

# The Incompatible Desiderata of Gene Cluster Properties

Rose Hoberman<sup>1</sup> and Dannie Durand<sup>2</sup>

<sup>1</sup> Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, USA.  
E-mail: [roseh@cs.cmu.edu](mailto:roseh@cs.cmu.edu)

<sup>2</sup> Departments of Biological Sciences and Computer Science, Carnegie Mellon University. E-mail: [durand@cmu.edu](mailto:durand@cmu.edu)

**Abstract.** There is widespread interest in comparative genomics in determining if historically and/or functionally related genes are spatially clustered in the genome, and whether the same sets of genes reappear in clusters in two or more genomes. We formalize and analyze the desirable properties of gene clusters and cluster definitions. Through detailed analysis of two commonly applied types of cluster,  $r$ -windows and max-gap, we investigate the extent to which a single definition can embody all of these properties simultaneously. We show that many of the most important properties are difficult to satisfy within the same definition. We also examine whether one commonly assumed property, which we call *nestedness*, is satisfied by the structures present in real genomic data.

## 1 Introduction

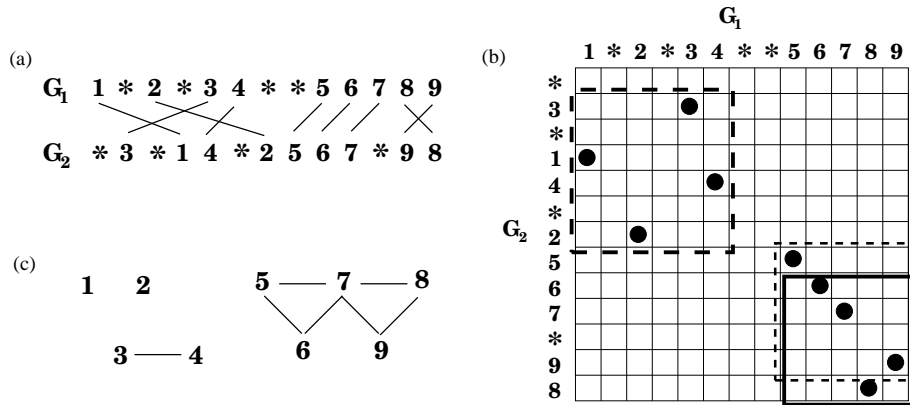
Comparisons of the spatial arrangement of genes within a genome offer insight into a number of questions regarding how complex biological systems evolve and function. Spatial analyses of orthologous genomes focus on elucidating evolutionary processes and history, and on constructing comparative maps that facilitate the transfer of knowledge between organisms [1, 2]. Conserved segments between different genomes have been used extensively to reconstruct the history of chromosomal rearrangements and infer an ancestral genetic map for a diverse group of species [3, 4], as well as to provide novel features for new phylogenetic approaches. Genome self-comparisons reveal ancient large-scale or whole-genome duplication events [5]. Finally, spatial comparative genomics can also help predict protein function and regulation. In bacteria, conserved gene order and content have been used for prediction of operons, horizontal transfers, and more generally to help understand the relationship between spatial organization and functional selection [6–11].

A prerequisite to all of these tasks is the identification of genomic regions that share a common ancestor. Although offspring genomes immediately following speciation or a whole-genome duplication will have identical gene content and order, over time large and small scale rearrangements will obscure this relationship, leading to pairs of regions, or *gene clusters*, that share a number of homologous genes, but where neither order nor gene content is strictly conserved.

To identify such diverged homologous regions it is necessary to define the spatial patterns suggestive of common ancestry, and then design a search algorithm to find such patterns. The exact definition of the structures of interest is critical for sensitive detection of ancient homologies without inclusion of false positives. It is difficult to characterize what such regions will look like, however, since in most cases evolutionary histories are not known. Consequently, cluster definitions are generally based upon intuitive notions, derived either from small, well-studied examples (*e.g.*, such as the MHC region [12–14]), or from ideas about how rearrangements of genomes proceed. However, not much is known about the rates at which different evolutionary processes occur, and the little that is known is often based (somewhat circularly) on inferred homology of chromosomal segments.

The properties underlying existing cluster definitions are generally not stated, and the dimensions along which they differ have been analyzed in only a cursory manner. As a result, the formal tradeoffs between different models have been difficult to understand or compare in a rigorous way. Most cluster definitions are constructive, in the sense that they supply an algorithm to find clusters but do not specify explicit cluster criteria. In order to verify that an algorithm will identify all clusters satisfying the underlying intuitive criteria, however, these criteria must be stated formally. A few attempts have been made to formally define a gene cluster, but in these cases the focus tends to be on the design of an efficient and correct search algorithm, rather than on selecting a definition that captures those underlying intuitions. In addition to the cluster definition, the design of the search procedure may implicitly lead to additional unexpected or even undesirable properties, which would not be detected without explicit consideration of the cluster criteria. Finally, analysis of cluster properties can be useful for determining which characteristics actually reflect the types of structures found in real genomes, and thus which will best discriminate truly homologous regions from background noise (clusters of genes that occur by chance).

The goal of this paper is to characterize desirable properties of clusters and cluster definitions, in order to develop a more rigorous understanding of how modeling choices determine the types of clusters we are able to find, and how such choices influence the statistical power of tests of segmental homology. In Section 2, we describe the formal models and definitions discussed in this work. In Section 3, we present a set of properties upon which many existing gene cluster definitions, algorithms, and statistical tests are explicitly or implicitly based. We also propose additional properties that we believe are desirable, but are rarely stated explicitly. Through detailed analysis of two commonly applied types of cluster,  $r$ -windows and max-gap, we investigate the extent to which a single definition can embody all of these properties simultaneously. In Section 4, we examine whether one property that is implicitly assumed in many analyses, which we call *nestedness*, is actually satisfied by the structures present in real genomic data.



**Fig. 1.** Three ways in which to visualize a whole-genome comparison. Integers and stars denote genes, with stars denoting genes with no homolog in the other genome. (a) A comparative map. Lines show the mapping between homologous genes. (b) A dot plot showing the same information in a matrix format. Columns represent genes in  $G_1$  and rows represent genes in  $G_2$ . A matrix element is filled with a black circle if the genes are homologous, and empty otherwise. (c) A graph in which vertices represent homologous gene pairs, and edges connect vertices if the corresponding genes are close together in both genomes. In this example, edges connect genes if the sum of the distances between the genes in both genomes is no greater than two.

## 2 Models and Cluster Definitions

### 2.1 Models

We employ a commonly used model in which a genome is represented as an ordered set of  $n$  genes:  $G = (g_1, \dots, g_n)$ . We assume a single unbroken chromosome, in which genes do not overlap. The distance between two genes in this model is simply the number of genes between them. In a *whole-genome comparison*, we are given two genomes  $G_1$  and  $G_2$ , and a mapping between homologs in  $G_1$  and  $G_2$ , where  $m$  of the genes in  $G_1$  have homologs in  $G_2$  (and vice versa). In this paper, we assume that each gene has at most one homolog in the other genome. We are interested in finding sets of homologs found in proximity in two different genomes (or possibly in two distinct regions of the same genome).

This model can be conceptualized in a number of ways, shown in Figure 1. Consider two genomes  $G_1 = 1*2*34**56789$  and  $G_2 = *3*14*2567*98$ , where the integers correspond to homologous gene pairs, and the stars indicate genes with no homolog in the other genome. Figure 1(a) shows a *comparative map* representation, in which homologous pairs are connected by a line. Alternatively, in a *dot-plot* (Figure 1(b)), the horizontal axis represents  $G_1$ , the vertical axis represents  $G_2$ , and homologous pairs are represented as dots in the matrix. Finally, this data can be converted into an *undirected graph* (Figure 1(c)), where vertices correspond to homologous gene pairs. Two vertices are connected by an edge if the corresponding genes are close together in both genomes, where “close” is determined based on a user-defined distance function and threshold.

## 2.2 Cluster Definitions

A number of cluster definitions and algorithms have been proposed. In this paper we primarily focus on  $r$ -windows and max-gap clusters, two cluster definitions that are used in practice [6, 7, 9, 10, 15–17], but briefly describe other definitions as well.

An  $r$ -window cluster is defined as a pair of windows of  $r$  genes, one in each genome under consideration, in which at least  $k$  genes are shared [15, 18, 19]. This corresponds to a square in the dot-plot with sides of length  $r$ , which contains at least  $k$  homologs. For example, for a window model with  $r = 5$  and  $k = 4$ , two clusters can be found in the example genome in Figure 1(b):  $\{5, 6, 7, 9\}$  (dotted box) and  $\{6, 7, 8, 9\}$  (solid box). We distinguish between the homologs shared in both instances of the cluster (the “marked” genes) and the intervening “unmarked” genes that occur in only one instance of the cluster (but which may have a homolog elsewhere in the genome).

The *max-gap* cluster definition also ignores gene order and allows insertions and deletions, but does not constrain the maximum length of the cluster to  $r$  genes. Instead, a max-gap cluster is described by a single parameter  $g$ , and is defined as a set of marked genes where the distance (or *gap*) between adjacent marked genes in each genome is never larger than a given distance threshold,  $g$  [20, 21]. When  $g = 0$ , max-gap clusters are referred to as *common intervals* [22–24]. When the maximum gap allowed is  $g = 1$ , two maximal max-gap clusters are found in the example genome in Figure 1(b):  $\{1, 2, 3, 4\}$  (dashed box) and  $\{5, 6, 7, 8, 9\}$  (not shown). A max-gap cluster is *maximal* if it is not contained within any larger max-gap cluster. Correct search algorithms for this definition require some sophistication. Bergeron *et al.* originally developed a divide-and-conquer algorithm to conduct a whole-genome comparison, and efficiently detect all maximal max-gap clusters [20]. Many groups design heuristics to find max-gap clusters, but such methods are not guaranteed to find all maximal max-gap clusters.

Other definitions include that of Calabrese *et al.* [25], in which the distance between each pair of homologs is evaluated as a function of the gap size in *both* genomes. Unlike the max-gap definition, which only requires that in both genomes the distance to *some* other marked gene in the cluster is small, this method requires that all marked genes that are adjacent in genome  $G_1$  *also* be close in genome  $G_2$ , but not vice versa. A very different approach by Sankoff *et al.* [26] explicitly evaluates a cluster (or segment) by a weighted measure of three properties: compactness, density, and integrity. They seek a global partition of the genome into segments such that the sum of segment scores is minimized. Clusters have also been defined in terms of graph-theoretic structures (*e.g.*, Figure 1(c)), such as connected components [27] or high-scoring paths [28, 29]. Finally, a variety of heuristics have been proposed to search for gene clusters [11, 25, 29–34], the majority of which are specifically designed to find sets of genes in approximately collinear order (*i.e.*, forming a rough diagonal on the dot-plot).

### 3 Cluster Properties

Many of the cluster properties underlying existing definitions derive from the processes that lead to genome rearrangements. As genomes diverge, large-scale rearrangements break apart homologous regions, reducing the size and length of clusters. Gene duplications and losses cause the gene complement of homologous regions to drift apart, so that many genes will not have a homolog in the other region, and gene clusters will appear less dense. Smaller rearrangements will disrupt the gene order and orientation within homologous regions. Thus, clusters are often characterized according to their size, length, density, and the extent to which order and orientation are conserved. We discuss these properties in more detail below, as well as a number of additional properties that are rarely stated explicitly, but that we argue are nonetheless desirable.

**Size:** Almost all methods to evaluate clusters consider the size of a cluster, *i.e.*, the number of marked genes contained within it. In general it is assumed that the more homologs in a cluster, the more likely it is to indicate common ancestry rather than chance similarities. An appropriate minimum size threshold will depend, however, on the specific cluster definition. For example, a cluster of four homologs in which order is conserved may be less likely to occur by chance, and thus more significant than an unordered cluster of size four.

**Length:** The length of a cluster, defined with respect to a particular genome, is the total number of marked and unmarked genes contained within it. For example, in Figure 1(b), the upper left cluster is of size four, and spans two unmarked genes, so is of total length six. In a whole-genome comparison, the number of unmarked genes spanned by the cluster in each genome may differ. However, if the processes that degrade a cluster are operating uniformly, then the length of the cluster in both genomes should be similar. This similarity of lengths is implicitly sought by the length constraint of  $r$ -windows, and explicitly sought in the clustering method of Hampson *et al.* [33].

**Density:** Although over time gene insertions and losses will cause the gene content of homologous regions to diverge, in most cases we expect that significant similarity in gene content will be preserved. Thus, the majority of existing approaches attempt to find regions that are densely populated with homologs. We define the *global density* of a cluster as its size divided by its length. For example, in Figure 1(b), the first max-gap cluster is of size four and length six, so has a density of  $2/3$ . For a fixed value of  $r$ , the minimum global density of an  $r$ -window is set by choosing the parameter  $k$ . The only way to set a constraint on the global density of a max-gap cluster, on the other hand, is to reduce  $g$ , which will also reduce the maximum length of a cluster.

Even when a minimum global density is required, regions of a cluster may not be locally dense: a cluster could be composed of two very dense regions separated by a large region with no homologs. In this case, it might seem more natural to break the cluster into two separate clusters. Density as we have defined it here reflects the average gap size, but does not reflect the *variance* in gap sizes. The gap between adjacent marked genes in an  $r$ -window can be as large as  $r-k$ , whereas max-gap clusters guarantee that the maximum gap will be no more

than  $g$ . Note that the two definitions have switched roles: the local density is easily controlled by the parameter  $g$  for max-gap clusters but there is no way to constrain the local density of  $r$ -window clusters without also further constraining the maximum cluster length. This trade-off between global and local density gives a simple illustration of how it can be difficult to design a cluster definition that satisfies our basic intuitions about cluster properties.

**Order:** For whole-genome comparison, a cluster is considered ordered if the homologs in the second genome are in the identical or opposite order of the homologs in the first genome. For example, consider the two genomes shown in Figure 1(b). The clusters  $\{5, 6, 7\}$  and  $\{8, 9\}$  are ordered, but  $\{1, 2, 3, 4\}$  is not. Many cluster definitions require a strictly conserved gene order [6, 11, 31]. Over time, however, inversions will cause rearrangements, and thus conserved gene order is often considered too strict a requirement. In order to allow some short inversions, Hampson *et al.* [32] explicitly parameterize the number of order violations that are allowed in a cluster. A number of groups use heuristic, constructive methods that either implicitly enforce certain constraints on gene order, or explicitly bias their method to prefer clusters that form near-diagonals in the dot plot [17, 25, 29, 34]. The remainder, including  $r$ -windows and max-gap clusters, completely disregard gene order. As we will see, however, though a number of groups *state* that they ignore gene order, constraints on gene order are often unintended consequences of algorithmic choices (see nestedness).

**Orientation:** Conserved spatial organization in bacterial genomes often points to functional associations between genes. In particular, clusters of genes in close proximity, with the same orientation, often indicate operons. In whole-genome comparison of eukaryotes, similarities in gene orientation can provide additional evidence that two regions share a common ancestor. To the best of our knowledge, however, except for the method of Vision *et al.* [29], in which changes in orientation decrease the cluster score, existing definitions either require all genes in a cluster to have the same orientation, or disregard orientation altogether.

**Temporal Coherence:** Temporal information can be used to evaluate the significance of a putative homologous region identified through whole-genome comparison. If a set of homologous genes all arose through the same speciation or duplication event, then the points in time at which each homolog pair diverged will be identical, and consequently we would expect our estimates of these divergence times to be similar. However, all existing methods to find clusters are based solely on spatial information, and divergence times have been used only to estimate the age of a duplicated block identified based on spatial organization [6, 35], but not to assess the statistical significance of a cluster. In theory, combined analysis of temporal and spatial information could be used, for example, to increase our confidence that a region is the result of a single large-scale duplication event. However, due to the large error bounds that must be associated with any sequence-based estimate of divergence times [36–38], the practicality of such an approach is as yet unclear.

**Nestedness:** For whole-genome comparison, one cluster property that is generally not considered explicitly, but may be assumed implicitly, is nestedness.

A cluster of size  $k$  is *nested* if for each  $h \in 1 \dots k - 1$  it contains a valid cluster of size  $h$ . Intuitively it may seem that any reasonable cluster definition should have this property. In fact, clusters with no ordering constraints are not necessarily nested. For example, Bergeron *et al.* [20] state a formal definition of max-gap clusters, and prove that there are maximal max-gap clusters of size  $k$  which do not contain any valid sub-cluster of size  $2..k-1$ . For example, when  $g = 0$  they present a non-nested max-gap cluster with only four genes. The sequence of genes 1234 on one genome and 3142 on the other form a max-gap cluster of size four which does not contain any max-gap cluster of size two or three. Thus, nested max-gap clusters comprise only a subset of general max-gap clusters found through whole-genome comparison.

There are no definitions that explicitly require that clusters be nested; rather, greedy search algorithms implicitly limit the results to nested clusters. Greedy algorithms use a bottom-up approach: each homologous gene pair serves as a cluster seed, and a cluster is extended by looking in its chromosomal neighborhood for another homologous gene pair close to the cluster on both genomes [25, 31, 33, 39]. It can be shown that any greedy search algorithm that constructs max-gap clusters iteratively, *i.e.*, by constructing a cluster of size  $k$  by adding a gene to a cluster of size  $k - 1$ , will find *exactly* the set of all maximal nested max-gap clusters, as long as it considers each homologous gene pair as a seed for a potential cluster. In such cases, although order is not explicitly constrained, the search algorithm enforces implicit constraints on gene order: nested clusters can only get disordered to a limited degree. In most cases, however, such constraints are not acknowledged, and perhaps not even recognized.

**Disjointness:** If two clusters are not disjoint, *i.e.*, the intersection of the marked genes they contain is not empty<sup>3</sup>, our intuitive notion of a cluster may correspond more closely to the single island of overlapping windows than to the individual clusters. For example, Figure 1(b) shows two windows for which  $r = 5$  and  $k = 4$ :  $\{5, 6, 7, 9\}$  and  $\{6, 7, 8, 9\}$ . Although both clusters contain genes 6, 7, and 9, there is no window of length five that contains all five of the genes. Thus,  $r$ -windows are not always disjoint. Indeed, it is surprisingly hard to find a cluster definition that guarantees that all clusters will be disjoint. The majority of definitions lead to overlapping clusters that must be merged or separated in an ad-hoc post-processing step for use by algorithms that require a unique tiling of regions. The only definition for which maximal clusters have been shown to be disjoint is the max-gap cluster [20]. However, when adding additional constraints in addition to the maximum gap size, disjointness is quickly forfeited. For example, consider the consequences of requiring conserved order when looking for max-gap clusters in Figure 1(a). With a maximum gap of  $g = 2$ , three clusters with conserved order are identified  $\{1, 2\}$ ,  $\{3, 4, 5, 6, 7, 8\}$ ,  $\{3, 4, 5, 6, 7, 9\}$ . Although the last two clusters overlap, they cannot be merged without breaking the ordering constraint (due to the inversion of the segment containing genes 8 and 9).

---

<sup>3</sup> Note that it is possible, however, for two disjoint clusters to have overlapping spans in one of the genomes, as long as they do not share any homologs.

$G_1$	$G_2$	$n_1$	$n_2$	$m$
<i>E. coli</i>	<i>B. subtilis</i>	4, 108	4, 245	1, 315
Human	Mouse	22, 216	25, 383	14, 768
Human	Chicken	22, 216	17, 709	10, 338

**Table 1.** The genomes compared ( $G_1$  and  $G_2$ ), the total number of genes in each genome ( $n_1$  and  $n_2$ , respectively), and the number of orthologs identified, excluding ambiguous orthologs ( $m$ ).

More generally, a lack of disjointness strongly suggests that the cluster definition is too constrained. In the  $r$ -window example, these clusters are not disjoint *precisely* because the definition artificially constrains the length of a cluster. In the second example, the clusters were not disjoint because a definition with a strict ordering constraint was not able to capture the types of processes, such as inversions, that created the cluster.

**Isolation:** If we observe a cluster with some additional homologous pairs in close proximity to its borders we might feel that the cluster border was arbitrary, and should extend to cover the neighboring island of genes. Thus, we propose that cluster definitions should guarantee that clusters will be *isolated*, that is: the maximum distance between marked genes in a cluster should always be less than the minimum distance between two clusters. A maximum-gap constraint guarantees that clusters will be isolated, but only barely—the gap within a cluster may be as large as  $g$ , whereas the gap separating two clusters may be just  $g+1$ .

**Symmetry:** For whole-genome comparison, a desirable property that is rarely considered explicitly is whether the definition is symmetric with respect to genome. In some cases, such as the definition proposed by Calabrese *et al.* [25], a cluster is defined in such a way that whether a set of genes form a valid cluster may depend on whether genome  $G_1$  or genome  $G_2$  is represented by the vertical axis in the dot-plot. Put another way, the set of clusters identified will differ depending on which genome is designated as the reference genome. A surprisingly large proportion of constructive definitions are not symmetric. These clustering algorithms require the selection of a reference genome even when there is no clear biological motivation for this choice. Definitions that are symmetric with respect to genome include  $r$ -windows and max-gap cluster definitions, as well as algorithms that represent the dot-plot as a graph and use a symmetric distance function [27, 29].

## 4 Are Max-Gap Clusters in Genomic Data Nested?

Cluster definitions that constrain the gap size between marked genes are widely used in genomic studies [6, 7, 9, 10, 16, 17, 29, 30, 34, 40, 41]. In the majority of cases, however, clusters are detected with a greedy algorithm, whereby larger clusters are identified by extending smaller clusters. Remember that greedy methods find the subset of max-gap clusters that are nested and that nestedness implies



```

for i= 1..n do    // i iterates through all genes in G1
  C = {i};      // C is the cluster being constructed
  L1 = R1 = i; // Li and Ri are the left/rightmost positions in C on Gi
  L2 = R2 = p(i); // p(i) indicates the position of gene i's homolog in G2
  j = L1-g-1; // j iterates through all genes close to C on G1
  while (L1-g-1 ≤ j ≤ R1+g+1) do
    if j ∉ C and p(j) ∈ {L2-g-1, ..., R2+g+1} // if j is close to C in G2
      C = C ∪ j; // add it to C
      L1 = min(L1, j); L2 = min(L2, p(j));
      R1 = max(R1, j); R2 = max(R2, p(j));
      j = L1-g-1; // start the search over
    else
      j++;
    end
  end
  clusters = clusters ∪ C;
end

```

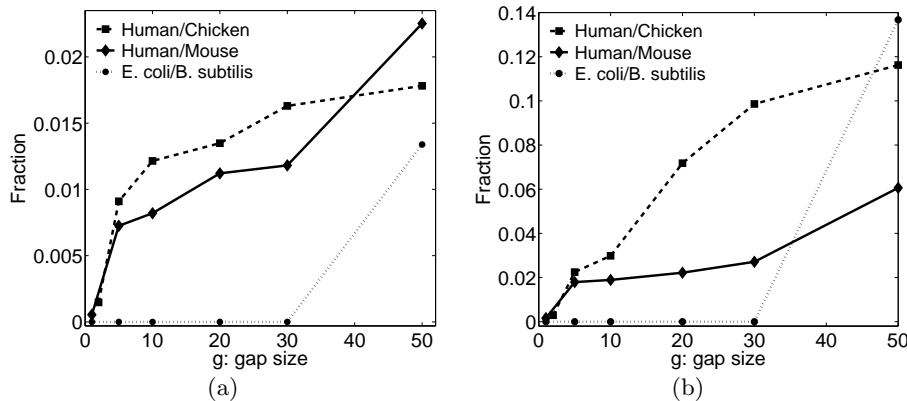
**Fig. 2.** Pseudo-code for a greedy, bottom-up algorithm to find nested max-gap clusters.

a certain degree of ordering. It is not clear whether greedy methods are used for computational convenience or because researchers believe that nested clusters better capture the biological processes of interest. In this section, we investigate the practical consequences of choosing one search procedure over the other. We compare three pairs of genomes to determine the proportion of max-gap clusters in real genomes that are actually nested.

Whole-genome comparisons of three pairs of genomes at varying evolutionary distances were conducted. The first comparison was of *E. coli* and *B. subtilis*, with a mapping of orthologs between the two genomes obtained from the GOLDIE database [30]. The other two comparisons were of human and mouse, and human and chicken, with ortholog mappings obtained from the InParanoid database [42]. The total number of genes in each genome, and the number of orthologs identified, is given in Table 1.

The GeneTeams software, an implementation of the top-down algorithm of Bergeron *et al.* [20], was used to identify all maximal max-gap clusters shared between the two genomes, for  $g \in \{1, 5, 10, 15, 20, 30, 50\}$ . In addition, we designed a simple bottom-up, greedy algorithm to identify all maximal *nested* max-gap clusters (Figure 2). This algorithm considers each pair of orthologs in turn, treating each as a cluster seed from which a greedy search for additional orthologs is initiated. Occasionally different seeds may yield identical clusters. Any such duplicate clusters are filtered out, as are non-maximal nested clusters (clusters strictly contained within another nested cluster). However, overlapping clusters (*e.g.*, properly intersecting sets) are not merged together, since the resulting merged clusters would not be nested.<sup>4</sup>

<sup>4</sup> It is unclear whether those who employ a greedy heuristic merge all overlapping clusters or not, since such heuristics are generally specified quite vaguely, if at all.



**Fig. 3.** Comparison of the set of nested clusters to the set of gene teams, for for  $g \in \{1, 5, 10, 15, 20, 30, 50\}$ . (a) The fraction of gene teams that are nested. (b) The fraction of maximal nested clusters that are *not* gene teams.

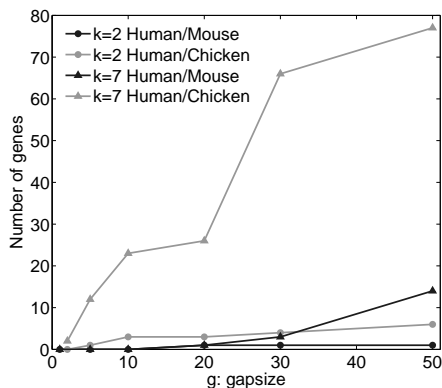
For the bacterial comparison, for all gap values except  $g=50$ , both methods found the same set of clusters, *i.e.*, all gene teams were nested. In all eukaryotic comparisons, however, at least one non-nested gene team was identified. Nonetheless, the percentage of teams that were not nested remained low for all comparisons, ranging from close to 0% to about 2% as the gap size was increased (Figure 3(a)). The percentage of nested clusters that were not gene teams (in other words, clusters that could have been extended further if a greedy algorithm had not been used), was also close to zero for small gap sizes, but increased more quickly, peaking at almost 15% for a gap size of  $g=50$  (Figure 3(b)). In contrast, in randomly ordered genomes, although large gene-teams are much rarer, a much higher percentage are not nested (data not shown).

Another quantity of interest is the number of *genes* that would be missed altogether if a greedy approach is used rather than a top-down algorithm; that is, the number of genes that are found in a large gene team but not in a large nested cluster. For a minimum cluster size of two, very few genes are missed: the number of genes missed remains under 20 for both eukaryotic datasets, no matter how large the gap size (Figure 4, circles). For a more realistic minimum cluster size of seven, however, the number of missed genes rises more quickly, peaking near 80 for the human/chicken comparison (Figure 4, triangles), and near 120 for the bacterial comparison (data not shown).

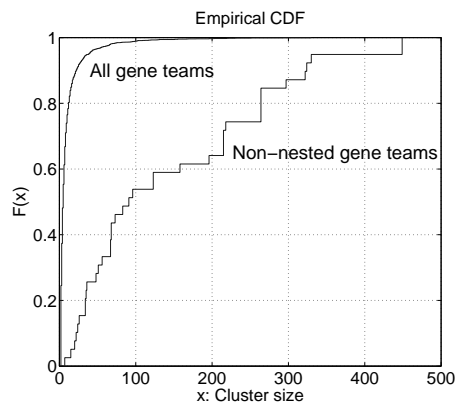
The gene teams that are not nested tend to be the larger clusters. For example, Figure 5 compares the distribution of gene teams sizes to the distribution of non-nested gene teams sizes, for the human/chicken comparison, for the complete set of clusters identified at any gap size. The gene team size distribution

---

However, in our datasets, only a small percentage of clusters detected with the greedy algorithm overlapped (*e.g.*, 2% in the human/chicken comparison).



**Fig. 4.** The number of genes in a gene team of size  $k \geq 2$ , that are not in *any* nested max-gap cluster of size  $k \geq 2$  (circles). The triangles show the number of genes when the the minimum cluster size is seven.



**Fig. 5.** A CDF comparing the distribution of gene team sizes to the distribution of nested gene team sizes, for human vs chicken, for all gap sizes tested.

peaks very quickly: over 80% of gene teams contain fewer than ten genes. The sizes of non-nested gene teams, however, peak much more slowly: only about 10% of non-nested gene teams contain fewer than ten genes. It is not until the size reaches 270 genes that the CDF reaches 0.8.

In summary, when comparing *E. coli* with *B. subtilis* with reasonable gap sizes, the nestedness assumption does not exclude any clusters from the data. For the eukaryotic datasets, these results also suggest that for smaller gap sizes few clusters are missed when using a greedy search strategy. For larger gap values, the nestedness assumption does appear to lead to some loss of signal, especially in the human/chicken comparison: large clusters are identified only in fragments, and the spatial clustering of many genes is not detected at all. For more diverged genome pairs, as clusters become more disordered, this loss of signal may be exacerbated. This remains to be investigated, as do the practical implications of the nestedness assumption on the detection of duplicated segments through genome self-comparison.

## 5 Discussion

We have characterized desirable properties of cluster definitions, and compared a number of existing definitions with respect to these properties. The detailed catalog of cluster properties presented here will be useful for assessing whether definitions satisfy the intuitive notions upon which they are implicitly based, and whether these notions actually correspond to the types of structures present in real-genomic data. Analyses of desirable cluster properties may also pave the way for new, possibly more powerful cluster definitions.

Our analysis of cluster properties reveals that existing approaches to identifying gene clusters differ both in terms of the characteristics of the clusters they were explicitly designed to find, and in terms of the properties that emerge as unintended consequences of modeling choices. We show that the search procedure, in addition to the cluster definition, often implicitly enforces additional types of constraints. Such implicit constraints may be particularly problematic when the goal is to characterize the properties of homologous regions. For example, although the CloseUp algorithm was ostensibly designed to identify chromosomal homology using “shared-gene density alone” [33], the greedy nature of the search algorithm means that all clusters with a minimum gene density may not actually be detected. If such an approach was used to evaluate the extent to which order is conserved in homologous regions, incorrect inferences could be made. For example, if clusters with highly scrambled gene order were not found, one might erroneously conclude that no such clusters exist, rather than that the clustering algorithm was simply not capable of finding them. Without a clear understanding of which properties are constrained by the method, and which properties are inherent in the data, it can be difficult to interpret such results.

Our results also show that, for the datasets considered here, a greedy search strategy for max-gap clusters may actually improve statistical power, at least for small gap sizes. A test of cluster significance will have increased power (*i.e.*, a reduced number of false negatives) when the cluster definition is as narrow as possible, while still capturing the properties exhibited by diverged homologous regions. These properties, however, are generally not known, since there is little data about evolutionary histories or processes. In some cases, however, the appropriateness of a particular property can be evaluated even without full knowledge of evolutionary histories. For example, if adding an additional constraint to the cluster definition does not eliminate any of the clusters identified in the data, then we argue that it is not only acceptable to include such a property in the cluster definition, but desirable, in order to increase statistical power. Thus, when comparing *E. coli* with *B. subtilis* with reasonable gap sizes, a nested cluster definition appears to be a good choice: the nestedness assumption does not exclude any clusters from the data, but significantly reduces the probability of observing a cluster by chance, thereby strengthening the measurable significance of detected clusters.

These results also suggest that in the three datasets we studied most clusters remain quite ordered. Although an assumption of nestedness does implicitly constrain gene order, more quantitative measures of order conservation may be found that increase statistical power still further. How to best quantify the degree to which order is conserved, however, remains an open question.

Although there is often overlap among the properties of different definitions, there is as yet no consensus on what criteria best reflect biologically important features of gene clusters. This lack of consensus reflects the sparsity of data about evolutionary histories and evolutionary processes, and also that the relevance of particular properties depends to a large degree on the dataset being analyzed, as well as the researcher’s goals. For example, physical distances between genes

and gene orientation may not be very helpful for identifying homology between eukaryotic genomes, but may be important for identifying functional clusters in bacteria. For identifying gene duplications, which are often followed by significant differential gene loss of the homologs on each duplicated segment [43], gene density may be of reduced importance than for identifying paralogous segments. In addition, when clusters are being identified as a pre-processing step for reconstructing rearrangement histories, the exact boundaries and sizes of the cluster may be quite important [44]. In other cases, a researcher may be trying to test a global hypothesis (such as finding evidence for one or two rounds of whole-genome duplication), and may not necessarily care about the significance or boundaries of any specific cluster.

Even if it were known which properties reflect biologically relevant features, designing a definition to satisfy those properties may not be straightforward because, in many cases, properties are not independent. Properties may interact in subtle ways—a definition that guarantees one desirable property will often fail to satisfy another. For example, one of the nice properties of the max-gap definition is that clusters are always disjoint. However, as shown in Section 3, adding additional constraints on order or length results in clusters that are no longer guaranteed to be disjoint. The subtle and sometimes undesirable interplay of some of these properties makes it difficult to devise a definition that satisfies them all. In fact, many of the most important properties are difficult to satisfy with the same definition. Thus, it remains an open question to what extent a single definition can capture all of these properties simultaneously.

## Acknowledgment

D.D. was supported by NIH grant 1 K22 HG 02451-01 and a David and Lucille Packard Foundation fellowship. 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 and N. Raghupathy for comments on the manuscript, and David Sankoff for helpful discussion and for suggesting the title of the paper.

## References

1. Murphy, W.J., Pevzner, P.A., O'Brien, S.J.: Mammalian phylogenomics comes of age. *Trends Genet* **20** (2004) 631–9
2. O'Brien, S.J., Menotti-Raymond, M., Murphy, W.J., Nash, W.G., Wienberg, J., Stanyon, R., Copeland, N.G., Jenkins, N.A., Womack, J.E., Graves, J.A.M.: The promise of comparative genomics in mammals. *Science* **286** (1999) 458–81
3. Sankoff, D.: Rearrangements and chromosomal evolution. *Curr Opin Genet Dev* **13** (2003) 583–7
4. Sankoff, D., Nadeau, J.H.: Chromosome rearrangements in evolution: From gene order to genome sequence and back. *PNAS* **100** (2003) 11188–9
5. Simillion, C., Vandepoele, K., de Peer, Y.V.: Recent developments in computational approaches for uncovering genomic homology. *Bioessays* **26** (2004) 1225–35

6. Blanc, G., Hokamp, K., Wolfe, K.H.: A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res* **13** (2003) 137–144
7. Chen, X., Su, Z., Dam, P., Palenik, B., Xu, Y., Jiang, T.: Operon prediction by comparative genomics: an application to the *Synechococcus* sp. WH8102 genome. *Nucleic Acids Res* **32** (2004) 2147–2157
8. Lawrence, J., Roth, J.R.: Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* **143** (1996) 1843–60
9. Overbeek, R., Fonstein, M., D’Souza, M., Pusch, G.D., Maltsev, N.: The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A* **96** (1999) 2896–2901
10. Tamames, J.: Evolution of gene order conservation in prokaryotes. *Genome Biol* **6** (2001) 0020.1–11
11. Wolf, Y.I., Rogozin, I.B., Kondrashov, A.S., Koonin, E.V.: Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res* **11** (2001) 356–72
12. Endo, T., Imanishi, T., Gojobori, T., Inoko, H.: Evolutionary significance of intra-genome duplications on human chromosomes. *Gene* **205** (1997) 19–27
13. Smith, N.G.C., Knight, R., Hurst, L.D.: Vertebrate genome evolution: a slow shuffle or a big bang. *BioEssays* **21** (1999) 697–703
14. Trachtulec, Z., Forejt, J.: Synteny of orthologous genes conserved in mammals, snake, fly, nematode, and fission yeast. *Mamm Genome* **3** (2001) 227–231
15. Friedman, R., Hughes, A.L.: Gene duplication and the structure of eukaryotic genomes. *Genome Res* **11** (2001) 373–81
16. Luc, N., Risler, J., Bergeron, A., Raffinot, M.: Gene teams: a new formalization of gene clusters for comparative genomics. *Comput Biol Chem* **27** (2003) 59–67
17. McLysaght, A., Hokamp, K., Wolfe, K.H.: Extensive genomic duplication during early chordate evolution. *Nat Genet* **31** (2002) 200–204
18. Cavalcanti, A.R.O., Ferreira, R., Gu, Z., Li, W.H.: Patterns of gene duplication in *Saccharomyces cerevisiae* and *Caenorhabditis elegans*. *J Mol Evol* **56** (2003) 28–37
19. Durand, D., Sankoff, D.: Tests for gene clustering. *Journal of Computational Biology* (2003) 453–482
20. Bergeron, A., Corteel, S., Raffinot, M.: The algorithmic of gene teams. In Gusfield, D., Guigo, R., eds.: WABI. Volume 2452 of Lecture Notes in Computer Science. (2002) 464–476
21. Hoberman, R., Sankoff, D., Durand, D.: The statistical significance of max-gap clusters. In Lagergren, J., ed.: Proceedings of the RECOMB Satellite Workshop on Comparative Genomics, Bertinoro, Lecture Notes in Bioinformatics, Springer Verlag (2004)
22. Didier, G.: Common intervals of two sequences. In: WABI. Volume 2812., Lecture Notes in Computer Science (2003) 17–24
23. Heber, S., Stoye, J.: Algorithms for finding gene clusters. In: WABI. Volume 2149 of Lecture Notes in Computer Science. (2001) 254–265
24. Uno, T., Yagiura, M.: Fast algorithms to enumerate all common intervals of two permutations. *Algorithmica* **26** (2000) 290–309
25. Calabrese, P.P., Chakravarty, S., Vision, T.J.: Fast identification and statistical evaluation of segmental homologies in comparative maps. *ISMB (Supplement of Bioinformatics)* (2003) 74–80
26. Sankoff, D., Ferretti, V., Nadeau, J.H.: Conserved segment identification. *Journal of Computational Biology* **4** (1997) 559–565
27. Pevzner, P., Tesler, G.: Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Res* **13** (2003) 37–45

28. Haas, B.J., Delcher, A.L., Wortman, J.R., Salzberg, S.L.: DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics* **20** (2004) 3643–6
29. Vision, T.J., Brown, D.G., Tanksley, S.D.: The origins of genomic duplications in *Arabidopsis*. *Science* **290** (2000) 2114–2117
30. Bansal, A.K.: An automated comparative analysis of 17 complete microbial genomes. *Bioinformatics* **15** (1999) 900–908 <http://www.cs.kent.edu/~arvind/orthos.html>.
31. Cannon, S.B., Kozik, A., Chan, B., Michelmore, R., Young, N.D.: DiagHunter and GenoPix2D: programs for genomic comparisons, large-scale homology discovery and visualization. *Genome Biol* **4** (2003) R68
32. Hampson, S., McLysaght, A., Gaut, B., Baldi, P.: LineUp: statistical detection of chromosomal homology with application to plant comparative genomics. *Genome Res* **13** (2003) 999–1010
33. Hampson, S.E., Gaut, B.S., Baldi, P.: Statistical detection of chromosomal homology using shared-gene density alone. *Bioinformatics* **21** (2005) 1339–48
34. Vandepoele, K., Saeys, Y., Simillion, C., Raes, J., Peer, Y.V.D.: The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between *Arabidopsis* and rice. *Genome Res* **12** (2002) 1792–801
35. Raes, J., Vandepoele, K., Simillion, C., Saeys, Y., de Peer, Y.V.: Investigating ancient duplication events in the *Arabidopsis* genome. *J Struct Funct Genomics* **3** (2003) 117–29
36. Graur, D., Martin, W.: Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. *Trends Genet* **20** (2004) 80–6
37. Nei, M., Kumar, S.: *Molecular Evolution and Phylogenetics*. Oxford University Press (2000)
38. Zhang, L., Vision, T.J., Gaut, B.S.: Patterns of nucleotide substitution among simultaneously duplicated gene pairs in *Arabidopsis thaliana*. *Mol Biol Evol* **19** (2002) 1464–73
39. Hokamp, K.: *A Bioinformatics Approach to (Intra-)Genome Comparisons*. PhD thesis, University of Dublin, Trinity College (2001)
40. Bourque, G., Zdobnov, E., Bork, P., Pevzner, P., Telser, G.: Genome rearrangements in human, mouse, rat and chicken. *Genome Research* (2004)
41. Simillion, C., Vandepoele, K., Montagu, M.V., Zabeau, M., de Peer, Y.V.: The hidden duplication past of *Arabidopsis thaliana*. *PNAS* **99** (2002) 13627–32
42. O’Brien, K.P., Remm, M., Sonnhammer, E.L.L.: Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res* **33** (2005) D476–80 Version 4.0, downloaded May 2005.
43. Lynch, M., Conery, J.S.: The evolutionary fate and consequences of duplicate genes. *Science* **290** (2000) 1151–1155
44. Trinh, P., McLysaght, A., Sankoff, D.: Genomic features in the breakpoint regions between syntenic blocks. *Bioinformatics* **20 Suppl 1** (2004) I318–I325