BIOINFORMATICS ORIGINAL PAPER
Vol. 23 no. 9 2007, pages 1053–1060
doi:10.1093/bioinformatics/btl673

Genome analysis

Identifying clusters of functionally related genes in genomes

Gangman Yi[1], Sing-Hoi Sze[1,2] and Michael R. Thon[1,3,*]

[1]Department of Computer Science, [2]Department of Biochemistry & Biophysics and [3]Department of Plant Pathology & Microbiology, Texas A&M University, College Station, TX 77845, USA

Received on August 23, 2006; revised on January 2, 2007; accepted on January 17, 2007

Advance Access publication January 19 , 2007

Associate Editor: Alex Bateman

ABSTRACT

Motivation: An increasing body of literature shows that genomes of eukaryotes can contain clusters of functionally related genes. Most approaches to identify gene clusters utilize microarray data or metabolic pathway databases to find groups of genes on chromosomes that are linked by common attributes. A generalized method that can find gene clusters regardless of the mechanism of origin would provide researchers with an unbiased method for finding clusters and studying the evolutionary forces that give rise to them.

Results: We present an algorithm to identify gene clusters in eukaryotic genomes that utilizes functional categories defined in graph-based vocabularies such as the Gene Ontology (GO). Clusters identified in this manner need only have a common function and are not constrained by gene expression or other properties. We tested the algorithm by analyzing genomes of a representative set of species. We identified species-specific variation in percentage of clustered genes as well as in properties of gene clusters including size distribution and functional annotation. These properties may be diagnostic of the evolutionary forces that lead to the formation of gene clusters.

Availability: A software implementation of the algorithm and example output files are available at http://fcg.tamu.edu/C_Hunter/.

Contact: mthon@tamu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

It is well known that genes in bacterial genomes are usually not distributed randomly in the genome but are organized into groups of transcriptionally linked genes called operons. Unlike their prokaryotic counterparts, genes in eukaryotic genomes are traditionally thought of as being randomly distributed among the chromosomes. However, an increasing number of functional and comparative genomic studies are revealing that, in fact, gene clusters may be common in eukaryotic species (Hurst *et al*., 2004; Lee and Sonnhammer, 2003). Furthermore, these studies suggest that multiple mechanisms may be responsible for forming gene clusters leading to levels of organization that range from small clusters comprised of only a few genes to large clusters spanning hundreds of genes.

Operon-like gene clusters are known to occur in *Caenorhabditis elegans* and share many similarities with their prokaryotic counterparts. Fungi also contain metabolic pathway clusters though their structure differs considerably from operons in *C.elegans* (Blumenthal, 1998; Spieth *et al*., 1993; Zorio *et al*., 1994). Some fungal metabolic pathway clusters have been shown to have coordinated gene transcription through the action of *cis*-acting regulatory elements (Herbert and Donald, 1975; Sophianopoulou *et al*., 1993). The yeast (*Saccharomyces cerevisiae*) genome contains a number of well-documented clusters, including the DAL and GAL clusters, which contain six and three genes, respectively (Cooper,1996; Hittinger *et al*., 2004). Filamentous fungi also contain a number of metabolic pathway clusters that consist of genes for biosynthesis of primary or secondary metabolites (Keller and Hohn, 1997). In all of these cases, the gene clusters are relatively small in size, often containing less than 15 genes arranged adjacent to one another on the chromosome.

One of the first genome-wide analyses of metabolic pathway clustering in eukaryotes revealed that gene clusters may span large segments of the genome (Lee and Sonnhammer, 2003). Their method examined genes linked to the same pathway described in the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000). The average distances of gene pairs within the pathway were compared to the distances calculated from randomized gene order. Two important conclusions could be drawn from this study. First, in every species examined, statistically significant clusters of metabolic pathway genes were found, suggesting that gene clusters are widespread in eukaryotes. Second, gene clusters were not necessarily comprised of sets of adjacent genes. Many clusters were sparse, i.e. they were comprised of genes belonging to the same metabolic pathway that were spread out over large segments of the genome but were nevertheless much closer to each other than expected by chance. In fact, a large number of gene expression studies are now showing that co-expressed genes have a tendency to be clustered and that the genes in these clusters tend to have related functions (for a review, see Hurst *et al*., 2004). It is important to note, however, that gene clusters are not always comprised of genes belonging to the same metabolic pathway, nor do they necessarily have coordinated gene expression. In this article, we define a gene cluster as a set

[*] To whom correspondence should be addressed. Present address: Department of Microbiology and Genetics. University of Salamanca, Salamanca, 37007, Spain

of genes with a common function that are closer to one another than is expected by chance.

The presence of gene clusters implies that clustering confers a selective advantage and that some evolutionary mechanism exists to promote the formation and maintenance of clusters. Genes in clusters may belong to common metabolic pathways, in which each gene encodes a protein (a gene product) that functions as an enzymatic step in a cellular metabolic process. Alternatively, gene products may form interaction networks in which proteins interact directly with each other to form multimeric proteins or serve as ligands and receptors in signaling cascades. Clusters of interacting proteins have been reported in *S.cerevisiae* (Teichmann and Veitia, 2004), and it has also been suggested that human protein ligands may be genetically linked to their receptors (Hurst *et al.*, 2004). In either case, there must be selective pressure to promote clustering. Such selective pressure may arise through coordinated gene expression, and it is believed that this is the most common force that drives clustering. Alternatively, coinheritance may provide the motive force for driving the clustering of genes. This theory states that natural selection will favor genetic linkage among genes that interact in some way, and they will tend to be inherited as a group (Fisher, 1930; Nei, 2003). It was recently demonstrated that among inbred mouse lines, extensive regions of linkage disequilibrium exist that are correlated with biological function (Petkov *et al.*, 2005). These observations are consistent with the concept of coinheritance and such a mechanism might also explain the clustering of metabolic pathway genes reported by Lee and Sonnhammer (2003).

Another mechanism by which gene clusters may form is through the tandem duplication of genes. Such homologous gene clusters are widespread in eukaryotes (Thomas, 2006). In *C.elegans*, Thomas (2006) showed that clusters of homologous genes tend to be formed of species-specific gene families that play roles in detoxification and immunity, and are found in chromosomal regions that undergo rapid evolution and reorganization. Further study of the content, function and distribution of homologous gene clusters will likely reveal important processes that regulate the formation of gene families.

Computational approaches to identify gene clusters are usually aimed at identifying specific cluster types, such as those that correspond to metabolic pathways or those that represent sets of co-expressed genes. A generalized approach that can identify all clusters in a genome would be of great value for the study of eukaryotic genome organization and evolution. In addition, identification of gene clusters may help to identify functional relationships among genes, and aid in the discovery of metabolic pathways and protein interactions.

In this article, we describe a method for finding clusters of genes that are annotated to common functional categories described in the Gene Ontology (GO) (Ashburner *et al.*, 2000). The Gene Ontology is a common controlled vocabulary of terms and phrases describing the function of genes and gene products. The terms and relationships among the terms are represented by a directed acyclic graph (DAG) in which vertices represent GO terms and edges represent relationships among similar terms. Genes can be annotated with GO terms creating gene associations that can be used for whole genome analyses.

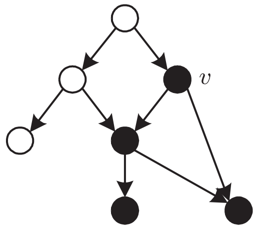

Fig. 1. Illustration of the set $R(v)$ of all reachable vertices from a given vertex v in a directed acyclic graph G. Filled circles denote vertices in $R(v)$, while hollow circles denote other vertices.

The Gene Ontology provides a rich framework for identifying gene clusters, regardless of the evolutionary mechanisms responsible for their formation. Our method can identify all possible clusters of genes annotated to the same GO term or a common parent term, and assigns p and e statistics that enable statistical evaluation of the clusters. We also describe an implementation of the algorithm and statistical test called C-Hunter. To demonstrate the utility of our method, we apply C-Hunter to the genomes of *Escherichia coli* and *S.cerevisiae*, and show that clusters identified with C-Hunter correspond to well-documented clusters in these species. We also perform a comparative analysis of gene clusters in several eukaryotic species and find species-specific variation in the number, size, function and putative evolutionary origin of the clusters.

2 METHODS

2.1 Preliminaries

A gene cluster is defined as a group of genes that are annotated with the same GO term or have the same parent term, and are also found within close proximity to each other on a chromosome. Cluster size refers to the number of genes in the cluster having the same GO term or parent term. Cluster length refers to the chromosomal length occupied by the cluster, including intervening genes that are not members of the cluster.

2.2 Algorithm and statistical evaluation of clusters

We represent each chromosome c by an ordered sequence of genes $(g_1, g_2, \ldots, g_n)$ while ignoring the orientation of each gene g_i on c. For genomes with multiple chromosomes, we concatenate all chromosomes together into a single sequence while disallowing clusters to span across chromosomes. To investigate functional assignments of these genes, we use the GO database (Ashburner *et al.*, 2000), in which three rooted DAGs are used to define hierarchical structures of increasingly specific functional categories, with top level categories being biological process, cellular component and molecular function. In each graph $G = (V, E)$, each vertex $v \in V$ represents a functional category (called a GO term) and each edge $(u, v) \in E$ represents that u is functionally less specific than v.

Since each gene g_i can have more than one functional assignment, let $F(g_i) \subseteq V$ be the set of all GO terms that are associated with g_i. Although these associations are typically on the bottom level of G, we are also interested in investigating the clustering of genes that belong to less specific functional categories. We consider each vertex $v \in V$ and let $R(v)$ be the set of all vertices that are reachable from v in G (Fig. 1), which gives all GO terms that are more specific than v in addition to v. We study the clustering of genes that belong to this category by finding

$c =$	g_1	g_2	g_3	g_4	g_5	g_6	$n = 6$	
$c(v) =$			g'_1		g'_2	g'_3	$n' = 3$	
Cluster 1			g_3		g_5		$k = 3$	$k' = 2$
Cluster 2					g_5	g_6	$k = 2$	$k' = 2$
Cluster 3			g_3		g_5	g_6	$k = 4$	$k' = 3$

Fig. 2. Illustration of all clusters of size greater than one that are associated with a vertex *v* in *G*.

```
Algorithm C-Hunter($G$,$c$,$F$) {
  for each vertex $v$ in $G$ do {
    $R(v) \leftarrow$ set of all vertices that are reachable from $v$ in $G$;
    $c(v) \leftarrow$ subsequence $(g'_1, g'_2, \ldots, g'_{n'})$ of $c =$
       $(g_1, g_2, \ldots, g_n)$ so that $R(v) \cap F(g'_j) \neq \emptyset$ for each $j$;
    for $k' \leftarrow 1$ to $n'$ do {
      for $j \leftarrow 1$ to $n'–k' + 1$ do {
        compute $e(n, n', k, k')$ of the cluster
          $(g'_j, g'_{j+1}, \ldots, g'_{j+k'-1})$ on $c(v)$ that spans the
          region $(g_i, g_{i+1}, \ldots, g_{i+k–1})$ on $c$, where
          $g_i = g'_j$ and $g_{i+k-1} = g'_{j+k'-1}$; } } } }
```

Fig. 3. Algorithm to find all functionally related gene clusters on a chromosome *c* which belong to each functional category that is represented by each vertex *v* in *G*. The function *F* defines the set of all vertices in *G* that are associated with each gene on *c*.

all genes on the given chromosome *c* that are associated with at least one GO term in $R(v)$. This defines a subsequence $c(v) = (g'_1, g'_2, \ldots, g'_{n'})$ of *c* so that $R(v) \cap F(g'_j) \neq \emptyset$ for each *j*. We think of each substring $(g'_j, g'_{j+1}, \ldots, g'_{j+k'-1})$ on $c(v)$ between the *j*th gene and the $(j + k' - 1)$th gene as a potential gene cluster that spans the region $(g_i, g_{i+1}, \ldots, g_{i+k-1})$ on *c* between the *i*th gene and the $(i + k - 1)$th gene, where $g_i = g'_j$ and $g_{i+k-1} = g'_{j+k'-1}$ (Fig. 2 and 3). The probability of finding such a cluster of size at least *k'* is given by the hypergeometric distribution as

$$p(n, n', k, k') = \sum_{i=k'}^{k} \frac{\binom{n'}{i}\binom{n-n'}{k-i}}{\binom{n}{k}},$$

where

$$\begin{cases} n = \text{Number of genes in a genome} \\ n' = \text{Number of genes associated with a common parent term} \\ k = \text{Cluster length} \\ k' = \text{Number of genes in a cluster.} \end{cases}$$

We evaluate its statistical significance by finding the expected number of such clusters that span a region of length *k* on *c*, which is given by

$$e(n, n', k, k') = (n - k + 1)p(n, n', k, k').$$

The details of the algorithm are given in Figure 3. To compute $c(v)$, first initialize its set of genes according to the function *F*. Then consider each vertex *u* in reversed topological order (which can be obtained by depth-first search in $O(|E|)$ time (Cormen *et al.*, 2001), and update $c(u)$ by considering each edge (u, v) in *G* and adding genes from $c(v)$ to obtain all the qualifying genes. Since there are at most *n* genes to add along each edge and at most *n* genes to store in each vertex, the above procedure takes $O(|E|n)$ time and $O(|V|n)$ space (there is no need to compute $R(v)$ explicitly). For a fixed vertex *v* and a fixed *k'*, since each cluster $(g'_j, g'_{j+1}, \ldots, g'_{j+k'-1})$ can be obtained from the previous one in constant time by removing g'_{j-1} and adding $g'_{j+k'-1}$ (except for the leftmost cluster), the time to obtain all the clusters is proportional to the total number of clusters. To compute the *e*-value of each cluster, for fixed *n'*, we preprocess and store all the $O(n)$ binomial coefficients. For fixed *n'* and *k'*, we use $O(n)$ space to store $p(n, n', k, k')$ for all *k* and obtain $p(n, n', k, k')$ from $p(n, n', k, k' - 1)$ in constant time. For each vertex *v*, it then takes $O(n^2)$ time to compute all the *e*-values. The overall time complexity for the entire algorithm is thus $O(|E|n + |V|n^2)$. Since it is only necessary to store clusters that have *e*-value below a cutoff, the space requirement is not prohibitively large.

GO terms near the root of the GO graphs are considered to be more generic while terms near the leaves are more specific, however, our method does not consider the depth (i.e. distance from the root or distance from a leaf) of the GO terms when computing the significance of the clusters. Our rationale for this approach was 2-fold. First, GO terms are created by annotators based on their knowledge of a particular function. Thus, differences in the relative depth of 'sibling' terms may reflect the current state of knowledge of a term and not the relative level of specificity. Second, we did not make an a priori assumption that the GO terms with the most relevance to gene clustering are farthest from the root.

2.3 Data sets

We selected species that represent a broad phylogenetic diversity, and also had significant percentages of genes annotated with GO terms. The genomes varied in the level of annotation, ranging from 25.1% in *D.rerio* to 96.2% in *S.cerevisiae* (Table 1). Proteins annotated with GO terms and files describing the order of genes within each chromosome were obtained from NCBI (http://www.ncbi.nlm.nih.gov/). We used gene2accession files and gene2go files, both from the NCBI ftp site to obtain the ordered gene sequence for a given chromosome and the GO term assignments for its genes, respectively.

2.4 Comparative analysis of gene clusters

Gene clusters that originated by gene duplication, selection for genetic linkage of interacting proteins, or selection for metabolic pathway clustering may be identified by comparing C-Hunter clusters to clusters found in public databases or identified by various other clustering algorithms. To identify clusters containing interacting proteins, we compared C-Hunter clusters to the networks in the database of interacting proteins (DIP) (Xenarios *et al.*, 2000), which represent protein interactions by undirected graphs in which nodes represent proteins and edges represent interactions between two proteins. For each C-Hunter cluster, we computed the mean minimum distance between all possible protein pairs within each cluster. Clusters with mean distance of less than two were considered putative interacting protein clusters.

In order to identify putative homologous gene clusters, we compared C-Hunter clusters to those formed by TribeMCL, a method for clustering proteins into groups related by sequence similarity (Enright *et al.*, 2002). We used the default TribeMCL options with a BLAST *e*-value cutoff of 1e−05. C-Hunter clusters corresponded to TribeMCL clusters if they exactly matched or were a subset of a TribeMCL cluster.

Lastly, we searched for correspondence between C-Hunter and KEGG (Kanehisa and Goto, 2000) to identify whether genes within a cluster belong to a common metabolic pathway. We assume that a C-Hunter cluster represents a metabolic pathway if all proteins in the cluster are annotated to the same KEGG pathway.

Table 1. Summary of gene clusters identified in eight species

Species	Number of genes in genome	Percent of genes annotated	Percent of annotated genes in cluster	Average number of genes per cluster	Average density	Number of clusters
Arabidopsis thaliana	26518	90.4	6.17	8.83	0.80	208
Caenorhabditis elegans	3227	36.8	27.42	32.36	0.78	11
Danio rerio	17636	25.1	3.36	5.07	0.89	30
Drosophila melanogaster	7580	63.8	24.60	12.35	0.76	122
Escherichia coli	4237	57.6	40.53	24.28	0.77	54
Homo sapiens	20282	65.6	17.55	14.58	0.76	185
Mus musculus	29493	50.9	22.02	93.22	0.56	40
Saccharomyces cerevisiae	6150	96.2	2.25	5.32	0.79	25

Among clusters with *e*-value ≤ 0.001 and group threshold of 50%. Density $= k'/k$.

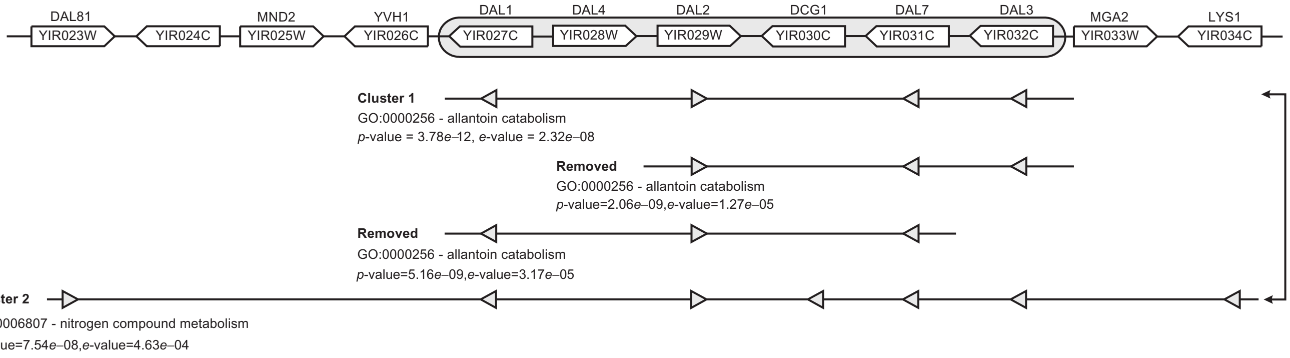
DAL81
YIR023W
YIR024C
MND2
YIR025W
YVH1
YIR026C
DAL1
YIR027C
DAL4
YIR028W
DAL2
YIR029W
DCG1
YIR030C
DAL7
YIR031C
DAL3
YIR032C
MGA2
YIR033W
LYS1
YIR034C
Cluster 1
GO:0000256 - allantoin catabolism
p-value = 3.78e–12, e-value = 2.32e–08
Removed
GO:0000256 - allantoin catabolism
p-value=2.06e–09,e-value=1.27e–05
Removed
GO:0000256 - allantoin catabolism
p-value=5.16e–09,e-value=3.17e–05
Cluster 2
GO:0006807 - nitrogen compound metabolism
p-value=7.54e–08,e-value=4.63e–04
Group 1

Fig. 4. Gene clusters identified in the region of the *S.cerevisiae* DAL cluster illustrating the filtering steps. 'Removed Clusters' were removed from the report during the filtering step 1 because they are subsets of Cluster 1 and have larger *e*-values than Cluster 1. Cluster 1 cannot be removed because its *e*-value is smaller than that of Cluster 2. Clusters 1 and 2 overlap and they were placed in a group during filtering step 2.

3 RESULTS AND DISCUSSION

3.1 Implementation

We developed a software package called C-Hunter that implements the above described algorithm and provides output of the clusters and statistical test in human readable format as well as comma-separated format suitable for import into other applications. Our algorithm finds all gene clusters that have an *e*-value below a user-specified cutoff and as such, numerous overlapping gene clusters are often reported. To improve readability of the output and facilitate comparative analyses of multiple genomes, we also apply several filtering steps. The standard filtering step consists of the removal of clusters that are subsets of a larger cluster that has a lower *e*-value (Fig. 4). We also implement a second optional filtering step that either masks or removes highly similar, overlapping clusters. In the second filtering step, the clusters are first sorted by *e*-value. Then, starting with the cluster with the lowest *e*-value, all other clusters that overlap by a user-specified threshold and are annotated with the same or a child GO term are labeled as members of a group of overlapping clusters. This process is repeated for each cluster that has not yet been labeled as a member of another group. A user-supplied parameter defines whether the labeled groups are reported in the output file or are ignored. For example, C-Hunter report files can be found in the Supplementary Material.

The running time for whole genome analyses depends on the number of genes in the genome and the number of annotated genes. For instance, the *S.cerevisiae* data set that we used for this analysis contains 6150 genes of which 96.2% are annotated with GO terms (Table 1). The running time for this data set including all filtering steps was 4 minutes on a system equipped with a 2.8 GHz Pentium IV processor and 2 GB of RAM.

We postulated that the primary limitation of our approach to finding gene clusters would be in the quality and quantity of the protein sequence annotation, and that there would be a tendency to find more gene clusters in species with more richly annotated genomes. The species we selected for our analyses vary widely in the percentage of genes with functional annotations, and the *D.rerio* and *S.cerevisiae* genomes contain the least and most annotated proteins, respectively (Table 1). Surprisingly, we found nearly identical percentage of genes in these species within clusters. Furthermore, we found no obvious tendency for level of annotation to be correlated with percentage of genes in clusters or number of clusters in the other species examined (Table 1).

3.2 Validation of known gene clusters

We evaluated the sensitivity of our approach by using C-Hunter to search for clusters in the *E.coli* genome and confirming the presence of documented operons in the C-Hunter output files. Most bacterial operons contain less than 10 genes and when the default C-Hunter parameters are used, large, sparse clusters predominate the search results. Clusters representing operons are either not present in the output because they have been removed by the first filtering step, or are hidden among a long list of larger clusters. By modifying the C-Hunter parameters to eliminate large clusters, smaller operon-sized clusters can be more easily identified. Therefore, we limited the search space to clusters containing 10 genes or less, and manually inspected the top 10 clusters in the output for known *E.coli* operons according to the Yale CGSC database (http://cgsc.biology.yale.edu/). Seven of the 10 clusters correspond to operons containing less than 10 genes while the remaining three correspond to operons with more than 10 genes (see Supplementary Fig. S1). Considering only the seven clusters that matched operons below our search threshold, we computed an error rate of 1.77 genes per operon, where an error is either a gene in a C-Hunter cluster that is not part of an operon or a gene in an operon that was not identified by C-Hunter.

The clusters identified with C-Hunter tended to include complete operons with additional flanking genes that have similar functions. In the case of the *his* operon, an additional flanking gene was identified as part of the cluster but was not reported by the CGSC database. The *his* operon entry in the database contains eight genes, while the C-Hunter cluster corresponding to this operon contains nine genes. Further inspection revealed that the additional gene, *hisL*, encodes the *his* operon leader peptide, which plays a regulatory role in the operon. We also found an overlapping cluster spanning a genomic interval (cluster length) of 281 genes, that contains 10 genes annotated to 'histidine biosynthesis' (*e*-value 3.27e−07). Since the search was limited to clusters containing 10 genes, we postulated that a larger 'histidine biosynthesis' cluster might be identified if the search was not restricted. By performing the search again with unrestricted cluster size, we identified a cluster spanning a genomic interval of 621 genes containing 12 genes annotated to 'histidine biosynthesis' (*e*-value 3.24e−07). This cluster may represent a level of organization in the *E.coli* genome that is on a much larger scale than that of operons.

We also validated the presence of well-documented gene clusters in *S.cerevisiae*. While the *S.cerevisiae* genome does not contain operons *per se*, it is known to contain clusters of genes belonging to metabolic pathways. Gene clusters in *S.cerevisiae* are not as well described as they are in bacteria, however, two well documented examples are known, namely, the DAL and GAL clusters. Therefore, we evaluated whether C-Hunter could identify these clusters. Using the default parameters (*e*-value cutoff 0.001 and no limits on cluster size), we identified the presence of both the *S.cerevisiae* DAL and GAL clusters as the first and sixth clusters in the result. Our algorithm identified four of the six genes that make up the allantoin cluster (Wong and Wolfe, 2005) within a genomic interval that contains six genes (Fig. 4). Two of the six genes were not identified as members of the cluster because their GO annotations did not share a common vertex in the GO graph with the other members of the cluster. The GAL cluster is comprised of three genes and was found in its entirety in our analysis (not shown).

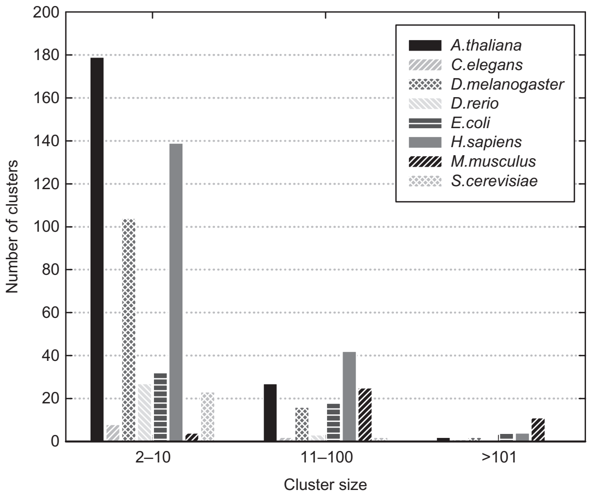

Fig. 5. Size distribution of clusters identified in each species.

3.3 Identification and comparative analysis of eukaryotic gene clusters

We used the C-Hunter application to find gene clusters in eight model organism genomes. For comparative analyses, we employed an *e*-value cutoff of 0.001 and applied the optional filtering step to remove clusters that overlap by 50% or more. We retained the cluster with the lowest *e*-value within each group for comparative analyses. Average cluster size varied considerably among species (Table 1) with *M.musculus* containing the largest clusters. The smallest clusters were found in *D.rerio*, averaging 5.07 genes per cluster. The gene clusters identified varied not only in size but in density as well, with *M.musculus* containing the most sparse clusters. Small clusters, such as bacterial operons and the *S.cerevisiae* DAL and GAL clusters, contained small numbers of genes with few intervening genes that were not part of the cluster. At the other end of the spectrum, many of the large clusters frequently found in vertebrate genomes were interspersed with genes that were not members of the cluster. For example, the top cluster in *H.sapiens* (GO:0006334 — nucleosome assembly) spans a genomic interval containing 84 genes, of which 26 are annotated to the function. The distribution of the number of genes per cluster in each species can be found in Supplementary Figure S2.

Table 2. Percentage of genes in each genome that were found in clusters assigned to each of the three categories within the Gene Ontology

Species	Biological process	Cellular component	Molecular function
A.thaliana	34.74	21.86	43.41
C.elegans	89.92	1.46	8.62
D.rerio	26.85	2.68	70.47
D.melanogaster	24.37	3.19	72.44
E.coli	71.36	25.38	3.26
H.sapiens	31.25	42.17	26.58
M.musculus	34.12	28.08	37.80
S.cerevisiae	26.49	8.65	64.86

Among clusters with *e*-value ≤0.001 and overlap threshold of 50%.

Table 3. Percentage of genes and gene clusters comprised of duplicated genes

Species	Percent of clusters*	Percent of genes
A.thaliana	79.51	60.49
C.elegans	18.18	65.24
D.rerio	60.00	81.55
D.melanogaster	51.64	44.48
E.coli	0.00	22.68
H.sapiens	50.81	58.77
M.musculus	22.50	55.58
S.cerevisiae	40.00	7.59

*Filtering was applied using a 50% overlap threshold.

The size distribution of clusters varied between species as well (Fig. 5). In all species examined, the majority of clusters were small in size, often with less than 10 genes, however, some contain large clusters with hundreds of genes. Large clusters comprised of more than 100 genes were found in most species, but were much less common than clusters with less than 10 genes. One exception is *M.musculus* which, unlike any of the other species examined, contains predominantly large gene clusters. The *M.musculus* genome has approximately the same proportion of annotated genes as *H.sapiens* (50.9% versus 65.6%) (Table 1), yet has six times more genes per cluster and four times less clusters. These differences are somewhat unexpected since mouse and human have strongly conserved gene order (Waterson *et al.*, 2002; Zhao *et al.*, 2004). While this result may indicate differences in the evolutionary processes that drive gene clustering in these species, it may also be due to variation in the methods that were used to annotate the genomes.

The Gene Ontology is divided into three separate graphs reflecting three general functional categories that describe gene function. To aid in identifying the functional constraints that may be important in forming gene clusters, we investigated whether there was a tendency in any of the species we examined for functional gene clusters to be annotated to terms within the three general categories. We considered number of clustered genes annotated to each ontology rather than number of clusters since the former can be compared directly to the annotations represented in the whole genome. All of the analyzed genomes contain genes annotated to GO terms from all three of the ontologies at roughly equivalent levels, however, in some species we found considerable bias in the representation of clustered genes among the three ontologies (Table 2). The most striking examples are in the *C.elegans* and *E.coli* genomes where 89.92 and 71.36% of the genes respectively were found in the biological process ontology. *C.elegans* is unique among eukaryotes in that, like bacteria, its genome contains operons and some analyses suggest that as many as 15% of the genes in this species are arranged in this manner (Spieth *et al.*, 1993). The biological process ontology contains terms describing metabolic processes, and it is likely that the relatively high proportion of genes annotated to this ontology reflects a trend towards clustering of metabolic pathways.

To gain insight into the evolutionary forces that may play roles in the formation of the gene clusters that were identified with C-Hunter, we assigned the clusters to categories, depending on evidence available to suggest relationships among the proteins. Homologous gene clusters were identified by determining whether genes corresponded to a cluster of highly similar proteins identified with TribeMCL. All species examined except for *E.coli* contained some percentage of homologous gene clusters (Table 3). There was no clear association to the overall percentage of duplicated genes in each genome, suggesting that the presence of homologous gene clusters is not merely a function of the rate of gene duplication. The human genome contained more than twice as many homologous gene clusters than mouse, which is consistent with the overall larger number of clusters found in human (Table 1) and suggesting that the increased number of clusters in human are predominantly clusters of homologous genes.

C-Hunter clusters representing groups of interacting proteins or metabolic pathways were identified by searching for corresponding clusters in DIP and KEGG, respectively. This analysis was only performed with *S.cerevisiae* since it was the only species that is relatively completely represented in the DIP and KEGG databases. We found three clusters that contained evidence of genes encoding interacting proteins (Table 4). Cluster 8 contains four histone proteins that make up the yeast nucleosome. Clusters 13 and 14 both encode genes with products that are involved with thiamine biosynthesis. Cluster 13 encodes SNZ3, SNO3 and THI5, while cluster 14 encodes SNZ2, SNO2 and THI12 and comprise two clusters of homologous sets of genes (Rodríguez-Navarro *et al.*, 2002).

We found six clusters that contained genes annotated to the same KEGG metabolic pathway (Table 4). Four were also identified as homologous gene clusters, so it is likely that the cluster members represent redundant components of the metabolic pathways. Two clusters, however, are not homologous gene clusters and correspond to known metabolic pathway clusters in yeast, including the biotin biosynthesis cluster (Wu *et al.*, 2005) (cluster 4) and the GAL cluster (cluster 6). Absent from this list is cluster 1, the DAL cluster, because

Table 4. C-Hunter clusters found in *S.cerevisiae*

Cluster ID	GO term	*e*-value	Chromosome number	Number of genes on chromosome	Cluster length (*k*)	Cluster size (*k*′)	Number of genes with GO term (*n*′)	TribeMCL	DIP	KEGG
1	0000256 - allantoin catabolism	2.32e−08	9	226	6	4	6			
2	0006814 - sodium ion transport	1.59e−07	4	781	3	3	3	•		•
3	0006530 - asparagine catabolism	3.68e−07	12	532	13	4	5	•		•
4	0009102 - biotin biosynthesis	6.35e−07	14	408	3	3	4			•
5	0015392 - cytosine-purine permease activity	5.55e−06	5	293	7	3	3	•		
6	0006012 - galactose metabolism	8.89e−06	2	419	3	3	8			•
7	0005488 - binding	2.26e−05	4	781	118	47	1059			
8	0000788 - nuclear nucleosome	3.61e−05	2	419	13	4	12		•	
9	0015891 - siderophore transport	5.33e−05	15	557	4	3	9			
10	0019541 - propionate metabolism	5.54e−05	16	481	7	3	5			
11	0005353 - fructose transporter activity	7.22e−05	4	781	3	3	15	•		
12	0005353 - fructose transporter activity	7.22e−05	8	292	3	3	15	•		
13	0009228 - thiamin biosynthesis	1.54e−04	6	136	3	3	19		•	
14	0009228 - thiamin biosynthesis	1.54e−04	14	408	3	3	19		•	
15	0006790 - sulfur metabolism	2.02e−04	12	532	5	4	57			
16	0016070 - RNA metabolism	2.26e−04	8	292	34	14	450			
17	0000943 - retrotransposon nucleocapsid	3.15e−04	7	561	4	4	94	•		
18	0000943 - retrotransposon nucleocapsid	3.15e−04	10	226	4	4	94	•		
19	0000943 - retrotransposon nucleocapsid	3.15e−04	16	481	4	4	94			
20	0019483 - beta-alanine biosynthesis	3.25e−04	13	482	2	2	2	•		•
21	0003850 - 2-deoxyglucose-6-phosphatase activity	3.25e−04	8	292	2	2	2	•		
22	0015291 - porter activity	3.72e−04	2	419	8	4	35			
23	0004099 - chitin deacetylase activity	9.75e−04	12	532	3	2	2	•		•
24	0008863 - formate dehydrogenase activity	9.76e−04	16	481	2	2	3			
25	0003941 - L-serine ammonia-lyase activity	9.76e−04	9	226	2	2	3			

The putative evolutionary forces that formed and/or maintain the clusters were inferred by searching for corresponding clusters in three different data sources. Homologous gene clusters were inferred from clusters formed by TribeMCL; interacting protein clusters by DIP and metabolic pathway clusters by KEGG. Total number of genes in *S.cerevisiae* genome is 6150.

only three of the four genes from this cluster were identified as components of the KEGG Purine metabolic pathway.

One interesting result is that the relative level of gene clustering and average cluster sizes that we observed among the species we examined was quite different from that reported by Lee and Sonnhammer (2003). These authors reported the presence of large metabolic pathway clusters in several species, including *S.cerevisiae*, whereas our analysis identified predominantly small clusters in this species. These differences can be attributed to differences in the nature of the functional annotation methods, the search algorithm or statistical tests that were employed. Another important distinction is that our approach utilizes all of the functional categories within the Gene Ontology, and as such, it may report clusters that are assigned to categories that do not immediately point to a reason why the genes are clustered. For example, the functional categories of gene clusters within the cellular component ontology do not reveal whether the proteins are members of a metabolic pathway or have coordinated regulation of expression, but such clusters may indicate that the proteins may have common functions that are not represented within the biological process or molecular function ontologies. Another example is *S.cerevisiae* cluster 8, ‘nuclear nucleosome’, which represents a group of histone proteins required for chromatin assembly. While all four proteins of the cluster are also annotated to ‘chromatin assembly or disassembly’, this GO term did not result in clusters with *e*-values below our cutoff. In this case, while ‘nuclear nucleosome’ refers to a cellular component, it also implies a molecular function that is more specific than ‘chromatin assembly’. It may be that the proteins in this cluster are either under-annotated or an appropriate term in the molecular function ontology does not exist. Still another example is cluster 7, ‘binding’, which is a large cluster containing genes encoding proteins that bind to other substrates, including DNA, RNA and proteins. The genes in this cluster are annotated to a variety of other functions, such as cytoskeleton organization, cell division and protein processing. The functional relationship among these proteins, beyond that of ‘binding’, is not immediately evident, nor is the mechanism that drives the clustering, however, all of the genes appear to have roles in core cellular functions. Clustering of essential genes has been demonstrated in *S.cerevisiae*, although the reasons why these genes are clustered have yet to be established (Pál, 2003). Both clusters 7 and 8 are good examples of how a broad-based search may uncover associations of clustered genes that may not be evident with searches restricted to a more specific set of terms.

4 CONCLUSIONS

We have developed an algorithm and application to identify clusters of functionally related genes in eukaryotic and prokaryotic genomes. Our approach finds all gene clusters in the data set and ranks them by their likelihood of occurrence by chance. *Post-hoc* filtering and sorting options create a report that is easy to read and enables researchers to evaluate the biological relevance of the results.

We identified a cluster corresponding to four of the six genes that make up the *S.cerevisiae* DAL cluster. The remaining two genes, while annotated with GO terms, did not share a common node in the GO graph with the other genes in the cluster. While a new node representing all members of the DAL cluster may eventually be added to the GO, its absence does not preclude the identification of the cluster and indicates that new gene clusters may be identified, despite the lack of a unifying term in the GO graph.

The clusters identified with C-Hunter may be annotated to functional categories that, like the DAL and GAL clusters, provide clues as to the mechanisms that may play roles in the formation of the clusters, but many functional categories do not easily suggest a reason for the clustering. By combining C-Hunter clustering results with information from other sources, such as metabolic pathway or interacting protein databases, we may begin to identify the evolutionary processes or mechanisms that lead to the formation of these clusters.

Our comparative analysis revealed species-specific differences in gene cluster content, size distribution and functional annotations. Variation in the level of completeness of the functional annotation could lead to differences in the number and size of gene clusters and should be taken into consideration when performing comparative studies. Despite this, some of the differences in cluster properties are likely to result from species-specific differences in the evolutionary processes that drive the functional clustering of genes.

ACKNOWLEDGEMENTS

This work was supported by funds from the Texas Agricultural Experiment Station and by the National Science Foundation under grants CCR-0311590 and DBI-0421815.

Conflict of Interest. none declared.

REFERENCES

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet*., **25**, 25–29.

Blumenthal,T. (1998) Gene clusters and polycistronic transcription in eukaryotes. *Bioessays*, **20**, 480–487.

Cooper,T.G. (1996) Regulation of allantoin catabolism in *Saccharomyces cerevisiae*. In Marzluf,G.A.(ed.), *The Mycota III: Biochemistry and Molecular Biology*, Springer, Berlin, pp. 139–169.

Cormen,T.H. *et al*. (2001) *Introduction to Algorithms*, 2nd edn. The MIT Press Cambridge, MA.

Enright,A.J. *et al*. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*., **30**, 1575–1584.

Fisher,R.A. (1930) *The Genetical Theory of Natural Selection*. Clarendon Press Oxford.

Herbert,N.A. and Donald,W.M. (1975) A gene cluster in *Aspergillus nidulans* with an internally located *cis*-acting regulatory region. *Nature*, **254**, 26–31.

Hittinger,C.T. *et al*. (2004) Parallel inactivation of multiple GAL pathway genes and ecological diversification in yeasts. *Proc. Natl Acad. Sci. USA*, **101**, 14144–14149.

Hurst,L.D. *et al*. (2004) The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet*., **5**, 299–310.

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*., **28**, 27–30.

Keller,N.P. and Hohn,T.M. (1997) Metabolic pathway gene clusters in filamentous fungi. *Fungal Genet. Biol*., **21**, 17–29.

Lee,J.M. and Sonnhammer,E.L.L. (2003) Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res*., **13**, 875–882.

Nei,M. (2003) Genome evolution – Let's stick together. *Heredity*, **90**, 411–412.

Pál,C. (2003) Evidence for co-evolution of gene order and recombination rate. *Nature*, **33**, 392–395.

Petkov,P.M. *et al*. (2005) Evidence of a large-scale functional organization of mammalian chromosomes. *PLoS Genet*., **1**, e33.

Rodríguez-Navarro,S. *et al*. (2002) Functional analysis of yeast gene families involved in metabolism of vitamins B1 and B6. *Yeast*, **19**, 1261–1276.

Sophianopoulou,V. *et al*. (1993) Operator derepressed mutations in the proline utilisation gene cluster of *Aspergillus nidulans*. *Mol. Genet. Genom*., **236**, 209–213.

Spieth,J. *et al*. (1993) Operons in *C. elegans*: polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions. *Cell*, **73**, 521–532.

Teichmann,S.A. and Veitia,R.A. (2004) Genes encoding subunits of stable complexes are clustered on the yeast chromosomes: an interpretation from a dosage balance perspective. *Genetics*, **167**, 2121–2125.

Thomas,J.H. (2006) Analysis of homologous gene clusters in *Caenorhabditis elegans* reveals striking regional cluster domains. *Genetics*, **172**, 127–143.

Waterston,R.H. *et al*. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.

Wong,S. and Wolfe,K.H. (2005) Birth of a metabolic gene cluster in yeast by adaptive gene relocation. *Nature*, **37**, 777–782.

Wu,H., Ito,K. and Shimoi,H. (2005) Identification and characterization of a novel biotin biosynthesis gene in *Saccharomyces cerevisiae*. *Appl. Environ. Microbiol*., **71**, 6845–6855.

Xenarios,I. *et al*. (2000) DIP: the database of interacting proteins. *Nucleic Acids Res*., **28**, 289–291.

Zhao,S. *et al*. (2004) Human, mouse, and rat genome large-scale rearrangements: stability versus speciation. *Genome Res*., **14**, 1851–1860.

Zorio,D.A. *et al*. (1994) Operons as a common form of chromosomal organization in *C. elegans*. *Nature*, **372**, 270–272.