Individual gene cluster statistics in noisy maps

Narayanan Raghupathy[1], Dannie Durand[2]

[1] Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA. E-mail: rnarayan@cmu.edu
[2] Departments of Biological Sciences and Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. E-mail: durand@cmu.edu

Abstract. Identification of homologous chromosomal regions is important for understanding evolutionary processes that shape genome evolution, such as genome rearrangements and large scale duplication events. If these chromosomal regions have diverged significantly, statistical tests to determine whether observed similarities in gene content are due to history or chance are imperative. Currently available methods are typically designed for genomic data and are appropriate for whole genome analyses. Statistical methods for estimating significance when a single pair of regions is under consideration are needed. We present a new statistical method, based on generating functions, for estimating the significance of orthologous gene clusters under the null hypothesis of random gene order. Our statistics is suitable for noisy comparative maps, in which a one-to-one homology mapping cannot be established. They are also designed for testing the significance of an individual gene cluster in isolation, in situations where whole genome data is not available. We implemented our statistics in Mathematica and demonstrate their utility by applying them to the MHC homologous regions in human and fly.

1 Introduction

Identification of pairs of homologous chromosomal regions is an important step in solving a broad range of evolutionary and functional problems that arise in comparative mapping and genomics. Closely related homologous regions will be characterized by conserved gene order and content and may have substantial similarity in non-coding regions as well. However, in more distantly related regions, significant sequence similarity will typically only be observable in coding regions. In this case, genes are frequently treated as markers and putative homologous regions are identified by searching for *gene clusters*, regions that share similar gene content but where neither content nor order are preserved. Statistical tests to distinguish significant clusters from chance similarities in gene organization become essential as gene content and order diverge.

Conserved regions in whole genome comparisons are the basis of comparative map construction, studies of genome rearrangements [15, 40, 41] and gene order conservation [23, 45, 47], alternative approaches to phylogeny reconstruction [7, 10, 18, 38, 39, 47] and operon prediction in prokaryotes [9, 46]. Genome self-comparison is used to test hypotheses of whole genome duplication [42, 52].

Studies such as these consider gene clusters in a genomic context, focusing on large scale evolutionary processes and chromosomal organization.

In addition, many evolutionary and functional studies are based on studies of a single linkage groups [3, 11, 16, 22, 22, 24–27, 34, 37, 43, 44, 49]. Some studies examine the evolutionary history of a particular conserved region and the selective forces that hold it together. Others seek to exploit local similarities in gene organization for functional inference, gene annotation or to disambiguate orthology identification. For organisms whose lifestyle (e.g., lamprey) or longevity (e.g., fig) precludes construction of a genetic map, further research depends on identification of a homologous region in a species that is more suited to genetic manipulation or metabolic studies.

Analyses of such *individual clusters* often cannot take broader genomic context into consideration. For example, many researchers in fields such as ecology and organismal, behavioral and evolutionary biology work on species which have not been sequenced and are unlikely to be sequenced in the foreseeable future. In such cases, the amount of information about a region of interest, is limited by the laboratory's sequencing budget and available sequences in public data bases.

Our goal is to develop methods for estimating the statistical significance of individual gene clusters that can be carried out with knowledge of a local region plus aggregate properties such as an estimate of total gene number.

1.1 Related work

While statistical models for testing cluster significance are beginning to appear in the literature [8, 12–14, 49, 50], none are currently suitable for testing the significance of individual clusters. The lack of genomic context imposes a number of constraints on the statistical approach. Monte Carlo methods typically involve randomization of the entire genome, which is possible only with a complete genomic data set. When this information is not available, analytical tests that are parameterized by aggregate properties (e.g., the size of the genome, the size and number of gene families, etc.) are required.

The statistical tests for individual clusters must be based on an appropriate cluster definition. The intuitive notion that gene clusters share similar, but not conserved, gene content has been translated into a number of different formal models for finding and testing clusters [4–6,8,17,19,20,30,31,33,36,45,50,51], yet most of these are not suitable for individual clusters. Cluster statistics depend on the size of the search space. A number of statistical tests have been developed for a *reference region* model [13,21,49,50], in which an investigator is interested in a particular genomic region and searches the entire genome for additional regions containing the same genes. This is equivalent to considering $O(n)$ pairs of regions. However, when the investigator selects one or more pairs of homologous anchor genes and searches their genomic neighborhoods for additional homologs, the search space is $O(1)$. For such studies, the $O(n)$ reference region approach will underestimate the significance of the cluster.

Furthermore, the cluster definition must not require whole genome context to make sense. Many studies are based on the *max-gap* cluster, a maximal set

of homologous pairs where the distance between adjacent homologs on the same chromosome is no greater than a pre-specified parameter, g. The nature of the max-gap definition creates a "look-ahead" problem [6, 21], such that maximal max-gap clusters cannot be found using local, greedy heuristics. Although a local search may suggest that a particular region does not contain a max-gap cluster, a whole genome search is required to verify that a cluster meeting the max-gap criterion does not exist. Thus, while statistical tests for max-gap clusters based on the reference region model have been developed [21], these are not appropriate for individual clusters found by local search.

Finally, most current statistical tests do not take gene families into account, yet the significance of a cluster decreases as gene family size grows, because a given gene in one genome can be homologous to more than one gene in the other. As the number of possible matches increases, so do chance occurrences of gene clusters. As a result, tests that do not take gene family size into account risk overestimation of cluster significance.

1.2 Results

We propose a statistical test for individual clusters, under the null hypothesis of random gene order. These may be used without complete genomic context, are suitable for individual clusters found by local search, incorporate gene family size and are computationally tractable. In previous work [13], we proposed a test for individual clusters based on a window sampling model. Given two genomes with gene families, our measure of significance was the probability of observing a conserved set of linked genes in close proximity on both genomes. However, the treatment presented was mainly of theoretical interest since it did not lend itself to a computationally tractable implementation.

In the current paper, we recast this model in terms of generating functions, allowing us to obtain a general expression for our test statistic under the assumption of arbitrary gene family sizes. This statistic requires only the size of the conserved region, number of homologous genes in the linkage group and estimates of the distribution of gene family sizes and of the total number of genes in the genome. No information about the spatial organization of the genome outside the conserved region is needed. Under the additional assumption of fixed gene family sizes, we use the generating function model to obtain closed form expressions approximating cluster probabilities that can be calculated efficiently using Mathematica.

We describe our model and give a formal statement of the problem in Section 2. In Section 3, we derive the probability of observing an individual cluster under the null hypothesis of random gene order. In Section 4, we demonstrate how our model may be applied, using the heavily studied conserved homologous regions associated with Major Histocompatibility Complex (MHC) in human and fly.

2 Model

We develop tests for individual clusters based on the probability of observing a cluster in a genome with uniform random gene order (a "random genome"). A genome, $G_i = (1, \ldots, n_i)$ is modeled as an ordered set of n_i genes, ignoring chromosome break and physical distances between genes. We assume that genes do not overlap.

2.1 The 1-1 model

We begin with a simple model of two genomes, G_1 and G_2, with identical gene content and a one-to-one mapping between genes in G_1 and genes in G_2. That is, every gene in G_1 has exactly one homolog in G_2 and vice versa. We define an orthologous cluster as a pair of windows, W_1 and W_2, of length r_1 and r_2 selected from genomes G_1 and G_2, respectively, that share m homologous gene pairs. Figure 1 shows a cluster of three genes in a window of size five.

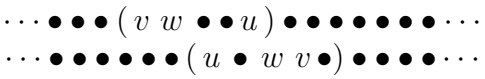

Fig. 1. A cluster with $r = 5$, $m = 3$ in the *1-1* model . Genes without homologs in this region shown as dots.

In this simple model, the probability that a pair of windows, of length r_1 and r_2, have exactly m genes in common is simply the probability that m of the r_1 genes in W_1 also appear in W_2 and can be calculated using a hypergeometric distribution:

$$p_{1-1}(m) = \frac{\binom{r_1}{m}\binom{n_2-r_1}{r_2-m}}{\binom{n_2}{r_2}}$$

The probability that the windows share *at least* m genes is then $\sum_{i=m}^{r} p_{1-1}(i)$. The 1-1 model requires a perfect, unambiguous homology mapping between G_1 and G_2. This may be possible after a recent speciation or polyploidization event. In general, however, because of variations in mutation rates, convergent evolution, non-homologous gene displacement and multi-domain proteins generated by exon shuffling, it is not possible to identify a unique match.

In this case, a many-to-many model is required. Genes are partitioned into families, such that any gene in a given family in G_1 can match any gene in the same family in G_2. The probability of finding a cluster by chance increases with family size. Consider, for example, the simple scenario where just one of the genes in W_1 matches f genes in G_2. The probability of finding m matches to the genes in a fixed size in G_2 increases since there are f possible matches for this one gene. However, it is surprisingly difficult to obtain a straightforward closed formula expressing this probability, even for this simple scenario. Therefore, accurate statistics require a model of gene family size. However, this raises the challenge that once gene families are incorporated in the model, it is no longer easy to determine the expected number of matches in a window of size r.

2.2 General gene family model

The problem of identifying true homologs has been much debated and numerous of solutions have been proposed (e.g., [2, 28, 33, 48, 50]). The first step is typically sequence comparison. A variety of approaches are applied to rule out false positives or negatives due to weak sequence similarity and/or matches based on homologous domains in otherwise unrelated sequences. These include bi-directional best hits, imposing a minimum alignment length requirement and phylogeny reconstruction. Despite these efforts, homology frequently remains unresolved. Furthermore, gene duplications that occur after the speciation separating G_1 and G_2, result in situations where a gene in one genome has two or more legitimate orthologs in the other.

We therefore extend our model to include gene families. A gene family is a set of *homologous genes*; that is, genes that share a common ancestor, through either duplication (paralogs) or speciation (orthologs). Gene family membership in our model does not depend on inherent functional or structural properties of the family but rather on what type of information the user brings to bear on identification of homology relationships. We define a gene family to be the set of *indistinguishable* homologous genes; i.e., homologous genes, where subfamily classification cannot be further disambiguated.

This is illustrated by the tyrosine kinases, a large multi-domain family of eukaryotic signaling proteins with 90 members in human [29]. While sequence similarity in the kinase domain shows that all tyrosine kinases are related, the domain composition of these sequences varies greatly, so that domain architecture can be used to disambiguate orthology. For instance, the Insulin Receptor (IR), in addition to the kinase domain, has two Furin-like domains, a Leucine-rich domain and two fibronectin domains, a domain architecture shared only with two other human genes: the Insulin Growth Factor 1 Receptor and the Insulin-Receptor Related Receptor. Thus, while an analysis based on sequence comparison alone, might map mouse IR to almost 100 kinases in human, an analysis based on domain architecture would associate mouse IR with only three human homologs.

We will assume that the set of genes in genomes G_1 and G_2 can be partitioned into non-intersecting gene families. Let $f_{ij} \subset G_i$ denote the members of the jth gene family in genome i. Then, the jth gene family, $f_j = f_{1j} \cup f_{2j}$, is a set of genes such that each gene in f_j is homologous to all other genes in f_j and only those genes. There are $\phi_{ij} = |f_{ij}|$ genes in the jth family in genome G_i. Let $\mathcal{F} = \{f_j\}$ be the set of all gene families in both genomes. In the gene family model, we define an orthologous gene cluster to be a pair of windows of length r_1 and r_2, drawn from G_1 and G_2, respectively, that have m *gene families* in common.

3 Cluster Statistics

We develop a test for individual clusters based on the probability of observing a cluster in two genomes with uniform random gene order (a "random genome").

In calculating cluster probabilities for the general case, we will need to count the number of ways that a window of a particular size can be filled with a given set of gene families in several contexts. We therefore derive a general solution to this problem using generating functions, a powerful combinatorial approach which can be used to determine a sequence of interest from the coefficients of a power series(see, for example, [35]). Here the sequence of interest is the number of ways filling the window. It is this formalism that allows us to compute cluster probabilities efficiently.

3.1 Window packings

Define $\mathcal{T}$ to be a set of λ gene families of arbitrary size $\phi_1 \dots \phi_\lambda$. Given the sample space of all sets of w genes sampled from a genome of size n, we wish to enumerate those that contain at least one gene from each family in $\mathcal{T}$. Since we do not take into account the order of genes in a window, this enumeration is equivalent to finding all window packings. The generating function formulation allows us to determine the number of such window packings, denoted by $\mathcal{N}(w, \lambda, \mathcal{T})$.

We represent contribution of the jth family in $\mathcal{T}$ by the generating function

$$\alpha_j(t) = \binom{\phi_j}{1} t + \binom{\phi_j}{2} t^2 + ... + \binom{\phi_j}{\phi_j} t^{\phi_j}. \tag{1}$$

The coefficient of t^i in $\alpha_j(t)$, denoted by $[t^i]\alpha_j(t)$, represents the number of ways of choosing i genes from j^{th} family. The contributions of all λ families to the window can then be derived from the product of their generating functions:

$$\alpha(t) = \prod_{j=1}^{\lambda} \Big[\binom{\phi_j}{1} t + \binom{\phi_j}{2} t^2 + ... + \binom{\phi_j}{\phi_j} t^{\phi_j} \Big]. \tag{2}$$

The coefficient $[t^w]\alpha(t)$ gives the number of ways of filling w slots with genes from the λ families, which is just $\mathcal{N}(w, \lambda, \mathcal{T})$. Note that the t^w term in $\alpha(t)$ will be a sum of products of the form $\beta_1 t^{x_1} \cdot \beta_2 t^{x_2} \cdots \beta_\lambda t^{x_\lambda} = (\prod_j \beta_j) t^w$, where the exponents of the dummy variable, t, sum to w. By inspecting Equation (2), we see that since β_j is the coefficient of t^{x_j}, it must be of the form $\beta_j = \binom{\phi_j}{x_j}$. The term $[t^w]\alpha(t)$ corresponds to packings containing x_1 genes from the first family, x_2 genes from the second family and so forth, where β_j corresponds to the number of ways of choosing x_j genes from the jth gene family. Summing over all packings, we obtain

$$\mathcal{N}(w, \lambda, \mathcal{T}) = \sum_{(x_1, \cdots x_\lambda)} \binom{\phi_1}{x_1} \binom{\phi_2}{x_2} \cdots \binom{\phi_\lambda}{x_\lambda}, \tag{3}$$

where the sum is over the set of all λ-tuples $(x_1, \dots, x_\lambda)$ such that

$$\sum_{j=1}^{\lambda} x_j = w, \tag{4}$$

and $1 \leq x_j \leq \phi_j$, $\forall j$.

Let us illustrate the window packing problem with a simple example. Suppose we wish to find the number of ways a window of size $w = 7$ can be packed with four gene families ($\lambda = 4$), such that the window has at least one gene from each gene family. Let the gene family sizes of $\mathcal{T}$ be $\phi_1 = 1$, $\phi_2 = 2$, $\phi_3 = 3$ and $\phi_4 = 4$ and the 4-tuple (x_1, x_2, x_3, x_4) refers to a window packing that has x_1 genes from the first gene family, x_2 genes from the second gene family, x_3 genes from third gene family and x_4 genes from the fourth gene family.

In order to find all possible packings, we need to find all 4-tuples satisfying Equation (4); in this example $\sum_{j=1}^{4} x_j = 7$. Since j^{th} gene family can contribute x_j genes in $\binom{\phi_j}{x_j}$ ways, the 4-tuple (x_1, x_2, x_3, x_4) can contribute $\binom{\phi_1}{x_1}\binom{\phi_2}{x_2}\binom{\phi_3}{x_3}\binom{\phi_4}{x_4}$ window packings. For example, the tuple $(1, 1, 1, 4)$ can contribute $\binom{1}{1}\binom{2}{1}\binom{3}{1}\binom{4}{4} = 6$ window packings. Table 1 lists the set of all possible 4-tuples and the number of packings associated with each 4-tuple. By adding the number of packings for each 4-tuple, we get the total number of ways the window can be filled with genes from the four gene families as given in Equation (3). Here, $\mathcal{N}(7, 4, \mathcal{T}) = 76$.

λ-tuple (x_1, x_2, x_3, x_4)	Number of packings
$(1, 1, 1, 4)$	6
$(1, 1, 2, 3)$	24
$(1, 1, 3, 2)$	12
$(1, 2, 1, 3)$	12
$(1, 2, 3, 1)$	4
$(1, 2, 2, 2)$	18
$\mathcal{N}(7, 4, \mathcal{T})$	76

Table 1. Number of ways packing a window of size $w = 7$ with four gene families of size {1,2,3,4}

3.2 Orthologous clusters with arbitrary gene families

We estimate the significance of a gene cluster using the probability that two windows, arbitrarily chosen from two random genomes, share at least m gene families. We enumerate over all sets of k gene families, for each value of k from m to r. For each such set, F, we determine the probability that W_1 contains only genes in families in F, including at least one from each family, followed by the conditional probability that at least l of the families in F also appear in W_2.

Expressed formally, the probability that W_1 and W_2 share *at least* m gene families is

$$q_o(m) = \sum_{k=m}^{r} \left[\sum_{F \in \mathcal{F}^k} p_1(F) \sum_{l=m}^{k} \sum_{\substack{E \in \mathcal{F}^l \\ E \subseteq F}} p_2(E) \right], \tag{5}$$

where $\mathcal{F}$ is the set of gene families in G_1 and G_2.

The probability that a given set, F, of k gene families is seen in W_1 is

$$p_1(F) = \binom{n_1}{r_1}^{-1} \mathcal{N}(r_1, k, F) \tag{6}$$

where $\mathcal{N}(r, k, F)$ is the number of window packings given by Equation (3). To determine, $p_2(E)$, we enumerate over all subsets of F of size l, where l ranges from m to k. For each subset, E, we seek the probability that each family in E is represented in W_2 at least once and that no other family in F appears in W_2. We exclude all other families in F to avoid overcounting.

At least l slots in W_2 must be filled with genes in E. The remaining $r_2 - l$ slots may be filled either from families in E or from families that do not appear in W_1; i.e., genes from $\mathcal{F} \setminus F$. Let z be the number of slots filled with genes from $\mathcal{F} \setminus F$. By considering all possible values of z, we obtain

$$p_2(E) = \binom{n_2}{r_2}^{-1} \sum_z \mathcal{N}(r_2 - z, l, E) \binom{n_2 - \Phi(F)}{z} \tag{7}$$

where $\Phi(F) = \sum_{j \in F} \phi_{2j}$. The parameter z ranges from $\max\{0, r_2 - \Phi(E)\}$ to $r_2 - l$ where $\Phi(E)$ is defined as above. The first term in the numerator is the number of ways of filling $r_2 - z$ slots with genes from the l families in E. The second term corresponds to all the ways of choosing the z outsiders from the set of genes not included in any gene family in W_1.

By substituting the expression in Equation (3) in Equations (6) and (7), we get a statistic for individual clusters in terms of n_1, r_1, n_2, r_2, m and the set of the gene families in G_1 and G_2. However, calculating this probability requires the enumeration of all subsets of k gene families. For each such subset, we must enumerate all packings satisfying Equation (4) and calculate a product of binomials for each packing. Computing this probability is prohibitively slow.

3.3 Orthologous clusters with fixed size families

The complexity of calculating $q(m)$ can be substantially reduced under the assumption that all gene families are of equal size, ϕ. When gene families are of equal size, it is not necessary to enumerate $\mathcal{F}^k$, since all subsets of k gene families are indistinguishable. We can simply replace the first term, $\sum_F p_1(F)$, in Equation (5) with the product of the number of sets of k gene families times $p_1(k)$, the probability that exactly k gene families of size ϕ are represented in the window:

$$\sum_{F \in \mathcal{F}^k} p_1(F) = \binom{|\mathcal{F}|}{k} p_1(k). \tag{8}$$

Invoking a similar transformation of the second term in Equation (5), the probability that W_1 and W_2 share at least m gene families simplifies to

$$q_o(m) = \sum_{k=m}^{r} \left[\binom{n_f}{k} p_1(k) \sum_{l=m}^{k} \binom{k}{l} p_2(l) \right]. \tag{9}$$

Under the fixed size assumption, $p_1(k)$ and $p_2(l)$ correspond to the probability that exactly k families appear in W_1 and exactly l families appear in W_2, respectively. To calculate $p_1(k)$ and $p_2(l)$, we require an expression for $\mathcal{N}'(w, \lambda, \phi)$, the number of window packings when all families are of fixed size. When $\phi_j = \phi$, β_j reduces to $\binom{\phi}{x_j}$ and Equation (2) becomes

$$\alpha'(t) = \Big[\binom{\phi}{1}t + \binom{\phi}{2}t^2 + ... + \binom{\phi}{\phi}t^{\phi}\Big]^{\lambda}. \tag{10}$$

The number of ways of observing λ gene families in a window of size w is given by $[t^w]\alpha'(t)$, yielding

$$\mathcal{N}'(\omega, \lambda, \phi) = \sum_{(x_1,\cdots x_\lambda)} \binom{\phi}{x_1}\binom{\phi}{x_2}\cdots\binom{\phi}{x_\lambda}, \tag{11}$$

where the sum is over the set of all λ-tuples $(x_1, \ldots, x_\lambda)$ satisfying Equation(4), under the constraint that $0 < x_i \leq \phi$, $\forall i$.

In this case, we can avoid enumerating the λ-tuples using the following simplification. Note that the right hand side of Equation (10) is a binomial series of the form $[(1+t)^{\phi} - 1]^{\lambda}$. By applying two binomial expansions, we obtain

$$\alpha'(t) = (-1)^{\lambda}\sum_{i=0}^{\lambda}\big[(-1)^i\binom{\lambda}{i}\big(\sum_{j=0}^{i*\phi}\binom{i*\phi}{j}t^j\big)\big]. \tag{12}$$

The number of ways of filling w slots with genes from the λ *fixed size* families is just $[t^w]\alpha'(t)$, yielding

$$\mathcal{N}'(w, \lambda, \phi) = (-1)^{\lambda}\sum_{i=0}^{\lambda}\Big[(-1)^i\binom{\lambda}{i}\binom{i*\phi}{w}\Big]. \tag{13}$$

Notice that at least $\lceil\frac{w}{\phi}\rceil$ gene families are required to fill a window of size w. Substituting the expression for $\mathcal{N}'(w, \lambda, \phi)$ in Equation (6) and restricting the lower bound on the dummy variable i to $\lceil\frac{r_1}{\phi}\rceil$, we obtain

$$p_1(k) = \binom{n_1}{r_1}^{-1}(-1)^k\sum_{i=\lceil\frac{r_1}{\phi}\rceil}^{k}\Big[(-1)^i\binom{k}{i}\binom{i*\phi}{r_1}\Big]. \tag{14}$$

Similarly, $p_2(l)$, the probability that W_2 contains exactly l gene families is

$$\binom{n_2}{r_2}^{-1}\sum_{z}(-1)^l\sum_{i=\lceil\frac{r_2-z}{\phi}\rceil}^{l}\Big[(-1)^i\binom{l}{i}\binom{i*\phi}{r_2-z}\Big]\binom{n_2-k\phi}{z} \tag{15}$$

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 enumerate window packings result in an efficient approximation to the probability that two windows, arbitrarily chosen from two random genomes, share at least m gene families. The general gene family model, Equations (6) and (7), requires implementation of an algorithm to enumerate all λ-tuples satisfying Equation (4). Furthermore, it is necessary to compute the product of λ binomial terms for each of the tuples in the enumeration. In contrast, Equations (14) and (15) require only a simple summation and can be easily computed in Mathematica. We can compute Equation (9) using the number of genes in each genome, the window sizes, gene family sizes and the number of gene families shared between the windows. Therefore, we only need information about the local regions and the aggregate properties of the genomes to determine significance of individual clusters.

4 Experimental results

In this section, we demonstrate how the results derived in Section 3 can be applied to test the validity of a pair of putative homologous chromosomal regions. As an example, we applied our models, implemented in Mathematica, to the MHC-like region, so called because it contains a conserved linkage group that resides near the human Major Histocompatibility Complex. This conserved homologous region, which has four copies in mammalian genomes, has been discussed in the molecular evolution literature extensively [1,12,16,22,24,25,43,49].

In recent literature, there have been many papers about conserved linkage groups observed in eukaryotes that appear to be duplicated and, in some cases, also conserved across several distantly related species (surveyed in [1, 12, 13]). These include the mammalian MHC region, the regions surrounding the Hox clusters [3], a region on chromosome 8 in human (FGR) [44] and a region containing a Tbox subfamily on chromosomes 5 and 11 in mouse [37]. These clusters typically contain five to fifteen genes spread over a window of 15 to 300 slots. Most of these studies do not present any statistical analysis testing the significance of the clusters. A few use simple statistical tests based on a reference region model with no correction for gene family size [1, 12, 49].

Several of these conserved regions have been the focus of particular interest because four paralogous copies have been observed in mammals, leading to the speculation that they could have arisen through the early vertebrate tetraploidization postulated by Ohno [32]. The MHC-like region contains a conserved linkage group of roughly a dozen genes (depending on which analysis you look at) on chromosome 6p21.3 in human. Paralogous subsets of these genes are also found on human chromosomes 1, 9 and 19. The four putative paralogous regions in human and mouse have been studied extensively [16, 22, 24, 25, 43] as new sequence and mapping data has provided additional insights into the evolutionary implications of the regions. The increasing availability of whole sequence data has also led to the investigation of regions in other species with orthologous gene content and organization that is suggestive of common ancestry for

the entire region. These include mouse, *C. elegans*, *D. melanogaster*, *S. Pombe* and several species of amphioxus [1, 12, 49].

We use a recent comparative analysis [12] of a chromosomal region on *Drosophila* chromosome X and the human MHC-like regions as an example for demonstrating the application of our statistical tests. Danchin *et. al.* [12] investigated a region delimited by *Drosophila* genes USP and Notch. USP is homologous to human RXRA, RXRB and RXRG. RXRA and Notch are "anchor" genes that bracket the MHCII region on human chromosome 6, a region also containing COL, ABC, RING and PSMB genes. Their analysis of *Drosophila* contigs in the public databases turned up 183 non-redundant transcripts in this region. Of these, 161 had significant matches in human, 32 of which included at least one significant hit in one or more of the MHC-like regions. Based on phylogenetic analysis, they [12] concluded that 19 of these were reliable orthologs. The two original anchor genes used to identify the region were eliminated from the study, since these were used to find the region do not constitute independent observations. Of the remaining 17 *Drosophila* genes in the study that had trusted human orthologs within one or more MHC-like regions, four fell into a region containing 44 genes on chromosome 6p21.3 in human.

We investigated the probability of observing such a cluster by chance using Equations (9), (14) and (15) and the following parameters $n_{hs} = 24194, n_{dm} = 13833, r_{hs} = 44, r_{dm} = 161$ and $m = 4$. Genome sizes were obtained from the *ensembl* database (*www.ensembl.org*). Note that our method to determine the significance of a gene cluster is not symmetric. When the sizes of the genomes and/or the windows are different ($n_1 \neq n_2$ and/or $r_1 \neq r_2$) the results will depend on which genome is designated G_1. Therefore, we estimated the cluster significance twice, using both the *Drosophila* and human regions as references.

The dependence of cluster statistics on gene family size is shown in Figure 2. These results show that cluster significance decreases rapidly with gene family size. Note that the probability of observing a gene cluster is slightly lower when *Drosophila* is used as reference. Since the second term in Equation (15) depends on n_2, the significance will decrease when G_2 is the larger genome. The probability of observing a homologous cluster structured like the human MHC-like cluster under the null hypothesis is greater than 0.1 for gene family sizes of four or greater and is close to one by the time ϕ reaches ten. While these numbers, taken alone, would suggest that the observed gene cluster is not statistically significant, a comprehensive analysis would require comparison of the chromosomal region in *Drosophila* with all four paralogous regions in human using a multiple testing approach. Our intent here is not to reanalyze the data or question the conclusions of the studies cited above, but rather to provide a concrete example of how our models can be put to practical use in real biological studies.

5 Conclusion

We have presented a new combinatorial approach to determine the significance of individual gene clusters. Our method takes gene family size into account and can be used to determine the significance of gene clusters in the absence of complete

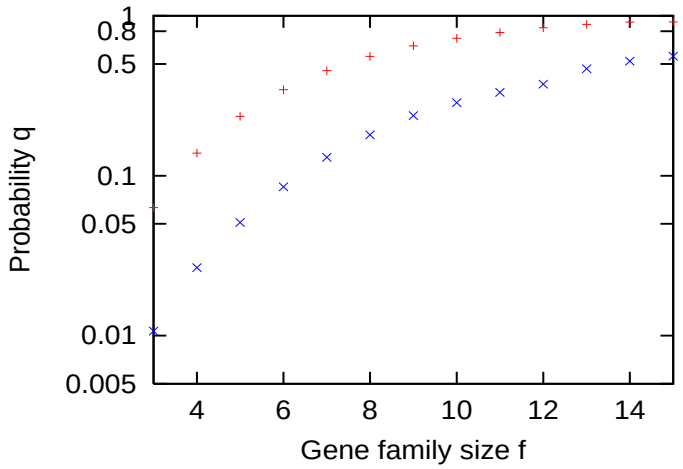

Fig. 2. Significance of MHC-like cluster using the fly genome as reference (×) and the human genome as reference (+)

genomic context. We estimate the significance of gene clusters by determining the probability that two regions, containing r_1 and r_2 genes respectively, share at least m gene families. By using generating functions, we have developed tractable expressions for the estimating the probability of observing orthologous gene clusters in two genomes. To demonstrate the utility of the method, we have applied it to estimate the significance of a well-studied conserved region in the fly and human genome.

Acknowledgment

We thank P. Chebolu, A. Frieze, A. Goldman, R. A. Hoberman, N. Song and B. Vernot for helpful discussions and T. Kalra for nutritional support. This research was supported by NIH grant 1 K22 HG 02451-01 and a David and Lucille Packard Foundation fellowship.

References

1. Laurent Abi-Rached, Andre Gilles, Takashi Shiina, Pierre Pontarotti, and Hidetoshi Inoko. Evidence of en bloc duplication in vertebrate genomes. *Nat Genet*, 31(1):100–5, May 2002.
2. M. D. Adams et al. The genome sequence of *Drosophila melanogaster*. *Science*, 287(5461):2185–9, 2000.
3. A. Amores, A. Force, Y. l. Yan, L. Joly, C. Amemiya, A. Fritz, R.K. Ho, J. Langeland, V. Prince, Y. L. Wang, M. Westerfield, M. Ekker, and J. H. Postlethwait. Zebrafish hox clusters and vertebrate genome evolution. *Science*, 282:1711–1714, 1998.
4. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408:796–815, 2000.

5. A. K. Bansal. An automated comparative analysis of 17 complete microbial genomes. *Bioinformatics*, 15:900–908, 1999.
6. A. Bergeron, S. Corteel, and M. Raffinot. The algorithmic of gene teams. In D. Gusfield and R. Guigo, editors, *Algorithms in Bioinformatics, Second International Workshop WABI2002*, Lecture Notes in Computer Science 2452, pages 464–476, 2002.
7. M. Blanchette, T. Kunisawa, and D. Sankoff. Gene order breakpoint evidence in animal mitochondrial phylogeny. *Journal of Molecular Evolution*, 49:193–203, 1999.
8. P. P. Calabrese, S. Chakravarty, and T. J. Vision. Fast identification and statistical evaluation of segmental homologies in comparative maps. *ISMB (Supplement of Bioinformatics)*, pages 74–80, 2003.
9. X Chen, Z Su, P Dam, B Palenik, Y Xu, and T Jiang. Operon prediction by comparative genomics: an application to the Synechococcus sp. WH8102 genome. *Nucleic Acids Res*, 32(7):2147–2157, 2004.
10. 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.
11. 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.
12. Etienne G J Danchin, Laurent Abi-Rached, Andre Gilles, and Pierre Pontarotti. Conservation of the MHC-like region throughout evolution. *Immunogenetics*, 55(3):141–8, Jun 2003.
13. D. Durand and D. Sankoff. Tests for gene clustering. *Journal of Computational Biology*, pages 453–482, 2003.
14. J. Ehrlich, D. Sankoff, and J.H. Nadeau. Synteny conservation and chromosome rearrangements during mammalian evolution. *Genetics*, 147(1):289–96, 1997.
15. Evan E Eichler and David Sankoff. Structural dynamics of eukaryotic chromosome evolution. *Science*, 301(5634):793–7, Aug 2003.
16. 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.
17. L. A. Goldberg, P. W. Goldberg, M. S. Paterson, P. Pevzner, S. C. Sahinalp, and E. Sweedyk. The complexity of gene placement. *Journal of Algorithms*, 41(2):225–2435, 2001.
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 *Proceedings of WABI01*, Lecture Notes in Computer Science 2149, pages 254–265, 2001.
20. S. Heber and J. Stoye. Finding all common intervals of k permutations. In *Proceedings of CPM01*, Lecture Notes in Computer Science 2089, pages 207–218, 2001.
21. R. Hoberman, D. Sankoff, and D. Durand. The statistical significance of max-gap clusters. "Proceedings of the RECOMB Satellite Workshop on Comparative Genomics", J. Lagergren, ed., Lecture Notes in Bioinformatics, Springer Verlag, in press., 2005.
22. 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.

23. Laurence D Hurst, Csaba Pal, and Martin J Lercher. The evolutionary dynamics of eukaryotic gene order. *Nat Rev Genet*, 5(4):299–310, Apr 2004.
24. 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.
25. N. Katsanis, J. Fitzgibbon, and E.M. Fisher. Paralogy mapping: identification of a region in the human MHC triplicated onto human chromosomes 1 and 9 allows the prediction and isolation of novel PBX and NOTCH loci. *Genomics*, 35(1):101–8, 1996.
26. L. Lipovich, E. D. Lynch, M. K. Lee, and M-C. King. A novel sodium bicarbonate cotransporter-like gene in an ancient duplicated region: *SLC4A9* at 5q31. *Genome Biology*, 2(4):0011.1–0011.13, 2001.
27. 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.
28. M.A.Huynen and P.Bork. Measuring genome evolution. *Proc. Natl. Acad. Sci. U.S.A.*, 95:5849, 1998.
29. G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam. The protein kinase complement of the human genome. *Science*, 298(5600):1912–1934, Dec 2002.
30. J.H. Nadeau and D. Sankoff. Counting on comparative maps. *Trends Genet*, 14(12):495–501, 1998.
31. S. J. O'Brien, J. Wienberg, and L. A. Lyons. Comparative genomics: lessons from cats. *Trends Genet*, 10(13):393–399, Oct 1997.
32. S. Ohno. *Evolution by genome duplication.* Berlin: Springer Verlag, 1970.
33. R. Overbeek, M. Fonstein, M. D'Souza, G. D. Pusch, and N. Maltsev. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A*, 96(6):2896–2901, Mar 1999.
34. 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.
35. G. Polya. *Notes on introductory combinatorics*. Birkhauser, 1983.
36. C. P. Ponting, J. Schultz, R. R. Copley, M. A. Andrade, and P. Bork. Evolution of domain families. *Adv Protein Chem*, 54:185–244, 2000.
37. 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.
38. 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.
39. 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.
40. David Sankoff. Rearrangements and chromosomal evolution. *Curr Opin Genet Dev*, 13(6):583–7, Dec 2003.
41. David Sankoff and Joseph H Nadeau. Chromosome rearrangements in evolution: From gene order to genome sequence and back. *Proc Natl Acad Sci U S A*, 100(20):11188–9, Sep 2003.
42. Cathal Seoighe. Turning the clock back on ancient genome duplication. *Curr Opin Genet Dev*, 13(6):636–43, Dec 2003.

43. 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.
44. J. Spring. Genome duplication strikes back. *Nature Genetics*, 31:128–129, 2002.
45. J. Tamames. Evolution of gene order conservation in prokaryotes. *Genome Biol*, 6(2):0020.1–11, 2001.
46. J. Tamames, G. Casari, C. Ouzounis, and A. Valencia. Conserved clusters of functionally related genes in two bacterial genomes. *JME*, 44::66–73, 1997.
47. J. Tamames, M. Gonzalez-Moreno, A. Valencia, and M. Vicente. Bringing gene order into bacterial shape. *Trends Genet*, 3(17):124–126, Mar 2001.
48. R. L. Tatusov, E. V. Koonin, and D. Lipman. A genomic perspective on protein families. *Science*, 278:631–637, 1997.
49. 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.
50. J. C. Venter et al. The sequence of the human genome. *Science*, 291(5507):1304–51, 2001.
51. K. H. Wolfe and D. C. Shields. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387:708–713, 1997.
52. K.H. Wolfe. Yesterday's polyploids and the mystery of diploidization. *Nature Rev. Genet.*, 2:33–41, 2001.