The Statistical Significance of Max-Gap Clusters

Rose Hoberman[1]*, David Sankoff[2], and Dannie Durand[3]

[1] Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, USA. E-mail: roseh@cs.cmu.edu

[2] Department of Mathematics and Statistics, University of Ottawa, Ontario, Canada. E-mail: sankoff@uottawa.ca

[3] Departments of Biological Sciences and Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. E-mail: durand@cmu.edu

Abstract. Identifying *gene clusters*, genomic regions that share local similarities in gene organization, is a prerequisite for many different types of genomic analyses, including operon prediction, reconstruction of chromosomal rearrangements, and detection of whole-genome duplications. A number of formal definitions of gene clusters have been proposed, as well as methods for finding such clusters and/or statistical tests for determining their significance. Unfortunately, there is very little overlap between previously published rigorous analytical statistical tests and the definitions used in practice. In this paper, we consider the *max-gap* cluster: a contiguous region containing a maximal set of homologs, where the number of non-homologous genes between pairs of adjacent homologs is never greater than a predefined, fixed parameter, g. Although this is one of the models most widely used in practice, currently the statistical significance of max-gap clusters can only be evaluated using Monte Carlo simulations because no analytical statistical tests have been developed for it. We give exact expressions for the probability of observing such a cluster by chance, assuming a simple reference-region scenario and random gene order, as well as more efficient methods for approximating this probability. We use these methods to identify which regions of the parameter space yield clusters that are statistically significant. Finally, we discuss some of the challenges in extending this model to whole-genome comparison.

1 Introduction

Identification of conserved chromosomal segments is an essential first step for many different types of genomic analyses. Regions of similar gene content in related genomes can provide evidence for evolutionary relatedness or functional selection on gene order. For example, within a single genome the pattern of duplicated regions can provide evidence for large-scale or whole-genome duplication [2, 17, 18, 40, 53, 54, 66–68]. Conserved segments between different genomes, on the other hand, have been used extensively to reconstruct the history of chromosomal rearrangements and infer an ancestral genetic map for a diverse group of

* Contact author

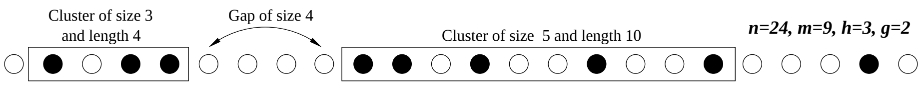


Fig. 1. A sample genome ($n = 24$), with $m = 9$ genes of interest shown in black. When the maximum gap allowed is $g = 2$ and the minimum cluster size is $h = 3$, then two clusters are found. The rightmost black gene is not part of any cluster.

species [8, 11, 16, 43, 41, 55, 47, 51], as well as to provide coarse-grain features for new phylogenetic approaches [6, 12, 26, 49, 50, 62]. In bacteria, conserved gene order and content have been used for prediction of operons [7, 20], horizontal transfers [36], and more generally to help understand the relationship between spatial organization and functional selection [31, 35, 45, 60, 61].

The common goal in all of these analyses is either to detect regions that share a common ancestor or where gene content is under functional selection. The signature of such conserved regions, which we call *gene clusters*, will be similar gene content, but we do not require gene content or order be strictly conserved as this would rule out many more distantly related regions.

It is not obvious how to choose a formal definition that best captures our intuitive notions about gene clusters. A number of definitions have been proposed, as well as algorithms for finding clusters which meet these definitions and statistical tests to evaluate their significance [3, 22, 23, 29]. The most stringent of these define conserved segments as two or more contiguous regions that contain the same genes in the same order [42, 44] and sometimes orientation [45, 60, 68]. However, such stringent definitions will invariably lead to the exclusion of many regions that did indeed descend from a single ancestral region but have since undergone small rearrangements. More flexible definitions allow for some amount of divergence and rearrangement.

Many of these more flexible definitions are based on a simple model in which a genome is represented as an ordered set of n genes: $G = (g_1, \ldots, g_n)$. Chromosome breaks are ignored and it is assumed that genes do not overlap. We start with a simple abstraction in which m genes ("the black genes") are pre-specified as interesting. These m genes may be of interest because their homologs are contiguous in another region or genome (the "reference region") or because they share some functional properties. We are interested in finding a large group of black genes that appear in close proximity. The *size* of the cluster is usually quantified as the total number of black genes in the cluster, where a *complete* cluster contains all m black genes and an *incomplete* cluster contains only a subset of the black genes. For example, a short genome with $n = 24$ genes is illustrated in Figure 1. The $m = 9$ black genes are shown grouped into two incomplete clusters, of size three and five respectively.

Although it is quite clear how to characterize cluster size, there is no agreed upon definition of "close proximity." Some definitions restrict the total *length* of the cluster [15] (the total number of genes from the first to the last black gene in the cluster). Others constrain the cluster density (the proportion of black genes in the cluster, or size / length). Others require only that clusters be compact [52], where compactness is determined by the distance, or *gap*, between adjacent black genes, that is, the number of white genes between them. For example, in Figure 1 the gap between the first and second black genes is one and the gap between the second and third black genes is zero. Of the definitions that constrain the gap sizes, some allow no gaps in a cluster [27, 28], others limit the sum of all gaps, while the majority constrain the size of the largest gap observed [5, 10, 40, 45, 56, 60, 65, 67].

In addition to cluster size and length, many cluster definitions constrain gene order, with some requiring a strictly conserved gene order, while others allow only a fixed number of order violations [25]. The majority ignore gene order altogether.

Although a number of formal definitions of gene clusters have been proposed, there is unfortunately very little overlap between cluster definitions used in analyses of genomic data and the definitions upon which rigorous analytical statistical tests are based. In this paper, we focus on a particular cluster definition that is widely used in genomic studies, including the identification of large-scale duplications in *Arabidopsis* [5] and the chordate lineage [40], the assignment of functions to uncharacterized genes in prokaryotes [45, 60], and the prediction of putative operons in newly sequenced bacterial genomes [10]. According to this definition, gene order is disregarded, and there is no limit on the total number of gaps as long as the maximum gap between adjacent black genes in the cluster is not too large. To distinguish these clusters from our informal notion of a cluster we call them *max-gap* clusters. A max-gap cluster is a maximal set of black genes where the gap between adjacent black genes is never larger than g. For example, when the maximum gap allowed is $g = 2$, three clusters can be found in the example genome in Figure 1. The first has size three and length four, the second has size five and length ten, and the third is a singleton.

The max-gap cluster definition has a number of desirable properties. It is flexible in that it does not require that every gene in the cluster have a homolog, yet it guarantees that the gap between adjacent homologs will not be too large. As a result, the density of a cluster is guaranteed to be no less than $1/(g + 1)$. This definition does not arbitrarily constrain the cluster length, but instead lets clusters grow to their "natural" size. Consequently, clusters will never overlap: unlike some other cluster definitions [15, 9], a gene can never be considered part of two distinct clusters that cannot be merged. On the other hand, two max-gap clusters containing the same number of homologs may have significantly different densities. For example, the length of a cluster of size m can range from m (density of one) to $g(m - 1) + m$ (a density close to $1/(g + 1)$). Finally, an algorithm has been developed for finding max-gap clusters efficiently [4]. However, most groups do not describe in detail the algorithm they use for finding max-gap clusters, so it is not clear whether they are using an efficient or even a correct algorithm.

Analytical statistical models in the literature are designed for other definitions of gene clusters [9, 14–16, 63, 66] and it is not obvious how to extend them to apply to this commonly used cluster model. Studies based on the max-gap cluster model usually use randomization to estimate the significance of clusters [5, 40, 45, 56, 65, 67]. However, this approach "is computationally expensive and does not permit very precise estimation of the probabilities of rare events" [9]. In addition, parameter values such as the maximum gap and minimum cluster size are generally selected in an ad-hoc manner. A formal, rigorous mathematical model of gene clusters will allow us to evaluate cluster significance more accurately and more quickly, and to choose parameter values in a principled manner.

Our goal in this paper is to try to close the gap between rigorous mathematical models and models used in the analysis of real genomes by developing formal statistical tests for max-gap clusters. We first present an exact expression for the probability of observing a complete max-gap cluster containing all m genes of interest within a randomly ordered genome of size n. We also provide an approximation for faster analysis. Next we extend this analysis to evaluate the probability of observing a cluster containing only a subset of the black genes. We present a simple dynamic programming algorithm that exactly calculates the probability of observing an incomplete cluster of size $h < m$, as well as an analytic solution for the case where $h > \frac{m}{2}$. We then use these equations to calculate the probability of clusters for a range of different genome sizes and parameter values. We discuss the influence of the parameters n, m, g and h on cluster significance and determine which regions of the parameter space yield clusters that are statistically significant. Finally we discuss some of the challenges that arise in extending this statistical model to whole-genome comparison.

2 Probabilities of Max-Gap Clusters

Our analytical tests of max-gap cluster significance are based on the probability of observing a cluster by chance in a genome with random gene order, the most basic null hypothesis we can consider. If we cannot reject that null hypothesis, no more complex, biologically motivated null hypothesis need be considered.

When calculating the probability of max-gap clusters it will be useful to know the number of ways of arranging m black genes to form a max-gap cluster within a window of length l. When both endpoints of the window contain a black gene the cluster will be of length *exactly* l and the problem is equivalent to a well-known sum-of-dice combinatorics problem [64]. Let

$$d_c(m, g, l) = \sum_{i=0}^{\lfloor (l-m)/(g+1) \rfloor} (-1)^i \binom{m-1}{i} \binom{l - i(g+1) - c}{m - c}.$$

When $c = 2$, $d_c(m, g, l)$ corresponds to the the number of ways of rolling $m - 1$ dice, each with faces numbered 0 to g, such that the sum of their faces is equal to $l - m$.[4] This is equivalent to the number of ways of creating a max-gap cluster

[4] 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 $l - 1$, due to Uspensky [64].

of size m and length l since such a cluster has $m-1$ gaps with a cumulative sum of $l-m$.

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 black gene. This is simply: $\sum_{r=m}^{l} d_2(m,g,r)$, which can be shown to be equivalent to $d_1(m,g,l)$ (see the Appendix for the derivation). 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 $\sum_{r=m}^{l} d_1(m,g,r) = d_0(m,g,l)$.

These expressions will be used in the subsequent sections in various situations in which the length of a cluster is constrained. Note that an efficient implementation of d_c can be obtained by pre-computing all necessary factorials, allowing the entire summation to be computed in $O(l)$ time.

2.1 Exact and Approximate Probabilities for Complete Max-Gap Clusters

We begin by calculating the probability of observing a *complete* max-gap cluster. More formally, given a random genome of size n, what is the probability of observing all m black genes (in any order), such that the gap between adjacent black genes does not exceed g. We determine the probability by counting the number of ways to place all m genes in a genome of size n so that they form a max-gap cluster. We enumerate the clusters by the position of the leftmost black gene in the cluster. Given the position of the first black gene, there are $(g+1)^{m-1}$ ways to place the remaining black genes so that they form a max-gap cluster, which is simply the number of ways of choosing $m-1$ gaps so that the length of each gap is between 0 and g. The maximum possible length of a max-gap cluster is $w = m + g(m-1)$, and thus there are $n - w + 1$ ways of placing the first black gene so that a cluster of maximal length can be accommodated. In addition, the leftmost black gene could also be positioned within the $w-1$ genes at the end of the genome. The number of ways of placing m black genes to form a max-gap cluster in the last $w-1$ slots is precisely the quantity we derived in the previous section. Combining these terms, the probability of observing a complete max-gap cluster of m genes in a genome of size n is

$$P_M(n,m,g) = \frac{\max(0, n-w+1) \cdot (g+1)^{m-1} + d_0(m,g,\min(n,w-1))}{\binom{n}{m}}. \quad (1)$$

When $m \ll n$, the total number of permutations can be approximated in constant time using Stirling's approximation, and then the complexity of computing P_M is simply $O(w) = O(mg)$. Except when $w \geq n$, 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 a more efficient running time is required, we can construct a lower bound on the probability of observing a cluster by simply eliminating the final term that takes edge effects into account.

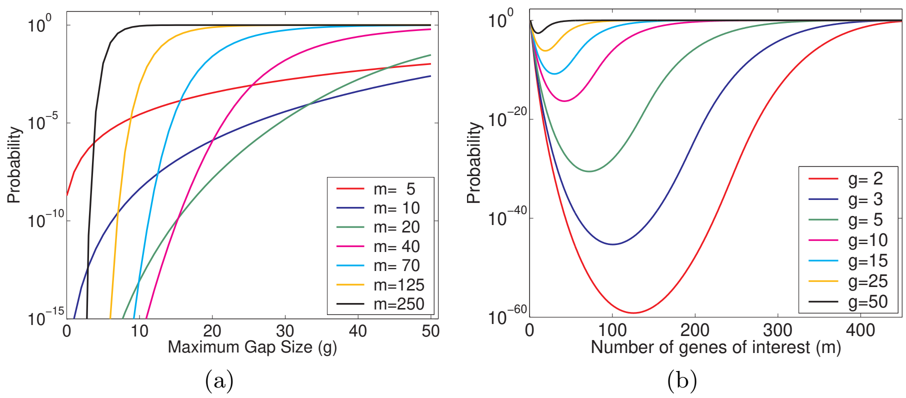


Fig. 2. Probability of a complete max-gap cluster of m black genes in a genome of size $n = 500$ as a function of g (a), and as a function of m (b).

We can compute an upper bound by instead assuming that all but the last $m-1$ positions in the genome can accommodate a cluster of maximal length:

$$\frac{\max(0, n-w+1) \cdot (g+1)^{m-1}}{\binom{n}{m}} \leq P_M(n, m, g) \leq \frac{(n-m+1) \cdot (g+1)^{m-1}}{\binom{n}{m}}.$$

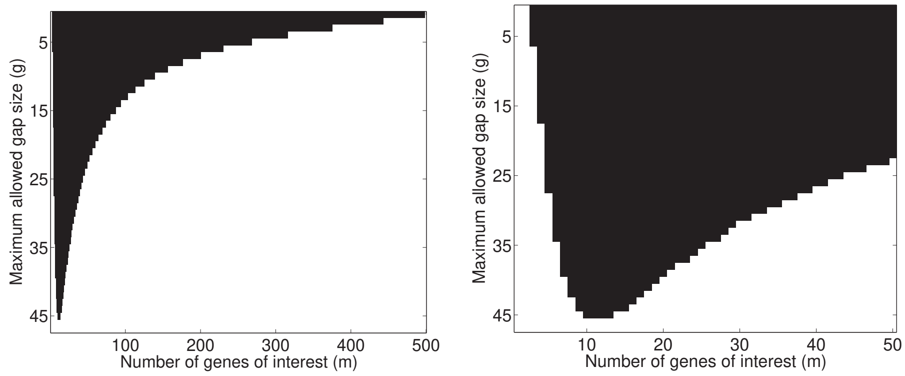
Both bounds can be computed in constant time using Stirling's approximation to estimate the denominator. We have verified empirically that when n is large in relation to w, the upper bound is only a slight overestimate of P_M (data not shown).

In some cases we may wish to constrain the total length of the cluster, by adding the restriction that all m genes must appear in a window of size at most r. The limit on window size ensures a minimum cluster density, while the max-gap property prevents the gaps between black genes from becoming too large. More formally, given a genome of size n, the probability of finding all m black genes (in any order) in a window of size at most r such that the gap between adjacent black genes is never more than g, is simply

$$P_{MR}(n, m, g, r) = \frac{1}{\binom{n}{m}} \left[(n-r+1) \cdot d_1(m, g, r) + d_0(m, g, r-1)\right],$$

where we have replaced $(g+1)^{m-1}$ in Equation 1 with $d_1(m, g, r)$ in order to constrain the maximum length of the cluster.

The probability of finding a complete cluster for varying values of n, m, and g was calculated from Equation 1 using Mathematica. 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 [60] to about thirty in human [40]. We calculated probabilities for genomes sizes of 500, 1000, 5000, 20,000, and 25,000, corresponding to typical gene sets for bacteria, yeast, worm,



(a) (b)

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

and higher eukaryotes like human and *Arabidopsis*. For complete clusters we tested all values of m ranging from 2 to n.

Figure 2(a) shows the probability of observing a complete cluster containing all m black genes in a genome of size $n = 1000$, as m ranges from 1 to 250 and g increases from 2 to 50. The probability of finding a complete cluster increases monotonically with g. We might also expect that this probability will increase monotonically with m, 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 1, and the values of m in between these two extremes will have probabilities of less than one.

One question of interest is the range of values of m and g for which is it possible to obtain a significant cluster. Figure 3 shows the parameter values for which the probability of observing a 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.

As the genome size n increases the probabilities decrease but the general trends seen in Figure 2 remain the same (data not shown).

2.2 Exact Probabilities for Incomplete Max-Gap Clusters

Requiring all m genes of interest to appear in a single cluster is often too strict a requirement. In practice, researchers often look for clusters that contain a subset of the genes of interest [1, 13, 19, 21, 30, 33, 34, 37, 39, 46, 48, 58, 59, 63]. 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 black genes). Unlike complete clusters, there can be more than one incomplete cluster in the same genome. A simple extension of Equation 1 to incomplete clusters would therefore lead to overcounting permutations containing more than one cluster. Instead, we present a simple dynamic programming algorithm 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 moves along the genome, adding a black or white gene at each step. It keeps track of runs of black genes that satisfy the max-gap cluster criterion and avoids creating a cluster of size h or larger by judicious placement of white genes.

The quantity $n_{\bar{H}}[n, m, j, c]$ represents the number of ways to place m black 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 black gene and c is the size of any cluster created so far. It is defined recursively as follows:

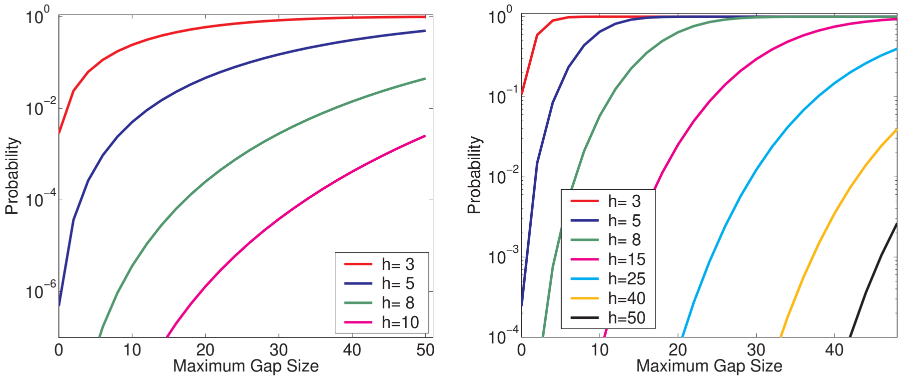
$$n_{\bar{H}}[n, m, j, c] = \begin{cases} 0, & \text{if } c = h \\ 0, & \text{else if } n < m \\ 1, & \text{else if } m = 0 \\ n_{\bar{H}}[n{-}1, m, j{+}1, c] + n_{\bar{H}}[n{-}1, m{-}1, 0, c{+}1], & \text{else if } j \leq g \\ n_{\bar{H}}[n{-}1, m, j{+}1, c] + n_{\bar{H}}[n{-}1, m{-}1, 0, 1], & \text{otherwise.}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

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

$$P_H(n, m, h, g) = 1 - \frac{n_{\bar{H}}[n, m, g+1, 0]}{\binom{n}{m}}. \tag{2}$$

The complexity of computing P_H 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 [57], 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 [69]. Vertebrate genomes can be much larger. For example, the estimated size of the human genome is around 25,000 genes [32], 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 [1, 13, 19, 21, 30, 33, 34, 37, 39, 46, 48, 58, 59, 63]. 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 $h > \frac{m}{2}$, the probability can be computed directly because we do not have to worry about overcounting genomes containing more than one cluster. We count the number of permutations containing a cluster, enumerating them by the position of the leftmost black gene in the leftmost cluster, just as we did for complete clusters. Unlike the complete case, however, we have to be careful not to overcount clusters of size greater than h. We accomplish this by considering each possible cluster length (for the first h black genes in the cluster) individually



(a) (b)

Fig. 4. Probability of an incomplete cluster of size at least h as a function of gap size in (a) a genome of 500 genes with $m = 10$ black genes, (b) a genome of 1000 genes with $m = 50$ black genes.

and placing $g + 1$ white genes before the start of the cluster to ensure that it cannot be extended to the left. This yields a probability of finding an incomplete cluster of size at least h of

$$\frac{1}{\binom{n}{m}} \sum_{l=h}^{h+g(h-1)} \left[(n-l-g) \cdot d_0(h,g,l) \cdot \binom{n-l-g-1}{m-h} + E \right], \tag{3}$$

where l ranges over all possible lengths of a cluster of size h and E is a term to address edge effects. The first term is the number of positions in which to start the cluster. The second term is the number of ways to choose the gaps to obtain a cluster length of exactly l. The third term is the number of ways to place the remaining $m - h$ genes outside the cluster. The final term counts clusters close to the beginning of the genome before which it is only possible to place $i < g+1$ white genes. It is calculated as

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

where the binomials are defined to be zero when the upper value is smaller than the lower value and the simplification is by application of the upper summation identity [24]. The complexity of computing Equation 3 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 2 (data not shown).

We calculated the probability of finding an incomplete cluster from Equations 2 and 3 using Mathematica for the values of n and g given in Section 2.1. 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 [1, 13, 19, 21, 30,

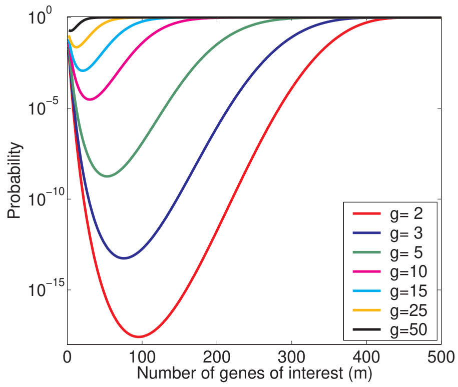
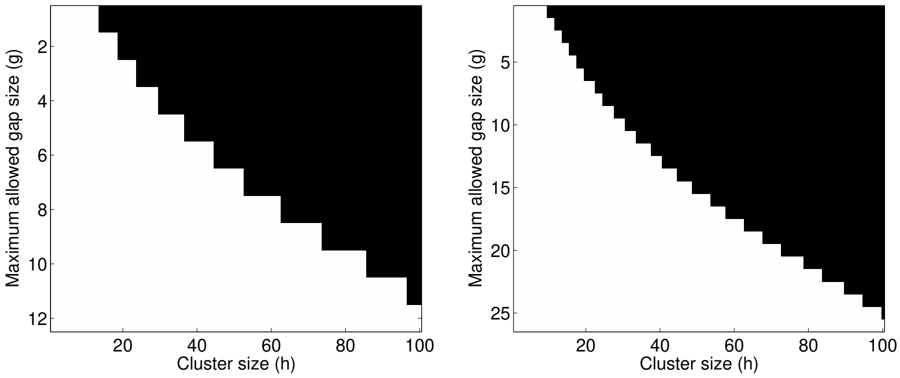


Fig. 5. Probability of observing a cluster that includes at least half of all m black genes in a genome of size 500.

$33, 34, 37, 39, 46, 48, 58, 59, 63$], and values of h ranging from 3 to $m/2$. Figure 4 shows the probability of observing a cluster of a subset of 50 black genes in a genome of size 500 for varying values of g and h. 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 5 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. Finally, Figure 6 shows the region of parameter space for which it is possible to find a significant cluster at a significance level of $\alpha = 0.0001$, when $m = 100$, for genomes of size $n = 500$ and $n = 1000$. Probabilities were also calculated for larger genome sizes as in Section 2.1. Again, as n increases the probabilities decrease but the general trends are similar (data not shown).

3 Discussion

The work presented here was motivated by the gap that currently exists between mathematical cluster models and models used in analysis of real genomes. We provide analytical statistical tests for max-gap clusters, a model widely used in practice [5, 10, 38, 40, 45, 60]. We determine the probability of observing a max-gap cluster containing a set of m pre-specified genes of interest, assuming a genome with random gene order. We also consider incomplete clusters, where a subset of the pre-specified genes satisfies the max-gap criterion. This scenario corresponds to a reference-region approach in which a particular chromosomal region in one genome is of interest, and another region containing a similar set of genes is sought. We have presented exact expressions for the probabilities of finding complete and incomplete max-gap clusters under this simple model. We have also provided an efficient approximation for the probability of finding a complete cluster, which is highly accurate when n is large in relation to mg.



(a) (b)

Fig. 6. Region of the parameter space that is statistically significant (shown in black) at the $\alpha = 0.0001$ level for $m = 100$ black genes in a genome of size $n = 500$ (a) and $n = 1000$ (b).

Our calculations show that the probability of finding a cluster increases monotonically with g, and that as the gap size increases, the range of values of m for which it is possible to obtain a significant cluster becomes more and more restricted. For a fixed value of m, increasing the required size (h) of an incomplete cluster decreases its probability of occurring by chance. However, the behavior of cluster probabilities with respect to m is more complex. There is a high probability that all m black genes will form a cluster when m is small in relation to n, and this probability decreases as m grows larger. As m approaches n, however, the majority of genes in the genome will be black, and the probability that they cluster together begins to increase again. This behavior is also observed for incomplete clusters when h is chosen to be a fixed percentage of m.

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 often differ substantially between organisms. 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. These extensions remain as future work.

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.

An important open problem is the development of statistical tests for max-gap clusters in whole genome comparisons. More formally, given two genomes $G = (g_1,\ldots,g_n)$ and $H = (h_1,\ldots,h_n)$, and a mapping between homologs in G and H, we wish to find all maximal max-gap clusters containing at least k homologs. It is not obvious how to calculate max-gap cluster probabilities in the case of whole-genome comparison because, unlike the abstraction of white and black genes presented here, in whole-genome comparison there is no specific set of genes that is of interest. Consider the simple model of whole genome comparison in which the genomes are assumed to have identical gene complements, and can therefore be treated as two permutations of the numbers $1, \ldots, n$. Although this model appears quite natural, max-gap clusters found under this approach to genome comparison have some surprising properties[5]:

1. Under this simple model of genome comparison with identical gene content, there will always be a cluster of size n and hence, the probability of finding a max-gap cluster of at least size k when comparing two genomes is always one. For example, consider these two genomes:

```
G = 1 2 3 4 5 6
H = 3 4 6 5 1 2
```

Suppose we wish to find the largest max-gap cluster that can be formed around gene 3, when $g = 0$. If we attempt to construct a cluster in a greedy fashion, the cluster will only include genes 3 and 4. However, if we look ahead a bit, it is possible to find the cluster `[3 4 5 6]`. In both genomes there are zero gaps between these four genes. Extending this look-ahead idea, we can see that under this model, regardless of the value of g, a pair of genomes always contain a max-gap cluster of size n. Since $n \geq k$, the probability of finding a cluster of size at least k is one.

2. In the reference region model discussed in this paper, as well as the gene cluster models of Durand and Sankoff [15] and Calabrese *et al.* [9], a cluster that contains k genes will always contain at least one valid cluster of size each from 1 to $k - 1$. However, this property does not hold when applying the max-gap

[5] Bergeron and colleagues [4] 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 model to whole genome comparison. For example, consider the following two genomes:

```
G = ... 0  1  2  3  4  5  6  7  8  9 10 11 12 ...
H = ... 2  4 27 30  9 12 53 81  0  8 99 72  7 ...
```

With a maximum allowed gap of $g = 2$, the size of the largest max-gap cluster is seven: `[0 2 4 7 8 9 12]`. However, this cluster does not contain any valid max-gap clusters of size three to six. Indeed, it contains only sub-clusters of size two (`[2 4]`, `[9 12]`, and `[7 8]`). This issue is related to point (1). There may be a higher probability of finding a larger cluster than a smaller cluster. To see why this is the case, note that increasing the size of the cluster essentially increases the maximum allowed window size. As a result, as the size of the cluster sought increases, the number of clusters found may grow substantially.

When looking for evidence of whole-genome duplication, a genome is compared with itself, and the gene sets will indeed be identical. In the comparison of two different genomes, however, point (1) will not be an issue, because gene sets are never identical in practice. This problem can be partially addressed by a more realistic model, where only a subset of the gene sets of the two genomes are shared. We assume that only m genes in each genome have homologs in the other genome, and the non-homologous genes are randomly distributed throughout the genome. When $g = 0$, the non-matching genes will create a natural barrier to unlimited extension of a cluster, preventing the formation of a max-gap cluster of size m. However, if g is greater than the longest contiguous run of non-matching genes then it will still be possible to form a cluster of size m. Furthermore, this more realistic model does not circumvent the second issue of non-monotonic cluster sizes. These two issues have implications for the development of analytical statistical models of max-gap clusters found through whole-genome comparison, and remain exciting problems for the future.

4 Acknowledgments

D.D. was supported by NIH grant 1 K22 HG 02451-01 and a David and Lucille Packard Foundation fellowship. D.S. was supported in part by grants from the Natural Sciences and Engineering Research Council of 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.

References

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

2. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408:796–815, 2000.
3. A. K. Bansal. An automated comparative analysis of 17 complete microbial genomes. *Bioinformatics*, 15:900–908, 1999.
4. A. Bergeron, S. Corteel, and M. Raffinot. The algorithmic of gene teams. In D. Gusfield and R. Guigo, editors, *Algorithms in Bioinformatics, Second International Workshop WABI2002*, Lecture Notes in Computer Science 2452, pages 464–476, 2002.
5. G. Blanc, K. Hokamp, and K.H. Wolfe. A recent polyploidy superimposed on older large-scale duplications in the arabidopsis genome. *Genome Res*, 13(2):137–44, 2003.
6. M. Blanchette, T. Kunisawa, and D. Sankoff. Gene order breakpoint evidence in animal mitochondrial phylogeny. *Journal of Molecular Evolution*, 49:193–203, 1999.
7. P Bork, B. Snel, G. Lehmann, M. Suyama, T. Dandekar, W. Lathe III, and M. Huynen. Comparative genome analysis: exploiting the context of genes to infer evolution and predict function. In D. Sankoff and J. H. Nadeau, editors, *Comparative Genomics*, pages 281–294. Kluwer Academic Press, Dordrecht, NL, 2000.
8. G. Bourque and P.A. Pevzner. Genome-scale evolution: Reconstructing gene orders in the ancestral species. *Genome Res*, 12(1):26–36, 2002.
9. P. P. Calabrese, S. Chakravarty, and T. J. Vision. Fast identification and statistical evaluation of segmental homologies in comparative maps. *ISMB (Supplement of Bioinformatics)*, pages 74–80, 2003.
10. X Chen, Z Su, P Dam, B Palenik, Y Xu, and T Jiang. Operon prediction by comparative genomics: an application to the Synechococcus sp. WH8102 genome. *Nucleic Acids Res*, 32(7):2147–2157, 2004.
11. A. Coghlan and K. H. Wolfe. Fourfold faster rate of genome rearrangement in nematodes than in *Drosophila*. *Genome Research*, 12(6):857–867, 2002.
12. M. E. Cosner, R. K. Jansen, B. M. E. Moret, L. A. Raubeson, L.-S. Wang, T. Warnow, and S. Wyman. An empirical comparison of phylogenetic methods on chloroplast gene order data in *Campanulaceae*. In D. Sankoff and J. H. Nadeau, editors, *Comparative Genomics*, pages 99–121. Kluwer Academic Press, Dordrecht, NL, 2000.
13. F. Coulier, P. Pontarotti, R. Roubin, H. Hartung, M. Goldfarb, and D. Birnbaum. Of worms and men: An evolutionary perspective on the fibroblast growth factor (FGF) and FGF receptor families. *J. Mol Evol*, 44:43–56, 1997.
14. E.G. Danchin, L.Abi-Rached, A. Gilles, and P. Pontarotti. Abstract conservation of the mhc-like region throughout evolution. *Immunogenetics*, 5(3):141–8, 2003.
15. D. Durand and D. Sankoff. Tests for gene clustering. *Journal of Computational Biology*, 10(3/4):453–482, 2003.
16. J. Ehrlich, D. Sankoff, and J.H. Nadeau. Synteny conservation and chromosome rearrangements during mammalian evolution. *Genetics*, 147(1):289–96, 1997.
17. N. El-Mabrouk, J. H. Nadeau, and D. Sankoff. Genome halving. In Springer-Verlag, editor, *Combinatorial Pattern Matching*, pages 235–250, 1998.
18. N. El-Mabrouk and D. Sankoff. The reconstruction of doubled genomes. *SIAM Journal of Computing*, 32:754–792, 2003.
19. T. Endo, T. Imanishi, T. Gojobori, and H. Inoko. Evolutionary significance of intra-genome duplications on human chromosomes. *Gene*, 205(1–2):19–27, 1997.
20. M. D. Ermolaeva, O. White, and S. Salzberg. Prediction of operons in microbial genomes. *Nucleic Acids Res*, 5(29):1216–1221, Mar 2001.

21. T.J. Gibson and J. Spring. Evidence in favour of ancient octaploidy in the vertebrate genome. *Biochem Soc Trans*, 2:259–264, Feb 2000.
22. D. Goldberg, S. McCouch, and J. Kleinberg. Algorithms for constructing comparative maps. In D. Sankoff and J. H. Nadeau, editors, *Comparative Genomics*, pages 281–294. Kluwer Academic Press, Dordrecht, NL, 2000.
23. L. A. Goldberg, P. W. Goldberg, M. S. Paterson, P. Pevzner, S. C. Sahinalp, and E. Sweedyk. The complexity of gene placement. *Journal of Algorithms*, 41(2):225–2435, 2001.
24. Graham, Knuth, and Patashnik. *Concrete Mathematics*. Addison-Wesley, 1989.
25. S. Hampson, A. McLysaght, B. Gaut, and P. Baldi. LineUp: statistical detection of chromosomal homology with application to plant comparative genomics. *Genome Res*, 13(5):999–1010, 2003.
26. S. Hannenhalli, C. Chappey, E. V. Koonin, and P. A. Pevzner. Genome sequence comparison and scenarios for gene rearrangements: A test case. *Genomics*, 30:299 – 311, 1995.
27. S. Heber and J. Stoye. Algorithms for finding gene clusters. In *Proceedings of WABI01*, Lecture Notes in Computer Science 2149, pages 254–265, 2001.
28. S. Heber and J. Stoye. Finding all common intervals of k permutations. In *Proceedings of CPM01*, Lecture Notes in Computer Science 2089, pages 207–218, 2001.
29. E. A. Housworth and J. Postlethwait. Measures of synteny conservation between species pairs. *Genetics*, 162(1):441–8, 2002.
30. A. L. Hughes. Phylogenetic tests of the hypothesis of block duplication of homologous genes on human chromosomes 6, 9, and 1. *MBE*, 15(7):854–70, 1998.
31. M. Huynen and P. Bork. Measuring genome evolution. *Proc Natl Acad Sci U S A*, 95:5849–56, 1998.
32. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(682):860–921, 2001.
33. M. Kasahara. 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(1–2):59–65, 1997.
34. N. Katsanis, J. Fitzgibbon, and E.M. Fisher. 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(1):101–8, 1996.
35. A. B. Kolsto. Dynamic bacterial genome organization. *Molecular Microbiology*, 24:241–8, 1997.
36. J.G. Lawrence and J. R. Roth. Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics*, 143:1843–60, 1996.
37. L. Lipovich, E. D. Lynch, M. K. Lee, and M-C. King. A novel sodium bicarbonate cotransporter-like gene in an ancient duplicated region: *SLC4A9* at 5q31. *Genome Biology*, 2(4):0011.1–0011.13, 2001.
38. N. Luc, J.L. Risler, A. Bergeron, and M. Raffinot. Gene teams: a new formalization of gene clusters for comparative genomics. *Comput Biol Chem.*, 27(1):59–67, 2003.
39. L. G. Lundin. Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. *Genomics*, 16(1):1–19, 1993.
40. A. McLysaght, K. Hokamp, and K. H. Wolfe. Extensive genomic duplication during early chordate evolution. *Nat Genet.*, 31(2):200–204, 2002.
41. J. H. Nadeau and B. A. Taylor. Lengths of chromosomal segments conserved since the divergence of man and mouse. *Proc.Natl.Acad.Sci. USA*, 81:814–818, 1984.
42. J.H. Nadeau and D. Sankoff. Counting on comparative maps. *Trends Genet*, 14(12):495–501, 1998.

43. J.H. Nadeau and D. Sankoff. The lengths of undiscovered conserved segments in comparative maps. *Mamm Genome*, 9(6):491–5, 1998.
44. S. J. O'Brien, J. Wienberg, and L. A. Lyons. Comparative genomics: lessons from cats. *Trends Genet*, 10(13):393–399, Oct 1997.
45. R. Overbeek, M. Fonstein, M. D'Souza, G. D. Pusch, and N. Maltsev. The use of gene clusters to infer functional coupling. *PNAS*, 96:2896–2901, 1999.
46. M.-J. Pebusque, F. Coulier, D. Birnbaum, and P. Pontarotti. Ancient large-scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. *MBE*, 15(9):1145–59, 1998.
47. Pavel A. Pevzner. *Computational Molecular Biology: An Algorithmic Approach*. MIT Press, Cambridge, MA, 2000.
48. I. Ruvinsky and L. M. Silver. Newly indentified paralogous groups on mouse chromosomes 5 and 11 reveal the age of a t-box cluster duplication. *Genomics*, 40:262–266, 1997.
49. D. Sankoff, D. Bryant, M. Deneault, B. F. Lang, and G. Burger. Early eukaryote evolution based on mitochondrial gene order breakpoints. *J Comput Biol*, 3–4:521–535, 2000.
50. D. Sankoff, M. Deneault, D. Bryant, C. Lemieux, and M. Turmel. Chloroplast gene order and the divergence of plants and algae from the normalized number of induced breakpoints. In D. Sankoff and J. H. Nadeau, editors, *Comparative Genomics*, pages 89–98. Kluwer Academic Press, Dordrecht, NL, 2000.
51. D Sankoff and N. El-Mabrouk. Genome rearrangement. In T. Jiang, T. Smith, Y. Xu, and M. Zhang, editors, *Current Topics in Computational Biology*, pages 135–155. MIT Press, 2002.
52. D. Sankoff, V. Ferretti, and J. H. Nadeau. Conserved segment identification. *Journal of Computational Biology*, 4:559–565, 1997.
53. C. Semple and K. H. Wolfe. Gene duplication and gene conversion in the *Caenorhabditis elegans* genome. *JME*, 48(5):555–64, 1999.
54. C. Seoighe and K. H. Wolfe. Updated map of duplicated regions in the yeast genome. *Gene*, 238:253–261, 1999.
55. C. Seoighe and K.H. Wolfe. Extent of genomic rearrangement after genome duplication in yeast. *Proc Natl Acad Sci U S A*, 95(8):4447–52, 1998.
56. C. Simillion, K. Vandepoele, M.C. Van Montagu, M. Zabeau, and Y. Van de Peer. The hidden duplication past of arabidopsis thaliana. *Proc Natl Acad Sci U S A*, 99(21), 2002.
57. M Skovgaard, L J Jensen, S Brunak, D Ussery, and A Krogh. On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet*, 17(8):425–428, Aug 2001.
58. N. G. C. Smith, R. Knight, and L. D. Hurst. Vertebrate genome evolution: a slow shuffle or a big bang. *BioEssays*, 21:697–703, 1999.
59. J. Spring. Genome duplication strikes back. *Nature Genetics*, 31:128–129, 2002.
60. J. Tamames. Evolution of gene order conservation in prokaryotes. *Genome Biol*, 6(2):0020.1–11, 2001.
61. J. Tamames, G. Casari, C. Ouzounis, and A. Valencia. Conserved clusters of functionally related genes in two bacterial genomes. *JME*, 44::66–73, 1997.
62. J. Tamames, M. Gonzalez-Moreno, A. Valencia, and M. Vicente. Bringing gene order into bacterial shape. *Trends Genet*, 3(17):124–126, Mar 2001.
63. Z. Trachtulec and J. Forejt. Synteny of orthologous genes conserved in mammals, snake, fly, nematode, and fission yeast. *Mamm Genome*, 3(12):227–231, Mar 2001.
64. J. V. Uspensky. *Introduction to Mathematical Probability*, pages 23–24. McGraw-Hill, New York, 1937.

65. K. Vandepoele, Y. Saeys, C. Simillion, J. Raes, and Y. Van De Peer. The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between arabidopsis and rice. *Genome Res*, 12(11):1792–801, 2002.
66. J. C. Venter et al. The sequence of the human genome. *Science*, 291(5507):1304–51, 2001.
67. T. J. Vision, D. G. Brown, and S. D. Tanksley. The origins of genomic duplications in Arabidopsis. *Science*, 290:2114–2117, 2000.
68. K. H. Wolfe and D. C. Shields. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387:708–713, 1997.
69. Yu Zheng, Joseph D Szustakowski, Lance Fortnow, Richard J Roberts, and Simon Kasif. Computational identification of operons in microbial genomes. *Genome Res*, 12(8):1221–1230, Aug 2002.

A Derivation of $d_1(m, g, r)$ from $d_2(m, g, r)$

In Section 2 we gave an expression $d_2(m, g, l)$ for the number of ways of arranging m black genes into a max-gap cluster of length *exactly* l. The number of ways $d_1(m, g, l)$ of arranging m black genes in a max-gap cluster of length *no greater* than l is as follows:

$$\sum_{r=m}^{l} d_2(m, g, r) = \sum_{r=m}^{l} \sum_{i=0}^{\lfloor (r-m)/(g+1) \rfloor} (-1)^i \binom{m-1}{i} \binom{r - i(g+1) - 2}{m-2},$$

The r in the upper bound of the second summation can be replaced by l because when $i > \lfloor (r-m)/(g+1) \rfloor$ the final binomial will be zero, which gives

$$\sum_{r=m}^{l} \sum_{i=0}^{\lfloor (l-m)/(g+1) \rfloor} (-1)^i \binom{m-1}{i} \binom{r - i(g+1) - 2}{m-2}.$$

Now the upper bound of the second summation is no longer dependent on r, and so the outer summation can be moved inward:

$$\sum_{i=0}^{\lfloor (l-m)/(g+1) \rfloor} (-1)^i \binom{m-1}{i} \sum_{r=m}^{l} \binom{r - i(g+1) - 2}{m-2}.$$

Rewriting the bounds of the inner summation gives:

$$\sum_{i=0}^{\lfloor (l-m)/(g+1) \rfloor} (-1)^i \binom{m-1}{i} \sum_{r=m-i(g+1)-2}^{l-i(g+1)-2} \binom{r}{m-2}.$$

Decreasing the lower bound to $r = 0$ does not affect the probability because when $0 \leq r < m - 2$ the binomial is zero. We apply the upper summation identity [24] to eliminate the inner summation, which yields

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

which is exactly $d_1(m, g, r)$. The derivation of $d_0(m, g, r)$ from $d_1(m, g, r)$ is identical.