Tests for Gene Clustering

Dannie Durand
Department of Biological Sciences
Carnegie Mellon University
durand@cmu.edu

David Sankoff
Centre de recherches mathématiques
Université de Montréal
sankoff@courrier.umontreal.ca

ABSTRACT

Comparing chromosomal gene order in two or more related species is an important approach to studying the forces that guide genome organization and evolution. Linked clusters of similar genes found in related genomes are often used to support arguments of evolutionary relatedness or functional selection. However, as the gene order and the gene complement of sister genomes diverge progressively due to large scale rearrangements, horizontal gene transfer, gene duplication and gene loss, it becomes increasingly difficult to determine whether observed similarities in local genomic structure are indeed remnants of common ancestral gene order, or are merely coincidences.

A rigorous comparative genomics requires principled methods for distinguishing chance commonalities, within or between genomes, from genuine historical or functional relationships. In this paper, we construct tests for significant groupings against null hypotheses of random gene order, taking incomplete clusters, multiple genomes and gene families into account. We consider both the significance of individual clusters of pre-specified genes, and the overall degree of clustering in whole genomes.

1. INTRODUCTION

Comparison of gene order and content in related genomes is a rich source of information concerning genome evolution and function. The biology literature is rife with articles in which local similarities in two or more genomes are presented as evidence of evolutionary relatedness or functional selection on gene order. To be convincing, such reports should reject the hypothesis that the observed similarities could have occurred by chance, yet many of those reports present no statistical analysis and those that do usually rely on intuitive criteria, *ad hoc* tests or, at best, randomization simulations. Very few formal probabilistic analyses of gene clustering have been presented and there is no consensus among them on what criteria best reflect biologically important features of gene clusters.

RECOMB '02 Washington, DC USA

Biological background and significance: Speciation results in offspring genomes that initially have identical gene content and order. Similarly, whole genome duplication creates a new genome with two identical copies of the ancestral genome embedded in it. In both cases, the gene complement and gene order of the offspring genomes will diverge over time. Gene duplication and loss and horizontal gene transfer result in changes in gene complement, while gene order is disrupted by large scale rearrangements, including translocation, transposition, inversion and chromosome fission and fusion. Intuitively, rearrangement processes should result in a pattern of *conserved segments*, pairs of chromosomal regions, one in each genome, that are descended from a single, contiguous region in the ancestral genome. Because rearrangements may involve arbitrarily long chromosomal fragments, conserved segments that are adjacent in one genome will not necessarily be close to each other in the sister genomes.

In the absence of selective pressure on gene order, successive rearrangement will lead to randomization of gene order. Therefore, similarity in genomic organization is a source of evidence for inferring evolutionary relationships and/or for predicting the functional roles of gene clusters. For example, comparison of genetic maps of two or more species has been used to infer patterns of chromosomal rearrangement [10, 35, 37, 52] and as a basis for alternative approaches to phylogeny reconstruction [4, 8, 18, 46, 47, 58]. Comparison of a genetic map from a single species with itself has been used to analyze patterns of gene duplication in genome evolution [2, 11, 12, 50, 53, 61, 62, 64]. Interest in such questions has spawned a growing body of research in algorithms for inferring the history of rearrangements (see, for example, [42, 48] for surveys.) In microbial genomics, comparisons of gene content and order have also been used to study the importance of spatial organization in genome function including functional selection [23, 27, 40, 57, 56], operon formation [5, 14] and horizontal transfer [29].

The identification of conserved segments is a basic building block of all these analyses. According to the most stringent definitions, conserved segments are defined to be two or more contiguous regions that contain the same genes in the same order [34, 38] and, in some cases, in the same orientation [40, 56, 64]. However, it is common practice in indicating conserved segments in comparative genomic maps to disregard small deviations from strict conservation of gene order [17, 49]. For example the human-mouse comparison in [24] indicates only around 200 segments, many of which are known to contain small inversions and other inconsistencies.

For studies that focus on large scale genome organization and rearrangements, more generally defined *gene clusters* are the units of interest. Under some rearrangement regimes (e.g. short inversions, single gene insertion, loss or duplication), a high degree of gene proximity is conserved, even while gene order is rapidly scrambled [45]. The notion of conserved segment loses significance when it is reduced to span one or two genes (e.g., [28]). The strict definition of conserved segments is also inappropriate in the presence of positional errors. These observations have led to formalizations of a more flexible concept of gene cluster, as well as algorithms for finding gene clusters given these more relaxed definitions [2, 3, 5, 19, 20, 34, 38, 40, 56, 61].

Our Results: Given a method for identifying them, how can we assess whether gene clusters are statistically meaningful? The issue of cluster significance arises in two types of analysis: detailed study of the history or function of a particular set of genes and large scale studies of the selective forces acting on the genome as a whole. This leads to two statistical questions:

> **Individual clusters:** Given a *particular* set of genes of interest, is it significant to find these genes in close proximity in a previously unexamined genome?
>
> **Whole genome clustering:** Given two random genomes, is the observation that a certain number of gene clusters appear in both genomes significant?

The problem of significance testing for gene clusters has been introduced by previous authors for a limited set of conditions. Their results, based on combinatorial analysis, are described in detail in Section 4. In the current paper, we model a broad range of scenarios. In Section 2, we focus on a single cluster of pre-specified genes, providing exact expressions for the probability of finding a given set of m genes in a window of size r. In Section 3, we extend these results to take into account incomplete clusters, gene families and clusters found in several genomes. Probabilistic models for detecting clusters in whole genome comparisons are presented, including both genome self-comparison and comparison of genomes from different species. The application of these results to specific biological problems is discussed in Section 5.

2. SIMPLE CLUSTER PROBABILITIES

We begin by introducing a simple definition of a gene cluster and calculate the probability of observing such a cluster in a genome with uniform random gene order (a "random genome"). Let genome, $G = (1, \ldots, n)$, be an ordered set of n genes and let M be a pre-selected set of m genes of interest. These m genes may be of interest because they are contiguous in some other genome ("the reference genome") or because they share a functional property. In any case, the spatial organization of the genes on the reference genome does not enter into the analysis at this point.

Consider the case where the genes in M are found in any order in a window of *exactly* r slots in G. In this case, the first and last of the r slots contain two of the m genes and the remaining $r-2$ slots contain the remaining $m-2$ genes plus $r-m$ intruders. The probability of this event is

$$\frac{\binom{r-2}{m-2}}{\binom{n-1}{m-1}}. \tag{1}$$

In the event these m genes span *at most* r slots in G, it suffices that one of the end points of the window be occupied by one of the m genes. In this case, the probability is

$$q(n,m,r) = \frac{\binom{r-1}{m-1}}{\binom{n-1}{m-1}}. \tag{2}$$

If we require that the genes in M appear in a given order, then the probability of observing the cluster is $q(n,m,r)/m!$.

Intuitively, we expect the probability of finding a cluster by chance will depend on the size of the window, relative to the genome size, and the fraction of slots in the window that are occupied by intruders. For large n, we can make this intuition explicit by applying Stirling's approximation to Equation 2 to obtain

$$q(n,m,r) \approx \left(\frac{w\theta}{e}\right)^{m-1} \theta^{-(r-\frac{1}{2})}\mathcal{O}(1), \tag{3}$$

where $w = \frac{r-1}{n-1}$ and $\theta = 1 - \frac{m-1}{r-1}$ are two parameters introduced to represent *window proportion* and *window sparsity*, respectively. Formulae 2 and 3 can be used to test whether a specific set of m genes is more highly clustered than by chance.

3. GENE CLUSTER STATISTICS

In the previous section we introduced the our basic notion of a cluster, m genes in a window of size at most $r \geq m$, and calculated the probability of finding such a cluster in a random genome. We now use this probability to develop tests for a variety of more complex, biologically motivated, clustering scenarios:

In many cases, authors report finding incomplete clusters: given a cluster of m genes in one genome, a subset of those genes is found close together in another genome. For a given $h < m$, is a cluster of h of the m genes significant?

In our initial analysis, we assumed that each gene in M has exactly one homolog in G[1]. When G includes families of paralogous genes, a given gene in M may now correspond to several genes in G. In this case, a cluster is observed if any set of m genes corresponding to the genes in M is found within a window of length at most r. What is the significance of observing a cluster under these conditions?

The significance of observing a cluster in several genomes arises in the context of comparative maps. For each genome in a comparative map, a mapping is established between each gene in the genome and its homologs in other genomes. Given such a map, what is the probability of observing the same clusters in several genomes?

The advent of comparative maps also introduces the question of the significance of multiple shared clusters in the context of whole genome comparison. When comparing entire

[1]Homologs are genes descended from a single gene in the common ancestral species, resulting from speciation (orthologs) or gene duplication (paralogs).

genomes, how many pairs of homologous clusters should we expect to find by chance alone? Aggregate clustering properties have been used to study the functional and evolutionary implications of large-scale genomic organization, including rates of rearrangement [10, 32, 52, 51], the distribution of breakpoints, conservation of gene order [56], and the duplication processes (e.g., tandem duplication, whole genome duplication, duplication of subchromosomal segments) that dominate in a given lineage [7, 15, 50]. In order to interpret such data correctly, the significance of a certain number of observed shared clusters must be determined.

Whole genome clustering statistics are also relevant to the analysis of individual clusters. The discussion in Section 2 focussed on clusters of pre-specified genes. In practice, gene clusters are often found through comparison of whole genomes. This is often a serendipitous finding, and the circumstances of the discovery may not even be reported. Indeed, the description of the cluster may read as if the genes involved were the original focus of interest, i.e., were prespecified. But such presentation can be misleading as to the significance of the clusters. Signicance tests based on whole genome comparison models (given in Sections 3.4 and 3.5) should be used in this case rather than Equation 2.

In the following sections, we derive a variety of statistical tests for determining the significance of gene clusters by rejecting the hypothesis that such a cluster could have occurred by chance in a genome with uniform 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.

The probability of observing a single cluster of pre-specified genes (Equation 2) is sufficient to test its signficance. However, for the more complex biological scenarios described above, there will typically be more than one cluster of genes that meet the criterion under consideration. In this case, the probability of observing at least one such cluster may be used to test signficance. Unfortunately, in many instances this probability is difficult to calculate because some sets of genes that meet the criterion intersect so that the events under consideration are not independent.

It is generally easier to calculate the expected number of clusters of a given type. Such a result can be used as a benchmark or informal test; if the number of observed clusters, ν, is much greater than the expected number, S, we can assume that about $\nu - S$ of them represent evolutionary or functionally derived clusters. Markov's inequality provides a formal, albeit weak, test: if $\nu > S/\alpha$, then the number of observed clusters exceeds the null hypothesis at a significance level of α.

The above approach assumes that it is possible to calculate the number of observed clusters, ν, from experimental data. In some cases, enumerating all observed clusters may be difficult and it is more convenient to use an approach based on sampling windows from the genome. In this case, significance tests focus on the expected number of windows in the sample that contain a cluster of interest and on the probability that the sample contains at least one such window.

3.1 Incomplete clusters:

Frequently, only a subset of the m genes of interest are found in close proximity in the genome. When is this event significant? To model this scenario, let H be the set of all subsets of M of size $h < m$. The probability that a specific subset in H appears in a window spanning at most r slots is $q(n, h, r)$ and the expected number of such subsets is

$$S_H(n, h, m, r) = \binom{m}{h} q(n, h, r). \tag{4}$$

Notice that these subsets may intersect. For example, if all m genes were found in a single window of length r, then G contains all the incomplete clusters in H but only one biologically interesting cluster.

For this reason, significance tests based on the probability of observing at least one incomplete cluster are easier to interpret. This probability is $P_H(n, h, m, r) = \text{Prob}(\cup_{i=1}^{n} E_i)$, where E_i is the event that the ith subset in H is found in a window of size r in G. Since many of the subsets intersect, the events $\{E_i\}$ are not independent. Let $E_{i_1,\cdots,i_g}$ be the event that each of the g subsets $i_1, \cdots, i_g$ appears in a window of size at most r in G. Then, by the inclusion-exclusion rule,

$$P_H(n, h, m, r) = \sum_{i=1}^{n} \text{Prob}(E_i) - \tag{5}$$
$$\sum_{i_1 \neq i_2}^{n} \text{Prob}(E_{i1,i2}) + \sum_{i_1 \neq i_2 \neq i_3}^{n} \text{Prob}(E_{i1,i2,i3}) - \cdots$$

The first term of this equation is $S_H(n, h, m, r)$ and the remaining terms correct for intersecting subsets. In the genomic context, i.e. for large n, the dominant term of this correction will be due to pairs of subsets whose intersections are as large as possible, namely of size $h - 1$. The windows containing such a pair must overlap by at least $h - 1$ positions. Thus we can estimate the dominant term of Equation 5 by calculating

$$S'_H(n, h, m, r) = \binom{m}{h+1} q(n, h+1, 2r-h+1),$$

the expected number of windows of size $2r-h+1$ containing $h+1$ of the m genes. (Note that this is not exact because not every such window will be the union of two windows of size at most r each containing h members of M.) Then

$$P_H(n, h, m, r) \approx S_H(n, h, m, r) - S'_H(n, h, m, r) \tag{6}$$

represents a first order approximation to the probability that at least one incomplete cluster of size h appears in G.

An upper bound on the probability of finding at least one incomplete cluster can be derived using a sampling approach. Given a *particular* window of size $r \geq h$ of G, the probability that *exactly* h of the m genes fall into that window is just the hypergeometric probability

$$\check{q}_{HW}(n, h, m, r) = \frac{\binom{m}{h}\binom{n-m}{r-h}}{\binom{n}{r}}. \tag{7}$$

The probability that *at least* h of the m genes fall into that window is then

$$q_{HW}(n, h, m, r) = \sum_{i=h}^{\min(r,m)} \check{q}_{HW}(n, h, m, r). \tag{8}$$

The probability of finding at least one incomplete cluster from H anywhere in the genome can now be bounded above

by sampling all windows of size r in G that have a gene from M in the first position:

$$P_H(n,h,m,r) \leq m\, q_{HW}(n,h-1,m-1,r-1). \quad (9)$$

When $S_H() \ll 1$, Equation 4 can be used to test whether the observation that h genes of some pre-specified set of m genes fall into a window of size $r \geq h$ of G is significant. Alternatively, Formulae 6 and 9 can be used to show that the probability of observing at least one such cluster is small. In the usual case where n is very large and either r or m is small, the combinatorial terms involving n may be approximated, and q_{HW} rapidly calculated. In the case of larger m and r, we may use the binomial approximation to the hypergeometric:

$$q_{HW}(n,h,m,r) \approx \sum_{i=h}^{\min(r,m)} \binom{r}{i} \left(\frac{m}{n}\right)^i \left(1-\frac{m}{n}\right)^{r-i}, \quad (10)$$

or a normal approximation with mean $\frac{rm}{n}$ and variance

$$\left(\frac{n-r}{n-1}\right)\left(\frac{rm}{n}\right)\left(1-\frac{m}{n}\right). \quad (11)$$

These approximations improve as m increases with respect to r.

3.2 Allowing for gene families

So far, our analysis has assumed that each gene has exactly one homolog in each genome, a highly unrealistic assumption in most organisms. We now consider genomes with gene families, assuming that the set of genes in G can be partitioned into non-intersecting gene families (i.e., each gene in the family is homologous to all other genes in the family and to no genes outside the family.)

As the size of gene families increases, so do chance occurrences of gene clusters. For example, if we are looking for a cluster containing the m genes in M, and if just one of the genes in M has two homologs, then there are two different sets of m genes that qualify as clusters and the probability of finding the cluster by chance doubles (almost). In general, if each gene $j \in M$ has f_j homologs in G, then there are $\phi(M) = \Pi_{j\in M} f_j$ distinct sets of genes that are homologous to M[2]. For each of these, the probability that it spans at most r slots is $q(n,m,r)$. Thus, the expected number of homologous clusters is

$$S_\phi(n,m,r) = \phi(M) q(n,m,r). \quad (12)$$

The probability that there is at least one such cluster is $P_\phi(n,m,r) \approx 1-[1-q(n,m,r)]^{\phi(M)}$. The latter expression can be used for rough tests, but it is based on an unwarranted assumption of independence of occurrence among the $\phi(M)$ possible clusters. A better approximation of $P_\phi(n,m,r)$ can be estimated by using the inclusion-exclusion rule (Equation 5) to correct for overlapping clusters. For large n, the dominant term of this correction will be due to pairs of clusters that share identical genes in all but one of the m families. The windows containing such a pair must overlap by at least $m-1$ positions. Thus we can estimate the second term of Equation 5 by calculating

$$S'_\phi(n,m,r) = q(n,m+1,2r-m+1) \sum_{j\in M} \frac{\phi(M)}{f_j} \binom{f_j}{2},$$

the expected number of windows of size $2r-m+1$ containing a cluster plus an extra member of one of the m families. (As above, this is only an estimate of the second term because not every such window will be the union of two windows of size at most r each containing a complete cluster.) Then

$$P_\phi(n,m,r) \approx S_\phi(n,m,r) - S'_\phi(n,m,r) \quad (13)$$

represents a first order approximation to the probability that at least one cluster appears. As in the case of incomplete clusters, significance tests for gene clusters in genomes with gene families can be performed using either the expected number of clusters (Equation 12) or the probability of observing at least one cluster (Equation 13).

3.3 Clusters in k genomes:

The probability that a gene cluster is a chance occurrence decreases if found in more than one genome. For k genomes of same gene content with no gene families, the probability that a specific set of m genes appear in all these genomes in windows spanning at most r slots is $q^k = q(n,m,r)^k$. The probability that it appears in at least $k' \leq k$ of these genomes, spanning at most r slots in each case, is

$$P_K = \sum_{j=k'}^{k} \binom{k}{j} q^j (1-q)^{k-j}, \quad (14)$$

which can be used for testing purposes.

The more typical case reported in the literature, however, is that different subsets of M are found in different genomes. Consider k random genomes of size $n_1, \cdots, n_k$ containing subsets of M of sizes $m_1, \cdots, m_k$, respectively. The probability that, for each genome G_i, at least one subset of M of size h_i appears in G_i in a window spanning at most r_i slots is $\Pi_{i=1}^k P_H(n_i,h_i,m_i,r_i)$, where each $h_i \leq \min[m_i,r_i]$. An *ad hoc* test based on all the h_i and r_i is not, however, rigorous. For fairness, the test should be based on $h = \min[h_1,\cdots,h_k]$ and $r = \max[r_1,\cdots,r_k]$ and the test distribution becomes

$$P_K = \Pi_{i=1}^k P_H(n_i,h,m_i,r). \quad (15)$$

For uniform n_i and m_i, this is $P_H(n,h,m,r)^k$.

If the cluster is missing from any of the genomes, then it is fair to use $n = \min[n_1,\cdots,n_k]$ and $m = \max[m_1,\cdots,m_k]$. The probability that subsets of M of size h appear in at least $k' \leq k$ of these genomes, spanning at most r slots in each case, is obtained by substituting $P_H(n,h,m,r)$ for q in Equation 14.

3.4 Comparing two different genomes

The previous sections focussed on the event that a single set of pre-specified genes is observed in a cluster under various conditions. We now address the event that two genomes, G_1 and G_2, from different species share a certain number of gene clusters. Initially, we treat the case where there are no gene families. Each gene in G_1 has exactly one homolog in G_2 and vice versa. We define a paired cluster to be a set of m genes observed in two windows of length at most r, one in G_1 and one in G_2. The expected number of such paired clusters is

$$S^o_C(n,m,r) = \binom{n}{m} q(n,m,r)^2, \quad (16)$$

where $q()$ is defined in Equation 2.

[2]We assume that no two genes in M are members of the same gene family.

While this expression provides a measure of the degree of shared clustering between G_1 and G_2, it is not a convenient basis for data analysis because it requires enumerating all paired clusters. An alternate approach, based on sampling windows from the genome at random, may be preferable. Given a pair of windows of length r, one from each genome, the probability that these windows share at least m homologous gene pairs is

$$q^o_W(n,r,m) = \sum_{i=m}^{r} \frac{\binom{r}{i}\binom{n-r}{r-i}}{\binom{n}{r}}. \qquad (17)$$

Given a random sample of n_w pairs of windows, such that no window in the sample overlaps with any other window in the sample, the expected number of pairs that share at least m genes is

$$S^o_W = n_w \cdot q^o_W(n,m,r).$$

Given a random sample of non-identical, but possibly overlapping, windows, the above expressions can be used to estimate the expected number of pairs that share m homologous pairs, since the fraction of overlapping pairs is $\mathcal{O}(n^{-1})$, when $r \ll n$.

The probability of finding at least one pair of windows in the sample that share at least m genes can be approximated by the equation

$$P^o_W(n_w,n,m,r) = 1 - [1 - q^o_W(n,m,r)]^{n_w}. \qquad (18)$$

but since it is based on an unwarranted assumption that the events of finding clusters in the various pairs of windows are independent, it provides only a rough estimate.

3.4.1 Comparison with a reference genome

A better estimate can be obtained by designating one genome as the reference genome (without loss of generality, G_1) and considering the set of $n-m+1$ contiguous runs of m genes in that genome. The expected number of those runs that will appear in a window of length r in G_2 is

$$S^o_R(n,r,m) = (n-m+1)\, q(n,m,r). \qquad (19)$$

What is the probability that at least one of those runs will be clustered in the second genome? Let $E_i(m,n,r)$ be the event that the m consecutive genes starting at gene i in G_1 appear in a window of size at most r in the second genome. (Note that $\mathrm{Prob}(E_i) = 0$ if $i > n-m+1$, since there are only n genes in the genome.) Let $E_{i_1,\cdots,i_g}(m,n,r)$ be the event that *all* of the g runs of m consecutive genes in G_1 starting at genes $i_1,\cdots,i_g$, respectively, appear in windows of size at most r in G_2. Note that some of the runs may overlap.

The probability that at least one cluster of m (or more) consecutive genes in G_1 appears in a window of size at most r in G_2 is $P^o_R(m,n,r) = \mathrm{Prob}(\cup_{i=1}^{n} E_i)$ and can be calculated using the inclusion-exclusion rule (Equation 5). For large n, we may neglect third and higher order terms and even those second-order terms where $i_2 > i_1+1$, yielding the approximation

$$P^o_R \approx (n-m+1)q(m,n,r) - (n-m)\mathrm{Prob}(E_{1,2}).$$

We can calculate $\mathrm{Prob}(E_{1,2})$ exactly, by considering all the ways in which a set of genes, $1 \ldots m+1$, can appear in two windows that overlap at $m-1$ positions. Stated formally, we compute the probability of the event that genes $1,\ldots,m$ appear in a window of size exactly $r_1 \geq m$ and that genes $2,\ldots,m+1$ appear in a window of size exactly $r_2 \geq m$, where the leftmost positions of the two windows are a_1 and a_2, respectively. Figure 1 lists all possible configurations for two overlapping windows W_1 and W_2 with endpoints a_1+1, a_1+r_1 and a_2+1, a_2+r_2, respectively, that can satisfy these conditions. Let $P_a, P_b, \cdots, P_g$ be the probabilities that the seven configurations in Figure 1 occur. Then

$$\mathrm{Prob}(E_{1,2}) = \sum_{r_1,r_2=m}^{r} \sum_{a_1,a_2} P_a+P_b+P_c+P_d+P_e+P_f+P_g \qquad (20)$$

Due to space limitations, we do not derive these probabilities here, but simply state the results:

$$\sum_{r_1,r_2,a_1,a_2} P_a = q(n,m,r)\frac{(r-m+2)r(r+1)}{m(m+1)},$$

$$\sum_{r_1,r_2,a_1} P_c = q(n,m,r)\,(m-1)\left[\frac{r(r+1)}{m(m+1)} - \frac{1}{\binom{r-1}{m-1}}\right],$$

$$\sum_{r_1,a_1} P_g = q(n,m,r)\frac{(r-m)(m-2)}{m}.$$

Due to symmetry, the term for case (b) is identical to (a), while the terms for (d), (e) and (f) are identical to the term for (c). Collecting terms, Equation 20 becomes

$$\mathrm{Prob}(E_{1,2}) = 2\sum_{r_1,r_2,a_1,a_2} P_a + 4\sum_{r_1,r_2,a_1} P_c + \sum_{r_1,a_1} P_g,$$

which may be calculated rapidly with the help of approximation 3.

3.4.2 Gene families

Let G_1 and G_2, be two random genomes containing n genes and n_f gene families. The number of sets of m gene families is defined by

$$\mu = \binom{n_f}{m} \qquad (21)$$

In a given genome, G_i, gene family j has f_{ij} members. Given a particular set, M, of m genes, the number of distinct sets of genes that are homologous to M is $\phi_i(M) = \Pi_{j=1}^{m} f_{ij}$ in G_i. In the presence of gene families, the expected number of paired clusters found when comparing G_1 and G_2 is

$$S^o_F(n,m,r) = \left[\sum_{k=1}^{\mu} \phi_2(M_k)\phi_1(M_k)\right] q(n,m,r)^2. \qquad (22)$$

To compute $S^o_F(n,m,r)$ requires a complete catalog of all gene families and their sizes for the genome in question. For a few fully sequenced species, it is currently possible to calculate $S^o_F(n,m,r)$, but requires enumerating all sets of m gene families. Distributions of gene family sizes, under various assumptions, have been published for a number of

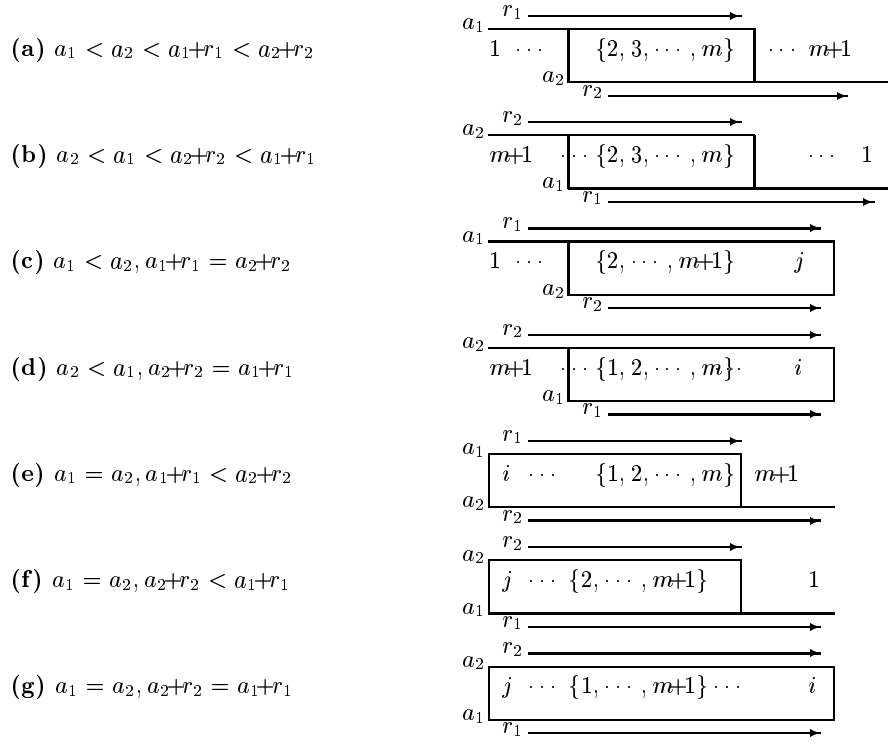


Figure 1: All possible configurations for two overlapping windows W_1 and W_2 such that genes $1, ..., m$ appear in W_1 and genes $2, ..., m{+}1$ appear in W_2. Genes 1 and $m{+}1$ must be present where indicated at the endpoints of windows. Genes i and j may be any of the genes within the braces, except gene 1 or gene $m{+}1$. Except for these constraints, genes within the braces may occur in any order within the overlap between the windows, and may be intermingled with other genes.

species (e.g., [15, 30, 36, 59]). In future work, we plan to derive approximations that are more easily calculated based on parameterized models of such distributions. We seek quantities $\tilde{f}_i(m)$ that depend only on m and the model parameters such that

$$\binom{n_f}{m} \tilde{f}_1(m)\tilde{f}_2(m)$$

is a good approximation for $\prod_{k=1}^{\mu}[\phi_2(M_k)\phi_1(M_k)]$. A first approach is to assume that each gene family has the same number of paralogs, f. Then $n_f = n/f$, $\tilde{f}(m) = f^m$ and

$$S_F^o(n, m, r) \approx \binom{n_f}{m} [f^m q(n, m, r)]^2. \tag{23}$$

While this assumption will rarely, if ever, be true, Equation 30 is easy to calculate and provides a useful tool for exploring how the significance of genome-wide clustering varies with gene family size.

For the case where all gene families are of equal size, we can also derive measures that can be used in sampling tests. Given a pair of windows, W_1 and W_2, of length r selected from G_1 and G_2, respectively, the probability that these windows share at least m gene families is

$$q_{FW}^o(m) = \sum_{k=m}^{r} p_1(k) \cdot p_2(k, m), \tag{24}$$

where $p_1(k)$ is the probability that there are k distinct gene families in W_1 and $p_2(k, m)$ is the probability that at least m of those k families appear in W_2.

The first term is

$$p_1(k) = \binom{n_f}{k} \frac{\sum_{\mathcal{S}} \binom{f}{x_1} \cdot \binom{f}{x_2} \cdots \binom{f}{x_k}}{\binom{n}{r}}, \tag{25}$$

where $\mathcal{S}$ is the set of all ensembles $\{x_1 \ldots x_k\}$ such that

$$\sum_{j=1}^{k} x_j = r$$
$$0 < x_j \leq f.$$

The second term, $p_2()$, is the probability that m of those k gene families appear in W_2. This is equivalent to the requirement that $k{-}m{+}1$ families be excluded from the window or

$$p_2(k, m) = 1 - \bar{p}(k, k{-}m{+}1). \tag{26}$$

The probability that at least j gene families will be absent from a window of size r selected at random is

$$\bar{p}(k, j) = \binom{k}{j} \frac{\binom{n - jf}{r}}{\binom{n}{r}},$$

where the first term is the number of ways of excluding j of the k families from the set of all genes and the second term is the number of ways of selecting r genes from the remaining

genes normalized by the number of ways of sampling r genes from the entire gene set.

Given a random sample of n_w pairs of non-overlapping windows, the expected number of pairs that have m gene families in common is

$$S^o_{FW}(n_w, n, m, r) = n_w \; q^o_{FW}(n, m, r), \tag{27}$$

and the probability of finding at least one such pair is approximately

$$P^o_{FW}(n_w, n, m, r) \approx 1 - [1 - q^o_{FW}(n, m, r)]^{n_w}. \tag{28}$$

3.5 Genome self-comparison

Clusters of paralogs in the same genome are often presented as evidence of whole genome duplication or duplication of large sub-chromosomal segments. Thus, the goal in genome self-comparison is to determine the degree of clustering among duplicated genes. In this case, genes with no paralogs may be ignored. Let G_p be the set of $n_p \leq n$ genes in G that have been duplicated. For convenience, we will describe G_p as an ordered set, $1, \ldots, n_p$, keeping in mind that the numbering scheme in G_p is different from that in G (e.g., the first gene in G_p is not necessarily the first gene in G.) Let f_j be the number of paralogs in gene family j and n_f be the number of gene families. Let $\mathcal{M}$ be a particular set of m different gene families in G_p and let $M_i(\mathcal{M})$ be a set of m genes, one from each family in $\mathcal{M}$. The total number of pairs of non-intersecting sets $(M_i(\mathcal{M}), M_l(\mathcal{M}))$ is

$$\psi(\mathcal{M}) = \Pi_{j \in \mathcal{M}} \binom{f_j}{2}.$$

In the paralogous case, we define a paired cluster to be two non-intersecting sets $M_i(\mathcal{M})$ and $M_j(\mathcal{M})$ found in two, possibly overlapping, windows of length at least r in the same genome. The expected number of paired clusters is then

$$S^p_F(n_p, m, r) = \left[\sum_{i=1}^{\mu} \psi(\mathcal{M}_i) \right] q(n_p, m, r)^2, \tag{29}$$

where μ is defined in Equation 21. If all genes in G_p have the same number of paralogs, f, then

$$S^p_F(n_p, m, r) = \binom{n_f}{m} \binom{f}{2}^m q(n_p, m, r)^2. \tag{30}$$

Calculating $q^p_{FW}(m)$, the probability that two non-overlapping windows selected at random share at least m gene families, is analogous to the two genome case (Equation 24). The probability, $p_1(k)$, that there are k distinct gene families in W_1 is again given by Equation 25. However, in calculating the probability that at least m of those families appear in W_2, we must take into account the fact that W_1 and W_2 are in the same genome and competing for the same genes. If W_1 contains c_i genes from family i, then at most $f - c_i$ genes from that family can appear in W_2. We estimate that $c_i \approx r/k$ for each of the k gene families in W_1. Then the probability that at least j gene families will be absent from W_2 is approximately

$$\bar{p}'(k, j) \approx \binom{k}{j} \frac{\binom{n - r - j(f - \frac{r}{k})}{r}}{\binom{n-r}{r}}, \tag{31}$$

yielding

$$q^p_{FW}(m) \approx \sum_{i=m}^{r} p_1(k) \cdot (1 - \bar{p}'(k, k-m+1)). \tag{32}$$

The degree of clustering of duplicated genes in a genome can be estimated by counting the number of pairs of windows that share a given number of gene families. Given a random sample of n_w pairs of non-overlapping windows taken from G_p, the expected number of pairs that have m gene families in common is

$$S^p_{FW}(n_w, n_p, m, r) = n_w \; q^p_{FW}(n_p, m, r) \tag{33}$$

and

$$P^p_{FW}(n_w, n_p, m, r) \approx 1 - [1 - q^p_{FW}(n_p, m, r)]^{n_w} \tag{34}$$

yields a rough approximation for the probability of finding at least one such pair.

4. PREVIOUS WORK

In their analysis of the significance of conserved synteny, Trachtulec and Forejt [60] estimate the probability of finding m genes in a window of exactly r slots by chance to be $(r/n)^m$, where n is the number of genes in the genome. If $m \ll r$ and $m, r \ll n$, this formula approximates our exact expression 1.

As part of their analysis of gene duplication in the human genome, Venter *et al.* [61] suggest that the probability of a fixed set of m genes occurring in a given order within an interval of r successive gene positions in a random genome of length n is

$$u_1(n, m, r) = \frac{\sum_{i=m-2}^{r-2} \binom{i}{m-2}}{n^{m-1}} \tag{35}$$

For a large genome, where $n^{m-1} \approx (n-1)!/(n-m+1)!$, and neglecting end effect (or assuming a circular genome), Equation 35 is essentially correct. An exact expression for this quantity is $u(n, m, r) = q(n, m, r)/m!$, where $q(n, m, r)$ is defined in Equation 2

They further consider the case of two sets of m genes that are pairwise paralogous and state a probability "allowing for" the two sets "to be spread across r positions" in two separate locations:

$$u_2(n, m, r) = \frac{\left[\sum_{i=m-2}^{r-2} \binom{i}{m-2}\right]^2}{n^{m-1}} \tag{36}$$

However, it is not clear what event has this probability, even approximately. Indeed, for $m = 3$ and $r = \frac{n}{2}$, for example, Equation 36 is $O(n^2)$ and thus cannot be a probability.

5. APPLICATION TO BIOLOGICAL DATA

There is a broad literature in which gene cluster analysis has been used to interpret the evolutionary or functional implications of gene order in species ranging from viruses and bacteria to mammals, based on data derived from both whole genome sequencing and linkage mapping. To show the utility of the models developed in the previous sections, we apply our results to a few examples from this literature. Our intent here is not to reanalyze the data or question the conclusions of the studies cited below, but rather to provide

Region	*Gene families found in region*	*References*
MHC	Abc, C3/4/5, Col, Hsp, Notch, Pbx, Psmb, Rxr, Ten	[13, 21, 25, 26, 55, 60]
HOX	Achr, Ccnd, Cdc, Cdk, Dlx, En, Evx, Gli, Hh, Hox, If, Inhb, Nhr, Npy/Ppy, Wnt	[1, 21]
FGR	Adr, Ank, Egr, Fgfr, Pa, Vmat, Lpl	[9, 31, 41]
TBOX	Cryb, Lhx, Nos, Tbx, Tcf, Prkar	[43]
MATN	Eya, Hck, Matn, Myb , Myc, Sdc, Src	[16]

Table 1: Paralogous gene clusters in vertebrate genomes recently reported in the literature. Many of these clusters appear in several vertebrate species and have also been found in invertebrate genomes.

concrete examples of how our models can be put to practical use in real biological studies.

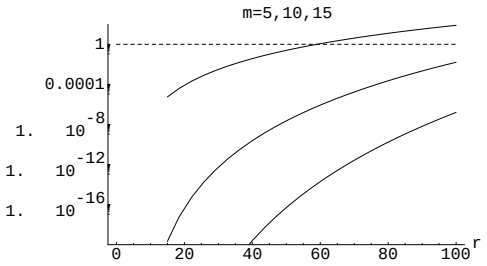
Individual clusters: Since Ohno [39] first hypothesized two whole genome duplications in early vertebrates, the role of large scale duplication in vertebrate evolution has been much debated [22, 44, 54, 63]. One type of evidence that is offered in these debates is the presence of linkage groups that appear to be duplicated and also to be conserved across several species. At least a dozen papers analyzing such regions, summarized in Table 1, have appeared in the last decade. These clusters typically contain five to fifteen genes spread over a window of 15 to 100 slots. Are conserved clusters of this sparsity truly significant? Figure 2 shows the expected number of clusters in a random genome of size $n = 3000$, calculated using Equation 30. Since these studies were performed on linkage data, n refers to the number of genes in the data set not the number of genes in the organism. The Jackson Laboratories Mouse Genome Database [33] (MGD), currently contains roughly 3000 mapped genes. The curves in Figure 2(a) suggest that when gene family sizes are small ($f = 2$), a cluster of size ten or larger is significant even if spread over a large window. However, as f increases, clusters found in larger windows are no longer significant (Figure 2(b)).

Let us consider one of these examples, the TBOX cluster, in detail. Ruvinsky and Silver [43] observed paralogous gene clusters on mouse chromosomes 5 and 11, shown in Figure 3, and explored the hypothesis that these genes were duplicated in a single event. The central cluster is quite compelling but it is more difficult to decide whether the more distant *Prkar* paralogs should be included in this candidate duplicated region. A statistical test can help resolve this question. We extend Equation 4 to estimate the expected number of incomplete clusters of a pre-specified set of genes, obtaining

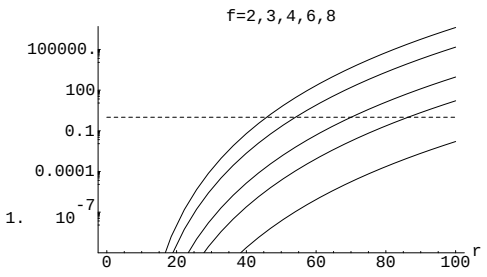
$$S_{FH}(n, h, m, r, f) = \binom{m}{h} (f\text{–}1)^h q_H(n\text{–}m, h, m, r), \qquad (37)$$

for a uniform gene family size, f. In a dataset of 2888 mapped genes extracted from MGD [33], genes from the *Cryb*, *Lhx*, *Nos*, *Tbx* and *Tcf* families were found in a window of 15 slots on chromosome 5 and a window of 48 slots on chromosome 11. The inclusion of the *Prkar* genes, yields a cluster of seven genes in windows of 47 and 65 slots, respectively. If we take the six gene cluster on chromosome 5 as the reference, assuming $f = 3$, the expected number of such clusters in a random genome is $S_{FH}(2888, 6, 15, 48, 3) = 3.0 \times 10^{-4}$, suggesting that the six gene cluster is significant. Adding *Prkar1b* to the reference cluster, yields $h = 7$, $m = 47$ and $r = 65$, and the expected number of clusters becomes 0.83. In this case, it is no longer possible to reject chance as a possible explanation for the seven gene cluster with confidence. Moreover, if we select chromosome 11 as the reference, then the expected numbers of comparable six and seven gene clusters in a random genome are 1.0×10^{-3} and 1.2, respectively, leading to the same conclusions.

Whole genome analysis: In large scale studies of conserved regions, the intent is to characterize processes of duplication, rearrangement and conservation on a genome wide scale rather than detailed study of a particular region. In one example of such an analysis [56], Tamames compaired pairs of bacterial genomes in a study of gene order conservation in prokaryotes. His approach uses a parameterized method for identifying pairs of runs of orthologs (one in each species), in which the user must specify two parameters: m_0, the minimum number of pairs of orthologs in the



(a)



(b)

Figure 2: Expected number of paralogous clusters in a random genome, where $n = 3000$ and $m = r$. The threshold, $S_F^p() = 1$, is shown as a dashed line. (a) Each gene has one paralog ($f = 2$). The number of genes in the cluster ranges from $m = 5$ (top curve) to $m = 15$ (bottom curve). (b) Cluster size $m = 10$. Gene family sizes range from $f = 2$ (bottom curve) to $f = 8$ (top curve).

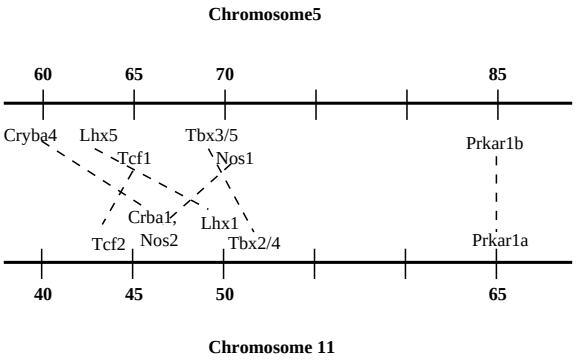


Figure 3: Clusters of paralogous mouse genes on chromosomes 5 and 11. Adapted from Ruvinsky and Silver [43].

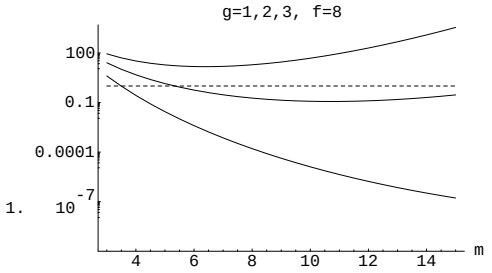
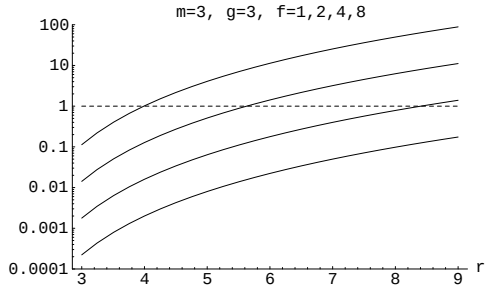
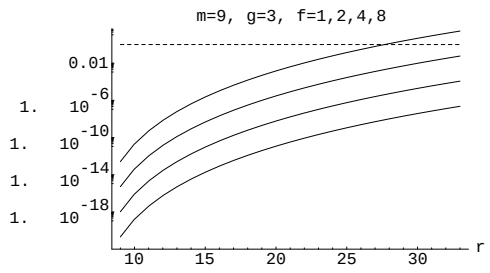


Figure 5: Expected number of orthologous clusters of m genes in a window of size $r = (g+1)(m-1)+1$, $f = 8$ and $n = 3000$. Gap sizes range from $g = 1$ (bottom curve) to $g = 3$ (top curve). The threshold, $S_F^o() = 1$, is shown as a dashed line.

run and, g, the maximum number of intruders found between any pair of orthologs in the run. Thus, a run with m orthologs can be at most $(g+1)(m-1)+1$ slots long. For the purposes of the study [56], m_0 and g were both set to three. Our model provides a rational basis for selecting parameters of clustering algorithms such as this. In this example, we seek the minimum m_0 and maximum g such that the clusters obtained are still significant. Figures 4 and 5 show the expected number of clusters with m gene families in common, calculated using Equation 23. Since bacterial genomes range from roughly 500 to 7000 genes, an intermediate size of $n = 3000$ was used. Figure 4(a) shows that with a gap of three, clusters of three genes are significant for small gene family sizes but cease to be for $f \geq 4$. When $m = 9$, however, most clusters are significant except when $f = 8$ and r approaches its maximum range (Figure 4(b)). The dependence of cluster significance on gap size is demonstrated in Figure 5. When $f = 8$, clusters in windows of maximum size with $g = 3$ are not significant for any value of m, no matter how large. However, if $g = 2$ clusters are signficant for $m \geq 6$ and for $g = 1$ most clusters are significant. In the absence of additional biological information that can be used to determine cluster significance, such as gene orientation, these results suggest that a slightly higher value of m_0 and a gap size of $g \leq 2$ would guarantee the significance of clusters found with this algorithm. This example demonstrates that statistical models of gene clusters are useful not only in data analysis but also in algorithm design.



(a)



(b)

Figure 4: Expected number of orthologous clusters of m genes in a window of size r, where r ranges from m to $(g+1)(m-1)+1$ and $n = 3000$. Gene family sizes are uniform and range from $f = 1$ (bottom curve) to $f = 8$ (top curve). A maximum of $g = 3$ genes is allowed between any pair of genes in the cluster. The threshold, $S_F^o() = 1$, is shown as a dashed line. (a) m = 3. (b) m = 9.

6. DISCUSSION AND FUTURE WORK

We have presented probabilistic models for determining the significance of local gene clusters in both paralogous and orthologous settings. Under a model of uniform random gene order, we consider the probability of finding a cluster of a particular set of genes, as well as the expected number of clusters observed in whole genome comparison. Our models take multiple genomes and gene families into account. Despite a fairly simple and abstract model, we have demonstrated that our results can be applied to range of problems from the biology literature.

In future work, we plan to develop more detailed, biologically motivated models. The current model treats the genome as an ordered set of genes. An extended analysis would model the chromosomal positions of genes, and would take tandem duplications and gene rich and gene poor regions into account. A parameterized model of gene family sizes that yields realistic, computationally tractable approximations is also needed. Finally, other types of biological information besides gene order can be brought to bear on the assessment of significance including gene orientation (e.g., [56, 64]) and divergence times (e.g., [6, 15, 43].)

7. ACKNOWLEDGMENTS

The authors wish to thank the reviewers for helpful comments. 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 is a Fellow of the Evolutionary Biology Program of the Canadian Instiute for Advance Research.

8. REFERENCES

[1] A. Amores, A. Force, Y. l. Yan, L. Joly, C. Amemiya, A. Fritz, R. 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] M. Blanchette, T. Kunisawa, and D. Sankoff. Gene order breakpoint evidence in animal mitochondrial phylogeny. *Journal of Molecular Evolution*, 49:193–203, 1999.

[5] P. Bork, B. Snel, G. Lehmann, M. Suyama, T. Dandekar, W. L. 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.

[6] K. Chen, D. Durand, and M. Farach-Colton. Notung: A program for dating gene duplications and optimizing gene family trees. *Journal of Computational Biology*, 7(3/4):429–447, 2000.

[7] S. Chervitz, L. Aravind, G. Sherlock, C. Ball, E. Koonin, S. Dwight, M. Harris, K. Dolinski, S. Mohr, T. Smith, S. Weng, J. Cherry, and D. Botstein. Comparison of the complete protein sets of worm and yeast: Orthology and divergence. *Science*, 282:2022–2028, 1998.

[8] 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.

[9] 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.

[10] J. Ehrlich, D. Sankoff, and J. Nadeau. Synteny conservation and chromosome rearrangements during mammalian evolution. *Genetics*, 147(1):289–96, 1997.

[11] N. El-Mabrouk, D. Bryant, and D. Sankoff. Reconstructing the pre-doubling genome. In *RECOMB99, Third Annual International Conference on Computational Molecular Biology*, 1999.

[12] N. El-Mabrouk, J. H. Nadeau, and D. Sankoff. Genome halving. In Springer-Verlag, editor, *Combinatorial Pattern Matching*, pages 235–250, 1998.

[13] 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.

[14] M. D. Ermolaeva, O. White, and S. Salzberg. Prediction of operons in microbial genomes. *Nucleic Acids Res*, 5(29):1216–1221, Mar 2001.

[15] R. Friedman and A. Hughes. Gene duplication and the structure of eukaryotic genomes. *Genome Research*, 11:373–381, 2001.

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

[17] 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.

[18] 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.

[19] S. Heber and J. Stoye. Algorithms for finding gene clusters. In *WABI01*, 2001.

[20] S. Heber and J. Stoye. Finding all common intervals of k permutations. In *CPM01*, 2001.

[21] 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.

[22] A. L. Hughes. *Adaptive Evolution of Genes and Genomes*. Oxford University Press, New York, 1999.

[23] M. Huynen and P. Bork. Measuring genome evolution. *PNAS*, 95:5849–56, 1998.

[24] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(682):860–921, 2001.

[25] 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.

[26] N. Katsanis, J. Fitzgibbon, and E. 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.

[27] A. B. Kolsto. Dynamic bacterial genome organization. *Molecular Microbiology*, 24:241–8, 1997.

[28] S. Kumar, S. R. Gadagkar, A. Filipski, and X. Gu. Determination of the number of conserved chromosomal segments between species. *Genetics*, 157:1387–1395, 2001.

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

[30] W.-H. Li, Z. Gu, H. Wang, and A. Nekrutenko. Evolutionary analyses of the human genome. *Nature*, 409:847–9, 2001.

[31] 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.

[32] A. McLysaght, C. Seoighe, and K. H. Wolfe. High frequency of inversions during eukaryote gene o. In D. Sankoff and J. H. Nadeau, editors, *Comparative Genomics*, pages 281–294. Kluwer Academic Press, Dordrecht, NL, 2000.

[33] Mouse Genome Database. *http://www.informatics.jax.org*.

[34] J. Nadeau and D. Sankoff. Counting on comparative maps. *Trends Genet*, 14(12):495–501, 1998.

[35] J. Nadeau and D. Sankoff. The lengths of undiscovered conserved segments in comparative maps. *Mamm Genome*, 9(6):491–5, 1998.

[36] J. H. Nadeau and D. Sankoff. Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. *Genetics*, 147:1259–1266, November 1997.

[37] 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.

[38] S. J. O'Brien, J. Wienberg, and L. A. Lyons. Comparative genomics: lessons from cats. *Trends Genet*, 10(13):393–399, Oct 1997.

[39] S. Ohno. *Evolution by Gene Duplication*. Springer-Verlag, 1970.

[40] 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.

[41] 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.

[42] P. A. Pevzner. *Computational Molecular Biology: An Algorithmic Approach*. MIT Press, Cambridge, MA, 2000.

[43] 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.

[44] D. Sankoff. Gene and genome duplication. *Current Opinion in Genetics and Development*, 11:681–4, 2001.

[45] D. Sankoff. Short inversions and conserved gene clusters. In *Proceedings of the Society for Applied Computing(SAC 2002)*, 2002.

[46] 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.

[47] 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.

[48] D. Sankoff. and N. El-Mabrouk. Genome rearrangement. In T. Jiang, T. Smith, Y. Xu, and M. Zhang, editors, *Current Topics in Computational Biology*. MIT Press, in press, 2002.

[49] D. Sankoff, V. Ferretti, and J. H. Nadeau. Conserved segment identification. *Journal of Computational Biology*, 4:559–565, 1997.

[50] C. Semple and K. H. Wolfe. Gene duplication and gene conversion in the *Caenorhabditis elegans* genome. *JME*, 48(5):555–64, 1999.

[51] C. Seoighe et al. Prevalence of short inversions in yeast genome evolution. *PNAS*, 97:14433–14437, 2000.

[52] C. Seoighe and K. Wolfe. Extent of genomic rearrangement after genome duplication in yeast. *Proc Natl Acad Sci U S A*, 95(8):4447–52, 1998.

[53] C. Seoighe and K. H. Wolfe. Updated map of duplicated regions in the yeast genome. *Gene*, 238:253–261, 1999.

[54] L. Skrabanek and K. Wolfe. Eukaryote genome duplication - where's the evidence? *Curr Opin Genet Dev*, 8(6):559–565, 1998.

[55] 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.

[56] J. Tamames. Evolution of gene order conservation in prokaryotes. *Genome Biol*, 6(2):0020.1–11, 2001.

[57] J. Tamames, G. Casari, C. Ouzounis, and A. Valencia. Conserved clusters of functionally related genes in two bacterial genomes. *JME*, 44:(66–73), 1997.

[58] J. Tamames, M. Gonzalez-Moreno, A. Valencia, and M. Vicente. Bringing gene order into bacterial shape. *Trends Genet*, 3(17):124–126, Mar 2001.

[59] J. Tiuryn, J. P. Radomski, and P. P. S. . A formal model of genomic DNA multiplication and amplification. In D. Sankoff and J. H. Nadeau, editors, *Comparative Genomics*, pages 503–5013. Kluwer Academic Press, Dordrecht, NL, 2000.

[60] 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.

[61] J. C. Venter et al. The sequence of the human genome. *Science*, 291(5507):1304–51, 2001.

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

[63] K. H. Wolfe. Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genetics*, 2(5):333–41, 2001.

[64] K. H. Wolfe and D. C. Shields. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387:708–713, 1997.

This research was sponsored in part by National Science Foundation (NSF) grant no. CCR-0122581.