RESEARCH ARTICLES

Gene Cluster Statistics with Gene Families

Narayanan Raghupathy[1] *and Dannie Durand*†

*Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA; and †Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA

Identifying genomic regions that descended from a common ancestor is important for understanding the function and evolution of genomes. In distantly related genomes, clusters of homologous gene pairs are evidence of candidate homologous regions. Demonstrating the statistical significance of such "gene clusters" is an essential component of comparative genomic analyses. However, currently there are no practical statistical tests for gene clusters that model the influence of the number of homologs in each gene family on cluster significance. In this work, we demonstrate empirically that failure to incorporate gene family size in gene cluster statistics results in overestimation of significance, leading to incorrect conclusions. We further present novel analytical methods for estimating gene cluster significance that take gene family size into account. Our methods do not require complete genome data and are suitable for testing individual clusters found in local regions, such as contigs in an unfinished assembly. We consider pairs of regions drawn from the same genome (paralogous clusters), as well as regions drawn from two different genomes (orthologous clusters).

Determining cluster significance under general models of gene family size is computationally intractable. By assuming that all gene families are of equal size, we obtain analytical expressions that allow fast approximation of cluster probabilities. We evaluate the accuracy of this approximation by comparing the resulting gene cluster probabilities with cluster probabilities obtained by simulating a realistic, power-law distributed model of gene family size, with parameters inferred from genomic data. Surprisingly, despite the simplicity of the underlying assumption, our method accurately approximates the true cluster probabilities. It slightly overestimates these probabilities, yielding a conservative test. We present additional simulation results indicating the best choice of parameter values for data analysis in genomes of various sizes and illustrate the utility of our methods by applying them to gene clusters recently reported in the literature. Mathematica code to compute cluster probabilities using our methods is available as supplementary material.

Introduction

Identifying homologous genomic regions is an important step for many comparative genomic analyses. Evidence of spatial gene conservation is used in comparative map construction, function prediction, operon detection, protein interaction, phylogeny reconstruction based on breakpoint or rearrangement distance, reconstruction of ancestral gene order, and identification of horizontal gene transfer events (Dandekar et al. 1998; Huynen and Bork 1998; Overbeek et al. 1999; Tamames 2001; Tamames et al. 2001; Zheng et al. 2002; Chen et al. 2004; Hurst et al. 2004; Bourque et al. 2005; Murphy et al. 2005; Homma et al. 2007; von Mering et al. 2007). Identifying homologous regions is straightforward when the two genomic regions are so closely related that gene content and order are preserved (fig. 1*b*). However, the task becomes more challenging when homologous regions have been scrambled by genome rearrangement events. In more diverged genomes, "gene clusters," regions in which gene content is similar but neither gene content nor order is completely preserved, are signatures of shared ancestry (fig. 1*c*). Before accepting such clusters as evidence of regional homology, it is essential to rule out the possibility that the observed spatial arrangement occurred by chance.

Tests of gene cluster significance require estimation of the probability of observing a given gene cluster under a suitable null hypothesis. This probability depends on the number of conserved genes in the cluster, the number of unmatched genes between conserved genes in the cluster (gap size), the number of genes in each genome, how the gene cluster was found, gene order conservation, the number of genomes in which the cluster was observed, and the distribution of gene family sizes in the genome.

Statistical models have been developed that consider various subsets of these properties, although currently none take all into account (Wolfe and Shields 1997; Vision et al. 2000; Trachtulec and Forejt 2001; Venter et al. 2001; McLysaght et al. 2002; Vandepoele et al. 2002; Calabrese et al. 2003; Durand and Sankoff 2003; Hoberman et al. 2005; Sankoff and Haque 2005; Raghupathy et al. 2008). Typically these approaches prove significance by showing that a value of test statistic, for example, the number/density of homologous gene pairs in the cluster, as extreme as the observed value is unlikely to occur under the null hypothesis. Random gene order is the most widely used null hypothesis for gene cluster statistics.

Most methods model a genome as an ordered list of genes. Physical distances and chromosome breaks are ignored. When two genomes, G_1 and G_2, are compared, the input is a mapping between homologous genes in G_1 and G_2. Genes that are homologous are considered to be in the same gene family. If each gene in G_1 maps to at most one gene in G_2 and vice versa, then the mapping is one-to-one and represents orthology. Otherwise, the mapping is one-to-many or many-to-many. In this case, the genes in the same family may be either orthologous or paralogous. In a genome self-comparison, the input is a single genome,

[1] Present address: Lewis-Sigler Institute for Integrative Genomics, Princeton University.

Key words: gene cluster, gene family, statistical significance, spatial comparative genomics, genome evolution, combinatorics.

E-mail: nraghupa@cs.cmu.edu

Mol. Biol. Evol. 26(5):957–968. 2009
doi:10.1093/molbev/msp002
Advance Access publication January 15, 2009

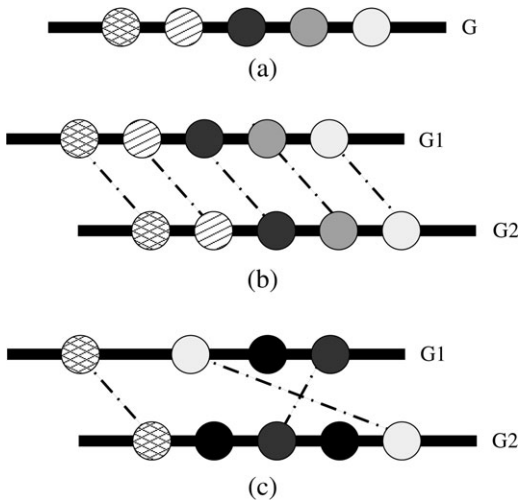

FIG. 1.—Evolution of a hypothetical gene cluster: (*a*) Ancestral genome. (*b*) Genomic regions immediately after a speciation or a whole genome duplication. Gene content and order are preserved. (*c*) Diverged genomic regions with similar gene content but in which neither gene content nor order is preserved. Black circles represent genes that do not have homologs in the depicted regions.

G, and a mapping between homologous genes within G. In this case, all genes in the same family are paralogs.

A gene cluster is a set of homologous gene pairs that satisfy a set of mathematical constraints that enforce compactness (Hoberman and Durand 2005). The most frequently used gene cluster definitions fall into one of two frameworks:

- The "r-window cluster": Two genomic regions, each containing r adjacent genes, that share "at least" m homologous gene pairs.
- The "max-gap cluster": Two genomic regions sharing at least m homologous gene pairs such that the number of unmatched, intervening genes between adjacent homologs in the same region is never larger than a given threshold, g.

A gene cluster is orthologous if each homologous gene pair consists of one gene in G_1 and one gene in G_2 and paralogous if the genes are drawn from nonoverlapping regions of the same genome.

Durand and Sankoff (2003) showed that in addition to the parameters that characterize the gene cluster, the significance of a gene cluster depends on how the cluster was found. Typically, one of the following strategies is used to find gene clusters:

- The "reference region" approach: A researcher is interested in a set of genes, called the reference genes, and searches the genome for regions containing either all the reference genes or a subset of them.
- The "window sampling" approach: A researcher is interested in studying the evolution of regions surrounding a pair of homologous genes of interest and searches for more homologs in their immediate vicinity. Typically, two windows, W_1 and W_2, containing r_1 and r_2 genes, respectively, are compared.
- The "whole genome" approach: A researcher is interested in studying spatial organization on a genomic scale and finds the set of all gene clusters by scanning both genomes.

Among these, the reference region and window sampling approaches are natural models for local studies focusing on specific regions or genes (Lundin 1993; Katsanis et al. 1996; Coulier et al. 1997; Endo et al. 1997; Kasahara 1997; Ruvinsky and Silver 1997; Amores et al. 1998; Hughes 1998; Pebusque et al. 1998; Smith et al. 1999; Lipovich et al. 2001; Abi-Rached et al. 2002; Spring 2002; Danchin et al. 2003; Vienne et al. 2003). The whole genome approach is appropriate for global questions, such as the nature of duplication events that shaped the evolution of the genome (e.g., McLysaght et al. 2002; Panopoulou et al. 2003; Dehal and Boore 2005). The probability of finding a cluster increases with the number of searches required to find the cluster. In the reference region approach, the number of searches is proportional to the number of genes in the genome, n. In the window sampling approach, the search space depends on the window size $r = \max(r_1, r_2)$. Typically, $r \ll n$. In the whole genome approach, the search space is proportional to n^2 because all combinations of starting positions in both genomes must be considered. Note that a cluster found using window sampling might be significant, whereas a cluster with identical properties found using a reference region approach might not be significant.

Monte Carlo methods, where the distribution of test statistic is estimated by randomization, are widely used for assessing statistical significance of gene clusters (e.g., Wolfe and Shields 1997; Vision et al. 2000; McLysaght et al. 2002; Vandepoele et al. 2002). Randomization has the advantage that it preserves genomic properties (other than gene order) including gene family sizes. However, randomization methods are computationally expensive and are suitable mainly for genome scale, rather than local, analyses because a complete comparative map is needed in order to carry out randomization.

The simplest analytical methods are based on the r-window model using the reference region approach and consider only the number of homologous pairs, the number of genes in the window, and the total number of genes in the genome (Trachtulec and Forejt 2001; Venter et al. 2001). Durand and Sankoff (2003) generalized r-window gene cluster statistics to window sampling and whole genome approaches using a combinatorial framework. Hoberman et al. (2005) presented the first statistical tests for determining significance of max-gap gene clusters. They developed tests for the reference region and whole genome comparison approaches, assuming that the mapping between the genes is one-to-one. To our knowledge, there are no gene cluster statistics for the max-gap model with many-to-many mapping.

Cluster significance increases when gene order is conserved. If the order of all m homologs in the cluster is preserved, then it is straightforward to incorporate gene order into an existing test (Venter et al. 2001; Durand and Sankoff 2003). However, combining partial gene order with other cluster properties in a single test is challenging (Sankoff and Haque 2005). Although recently a new gene cluster model has been proposed that is more sensitive to partial order conservation (Zhu et al. 2008).

Most approaches to computing the significance of clusters spanning three or more regions combine pairwise tests in various ways (Durand and Sankoff 2003; Simillion et al. 2004). However, combining pairwise comparisons does not capture the additional significance of genes that are conserved in more than two regions resulting in underestimation of cluster significance (Raghupathy and Durand 2008). In a departure from the current approaches Raghupathy and Durand (2008) presented tests for r-window gene clusters spanning exactly three regions that combine evidence from genes shared among all three regions, as well as genes shared between pairs of regions.

Realistic tests require a gene family model that captures many-to-many homology. Although this assumption may be appropriate when comparing two closely related genomes, in general establishing one-to-one homology can be difficult (Chen et al. 2007). Moreover, orthology is not always a one-to-one relationship (Fitch 2000). In many cases, such as when lineage-specific duplications have occurred, a many-to-many mapping correctly represents the underlying biology. Even when the true relationship is one-to-one homology, computational methods cannot always unambiguously identify unique homologous pairs (Chen et al. 2007).

Most currently available tests assume that the mapping between homologous genes is one-to-one. The few methods that consider a many-to-many model are not suitable for practical data analysis. Durand and Sankoff (2003) presented r-window cluster statistics that consider gene family size for both the reference region and window sampling approaches. They provide upper bounds on cluster probabilities for the reference region model under the fixed-size family assumption that are computationally tractable. However, their statistical tests for the window sampling approach do not admit a computationally feasible implementation. Calabrese et al. (2003) presented an approach that combines gene cluster identification and significance testing. This method implicitly assumes a binomial gene family distribution. However, empirical studies have shown that gene family sizes follow a power-law distribution (Qian et al. 2001; Rzhetsky and Gomez 2001; Karev et al. 2002; Koonin et al. 2002; Kaplan et al. 2004). Moreover, because the test is coupled with a cluster finding algorithm, it cannot be used to test the significance of clusters found by other methods. Danchin and Pontarotti (2004) recognized the importance of accounting for multiple co-orthologs when testing the significance of conserved genomic regions and proposed a strategy for downweighting homologs in gene families, in a reference region framework. However, because the weights assigned are not based on a formal model relating gene family size to the probability of observing a cluster under the null hypothesis, their approach is not appropriate to general data analysis problems.

Accurate, computationally tractable analytical methods for determining cluster significance in the presence of gene families remain a major challenge. Considering gene families is important because cluster significance is sensitive to gene family size. Given a one-to-one mapping between homologous genes in G_1 and G_2, each gene in G_1 matches at most one gene in G_2. As the gene family size increases, so does the number of possible matches and, hence, the chance occurrence of gene clusters. Failure to account for gene families can lead to overestimation of gene cluster significance.

In this paper, we develop computationally tractable statistical tests for r-window clusters obtained by window sampling, assuming a many-to-many mapping between homologous genes. We focus on the case where the goal is to establish the homology of a specific pair of genomic regions. One of the key features of our model is that it does not require a complete comparative map to calculate cluster significance. Detailed information about gene content is only required within the regions of interest. Outside of the gene cluster, only global properties of genomes are needed.

In developing gene cluster probabilities that depend on the distribution of gene family sizes, we show that allowing arbitrary gene family size distributions leads to an expression of cluster significance that reduces to an NP-complete problem (Garey and Johnson 1979). In order to obtain tractable analytic expressions of cluster significance, we impose the assumption that all gene families are of a fixed size, ϕ. We present easily computable expressions for determining the significance of both orthologous and paralogous gene clusters under this assumption.

The fixed-size gene family assumption is highly unrealistic. Gene families clearly vary in size and a number of studies have presented evidence that gene family size follows a power-law distribution (Qian et al. 2001; Rzhetsky and Gomez 2001; Karev et al. 2002; Koonin et al. 2002; Kaplan et al. 2004). In order to test the utility of the fixed-size approximation, we used Monte Carlo simulation to estimate cluster probabilities in genomes containing power-law distributed gene families. Surprisingly, the probabilities of simulated clusters based on this more realistic gene family model are well approximated by the probabilities obtained using our analytical expressions for the fixed-size gene families. Thus, our approximations offer a satisfactory balance between speed and accuracy and are suitable for practical analyses.

Methods

In our model, the genes in a genome are partitioned into nonintersecting, fully connected families. In other words, every gene in family f_i is homologous to all the other genes in f_i and only to those genes. We define the gene family size, ϕ_{ij}, as the number of genes in G_j belonging to family, f_i. In the presence of gene families, we extend the definition of an r-window gene cluster to be a pair of windows, W_1 and W_2, of size r_1 and r_2, respectively, sharing m "gene families," where $m \leq \min(r_1, r_2)$. Note that two windows share a gene family if each window has at least one member from the gene family. No additional weight is given to multiple shared pairs from the same gene family.

Orthologous Clusters for Arbitrary Gene Families

Here, we derive analytical expressions for computing the probability of observing a gene cluster under the null hypothesis of random gene order, using the number of shared gene families as our test statistic. We first consider the orthologous gene cluster scenario, where a pair of windows is sampled from two different genomes (Durand and Sankoff 2003). We then extend this model to the paralogous case, where both windows are sampled from a single genome. Let $\mathcal{F} = \{\mathcal{F}_i\}$ be the set of gene families represented

in genomes G_1 and G_2 and let $n_j = \sum_i \phi_{ij}$ be the number of genes in G_j. Let $\mathcal{F}^k = \{F\}$ be the set of all subsets of $\mathcal{F}$ containing "exactly" k gene families. The probability that W_1 and W_2 share at least m gene families is

$$q^o(m) = \sum_{k=m}^{\min(r_1,r_2)} \left[\sum_{F \in \mathcal{F}^k} p_1(F) \sum_{l=m}^{k} \sum_{\substack{E \in \mathcal{F}^k \\ E \subseteqq F}} p_2^o(E) \right]. \quad (1)$$

The first term $p_1(F)$ is the probability that a given set, F, of k gene families is seen in W_1, and the second term, $p_2^o(E)$, is the conditional probability that at least l of the families in F also appear in W_2. The superscript o in $q_2^o(m)$ and $p_2^o(E)$ indicates that these terms refer to orthologous cluster probabilities. The superscript p will be used for paralogous clusters. Because the first term, $p_1(F)$, is the same for both orthologous and paralogous cluster probabilities, it does not require a superscript.

Note that both probability terms in equation (1) depend on the probability that a window of a given size contains a certain number of gene families. Let w be the window size and λ be the number of gene families. This probability depends on the number of ways of selecting w genes from genome G_j, such that λ gene families are represented. Let x_i be the number of genes from ith gene family, f_i. We seek the number of possible ensembles $\{x_1, x_2, \ldots, x_\lambda\}$, such that

$$\sum_{i=1}^{\lambda} x_i = w, \quad (2)$$

and $1 \leq x_i \leq \phi_{ij}$, $\forall i$.

In preliminary work, we derived a general solution using generating functions (Raghupathy and Durand 2005). This solution makes no restriction on the gene family size distribution. However, there is little hope of finding an efficient, exact solution, as the problem of enumerating all ensembles that satisfy equation (2) can be stated as a variant of the subset sum problem, which is NP-complete (Cormen et al. 1990). In the subset sum problem, given a finite set $S_{\mathbb{N}}$ and a target $t \in \mathbb{N}$, we ask whether there exists a subset $S' \subseteqq S$ whose elements sum to t. In the problem considered here, the set S corresponds to the cardinalities of the possible contributions of the gene families to the window, and the target corresponds to the window size, w. In addition to the solution to equation (2), other aspects of computing the probability for general gene families are computationally demanding. In particular, enumeration of the set of all subsets of $\mathcal{F}$ containing exactly k gene families is prohibitively slow.

In this section, we address this problem by deriving computationally tractable, approximate methods to estimate cluster significance, assuming that the gene family size is fixed. In Results, we demonstrate, using simulation that, despite the extreme nature of the assumption, our expressions yield a good approximation of cluster probabilities under a more realistic gene family model.

Orthologous Clusters, Fixed-Size Families

The complexity of calculating $q^o(m)$ can be substantially reduced under the assumption that all gene families are of equal size, ϕ. Under this assumption, it is not necessary to enumerate $\mathcal{F}^k$, because all subsets of k gene families are indistinguishable. We can instead replace the first term, $\sum_F p_1(F)$, in equation (1) with the product of two quantities. The first quantity is the number of sets of k gene families, $\binom{n_f}{k}$, where $n_f = |\mathcal{F}|$. The second is the probability that exactly k gene families of size ϕ are represented in the window, $p_1(k)$. Invoking a similar transformation of the second term in equation (1), the probability that W_1 and W_2 share at least m gene families simplifies to

$$q^o(m) = \sum_{k=m}^{\min(r_1,r_2)} \left[\binom{n_f}{k} p_1(k) \sum_{l=m}^{k} \binom{k}{l} p_2^o(l|k) \right]. \quad (3)$$

Under the fixed-size assumption, $p_1(k)$ and $p_2^o(l)$ correspond to the probability that exactly k families appear in W_1 and exactly l families appear in W_2, respectively. Because, in both cases, the probability of observing a certain number of families in a window of a particular size is required, we first derive a general expression for the probability that a window of size w contains exactly λ gene families. Let $\mathcal{T}$ be a set of λ gene families of fixed size, ϕ. Given the sample space of all sets of w genes sampled from G, we wish to determine the number of sets that contain at least one gene from each family in $\mathcal{T}$. Because our cluster definition does not take into account the order of genes in a window, this enumeration is equivalent to $\mathcal{N}(w, \lambda, \mathcal{T})$, the number of ensembles satisfying equation (2), when all families are of fixed size, ϕ.

To determine the number of such ensembles, we note that the contribution of the ith family in $\mathcal{T}$ to the window is represented by the generating function

$$\alpha_i(t) = \binom{\phi}{1} t + \binom{\phi}{2} t^2 + \cdots + \binom{\phi}{\phi} t^\phi. \quad (4)$$

The coefficient of t^x in $\alpha_i(t)$, denoted by $[t^x]\alpha_i(t)$, represents the number of ways of choosing x genes from ith family. The contributions of all λ families to the window can then be derived from the product of their generating functions, $\alpha(t) = \prod_i \alpha_i(t)$. Because the generating functions for all $\alpha_i(t)$ are identical, this product is

$$\alpha(t) = \left[\binom{\phi}{1} t + \binom{\phi}{2} t^2 + \cdots + \binom{\phi}{\phi} t^\phi \right]^\lambda. \quad (5)$$

The coefficient $[t^w]\alpha(t)$ gives the number of ways of selecting w genes such that at least one gene from each of the λ families is represented in the sample:

$$[t^w]\alpha(t) = \sum_{\{(x_1,\ldots,x_\lambda)\}} \binom{\phi}{x_1} \binom{\phi}{x_2} \cdots \binom{\phi}{x_\lambda}, \quad (6)$$

where the sum is over the set of all λ-tuples $(x_1, \ldots, x_\lambda)$ satisfying equation (2), under the constraint that $0 < x_i \leq \phi$, $\forall i$. Note that the right-hand side of equation (6) is the number of ensembles containing x_1 genes from the first family, x_2 genes from the second family, and so forth, where $\binom{\phi}{x_i}$ corresponds to the number of ways of choosing x_i genes

from the ith gene family. This is exactly the quantity $\mathcal{N}(\omega, \lambda, \mathcal{T})$.

We can avoid enumerating the set of all λ-tuples using the following simplification: Because a binomial series is of the form

$$(1 + t)^{\phi} = 1 + \binom{\phi}{1} t + \binom{\phi}{2} t^2 + \cdots + \binom{\phi}{\phi} t^{\phi}, \tag{7}$$

the right-hand side of equation (5) is equivalent to a binomial series of the form $[(1 + t)^{\phi} - 1]^{\lambda}$. Applying two binomial expansions yields

$$\begin{aligned}\alpha(t) &= \left[(1 + t)^{\phi} - 1\right]^{\lambda} = (-1)^{\lambda}\left[1 - (1 + t)^{\phi}\right]^{\lambda} \\ &= (-1)^{\lambda}\sum_{i=0}^{\lambda}\left[(-1)^{i}\binom{\lambda}{i}(t + 1)^{i\phi}\right] \\ &= (-1)^{\lambda}\sum_{u=0}^{\lambda}\left[(-1)^{u}\binom{\lambda}{u}\left(\sum_{v=0}^{u*\phi}\binom{u*\phi}{v}t^{v}\right)\right].\end{aligned} \tag{8}$$

Therefore, we obtain

$$\begin{aligned}\mathcal{N}(w, \lambda, \mathcal{T}) &= [t^{w}]\alpha(t) \\ &= (-1)^{\lambda}\sum_{u=0}^{\lambda}\left[(-1)^{u}\binom{\lambda}{u}\binom{u*\phi}{w}\right].\end{aligned} \tag{9}$$

Because no gene family can contribute more than ϕ genes to the window, at least $\lceil \frac{w}{\phi} \rceil$ gene families are required to fill the window. Using the above expression for $\mathcal{N}(w, \lambda, \phi)$ and restricting the lower bound on the dummy variable u to $\lceil \frac{r_1}{\phi} \rceil$, we obtain

$$p_1(k) = \frac{(-1)^{k}\sum_{u=\lceil \frac{r_1}{\phi} \rceil}^{k}\left[(-1)^{u}\binom{k}{u}\binom{u*\phi}{r_1}\right]}{\binom{n_1}{r}}. \tag{10}$$

We now derive a similar simplification for $p_2^{o}(l|k)$, the probability that W_2 contains exactly l of the k gene families in W_2. In this case, we seek ensembles of $r_2 = y + z$ genes, where y genes are selected from the subset of l gene families. The remaining z genes are selected from families not included in W_1. We must ensure that no genes from the remaining $k - l$ gene families appear in W_2, in order to guarantee that "exactly" l families are represented in W_2. The probability of this event is

$$p_2^{o}(l|k) = \frac{\sum_{z}(-1)^{l}\sum_{u=\lceil \frac{r_2 - z}{\phi} \rceil}^{l}\left[(-1)^{u}\binom{l}{u}\binom{u*\phi}{r_2 - z}\right]\binom{n_2 - k\phi}{z}}{\binom{n_2}{r_2}}. \tag{11}$$

The first term in the inner summation represents the number of ways of selecting $y = r_2 - z$ genes such that equation (2) is satisfied when $\lambda = l$ and $w = y$. The second term represents the number of ways of selecting the remaining z genes. The value of z in the outer summation ranges from $\max\{0, r_2 - k\phi\}$ to $r_2 - l$.

Paralogous Clusters, Fixed-Size Gene Families

We can extend these results to obtain a measure of significance for paralogous gene clusters. Recall that in the paralogous case, two windows, W_1 and W_2, are nonoverlapping windows sampled from the same genome G. The probability that they share at least m gene families is

$$q^{p}(m) = \sum_{k=m}^{r}\left[\binom{n_f}{k}p_1(k)\sum_{l=m}^{k}\binom{k}{l}p_2^{p}(l|k)\right], \tag{12}$$

where $p_1(k)$ is given in equation (10). The probability for the second window differs from equation (11), however, because the fact that both windows are sampled from the same genome further constrains the set of possible ensembles for the second window.

To calculate $p_2^{p}(l|k)$, we need an expression for the number of ensembles of y genes containing exactly l gene families. Let x_{ij} denote the number of genes from the ith family that appear in W_j. (Unlike the orthologous case, x_{i1} and x_{i2} are drawn from the same genome but different windows.) Because W_2 is in the same genome as W_1, only $\phi'_j = \phi - x_{i1}$ genes from the ith family are available to fill the second window. Thus, in the paralogous case, equation (6) becomes

$$\begin{aligned}t^{w}[\alpha(t)] &= \mathcal{N}_p(\omega, l, \mathcal{T}) \\ &= \sum_{\mathcal{S}}\binom{\phi - x_{11}}{x_{12}}\binom{\phi - x_{21}}{x_{22}}\cdots\binom{\phi - x_{l1}}{x_{l2}},\end{aligned} \tag{13}$$

where $\mathcal{S}$ is the set of all l-tuples $(x_{12}, \ldots, x_{l2})$ such that

$$\sum_{i=1}^{l} x_{i2} = w,$$

and $0 < x_{i2} \leq \phi - x_{i1}$.

In the case of orthologous clusters, the assumption of fixed-size gene families simplified the enumeration of ensembles for both W_1 and W_2. However, for paralogous clusters, the fixed-size assumption does not hold when calculating the probability for W_2, as we do not know how many genes from a given gene family are represented in W_1. Because the number of genes available from each family to fill W_2 is no longer fixed, we cannot simplify the enumeration of $\mathcal{S}$ in the same way. In order to obtain a tractable expression, we make the further assumption that the number of genes from each gene family available to fill W_2 is fixed as well. We define, $\phi' = \phi - \bar{x}$, where $\bar{x}$ is an estimate of the mean number of genes contributed by each gene family in W_1.

$$p_2^{\mathrm{p}}(l|k) = \frac{\sum_z N(r_2 - z, l, \phi') \binom{n - k\phi}{z}}{\binom{n - r_1}{r_2}} = \frac{\sum_z (-1)^l \sum_{u=\lceil \frac{r_2 - z}{\phi'} \rceil}^{l} \left[(-1)^u \binom{l}{u} \binom{u * \phi'}{r_2 - z} \right] \binom{n - k\phi}{z}}{\binom{n - r_1}{r_2}}, \quad (14)$$

Applying the same argument used to derive equation (9) from equation (6), we obtain the following approximation: where z ranges from $\max\{0, r_2 - k\phi\}$ to $r_2 - l$.

The fixed-size approximation and the use of generating functions to determine the number of ensembles satisfying equation (2) result in expressions that require only simple summations (eqs. 10, 11, and 14). These expressions constitute an efficient approximation to the probability that two arbitrarily chosen windows share at least m gene families under the null hypothesis of random gene order. The results provide statistical tests for orthologous and paralogous gene clusters that depend only on the size of the conserved regions, the number of shared homologous genes, and the total number of genes in the genome. They do not require information about the spatial organization of the genome outside the regions of interest. The equations required to estimate cluster probabilities are summarized in table 1. In the following section, we compare these general fixed-size probabilities with cluster probabilities obtained using a more realistic power-law model and show that our approximations are not only tractable but also accurate.

Results

Effect of Simplifying Assumptions on Cluster Significance

We first investigated the effect of the simplifying assumptions used to derive our significance tests. In order to determine the accuracy of our tests as an estimate for statistical significance, we compared the probabilities obtained under our fixed-size model (table 1) with probabilities of clusters in simulated genomes with a power-law gene family distribution.

The key step in performing the simulation is to construct random genomes with gene family sizes that approximate the observed distribution in real genomes. A number of studies have shown that the gene family size distribution follows a power-law of the form

$$f(x) = ax^{-b}, \quad (15)$$

where $f(x)$ is the number of gene families of size x, and a and b are constants (Qian et al. 2001; Rzhetsky and Gomez 2001; Karev et al. 2002; Koonin et al. 2002; Kaplan et al. 2004). We modeled our simulated genomes on three eukaryotic genomes, yeast, fly, and human. These genomes represent a wide range of genomes sizes and include a single cell organism, an invertebrate, and a vertebrate. We focused on eukaryotes over prokaryotes because of the greater size and complexity of gene families in eukaryotes.

To determine the gene family distribution, we obtained nonredundant, full length amino acid sequences from the yeast, fly, and human genomes from the Swiss-Prot database Version 50.9 (Gasteiger et al. 2003). For each genome, all-against-all Blast was performed to find a set of significantly similar sequence pairs (Altschul et al. 1997). E-values obtained from the Blast results were used to cluster the sequences into families using both the single and complete linkage clustering methods implemented in the hierarchical clustering package in R (R Development Core Team 2005). In both cases, we used an E-value threshold of 10^{-4}, which is suitable for identifying both orthologs and paralogs. Consistent with previous studies (Qian et al. 2001; Rzhetsky and Gomez 2001; Karev et al. 2002; Koonin et al. 2002; Kaplan et al. 2004), the resulting gene family size distributions approximated a power-law. Because distributions obtained using single and complete linkage clustering were similar, we only present results obtained by single linkage here. Power-law distribution parameters were obtained by fitting the observed gene family size distributions to equation (15) and are given in table 2.

For orthologous cluster probabilities, for each genome size, an artificial genome with n genes was constructed such that the gene family size distribution follows the power-law parameters given in table 2. We used genome sizes of $n =$ 5,000, 14,000, and 22,000, which correspond roughly to the number of genes in yeast, fly, and human genomes, respectively (http://www.ensembl.org). We next generated a pool of $N = 35{,}000$ random permutations for each of these genomes. In each simulation, two genomes were selected randomly from the pool of random permutations. A window of size r was then chosen at random from each of these two genomes, and the number of gene family matches (m) between the windows was tabulated. The probability of observing exactly m matches was estimated by sampling 25,000 window pairs. The probability of observing at least m matches was calculated from the resulting distribution.

The probabilities obtained using simulation were compared with the results of our analytical approximation for $r = 50$, $r = 100$, and $r = 150$, typical window sizes in empirical studies (Lundin 1993; Katsanis et al. 1996; Coulier et al. 1997; Endo et al. 1997; Kasahara 1997; Ruvinsky and Silver 1997; Amores et al. 1998; Hughes 1998; Pebusque

Table 1
Expressions for Computing Cluster Significance with Constant Gene Family Size

(a) Orthologous gene clusters		
$q^{o}(m)$	Equation (3)	$\sum_{k=m}^{\min(r_1,r_2)}\left[\binom{n_f}{k}p_1(k)\sum_{l=m}^{k}\binom{k}{l}p_2^{o}(l\|k)\right]$
$P_1(k)$	Equation (10)	$\binom{n_1}{r_1}^{-1}(-1)^k\sum_{u=\lceil\frac{r_1}{\phi}\rceil}^{k}\left[(-1)^u\binom{k}{u}\binom{u*\phi}{r_1}\right]$
$P_2{}^{o}(l\|k)$	Equation (11)	$\binom{n_2}{r_2}^{-1}\sum_z(-1)^l\sum_{u=\lceil\frac{r_2-z}{\phi}\rceil}^{l}\left[(-1)^u\binom{l}{u}\binom{u*\phi}{r_2-z}\right]\binom{n_2-k\phi}{z}$
(b) Paralogous gene clusters		
$q^{p}(m)$	Equation (12)	$\sum_{k=m}^{\min(r_1,r_2)}\left[\binom{n_f}{k}p_1(k)\sum_{l=m}^{k}\binom{k}{l}p_2^{p}(l\|k)\right]$
$P_1(k)$	Equation (10)	$\binom{n_1}{r_1}^{-1}(-1)^k\sum_{u=\lceil\frac{r_1}{\phi}\rceil}^{k}\left[(-1)^u\binom{k}{u}\binom{u*\phi}{r_1}\right]$
$p^{p}{}_2(l\|k)$	Equation (14)	$\binom{n-r_1}{r_2}^{-1}\sum_z(-1)^l\sum_{u=\lceil\frac{r_2-z}{\phi'}\rceil}^{l}\left[(-1)^u\binom{l}{u}\binom{u*\phi'}{r_2-z}\right]\binom{n-k\phi}{z}$

et al. 1998; Smith et al. 1999; Lipovich et al. 2001; Spring 2002). For each value of n, the orthologous gene cluster probabilities under the fixed-gene family size assumption were computed using the equations in table1*a*, with $\phi = 2$ and $\phi = 3$.

The results, given in figure 2, supplementary figures S1, S2, and S3, Supplementary Material online, show that the probabilities obtained using our simplifying assumptions closely approximate the simulated cluster probabilities obtained with the power-law size distribution. When $n = 5{,}000$ and $n = 14{,}000$, the probabilities obtained with $\phi = 2$ slightly overestimate the simulated cluster probabilities, for all window sizes considered. Similarly, when both genomes have 22,000 genes, the estimated probabilities obtained with $\phi = 3$ slightly overestimate the power-law based probabilities. Moreover, the estimated probabilities obtained with $\phi = 2$ and with $\phi = 3$ give lower and upper bounds on the true probability, making it possible to estimate the magnitude of the error that can result from using this approximation. These guidelines hold for genomes of different sizes (supplementary fig. S3, Supplementary Material online). It is only necessary to use $\phi = 3$ when both genomes have more than 25,000 genes. If one or both of the genomes is smaller, $\phi = 2$ suffices.

These results suggest that an accurate, conservative approximation can be obtained using the equations in table 1*a* with $\phi = 2$ for small to medium-sized genomes and with $\phi = 3$ for larger genomes. These approximations slightly underestimate the significance, guarding against false positives. Moreover, for the parameter values we considered, the set of clusters deemed significant is frequently unaffected by the use of approximation. Even though the probabilities obtained with the power-law and fixed-size models differ, for many values of m and significance thresholds α, the same clusters will be rejected by both models. For example, when $n = 22{,}000$ and $r = 50$, both models will reject the null hypothesis for clusters of size 4, but not size 3, at the $\alpha = 0.001$ significance level. In the remaining cases, the number of matches required to make a cluster significant for a given window size is overestimated by at most one. For example, when $n = 22{,}000$, $r = 100$, and $\alpha = 0.001$, the fixed-size approximation would rule out a cluster of size 6, although the power-law model indicates it is significant. However, the fixed-size approximation will accept clusters of size 7 and greater.

Table 2
Power-Law Parameters of Gene Family Size Distributions of the Three Genomes Species Obtained Using Single-Linkage Clustering under *E* Value Threshold 10^{-04}

Genome	a	b
Yeast	2,435	2.73
Fly	803	2.76
Human	2,300	2.28

We also evaluated the accuracy of the approximations for paralogous clusters given in table 1*b*. The paralogous case required a second simplifying assumption, namely, replacing $\phi - x_{i1}$ with $\phi' = \phi - \bar{x}$ in equation (14). As described in the supplementary text, Supplementary Material online, we confirmed by simulation that using $\phi' = \phi - 1$ has little effect on the estimated cluster probability for the parameter values investigated. The use of $\bar{x} = 1$ reflects the assumption that in a random genome, the appearance of more than one gene from a given family in a window of size r is a rare event when $r \ll n$ and $\phi \ll n$. We next investigated the impact of the fixed-size approximation on paralogous cluster probabilities, by comparing the approximation in table 1*b* with the simulated power-law model. The simulation procedure for estimating paralogous cluster probabilities was identical to that for orthologous clusters, with the exception that in each simulation two random nonoverlapping windows were sampled from a single random genome chosen from a pool of N random genomes.

Figure 3 and supplementary figures S4 and S5, Supplementary Material online, show the paralogous gene cluster significance obtained by the simulation compared with that obtained using the equations in table 1*b* with $\phi = 2$ and $\phi = 3$. The results are similar to those observed with

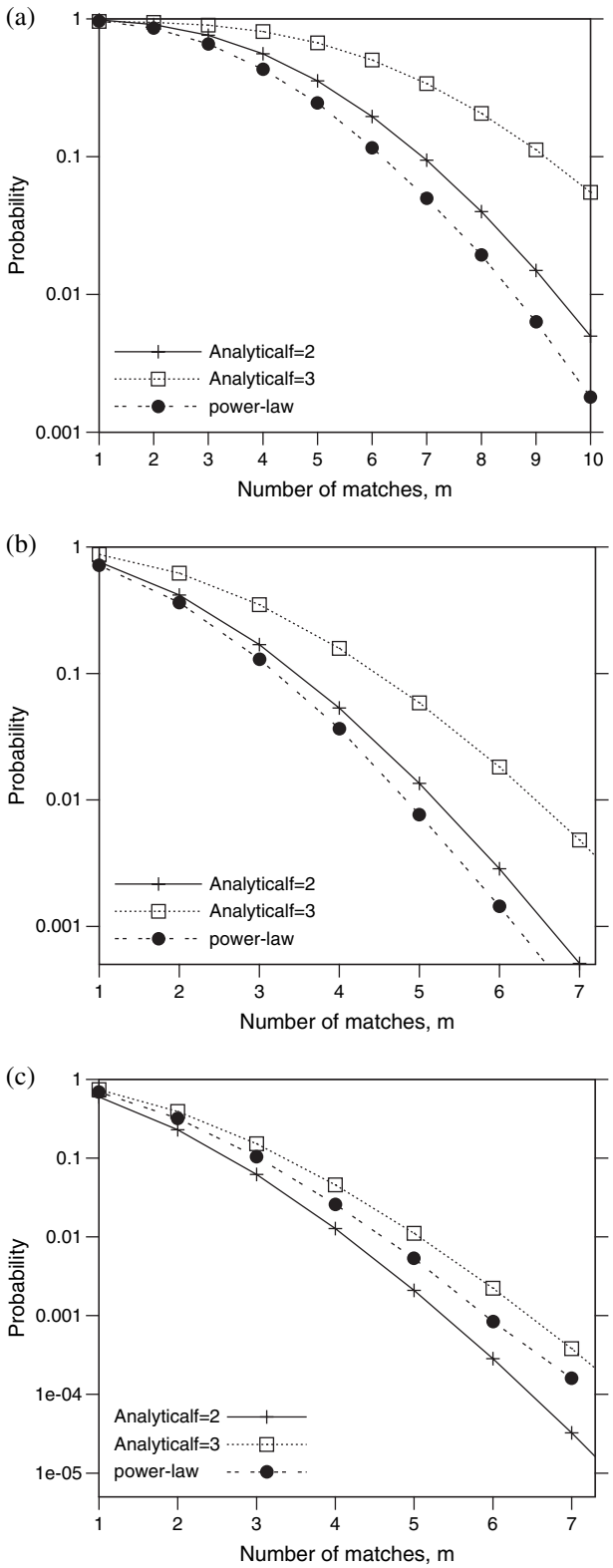

FIG. 2.—Comparison of orthologous cluster probabilities for power-law distributed and fixed-gene family sizes. The probabilities of observing at least one cluster of size m in a window of size r = 100 when genome size (*a*) n = 5,000, (*b*) n = 14,000, and (*c*) n = 22,000.

orthologous clusters. The probabilities obtained using ϕ = 2 provide an accurate, conservative approximation when n = 5,000 and n = 14,000. For larger genomes, an accurate, conservative approximation can be obtained with ϕ = 3.

Importance of a Many-to-Many Homology Model

Most previously published tests for gene clustering do not model gene families, instead assuming that each gene is homologous to exactly one other gene. We investigated the importance of many-to-many homology model in estimating cluster significance by comparing probabilities obtained with the one-to-one and power-law distributed homology models. Cluster probabilities for one-to-one homology were calculated using the test based on the hypergeometric function proposed by Durand and Sankoff (2003) (eq. 22 in that paper). These were compared with the cluster probabilities obtained using simulation with power-law distributed gene family sizes, described above.

Figure 4 and supplementary S6, Supplementary Material online, show that the one-to-one mapping assumption underestimates cluster probabilities and, hence, overestimates cluster significance. This problem is particularly severe for larger genome and window sizes. For example, compare the number of matches required to reject the null hypothesis at a significance level of 0.001, when n = 22,000 and r = 150 (fig. 4*b*). Under the one-to-one assumption, an experimenter observing six homologous pairs would erroneously conclude that the cluster was significant. In contrast, at least nine homologous pairs are required in order to reject the null hypothesis under the more realistic power-law model. Therefore, unless the experimenter is able to unambiguously identify a unique homolog for each match, the use of the one-to-one homology model will lead to false positives. This effect also occurs for smaller genome and/or window sizes, but is less pronounced.

The Influence of Window Size on Significance

In addition to data analysis, our equations can be used to analyze trends in cluster significance, evaluate the impact of parameter choices on cluster probabilities, and to design data analysis protocols. For example, how should the window size, r, be selected in a window sampling analysis? We studied the effect of window size, r, on orthologous gene cluster significance by computing the probability of observing a gene cluster for various values of r using the equations in table 1*a*. Because paralogous gene cluster probabilities follow similar trends, only results for orthologous clusters are given.

Figure 5 shows that for given values of n, m, and ϕ, the cluster significance decreases as the size of windows increases. When the number of conserved homologous pairs in a cluster is small, the cluster is significant only for a small range of window sizes. For example, when n = 5,000, ϕ = 2, and m = 5, at a significance threshold of α = 0.0001, clusters are significant only when $r \leq 36$ (fig. 5*a*). As the number of conserved homologous pairs grows, the clusters are significant for a wider range of window sizes.

When the cluster contains 10 homologous pairs, it is significant when $r \leq 79$. Similar trends were found for the

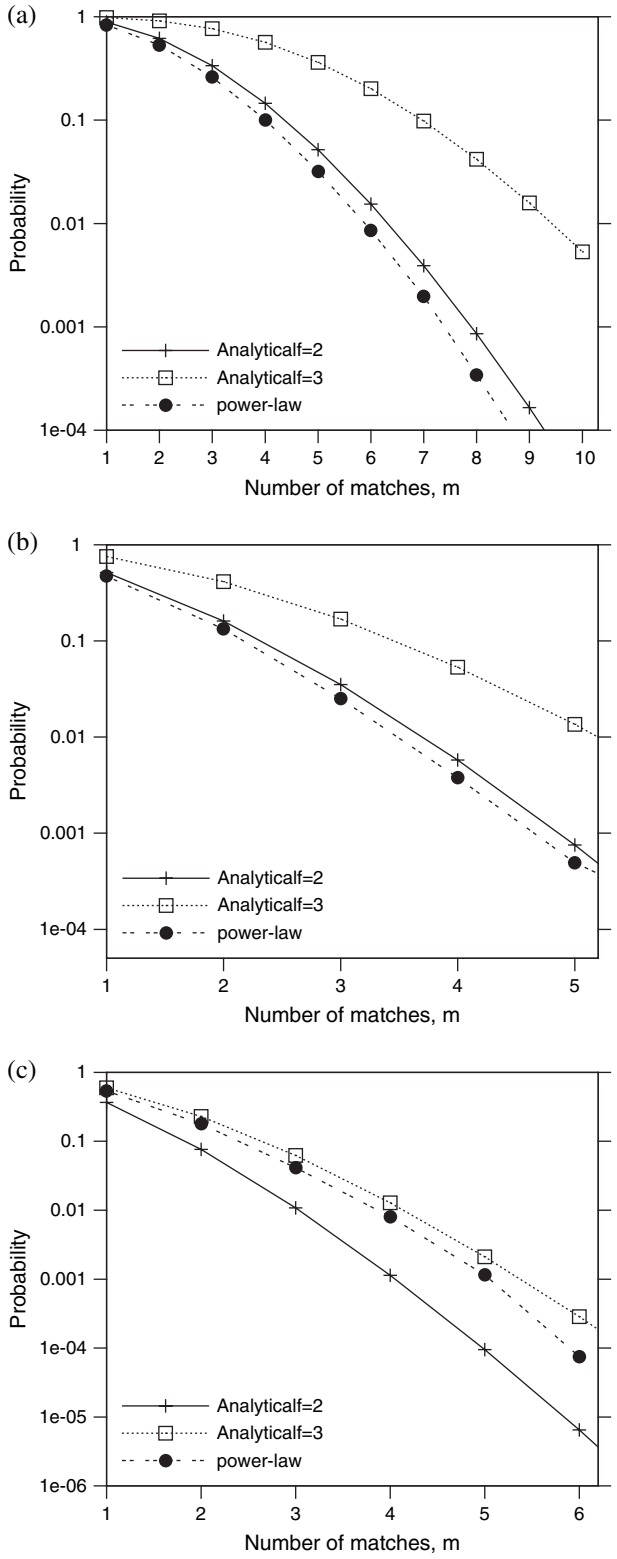

FIG. 3.—Comparison of paralogous cluster probabilities for power-law distributed and fixed-size gene families. The probabilities of observing at least one cluster of size m in a window of size $r = 100$ when genome size (*a*) $n = 5{,}000$, (*b*) $n = 14{,}000$, and (*c*) $n = 22{,}000$.

larger genome sizes as shown in figure 5*b*. In addition, when the genome size is large, fewer homologous pairs are required to make a cluster significant for a given window size. For example, when $n = 22{,}000$, a cluster of five genes is significant when $r \leq 60$, approximately.

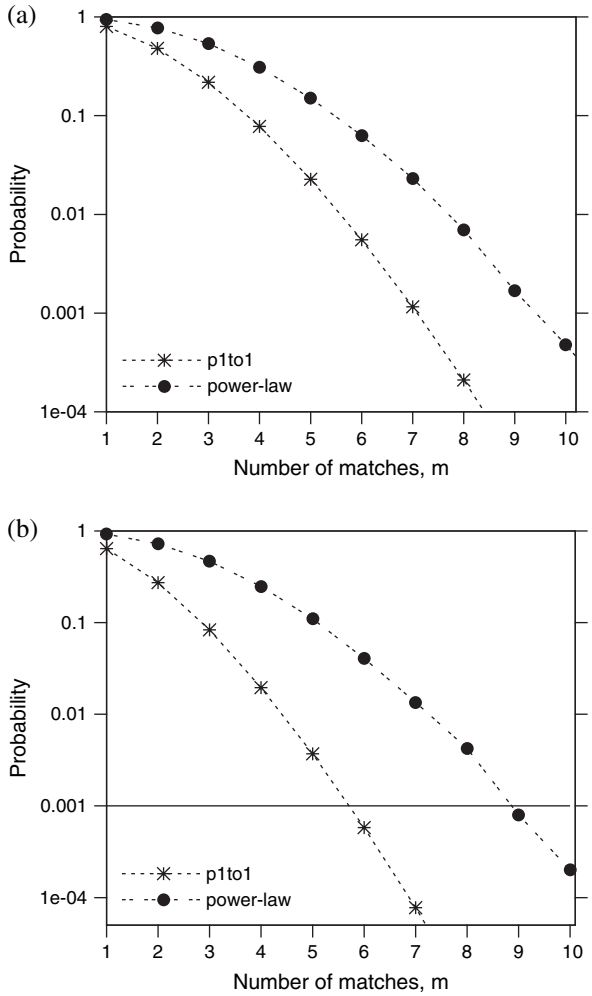

FIG. 4.—Comparison of orthologous cluster probabilities for power-law distributed gene families and the one-to-one homology model. The probabilities of observing at least one cluster of size m in a window of size $r = 150$ when genome size (*a*) $n = 14{,}000$ and (*b*) $n = 22{,}000$.

In the last 15 years, many reports of both paralogous and orthologous conserved gene clusters have appeared (surveyed in Abi-Rached et al. 2002; Danchin et al. 2003; Durand and Sankoff 2003). These clusters typically include 5 to 15 homologous pairs, with window sizes ranging from 15 to 300. The results in figure 5*b* suggest that for the larger window sizes even 15 homologous pairs may not be sufficient to reject the null hypothesis.

Application to a Real Example

Geddy and Brown (2007) used spatial genomic analysis in a recent study of the evolution and functional diversification of genes encoding pentatricopeptide repeat proteins (PPR) in plant genomes. PPR proteins are associated with various RNA processing functions, including processing of RNA transcripts, RNA editing, and initiation of translation. Some are also implicated in plant-specific functions, such as restoration of fertility. Loss of fertility is due to mitochondrially encoded

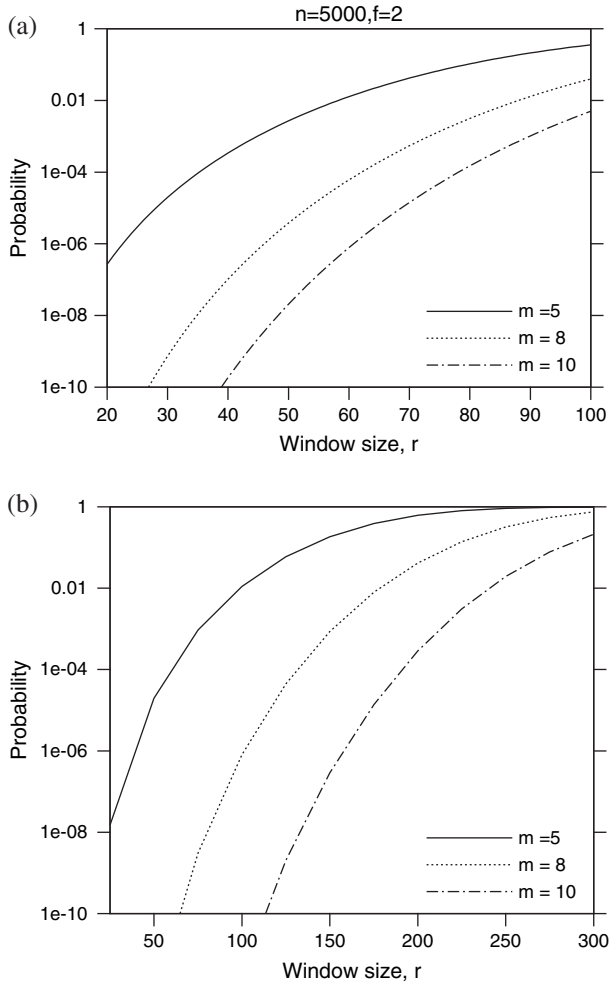

FIG. 5.—The effect of window size on orthologous cluster significance. The probabilities of observing at least one cluster of size m in a window of size r given a genome of size n when (*a*) $n = 5{,}000$, $\phi = 2$, and $m = 5$, $m = 8$, and $m = 10$; (*b*) $n = 22{,}000$, $\phi = 3$, and $m = 5$, $m = 10$, and $m = 15$.

cytoplasmic male sterility genes observed in a number of plant species, including radish and petunia.

The spatial organization of PPR genes is highly variable compared with the relatively stable syntenic organization of other genes in the genomic regions in which they are found (Geddy and Brown 2007). In their investigation of the genomic processes driving the distribution of these genes, Geddy and Brown (2007) present partially conserved gene clusters containing PPR genes. The hypothesis that these regions are descended from the same region in an ancestral genome could have been further supported by statistical validation using tests such as those presented here.

To demonstrate the relevance of our methods to current genomic studies, we applied our statistics to two of the clusters containing PPR genes identified by Geddy and Brown (2007). The first is an orthologous cluster (fig. 3 in Geddy and Brown 2007) comprised of regions from the *Arabidopsis* and Ogura radish genomes. The cluster spans 15 genes in the *Arabidopsis* genome and 6 genes in Ogura radish. Of these genes, four are homologous pairs appearing in both regions. We computed the probability of observing such an orthologous cluster using equation (3) with $\phi = 3$, assuming $n = 28{,}000$ (http://www.arabidopsis.org, RadishDB: http://radish.plantbiology.msu.edu/). The resulting probability is 6.45×10^{-11}, showing that the cluster is statistically significant.

The second cluster is a paralogous cluster (fig. 4 in Geddy and Brown 2007) of two genomic regions containing PPR genes in the *Arabidopsis* genome. These regions contain 19 and 18 genes, respectively, and share 8 homologous pairs. The probability of observing such a paralogous cluster under the null hypothesis is 8.96×10^{-20}, computed using equation (12) with $\phi = 3$. This suggests that the cluster is highly significant. Thus, our statistical analysis provides further evidence of the shared ancestry of the gene clusters reported by Geddy and Brown (2007).

This example underscores the importance of two key features of our tests: They can be applied when one-to-one homology cannot be determined and when whole genome data are not available. A many-to-many homology model is particularly important for the PPR genes because of the difficulty of determining exact homology relationships in this gene family. Evidence that these genes are under diversifying selection contradicts the usual expectation that genes that are most closely related will also be most similar. Moreover, the ability to obtain accurate sequence alignments is challenged by the presence of repeated sequence motifs within these genes. Our methods are also particularly well suited to analysis of these data because the radish genome is not completely sequenced. Thus, statistical tests based on randomization of gene order could not have been applied.

Discussion

Identification of homologous genomic regions is a fundamental component of genome evolution studies, as well as predictive methods that exploit spatial conservation for functional inference. In distantly related genomes, putative homologous regions are found through similarities in local gene content. When spatial organization has been disrupted by genome rearrangements, statistical tests are essential to exclude the possibility that such similarities arose by chance. Although there is a growing statistical methodology for validating gene clusters, practical significance tests that are applicable to noisy and incomplete data have not been realized.

Here, we present accurate, efficient statistical tests that meet these needs in two ways: First, our results are appropriate for studies that focus on a single pair of regions containing specific genes of interest. Because they do not require detailed genomic information outside the region of interest, our methods can be used to analyze homologous regions in species for which a genomic map is not available, either because genome sequencing and assembly has not been completed or because the organism under study has not been targeted for genome sequencing. Such data sets are not amenable to statistical tests based on randomization.

Second, our tests support a many-to-many homology model, applicable to genome self-comparison and data sets with large gene families. The challenge is to model the distribution of gene family sizes to obtain a test that is both efficient and accurate. Exact calculation of the cluster

probabilities assuming an arbitrary distribution is computationally intractable. By assuming that all gene families are of equal size, we obtain an efficient test that can be easily calculated in Mathematica. To evaluate the impact of this simplifying assumption, we used simulation to estimate gene cluster probabilities under the null hypothesis using a realistic model of gene family size distributions. These were obtained by fitting a power law to clustered sequences from the yeast, fly, and human genomes. Remarkably, the results show that our tests closely approximate the null hypothesis, despite the highly unrealistic assumption on which they are based. Comparing the simulated probabilities with our analytical model shows that our tests slightly overestimate cluster probabilities, yielding a test that is accurate and conservative. We also compared previously published tests (Durand and Sankoff 2003) that assume one-to-one homology with our simulation results. The probabilities obtained by assuming one-to-one homology substantially underestimate the cluster probabilities in the simulated genomes leading to erroneous rejection of the null hypothesis. This confirms the need for statistical tests that include a model of gene families.

Our results represent a practical balance between accuracy and efficiency that is well suited to analysis of real biological data sets. We demonstrate the utility of our results empirically by applying them to gene clusters in the *Arabidopsis* and radish genomes from a recent report on the evolution of the pentatricopeptide repeat (PPR) gene family in plants. This data set exemplifies the two practical advantages of our tests. A many-to-many homology model is required for the PPR genes because determining phylogenetic relationships is difficult in this family due to repeated motifs and diversifying selection. In addition, a method suitable for local regions is required because a complete assembly of the radish genome sequence is not yet available.

Supplementary Material

Supplementary figures S1–S6 and supplementary text are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank P. Chebolu and A. Frieze for helpful discussions and A. Goldman, R. Hoberman, N. Song, B. Vernot, R. Sedgewick, and J. Joseph for their help in performing the simulation experiments. This research was supported by NIH grant 1 K22 HG 02541-01, a David and Lucile Packard Foundation Fellowship, and a Pittsburgh Supercomputing Center Computational Facilities Access Grant MCB000010P.

Literature Cited

Abi-Rached L, Gilles A, Shiina T, Pontarotti P, Inoko H. 2002. Evidence of en bloc duplication in vertebrate genomes. Nat Genet. 31(1):100–105.

Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D. 1997. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res. 25(17):3389–3402.

Amores A, Force A, Yan Y, et al. (12 co-authors). 1998. Zebrafish hox clusters and vertebrate genome evolution. Science. 282:1711–1714.

Bourque G, Yasef Y, El-Mabrouk N. 2005. Maximizing synteny blocks to identify ancestral homologs. In: McLysaght A, Huson DH, editors. RECOMB 2005 Workshop on Comparative Genomics. Berlin Heidelberg: Springer Verlag. p. 21–34.

Calabrese P, Chakravarty S, Vision T. 2003. Fast identification and statistical evaluation of segmental homologies in comparative maps. Bioinformatics. 19(Suppl 1):i74–i80.

Chen F, Mackey AJ, Vermunt JK, Roos DS. 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS ONE. 2:e383.

Chen X, Su Z, Dam P, Palenik B, Xu Y, Jiang T. 2004. Operon prediction by comparative genomics: an application to the *Synechococcus* sp. WH8102 genome. Nucleic Acids Res. 32(7):2147–2157.

Cormen T, Leiserson C, Rivest R. 1990. Introduction to algorithms. Cambridge: MIT Press/McGraw-Hill.

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

Danchin E, Abi-Rached L, Gilles A, Pontarotti P. 2003. Conservation of the MHC-like region throughout evolution. Immunogenetics. 55(3):141–148.

Danchin E, Pontarotti P. 2004. Statistical evidence for a more than 800-million-year-old evolutionarily conserved genomic region in our genome. J Mol Evol. 59(5):587–597.

Dandekar T, Snel B, Huynen M, Bork P. 1998. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci. 23(9):324–328.

Dehal P, Boore J. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. 3(10):e314.

Durand D, Sankoff D. 2003. Tests for gene clustering. J Comput Biol. 10(3–4):453–482.

Endo T, Imanishi T, Gojobori T, Inoko H. 1997. Evolutionary significance of intra-genome duplications on human chromosomes. Gene. 205(1–2):19–27.

Fitch W. 2000. Homology: a personal view on some of the problems. Trends Genet. 16(5):227–231.

Garey M, Johnson D. 1979. Computers and Intractability: a guide to the theory of NP-completeness. New York: W.H. Freeman.

Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel R, Bairoch A. 2003. ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res. 31(13): 3784–3788.

Geddy R, Brown G. 2007. Genes encoding pentatricopeptide repeat (ppr) proteins are not conserved in location in plant genomes and may be subject to diversifying selection. BMC Genomics. 8:130.

Hoberman R, Durand D. 2005. The incompatible desiderata of gene cluster properties. In: McLysaght A, Huson DH, editors. Comparative genomics. RECOMB 2005 International Workshop, Vol. 3678 Lecture Notes in Bioinformatics. Springer-Verlag. p. 73–87.

Hoberman R, Sankoff D, Durand D. 2005. The statistical analysis of spatially clustered genes under the maximum gap criterion. J Comput Biol. 12(8):1081–1100.

Homma K, Fukuchi S, Nakamura Y, Gojobori T, Nishikawa K. 2007. Gene cluster analysis method identifies horizontally transferred genes with high reliability and indicates that they provide the main mechanism of operon gain in 8 species of gamma-Proteobacteria. Mol Biol Evol. 24(3):805–813.

Hughes A. 1998. Phylogenetic tests of the hypothesis of block duplication of homologous genes on human chromosomes 6, 9, and 1. Mol Biol Evol. 15(7):854–870.

Hurst L, Pál C, Lercher M. 2004. The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet. 5(4):299–310.
Huynen M, Bork P. 1998. Measuring genome evolution. Proc Natl Acad Sci USA. 95(11):5849–5856.
Kaplan N, Friedlich M, Fromer M, Linial M. 2004. A functional hierarchical organization of the protein sequence space. BMC Bioinformatics. 5:196.
Karev G, Wolf Y, Rzhesky A, Berezovskaya F, Koonin E. 2002. Birth and death of protein domains: a simple model of evolution explains power law behavior. BMC Evol Biol. 2:18–43.
Kasahara M. 1997. New insights into the genomic organization and origin of the major histocompatibility complex: role of chromosomal (genome) duplication in the emergence of the adaptive immune system. Hereditas. 127(1–2):59–65.
Katsanis N, Fitzgibbon J, Fisher E. 1996. Paralogy mapping: identification of a region in the human MHC triplicated onto human chromosomes 1 and 9 allows the prediction and isolation of novel PBX and NOTCH loci. Genomics. 35(1):101–108.
Koonin E, Wolf Y, Karev G. 2002. The structure of protein universe and genome evolution. Nature. 420:218.
Lipovich L, Lynch E, Lee M, King M. 2001. A novel sodium bicarbonate cotransporter-like gene in an ancient duplicated region: sLC4A9 at 5q31. Genome Biol. 2(4):0011.1–0011.13.
Lundin L. 1993. Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. Genomics. 16(1):1–19.
McLysaght A, Hokamp K, Wolfe K. 2002. Extensive genomic duplication during early chordate evolution. Nat Genet. 31(2):200–204.
Murphy WJ, Larkin DM, Everts-van der Wind A, et al. (25 co-authors). 2005. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. Science. 309:613–617.
Overbeek R, Fonstein M, D'Souza M, Pusch G, Maltsev N. 1999. The use of gene clusters to infer functional coupling. Proc Natl Acad Sci USA. 96(6):2896–2901.
Panopoulou G, Hennig S, Groth D, Krause A, Poustka A, Herwig R, Vingron M, Lehrach H. 2003. New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes. Genome Res. 13(6A):1056–1066.
Pebusque M, Coulier F, Birnbaum D, Pontarotti P. 1998. Ancient large-scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. Mol Biol Evol. 15(9):1145–1159.
Qian J, Luscombe N, Gerstein M. 2001. Protein family and fold occurrence in genomes: powerlaw behaviour and evolutionary model. J Mol Biol. 313:679–681.
R Development Core Team. R. 2005. A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
Raghupathy N, Durand D. 2005. Individual gene cluster statistics in noisy maps RECOMB 2005 Workshop on Comparative Genomics, Vol. 3678 LNBI. Springer-Verlag. p. 106–120.
Raghupathy R, Hoberman N, Durand D. 2008. Two plus two does not equal three: statistical tests for multiple genome comparison. J Bioinform Comput Biol. 6(1):1–22.
Ruvinsky I, Silver L. 1997. Newly indentified paralogous groups on mouse chromosomes 5 and 11 reveal the age of a T-box cluster duplication. Genomics. 40:262–266.
Rzhetsky A, Gomez SM. 2001. Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome. Bioinformatics. 17(10):988–996.
Sankoff D, Haque L. 2005. Power boosts for cluster tests. In: McLysaght A, Huson DH, editors. Comparative genomics. RECOMB 2005 International Workshop, Vol. 3678 Lecture Notes in Computer Science. Springer Verlag. p. 121–130.
Simillion C, Vandepoele K, Saeys Y, Van de Peer Y. 2004. Building genomic profiles for uncovering segmental homology in the twilight zone. Genome Res. 14(6):1095–1106.
Smith NGC, Knight R, Hurst L. 1999. Vertebrate genome evolution: a slow shuffle or a big bang. BioEssays. 21:697–703.
Spring J. 2002. Genome duplication strikes back. Nat Genet. 31:128–129.
Tamames J. 2001. Evolution of gene order conservation in prokaryotes. Genome Biol. 6(2):0020.1–0020.11.
Tamames J, Gonzalez-Moreno M, Valencia A, Vicente M. 2001. Bringing gene order into bacterial shape. Trends Genet. 3(17):124–126.
Trachtulec Z, Forejt J. 2001. Synteny of orthologous genes conserved in mammals, snake, fly, nematode, and fission yeast. Mamm Genome. 3(12):227–231.
Vandepoele K, Simillion C, Van de Peer Y. 2002. Detecting the undetectable: uncovering duplicated segments in *Arabidopsis* by comparison with rice. Trends Genet. 18(12):604–606.
Venter J, Adams M, Myers E, et al. (274 co-authors). 2001. The sequence of the human genome. Science. 291(5507):1304–1351.
Vienne A, Rasmussen J, Abi-Rached L, Pontarotti P, Gilles A. 2003. Systematic phylogenomic evidence of en bloc duplication of the ancestral 8p11.21-8p21.3-like region. Mol Biol Evol. 20(8):1290–1298.
Vision T, Brown D, Tanksley S. 2000. The origins of genomic duplications in *Arabidopsis*. Science. 290:2114–2117.
von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Krüger B, Snel B, Bork P. 2007. STRING 7–recent developments in the integration and prediction of protein interactions. Nucleic Acids Res. 35(Database issue).
Wolfe K, Shields D. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature. 387:708–713.
Zheng Y, Szustakowski J, Fortnow L, Roberts R, Kasif S. 2002. Computational identification of operons in microbial genomes. Genome Res. 12(8):1221–1230.
Zhu Q, Adam Z, Choi V, Sankoff D. 2008. Generalized gene adjacencies, graph bandwidth and clusters in yeast evolution. IEEE/ACM Trans CompBiol Bioinf.

Aoife McLysaght, Associate Editor

Accepted November 3, 2008