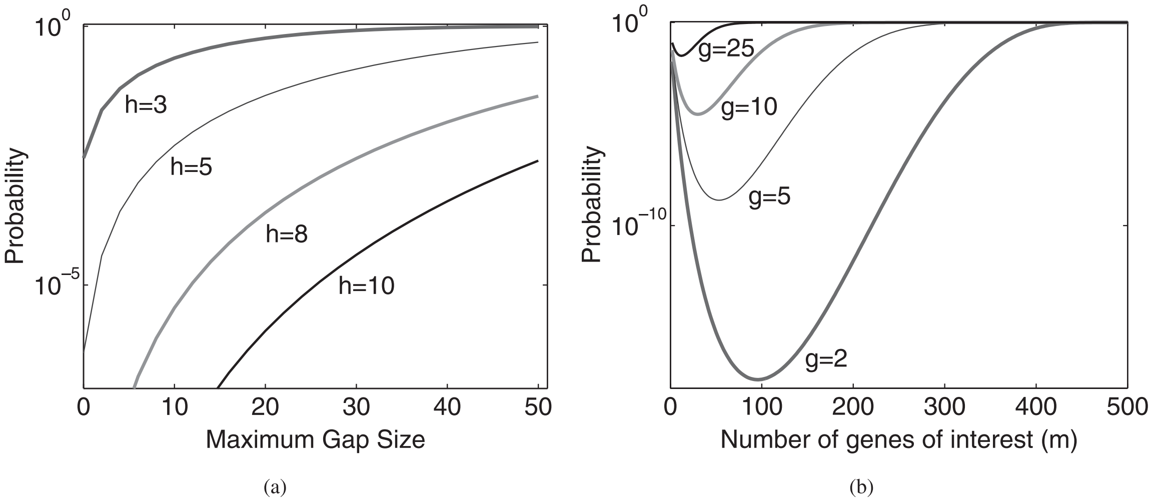


FIG. 4. (**a**) Probability of an incomplete cluster of size at least h when $n = 500$ and $m = 10$. (**b**) Probability of an incomplete cluster that contains at least half of all m marked genomes when $n = 500$.

Figure 4(a) shows that as the maximum gap size allowed increases, so does the probability of observing an incomplete cluster. Increasing the required size (h) of the cluster, on the other hand, decreases its probability of occurring by chance. Figure 4(b) shows the probability of max-gap clusters for varying values of m, when $h = \frac{m}{2}$. As in the case of complete clusters, the probabilities first decrease then increase with m. Probabilities were also calculated for larger genome sizes. Again, as n increases cluster probabilities decrease but the general trends are similar (data not shown).

4. WHOLE GENOME MODELS

In a *whole genome comparison*, we are given two genomes, G_1 and G_2, each of length n, and a mapping between the m homologs shared between G_1 and G_2. We use the term *homolog* either to refer to a gene in one genome that has a homolog in the other, or to refer to the pair, depending on context. A singleton is a gene without a homolog in the other genome. We are interested in assessing the significance of a cluster composed of a set of homologs found in proximity in both genomes, under the assumption that both the homologs and the singletons are uniformly randomly distributed throughout the genome.

Just as we had a formal definition of a max-gap cluster for a marked gene scenario, we now need to formulate a precise cluster definition of max-gap clusters found by whole genome comparison. Some cluster definitions are constructive, i.e., they specify an algorithm to find clusters, while others are formal, i.e., they specify precise criteria that a cluster must satisfy. Although a constructive definition makes it clear how to find clusters, a formal definition must be abstracted from the algorithm in order to develop statistical models. On the other hand, if a formal cluster definition is used, an additional search procedure to find clusters that satisfy the definition is required for data analysis. In both cases, it is necessary to verify that the constructive and formal definitions are equivalent.

Many groups state informally that they sought clusters where the maximum gap between genes is no greater than g, but provide neither a formal definition nor a precise constructive definition. This can be problematic, because an informal definition of a max-gap cluster could translate to more than one formal definition. Furthermore, different constructive definitions can identify very different sets of clusters. These clusters may have quite different properties and thus their statistical properties may differ substantially as well.

In this paper, we define a max-gap cluster *in the context of whole-genome comparison* as follows:

Definition 4.1. *A set of homologs form a max-gap cluster in two genomes if they form a chain in both genomes of interest and are not a subset of any larger cluster.*

An efficient divide-and-conquer algorithm for finding max-gap clusters (as defined here) via whole-genome comparison has been presented by Bergeron *et al.* (2002). Other groups use a greedy, bottom-up heuristic, in which larger clusters are built iteratively from smaller clusters (Calabrese *et al.*, 2003; Cannon *et al.*, 2003; Hampson *et al.*, 2005; Hokamp, 2001). In each step, the heuristic "looks ahead" a certain number of positions to see if additional homologs may be added to the cluster without violating the max-gap criterion. It can easily be shown that a simple greedy approach with a lookahead in either direction of size $g + 1$ will not find all max-gap clusters. For example, consider the two genomes $G_1 = \texttt{12*34*}$ and $G_2 = \texttt{31*4*2}$, where numbers indicate homologs and stars represent singletons. Regardless of the gene at which the algorithm starts, a greedy approach using a gap size of $g = 1$ will never find the max-gap cluster {1,2,3,4}. Unless the algorithm "looks ahead" all the way to the end of the genome—in which case it is no longer greedy—it is not guaranteed to find all max-gap clusters.[2]

It can be shown that the subset of max-gap clusters identified with a purely greedy heuristic are exactly the set of clusters that are nested, where a cluster of size h is *nested* if for each $k \in 1 \ldots h - 1$ it contains a valid (but not maximal) cluster of size k. For instance, in the example above, the cluster of size four was not found precisely because it did not contain a max-gap cluster of size exactly three.

How does this nested subset differ from general max-gap clusters (the set of all clusters that satisfy Definition 4.1)? First, the requirement that a max-gap cluster be nested implicitly introduces some constraints on gene order within the cluster. The general max-gap cluster definition allows gene order to be arbitrary—the order of genes in the cluster is irrelevant as long as the gaps between genes remain small. Gene order in a nested max-gap cluster, on the other hand, may become disordered only to a limited degree. Second, unlike general max-gap clusters, the set of nested max-gap clusters is not guaranteed to be disjoint; i.e., two maximal nested max-gap clusters may both contain the same homolog. For example, if $g = 0$, the genomes $G_1 = \texttt{12345}$ and $G_2 = \texttt{21354}$ contain two nested clusters of size three: {1,2,3} and {3,4,5}. These two clusters both contain gene 3, yet cannot be merged because {1,2,3,4,5} is not a nested cluster (it does not contain any max-gap cluster of size four). We will see in Section 5 that these two properties have implications for the computation and use of statistics as well as the design of algorithms.

5. WHOLE GENOME COMPARISON STATISTICS

In this section, we present a statistical model for *general* max-gap clusters identified through whole genome comparison. The probability of observing a complete max-gap cluster containing all m homologous gene pairs when comparing two random genomes of size n is $[P(n, m, g)]^2$, where $P(n, m, g)$, defined in Equation (2), is the probability of observing a complete cluster of m marked genes in one genome. For whole genome comparison, m is the number of shared homologs rather than the number of marked genes. Figure 2 shows how $P(n, m, g)$ varies as m ranges from 2 to n. Note that for whole genome comparison the percentage of homologous genes shared between two closely related genomes may be quite high. Thus, squaring the probabilities in Fig. 2 would result in many parameter values for which the probability of a complete cluster will approach one.

To understand this, first consider the simpler case in which the gene sets are identical; e.g., $m = n$. In this case $P(n, m, g)$ equals one: under a simple model of identical gene content, there will always be a max-gap cluster of size n, since a window that spans the entire genome will contain n genes with no gaps, and the n genes will be identical in both genomes. Even without assuming identical gene content, when $m \cdot g$ is large with respect to n, it will still be possible to create extremely large clusters. Indeed, a complete cluster will be found whenever g is greater than the longest contiguous run of singletons. This observation has implications for the design of statistical tests for such clusters.

The traditional approach in hypothesis testing is to determine the probability under the null hypothesis of obtaining a value of the test statistic that is at least as *extreme* (e.g., less likely) as the observed value.

[2] Bergeron *et al.* (2002) have made similar observations in the context of the development of efficient algorithms for finding max-gap clusters, as opposed to the statistical questions considered here.

Remember that for a marked gene scenario we calculated the probability of observing a cluster of size at least h. That will not work in this case, however, since the probability of observing a cluster by whole genome comparison may actually increase with the size of the cluster. For example, as Fig. 2 shows, the probability of a complete cluster is often greater than 0.5. Whenever this is the case, the probability of observing a cluster of size $m - 1$ must be *less than* 0.5. Thus, a larger cluster is not necessarily less likely to occur by chance, and so is not more "extreme" from a statistical viewpoint. Consequently, for whole-genome comparison, rather than calculate the probability of observing a cluster of size at least h, we determine the probability of a cluster of *exactly* size h. These calculations can then be used to determine the probability of observing a cluster whose size is more extreme than h.

5.1. Enumeration strategy

We calculate the probability of a cluster by counting the permutations of n genes that result in a max-gap cluster of exactly h homologs, then divide that by the total number of permutations possible. One strategy for counting all permutations that contain a shared cluster of size exactly h is to first count the ways of creating a cluster of h homologs and then count the number of ways of judiciously placing the remaining $m - h$ homologs so that they cannot extend the cluster to make it larger. The challenge is to determine which regions are "safe" for these $m - h$ *outer* genes.

To answer this question, it can be useful to think about a cluster shared between two genomes in a two-dimensional space, such as the dot plots in Fig. 5, where G_1 is on the vertical axis, G_2 is on the horizontal axis, and a cluster is represented by a black rectangle in the center. When $g = 1$, Figure 5(a) contains a maximal cluster of size $h = 3$, namely {124}. For this gap size, which configurations of the remaining three outer genes are "safe," i.e., do not extend the cluster of size three? Clearly the black rectangle defined by the cluster itself is unsafe, as is the dark gray "moat" of width $g + 1$ around its border, since any gene that lies in these regions will increase the size of the cluster beyond $h = 3$. Suppose outer genes are prohibited from the black and dark gray regions, but allowed in the white and light gray regions. At first, this may seem like the appropriate constraint. The arrangement of outer genes in Figs. 5(a) and 5(b) is considered safe in both cases, correctly classifying both as configurations containing a cluster of size three. However, in Fig. 5(c), this rule fails. Although neither genes 5 nor 6 can independently extend the cluster (since each is further than g away from the cluster on one of the genomes), together they successfully extend the cluster of size three to a cluster of size five ({12456}) (shown in Fig. 5(d)).

Although the light gray region may sometimes be unsafe, ruling it out entirely will cause us to exclude some configurations that contain a maximal cluster of size h. Indeed, since max-gap clusters are not nested, it is difficult to exactly specify the unsafe regions so that we count all permutations containing a maximal cluster of exactly size h, while at the same time *not* counting those permutations which contain only a nonmaximal cluster of size h. Instead, we use the above intuition to devise upper and lower bounds for the probability of observing a cluster of size exactly h, when conducting a whole-genome comparison. The upper bound is too lenient and will erroneously count the configuration of Fig. 5(c) as one containing a maximal cluster of size h. The lower bound is too strict. Although both Figs. 5(a) and 5(b) contain a maximal cluster of size h, the lower bound will only enumerate configurations such as Fig. 5(a). Configurations such as Fig. 5(b) will be unnecessarily excluded.

5.2. Upper bound for incomplete clusters

Our upper bound counts the number of configurations of the m homologs on both genomes which satisfy the following criteria: there exist h homologs that form a chain on both chromosomes, and there does not exist any other homolog that is within a distance g of the chain on *both* genomes. The sole constraint is that an outer gene is permitted within a distance g of a cluster in G_1 only if its homolog is located at least $g + 1$ genes from the cluster on G_2 (like genes 3 and 6 in Fig. 5(a)). This strategy is guaranteed to count all permutations that contain a maximal cluster of size h, but because of its limited lookahead (as discussed in Section 4) it will also incorrectly count some permutations which contain a cluster of size h, but for which that cluster is not maximal (such as the cluster of size three in Figure 5(c)). Thus, enumerating all configurations that satisfy this constraint provides an upper bound on the probability of observing a maximal cluster of h genes.

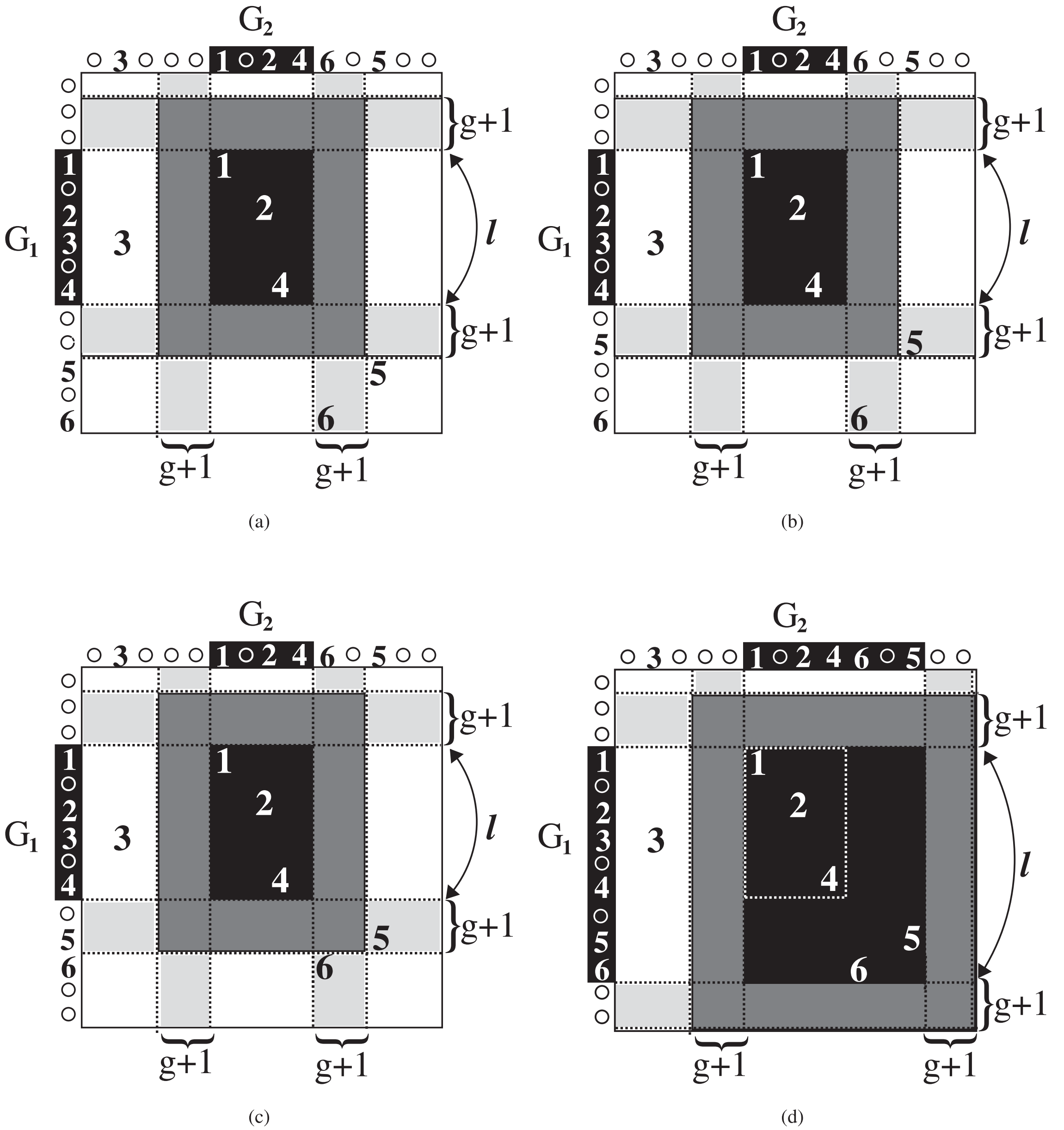


FIG. 5. Dot plots comparing two genomes—G_1 on the vertical and G_2 on the horizontal axis—that share $m = 6$ homologous gene pairs. Singletons are drawn on the axes as circles, but not shown in the dot plot. (**a**) and (**b**): A maximal cluster of size three. Genes 5 and 6 are placed so the cluster cannot be extended. (**c**) A non-maximal cluster of size three. The placement of genes 5 and 6 allows the cluster to be extended, resulting in (**d**) a larger cluster of size five. In all figures, the region of the cluster is shown in black, the dark gray area is unsafe, and the white area is safe. The light gray area is sometimes safe and sometimes unsafe.

Let M be the set of all m homologs shared between the genomes. We divide the set M into three subsets:

1. $H \subset M$ is the set of h homologs that form a chain in both genomes.
2. $T \subset M - H$ is the (possibly empty) set of t homologs that are located within a distance g from the cluster on G_1 but not G_2.
3. $R = M - H - T$ is the set of $r = m - h - t$ genes that are *not* within a distance g from the cluster on G_1.

The first subset comprises the h genes in the cluster, and the last two subsets together comprise the $m - h$ outer genes. Note that, although T and R are defined asymmetrically with respect to genome, we assume equal genome sizes, so the probability will be the same regardless of whether we restrict the position of genes with respect to G_1 or G_2.

The upper bound is the number of ways of placing these three subsets of genes on both genomes for all choices of t so that the constraints on each subset are satisfied, divided by the total number of ways to place the m homologs when their positions are unconstrained. To compute the upper bound on the probability of observing a cluster of size h, we must sum over all possible values of t, which yields

$$P_{up} = \sum_{t=0}^{m-h} \frac{h!t!r!}{m!} \cdot q_1[n, h, t, r, g] \cdot q_2[n, h, t, r, g], \tag{6}$$

where q_1 is the probability of "safely" placing the genes (according to the constraints on each set) in G_1 and q_2 is the probability of "safely" placing the genes in G_2. The factorials account for the different number of ways of ordering the genes within each subset (H, T, and R) versus the unrestricted case in which all m homologs can be permuted indistinguishably.

To compute q_1 and q_2, we present a general equation parameterized by the quantities X and Y, which represent the number of "safe" positions for the genes in T and R, respectively, and are set to different expressions depending on specific location constraints. We enumerate over all possible values of l, where l is the length of the cluster and ranges from h to $L = \min(n, (h-1)g + h)$. Thus, the general form of q_i is

$$q_i = \frac{1}{\binom{n}{m}} \sum_{l=h}^{L} \left[\max(0, n - l - 2g - 1) d_2(h, g, l) \binom{X(i)}{t} \binom{Y(i)}{r} + E \right], \tag{7}$$

where X is the number of legal positions for the genes in the set T, Y is the number of legal positions for the genes in the set R, and the last term handles edge affects. The expression E accounts for those clusters within a distance $j \leq g$ of either end of the genome. Generally, there are two possible clusters of length l within i of either end of the genome: one near the beginning of the genome and one near the end. However, when $l \geq n - j - g$, the cluster spans almost the entire genome and will be simultaneously close to both ends, so there is only one possible cluster. Putting these together yields the following expression:

$$E = \sum_{j=0}^{\min(g, n-l)} d_2(h, g, l) \binom{W(i)}{t} \binom{Z(i)}{r} \cdot \begin{cases} 1 & \text{if } l \geq n - j - g, \\ 2 & \text{otherwise.} \end{cases}$$

The values of X, Y, W, and Z needed to compute q_1 and q_2 are given in Table 1.

TABLE 1. THE VALUES OF q_i, FOR $i = 1..4$[a]

i	*Case*	$X(i)$	$Y(i)$	$W(i)$	$Z(i)$
1	q_1	$b - h$	$n - b$	$c - h$	$n - c$
2	q_2	$n - b$	$n - h - t$	$n - c$	$n - h - t$
3	q_3	$l - h$	$n - b$	$l - h$	$n - c$
4	q_4	$n - b$	$n - h - t - b + l$	$n - c$	$n - h - t - c + l$

[a] The length of the cluster plus its bounding moats (as shown in Fig. 5) is given by $b = l + 2(g + 1)$. Note that $W(i)$ and $Z(i)$, which account for edge effects, are identical to $X(i)$ and $Y(i)$, respectively, except that every instance of b is replaced by $c = \min(n, l + j + g + 1)$, which is the size of the cluster plus its bounding moats when the cluster is located within $j \leq g$ genes of either edge of the genome.

5.3. Lower bound for incomplete clusters

A similar approach can be used to calculate a lower bound on the probability of observing a maximal cluster of size h, for all $h > \frac{m}{2}$. To compute the upper bound, an outer gene was allowed to be within a distance g of the chain on G_1 *or* G_2 but not *both*. However, as explained previously, this constraint on the location of the outer genes is not sufficient to guarantee that the cluster is maximal. For example, both genes 5 and 6 in Fig. 5(c) are individually "safe," but together they extend the cluster. Consequently, the constraint is not strict enough and thus yields an upper bound rather than the exact probability.

To compute the lower bound, we strengthen the constraint on the placement of the outer genes. Let I_1 represent the set of configurations for which the genes in H form a max-gap cluster and no outer gene (a gene in $M - H$) is located within a distance g of the cluster on G_1, regardless of where it is located in G_2. Let I_2 represent those configurations for which the genes in H form a max-gap cluster and no outer gene is located within a distance g of the cluster on G_2. Both constraints are unnecessarily restrictive, but will guarantee that the outer genes do not extend the cluster and that the genes in H form a maximal cluster.

Our lower bound is the union of these two cases: $P_{low} = P(I_1 \cup I_2) = P(I_1) + P(I_2) - P(I_1 \cap I_2)$. Assuming equal genome sizes, I_1 and I_2 are symmetric, and consequently $P(I_1) = P(I_2)$. In this case, the lower bound can be written as $P_{low} = 2P(I_1) - P(I_1 \cap I_2)$.

$P(I_1)$ can be computed by Equation (6), replacing q_1 with q_3 (specified in Table 1). The intersection of the two constraints $P(I_1 \cap I_2)$ is the probability of observing a configuration in which no outside gene is within a distance g of the cluster in *either* genome. This probability can also be computed from Equation (6), except we again replace q_1 with q_3, and q_2 is replaced by q_4 (specified in Table 1). Combining these two applications of Equation (6) yields a lower bound on the probability of observing a cluster of exactly size h:

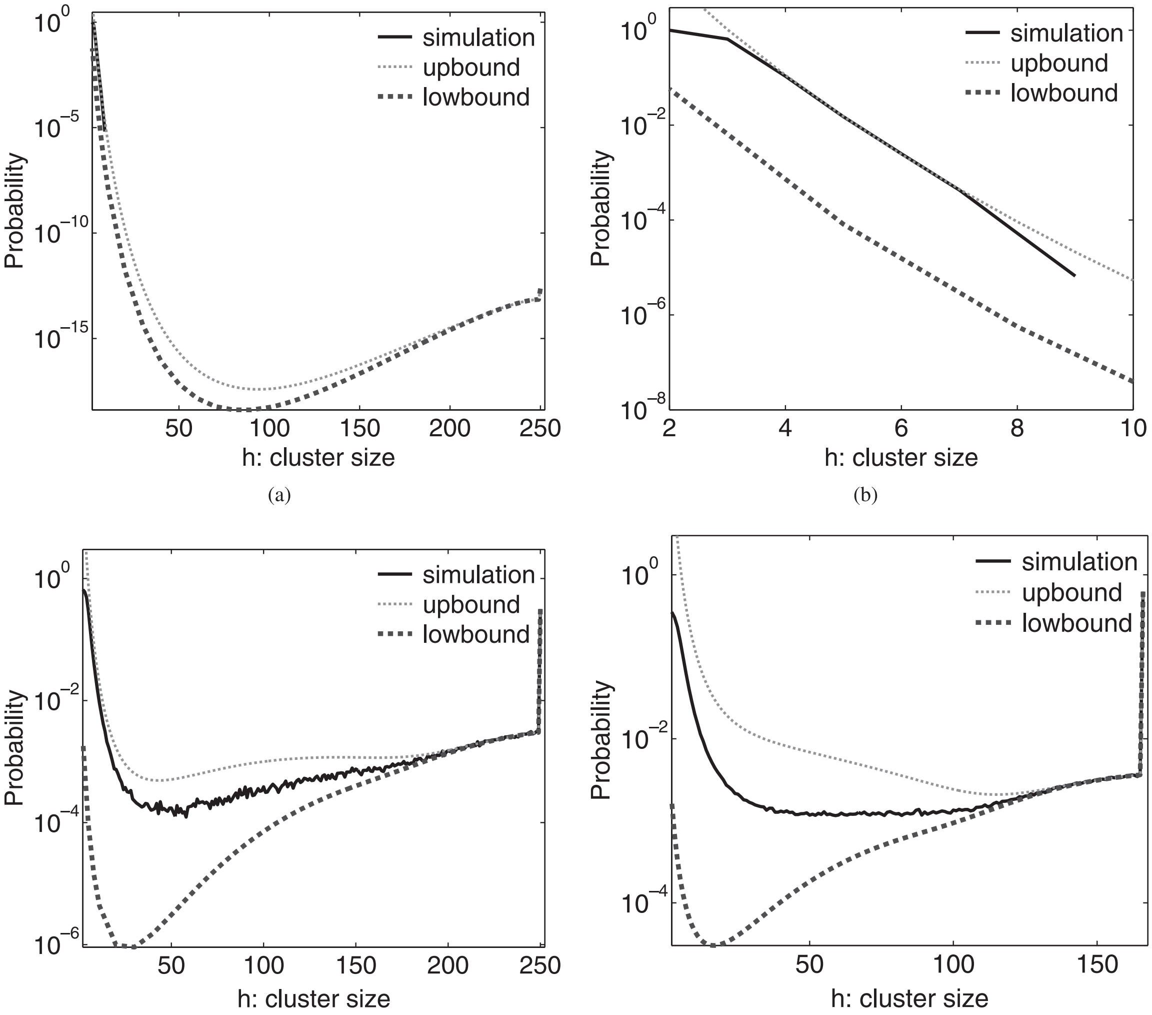
$$P_{low} = \sum_{t=0}^{m-h} \frac{h!t!r!}{m!} \cdot q_2[n,h,t,r,g] \cdot (2 \cdot q_3[n,h,t,r,g] - q_4[n,h,t,r,g]) \,. \tag{8}$$

When $h > m/2$, Equation (8) is a strict lower bound on the probability of observing a cluster of size h. However, when $h \leq m/2$ a permutation may contain more than one shared cluster of size h. The strategy described above enumerates clusters according to their position in the genome, so a permutation containing two clusters of size h at different locations will be double-counted. As h decreases, the percent of random genomes that share multiple clusters will increase, and the probability will be correspondingly overestimated. Thus, for small values of h, it is possible that the probability computed by Equation (8) will actually exceed the true probability. For all parameter values tested in Section 5.4, however, the probabilities obtained by Equation (8) were always lower than those estimated by sampling.

5.4. Experiments

In order to investigate the accuracy of the bounds in different regions of the parameter space, we compared them to the probability of observing max-gap clusters in randomly permuted genomes, estimated through simulation. A number of different parameter values and genome sizes were analyzed. For each set of parameter values, we generated one million random permutations of two genomes and used the Gene Teams software (Bergeron *et al.*, 2002) to find all maximal max-gap clusters. In Fig. 6, the upper bound (dashed line) and lower bound (dotted line) are compared to the probabilities estimated from the simulations (solid line). Notice that in Figs. 6(a) and 6(b) the simulated probabilities are shown only for $h < 10$ since only one million random trials were generated and that is the cluster size at which the probabilities drop below 10^{-6}.

First, we considered how the ratio of gap size to genome size affects the accuracy of the bound. Our experiments suggest that when the maximum gap size is small with respect to n (about 1%), the upper bound is extremely accurate for all values of h (as illustrated in Fig. 6(b)). However, when the maximum gap size is larger with respect to n (2% or 3%), then the bounds are exact only when estimating the probability of a large or complete max-gap cluster. This is illustrated in Fig. 6(d), which shows the behavior of the bounds when $n = 500$, $m = 166$, and $g = 15$. For these parameter values, the bounds are extremely



(a)

(b)

(c)

(d)

FIG. 6. The probability of observing at least one max-gap cluster of size h when (**a**) $n = 1{,}000$, $m = 250$, and $g = 10$, (**b**) detail of Fig. 6(a) for $h \leq 10$, (**c**) $n = 1{,}000$, $m = 250$, and $g = 20$, and (**d**) $n = 500$, $m = 166$, and $g = 15$. Solid lines indicate simulations results, dashed lines the upper bound, and dotted lines the lower bound.

accurate for large values of h, but begin to diverge significantly as h drops below 100. To what extent does the divergence of the upper bound effect the conclusions we may draw about cluster significance? At a significance level of $\alpha = 0.01$, for example, the error in the upper bound would lead to the unnecessary rejection of significant clusters of size 8 to 15. At a significance level of $\alpha = 0.001$, however, the upper bound could be used to correctly determine that no matter how large the cluster size, the null hypothesis cannot be rejected.

In addition to accuracy, we also considered the monotonicity of the probabilities with respect to cluster size. Our analysis shows that, under a null hypothesis of random gene order, the probabilities of observing a max-gap cluster are not always monotonic with respect to cluster size, but often decrease initially and then increase as h approaches m. For example, when $n = 1{,}000$, $m = 250$, and $g = 20$, Fig. 6(c) shows that the chance probability of observing a cluster of 50 genes is actually smaller than the chance probability of observing a cluster of 100 genes. This nonmonotonic behavior can be understood intuitively by observing that, as the size of the cluster increases, the max-gap criteria implicitly increases the maximum length of the cluster as well. As a result, as the size of the cluster sought increases, the probability of observing such a cluster may grow substantially.

Table 2. Number of Max-Gap Clusters of Varying Sizes Shared between *E. coli* and *B. subtilis* for a Range of Gap Values

	Cluster size				
Gap	*2–3*	*4–10*	*11–26*	*27–60*	*> 60*
1	108	21	1	0	0
5	112	26	1	0	0
15	144	32	2	0	0
50	165	50	6	2	1
100	0	0	0	0	2

In order to demonstrate the utility of our statistical tests, we conducted a whole genome comparison of the *E. coli* and *B. subtilis* genomes. A mapping of orthologs between the two genomes was obtained from the GOLDIE database (Bansal, 1999). The *E. coli* genome has $n = 4{,}108$ known genes and the *B. subtilis* genome has $n = 4{,}245$ known genes. After eliminating ambiguous orthologs, the map yields $m = 1{,}315$ homologous pairs. Using the Gene Teams software (Bergeron *et al.*, 2002), we identified all max-gap clusters shared between the two genomes for values of g ranging from 0 to 125. When $g = 110$, all homologs formed one complete cluster.

A subset of the results selected to illustrate the general trends is shown in Table 2. In addition, Fig. 7 shows the sizes of the clusters found with a range of different gap sizes. Our results fall into three regimes. When $g = 0..40$, cluster sizes range from 2 to 12, except for one larger cluster of size 20 to 30. When $g = 40..70$, clusters sizes grow from 2 to about 600. Finally, for gap sizes of $g \geq 70$, the homologs form only one or two large clusters.

The upper bound on the probability of observing at least one cluster of size h under the null hypothesis was calculated from Equation (6) for $n = 4{,}108$, $m = 1{,}315$, and a range of values of h and g. These probabilities are indicated by the dashed lines in Fig. 8. To assess the accuracy of our upper bound for this bacterial dataset, we again compared it with estimates of the probability of observing max-gap clusters in randomly permuted genomes of the same size, obtained through simulation. We generated one million random permutations of two genomes with $n = 4{,}108$ genes and $m = 1{,}315$ homologs and again used the Gene Teams software (Bergeron *et al.*, 2002) to find all maximal max-gap clusters with gap sizes ranging from $g = 0$ to $g = 100$. Figure 8 compares our upper bound (dashed lines) with the probabilities estimated from simulations (solid lines). The accuracy of the bound depends on both h and g. The bound

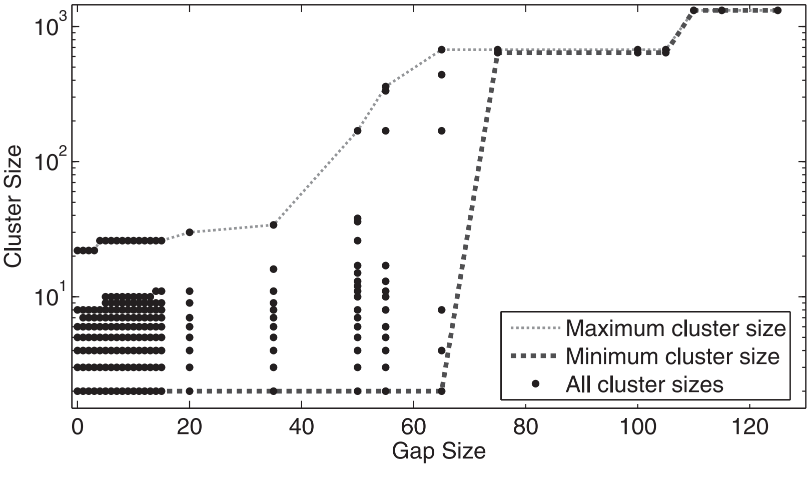


FIG. 7. The distribution of observed cluster sizes between *E. coli* and *B. subtilis* for g ranging from 0 to 125. The dashed line indicates the largest cluster found and the dotted line indicates the smallest cluster found for each value of g.

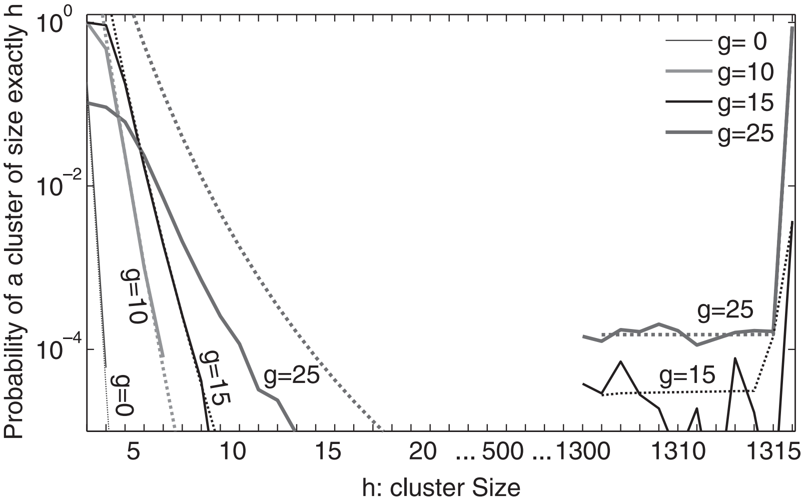


FIG. 8. Probability of observing max-gap clusters of size h in the *E. coli* and *B. subtilis* genome comparison for $g = 0$ to $g = 25$. Solid lines show probabilities estimated via simulation and dashed lines indicate the upper bound on the probabilities, calculated using Equation (6).

is extremely accurate when g is between one and fifteen, but as g becomes larger the bound diverges from the estimated probabilities for small values of h. However, as h approaches m, the bound provides a very accurate estimate of the probability even for large g. Note that although one million random permutations were carried out to estimate the cluster probabilities, clusters of size $20 \leq h \leq 1{,}314$ occurred only infrequently in randomly ordered genomes, and thus for $g = 15$ the probability estimates from randomized genomes are still noisy in this region. Although the upper bound appears to drop below the simulated probability for $h = 1{,}312$ and $h = 1{,}306$, this is due to the fact that one million iterates are insufficient to obtain a precise probability estimate in this region of the parameter space.

Since the upper bound is highly accurate for $0 \leq g \leq 15$, it can be used to evaluate the significance of clusters detected through whole genome comparison. If we consider a significance threshold of 0.001, then Fig. 8 shows that clusters of size three and larger are unlikely to be observed given random gene order when $g = 0$. When $g = 15$, however, only clusters of size seven or larger appear to be significant. Using these statistics, we find a total of 128 homologs in some significant cluster when $g = 0$, whereas 191 homologs are in a significant cluster when $g = 1$ and only 82 are in a significant cluster when $g = 15$. This suggests that using a gap value of $g = 1$ provides more discriminatory power than either $g = 0$ or $g = 15$ for this dataset.

For $g \leq 40$, most max-gap clusters contain 2 to 10 genes, which corresponds to the range of sizes for typical operons (Zheng *et al.*, 2002). We compared the clusters to the RegulonDB database of experimentally determined operons in *E. coli* (Salgado *et al.*, 2004) and verified that for gap sizes of 0 to 10, over 90% of the clusters are comprised entirely of genes from a single operon. The single large cluster of over 20 genes is composed entirely of ribosomal proteins, which together form the ribosomal "super-operon" in *E. coli.*

An intriguing observation is that the number of large clusters seems to be fewer than expected under the null model. When $g \geq 25$, the null model predicts that the probability that all genes will form a complete cluster is close to one. However, a gap size of $g > 100$ is required to obtain a complete max-gap cluster in the bacterial dataset. This discrepancy can be explained by the presence of operons. Since the genes in operons are densely clustered (Chen *et al.*, 2004), the singletons will be clustered more densely as well. These runs of singletons form large gaps and prevent large clusters from occurring as often as they would under a model of random gene order. This is one piece of evidence that the max-gap cluster definition is a good discriminator, since the frequency of both small and large clusters is clearly different than that expected under the null hypothesis, at least for this dataset. In eukaryotes, clusters will generally be due to shared ancestry rather than conserved operons, and so the difference between the observed and predicted cluster sizes remains to be seen.

6. CONCLUSION

We have presented the first rigorous statistical treatment of max-gap clusters, a definition that is frequently used in empirical studies (Blanc *et al.*, 2003; Bourque *et al.*, 2005; Chen *et al.*, 2004; Friedman and Hughes, 2001; Luc *et al.*, 2003; McLysaght *et al.*, 2002; Overbeek *et al.*, 1999; Simillion *et al.*, 2002; Tamames, 2001; Vandepoele *et al.*, 2002; Vision *et al.*, 2000). Analytical statistical models such as ours are useful for analyzing specific datasets, especially when only partial data is available, as well as choosing parameter values and evaluating the discriminatory power of different cluster definitions. We provide exact statistical tests for a marked gene scenario and an upper and lower bound on the probability of observing a cluster of size h through a pairwise whole genome comparison. Finally, we demonstrated the utility of our statistical model in a comparison of the *E. coli* and *B. subtilis* genomes, and confirm that for this dataset the max-gap cluster definition provides excellent discriminatory power.

The results presented here suggest a number of other interesting directions for future work. For example, the statistics presented in Section 4 are based on *general* max-gap clusters identified through whole genome comparison, but are not applicable to clusters found with a greedy heuristic or for studies in which only nested clusters are of interest. In particular, since nested max-gap clusters are a subset of general max-gap clusters, we expect to find fewer nested clusters than general under the null hypothesis. This is especially true for large clusters. In addition, the enumeration strategy we use to derive the statistics below relies on the fact that max-gap clusters are disjoint and that gene order is irrelevant. Thus, statistics for nested max-gap clusters remain an open problem.

The model considered here treats the genome as an ordered set of genes, disregarding actual distances between genes. This assumption can be advantageous because physical distances between genes often differ substantially across genomes. Furthermore, it eliminates the need to model the variation in gene density that can lead to gene-rich and gene-poor regions of chromosomes. A distance-based model would have to take into account the fact that a cluster that is surprising in a gene-poor region might easily occur by chance in a gene rich region. However, since prokaryotic genomes tend to be gene dense, it would not be difficult to modify the model used here to explicitly consider physical distances for bacteria. When analyzing clusters in bacterial genomes, statistical models that take into account the orientation of genes and the possibility of circular instead of linear chromosomes are also of interest.

The current model also disregards the presence of tandem duplications and gene families. Since tandem duplications can be detected easily in genomic data due to their regular spatial patterns, they can be taken into account by a preprocessing step in genomic analysis. Gene families are more problematic, however. Virtually all genomes contain gene families, sets of genes with similar sequence and function, that arose through duplication of genetic material. Large gene families will increase the likelihood of observing a conserved cluster by chance and, hence, can have a substantial impact on the statistical significance of a particular cluster. Unfortunately, factoring gene families into an analytical statistical model is difficult because the exact size of each gene family in a genome cannot be easily determined.

Finally, this work assesses the significance of spatial clusters of genes with respect to a null hypothesis of random gene order. When searching for genomic regions that share a recent common ancestor (either due to a duplication or speciation event), this null hypothesis is adequate. However, in order to distinguish significant gene clusters that are merely due to a shared common ancestor from sets of genes that remain clustered due to functional selection, the phylogenetic distance between the organisms will have to be incorporated into the null hypothesis. This remains as future work.

In our analysis of the statistical properties of spatially clustered genes under the maximum gap criterion, we have shown that the precise formalization of the max-gap cluster definition has surprising implications both for designing algorithms to find clusters and testing cluster significance. Most groups that choose a max-gap cluster definition explicitly state that they do not consider the order of genes within the cluster but only the distance between them. However, constructing max-gap clusters with a naive agglomerative algorithm implicitly introduces constraints on gene order within the cluster and thus will find only a subset of all max-gap clusters. In addition, our statistical analysis demonstrates that when a whole genome comparison is being conducted, the chance probability of observing a max-gap cluster is not monotonic with respect to cluster size. Although there is a widespread belief that cluster significance grows with the number of homologs in the cluster, it is critical to recognize that if only a max-gap criterion is used without additional constraints, larger clusters do not always imply greater significance.

APPENDIX: SIMPLIFIED EXPRESSION FOR $d_0(k, g, w_{kg} - 1)$

The following three lemmas are needed to obtain a closed-form expression for $d_0(k, g, w_{kg} - 1)$.

Lemma A.1. *For all x such that $k \le x \le w_{kg}$, $d_2(k, g, x) = d_2(k, g, w_{kg} + k - x)$.*

Proof. Let $S(k, g, x)$ be the set of max-gap clusters of size k and length x, with no gap greater than g. Clearly, $|S(k, g, x)| = d_2(k, g, x)$. Let $\langle x_1, \ldots, x_{k-1}\rangle$, where $0 \le x_i \le g$, denote a member of this set, i.e., a max-gap cluster of size k and length $x = k + \sum_{i=1}^{k-1} x_i$, with gap sizes $x_1, \ldots, x_{k-1}$. Define a function $f(\langle x_1, \ldots, x_{k-1}\rangle) = \langle y_1, \ldots, y_{k-1}\rangle$, where $y_i = g - x_i$. We claim f maps $S(k, g, x)$ to $S(k, g, w_{kg} + k - x)$. To see this, observe that $0 \le y_i \le g$ and the length of the cluster $\langle y_1, \ldots, y_{k-1}\rangle$ is

$$k + \sum_{i=1}^{k-1} y_i = k + \sum_{i=1}^{k-1} g - x_i = k + (k-1)g - \sum_{i=1}^{k-1} x_i$$

$$= k + (k-1)g - (x - k) = w_{kg} + k - x.$$

Since f is a bijection, $|S(k, g, x)| = |S(k, g, y)|$. ■

Lemma A.2. *For all x such that $k \le x \le w_{kg}$, $d_1(k, g, x) + d_1(k, g, w_{kg} + k - x - 1) = d_1(k, g, w_{kg})$.*

Proof. By definition,

$$d_1(k, g, x) + d_1(k, g, w_{kg} + k - x - 1) = \sum_{i=k}^{x} d_2(k, g, i) + \sum_{j=k}^{w_{kg}+k-x-1} d_2(k, g, j).$$

Lemma A.1 can be used to simplify the second term, yielding

$$\sum_{i=k}^{x} d_2(k, g, i) + \sum_{j=w_{kg}}^{x+1} d_2(k, g, j) = \sum_{j=k}^{w_{kg}} d_2(k, g, j) = d_1(k, g, w_{kg}). \quad ■$$

Lemma A.3. *If $(w_{kg} + k - 1)$ is even, then $2d_1(k, g, \frac{1}{2}(w_{kg} + k - 1)) = d_1(k, g, w_{kg})$.*

Proof.

$$d_1(k, g, w_{kg}) = \sum_{i=k}^{w_{kg}} d_2(k, g, i) = \sum_{i=k}^{\frac{1}{2}(w_{kg}+k-1)} d_2(k, g, i) + \sum_{j=\frac{1}{2}(w_{kg}+k+1)}^{w_{kg}} d_2(k, g, j).$$

Lemma A.1 can be used to simplify the second term, yielding

$$\sum_{i=k}^{\frac{1}{2}(w_{kg}+k-1)} d_2(k, g, i) + \sum_{j=\frac{1}{2}(w_{kg}+k-1)}^{k} d_2(k, g, j)$$

$$= \sum_{i=k}^{\frac{1}{2}(w_{kg}+k-1)} 2d_2(k, g, i) = 2d_1(k, g, \frac{1}{2}(w_{kg} + k - 1)) \quad ■$$

Theorem A.4.

$$d_0(k, g, w_{kg} - 1) = \frac{w_{kg} - k}{2}(g+1)^{k-1}.$$

Proof. Either $w_{kg} + k$ is even or it is odd. When it is even

$$d_0(k, g, w_{kg} - 1) = \sum_{i=k}^{w_{kg}-1} d_1(k, g, i) = \sum_{i=k}^{\frac{1}{2}(w_{kg}+k)-1} d_1(k, g, i) + \sum_{j=\frac{1}{2}(w_{kg}+k)}^{w_{kg}-1} d_1(k, g, j).$$

Rewriting the summation index on the second term yields

$$\sum_{i=k}^{\frac{1}{2}(w_{kg}+k)-1} d_1(k, g, i) + \sum_{j=\frac{1}{2}(w_{kg}+k)-1}^{k} d_1(k, g, w_{kg} + k - j - 1)$$

$$= \sum_{i=k}^{\frac{1}{2}(w_{kg}+k)-1} d_1(k, g, i) + d_1(k, g, w_{kg} + k - i - 1).$$

By Lemma A.2, this simplifies to

$$\sum_{i=k}^{\frac{1}{2}(w_{kg}+k)-1} d_1(k, g, w_{kg}) = \frac{w_{kg} - k}{2}(g+1)^{k-1},$$

as desired. Otherwise, if $w_{kg} + k$ is odd, then

$$d_0(k, g, w_{kg} - 1) = \sum_{i=k}^{\frac{1}{2}(w_{kg}+k-3)} d_1(k, g, i) + d_1(k, g, \frac{1}{2}(w_{kg} + k - 1)) + \sum_{j=\frac{1}{2}(w_{kg}+k+1)}^{w_{kg}-1} d_1(k, g, j).$$

The second term can be simplified by Lemma A.3, yielding

$$\frac{1}{2}d_1(k, g, w_{kg}) + \sum_{i=k}^{\frac{1}{2}(w_{kg}+k-3)} d_1(k, g, i) + \sum_{j=\frac{1}{2}(w_{kg}+k+1)}^{w_{kg}-1} d_1(k, g, j).$$

As for the even case, the last two terms can be combined and simplified by Lemma A.2:

$$\frac{1}{2}d_1(k, g, w_{kg}) + \frac{(w_{kg} - k - 1)}{2}d_1(k, g, w_{kg}) = \frac{w_{kg} - k}{2}(g+1)^{k-1},$$

as desired. ■

ACKNOWLEDGMENTS

D.D. was supported by NIH grant 1 K22 HG 02451-01 and a David and Lucille Packard Foundation fellowship. D.S. was funded by NSERC Canada. He holds the Canada Research Chair in Mathematical Genomics and is a Fellow of the Evolutionary Biology Program of the Canadian Institute for Advanced Research. R.H. was supported in part by a Barbara Lazarus Women@IT Fellowship, funded in part by the Alfred P. Sloan Foundation. We thank B. Vernot, A. Goldman, and N. Song for helpful comments and software support.

REFERENCES

Abi-Rached, L., Gilles, A., Shiina, T., Pontarotti, P., and Inoko, H. 2002. Evidence of en bloc duplication in vertebrate genomes. *Nat. Genet.* 31, 100–105.

Amores, A., Force, A., Yan, Y., Joly, L., Amemiya, C., Fritz, A., Ho, R., Langeland, J., Prince, V., Wang, Y.L., Westerfield, M., Ekker, M., and Postlethwait, J.H. 1998. Zebrafish hox clusters and vertebrate genome evolution. *Science* 282, 1711–1714.

Bansal, A.K. 1999. An automated comparative analysis of 17 complete microbial genomes. *Bioinformatics* 15, 900–908. *www.cs.kent.edu/∼arvind/orthos.html*.

Bergeron, A., Corteel, S., and Raffinot, M. 2002. The algorithmic of gene teams. *WABI*, *Lecture Notes in Computer Science* 2452, 464–476. *www-igm.univ-mlv.fr/∼raffinot/geneteam.html*.

Blanc, G., Hokamp, K., and Wolfe, K.H. 2003. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res.* 13, 137–144.

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

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

Calabrese, P.P., Chakravarty, S., and Vision, T.J. 2003. Fast identification and statistical evaluation of segmental homologies in comparative maps. *ISMB (Supplement of Bioinformatics)* 74–80.

Cannon, S.B., Kozik, A., Chan, B., Michelmore, R., and Young, N.D. 2003. DiagHunter and GenoPix2D: Programs for genomic comparisons, large-scale homology discovery and visualization. *Genome Biol.* 4, R68.

Chen, X., Su, Z., Dam, P., Palenik, B., Xu, Y., and Jiang, T. 2004. Operon prediction by comparative genomics: An application to the *Synechococcus* sp. WH8102 genome. *Nucl. Acids Res.* 32, 2147–2157.

Cosner, M.E., Jansen, R.K., Moret, B.M.E., Raubeson, L.A., Wang, L.-S., Warnow, T., and Wyman, S. 2000. An empirical comparison of phylogenetic methods on chloroplast gene order data in *Campanulaceae*, *in* Sankoff, D., and Nadeau, J.H., eds., *Comparative Genomics*, 99–121, Kluwer Academic Press, Dordrecht, NL.

Coulier, F., Pontarotti, P., Roubin, R., Hartung, H., Goldfarb, M., and Birnbaum, D., 1997. Of worms and men: An evolutionary perspective on the fibroblast growth factor (FGF) and FGF receptor families. *J. Mol. Evol.* 44, 43–56.

Danchin, E.G.J., Abi-Rached, L., Gilles, A., and Pontarotti, P. 2003. Conservation of the MHC-like region throughout evolution. *Immunogenetics* 55, 141–148.

Durand, D., and Sankoff, D. 2003. Tests for gene clustering. *J. Comp. Biol.* 453–482.

Endo, T., Imanishi, T., Gojobori, T., and Inoko, H. 1997. Evolutionary significance of intra-genome duplications on human chromosomes. *Gene* 205, 19–27.

Friedman, R., and Hughes, A.L. 2001. Gene duplication and the structure of eukaryotic genomes. *Genome Res.* 11, 373–381.

Gibson, T., and Spring, J. 2000. Evidence in favour of ancient octaploidy in the vertebrate genome. *Biochem. Soc. Trans.* 2, 259–264.

Graham, R.L., Knuth, D.E., and Patashnik, O. 1989. *Concrete Mathematics*, Addison-Wesley, Reading, MA.

Hampson, S.E., Gaut, B.S., and Baldi, P. 2005. Statistical detection of chromosomal homology using shared-gene density alone. *Bioinformatics* 21, 1339–1348.

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

Hoberman, R., Sankoff, D., and Durand, D. 2005. The statistical significance of max-gap clusters. *Proc. RECOMB Satellite Workshop on Comparative Genomics*, *Lecture Notes in Bioinformatics*, Springer Verlag. Editor: J. Lagergran, vol. 3388, pgs. 55–71.

Hokamp, K. 2001. *A Bioinformatics Approach to (Intra-)Genome Comparisons*. Ph.D Thesis, University of Dublin, Trinity College.

Hughes, A.L. 1998. Phylogenetic tests of the hypothesis of block duplication of homologous genes on human chromosomes 6, 9, and 1. *Mol. Biol. Evol.* 15, 854–870.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.

Kasahara, M. 1997. New insights into the genomic organization and origin of the major histocompatibility complex: Role of chromosomal (genome) duplication in the emergence of the adaptive immune system. *Hereditas* 127, 59–65.

Katsanis, N., Fitzgibbon, J., and Fisher, E. 1996. Paralogy mapping: Identification of a region in the human MHC triplicated onto human chromosomes 1 and 9 allows the prediction and isolation of novel PBX and NOTCH loci. *Genomics* 35, 101–108.

Lawrence, J., and Roth, J.R. 1996. Selfish operons: Horizontal transfer may drive the evolution of gene clusters. *Genetics* 143, 1843–1860.

Lipovich, L., Lynch, E.D., Lee, M.K., and King, M.-C. 2001. A novel sodium bicarbonate cotransporter-like gene in an ancient duplicated region: *SLC4A9* at 5q31. *Genome Biol.* 2, 0011.1–0011.13.

Luc, N., Risler, J., Bergeron, A., and Raffinot, M., 2003. Gene teams: A new formalization of gene clusters for comparative genomics. *Comp. Biol. Chem.* 27, 59–67.

McLysaght, A., Hokamp, K., and Wolfe, K.H. 2002. Extensive genomic duplication during early chordate evolution. *Nature Genet.* 31, 200–204.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., and Maltsev, N. 1999. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. USA* 96, 2896–2901.

Pebusque, M.-J., Coulier, F., Birnbaum, D., and Pontarotti, P. 1998. Ancient large-scale genome duplications: Phylogenetic and linkage analyses shed light on chordate genome evolution. *Mol. Biol. Evol.* 15, 1145–1159.

Ruvinsky, I., and Silver, L.M. 1997. Newly indentified paralogous groups on mouse chromosomes 5 and 11 reveal the age of a T-box cluster duplication. *Genomics* 40, 262–266.

Salgado, H., Gama-Castro, S., Martínez-Antonio, A., Díaz-Peredo, E., Sánchez-Solano, F., Peralta-Gil, M., Garcia-Alonso, D., Jiménez-Jacinto, V., Santos-Zavaleta, A., Bonavides-Martínez, C., and Collado-Vides, J. 2004. RegulonDB (version 4.0): Transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucl. Acids Res.* 32, D303–D306.

Sankoff, D. 2003. Rearrangements and chromosomal evolution. *Curr. Opin. Genet. Dev.* 13, 583–587.

Sankoff, D., Bryant, D., Deneault, M., Lang, B.F., and Burger, G. 2000a. Early eukaryote evolution based on mitochondrial gene order breakpoints. *J. Comp. Biol.* 3–4, 521–535.

Sankoff, D., Deneault, M., Bryant, D., Lemieux, C., and Turmel, M. 2000b. Chloroplast gene order and the divergence of plants and algae from the normalized number of induced breakpoints, *in* Sankoff, D., and Nadeau, J.H., eds., *Comparative Genomics*, 89–98, Kluwer Academic Press, Dordrecht, NL.

Sankoff, D., and Nadeau, J.H. 2003. Chromosome rearrangements in evolution: From gene order to genome sequence and back. *Proc. Natl. Acad. Sci. USA* 100, 11188–11189.

Simillion, C., Vandepoele, K., Montagu, M.V., Zabeau, M., and de Peer, Y.V. 2002. The hidden duplication past of *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* 99, 13627–13632.

Skovgaard, M., Jensen, L.J., Brunak, S., Ussery, D., and Krogh, A. 2001. On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet.* 17, 425–428.

Tamames, J. 2001. Evolution of gene order conservation in prokaryotes. *Genome Biol.* 6, 0020.1–0020.11.

Tamames, J., Gonzalez-Moreno, M., Valencia, A., and Vicente, M. 2001. Bringing gene order into bacterial shape. *Trends Genet.* 3, 124–126.

Trachtulec, Z., and Forejt, J. 2001. Synteny of orthologous genes conserved in mammals, snake, fly, nematode, and fission yeast. *Mamm. Genome* 3, 227–231.

Uspensky, J.V. 1937. *Introduction to Mathematical Probability*, 23–24, McGraw-Hill, New York.

Vandepoele, K., Saeys, Y., Simillion, C., Raes, J., and van de Peer, Y. 2002. The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between *Arabidopsis* and rice. *Genome Res.* 12, 1792–1801.

Vision, T.J., Brown, D.G., and Tanksley, S.D. 2000. The origins of genomic duplications in *Arabidopsis. Science* 290, 2114–2117.

Wolfe, K. 2001. Yesterday's polyploids and the mystery of diploidization. *Nature Rev. Genet.* 2, 33–41.

Zheng, Y., Szustakowski, J.D., Fortnow, L., Roberts, R.J., and Kasif, S. 2002. Computational identification of operons in microbial genomes. *Genome Res.* 12, 1221–1230.

Address correspondence to:
Rose Hoberman
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA, 15213

E-mail: roseh@cs.cmu.edu

ADDENDUM

While this paper was in press, three publications with relevance to gene cluster statistics have appeared: He and Goldwasser, *J. Comp. Biol.*, 12(6), 2005; Raghupathy and Durand, 2005 RECOMB Ws. on Comparative Genomics, published as Lecture Notes in *Bioinformatics*, volume 3678; and Hoberman and Durand, ibid. The first of these papers provides an approximation (page 654) to our Equation 2, but does not provide the closed form solution given here precisely because it disregards the edge cases.