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

Rose Hoberman,[*] Computer Science Department, Carnegie Mellon University,
Pittsburgh, PA, 15213, USA. roseh@cs.cmu.edu

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

Dannie Durand, Departments of Biological Sciences and Computer Science,
Carnegie Mellon University, Pittsburgh, PA, USA. durand@cmu.edu

March 25, 2005

[*]phone: (412) 268-3199, fax: (412) 268-7129

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 http://www.cs.cmu.edu/∼durand/Lab/software.html

Keywords: 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 in Wolfe, 2001) reconstructing chromosomal rearrangements (surveyed in Sankoff, 2003; 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 in Chen et al., 2004; Lawrence and Roth, 1996; 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., 2004; 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). 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 of size $h < m$, 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 finding max-gap clusters. In Section 5, we develop upper and lower bounds for the probability of finding a max-gap cluster of h homologs by 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. Figure 1

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 max-gap cluster is a set of marked genes that forms a chain but is not included within any larger cluster (i.e. is maximal).*

For example, when the maximum gap allowed is $g=2$, two max-gap clusters (indicated by boxes) are found in the example genome in Figure 1. The remaining gene cannot be clustered with any other genes. We use the the terms 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 Figure 1.

2.1 Preliminary computations

In the following sections it will be helpful to have an expression for the number of ways of arranging m 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 $m-1$ gaps with a cumulative sum of $l-m$, the number of ways of creating a max-gap cluster of size m and length l is equivalent to the number of ways of of rolling $m-1$ dice, each with faces numbered 0 to g, such that the sum of their faces is equal to $l-m$[1]. We can compute this quantity using the following expression:

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

where the parameter e is set to 2. The number of ways of generating a cluster with length no greater than l is equivalent to requiring that only one endpoint in the window contain a marked gene. This is simply $\sum_{r=m}^{l} d_2(m,g,r)$, which can be shown to be equivalent to $d_1(m,g,l)$. Similarly, the number of ways of arranging m genes so that they form a max-gap cluster anywhere within a window of size l is $\sum_{r=m}^{l} d_1(m,g,r) = d_0(m,g,l)$. Note that an efficient implementation of d_e can be obtained by pre-computing all necessary factorials, allowing the entire summation to be computed in $O(l)$ time.

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

3 Statistics for Marked Genes

In this search scenario, m genes of interest are selected, or marked. 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

Using the above calculations, we can determine the probability of observing a complete max-gap cluster that contains all m genes of interest. Let $w = (m-1)g + m$ be the maximum length of a cluster of size m. Given a random genome of size n, the probability of observing all m marked genes (in any order), such that the gap between adjacent marked genes does not exceed g is

$$P(n, m, g) = \frac{\max(0, n{-}w{+}1)(g{+}1)^{m-1} + d_0(m, g, \min(n, w{-}1)}{\binom{n}{m}}, \tag{2}$$

where $n{-}w{+}1$ is the 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 accounts for edge effects. When $m \ll n$, the total number of permutations can be approximated in constant time using Stirling's approximation. Then the complexity of computing $P(n, m, g)$ is simply $O(w) = O(mg)$. Often $w \ll n$, in which case the running time will be independent of the genome size since the only calculation that is not constant time is computing the number of ways of constructing a max-gap cluster within the last $w - 1$ genes in the genome. When n is large in relation to w we can omit the calculation for edge effects and still estimate the probability accurately.

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 $1/(g{+}1)$. Further 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 finding all m marked genes (in any order) in a max-gap cluster that is completely contained within a window of size r is

$$\binom{n}{m}^{-1}\left[(n-r+1)\; d_1(m,g,r) + d_0(m,g,r-1)\right], \tag{3}$$

where we have replaced $(g+1)^{m-1}$ in Equation (2) with $d_1(m,g,r)$ in order to constrain the maximum length of the cluster. 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, we relax the cluster definition to allow incomplete clusters of size at least h, for $h < m$ (maintaining the requirement that there is no gap greater than g between adjacent marked genes). Our test statistic is the probability of finding a cluster at least as extreme as h; e.g. a cluster of size $\geq h$, assuming random gene order.

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

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

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 non-marked 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 containing no incomplete clusters

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

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 5000 (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 $\geq h$, 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 $d_0(h,g,l)$. The remaining $m - h$ genes must be placed outside the cluster. Note that they also must be placed at a distance of at least $g + 1$ of 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 finding an incomplete cluster of size at least h is

$$\frac{1}{\binom{n}{m}} \sum_{l=h}^{h+g(h-1)} \left[(n-l-g) \cdot d_0(h,g,l) \cdot \binom{n-l-g-1}{m-h} + E \right], \tag{5}$$

where l ranges over all possible lengths of a cluster of size h and E addresses edge effects. E is the number of clusters in which the last marked genes 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. Thus the total number of clusters near the end of the genome is

$$E = \sum_{i=0}^{g} d_0(h,g,l) \cdot \binom{n-l-i}{m-h}.$$

Using the upper summation identity (Graham et al., 1989) this can be further simplified to

$$E = \sum_{i=0}^{g} d_0(h, g, l) \left[\binom{n-l+1}{m-h+1} - \binom{n-l-g}{m-h+1} \right],$$

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 sub-computations 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 three in bacteria (Tamames, 2001) to about thirty in human (McLysaght et al., 2002). We calculated probabilities for genomes sizes of 0.5K, 1K, 5K, 20K, and 25K, corresponding to typical gene sets for bacteria, yeast, worm, and higher eukaryotes like human and Arabidopsis.

3.4.1 Complete clusters

Figure 2

The probability of finding a complete cluster for varying values of n, m, and g was calculated from Equation (2). For complete clusters we tested all values of m ranging from 2 to the genome size n. These calculations will be useful in Section 4 as well, and the region in which m is close to n will be of particular interest.

Figure 2 shows the probability of observing a complete cluster containing all m marked genes in a genome of size $n = 500$, as m ranges from 5 to 250 and g increases from 0 to 50. The probability of finding a complete cluster increases monotonically with g. We might also expect that 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 Figure 2(b) shows, as m increases, the probabilities first decrease and then increase. When m is small, a small increase in the number of black genes will actually decrease the probability of finding a cluster. This makes sense intuitively if ones considers the extreme cases: when $m = 1$ or $m = n$ the probability of finding 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 show that as n increases the probabilities decrease but the general trends seen in Figure 2(b) remain the same. Figure 3

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 500 is no more than 0.0001. 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

Figure 4

We calculated the probability of finding an incomplete cluster from Equations (4) and (5) for the values of n and g as stated above. We chose to examine values of m ranging from 3 to 250, which covers the range of gene numbers found in typical reference regions of interest (cited above), and values of h ranging from 3 to $m/2$. Figure 4(a) shows that as the maximum gap size allowed increases, so does the probability of finding 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, where $h = \frac{m}{2}$. 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).

Finally, Figure 4(c) shows the parameter values for which the probability of observing an incomplete cluster in a genome of size $n = 1000$ with $m = 100$ marked gene is no more than 0.0001. For example, with a maximum gap size of $g = 5$, a cluster is not significant until it contains at least 20 marked genes.

4 Whole Genome Models

In a <u>whole genome comparison</u> 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 <u>homolog</u> 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 comprised of a set of homologs found in proximity in two different genomes, under the assumption that both the homologs and the singletons are randomly distributed throughout the genome.

Just as we had a formal definition of a max-gap cluster for a marked gene scenario, we need now need to formulate a precise cluster definition of a max-gap cluster 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 conditions that allow one to identify a cluster. 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 who state informally that they sought clusters where the maximum gap between genes is no greater than g, provide neither a formal definition nor a precise constructive definition. This can be problematic, however, 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 genes forms a chain in a single genome if the distance from any gene in the chain to at least one other gene in the chain is never more than g. A max-gap cluster is a set of homologs that forms a chain in both genomes of interest but is not included within any larger cluster.*

An efficient algorithm for finding maximal max-gap clusters (as defined above) via whole-genome comparison has been presented by Bergeron et al. (2002). However, other groups, (e.g. Hokamp, 2001), use a greedy heuristic, in which larger clusters are built iteratively from smaller clusters. In each step, the heuristic "looks ahead" a certain number of positions to see if additional homologs may be added to the clusters 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 maximal max-gap clusters. For example, if numbers indicate homologs and circles represent singletons, then given genomes $G_1 = 1\,2 \circ 3\,4 \circ$ and $G_2 = 3\,1 \circ 4 \circ 2$, regardless of the starting point, a greedy approach using a gap size of $g = 1$ will never find the (valid) max-gap cluster $\{1,2,3,4\}$. In fact, 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 is instructive to compare the properties of clusters found by this heuristic with those of general max-gap clusters (the set of all clusters that satisfy Definition 4.1). We will see in Section 5 that the properties described here have implications for the computation and use of statistics as well as the design of algorithms.

Nested: A cluster of size h is nested if for each $k \in 1 \ldots h-1$ it contains a valid (but not maximal)

[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.

cluster of size k. Obviously, not all max-gap clusters are nested. A greedy strategy, however, will find only nested max-gap clusters; in fact, the greedy approach of building large from small clusters entails an unstated assumption of nestedness. In the example above, the cluster of size four was not found precisely because it did not contain a valid cluster of size three. It can be shown that any greedy search algorithm that constructs clusters iteratively will find exactly the set of all maximal nested max-gap clusters as long as it considers all possible homologous genes as starting points for a cluster.

Order: For general max-gap clusters gene order can be arbitrary: the order of genes in the cluster is irrelevant as long as the gaps between genes remain small. On the other hand, although gene order in nested max-gap clusters need not be perfectly conserved, they may only be disordered to a limited degree. Requiring that a max-gap cluster be nested implicitly introduces some constraints on gene order within the cluster.

Disjoint: A set of clusters is disjoint if each homolog can be a member of at most one maximal cluster. Unlike general max-gap clusters, the set of nested max-gap clusters is not guaranteed to be disjoint. For example, if $g=0$, the genomes $G_1=12345$ and $G_2=21354$ 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).

5 Whole Genome Comparison Statistics

In this paper we present a statistical model for general max-gap clusters identified through whole genome comparison. The results presented here 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 statistics 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 probability of finding a complete max-gap cluster containing all m gene pairs when comparing two 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. Figure 2(b) shows how $P(n,m,g)$ varies as m ranges from 2 to n. Recall 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 Figure 2(b) 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 can be found whenever g is greater than the longest contiguous run of singletons. These observations have implications for the design of our statistics; in particular, our hypothesis test must be altered.

Remember that for a marked gene scenario we calculated the probability of observing a cluster of size $\geq h$. This conforms to the traditional approach in hypothesis testing of determining the probability under the null hypothesis of obtaining a value of the test statistic that is more extreme (e.g. less likely) than the observed value. However, the probability of finding a cluster by whole genome comparison may actually increase with the size of the cluster. For example, as Figure 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 finding a cluster of size $\geq h$, we determine the probability of a cluster of exactly size h. These calculations can then be used to determine the probability of finding a cluster whose size is more extreme than h.

5.1 Enumeration strategy

We calculate the probability of a cluster by counting the number of 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. Figure 5

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 plot in Figure 5, where G_1 is on the vertical axis, G_2 is on the horizontal axis, and the shared cluster is represented by the black rectangle in the center. In this example, a (non-maximal) cluster of size three ($\{124\}$) that is shared between both genomes is contained within a maximal cluster of size five ($\{12456\}$). For a gap size of $g = 1$, how many configurations of the remaining four 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 around its border, since any gene that lies in these regions will increase the size of the cluster beyond $h = 3$.

What about locating a gene within g positions from the cluster in only one of the genomes (e.g. the regions delineated by dotted lines in Figure 5)? This region is not necessarily unsafe. For example, gene 7 is within a distance g of the cluster in G_2 yet does not extend the cluster since it is far from the cluster in G_1. On the other hand, though 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. Thus it is not clear how to exactly specify the unsafe regions so that we count all valid permutations while at the same time <u>not</u> counting those permutations in which the cluster can be extended. Instead, we use the above intuition to devise upper and lower bounds for the probability of finding a shared cluster of size exactly h.

5.2 Upper bound for incomplete clusters

Our upper bound counts the number of ensembles 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 <u>both</u> genomes. The key observation is that an outer gene is permitted within a distance of g from the 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 5). 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). Thus, this approach provides an upper bound on the probability of observing a maximal cluster of h genes.

Let M be the set of all m homologs shared between the genomes. As stated previously, no gene is permitted in the dark gray region, since any gene in this region will extend the cluster. Thus, the set M can be divided into three subsets corresponding to the three legal regions indicated in Figure 5:

1. $H \subset M$ is the set of h homologs that form a chain in both genomes (e.g. the black region in Figure 5),

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 (e.g. the light gray regions), and

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 (e.g., the unshaded regions).

The upper bound is the number of ways of placing these three subsets of genes on both genomes so that all constraints are satisfied, divided by the total number of ways to place the m homologs. 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

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

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 $L = \min(n, (h-1)g+h)$. The general form of q_i is:

$$\frac{1}{\binom{n}{m}} \sum_{l=h}^{L} \max(0, n-l-2g-1)\, d_2(h, g, l) \binom{X}{t}\binom{Y}{r} \;+ \sum_{i=0}^{\min(g,n-1)} E \cdot d_2(h, g, l) \binom{W}{t}\binom{Z}{r}, \tag{7}$$

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 counts clusters within a distance $i \leq g$ 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. In this case $E = 2$. However, when $l \geq n - i - g$, the cluster spans almost the entire genome, and will be simultaneously close to both ends, so $E = 1$. The values of X, Y, W, and Z needed to compute q_1 and q_2 are given in Table 1. Table 1

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 placed 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 Figure 5 are individually "safe", but together they extend the cluster. Consequently, the constraint leads to overcounting, and thus the upper bound.

To compute the lower bound we strengthen the constraint so that no outsider is allowed within a distance g of the cluster on G_1, regardless of where it is located in G_2. This is unnecessarily restrictive but guarantees that a cluster is maximal. The choice of G_1, however, is arbitrary. A constraint that no outsider is allowed within a distance g of the cluster on G_2, regardless of where it is located in G_1 would also guarantee a maximal cluster. Our lower bound is the probability of an ensemble that satisfies either of the two constraints above, e.g. the union of the two constraints. By the inclusion-exclusion rule, the union is simply the sum of the probability that each constraint is satisfied minus the probability that both constraints are satisfied.

Assuming equal genome sizes, the first and second scenarios are symmetric, and consequently the probabilities are equal. The probability can be computed by Equation 6, replacing q_1 with q_3, for which $X = l - h$ instead of $b - h$. The intersection of the two constraints is not empty, i.e. the two scenarios are not independent; in enumerating all permutations that obey either constraint we will have double counted those permutations in which no homolog is within a distance g of the cluster in either genome. Thus we must subtract out the probability of observing a cluster of size h where there is no homolog 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:

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

Equation (8) is guaranteed to give a lower bound on the probability of observing a cluster of size h for all $h > m/2$. 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 with 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. For small values of h, it is possible that the probability computed by Equation (8) will actually exceed the true probability.

5.4 Experiments

Figure 6

In order to investigate the accuracy of the bounds in different regions of the parameter space, we compared them to the probability of finding 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 Figure 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 Figure 6(a) the simulated probabilities are only shown for $h \leq 10$ 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. As Figure 6(a) illustrates, when the maximum gap size is small with respect to n (about 1%), the upper bound is extremely accurate for all values of h. However, when the maximum gap size is larger with respect to n (2% or 3%), then the bounds are only exact when estimating the probability of a large or complete max-gap cluster. This is illustrated in Figure 6(c), which shows the behavior of the bounds when $n = 500$, $m = 166$, and $g = 15$. For these parameter values, the bounds are extremely 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 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 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 $n = 1000$, $m = 250$, and $g = 20$, Figure 6(b) shows that the chance probability of observing a cluster of fifty genes is actually smaller than the chance probability of observing a cluster of 100 genes. This non-monotonic behavior can be understood intuitively by observing that, as the size of the cluster increases, the max-gap criteria implicitly increases the maximum allowed window size. As a result, as the size of the cluster sought increases, the probability of observing such a cluster may grow substantially as well.

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 homologs between the two genomes was obtained from `http://www.cs.kent.edu/arvind/intellibio/database/orthologs` (Bansal, 1999). The E. coli genome has $n = 4108$ known genes and the B. subtilis genome has $n = 4245$ known genes. After eliminating ambiguous orthologs, the map yields $m = 1315$ 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.

Table 2

Figure 7

A subset of the results selected to show the general trends is shown in Table 2. In addition, Figure 6 shows the sizes of the clusters found with a range of different gap sizes. Our results fall into three regimes. When $g = 0 \ldots 40$, cluster sizes range from two to twelve, except for one larger cluster of size 20 to 30. When $g = 40 \ldots 70$, clusters sizes have a larger range, from two to about 600. Finally, for gap sizes of $g \geq 70$, the homologs form only one or two large clusters.

Figure 8

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 $n = 4108$, $m = 1315$, and a range of values of h and g. These probabilities are indicated by the dashed lines in Figure 8. To assess the accuracy of our upper bound for this bacterial dataset, we again compared it with estimates of the probability of finding max-gap clusters in randomly permuted genomes of the same size, obtained through simulation. We generated one million random permutations of two genomes with $n = 4108$ genes and $m = 1315$ 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 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 appears to be quite 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 $20 \leq h \leq 1314$ occurred only infrequently, 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 $h = 1312$ and $h = 1306$, 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 Figure 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 two to ten 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 zero to ten, over 90% of the clusters are comprised entirely of genes from a single operon. The single large cluster of over twenty genes is composed entirely of ribosomal proteins, which together from the ribosomal "super-operon" in E. coli.

An intriguing observation is that the number of large clusters seems to be fewer than expected under our null model. When $g \geq 25$, our 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 forming 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 may not be so extreme.

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., 2004; 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 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 finding the probability of 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 interesting directions for future work. 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 a model that explicitly considers distance 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 finding a conserved cluster by chance and, hence, can have a large impact on the statistical significance of a particular cluster. However, 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. 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 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 only find a subset of all max-gap clusters. In addition, our statistical analysis demonstrates that when conducting a whole genome comparison, 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 with only a max-gap criterion, larger clusters do not always imply greater significance.

Acknowledgment

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, 100–5.

Amores, A., Force, A., l. 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.

Bergeron, A., Corteel, S., and Raffinot, M., 2002. The algorithmic of gene teams. In Gusfield, D. and Guigo, R., eds., Algorithms in Bioinformatics, Second International Workshop WABI2002, Lecture Notes in Computer Science 2452, 464–476.

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

Blanchette, M., Kunisawa, T., and Sankoff, D., 1999. Gene order breakpoint evidence in animal mitochondrial phylogeny. Journal of Molecular Evolution 49, 193–203.

Bourque, G., Zdobnov, E., Bork, P., Pevzner, P., and Telser, G., 2004. Genome rearrangements in human, mouse, rat and chicken. Genome Research .

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. Nucleic 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 extitCampanulaceae. 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, 141–8.

Durand, D. and Sankoff, D., 2003. Tests for gene clustering. Journal of Computational Biology 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., 2001. Gene duplication and the structure of eukaryotic genomes. Genome Res 11, 373–81.

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

Graham, Knuth, and Patashnik, 1989. Concrete Mathematics. Addison-Wesley.

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.

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

Hughes, A. L., 1998. Phylogenetic tests of the hypothesis of block duplication of homologous genes on human chromosomes 6, 9, and 1. MBE 15, 854–70.

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, 101–8.

Lawrence, J. and Roth, J. R., 1996. Selfish operons: horizontal transfer may drive the evolution of gene clusters. Genetics 143, 1843–60.

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: extitSLC4A9 at 5q31. Genome Biology 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. Comput Biol Chem. 27, 59–67.

McLysaght, A., Hokamp, K., and Wolfe, K. H., 2002. Extensive genomic duplication during early chordate evolution. Nat. 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 U S A 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. MBE 15, 1145–59.

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., Martnez-Antonio, A., Daz-Peredo, E., Snchez-Solano, F., Peralta-Gil, M., Garcia-Alonso, D., Jimnez-Jacinto, V., Santos-Zavaleta, A., Bonavides-Martnez, C., and Collado-Vides, J., 2004. RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12. Nucleic Acids Res 32, D303–6.

Sankoff, D., 2003. Rearrangements and chromosomal evolution. Curr Opin Genet Dev 13, 583–7.

Sankoff, D., Bryant, D., Deneault, M., Lang, B. F., and Burger, G., 2000a. Early eukaryote evolution based on mitochondrial gene order breakpoints. J Comput 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 U S A 100, 11188–9.

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 U S A 99, 13627–32.

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, 0020.1–11.

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 Peer, Y. V. D., 2002. The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between arabidopsis and rice. Genome Res 12, 1792–801.

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.

List of Figures

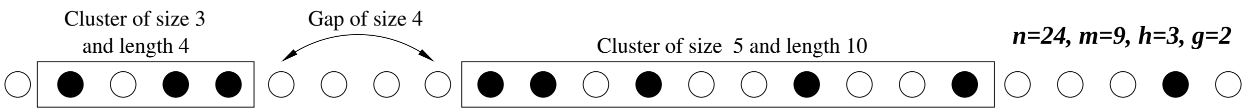
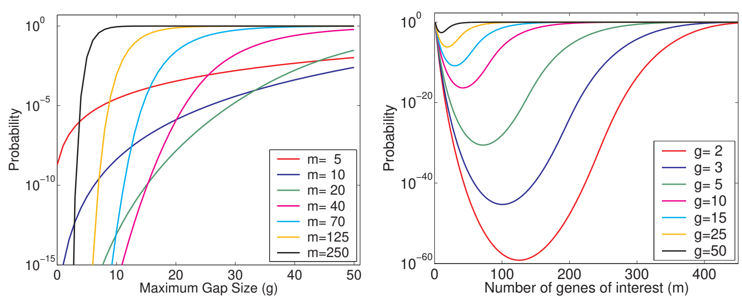
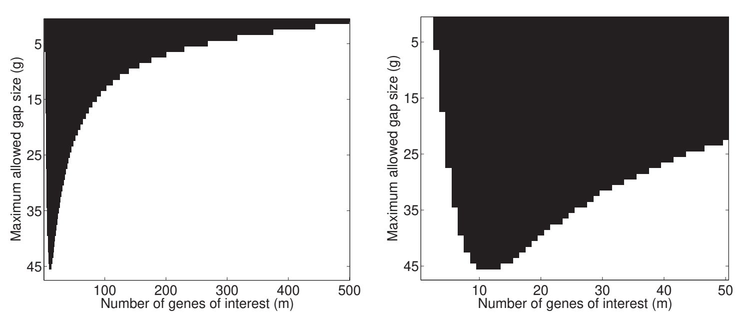


Figure 1: A sample genome ($n = 24$), with $m = 9$ marked genes shown in black. Two clusters are found when the maximum gap allowed is $g = 2$. The rightmost marked gene is not part of any cluster.



(a) (b)

Figure 2: Probability of a complete max-gap cluster of m marked genes in a genome of size $n = 500$ (a) as a function of g and (b) as a function of m.



(a) (b)

Figure 3: Region of the parameter space that is statistically significant (shown in black) at the $\alpha = 0.0001$ level for a complete cluster in a genome of size $n = 500$. (a) Complete parameter space where m ranges from 1 to 500. (b) Detail for $m \leq 50$.

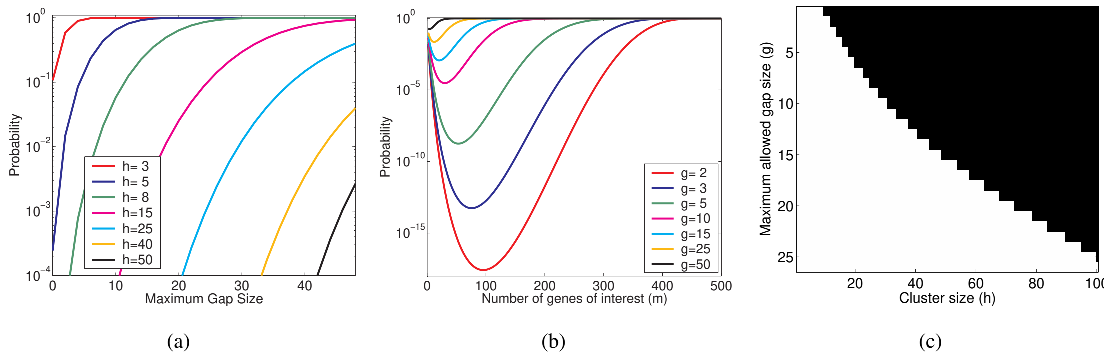


Figure 4: (a) Probability of an incomplete cluster of size at least h when $n = 1000$ and with $m = 50$. (b) Probability of an incomplete cluster that contains at least half of all m marked genomes when $n = 500$. (c) Region of the parameter space that is statistically significant (shown in black) at the $\alpha = 0.0001$ level for an incomplete cluster of $m = 100$ marked genes in a genome of size $n = 1000$.

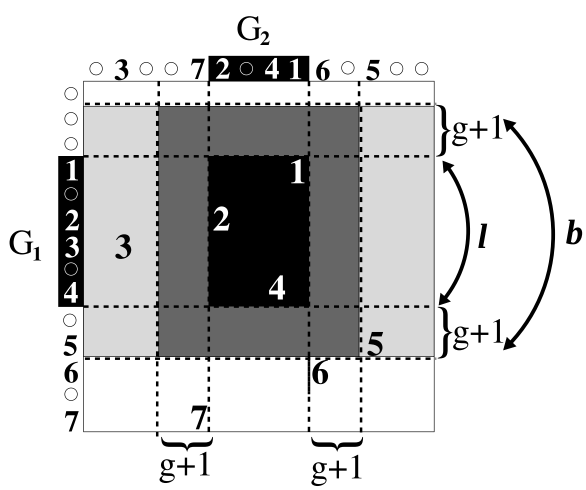
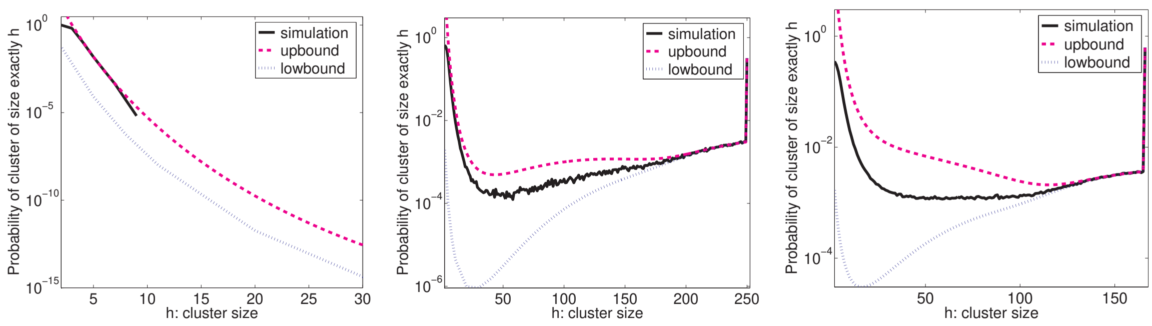


Figure 5: A dot plot comparing two genomes—G_1 on the vertical and G_2 on the horizontal axis—that share $m = 7$ homologous gene pairs. Singletons are drawn on the axes as circles, but not shown in the dot plot.



(a) (b) (c)

Figure 6: Comparison of simulation results (solid lines) to upper bound (dashed lines) and lower bound (dotted lines). Probability of finding max-gap clusters of size h when (a) $n = 1000$, $m = 250$, and $g = 10$, (b) $n = 1000$, $m = 250$, and $g = 20$, and (c) $n = 500$, $m = 166$, and $g = 15$.

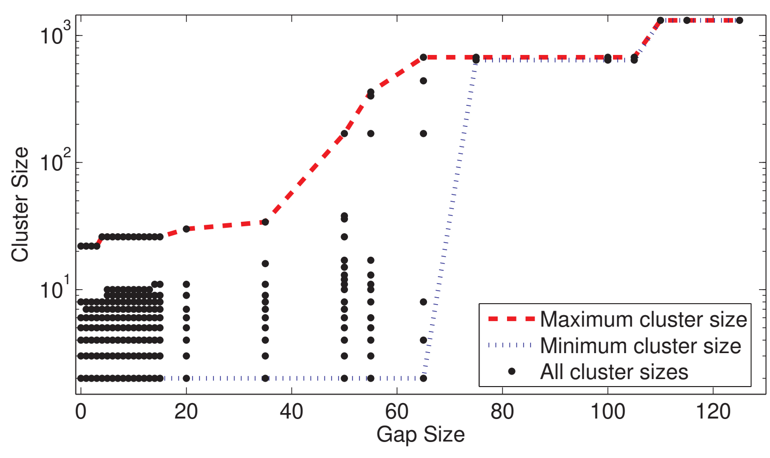


Figure 7: The distribution of observed cluster sizes between E. coli and B. subtilis for g ranging from 0 to 125. The dashed line indicates the largest cluster found and the dotted line indicates the smallest cluster found for each value of g.

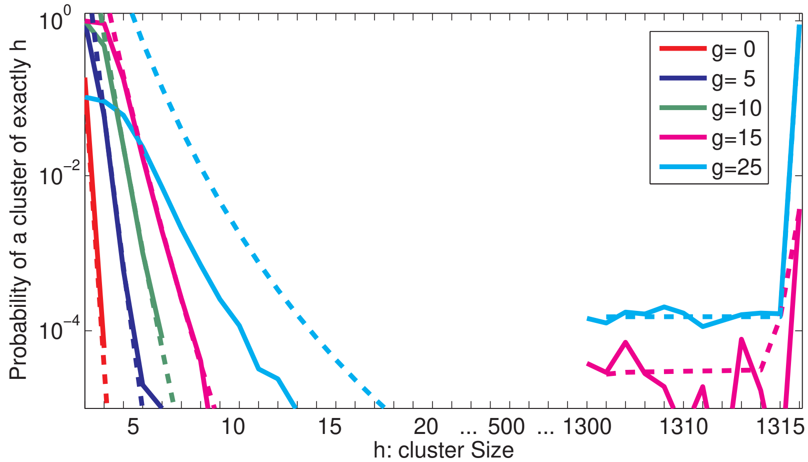


Figure 8: Probability of finding max-gap clusters of size h in the E. coli and B. subtilis genome comparison for $g = 0$ to $g = 25$. Solid lines show probabilities estimated via randomized genomes and dashed lines indicate the upper bound on the probabilities as calculated by Equation (6).

List of Tables

Case	X	Y	W	Z
q_1	$b-h$	$n-b$	$c-h$	$n-c$
q_2	$n-b$	$n-h-t$	$n-c$	$n-h-t$
q_3	$l-h$	$n-b$	$l-h$	$n-c$
q_4	$n-b$	$n-h-t-b+l$	$n-c$	$n-h-t-c+l$

Table 1: The values of q_1 through q_4. The length of the cluster plus its bounding moats (as shown in Figure 5) is given by $b = l + 2(g+1)$. Note that W and Z, which account for edge effects, are identical to X and Y respectively, except any instance of b is replaced by $c = \min(n, l+i+g+1)$, which is the size of the cluster plus its bounding moats when the cluster is located within $i \leq g$ genes of either edge of the genome.

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

Table 2: Number of max-gap clusters of varying sizes shared between E. coli and B. subtilis for a range of gap values.