AN ALGORITHM FOR IDENTIFYING CLUSTERS OF FUNCTIONALLY RELATED GENES IN GENOMES

A Thesis

by

GANG MAN YI

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

December 2006

Major Subject: Computer Science

AN ALGORITHM FOR IDENTIFYING CLUSTERS OF FUNCTIONALLY RELATED GENES IN GENOMES

A Thesis

by

GANG MAN YI

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Approved by:

Co-Chairs of Committee,	Michael Thon
	Sing-Hoi Sze
Committee Members,	Thomas Ioerger
	Christine Elsik
Head of Department,	Valerie E. Taylor

December 2006

Major Subject: Computer Science

ABSTRACT

An Algorithm for Identifying Clusters of Functionally Related Genes in Genomes.

(December 2006)

Gang Man Yi, B.S., Kangnung National University, South Korea

Co–Chairs of Advisory Committee: Dr. Michael Thon
Dr. Sing-Hoi Sze

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.

I present a basis of 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. I tested the algorithm by analyzing genomes of a representative set of species. I 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. The approach finds all gene clusters in the data set and ranks them by their likelihood of occurrence by chance. The method successfully identified clusters.

To my wife Jaehee and family for their love and encouragement.

ACKNOWLEDGMENTS

I would like to express my thanks and gratitude to Dr. Michael Thon for his research guidance, patience and understanding. I would also like to thank Dr. Sing-Hoi Sze for sharing his knowledge and experience, as well as for serving on my co-advisor. Additionally, I would like to thank Dr. Thomas Ioerger and Dr. Christine Elsik for their knowledge and advice as committee members

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

FIGURE Page

CHAPTER I

INTRODUCTION

A. Background

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 (Lee [1] and Hurst [2]). 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 [3], Zorio [4] and Spieth [5]). Some fungal metabolic pathway clusters have been shown to have coordinated gene transcription through the action of cis-acting regulatory elements (Herbert [6] and Sophianopoulou [7]). The yeast (*Saccharomyces cerevisiae*) genome contains a number of well documented clusters, including the DAL and GAL clusters, which contain 6 and 3 genes respectively (Hittinger [8] and Cooper [9]). Filamentous fungi also contain a number of metabolic pathway clusters that consist of genes for biosynthesis of primary or secondary metabolites

---

The journal model is *IEEE Transactions on Automatic Control.*

(Keller [10]). 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 [1]). Their method examined genes linked to the same pathway described in the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa [11]). The average distance of gene pairs within the pathway were compared to the distance 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 [2]). 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 thesis, I define a gene cluster as a set 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 gene 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 [12]) and it has also been suggested that human protein ligands may be genetically linked to their receptors (Hurst [2]). 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 [13] and Nei [14]). It was recently demonstrated that among inbred mouse lines, extensive regions of linkage disequilibrium exist that are correlated with biological function (Petkov [15]). 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 [1].

Another mechanism by which gene clusters may form is through the tandem duplication of genes. Such homologous gene clusters are widespread in eukaryotes (Thomas [16]). In *C. elegans*, Thomas [16] 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 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 aide in the discovery of metabolic pathways and protein interactions.

In comparison with Lee [1] paper, it only considers metabolic pathway, so the identified clusters are constrained within the metabolic pathway, and are not considered subsets of the metabolic pathway. The proposed method considers all functions of Gene Ontology, so the clusters are not constrained with specific properties. Clusters by the proposed method can obtain clusters with the broad functional meaning.

In this thesis, I describe a method for finding clusters of genes that are annotated to common functional categories described in the Gene Ontology (Ashburner [17]). The Gene Ontology (GO) 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. The Gene Ontology provides a rich framework for identifying gene clusters, regardless of the evolutionary mechanisms responsible for their formation. The proposed 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. I also describe an implementation of the algorithm and statistical test called C-Hunter. To demonstrate the utility of the proposed method, I apply C-Hunter to the genomes of *Escherichia coli* and *Saccharomyces cerevisiae*, and show that clusters identified with C-Hunter correspond to well-documented clusters in these species. I 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. A comparative analysis of clustering in several species revealed that species can be distinguished by a specific variation in percentage of clustered genes as well as in properties of gene clusters including size distribution and functional annotation.

CHAPTER II

METHODS

A. 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.

B. Algorithm and statistical evaluation of clusters

I present two algorithms 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. Clusters of genes annotated to a common GO term, or a common parent term, can be identified allowing for the identification of gene clusters regardless of the evolutionary mechanisms responsible for their formation. The algorithm is tested by analyzing genomes of a representative set of species.

I 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. To investigate functional assignments of these genes, I use the GO database (Ashburner [17]), in which three rooted directed acyclic graphs are used to define hierarchical structures of increasingly specific functional categories, with top level categories being biological process, cellular

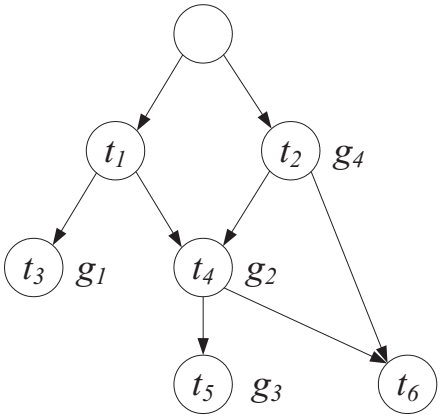


Fig. 1. Illustration of the set of all reachable vertices by a given gene list g_i in a directed acyclic graph G. t_j is GO term.

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 can have more than one functional assignment, the set of all GO terms that are associated with each gene is able to link to many genes (Figure 1). The difference between the two algorithms is the way in which putative genes clusters are selected. Algorithm 1 is focusing on genes on a chromosome and Algorithm 2 is focusing on GO terms. In order to select functionally related genes in Algorithm 1, fixed size clusters, which range in size from 2 to n where n is the number of genes in a chromosome, are selected sequentially. However, in Algorithm 2, functionally related genes, which span a region of a chromosome length, are selected by GO terms. GO terms are obtained by reversed topological search order using the depth-first search (DFS) in the DAG. The search takes in $O(|E|)$ time (Cormen [18])).

$c =$	g_1	g_2	g_3	g_4	g_5	g_6	$n = 6$
$c(w) =$	g_1	g_2	g_3				$k = 3$
$t(c(w)) =$	t_3	t_4	t_5				

cluster 1 [t_4]		g_2	g_3	$n' = 2$	$k' = 2$
cluster 2 [t_3,t_4]	g_1	g_2	g_3	$n' = 3$	$k' = 3$
cluster 3 [t_3,t_5]	g_1		g_3	$n' = 3$	$k' = 2$

Fig. 2. Illustration of all clusters of fixed size. The clusters are associated with genes linked to terms(defined in figure 1) by the combination of $t(c(w))$. c is a set of genes in a chromosome, $c(w)$ consists of genes of cluster size k=3. Cluster uses genes with the combination of terms associated in $c(w)$.

1. Algorithm 1

The first algorithm is finding gene clusters by a combination of functional categories annotating genes (see figures 1 and 2). Each chromosome c is represented by an ordered sequence of genes $(g_1, g_2 \ldots, g_n)$. For each candidate cluster $c(w)$, a set of terms is associated with genes in $c(w)$. The candidate cluster of size at least two and at most n moves from g_1 to $n - k + 1$, where n is a number of genes in a chromosome c, k is the number of genes associated with a common parent term, and k' is the number of genes in a cluster. The terms in $t(c(w))$ are determined by which a term is associated with at least one gene in $c(w)$. Terms to find gene clusters are selected by a combination of the term t_i in $t(c(w))$ (see figure 2). Only clusters, which have sub-node terms, are merged to clusters which have parent-node terms. For instance, in Table I, cluster $\{t_3, t_4, t_5\}$ and $\{t_4, t_5\}$ is merged to a cluster $\{t_3, t_4\}$, because t_4 is a parent term of t_5. Clusters $\{t_3\}$ and $\{t_5\}$ consist of only one gene, so they're

Table I. Illustration of a combination of terms. g_is are genes associated to terms t_j in clusters. The filled circle denotes 'removed cluster'

Cluster(terms)		Genes		Status
t_3	$\Rightarrow$	g_1	$\Rightarrow$	$\bullet$
t_4	$\Rightarrow$	g_2, g_3		
t_5	$\Rightarrow$	g_3	$\Rightarrow$	$\bullet$
t_3, t_4	$\Rightarrow$	g_1, g_2, g_3		
t_3, t_5	$\Rightarrow$	g_1, g_3		
t_4, t_5	$\Rightarrow$	g_2, g_3	$\Rightarrow$	$\bullet$
t_3, t_4, t_5	$\Rightarrow$	g_1, g_2, g_3	$\Rightarrow$	$\bullet$

removed, because only clusters consisting of at least two genes are considered.

Candidate clusters are created by the combination of all functional categories in $c(w)$ by selecting genes where more than two genes are in the cluster. The probability of each gene cluster with function t occurring by chance can be calculated by the hypergeometric distribution. It models the probability of observing at least k' from GO term within a cluster of size n'. This statistical test measures whether a cluster is enriched with genes from a particular term to a greater extent than would be expected by chance. p-value is given by

$$p(n, n', k, k') = 1 - \sum_{i=0}^{k'-1} \frac{\binom{k}{i}\binom{n-k}{n'-i}}{\binom{n}{n'}}$$

$$where \begin{cases} n &= \text{Total number of genes in a chromosome} \\ n' &= \text{Cluster length} \\ k &= \text{Number of genes associated with a common parent term} \\ k' &= \text{Number of genes in a cluster} \end{cases}$$

I check all genes n in a chromosome with various cluster size up to n, so it takes $O(n^2)$. In each cluster, $O(2^t)$ takes to calculate all possible terms. Thus, the overall time complexity takes $O(n^2 2^t)$.

The major problems in Algorithm 1 are "fixed cluster length" and "many computations". Gene clusters can be comprised of only a few genes to large clusters spanning hundred of genes without any considerations of a gene location within the cluster. But Algorithm 1 uses the fixed cluster length which is linearly increasing. Thus, only gene clusters which are comprised of sets of adjacent genes are found. The time complexity $O(n^2 2^t)$ in Algorithm 1 takes an exponential-time, thus, many computations are required. To make up for these problems in Algorithm 1, Algorithm 2 considers to trace each GO term in DAG and also considers genes associated with terms.

2. Algorithm 2

The basic condition of Algorithm 2 is same with Algorithm 1. I represent each chromosome by an ordered sequence of genes while ignoring the orientation of each gene on a chromosome. Functional annotations will be obtained from the gene ontology (GO) database. I combine three graphs (biological process, cellular component and molecular function) into a single directed acyclic graph $G = (V, E)$ with three connected components, in which each vertex $v \in V$ represents a functional category and

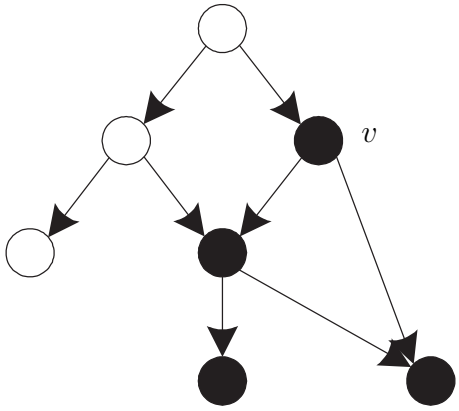


Fig. 3. Illustration of the set $R(v)$ of all reachable vertices from a given vertex v in a directed acyclic graph G. Filled circles denote vertices in $R(v)$, while hollow circles denote other vertices.

each edge $(u, v) \in E$ represents that v is functionally more specific than u.

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, I am also interested in investigating the clustering of genes that belong to less specific functional categories. I consider each vertex $v \in V$ and let $R(v)$ be the set of all vertices that are reachable from v in G (Figure 3), which gives all GO terms that are more specific than v in addition to v. I study the clustering of genes that belong to this category by finding 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. I think of each substring $(g'_j, g'_{j+1} \ldots, g'_{j+k'-1})$ on $c(v)$ between the jth 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 ith gene and the $(i + k - 1)$th gene, where $g_i = g'_j$ and $g_{i+k-1} = g'_{j+k'-1}$ (Figure 4). The probability of finding such a cluster of size at least k' is given by the hypergeometric

$$
\begin{array}{rcccccccc}
c = & g_1 & g_2 & g_3 & g_4 & g_5 & g_6 & n = 6 & \\
c(v) = & & & g'_1 & & g'_2 & g'_3 & n' = 3 & \\
 & & & & & & & & \\
\text{cluster 1} & & & g_3 & & g_5 & & k = 3 & k' = 2 \\
\text{cluster 2} & & & & & g_5 & g_6 & k = 2 & k' = 2 \\
\text{cluster 3} & & & g_3 & & g_5 & g_6 & k = 4 & k' = 3
\end{array}
$$

Fig. 4. Illustration of all clusters of size greater than one that are associated with a vertex v in G.

distribution as

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

$$
where \begin{cases}
n &= \text{Total number of genes in a chromosome} \\
n' &= \text{Number of genes associated with a common parent term} \\
k &= \text{Cluster length} \\
k' &= \text{Number of genes in a cluster}
\end{cases}
$$

I 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

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

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. 5. 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.

The details of the algorithm are given in Figure 5. 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 [18])), and update $c(u)$ by considering each edge (u, v) 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

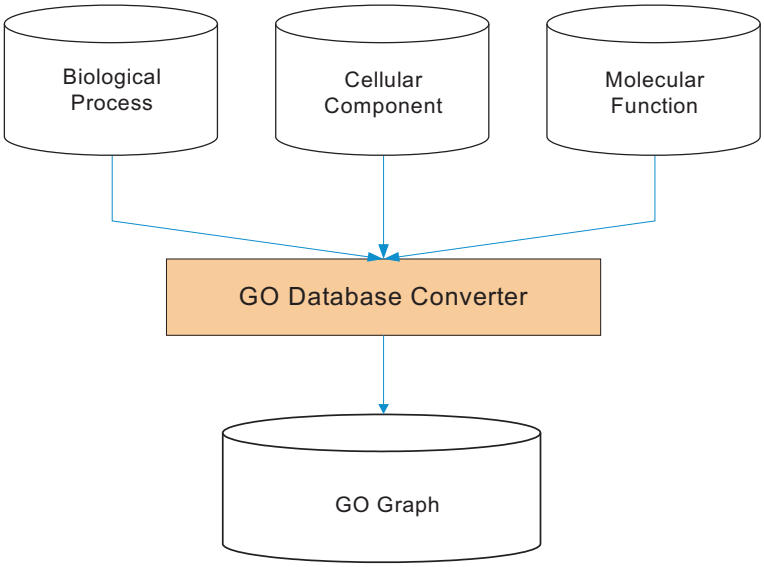


Fig. 6. The flowchart of converting GO Database to C-Hunter Database.

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 and n', I preprocess and store all the $O(n)$ binomial coefficients. For fixed n, n' and k', I 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.

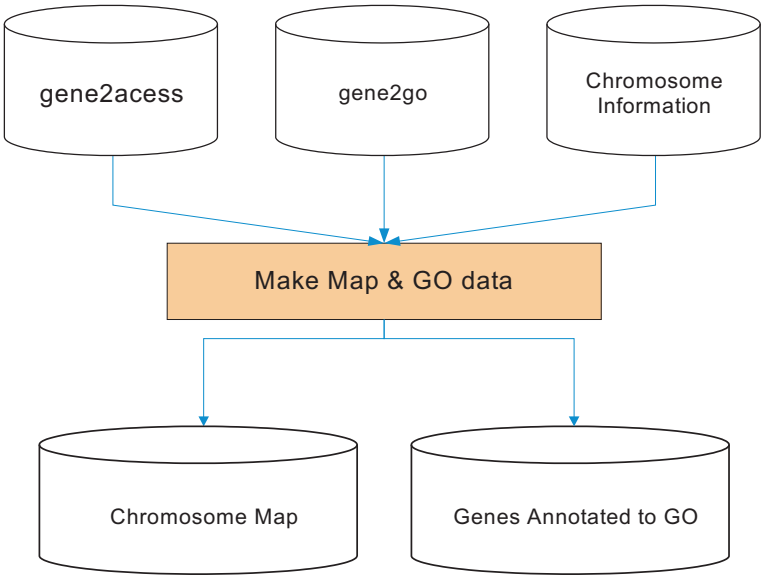


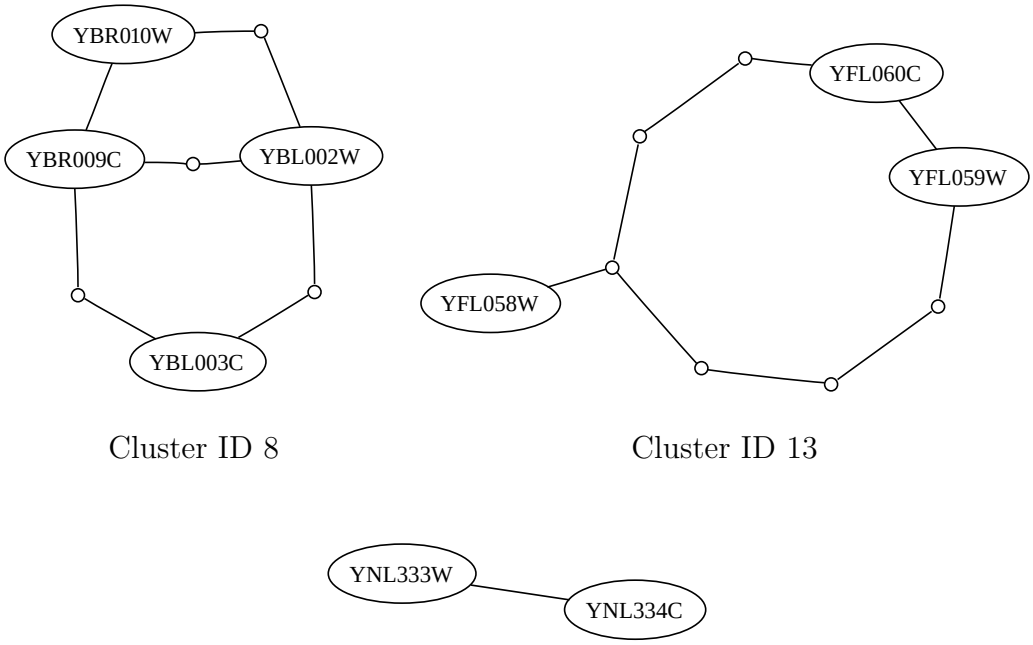
Fig. 7. The flowchart of making C-Hunter chromosome map file and GO data file.

C. Data set

I 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 II). 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/). I use 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 (Figures 6 and 7).

Table II. Summary of gene clusters and cluster groups identified in eight species (Among clusters with e-value ≤ 0.001 and group threshold of 50%).

Species	Number of Genes in Genome	Percent of Genes Annotated	Percent of Annotated Genes in Clusters	Avg. Number of Genes per Cluster	Number of Clusters
Arabidopsis thaliana	26518	90.4	6.17	8.83	208
Caenorhabditis elegans	3227	36.8	27.42	32.36	11
Danio rerio	17636	25.1	3.36	5.07	30
Drosophila melanogaster	7580	63.8	24.6	12.35	122
Escherichia coli	4237	57.6	40.53	24.28	54
Homo sapiens	20282	65.6	17.55	14.58	185
Mus musculus	29493	50.9	22.02	93.22	40
Saccharomyces cerevisiae	6150	96.2	2.25	5.32	25



Cluster ID 8

Cluster ID 13

Cluster ID 14

Fig. 8. Among top clusters in *S. cerevisiae*, gene clusters associated with DIP networks with mean distance of less than 2.

D. 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, I compared C-Hunter clusters to the Database of Interacting Proteins (DIP) (Xenarios [19]), which defines paths between proteins in terms of an undirected graph where a node represents a protein and an edge represents an interaction between two proteins. For each C-Hunter cluster, I computed the shortest path between each pair of proteins within the cluster. Using the minimum spanning tree (MST) algorithm

(Prim [20]), to find the shortest path between protein pairs, I computed the mean minimum distance between all possible protein pairs within each C-Hunter cluster. Clusters with mean distance of less than 2 were considered putative interacting protein clusters (Figure 8).

In order to identify putative homologous gene clusters, I compared C-Hunter clusters to those formed by TribeMCL, a method for clustering proteins into groups related by sequence similarity (Enright [21]). I 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 sub set of a TribeMCL cluster.

Lastly, I searched for correspondence between C-Hunter and KEGG (Kanehisa [11]) to identify whether genes within a cluster belong to a common metabolic pathway. I assume the C-Hunter cluster represents a metabolic pathway if all proteins in the cluster are annotated to the same KEGG pathway.

CHAPTER III

RESULTS AND DISCUSSION

A. Method use

As I discussed in the method section, the overall performance of Algorithm 2 is better than Algorithm 1 in terms of the algorithm efficiency and the speed. Algorithm 1 uses the fixed cluster length, but the actual cluster length can be smaller than the fixed cluster length if the genes are not annotated on the border. In addition, Algorithm 1 takes the exponential time due to the combination of computations. Therefore, I adopted Algorithm 2 for a series of experiments on actual data sets. Because Algorithm 2 is focusing on GO terms to select the putative genes, we will not miss clusters with the low *e*-value clusters due to the cluster length.The Algorithm 2 requires polynomial-time which is comparatively better than Algorithm 1. I developed a software package called C-Hunter that implements Algorithm 2 that can be used to find gene clusters in genomes that are annotated with GO terms.

B. Implementation

C-Hunter 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. The 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 (Figure 9). To improve readability of the output and facilitate comparative analyses of multiple genomes, I also apply several filtering steps (See figure 10). The standard filtering step consists of the removal of clusters that are exact subsets of a larger cluster that has a lower *e*-value. I also implement a second

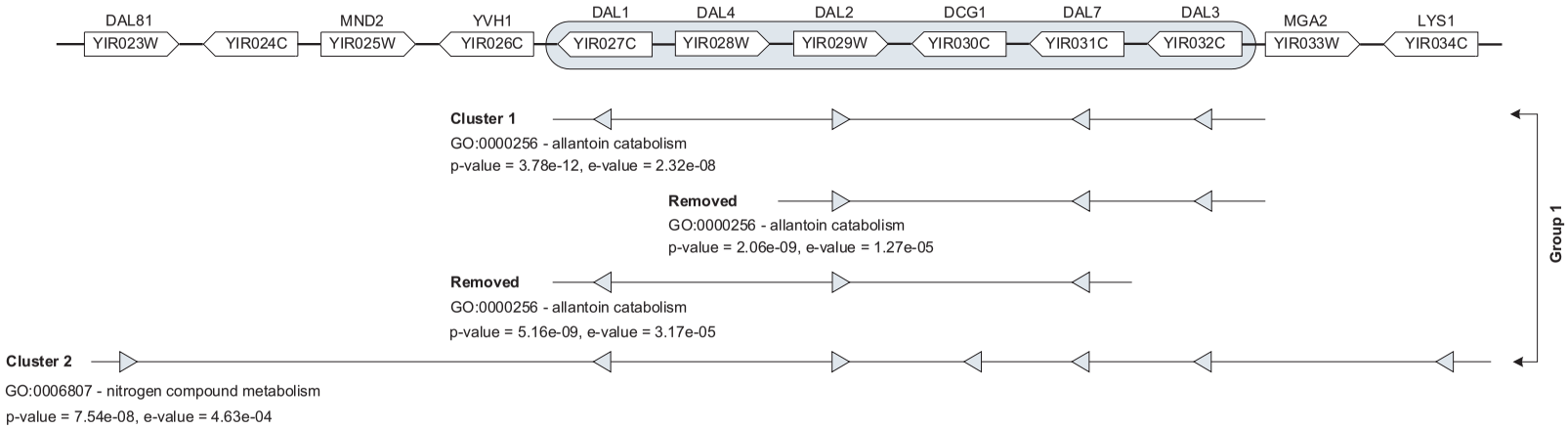


Fig. 9. Gene clusters identified in the region of the *S. cerevisiae* DAL cluster illustrating the filtering steps. "Removed Clusters" were removed from the report during the filtering step 1 because they are exact subsets and have larger *e*-values than Cluster 1. Cluster 1 cannot be removed because its *e*-value is smaller than that of Cluster 2. Clusters 1 and 2 overlap and during filtering step 2 they were placed in a group.

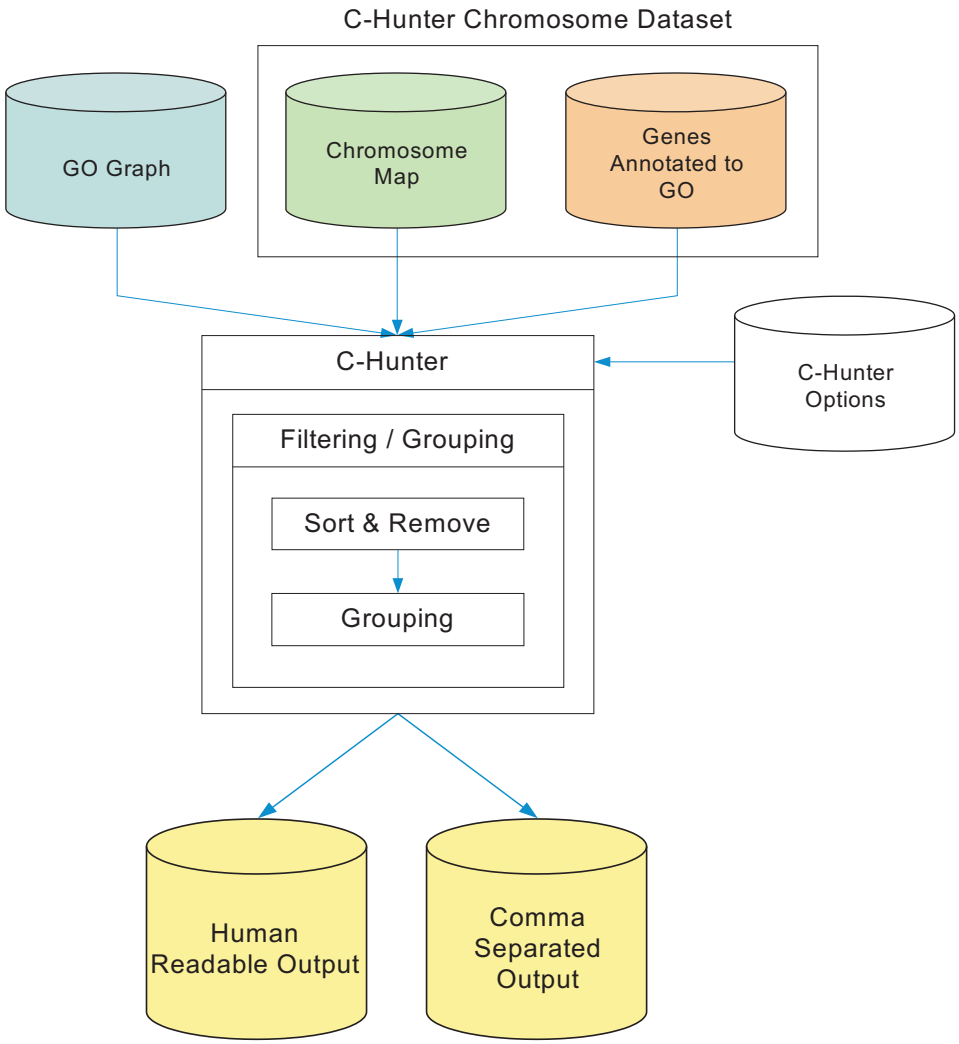


Fig. 10. C-Hunter Work Flow.

optional filtering step that either masks or removes highly similar, overlapping clusters. In the second filtering process, 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 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.

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 I used for this analysis contains 6150 genes of which 96.2% are annotated with GO terms (Table II). The running time for this data set was 4 minutes including all filtering steps when executed on a system equipped with 2.8 Ghz Pentium IV processor with 2 gigabytes of RAM.

I postulated that the primary limitation of the proposed 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 I selected for 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 II). Surprisingly, C-Hunter found nearly identical percentage of genes within clusters in these species. Furthermore, I 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 II).

C. Validation of known gene clusters

I evaluated the sensitivity of the proposed 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 are less than 10 genes in length 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 are more easily identified. Therefore, I 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/). Each of the top ten clusters correspond to known operons in the database and were identified as complete or nearly complete operons by C-Hunter. 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. I also found an overlapping cluster spanning a genomic interval (cluster size) of 281 genes, that contains 10 genes annotated to "histidine biosynthesis" (*e*-value 3.27e-7). Since the search was limited to clusters containing 10 genes, I 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, I identified a cluster spanning a genomic interval of 621 genes containing 12 genes annotated to "histidine biosynthesis" (*e*-value 3.24e-7). This cluster may represent a

level of organization in the *E. coli* genome that is on a much larger scale than that of operons.

I 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 GAL and DAL clusters. Therefore, I evaluated whether C-Hunter could identify these clusters. Using the default parameters (e-value cutoff 0.001, and no limits on cluster size) I identified the presence of both the *S. cerevisiae* DAL and GAL clusters as the first and sixth clusters in the result file. The proposed algorithm identified 4 of the 6 genes that make up the allantoin cluster (Wong [22]) within a genomic interval that contains 6 genes (Figure 9). 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 3 genes and was found in its entirety in analysis (not shown).

For a reference, C-Hunter identified 25 clusters in the *S. cerevisiae* output, but C-Hunter identified 18 clusters using GO terms without IEA (Inferred from Electronic Annotation) evidence code. The evidence code indicates how annotation to a particular term is supported, and is not necessarily a classification of an experiment. IEA is based on "hits" in sequence similarity searches, and curated by not a human but an electronic annotation (http://www.geneontology.org). In the *S. cerevisiae* output using GO terms with IEA, C-Hunter identified less cluster than clusters including IEA annotations, and I could not find GAL cluster which is one of well documented clusters in *S. cerevisiae* The average number of genes per cluster (5.72) is larger than the original output (5.32) (see Table II).

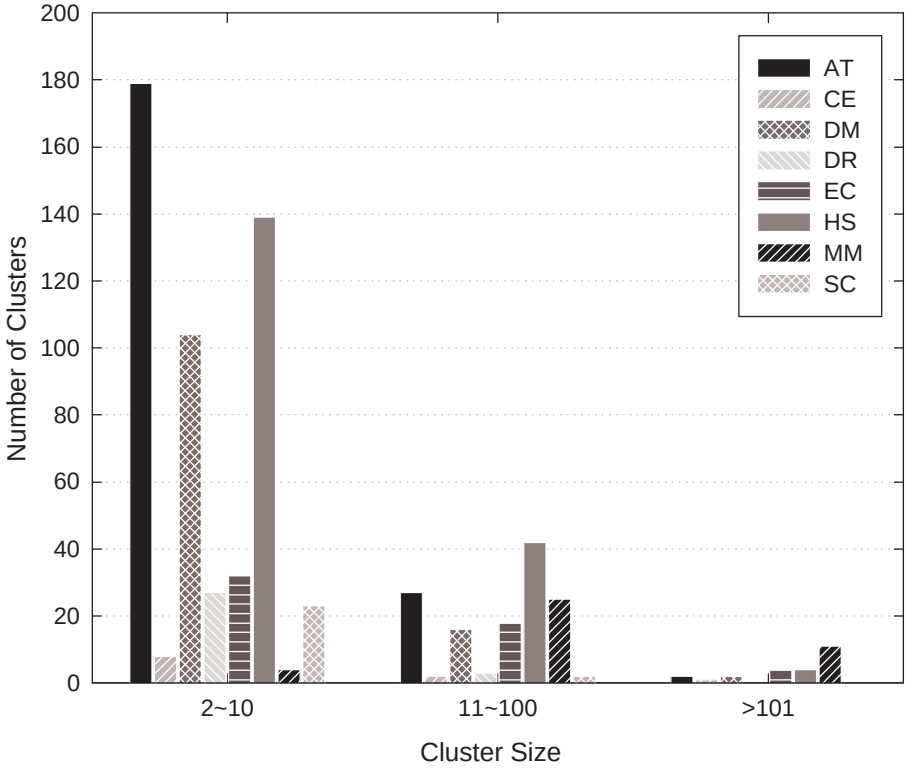


Fig. 11. Size distribution of clusters identified in each species.

D. Identification and comparative analysis of eukaryotic gene clusters

I used the C-Hunter application to find gene clusters in eight model organism genomes. For comparative analyses, I employed an *e*-value cutoff of 0.001 and applied the optional filtering step to remove clusters that overlap by 50% or more. I retained the cluster with the lowest *e*-value within each group for comparative analyses. Average cluster size varied considerably among species (Table II) with *M. musculus* containing the largest clusters. The smallest clusters were found in *S. cerevisiae*, with the top ten clusters averaging 5.3 genes per cluster. The gene clusters identified varied not

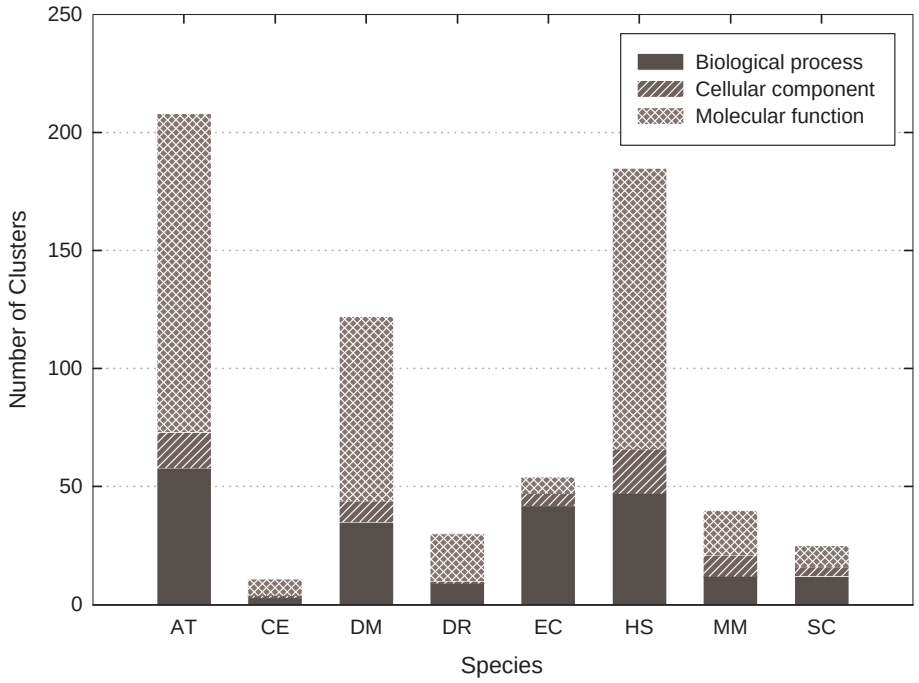


Fig. 12. Size distribution of clusters separated by three gene ontologies.

only in size but in density as well. Small clusters such as bacterial operons and the *S. cerevisiae* GAL and DAL 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* cluster 4 (GO:0006334 - nucleosome assembly) spans a genomic segment containing 84 genes, of which 26 are annotated to the function.

The size distribution of clusters varied between species as well (Figure 11). In all species examined, the majority of clusters were small in size, often smaller than 10 genes, however, some contain large clusters containing hundreds of genes. Large

Table III. Percentage of genes in each genome that were found in clusters assigned to each of the three sections within the Gene Ontology (Among clusters with e-value ≤ 0.001 and overlap threshold of 50%).

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

clusters comprised of more than 100 genes were found in most species, but were much less common than clusters with less than 10. 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% vs. 65.6%) (Table II) yet has 6.3X more genes per cluster and 4.6X less clusters. This result is somewhat unexpected since mouse and human have strongly conserved gene order (Zhao [23]).

The Gene Ontology is divided into three separate graphs reflecting three general functional categories that describe gene function. To aide in identifying the functional constraints that may be important in forming gene clusters, I investigated whether

there was a tendency in any of the species I examined for functional gene clusters to be annotated to terms within the three general categories. I 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 I found considerable bias in representation of clustered genes among the three ontologies (Figure 12 and Table III). The most striking examples are in the *C. elegans* and *E. coli* genomes where 89.92% and 71.36% respectively of the genes 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 at 15% of the genes in this species are arranged in this manner (Spieth [5]). 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.

The presence of gene clusters in eukaryotic genomes indicates that there is some selective pressure to form and maintain them. Several evolutionary processes have been proposed to provide the mechanism by which clusters of functionally related genes may arise. These are summarized below:

- Gene Duplication

  Duplicated genes are often found adjacent to each other, leading to clusters of genes with identical or similar functions. Gene duplication has been proposed as one of the major mechanisms by which new genes arise (Ohno [24]).

- Promoter-drives-expression (Hurst [2])

  Genes may share promoter elements that only function when they are in close

proximity to each other. This can enable co-expression of sets of genes that need to be regulated together. Examples include bidirectional promoters and polycistronic transcripts that have been described in some eukaryotes (Hurst [2]).

- Selection for Genetic Linkage.

  Two models have been proposed to provide selective pressure to genes with common functions to become genetically linked. Coordinated gene expression may be regulated at the chromatin level to regulate gene expression over chromosomal segments that span hundreds of kilobases (Hurst [2]). This can provide a selective pressure for genes to be located within close proximity to one another yet still separated by intervening genes. A second model suggests that proteins that interact with each other might be genetically linked. If certain combinations of alleles interact more strongly than others, selection might favor genetic linkage so that the interacting genes tend to be inherited as a group (Cooper [25] and Teichmann [12]).

The C-Hunter algorithm will enable us to identify all clusters of functionally related genes in genomes. By examining the clusters, we may be able to determine which of the above-described evolutionary process is responsible to forming the cluster. By comparing several species, we can gain insight into the relative contribution of each of these evolutionary processes into the formation of gene clusters. Clusters identified with C-Hunter can be compared to other data sets that can suggest whether genes are related by gene duplications, have coordinated transcription, or interact with each other. While such a study is beyond the scope of this thesis, I performed an analysis of the *S. cerevisiae* genome to demonstrate how such an analysis might be performed. I assigned C-Hunter clusters to categories, depending on evidence available to suggest

Table IV. Percentage of genes and gene clusters comprised of duplicated genes. *filtering was applied using a 50% overlap threshold

Species	Percent of Clusters*	Percent of Genes
A. thaliana	79.51	60.46
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

relationships among the proteins. Homologous gene clusters were identified by determining whether genes corresponded to a cluster of highly similar proteins identified with TribeMCL. TribeMCL is a method for clustering proteins into groups related by sequence similarities. All species examined, except for *E. coli* contained some percentage of homologous gene clusters (Table IV). 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, consistent with the overall larger number of clusters found in human (Table II) 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 path-

ways were identified by searching for corresponding clusters in DIP and KEGG, respectively. Database of Interacting Proteins (DIP) is useful for protein functions and protein-protein intreractions. 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. I found three clusters (containing 4, 3 and 2 genes) that contained evidence of genes encoding interacting proteins (Table V). 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 [26]).

I found six clusters that contained genes annotated to the same KEGG metabolic pathway. Four were also identified as homologous gene clusters, so it is likely that the cluster members represent redundant components of the metabolic pathways. Two, however, are not homologous gene clusters and correspond to known metabolic pathway clusters in yeast, the biotin biosynthesis cluster (Wu [27]) (cluster 4) and the GAL cluster (cluster 6). Absent from this list is cluster 1, the DAL cluster because only three of the four genes from this cluster were identified as components of the KEGG Purine metabolic pathway.

Table V. C-Hunter clusters found in *Saccharomyces cerevisiae*. The putative evolutionary forces that formed and / or maintains 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. B = Biological Process, C = Cellular Component, M = Molecular Function

Cluster ID	GO Term	*e*-value	Chromosome Number	Cluster Length	Cluster Size	TribeMCL	DIP	KEGG	Ontology
1	GO:0000256 - allantoin catabolism	2.32e-08	9	6	4				B
2	GO:0006814 - sodium ion transport	1.59e-07	4	3	3	•		•	B
3	GO:0006530 - asparagine catabolism	3.68e-07	12	13	4	•		•	B
4	GO:0009102 - biotin biosynthesis	6.35e-07	14	3	3			•	B
5	GO:0015392 - cytosine-purine permease activity	5.55e-06	5	7	3	•			M
6	GO:0006012 - galactose metabolism	8.89e-06	2	3	3			•	B
7	GO:0005488 - binding	2.26e-05	4	118	47				M
8	GO:0000788 - nuclear nucleosome	3.61e-05	2	13	4		•		C
9	GO:0015891 - siderophore transport	5.33e-05	15	4	3				B
10	GO:0019541 - propionate metabolism	5.54e-05	16	7	3				B
11	GO:0005353 - fructose transporter activity	7.22e-05	4	3	3	•			M
12	GO:0005353 - fructose transporter activity	7.22e-05	8	3	3	•			M
13	GO:0009228 - thiamin biosynthesis	1.54e-04	6	3	3		•		B
14	GO:0009228 - thiamin biosynthesis	1.54e-04	14	3	3		•		B
15	GO:0006790 - sulfur metabolism	2.02e-04	12	5	4				B
16	GO:0016070 - RNA metabolism	2.26e-04	8	34	14				B
17	GO:0000943 - retrotransposon nucleocapsid	3.15e-04	7	4	4	•			C
18	GO:0000943 - retrotransposon nucleocapsid	3.15e-04	10	4	4	•			C
19	GO:0000943 - retrotransposon nucleocapsid	3.15e-04	16	4	4				C
20	GO:0019483 - beta-alanine biosynthesis	3.25e-04	13	2	2	•		•	B
21	GO:0003850 - 2-deoxyglucose-6-phosphatase activity	3.25e-04	8	2	2	•			M
22	GO:0015291 - porter activity	3.72e-04	2	8	4				M
23	GO:0004099 - chitin deacetylase activity	9.75e-04	12	3	2	•		•	M
24	GO:0008863 - formate dehydrogenase activity	9.76e-04	16	2	2				M
25	GO:0003941 - L-serine ammonia-lyase activity	9.76e-04	9	2	2				M

CHAPTER IV

CONCLUSION

I have developed an algorithm and application to identify clusters of functionally related genes in eukaryotic and prokaryotic genomes. The proposed 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 (See figures 13 and 14). The proposed method successfully identified clusters in the *D.rerio* genome, which contains only 25.1% GO annotated genes indicating that gene clusters can be identified even in sparsely annotated genomes.

The 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.

I was also able to identify 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.

One interesting result is that the relative level of gene clustering and average

cluster sizes that I observed among the species I examined was quite different from that reported by Lee [1]. These authors reported the presence of large metabolic pathway clusters in several species, including *S. cerevisiae* whereas the analysis by C-Hunter identified predominantly small clusters in this species. These differences can be attributed to the use of different functional annotation methods as well as the nature of the statistical tests that were employed.

A comparative analysis of gene cluster content revealed that the mouse genome contains considerably fewer and larger gene clusters than human, despite the high level of conserved synteny between the two species (Zhao [23] and Waterson [28]). This may indicate that only subtle genome rearrangements are required for gene cluster formation. Alternatively, gene clusters may be concentrated in genomic segments that lack conserved synteny, and undergo frequent rearrangements, which would enable lineage specific changes in the forces that lead to gene clustering to shape clusters with such different properties. Integration of gene clusters identified with C-Hunter with other data types, such as synteny maps will shed new light on the evolutionary forces that lead to the formation and maintenance of functionally conserved gene clusters.

```
#================================================================================================#
        Date : Thu May  4 15:42:46 2006

        Total # of GO/FunCat nodes : 20436
        Total # of genes : 6150
        Total # of chromsomes : 16

        Minimum # of genes in a cluster : 2
        Maximum # of genes in a cluster ( 0 = No maximum size ) : 0
        Maximum cluster size ( 0 = No maximum size ) : 0

        E-value cutoff ( 0 = No consideration ) : 0.001
        Threshold of cluster overlap ( 0 = No consideration ) : 50%

        GO/FunCat scheme file : Scheme/scheme.GO.data.converted
        Map_file : Data_set/SaccharomycesCerevisiae/map_list
        FunCat data file : Data_set/SaccharomycesCerevisiae/data_list
        Output file : SC
#================================================================================================#

GROUP,NODE,#CHR,P-VALUE,E-VALUE,GENES,DEPTH,k,n',k',TERM

1,0000256,8,3.77651e-12,2.32067e-08,YIR027C YIR029W YIR031C YIR032C,7,6,6,4,allantoin catabolism
2,0006814,3,2.5807e-11,1.58662e-07,YDR038C YDR039C YDR040C,8,3,3,3,sodium ion transport
3,0006530,11,5.99655e-11,3.68068e-07,YLR155C YLR157C YLR158C YLR160C,9,13,5,4,asparagine catabolism
4,0009102,13,1.03228e-10,6.34646e-07,YNR056C YNR057C YNR058W,8,3,4,3,biotin biosynthesis
5,0015392,4,9.03246e-10,5.54954e-06,YER056C YER060W YER060W-A,9,7,3,3,cytosine-purine permease activity
6,0006012,1,1.44519e-09,8.88505e-06,YBR018C YBR019C YBR020W,9,3,8,3,galactose metabolism
7,0005488,3,3.74121e-09,2.25707e-05,YDR150W YDR151C YDR153C YDR159W YDR160W YDR164C YDR165W YDR166C
  YDR168W YDR170W-A YDR171W YDR172W YDR174W YDR176W YDR188W YDR189W YDR191W YDR194C YDR195W YDR207C
  YDR210W-B YDR210W-A YDR210C-D YDR210C-C YDR211W YDR212W YDR216W YDR217C YDR219C YDR224C YDR225W
  YDR227W YDR228C YDR229W YDR231C YDR235W YDR240C YDR243C YDR244W YDR252W YDR253C YDR254W YDR258C
  YDR261W-B YDR261W-A YDR261C-D YDR261C-C,2,118,1059,47,binding
8,0000788,1,5.88806e-09,3.61409e-05,YBL003C YBL002W YBR009C YBR010W,7,13,12,4,nuclear nucleosome
9,0015891,14,8.66481e-09,5.32626e-05,YOR381W YOR382W YOR383C,5,4,9,3,siderophore transport
10,0019541,15,9.02364e-09,5.54412e-05,YPR001W YPR002W YPR006C,7,7,5,3,propionate metabolism
11,0005353,3,1.17422e-08,7.2191e-05,YDR342C YDR343C YDR345C,7,3,15,3,fructose transporter activity
12,0005353,7,1.17422e-08,7.2191e-05,YHR092C YHR094C YHR096C,7,3,15,3,fructose transporter activity
13,0009228,5,2.5007e-08,0.000153743,YFL060C YFL059W YFL058W,8,3,19,3,thiamin biosynthesis
14,0009228,13,2.5007e-08,0.000153743,YNL334C YNL333W YNL332W,8,3,19,3,thiamin biosynthesis
15,0006790,11,3.29387e-08,0.000202441,YLL062C YLL060C YLL058W YLL057C,5,5,57,4,sulfur metabolism
16,0016070,7,3.69069e-08,0.000225759,YHR062C YHR065C YHR069C YHR070W YHR072W-A YHR077C YHR079C YHR081W
   YHR085W YHR086W YHR087W YHR088W YHR089C YHR091C,6,34,450,14,RNA metabolism
17,0000943,6,5.1211e-08,0.000314794,YGR161W-B YGR161W-A YGR161C-D YGR161C-C,5,4,94,4,retrotransposon
   nucleocapsid
18,0000943,9,5.1211e-08,0.000314794,YJR027W YJR026W YJR029W YJR028W,5,4,94,4,retrotransposon
   nucleocapsid
19,0000943,15,5.1211e-08,0.000314794,YPR158W-B YPR158W-A YPR158C-D YPR158C-C,5,4,94,4,retrotransposon
   nucleocapsid
20,0019483,12,5.28872e-08,0.000325203,YMR169C YMR170C,9,2,2,2,beta-alanine biosynthesis
21,0003850,7,5.28872e-08,0.000325203,YHR043C YHR044C,8,2,2,2,2-deoxyglucose-6-phosphatase activity
22,0015291,1,6.05632e-08,0.00037204,YBR291C YBR293W YBR296C YBR298C,5,8,35,4,porter activity
23,0004099,11,1.58662e-07,0.000975451,YLR307W YLR308W,6,3,2,2,chitin deacetylase activity
24,0008863,15,1.58662e-07,0.00097561,YPL276W YPL275W,6,2,3,2,formate dehydrogenase activity
25,0003941,8,1.58662e-07,0.00097561,YIL168W YIL167W,6,2,3,2,L-serine ammonia-lyase activity
```

Fig. 13. C-Hunter output of 25 top clusters in *S. cerevisiae*.

```
Group # : 1, #Clusters : 2, 0000256 : allantoin catabolism

      Cluster 1, 0000256 : allantoin catabolism
            E = 2.32067e-08, P = 3.77651e-12, Cluster size = 6
            Number of genes in this cluster = 4, chromosome # = 8
            Number of genes with this term (or a child term) = 6
            Genes:YIR027C
                        0000256 - allantoin catabolism
                        0004038 - allantoinase activity
                        0005622 - intracellular
                  YIR029W
                        0000256 - allantoin catabolism
                        0004037 - allantoicase activity
                        0008372 - cellular component unknown
                  YIR031C
                        0000256 - allantoin catabolism
                        0004474 - malate synthase activity
                        0008372 - cellular component unknown
                  YIR032C
                        0000256 - allantoin catabolism
                        0004848 - ureidoglycolate hydrolase activity
                        0016020 - membrane

      Cluster 2, 0006807 : nitrogen compound metabolism
            E = 0.000463027, P = 7.54239e-08, Cluster size = 12
            Number of genes in this cluster = 7, chromosome # = 8
            Number of genes with this term (or a child term) = 236
            Genes:YIR023W
                        0003704 - specific RNA polymerase II transcription factor activity
                        0005634 - nucleus
                        0006357 - regulation of transcription from RNA polymerase II promoter
                        0019740 - nitrogen utilization
                  YIR027C
                        0000256 - allantoin catabolism
                        0004038 - allantoinase activity
                        0005622 - intracellular
                  YIR029W
                        0000256 - allantoin catabolism
                        0004037 - allantoicase activity
                        0008372 - cellular component unknown
                  YIR030C
                        0005554 - molecular funtion unknown
                        0006807 - nitrogen compound metabolism
                        0008372 - cellular component unknown
                  YIR031C
                        0000256 - allantoin catabolism
                        0004474 - malate synthase activity
                        0008372 - cellular component unknown
                  YIR032C
                        0000256 - allantoin catabolism
                        0004848 - ureidoglycolate hydrolase activity
                        0016020 - membrane
                  YIR034C
                        0004754 - saccharopine dehydrogenase (NAD+\, L-lysine-forming) activity
                        0005737 - cytoplasm
                        0019878 - lysine biosynthesis via aminoadipic acid
```

Fig. 14. Human readable output of *S. cerevisiae*. Total number of genes = 6150, Number of chromosomes = 16. Minimum number of genes in a cluster = 2, e-value cutoff = 0.001, Threshold of cluster overlap = 50%.

CHAPTER V

FUTURE WORK

The current C-Hunter algorithm shows good results in several species. In order to increase the speed of C-Hunter, only new p-values are stored in the memory for each new computation. This method gives faster speed than many computations without storing data, but new species which can be more associated to GO terms than previous experiment data are able to consume many expensive hardware resources. In the next version, the novel memory management is needed.

C-Hunter finds many gene clusters qualitatively and quantitatively, but in order to improve the time complexity and enhance the quality of outputs in terms of finding gene clusters which are matched with real data sets (i.e. DAL or GAL clusters), increasing the sensitivity of the statistical test by applying a new statistical method would be a good idea. One approach can be to apply the bayesian statistical test. Using gene expression data or other data sources, the prior probability of candidate clusters can be obtained, and the posterior probability of set of genes annotating GO terms also can be measured so that the time complexity and the probability sensitivity can be improved, because the current measurement of the hypergeometric distribution takes many computations. Another way is using Hidden Markov Models (HMMs), which are statistical models that model a system being in a state at a particular time point and the transitions between states. It would be interesting to improve the current model to a new model using HMM in terms of finding gene clusters.

REFERENCES

[1] J. M. Lee and E. L. L. Sonnhammer, "Genomic gene clustering analysis of pathways in eukaryotes," *Genome Res.*, vol. 5, pp. 875–882, 2003.

[2] L. D. Hurst, C. Pál and M. J. Lercher, "The evolutionary dynamics of eukaryotic gene order," *Nat Rev Genet.*, vol. 5, pp. 299–310, 2004.

[3] T. Blumenthal, "Gene clusters and polycistronic transcription in eukaryotes,", *Bioessays*, vol. 20, pp. 480–487, 1998.

[4] D. A. Zorio, N. N. Cheng, T. Blumenthal and J. Spieth, "Operons as a common form of chromosomal organization in *C. elegans*," *Nature*, vol. 372, pp. 270–272, 1994.

[5] J. Spieth, G. Brooke, S. Kuersten, K. Lea and T. Blumenthal "Operons in *C. elegans*: Polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions," *Cell*, vol. 73, pp. 521–32, 1993.

[6] N. A. Herbert and W. M. Donald, "A gene cluster in *Aspergillus nidulans* with an internally located cis-acting regulatory region," *Nature*, vol. 254, pp. 26–31, 1975.

[7] V. Sophianopoulou, T. Suarez, G. Diallinas and C. Scazzocchio, "Operator derepressed mutations in the proline utilisation gene cluster of *Aspergillus nidulans*," *Molecular Genetics and Genomics*, vol. 236, pp. 209–213, 1993.

[8] C. T. Hittinger, A. Rokas and S. B. Carroll, "Parallel inactivation of multiple GAL pathway genes and ecological diversification in yeasts," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, pp. 14144–14149, 2004.

[9] T. G. Cooper, "Regulation of Allantoin Catabolism in *Saccharomyces cerevisiae*," in *The Mycota III: Biochemistry and Molecular Biology*, G. A. Marzluf, Ed. Berlin: Springer, 1996, pp. 139–169.

[10] N. P. Keller and T. M. Hohn, "Metabolic Pathway Gene Clusters in Filamentous Fungi," *Fungal Genet Biol.*, vol. 21, pp. 17–29, 1997.

[11] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Res.*, vol. 28, pp. 27–30, 2000.

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

[13] R. A. Fisher, *The Genetical Theory of Natural Selection,* Oxford: Clarendon Press, 1930.

[14] M. Nei, "Genome evolution - Let's stick together," *Heredity*, vol. 90, pp. 411–412, 2003.

[15] P. M. Petkov, J. H. Graber, G. A. Churchill, K. DiPetrillo, B. L. King and K. Paigen, "Evidence of a large-scale functional organization of mammalian chromosomes," *PLoS Genet.*, vol. 1, pp. 312–322, 2005.

[16] J. H. Thomas, "Analysis of homologous gene clusters in *Caenorhabditis elegans* reveals striking regional cluster domains," *Genetics*, vol. 172, pp. 127–143, 2006.

[17] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G.

M. Rubin and G. Sherlock, "Gene ontology: tool for the unification of biology," *Nat Genet.*, vol. 25, pp. 25-29, 2000.

[18] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms,* 2nd. ed. Cambridge: The MIT Press, 2001.

[19] I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte and D. Eisenberg "DIP: The Database of Interacting Proteins," *NAR*, vol. 28, pp. 289–91, 2000.

[20] R. C. Prim, "Shortest connection networks and some generalizations," *Bell System Technical Journal*, vol. 36, pp. 1389–1401, 1957.

[21] A. J. Enright, S. Van Dongen and C. A. Ouzounis, "An efficient algorithm for large-scale detection of protein families," *Nucleic Acids Res.*, vol. 30, pp. 1575–1584, 2002.

[22] S. Wong and K. H. Wolfe, "Birth of a metabolic gene cluster in yeast by adaptive gene relocation," *Nature*, vol. 7, pp. 777–782, 2005.

[23] S. Zhao, J. Shetty, L. Hou, A. Delcher, B. Zhu, K. Osoegawa, P. D. Jong, W. C. Nierman, R. L. Strausberg and C. M. Fraser, "Human, mouse, and rat genome large-scale rearrangements: stability versus speciation," *Genome Res.*, vol. 14, pp. 1851–1860, 2004.

[24] S. Ohno, *Evolution by Gene Duplication,* Berlin: Springer Verlag, 1970.

[25] D. N. Cooper, *Human Gene Evolution,* Oxford: BIOS Scientific, 1999.

[26] S. Rodríguez-Navarro, M. T. Rodriguez-Manzaneque, A. Ramne, G. Uber, D. Marchesan, B. Dujon, E. Herrero, P. Sunnerhagen and J. E. Perez-Ortin, "Func-

tional analysis of yeast gene families involved in metabolism of vitamins B1 and B6," *Yeast*, vol. 19, pp. 1261–1276, 2002.

[27] H. Wu, K. Ito and H. Shimoi, "Identification and characterization of a novel biotin biosynthesis gene in *Saccharomyces cerevisiae*," *Appl. and Environ. Microbiol.*, vol. 71, pp. 6845–6855, 2005.

[28] R. H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril *et al.*, "Initial sequencing and comparative analysis of the mouse genome," *Nature*, vol. 420, pp. 520–562, 2002.

VITA

Gang Man Yi

Education

- Master of Science, Computer Science, Texas A&M University, 2006
- Bachelor of Science, Computer Science, Kangnung National University, South Korea, 2002

Contact Address

- *Permanent mailing address* : 1/3, 101-6, Mansu-2-dong, Namdong-gu, Incheon, 405-242, South Korea
- *E-mail* : gangmanyi@tamu.edu, gangmanyi@hotmail.com, lekman@nate.com