1-1-2007

# Empirical study of Gene Ontology based Microarray clustering.

Tianbing Lin
*University of Windsor*

# Empirical Study of Gene Ontology based Microarray Clustering

by

Tianbing Lin

A Thesis
Submitted to the Faculty of Graduate Studies and Research
through Computer Science
in Partial Fulfillment of the Requirements for
the Degree of Master of Science at the
University of Windsor

Windsor, Ontario, Canada

2007

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

*Your file Votre référence*
*ISBN: 978-0-494-42325-7*
*Our file Notre référence*
*ISBN: 978-0-494-42325-7*

# ABSTRACT

This thesis project studies current similarity measures over Gene Ontology and introduces a new measure combined with Euclidean distance to perform Microarray analysis. New combined measures contain both the expression data and (known) biological information from Gene Ontology to express the biological relation between gene products. In order to adapt the similarity measure to the Gene Ontology, an On-The-Fly probability is initially defined to calculate the probability of a term in the current problem space. A similarity measure between a term and a set of terms is defined, as well as a similarity measure between sets. The performance of applying these similarity measures is compared by clustering a dataset of which the correct clustering scheme is known. The results of the comparison are analyzed and some conclusions are drawn about the similarity measure.

# ACKNOWLEDGEMENTS

First, I would like to take this opportunity to express my gratitude to my advisor, Dr. Alioune Ngom, for giving me this learning opportunity, his continuous support and generous encouragement all the way through my graduate study in University of Windsor. Also I want to thank my thesis committee, Dr. J. Lu, G. Zhang, and Dr. A.K. Aggarwal, for helping me finish my thesis with all your advices. I also acknowledge the help for my learning in University of Windsor from Dr. Kent, Dr. Tsin, Dr. Kobti and Dr. Boulos. You inspired my research interest.

Secondly, I want to give my sincere thanks to my friends in Windsor, X. Zhang and R. Hu, Y. Luo and M. Liang, for the warmth you brought me and the big help to me when I am alone in Windsor in the past months. Also many thanks to all my friends in Windsor, I can not list all here, for all the friendship and help you gave me in my life. I can not forget to give my thanks to my university friends, Wei Yang, Leo French, Yun Li, and Demin Yin.

And I appreciate the financial help from the University of Windsor.

Finally, my wife, Liping Ma, is the endless support for my life. This thesis can't be completed without her.

# TABLE OF CONTENTS

III. EXISTED SIMILARITY MEASURE OVER ONTOLOGY

IV. EMPIRICAL STUDY OF GENE ONTOLOGY BASED MICROARRAY CLUSTERING

V. EXPERIMENTS AND RESULT

VI. CONCLUSION AND FUTURE WORK

**APPENDICES**

# LIST OF TABLES

# LIST OF FIGURES

CHAPTER I

# INTRODUCTION

Bioinformatics is a discipline applying the knowledge of mathematics, statistics, and computer science into the study of biology at molecular level. In the science of biology, the successful completion of Human Genome Project and the emergence of new technologies such as Microarray greatly advance the development of gene-related research while creating large amount of data pending for analysis. Large scale of collaboration between researchers in the world is required to deal with data and acquire new biology knowledge. The development in information technology and Internet technically enables the large scale of collaboration. On the other side, Gene Ontology (GO) provides known gene-related common knowledge between large biological databases so that the collaborators can use a common language in communication. Gene Ontology is such a controlled vocabulary that it can interpret all the databases and promote the integration of them.

This thesis project promotes the application of Gene Ontology in Microarray analysis by introducing similarity measures over Gene Ontology to perform Microarray clustering. Several current similarity measures over ontology are studied. An On-The-Fly probability is defined to calculate the probability of a term in the current problem space. A new similarity measure between a term and a set is defined, and this measure is also used to define the similarity between sets. The performance of applying these similarity measures is compared by clustering a dataset of which the correct clustering scheme is known.

## 1.1 Organization of this thesis

In Chapter 1, the knowledge pertinent to this thesis topic, including Bioinformatics, Microarray, clustering algorithms and Gene Ontology is introduced in sequence. In Chapter 2, Gene Ontology based Microarray Clustering is explained and analyzed in detail by addressing the research topic of this thesis work. The previous research and study work about the topic is reviewed in Chapter 3, the approach to study GO based Microarray Clustering is proposed and the experimental methods are designed in Chapter 4. The following chapter is the experimental results and the analysis of the results. Finally, in Chapter 6 some conclusions from this empirical study are made, and the possible future work is discussed.

## 1.2 Basic of Biology

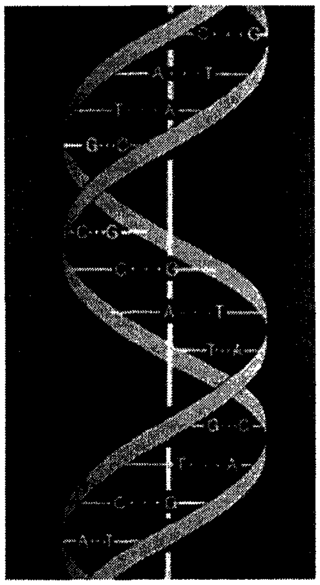


**Figure 1 DNA Double Helix**

The genetic information of every organism is stored in the molecule known as Deoxyribonucleic acid (**DNA**).

The structure of DNA is illustrated by a double helix (Figure 1 shows a segment of DNA double helix), with about 10 nucleotide pairs per helical turn. Each spiral strand is connected to a complementary strand by hydrogen bonding between paired bases, Adenine(A) with Thymine(T) and Guanine(G) with Cytosine(C), which means that each

DNA strand contains the template information for synthesis of a new copy of the other strand.

To better understand, we can regard DNA as a book, or even a library, storing all of the genetic information for synthesizing protein or RNA.

**RNA** is similar to DNA. RNA is formed by a single strand, while the DNA consists of two complementary strands attached to one another, forming a double helix. In the course of synthesizing proteins based on the genetic information on DNA, RNA is a molecular intermediary. Certain RNA is also a source for protein. The four bases in RNA nucleotides are Adenine(A), Guanine(G), Uracil(U), and Cytosine(C). A RNA segment is illustrated in Figure 2.

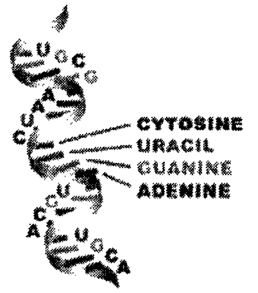


**Figure 2 RNA Single Strand**

There are three major types of RNA:

- **mRNA**, messenger-RNA, which transfers the information about the amino acid sequence from the DNA to the protein synthesis.
- **rRNA**, ribosomal-RNA, which builds up the ribosome together with proteins.
- **tRNA**, transfer-RNA, which transfer amino acids to the ribosome for protein synthesis.

The genetic information on DNA is divided into different segments -- genes. **Gene** is the basic unit of genetic function. One gene contains three parts:

- Regulatory segment, which contains information of initiation and regulating instructions;

- **Exon**, which is the coding part for protein or RNA;
- **Intron**, which is the non-coding part.

A gene is working as a recipe for a particular protein or RNA, in some cases. Usually a protein is synthesized with more than one gene. We call a gene is **expressed** by encoding this gene for synthesizing a protein.

**Protein** is a large 3-dimentional molecule playing structural and functional role as the basic building block for organisms. Huge number of different 3-dimentional structure of the molecules result in the variety of proteins.

A cell functions by using its genes to produce proteins [Coe]. And a gene is transcribed into mRNA before being translated into a protein. The production of mRNA is very exactly a reflection of the activity of a gene, and a lot of genetic information can be understood by studying it.

## 1.3 Development of Bioinformatics

**Bioinformatics** and **computational biology** involve the use of techniques including applied mathematics, informatics, statistics, computer science, artificial intelligence, chemistry and biochemistry to solve biological problems usually on the molecular level [Wiki]. The development of bioinformatics is the result of advances in both computer science and molecular biology over the past 40 years. The building of protein sequence database and the development sequence alignment algorithm in 1970s announced the establishment of this discipline. During 1980s and 1990s, more and more gene and protein sequence databases were built and related algorithms were developed to do the research. The completion of Human Genome Project (1990-2003) is a landmark of Bioinformatics which announces that this discipline had become mature. The emergence of World Wide Internet and the powerful inexpensive Personal Computer make it possible to implement large scale computation and collaboration of scientists, and this advances greatly the development of bioinformatics.

## 1.4 Basic of Microarray

Genes are continuous segments of genomic DNA constructed from four nucleotide blocks, named A, G, T, and C. Each gene can be used to encode a specific mRNA and then translate to a corresponding protein, which imparts biological function in the cell.

The process of converting genetic information at the DNA level into functional proteins is known as **gene expression**. Because cells express their genes only when they are required for a cellular process under specific physiological conditions, how many genes are expressed under this condition is an important clue to gene functions.

For many years, the study of a gene expression had to be done individually—looking at whether a specific gene is turned on (up-regulated, or over-expressed) or turned off (down-regulated, or under-expressed) under certain conditions. During the last half of the 20th century, the analysis of the regulation and function of genes has largely driven step-by-step studies of individual genes and proteins.

A Microarray [Schena 1995] is such a device that measures how many genes are expressed in experiments, with large scale number of genes simultaneously. Thousands of genes can be studied at one time.

A DNA Microarray consists of an orderly arrangement of DNA fragments representing the genes that we focus on. Each DNA fragment representing a gene is duplicated to be enough and assigned a specific location on the Microarray, usually a glass slide, silicon chips or nylon membrane, and then spotted (< 1 mm) to that location. Through the use of highly accurate robotic spotters, some Microarray experiments can contain up to 30000 [NCBI] target spots, allowing molecular biologists to analyze virtually every gene present in a genome.

The Microarray analysis cycle can be simplified into five basic steps: raising a biological question/guess, sample preparation, biochemical reaction, signal detection and data mining and analysis, and then updating the question and keep going to next step until we fully understand the biological question.

With the appearance of Microarray technique, some applications acquire a brilliant achievement in gene research, human disease, drug discovery, and genetic screening and diagnostics.

## 1.5 Microarray Analysis

Microarray is now able to produce large amounts of data about many genes in a highly parallel and rapidly serialized manner and allows scientists to study many, if not all, genes of an organism's at once. This high throughput achievement allows for the global study of changes in gene expression, giving us a complete cellular snapshot.

Microarray differs from traditional research in a number of striking ways [Wall 2001], one of which is the relationship between the amount of experimental time required and the amount of data obtained. Traditional experimental approaches based on gels and filter blots require a relatively large amount of experimental time to obtain a small volume of data, whereas Microarray analysis offers vast quantities of data with relatively little experimental time. Microarrays purchased commercially provide an extreme example, allowing a single researcher to generate millions of data points in a few weeks.

Analysis on Microarray is unique in the history of biology because no other technology has ever involved so much technology, combined expertise from so many different disciplines, including biology, chemistry, physics, engineering, mathematics, and computer science [Rocke 2003], and provided a quantitative and systematic view of a biological system.

How can we understand the role of the genes as a whole in biological function based on so large amount of data? In other words, how can we define the role of each gene (or sequence of genes) in some biological function and subsequently understand how the genes function as a whole?

Discovering patterns of gene expression can help to correlate genes to specific biological functions, and thereby understand the role of genes in biological functions on a genomic scale.

In order to properly comprehend and interpret expression data produced by Microarray technology, some computational and data mining techniques were developed during last decades.

The analysis and understanding of Microarray data is to group genes with similar or correlated patterns of expression together. Some clustering algorithms were employed very well in this field.

## 1.6 Clustering Algorithms

Clustering algorithms are generic tools for pattern recognition, grouping the data-points into groups. The data-points in same groups are very similar and those in different groups are quite distinct.

BSAS and MBSAS algorithms are the simplest clustering algorithms. Hierarchical clustering and k-Means clustering are two major classes of clustering algorithm applied on Microarray datasets. There are other clustering algorithms that can be used in this field, including SOM, and CAST [Jiang 2004].

### 1.6.1 Basic Sequential Algorithmic Scheme (BSAS)

BSAS is very easy to understand, as soon as we understand the threshold of dissimilarity. If the distance from one element to a defined cluster is smaller than the threshold of dissimilarity, we claim that element belong to the cluster. Otherwise, we search for other clusters, or create a new cluster for that element, if no near cluster is found.

One advantage of this algorithm is that the data is presented only once. It is a linear time algorithm with time complexity O(n). Another advantage is that we don't need to know a priori the number of cluster.

The disadvantage is that the threshold of dissimilarity must be deliberately adjusted to accommodate each case.

### 1.6.2 Modified Basic Sequential Algorithmic Scheme (MBSAS)

In BSAS, data is presented only once, so some clusters are fully defined even before some other clusters are created, when some elements might be better fit in the latter clusters than in the former clusters. Modified BSAS presents data twice: The first time is to find the kernel of all clusters, by letting a new element become a new kernel, if the distances from the element to all existed kernels are greater than the threshold of dissimilarity. Next time, all data can find a closest cluster to fit in.

Although data is presented twice, this algorithm is also linear time algorithm with time complexity O(n).

### 1.6.3 Hierarchical Clustering Algorithm

Clustering problem is considered as a sequence of partitions with n samples into k clusters based on the similarity matrix. The basic idea of hierarchical clustering algorithms [Wolkenhauer 2002] [Eisen 1998] is to build a tree as a sequence of partitions, by which the n samples are grouped into one cluster. This tree is called dendrogram. Based on the dendrogram tree and some prior knowledge, a leaf order with maximizing the sum of similarity of adjacent elements in this order can be presented.

For its simplicity, this algorithm becomes one of the most widely used algorithms on Microarray analysis. In [Eisen 1998], hierarchical clustering algorithm was shown to be an elegant one for the analysis on Microarray dataset. And in [Harrington 2001] [Dhanasekaran 2001] [Perou 2000], this algorithm was applied to analyze the Microarray datasets on the molecular classification of cancers and biological modeling. David [Eppstein 1998] developed data structures to obtain a faster hierarchical clustering algorithm.

A problem arising is how to use this dendrogram tree that resulted from Hierarchical clustering algorithm, and how to determine if a sub-tree is a cluster instead of a part of bigger cluster. We need additional algorithms to interpret the binary tree into a form that can be understood by analyzers.

For the convenience of analysis, usually, this binary tree is displayed with their leaves in a linear order. In [Bar-Joseph 2001], one optimal leaf-ordering algorithm was introduced. This algorithm makes the optimal leaf ordering maximize the sum of the similarity of adjacent elements in the ordering. And this algorithm also can help users identify and interpret the data.

The time complexity of Hierarchical clustering algorithm is at least $O(n^2)$

### 1.6.4 k-Means Clustering Algorithm

k-Means clustering algorithm [MacQueen 1967] is a clustering algorithm based on mixture model [Nurmi 2004]. It supposes the dataset is combined from multiple populations and split data point into these subpopulations.

First, k elements are randomly chosen as center of k clusters. Then all other elements can be group into k clusters by choosing the closest center. After all the elements are allocated into k clusters, the centers of the clusters are recalculated as the centers of established clusters, and then all the elements are allocated again. Repeat the center calculation and element allocation until the clusters are stable.

For its simplicity, k-Means and fuzzy k-Means are widely used on Microarray datasets. In [Futschik 2002], Futschik and Kasabov analyzed fuzzy k-Means clustering algorithm and cluster validity, and addressed the selection of parameters to gene expression data. In [Gasch 2002], Gasch *et al.* applied fuzzy k-Means clustering to identify overlapping clusters of yeast genes with environmental changes.

Although k-Mean and fuzzy k-Means clustering algorithms are used very well and widely, there are some weaknesses that these two algorithms can not avoid.

First, the result is based on the initialization of membership and mean. From the previous discuss, we can see that the basic idea of k-Means and fuzzy k-Means is to get the parameters known by a gradient descent. For the gradient descent, the algorithms only guarantee to get a local optimum. This is the main reason that initializing is very sensitive to the final result.

One way to solve this problem is to start randomly at different points. In [Bradley 1998] [Fayyad 1998], the authors introduced a better way to get a refined start point as initialization point.

Second, there is a need to have the prior knowledge on K. One of the ways to get the best K known is to run k-Means clustering algorithm on all of possible values of k, calculate the costs of objective function and choose the best one among them. This still is an open-problem in this field.

Although it is not verified, people believe the time complexity of k-Means algorithm is O(kn), when k is the number of clusters, and n is the number of elements.

### 1.6.5 Self-Organizing Map

Self-organizing map [Kohonen 1990] is a clustering algorithm very similar to k-Means clustering algorithm. It is also called SOM algorithm.

However, SOM is a two-level clustering algorithm. First, SOM projects N high-dimension data points onto a low-dimension map, usually a 2-dimension map, instead of dividing the original data points into the k clusters directly, and then to classify the units on this map into K clusters.

With this 2-level approach, there are two benefits [Veenman 2002]: decreasing the cost of computation and noise reduction.

In [Tamayo 1998], Tamayo *et al.* employed Self-organizing maps clustering algorithm to interpret the patterns of Microarray data sets

SOM is not only a good method to cluster data sets, but also a good tool to display Microarray data sets. Lee [Lee] presented a method to display the result of clustering.

This algorithm has the same time complexity with k-Means algorithm.

### 1.6.6 Clustering Affinity Search Technique

Using k-Means algorithm or Self-Organizing Map, you must decide the number of clusters first. An algorithm without the prior knowledge of k, named Clustering

Affinity Search Technique (CAST) [Singh 2002], was presented by Ben-Dor *et al.* first in Journal of Computational Biology [Ben-Dor 1999].

This algorithm groups data points into clusters based on the average similarity, or affinity, between the current cluster and un-clustered data points.

In CAST algorithm, we don't need the prior knowledge of k, but there is a fatal drawback, that is, we have to initialize a control parameter, threshold of dissimilarity. This parameter is affecting the shape of clustering structure. In [Bellaachia 2002], Abdelghani *et al.* proposed an enhanced CAST algorithm, in which, there is a dynamic threshold instead of the fixed threshold. And this value will be computed at the beginning of the generation of new cluster.

## 1.7 How to Evaluate the Clustering Result

When we have a clustering result, the first question we will ask is: How accurate the result is? When we know the actual clusters the items should be ("ground-truth"), the accuracy is calculated as

$$accuracy = \frac{number\ \ of\ \ correctly\ \ clustered\ \ items}{number\ \ of\ \ items}$$

To avoid random error of one clustering algorithm, the algorithm should be run several times to get the average accuracy.

$$\overline{accuracy} = \frac{1}{N}\sum_{i=1}^{N} accuracy_i$$

Another question being asked is: Is the result reproducible? Some algorithms, such as k-Means, involve random number during the calculation. So if I run the same algorithm using the same data, how likely that I will get the same or similar result? How stable the algorithm is? The experiments should be run several times, and the standard deviation of accuracy is calculated to evaluate the stability of the algorithm:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (accuracy_i - \overline{accuracy})^2}$$

## 1.8 Basic of Ontology

“Ontology” has been a philosophy jargon since Aristotle times, and it means the nature of existence. Computer scientists adopted this word to express a formally structured vocabulary in a discipline. In this vocabulary, items and relations between two items are well defined to present the knowledge in this discipline.

Ontologies offer a mechanism by which knowledge can be represented in a form capable of machine processing[Lord 2003]. Ontologies can be provided in Rational Database format or XML format.

Now ontology becomes the core of Semantic Web, because the geographically distributed Web forms information islands in the Internet, and the use of ontology can interpret meanings of information in different islands, reduce the confusion, and integrate data automatically. The decentralized infrastructure makes the communication and collaboration over Internet easy. Every one can focus on her own part of the project independently and integration of their work will be streamlined since every part of the collaboration follows the same ontology and plays her own role. Every one can also build new ontologies, and cooperate with the third part without the permission of her collaborators. The collaboration will be stronger and stronger as more and more collaborators join in and share their knowledge. This decentralized infrastructure breaks down the barrier between languages, geographical distance, automates the integration of knowledge, and leads to the evolution of knowledge.

## 1.9 Introduction to Gene Ontology

There are many biological databases emerged in the genome era, addressing different efforts in different biology communities. These biological databases are speaking different languages, so it's vital to have a common ontology, which can interpret all the databases and promote the integration of them.

The Gene Ontology project is a collaborative effort to address the need for consistent descriptions of gene products in different databases. Base on the common understanding that genes and proteins conserve their function in all eukaryotes, including fruitfly, mouse, human, or Arabidopsis thaliana [GO 2000], fourteen Databases organized the Gene Ontology Consortium and create a controlled vocabulary to describe the gene-related knowledge we have so far, and that is the Gene Ontology. The decentralized infrastructure of the collaboration enables every participant to develop its own database independently, and the knowledge, which is developed by one collaborator, can be shared by all other collaborators.

GO has three categories: biological process, cellular components and molecular functions. The structure reflects the biological knowledge we currently have, and it can help to understand and organize new knowledge. In that sense, GO is a dynamic vocabulary, because it will always change according to the new biological knowledge we learn from the development of biology study. In the GO project, researchers are interested in the following activities:

1. Creation and maintenance of the ontologies;

2. Making associates (annotations) between the ontologies and the genes and gene products in the collaborating databases;

3. Developing tools that facilitate the creation, maintenance and use of ontologies.

And the community has created many application tools in every category.

The terms within each category are linked in defined *is-a* relationships or *part-of* relationships that reflect current biological knowledge. GO is represented as a Directed Acyclic Graph. A term can have several parents, A typical DAG structure of GO is:

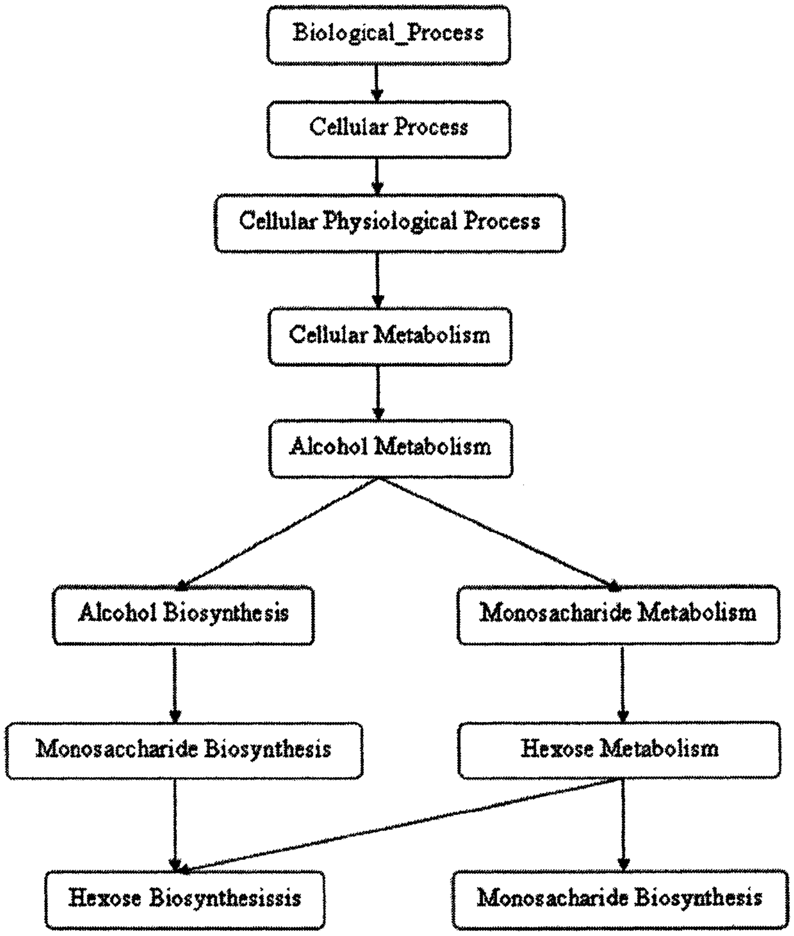


**Figure 3 The Gene Ontology DAG Structure**

In Figure 3, the biological process term hexose biosynthesis (GO:0019319) has two parents, monosaccharide biosynthesis (GO:0046364) and hexose metabolism (GO:0019318). The terms are used to annotate gene products. As they form a standard vocabulary across many biological resources, this shared understanding provides "a valuable, computationally accessible form of the community's knowledge". [Lord 2003] A program agent using the taxonomy of one database can understand the taxonomy of another database by means of the common ontology.

## 1.10 History and Future of Gene Ontology

There are many biological databases emerged in the genome era, addressing different efforts in different biology communities. In 1995 Schena *et al.* [Schena 1995] developed Microarray, a new technology that can analyze thousands of genes in very short time. This new technique greatly accesses the process of gene produces and creates big amount of data, highly increases the size of those biological databases.

Researchers found that genes and proteins conserve their function in all eukaryotes, including fruitfly, mouse, human, or Arabidopsis thaliana, so it is possible to automatically transfer "biological annotations from the experimentally tractable model organisms to the less tractable organisms" [GO 2000] based on gene and protein similarity. The building of **a common vocabulary between different databases** was imperative, and the new techniques of Computer Science including the development of Internet, ontology, and Relational Database Management System provided the best opportunity for the collaboration of biological database community.

There were some failed collaborations before the project of GO, and Lewis concluded that "the biggest impediment was getting the many people involved to agree on virtually everything" [Lewis 2004] when building a federated system. The decentralized infrastructure overcomes this impediment by allowing collaborators to keep their disagreement locally, but present the common knowledge in a controlled vocabullary.

In 1998, FlyBase, SGD (Sacharomyces Genome Database) and MGD (Mouse Genome Database and Gene Expression Database) started the GO project to create a common vocabulary and apply it to describe the biological process, molecular function and cellular component for every gene in their respective databases.

Gene Ontology has grown enormously after it was started, and it has become a big success. By the time writing this paper (December, 2006), there are fourteen members in the GO Consortium, including: Berkeley Bioinformatics and Ontology Project, dictyBase, FlyBase, GeneDB, Gene Ontology Annotation @ EBA, Gramene, MGD & GXD, Rat Genome Database, Reactome, SGD, The Arabidopsis Information Resource, The Institute for Genomic Research, WormBase and Zebrafish Information Network. Between them the Gene Ontology Annotation @ EBA (GOA) is another project that aims to collaborate GO terms with existed UniProt(Project Collaborators are UniProtKB, Swiss-Prot, TrEMBL, PIR-PSD) and InterPro(Project Collaborators are UniProt, PROSITE, Pfam, PRINTS, ProDom, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, Gene3D, PANTHER) databases, so the GO terms can be applied to most mainstream biological databases. In the gene community it has been the standard for functional annotation. There are 1,000 literatures in the PubMed, either referencing or utilizing GO. Khatri said the automatic ontological analysis approach is "the *de facto* standard for the secondary analysis of high throughput experiments"[Khatri 2005].

Gene Ontology applies computer science techniques into biology society, and it also provides feedback to the computer science community. GO has become one of the success stories of ontology. According to [GO 2006], GO has been used as a testbed of applying description logic approaches to building sound, complete and logically consistent ontologies, and has featured in research into machine-processable ontologies and into the automated checking of ontological consistency. The success of utilizing natural language processing, information extraction from texts, knowledge discovery in the building of GO has inspired computer scientists to put more effort in these areas.

Open Biomedical Ontologies (OBO) is an umbrella organization including well-structured controlled vocabularies for shared use across different biological and medical domains[OBO]. 56 ontologies are collected. The share vocabularies between different disciplines will promote the share knowledge and the collaboration in biological and medical community.

Sequence Ontology is a part of the Gene Ontology, and it provides terms and relationships for describing the features and attributes of biological sequences including DNA, RNA and proteins [Lewis 2004]. It has been mapped to homologous terms in other biological ontologies to facilitate the integration with existing genome annotation projects.

The Gene Ontology is a dynamic controlled vocabulary, and the terms will always be refined, reorganized with the discovery of new knowledge and techniques. In building new ontologies in biological and medical domains, more data are represented in a common base and experts in different areas can share their knowledge. The collaboration between different areas will be automated by machine-processable ontologies.

The Gene Ontology is far beyond completion. One of the shortcomings of GO is that it has 3 categories: biological process, cellular components and molecular functions, but there is no link between terms in different categories. For example, actin cortical patch (GO:0030479) is a term in cellular components, defined as the discrete actin-containing structure found at the plasma membrane in cells. actin cortical patch assembly (GO:0000147) is a term in biological process, defined as the assembly of an actin cortical patch. Apparently the actin-containing structure has strong relation with the assembly of itself, which should have been described in this controlled vocabulary. But because there is no link between terms in different categories, the only paths between these two terms are through the root of the ontology. Some researchers even described the three categories as three different ontologies [Kennedy 2003].

## 1.11 Gene Ontology Tools

Many tools have been developed to maintain and utilize GO. GO Tools are categorized as 4 types: searching and browsing tools, annotation tools, Microarray analysis tools and others. Searching and browsing tools are ontology-building tools used to browse and edit the ontology; Annotation tools are used to interpret every item, linked existed gene knowledge to the ontology. Microarray analysis tools have been developed actively to address the need of analyzing high throughput gene expression data in Microarray. Some other tools are also created to make full use of the Gene Ontology.

CHAPTER II

# GENE ONTOLOGY BASED MICROARRAY CLUSTERING

## 2.1 Data-driven Microarray Clustering

The analysis and understanding of Microarray data is to group genes with similar or correlated patterns of expression together, therefore clustering algorithms such as hierarchical clustering and k-Means algorithm have been deployed to cluster Microarray expression data so that the patterns of gene expression are discovered to correlate genes to specific biological functions, and hence, the role of genes in biological functions on a genomic scale can be derived. Since 1995 when the Microarray technology was developed [Schena 1995], clustering has become the major analysis tool in Microarray analysis. Euclidean distance is the major dissimilarity measure in clustering expression data.

$$\text{dis}_{\text{Eu}}(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

## 2.2 Applying Gene Ontology in Microarray Clustering

Gene Ontology is a controlled vocabulary that describes the gene-related knowledge we have so far. Many efforts have been made in applying GO to Microarray clustering.

One approach [Pavlidis 2002] making use of the ontology for analyzing Microarray-experiments is to annotate the "functional groups". After the genes are clustered by expression data, all genes in one single group are supposed to have a specific biological function. A *post hoc* analysis using Gene Ontology can label the group by annotated genes in that group, identify the predominant set of GO terms that describe the group, and then the unannotated genes in the same group are predicted to have the same or related labels.

A second approach is to search for over-representation of particular GO nodes or GO categories in a list of genes. Applications of this approach include FatiGO [Al 2004] and MAPPFinder [Doniger 2003]. Visualized analysis results are also provided by this approach to help researchers to inspect the results.

A third approach is to evaluate the gene expression clustering result using GO information [Datta 2006], [Bolshakova 2006]. The GO information is applied to calculate Biological Homogeneity Index in order to evaluate the biological similarity of the clustered result.

## 2.3 Similarity Measure over Ontology

In the practice of applying GO in Microarray clustering, different similarity measures over ontology are being used by different researchers. How well do these similarity measures present the similarity of (known) biology knowledge between genes and gene products, and how well do they perform in the clustering? Several similarity measures are studied in Chapter 3.

## 2.4 Directly Applying GO in Microarray Clustering

The similarity measure over Gene Ontology can be combined with Euclidean distance to perform Microarray clustering. The combined measure has both the expression data and (known) biological information from GO. A good measure can show the real biological relation between genes. A clustering algorithm using this measure should gain better performance than using bad measures.

## 2.5 Thesis Contribution

This thesis work studies the existed similarity measures over Gene Ontology and introduces measures in Chapter 3. In Chapter 4, different aspects of the similarity measure is discussed, a scheme of evaluating different measures is introduced. One method to combine the measures over GO with the Euclidean distance to perform Microarray analyses is proposed. In the Chapter 5, the measures are being put into empirical study to evaluate the pros and cons.

CHAPTER III

# EXISTED SIMILARITY MEASURE OVER ONTOLOGY

In the practice of utilizing Gene Ontology in Microarray clustering, some different similarity measures over ontology are used by different researchers.

Before we dive into literatures of similarity measure over Gene Ontology, let's look at a sample scheme which will be used through this chapter. In Figure 4 there is a sub-graph of Gene Ontology having 8 terms. Protein A is annotated by Term 5 and 7, while Protein B is annotated by Term 5, 6 and 8. Term 1 is the root of this DAG. The numbers in the parentheses are the probabilities of the terms appear in some context. Although some of the measures are defined as measures over common taxonomy, they can be easily applied to ontology.

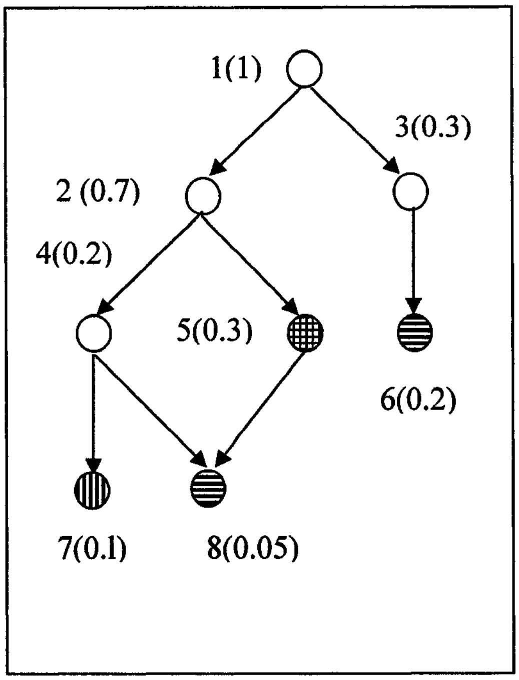


**Figure 4 A sample sub-graph of Gene Ontology. Protein A is annotated by Term 5 and 7, while Protein B is annotated by Term 5, 6 and 8**

## 3.1 Wu & Palmer's Measure

Wu & Palmer [Wu 1994] use the least general common ancestor of two terms to define the similarity.

$$\text{sim}(x, y) = \frac{2 \times N3}{N1 + N2 + 2 \times N3}$$

Assume c is the least general common ancestor of two terms, N1 is the number of nodes on the path from x to c, N2 is the number of nodes on the path from y to c, and N3 is the number of nodes on the path from c to root.

***[Example]****For example, to calculate the similarity between Term 5 and Term 7 in the sample scheme, the least general common ancestor is Term 2, so:*

*N1=2*

*N2=3*

*N3=2*

$$sim(5,\ 7) = \frac{2 \times 2}{2 + 3 + 2 \times 2} = 0.44$$

## 3.2 Resnik's Measure

Resnik [Resnik 1995] introduces the probability of encountering a term in an ontology, or taxonomy. The higher level a term is in the ontology, the more abstract it is, and the greater the probability to encounter the term is. Probability of the unique top node (if exists) is 1. Based on Shannon's information theory, the smaller the probability is, the more *information content* it has. The information content is quantified as negative log of probability.

$$\text{IC(c)} = -\log p(\text{c})$$

The more information two terms share in common, the more similar they are. So the similarity of two terms can be defined as the maximal information content of common ancestors:

$$\text{sim(x, y)} = \max_{c \in S(x,y)} [-\log p(c)]$$

where S(x, y) is the set of common ancestors of x and y. *p*(c) is the probability of term c. It is calculated as the relative frequency of term c in the context.

When a concept is associated with several terms, the similarity of two concepts is defined as the maximum similarity from a term of one concept to a term of another concept.

$$\text{SIM(X, Y)} = \max_{x \in X, y \in Y} sim(x, y)$$

***[Example]*** *The common ancestors of Term 5 and Term 7 are Term 1 and Term 2. Term 1 is the root of the ontology which has the biggest probability as 1,*

*-log p(Term 1) = -log1 = 0*

*-log p(Term 2) = -log 0.7 = 0.15*

$$sim(5,\ 7) = \max_{c \in S(5,7)} [-\log p(c)] = -\log p(2) = 0.15$$

***[Example]*** *The maximum similarity between two sets is the similarity between Term 7 and Term 8*

$$SIM(Protein\ A,\ Protein\ B) = \max_{x \in \Pr oteinA, y \in \Pr oteinB} sim(x, y)$$

$$= sim(7,\ 8) = -\log p(4) = 0.70$$

## 3.3 Lin's Measure

Lin [Lin 1998] used a measure similar to Resnik's measure.

$$\text{sim(x, y)} = \frac{2 \times \log p(c)}{\log p(x) + \log p(y)},$$

c is the least general common ancestor of x and y.

$$\text{SIM(X, Y)} = \frac{-2 \times \sum_{x \in X, y \in Y} sim(x, y)}{\sum_{x \in X} \log p(x) + \sum_{y \in Y} \log p(y)}$$

***[Example]** sim(5, 7)* $= \frac{-2 \times [-\log p(2)]}{\log p(5) + \log p(7)} = \frac{-2 \times 0.15}{-0.52 - 1} = 0.20$

## 3.4 Jiang & Conrath's Measure

The measure used by Jiang & Conrath [Jiang 1998] is of distance measure, which is the reverse of similarity measure:

$$\text{dis(x, y)} = -\log p(x) - \log p(y) - 2 \times \max_{c \in S(x,y)} [-\log p(c)]$$

Similarity measure can be defined as:

$$\text{sim(x, y)} = \frac{1}{dis(x, y) + 1}$$

***[Example]** sim(5, 7)* $= \frac{1}{-\log p(5) - \log p(7) - 2 \times [-\log p(2)] + 1}$

$$= \frac{1}{1 + 0.52 - 2 \times 0.15 + 1} = 0.45$$

## 3.5 Lord's Measure

Lord *et al* [Lord 2003] applied the Resnik's, Lin's and Jiang & Conrath's similarity measures between two terms above to calculate the semantical similarity of GO terms. The probability of a term is defined as the probability of this term occurring in the SWISS-PROT-Human database.

$$\text{sim(x, y)} = \max_{c \in S(x,y)} [-\log p(c)]$$

x and y are respectively the set of annotation terms of two gene products.

Similarity between two gene products is defined as average similarity of all annotation terms.

## 3.6 Kennedy's Measure

Kennedy *et al* [Kennedy 2003] use a similarity measure adapted from Tanimoto Measure. In Tanimoto Measure, the number of common members is divided by the number of all members to calculate the similarity of two sets.

$$\text{SIM(X, Y)} = \frac{n_{X \cap Y}}{n_{X \cup Y}}$$

In Kennedy's measure, X and Y are the set of annotation terms and their ancestors of two gene products. Because the higher level a term is in the ontology, the more general it is, and less important it is. A weighted measure is being used:

$$\mathrm{SIM'}(X, Y) = \frac{n'_{X\cap Y}}{n'_{X\cup Y}}, \text{ where } n_X = \sum_{i\in X} c^{d_i}, \; c \subset [0, 1],$$

$d_i$ is the distance of the term with index i from its associated descendent in the original set of terms, and c is the weight constant.

***[Example]*** *Term 5 and Term 7 annotate Protein A, and they have ancestors Term 1, Term 2 and Term 4. Term 5, Term 6 and Term 8 annotate Protein B, and they have ancestors Term 1, Term 2, Term 3, Term 4 and Term 5. The conjunction of two sets is {Term 1, Term 2, Term 4, Term 5}*

$$SIM(Protein\ A, Protein\ B) = \frac{n_{\Pr oteinA\cap \Pr oteinB}}{n_{\Pr oteinA\cup \Pr oteinB}} = \frac{4}{8} = 0.5$$

$$SIM'(Protein\ A, Protein\ B) = \frac{n'_{X\cap Y}}{n'_{X\cup Y}} = \frac{c^{d_1} + c^{d_2} + c^{d_4} + c^{d_5}}{\sum_{i=1}^{8} c^{d_i}}$$

The definition of $d_i$ is ambiguous, because one term can have more than one descendents in the original set of GO terms, Term 6 (Protein B) and Term 7 (Protein A) are the descendents of Term 1; The distance of Term 1 from Term 6 is 2, while the distance of Term 1 from Term 7 is 3. So we can't decide the value of $d_1$ by its definition.

CHAPTER IV

# EMPIRICAL STUDY OF GENE ONTOLOGY BASED MICROARRAY CLUSTERING

In the practice of applying Gene Ontology in Microarray clustering, different similarity measures over ontology are being used by different researchers. How well do these similarity measures present the similarity of (known) biology knowledge between genes, and how well do they perform in the clustering?

The Gene Ontology is a controlled vocabulary describing the gene-related knowledge we have so far. Its Directed Acyclic Graph (DAG) structure allows one term to have more than one parent, but a term is not an ancestor of itself.

According to [GO], there are two kinds of directed edges in the GO: *is_a* and *part_of*. *is_a* edge means relationship which indicates the child term is a subclass of the parent term. For example, nucleus *is_a* cellular_component. *part_of* edge means a relationship which explains when the child term is present, it is always a part of the parent term, but the child term does not always have to be present. For example, nucleus *part_of* cell means nuclei are always part of a cell, but not all cells have nuclei. From the description above, the two relations don't have big difference, so they should have the same weight in the similarity measure.

GO has a "true path rule". If a term describes one gene product, then all its ancestor terms must also apply to that gene product. So when one gene product is explicitly annotated by some GO terms, the gene product is implicitly annotated by all the ancestor terms of these GO terms.

## 4.1 On-The-Fly probability

In the Resnik's, Lin's and Jiang & Conrath's measure, the probability of a term is defined as the relative frequency of the term in the context. The higher level a term is in the ontology, the more abstract it is, and the bigger the probability is. Probability of the unique top node (if exists) has the biggest probability as 1.

Lord *et al.* [Lord 2003] use the GO annotation in the SWISS-PROT-Human database as context to define the probability of a term. Probability of each term is the frequency that this term is used to annotate proteins of SWISS-PROT-Human database. SWISS-PROT is one of the biggest annotated protein sequence databases in the world, and the SWISS-PROT-Human is the Human section of the whole database. So this probability can not be applied to other gene products, such as fruitfly proteins.

In this thesis project **On-The-Fly probability** is created to define the probability of terms in the current scene. The context is defined as all the terms occurred in the problem space. A term occurs if a term or any of its descendents occurs. The frequency of a term occur in the problem is the probability of this term. The more frequent a term is used in this scene, the less important it is to measure similarity. The probability of each term is changing dynamically based on the dataset, rather than a predefined value. For example, when 205 yeasts of the dataset we will use in the empirical study (it will be introduced in Chapter 5) is being clustered, the 604 GO terms form the context of the ontology, and the probability of each term is the frequency of that term being used to annotate the 205 yeasts. Probability of one term is different from case to case.

## 4.2 Similarity When Two Terms Are Identical

Usually when two terms are identical, we can expect to get the maximal value of similarity. But this principle can not simply apply to terms in an ontology. For example, if two proteins have one common annotation term, but the annotation term is on the top level of the ontology, that means the term is very general, therefore we can not assert that two proteins are "very similar". If the term is in the low level of the ontology, then it is very specific, we can think the two proteins as closely related, "very similar". So it is reasonable that the similarity is based on the depth or probability of the identical term.

For example, Resnik's measure between two terms is:

$$\mathrm{sim(x, y)} = \max_{c \in S(x,y)} [-\log p(c)]$$

So:

$$\mathrm{sim(5, 5)} = \text{-}\log\ p(5) = 0.52,$$

$$\mathrm{sim(7, 7)} = \text{-}\log\ p(7) = 1$$

## 4.3 Similarity of Two Sets

Wu & Palmer, Jiang & Conrath didn't define the similarity measure of two sets. Resnik defines it as the maximum similarity from a term in one set to a term in the other set. This measure performs well in the situation where there is few terms in each set. In the situation where a protein is annotated by 10 or 20 terms on the average, this measure loses a lot of information and can not express the fact (biological knowledge) correctly. So when Lord applies these similarity measures into Gene Ontology, he uses the average similarity of terms between two sets as the similarity of sets[Lord 2003].

Using the sample described in Chapter 3, Protein A is annotated by Term 5 and 7, while Protein B is annotated by Term 5, 6 and 8. The similarity of {5, 7} and {5, 6, 8} is:

SIM(Protein A, Protein B) = SIM({5, 7}, {5, 6, 8})

$$=\frac{1}{6}(\mathrm{sim}(5,5\}+\mathrm{sim}(5,6\}+\mathrm{sim}(5,8\}+\mathrm{sim}(7,5\}+\mathrm{sim}(7,6)+\mathrm{sim}(7,8))$$

When Term 5 in the set of Protein A is compared with Term 5 in the set of Protein B, we know that the two proteins have the same term, and the similarity is high. The next step is to compare Term 7 with the set of Protein B. We shouldn't compare Term 5 to other terms in the other set again. So a definition of similarity of one term to one set is defined as:

$$\mathrm{Sim(x, Y)} = \max_{y\in Y} sim(x,y)$$

And the similarity of two sets is defined as:

$$\mathrm{SIM(X, Y)} = \frac{1}{2}(\frac{1}{|X|}\sum_{x\in X} Sim(x,Y)+\frac{1}{|Y|}\sum_{y\in Y} Sim(y,X))$$

When applying this measure to the sample above,

SIM(Protein A, Protein B)

$$=\frac{1}{2}(\frac{1}{2}(sim(5,5)+sim(7,8))+\frac{1}{3}(sim(5,5)+sim(6,5)+sim(8,7)))$$

## 4.4 Modified Kennedy's Measure

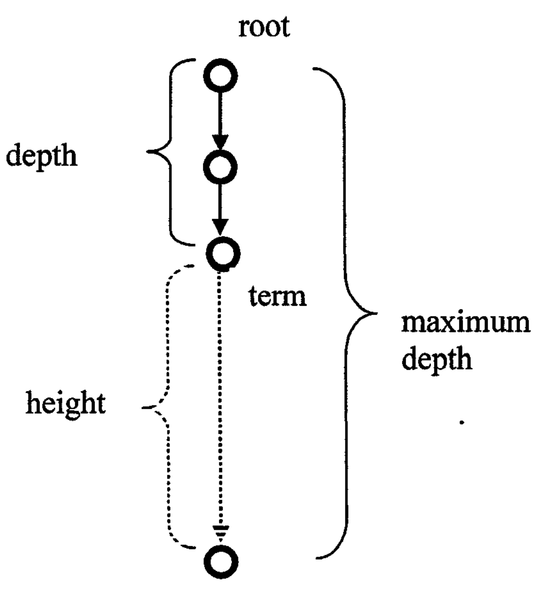


Figure 5 Height is the maximum depth of the graph minus depth

When we were trying to apply the sample scheme to Kennedy's Measure in Section 3.6, we found the definition of $d_i$ (the distance of the common ancestor with index i from its associated descendent in the original set of terms) is ambiguous, because one term can have more than one descendents in the original set of GO terms. But the value of distance of one term from the root is fixed. that is the depth of the term. The reverse of depth is **height**, which has value of the maximum depth of the graph minus depth of the term.

$$\text{height}_i = \text{depth}_{max} - \text{depth}_i$$

In Kennedy's measure, the bigger $d_i$ is, the further the common ancestor is from the descendent, and the less important it is in calculating the similarity. Height has the same property as $d_i$. Therefore in this thesis project, the height is used instead of $d_i$.

## 4.5 Combining Measure over GO with Euclidean Measure of Expression Data

Most Microarray analysis use Euclidean distance as dissimilarity measure in clustering Microarray. We can directly apply Gene Ontology in clustering Microarray expression data by combining measure over GO with the Euclidean measure of expression data. The simplest way is to use the weighted sum of two similarity measures. Be noted that the Euclidean distance is a dissimilarity measure. So the similarity can be:

$$\mathrm{SIM}_{\mathrm{combine}} = \mathrm{w}_1\mathrm{SIM} + \mathrm{w}_2 \frac{1}{1 + EuclidenDis\tan ce}$$

We can simplify this formula by letting $\mathrm{w}_2$=1-$\mathrm{w}_1$, while $\mathrm{w}_1$ is a float number between 0 and 1:

$$\mathrm{SIM}_{\mathrm{combine}} = \mathrm{w}_1\mathrm{SIM} + (1\text{-}\mathrm{w}_1) \frac{1}{1 + EuclidenDis\tan ce}$$

## 4.6 Three Categories of the Gene Ontology

GO has three categories: biological process, cellular components and molecular functions. There is no link between terms in different categories, and they can be regarded as three separate ontologies. Microarray is studying the biological process of gene expression when the environment is changed, so when we are applying Gene Ontology into clustering Microarray gene expression data, we have strong reason to belief that the category of biological process is more relevant to this topic than other categories in GO. We can use this category alone when calculating the similarity over the ontology. So in the empirical study of this thesis work, performance of using this category alone is compared with the performance of using the whole ontology.

## 4.7 List of All Similarity Measures

In this thesis project, the performance of these similarity measures is compared. To be clear, the following is all the similarity measures involved in the empirical study.

$$\mathrm{sim}_1(x, y) = \frac{2\times N3}{N1+N2+2\times N3} \qquad \text{(Wu \& Palmer's measure)}$$

$$\mathrm{Sim}_1(x, Y) = \max_{y\in Y} sim_1(x,y)$$

$$\mathrm{SIM}_1(X, Y) = \frac{1}{2}(\frac{1}{|X|}\sum_{x\in X} Sim_1(x,Y) + \frac{1}{|Y|}\sum_{y\in Y} Sim_1(y,X))$$

$$\mathrm{sim}_2(x, y) = \max_{c\in S(x,y)} [-\log p(c)] \qquad \text{(Resnik's measure)}$$

$$\mathrm{Sim}_2(x, Y) = \max_{y\in Y} sim_2(x,y)$$

$$\mathrm{SIM}_2(X, Y) = \frac{1}{2}(\frac{1}{|X|}\sum_{x\in X} Sim_2(x,Y) + \frac{1}{|Y|}\sum_{y\in Y} Sim_2(y,X))$$

$$\mathrm{sim}_3(x, y) = \frac{-2\times \max_{c\in S(x,y)} [-\log p(c)]}{\log p(x) + \log p(y)} \qquad \text{(Lin's measure)}$$

$$\mathrm{SIM}_3(X, Y) = \frac{-2\times \sum_{x\in X, y\in Y} sim_3(x,y)}{\sum_{x\in X} \log p(x) + \sum_{y\in Y} \log p(y)}$$

$$\mathrm{Sim}_3(x, Y) = \max_{y\in Y} sim_3(x,y)$$

$$\mathrm{SIM}_4(X, Y) = \frac{1}{2}(\frac{1}{|X|}\sum_{x\in X} Sim_3(x,Y) + \frac{1}{|Y|}\sum_{y\in Y} Sim_3(y,X))$$

(There is no $\mathrm{sim}_4(\mathrm{x}, \mathrm{y})$)

$$\mathrm{sim}_5(\mathrm{x}, \mathrm{y}) = \frac{1}{-\log p(x) - \log p(y) - 2 \times \max\limits_{c \in S(x,y)}[-\log p(c)] + 1} \quad \text{(J\&C's measure)}$$

$$\mathrm{Sim}_5(\mathrm{x}, \mathrm{Y}) = \max_{y \in Y} sim_5(x, y)$$

$$\mathrm{SIM}_5(\mathrm{X}, \mathrm{Y}) = \frac{1}{2}\left(\frac{1}{|X|}\sum_{x \in X} Sim_5(x, Y) + \frac{1}{|Y|}\sum_{y \in Y} Sim_5(y, X)\right)$$

$$\mathrm{SIM}_6(\mathrm{X}, \mathrm{Y}) = \frac{n_{X \cap Y}}{n_{X \cup Y}}, \quad \text{(Tanimoto measure)}$$

where $n_X$ is the number of elements in set X

$$\mathrm{SIM}_7(\mathrm{X}, \mathrm{Y}) = \frac{n'_{X \cap Y}}{n'_{X \cup Y}}, \quad \text{(Modified Kennedy's measure)}$$

where $n_X = \sum_{i \in X} c^{h_i}$ , $c \subset [0, 1]$, $h_i$ is the height of term i.

$p$(c) is the On-The-Fly probability of term c in the current scene

From $\mathrm{SIM}_1$ to $\mathrm{SIM}_7$ are original measures. Some other derivative measures are define as $\mathrm{SIM}_8$ to $\mathrm{SIM}_{28}$:

$\mathrm{SIM}_8$ to $\mathrm{SIM}_{14}$ are the weighted sum of above measure and Euclidean measure:

$$\mathrm{SIM}_{8\text{-}14}(\mathrm{X}, \mathrm{Y}) = \mathrm{w}_1\mathrm{SIM}_{1\text{-}7}(\mathrm{X}, \mathrm{Y}) + \mathrm{w}_2 \frac{1}{1 + EuclideanDis\tan ce(\mathrm{X}, \mathrm{Y})}$$

SIM15 to SIM21 have the same formula as SIM1 to SIM7, but only use the biological process category of the ontology.

$\mathrm{SIM}_{22}$ to $\mathrm{SIM}_{28}$ are the combined measure using $\mathrm{SIM}_{15}$ to $\mathrm{SIM}_{21}$:

$$\mathrm{SIM}_{22\text{-}28}(\mathrm{X}, \mathrm{Y}) = \mathrm{w}_1\mathrm{SIM}_{15\text{-}21}(\mathrm{X}, \mathrm{Y}) + \mathrm{w}_2 \frac{1}{1 + EuclideanDis\tan ce(\mathrm{X}, \mathrm{Y})}$$

## 4.8 Comparing the Similarity Measures

To compare the results of applying different similarity measures, a k-Means algorithm is implemented, and different similarity measures are applied to cluster a dataset. Because the correct clustering scheme is known already, we can check the performance of the clustering result with the correct answer to see how well each similarity measure performs.

Accuracy of each run is calculated as

$$\text{accuracy} = \frac{number\ \ of\ \ correctly\ \ clustered\ \ items}{number\ \ of\ \ items}$$

To avoid random error of one clustering algorithm, the algorithm is run several times to get the average accuracy.

$$\overline{accuracy} = \frac{1}{N}\sum_{i=1}^{N} accuracy_i$$

The standard deviation of accuracy is calculated to evaluate the stability of the algorithm:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (accuracy_i - \overline{accuracy})^2}$$

CHAPTER V

# EXPERIMENTS AND RESULT

## 5.1 Introduction

### 5.1.1 Dataset

Assessment of clustering algorithm requires dataset for which we independently know how the genes should be clustered. In addition, Microarray expression data of the genes should be available to perform clustering using combined measures.

The dataset involves 205 genes. They were chosen by Yeung *et al.* [Yeung 2003] from Ideker *et al.*'s yeast galactose-utilization pathway [Ideker 2001] in their study. The genes should be categorized into 4 groups. The Microarray expression data of these 205 genes is also available. 20 sets of expression data are provided, and each set has 4 time points. The first set is used in my study. The dataset can be downloaded from *http://expression.microslu.washington.edu/expression/kayee/cluster2003/yeunggb2003.html*

The dataset is applied in all the following experiments to cross check the clustering results.

### 5.1.2 Clustering algorithm

The k-Means algorithm described in Section 1.6.4 is used in all the experiments, and the "k" value, the number of clusters to be formed, is set according to the dataset. The "k" value is set as 4 in the dataset mentioned in Section 5.1.1.

### 5.1.3 Evaluation of results

As discussed in Section 4.8, average accuracy is used to describe the accuracy of results in each experiment. The standard deviation of accuracy is adopted to describe how stable the algorithm is. The lower the value of standard deviation is, the better performance the algorithm gain. 20 runs are performed in each experiment to get the average number and standard deviation.

## 5.2 Implementation

The program is implemented using the programming language Java, because of its several advantages: The object orientation of Java enable the clear structure of the program, while the Gene Ontology can be loaded as a graph object, and every item of the GO is treated as a GOElement object. The Java open source community provides Jena (*http://jena.sourceforge.net/index.html*), a Semantic Web framework, to support the operation of ontology. The platform-independence of Java language enable the program to be run in different Operating Systems.

### 5.2.1 Program Classes

The program consists of 8 classes: kMeans, ClusterAlgorithm, LOG, Group, Cluster, GOTerm, GOElement, and Element.

kMeans is a subclass of class ClusterAlgorithm. It implements the k-Means algorithm. The main() function of kMeans also perform the experiment task by calling the algorithm with different parameters repeatedly.

ClusterAlgorithm is the main class in this program. It provides functions to perform different clustering algorithms, including k-Means, BSAS and MBSAS. These functions includes:

- BuildDisMatrix(), builds distance matrix;
- Distance(element, cluster), gives the distance between element and cluster;
- NearestCluster(element), finds the nearest cluster for element;
- RecalculateCenterValue(), calculates center value of each clusters;
- PrintDisMatrix(), outputs the distance matrix for debugging;
- printCluster(), outputs all the cluster members.

This class also performs the manipulation of Gene Ontology and the similarity measure calculation.

LOG class prints debug information based on different demands: verbose, debug, info, notice, warn, error, critical, alert, and emergency. The traditional debug method is to print intermediate values of variable, and comment those printing command when the program is released. This class helps to output intermediate values according to different situation, and stop printing when the program is released.

Other classes are small classes used by the ClusterAlgorithm class.

### 5.2.2 Binding GO Information of Gene Products

The size of GO annotation file is 1.3 Giga bytes, so it will eat up the computer memory when it is fully loaded to find the annotation of genes. There are several public annotation libraries available, and Saccharomyces Genome Database (*http://www.yeastgenome.org*) is chosen in this study to query the gene annotation for the dataset. The query result is stored as a ".GoTerm" file. It is slow to query annotation information through Internet, but the annotation is relatively stable, and it can be stored locally to accelerate the loading next time.

The Gene Ontology is a graph with 111,672 terms. For the 205 genes in the dataset, each gene is annotated averagely by 5 GO terms on average, and each term has 15 ancestors. Totally 15023 terms are involved. But there are only 604 unique terms in the 15023 terms. It is a very small portion of the graph with 111,672 terms. So after the annotation terms and their ancestors are retrieved from the Gene Ontology, they can form a small graph with the size 1 percent of the size of Gene Ontology. This graph is saved as a ".GoPath" file. So next time we can reuse it to avoid loading the big Gene Ontology graph. Manipulation of the small graph, such as calculating the height of a term or locating the least common ancestor of two terms, is performed much faster than in the whole GO graph.

When a gene list is loaded at the first time, the GO information of genes is retrieved and saved as separate files using the same filename as the gene list but a different suffix names. By this means, the program will find the existence of the GO information files and load it without retrieving it from the Internet and GO graph again.

## 5.3 Clustering result of different similarity measurement

The measures listed in Section 4.7 are compared. The weights combining similarity measures with Euclidean distance are 0.5 and 0.5. The highest accuracy or lowest standard deviation of each column is highlighted.

| | Accuracy | Sd. Deviation |
|---|---|---|
| $SIM_1$ | 0.748 | 0.086 |
| $SIM_2$ | 0.671 | 0.102 |
| $SIM_3$ | 0.635 | 0.092 |
| $SIM_4$ | 0.675 | 0.072 |
| $SIM_5$ | 0.655 | 0.130 |
| $SIM_6$ | 0.647 | 0.118 |
| $SIM_7$ | 0.723 | 0.156 |
| Average | 0.679 | 0.108 |

**Table 1 Performance of Original Measures**

| | Accuracy | Sd. Deviation |
|---|---|---|
| $SIM_8$ | 0.688 | 0.083 |
| $SIM_9$ | 0.677 | 0.067 |
| $SIM_{10}$ | 0.634 | 0.097 |
| $SIM_{11}$ | 0.707 | 0.078 |
| $SIM_{12}$ | 0.668 | 0.091 |
| $SIM_{13}$ | 0.695 | 0.124 |
| $SIM_{14}$ | 0.689 | 0.122 |
| Average | 0.680 | 0.095 |

**Table 2 Performance of Combined Measures**

| | Accuracy | Sd. Deviation |
|---|---|---|
| $SIM_{15}$ | 0.705 | 0.091 |
| $SIM_{16}$ | 0.662 | 0.092 |
| $SIM_{17}$ | 0.750 | 0.079 |
| $SIM_{18}$ | 0.677 | 0.090 |
| $SIM_{19}$ | 0.651 | 0.058 |
| $SIM_{20}$ | 0.680 | 0.125 |
| $SIM_{21}$ | 0.629 | 0.066 |
| Average | 0.679 | 0.086 |

**Table 3 Measures based on Biological_Process Category of GO**

| | Accuracy | Sd. Deviation |
|---|---|---|
| $SIM_{22}$ | 0.669 | 0.062 |
| $SIM_{23}$ | 0.659 | 0.066 |
| $SIM_{24}$ | 0.780 | 0.082 |

| | | |
|---|---|---|
| $SIM_{25}$ | 0.711 | 0.064 |
| $SIM_{26}$ | 0.635 | 0.093 |
| $SIM_{27}$ | 0.688 | 0.094 |
| $SIM_{28}$ | 0.651 | 0.118 |
| Average | 0.685 | 0.083 |

**Table 4 Combined Measures based on Biological Process Category of GO**

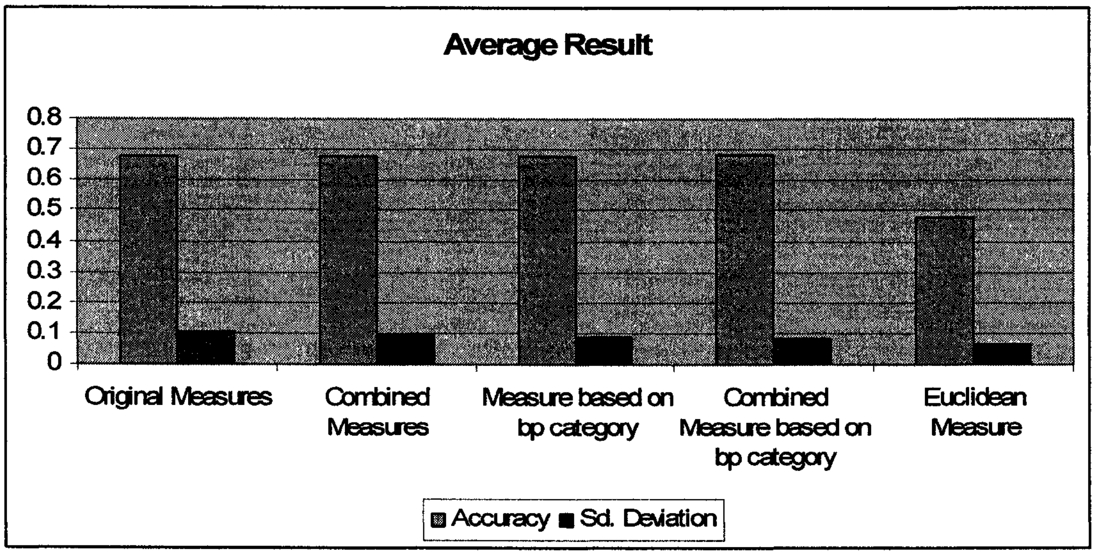
| | Accuracy | Sd. Deviation |
|---|---|---|
| Euclidean Measure | 0.478 | 0.065 |

**Table 5 Performance of Euclidean Measures**

Comparing the average result of each table:

| | Accuracy | Sd. Deviation |
|---|---|---|
| Original Measures | 0.679 | 0.108 |
| Combined Measures | 0.680 | 0.095 |
| Measure based on bp category | 0.679 | 0.086 |
| Combined Measure based on bp category | 0.685 | 0.083 |
| Euclidean Measure | 0.478 | 0.065 |

**Table 6 Average result of Table 1 to Table 5**



We can draw these conclusions:

- The accuracy of algorithm applying GO measures is much better than the accuracy of algorithm applying Euclidean measure. It is about 40% increasement. But the standard deviation of the accuracy is also increased by 40%.
- The performance of using the biological process category of the GO along is slightly better than using the whole GO, because the standard deviation of Table3 and Table 4 is 15% less than that of Table 1 and Table 2, when the accuracy remains the same.

## 5.4 Clustering result of different weight in combined measures

| | $w_1$=0.25 $w_2$=0.75 | | $w_1$=0.75 $w_2$=0.25 | |
|---|---|---|---|---|
| | Acc | St. Dev | Acc. | St. Dev |
| $SIM_8$ | 0.662 | 0.101 | 0.670 | 0.090 |
| $SIM_9$ | 0.623 | 0.065 | 0.700 | 0.117 |
| $SIM_{10}$ | 0.602 | 0.123 | 0.643 | 0.096 |
| $SIM_{11}$ | 0.586 | 0.097 | 0.636 | 0.104 |
| $SIM_{12}$ | 0.594 | 0.099 | 0.713 | 0.114 |
| $SIM_{13}$ | 0.706 | 0.099 | 0.722 | 0.110 |
| $SIM_{14}$ | 0.623 | 0.104 | 0.694 | 0.104 |

**Table 7 Different weight in Combing Measures**

| | $w_1$=0.25 $w_2$=0.75 | | $w_1$=0.75 $w_2$=0.25 | |
|---|---|---|---|---|
| | Acc | St. Dev | Acc. | St. Dev |
| $SIM_{22}$ | 0.630 | 0.094 | 0.672 | 0.078 |
| $SIM_{23}$ | 0.629 | 0.068 | 0.654 | 0.113 |
| $SIM_{24}$ | 0.630 | 0.096 | 0.745 | 0.094 |
| $SIM_{25}$ | 0.699 | 0.055 | 0.668 | 0.067 |
| $SIM_{26}$ | 0.590 | 0.095 | 0.670 | 0.070 |
| $SIM_{27}$ | 0.644 | 0.094 | 0.728 | 0.108 |
| $SIM_{28}$ | 0.584 | 0.070 | 0.689 | 0.091 |

**Table 8 Different Weights in Combing Derivative Measures**

The data in Table 1 and Table 3 is the result of using Ontology measures only; it can be regarded as the result of combined measure,

$$SIM_{1\text{-}7}(X, Y) = w_1 SIM_{1\text{-}7}(X, Y) + w_2 \frac{1}{1 + \mathit{EuclideanDis}\tan \mathit{ce}(X, Y)}$$

$$SIM_{15\text{-}21}(X, Y) = w_1 SIM_{15\text{-}21}(X, Y) + w_2 \frac{1}{1 + \mathit{EuclideanDis}\tan \mathit{ce}(X, Y)}$$

While $w_1$=1 and $w_2$=0

Data in Table 5 is the result of using Euclidean measure. It can also be regarded as the result of combined measure,

$$SIM_{Eu}(X, Y) = w_1 SIM_{1\text{-}28}(X, Y) + w_2 \frac{1}{1 + \mathit{EuclideanDis}\tan \mathit{ce}(X, Y)}$$

while $w_1$=0 and $w_2$=1.

Data in the previous tables is reorganized following the discussion above:

| | w1=0<br>w2=1 | w1=0.25<br>w2=0.75 | w1=0.5<br>w2=0.5 | w1=0.75<br>w2=0.25 | w1=1<br>w2=0 | Average |
|---|---|---|---|---|---|---|
| $SIM_8$ | 0.478 | 0.662 | 0.688 | 0.67 | 0.748 | 0.6492 |
| $SIM_9$ | 0.478 | 0.623 | 0.677 | 0.7 | 0.671 | 0.6298 |
| $SIM_{10}$ | 0.478 | 0.602 | 0.634 | 0.643 | 0.635 | 0.5984 |
| $SIM_{11}$ | 0.478 | 0.586 | 0.707 | 0.636 | 0.675 | 0.6164 |
| $SIM_{12}$ | 0.478 | 0.594 | 0.668 | 0.713 | 0.655 | 0.6216 |
| $SIM_{13}$ | 0.478 | 0.706 | 0.695 | 0.722 | 0.647 | 0.6496 |
| $SIM_{14}$ | 0.478 | 0.623 | 0.689 | 0.694 | 0.723 | 0.6414 |
| $SIM_{22}$ | 0.478 | 0.63 | 0.669 | 0.672 | 0.705 | 0.6308 |
| $SIM_{23}$ | 0.478 | 0.629 | 0.659 | 0.654 | 0.662 | 0.6164 |
| $SIM_{24}$ | 0.478 | 0.63 | 0.78 | 0.745 | 0.75 | 0.6766 |
| $SIM_{25}$ | 0.478 | 0.699 | 0.711 | 0.668 | 0.677 | 0.6466 |

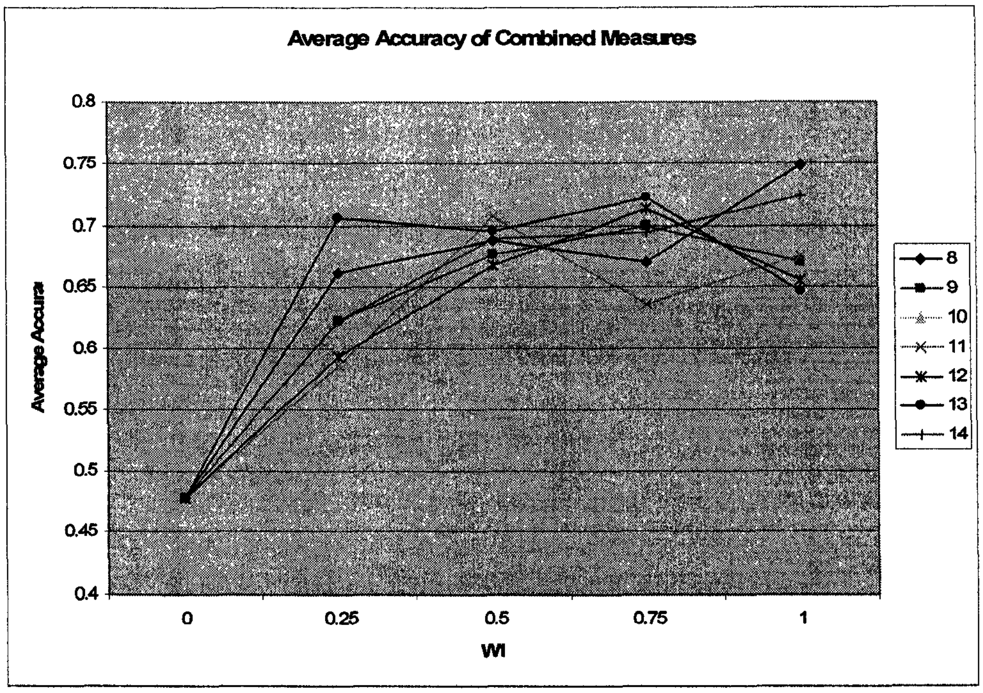
| | | | | | | |
|---|---|---|---|---|---|---|
| SIM$_{26}$ | 0.478 | 0.59 | 0.635 | 0.67 | 0.651 | 0.6048 |
| SIM$_{27}$ | 0.478 | 0.644 | 0.688 | 0.728 | 0.68 | 0.6436 |
| SIM$_{28}$ | 0.478 | 0.584 | 0.651 | 0.689 | 0.629 | 0.6062 |
| **Average** | **0.478** | **0.629** | **0.682** | **0.686** | **0.679** | **0.631** |

**Table 9 Average Accuracy of Combined Measure with Different Weights**

| | wl=0<br>w2=1 | wl=0.25<br>w2=0.75 | wl=0.5<br>w2=0.5 | wl=0.75<br>w2=0.25 | wl=1<br>w2=0 | Average |
|---|---|---|---|---|---|---|
| SIM$_{8}$ | 0.065 | 0.101 | 0.083 | 0.09 | 0.086 | 0.085 |
| SIM$_{9}$ | 0.065 | 0.065 | 0.067 | 0.117 | 0.102 | 0.0832 |
| SIM$_{10}$ | 0.065 | 0.123 | 0.097 | 0.096 | 0.092 | 0.0946 |
| SIM$_{11}$ | 0.065 | 0.097 | 0.078 | 0.104 | 0.072 | 0.0832 |
| SIM$_{12}$ | 0.065 | 0.099 | 0.091 | 0.114 | 0.13 | 0.0998 |
| SIM$_{13}$ | 0.065 | 0.099 | 0.124 | 0.11 | 0.118 | 0.1032 |
| SIM$_{14}$ | 0.065 | 0.104 | 0.122 | 0.104 | 0.156 | 0.1102 |
| SIM$_{22}$ | 0.065 | 0.094 | 0.062 | 0.078 | 0.091 | 0.078 |
| SIM$_{23}$ | 0.065 | 0.068 | 0.066 | 0.113 | 0.092 | 0.0808 |
| SIM$_{24}$ | 0.065 | 0.096 | 0.082 | 0.094 | 0.079 | 0.0832 |
| SIM$_{25}$ | 0.065 | 0.055 | 0.064 | 0.067 | 0.09 | 0.0682 |
| SIM$_{26}$ | 0.065 | 0.095 | 0.093 | 0.07 | 0.058 | 0.0762 |
| SIM$_{27}$ | 0.065 | 0.094 | 0.094 | 0.108 | 0.125 | 0.0972 |
| SIM$_{28}$ | 0.065 | 0.070 | 0.118 | 0.091 | 0.066 | 0.082 |
| **Average** | **0.065** | **0.090** | **0.089** | **0.097** | **0.097** | **0.087** |

**Table 10 Standard Deviation of Accuracy Combined Measures with Different Weights**

Table 9 is presented in Figure 7:


Average Accuracy of Combined Measures
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
Average Accura
0
0.25
0.5
0.75
1
WI
8
9
10
11
12
13
14

and Figure 8:

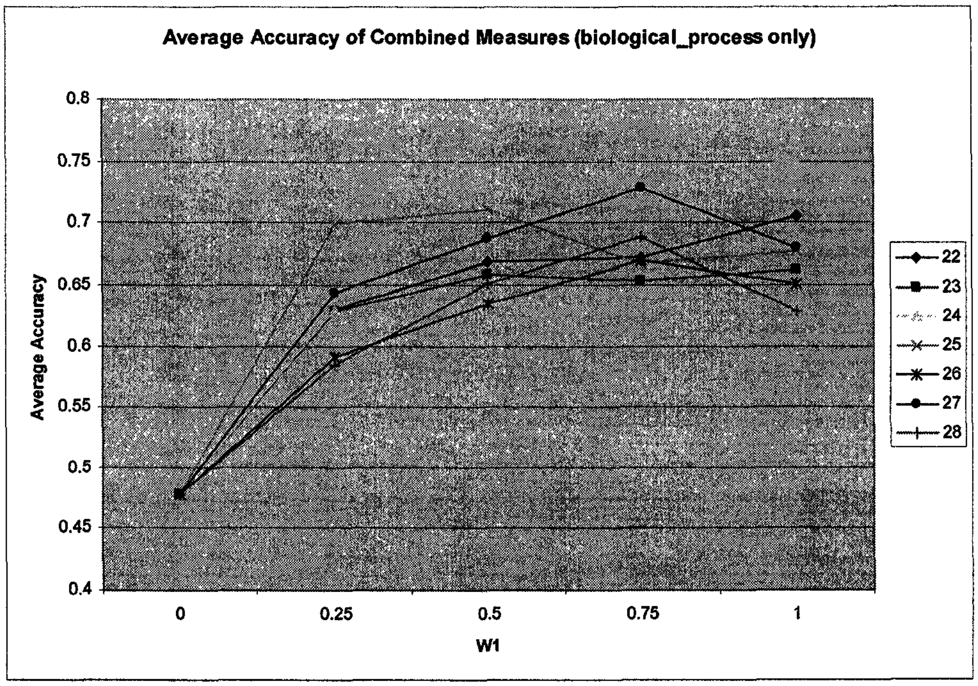
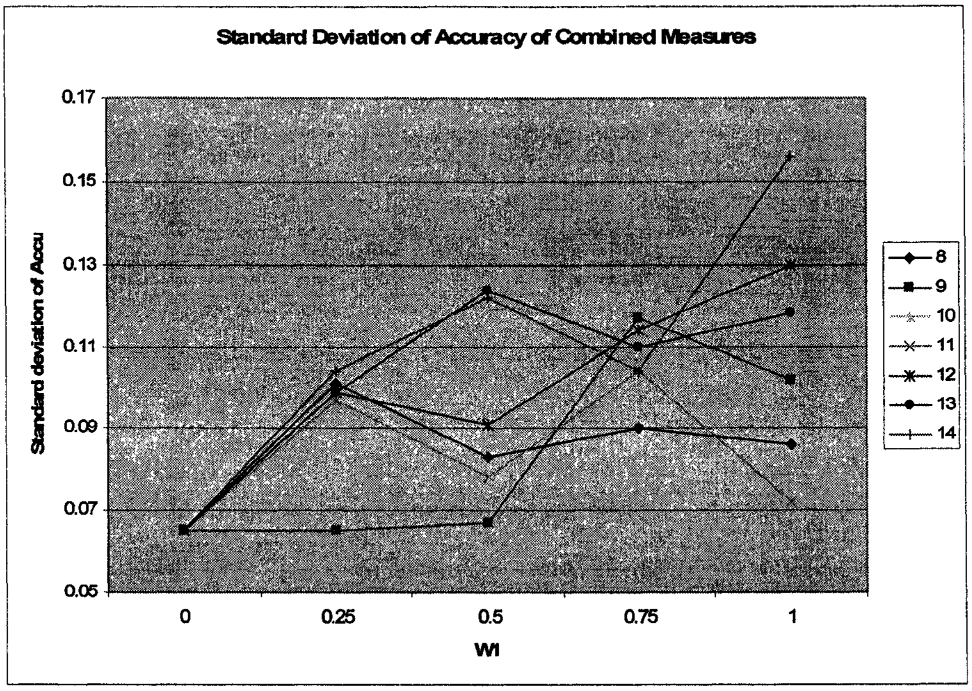
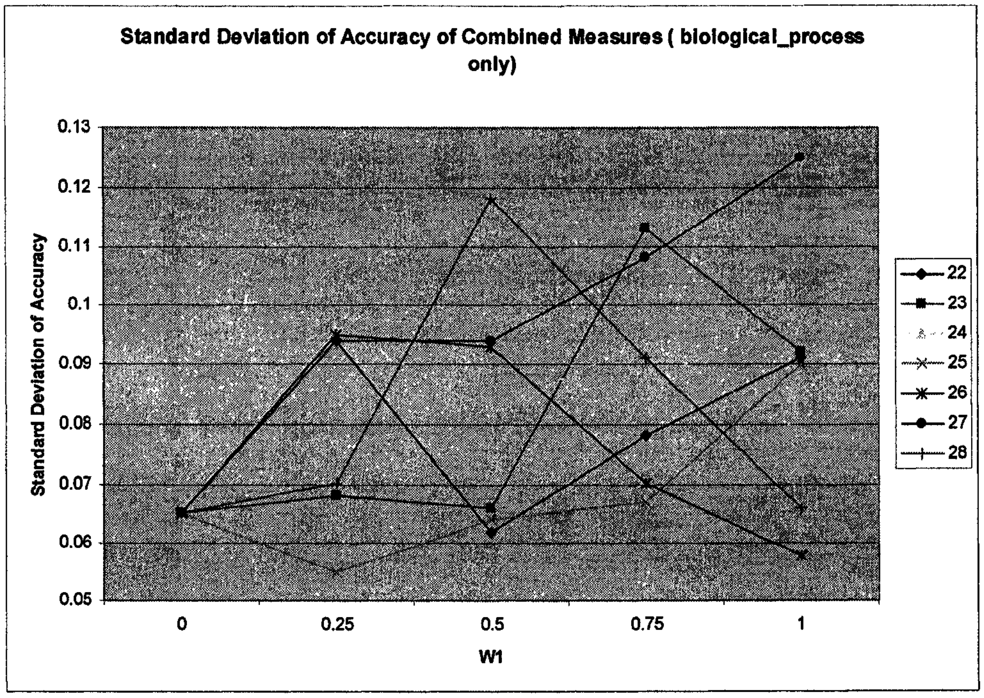

Average Accuracy of Combined Measures (biological_process only)
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
Average Accuracy
0
0.25
0.5
0.75
1
W1
22
23
24
25
26
27
28

Table 10 is presented in Figure 9:



and Figure 10:



CHAPTER VI

# CONCLUSION AND FUTURE WORK

## 6.1 Conclusion

In this thesis project, six currently-used methods for similarity measurement over ontology are studied, and then the On-The-Fly probability and the similarity between two sets of terms in ontology are initially defined in this project. The measures over ontology are combined with Euclidean distance to contain both the gene expression data and biological knowledge. In order to check which similarity measure performs better, experiment is conducted. The results of applying different measures into clustering Microarray expression data are compared with the known gene information and the following points came to conclusion:

- The accuracy of algorithm applying GO measures is much better than the accuracy of algorithm applying Euclidean measure. It is about 40% increasement. But the standard deviation of the accuracy is also increased by 40%. So the Gene Ontology based measures can greatly increase the accuracy.
- The performance of using the biological process category of the GO along is slightly better than using the whole GO, because the standard deviation of Table3 and Table 4 is 15% less than that of Table 1 and Table 2, when the accuracy remains the same. Using the biological process category can

also increase the efficiency of algorithm, because the size of this category is only one third of the size of whole ontology.

## 6.2 Contribution

In this thesis project, the On-The-Fly probability is newly defined as the frequency of a term occurring in the current problem space. It is used to calculate the importance of a term.

Secondly, based on the characteristics of Gene Ontology, the similarity between a term and a set of terms is defined, and which is further adapted to the similarity between two sets in order to better describe the similarity over Gene Ontology. The measures over ontology are combined with Euclidean distance so that they contain both gene expression data and biological knowledge to more fully express the gene information. The combined measure is applied in this empirical study to evaluate the performance of it.

## 6.2 Future Work

In this project, only accuracy of clustering result is used in comparing the performance of similarity measures. Computational time complexity and space complexity of each measure can be studied in the future.

The On-The-Fly probability can be compared with other probabilities based on different context. For example, the frequency of a term being used to annotate proteins in the SWISS-PROT database can form the probability of the term. Also, the number of gene products being annotated by a single term can also define the probability.

# APPENDICES

## APPENDIX A

### Yeasts Dataset

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| YHR010W | 0 | YPL143W | 0 | YOR167C | 0 | YJR050W | 2 | YMR213W | 2 |
| YOR182C | 0 | YLR333C | 0 | YGR240C | 1 | YJR052W | 2 | YDR073W | 2 |
| YHR021C | 0 | YLR325C | 0 | YGR254W | 1 | YJL115W | 2 | YBR279W | 2 |
| YBL072C | 0 | YLR264W | 0 | YGL253W | 1 | YFL001W | 2 | YMR270C | 2 |
| YBL087C | 0 | YLR185W | 0 | YGR192C | 1 | YHL009C | 2 | YMR277W | 2 |
| YHL033C | 0 | YER074W | 0 | YCR012W | 1 | YER162C | 2 | YBR253W | 2 |
| YNL178W | 0 | YER102W | 0 | YOR347C | 1 | YKL011C | 2 | YBR236C | 2 |
| YBR181C | 0 | YLR061W | 0 | YMR125W | 1 | YKL012W | 2 | YBR215W | 2 |
| YNL301C | 0 | YER117W | 0 | YJR009C | 1 | YKL015W | 2 | YBR193C | 2 |
| YBR189W | 0 | YER131W | 0 | YMR205C | 1 | YGR005C | 2 | YHR006W | 2 |
| YBR191W | 0 | YKR057W | 0 | YAL038W | 1 | YKL113C | 2 | YNL230C | 2 |
| YOR096W | 0 | YHL001W | 0 | YKL152C | 1 | YKL125W | 2 | YBL021C | 2 |
| YOL120C | 0 | YJR145C | 0 | YPL075W | 1 | YKL149C | 2 | YNL314W | 2 |
| YNL162W | 0 | YJR123W | 0 | YJL052W | 1 | YJR022W | 2 | YBL025W | 2 |
| YNL096C | 0 | YPL090C | 0 | YHR174W | 1 | YKR008W | 2 | YOL135C | 2 |
| YNL069C | 0 | YOR293W | 0 | YKL060C | 1 | YKR025W | 2 | YBR055C | 2 |
| YNL067W | 0 | YGL031C | 0 | YHR178W | 2 | YER159C | 2 | YOR148C | 2 |
| YMR242C | 0 | YOR312C | 0 | YGR075C | 2 | YLL036C | 2 | YHR058C | 2 |
| YKL006W | 0 | YGL076C | 0 | YGR091W | 2 | YER112W | 2 | YIL021W | 2 |
| YDL061C | 0 | YJL190C | 0 | YGR104C | 2 | YLR116W | 2 | YOR194C | 2 |
| YDL075W | 0 | YGL103W | 0 | YGR074W | 2 | YLR117C | 2 | YPR182W | 2 |
| YDL082W | 0 | YGL123W | 0 | YGR186W | 2 | YER032W | 2 | YPR168W | 2 |
| YDL083C | 0 | YGL135W | 0 | YGR047C | 2 | YER029C | 2 | YPR186C | 2 |
| YDL136W | 0 | YGL147C | 0 | YBR188C | 2 | YLR298C | 2 | YOR319W | 2 |
| YDL191W | 0 | YGL189C | 0 | YDL044C | 2 | YLR316C | 2 | YPR107C | 2 |
| YPR132W | 0 | YJL189W | 0 | YBR123C | 2 | YLR321C | 2 | YPR101W | 2 |
| YDR064W | 0 | YJL177W | 0 | YIR018W | 2 | YER022W | 2 | YPL213W | 2 |
| YMR230W | 0 | YHR141C | 0 | YGR200C | 2 | YEL056W | 2 | YGR289C | 3 |
| YMR193W | 0 | YJL136C | 0 | YGR056W | 2 | YDR473C | 2 | YHR092C | 3 |
| YMR143W | 0 | YIL133C | 0 | YJL127C | 2 | YDR397C | 2 | YHR094C | 3 |
| YMR142C | 0 | YGR027C | 0 | YDL030W | 2 | YML114C | 2 | YIL170W | 3 |
| YOR234C | 0 | YGR034W | 0 | YJL140W | 2 | YMR005W | 2 | YHR096C | 3 |
| YDR341C | 0 | YIL069C | 0 | YJL176C | 2 | YDR308C | 2 | YJL219W | 3 |
| YPR043W | 0 | YIL052C | 0 | YGL244W | 2 | YMR106C | 2 | YDR343C | 3 |
| YPL220W | 0 | YIL018W | 0 | YGL243W | 2 | YJR093C | 2 | YDR342C | 3 |
| YPL198W | 0 | YHR203C | 0 | YGL090W | 2 | YMR137C | 2 | YJL214W | 3 |
| YML026C | 0 | YPL079W | 0 | YJL203W | 2 | YDR243C | 2 | YDR345C | 3 |
| YDR450W | 0 | YOR369C | 0 | YGL070C | 2 | YDR240C | 2 | YDL194W | 3 |
| YDR471W | 0 | YGR118W | 0 | YFR037C | 2 | YMR182C | 2 | YFL011W | 3 |
| YLR344W | 0 | YGR148C | 0 | YGR013W | 2 | YDR088C | 2 | YMR011W | 3 |
| YDR500C | 0 | YPL081W | 0 | YIR015W | 2 | YER169W | 2 | YDR536W | 3 |

# REFERENCES

1. [Al 2004] F. Al-Shahrour, R. Diaz-Uriarte, and J. Dopazo, "FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes," Bioinformatics, Vol 20(4), pp. 578-580. 2004
2. [Bar-Joseph 2001] Z. Bar-Joseph, D.K. Gifford and T.S. Jaakkola, "Fast Optimal Leaf Ordering For Hierarchical Clustering," Bioinformatics, Vol. 17, 2001
3. [Bellaachia 2002] A. Bellaachia and D. Portnoy, "E-CAST: A Data Mining Algorithm For Gene Expression Data," 2nd Workshop on Data Mining in Bioinformatics, 2002
4. [Ben-Dor 1999] A. Ben-Dor, R. Shamir and Z. Yakhini, "Clustering gene expression patterns," Journal of Computational Biology, Vol 6, pp. 281-297, 1999
5. [Bolshakova 2006] N. Bolshakova, A. Zamolotskikh and P. Cunningham, "Comparison of the Data-based and Gene Ontology-based Approaches to Cluster Validation Methods for Gene Microarrays," Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems, 2006
6. [Bradley 1998] P.S. Bradley and U.M. Fayyad, "Refining Initial Point for K-Mean Clustering," Proc. 15th International Conf. Machine Learning, 1998
7. [Cheng 1998] T.W. Cheng, D.B. Goldgof and L.O. Hall, "Fast Fuzzy Clustering," Fuzzy Sets and Systems, Vol. 93, pp. 49-56, 1998
8. [Cho 1998] R.J. Cho, M.J. Campbell, E. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T.G. Wolfsberg, A.E. Gabrielian, D. Landsman, D.J. Lockhart and R.W. Davis, "A genome-wide transcriptional analysis of the mitotic cell cycle," Molecular Cell, Vol 12, pp. 65-73, 1993
9. [Coe] B. Coe and C. Antler, "Spot your Genes-An Overview of the MicroArray," [Online document], Available at http://bioteach.ubc.ca/MolecularBiology/microarray
10. [Couto 2005] F. Couto, M. Silva and P. Coutinho, "Semantic Similarity over the Gene Ontology: Family Correlation and Selecting Disjunctive Ancestors," ACM CIKM - Conference in Information and Knowledge Management, 2005
11. [Datta 2006] S. Datta and S. Datta, "Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes," BMC Bioinformatics, Vol 7:397, 2006
12. [Dhanasekaran 2001] S.M. Dhanasekaran, T.R. Barrete, D. Ghosh., R. Shah, S. Varambally, K. Kurachi, K. Pienta, M. Rubin and A. Chinnaiyan, "Delineation of Prognostic Biomarkers in prostate cancer," Nature, 2001
13. [Doniger 2003] S.W. Doniger, N. Salomonis, K.D. Dahlquist, K. Varnizan, S.C. Lawlor, and B.R. Conklin, "MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data," Genome Biology, Vol 4:R7, 2003
14. [Eisen 1998] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein, "Cluster Analysis and Display of Genome-Wide Expression Patterns," Proc. Natl. Acad. Sci, pp. 14863-14868, 1998
15. [Eppstein 1998] D. Eppstein, "Fast Hierarchical Clustering and Other Applications of Dynamic Closest pairs," Symposium on Discrete Algorithms, 1998

16. [Fayyad 1998] U. Fayyad, C. Reina and P.S. Bradley, "Initialization of Iterative Refinement Clustering Algorithms," ICML, pp. 194-198, 1998
17. [Futschik 2002] M.E. Futschik and N.K. Kasabov, "Fuzzy Clustering of Gene Expression Data," Proceedings of World Congress of Computational Intelligence, 2002
18. [Gasch 2002] A.P. Gasch and M.B. Eisen, "Exploring the conditional co-regulation of yeast gene expression through fuzzy k-mean clustering," Genome Biology Vol 3, 2002
19. [GO] The Gene Ontology Consortium, "the Gene Ontology," [Online document], Available at http://www.geneontology.org
20. [GO 2000] The Gene Ontology Consortium, "Gene Ontology: tool for the unification of biology," Nature Genet, Vol 25, pp. 25-29, 2000
21. [GO 2006] The Gene Ontology Consortium, "The Gene Ontology (GO) project in 2006," Nucleic Acids Research, Vol 34, 2006
22. [Harrington 2001] C.A. Harrington, C. Rosenow and J. Retief, "Monitoring gene expression using DNA microarrays," Current Opinion in Microbiology, Vol 3, pp. 285-291, 2001
23. [Ideker01] T. Ideker, V. Thorsson, J.A. Ranish, R. Christmas, J. Buhler, J.K. Eng, R.E. Bumgarner, D.R. Goodlett, R. Aebersold and L. Hood, "Integrated genomic and proteomic analyses of a systemically perturbed metabolic network," Science, pp. 929-934, 2001
24. [Jiang 1998] J.J. Jiang and D.W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," Proceedings of International Conference on Research in Computational Linguistics, 1998
25. [Jiang 2004] D. Jiang and A. Zhang, "Cluster Analysis for Gene Expression Data: a Survey," IEEE Trans Knowledge and Data Engineering, Vol 16(11), 2004
26. [Kennedy 2003] P.J. Kennedy and S.J. Simoff, "CONGO: Clustering on the Gene Ontology," Proceeding of 2nd Australasian Data Mining Workshop, 2003
27. [Khatri 2005] P. Khatri and S. Draghici, "Ontological analysis of gene expression data: current tools, limitations, and open problems," Bioinformatics, Vol 21(18), 2005
28. [Kohonen 1990] T. Kohonen, "The Self-Organizing Map," Proc IEEE, Vol 78(9), pp. 1464-1480, 1990
29. [Lam 2007] B.S.Y. Lam and H. Yan, "Assessment of microarray data clustering results based on a new geometrical index for cluster validity," Soft Comput, Vol 11, pp. 341-348, 2007
30. [Lee] S.K. Lee, "A Self-Organizing Map for Finding the Optimal Gene Order in Displaying Microarray Data"
31. [Lewis 2004] S.E. Lewis, "Gene Ontology: looking backwards and forwards," Genome Biology, Vol 6, pp. 103, 2004
32. [Lin 1998] D. Lin, An information-Theoretic Definition of Similarity, Proc. 15th International Conf. on Machine Learning, pp. 296-304, 1998
33. [Lord 2003] P.W. Lord, R.D. Stevens, A. Brass and C.A. Goble, "Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation," Bioinformatics. Vol 19(10), pp. 1275-1283, 2003

34. [MacQueen 1967] J. MacQueen, "Some methods for classification and analysis of multivariate observations," Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol 1, pp. 281-297, 1967
35. [NCBI] NCBI, "Just the Facts: A Basic Introduction to the Science Underlying NCBI Resources," [Online document], Available at http://www.ncbi.nlm.nih.gov/About/primer/microarrays.html
36. [Nurmi 2004] P. Nurmi, "Mixture Models," Helsinki Institute for Information Technology, 2004
37. [OBO] OBO webmaster, "The Open Biomedical Ontologies," [Online document], Available at http://obo.sourceforge.net
38. [Pavlidis 2002] P. Pavlidis, D.L. Lewis, and W.S. Noble, "Exploring gene expression data with class scores" Proceedings of the Pacific Symposium on Biocomputing, pp. 474-485, 2002
39. [Perou 2000] C.M. Perou, T. Sorlie, M.B. Eisen, M. Rijn et al., "Molecular Portraits of Human Breast Tumor," Nature, 2000.
40. [Resnik 1995] P. Resnik, "Using information content on evaluate semantic similarity in a taxonomy," Proceedings of the 14th International Joint Conference on Artificial Intelligence, pp. 448-453, 1995
41. [Robinson 2002] M.D. Robinson, J. Grigull, N. Mohammad and T.R. Hughes, "FunSpec: a web-based cluster interpreter for yeast," BMC Bioinformatics. Vol 3(1), pp. 35, 2002
42. [Rocke 2003] D.M. Rocke, "The emerging role of Mathematics, Statistics and Computation in Drug Discovery via High-throughput Assays," Business Briefing, 2003
43. [Schena 1995] M. Schena, D. Shalon, R.W. Davis and P.O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA Microarray", Science, Vol 270, pp. 467-470, 1995
44. [Shah 2004] N.H. Shah and N.V. Fedoroff, "CLENCH: a program for calculating Cluster ENRiCHment using the Gene Ontology," Bioinformatics, Vol 20(j), pp. 1196-1197, 2004
45. [Singh 2002] M. Singh, "Clustering Affinity Search Technique", University of Colorado, 2002
46. [Tamayo98] P. Tamayo et al., "Interpreting Patterns of Gene Expression with Self-Organizing Maps: Methods and Application to Hematopoietic Differentiation", Genetics, 1998
47. [Tavazoie 1999] S. Tavazoie, J. Hughes, M. Campbell, R. Cho and G. Church, "Systematic determination of genetic network architecture," Nat Genet, Vol 22, pp. 218-285, 1999
48. [Veenman 2002] C.J Veenman, M.J.T. Reinders and E. Backer, "A Maximum Variance Cluster Algorithm," IEEE Trans, Vol 24, 2002
49. [Wall 2001] M.E. Wall, P.A. Dyck and T.S. Brettin, "SVDMAN-Singular Value Decomposition Analysis of Microarray Data," Bioinformatics, Vol 17(6), pp. 566-568, 2001
50. [Wiki] Anonymous, "Bioinformatics", [Online document], available at http://en.wikipedia.org/wiki/Bioinformatics
51. [Wolkenhauer 2002] O. Wolkenhauer, "Cluster Analysis", UMIST, 2002

52. [Wu 1994] Z. Wu and M. Palmer, "Verb Semantics and Lexical Selection," 32nd. Annual Meeting of the Association for Computational Linguistics, pp. 133-138, 1994
53. [Yeung 2003] S.K. Yeung, M. Medvedovic and R.E. Bumgarner, "Clustering gene expression data with repeated measurements," Genome Biology, Vol 4(5), 2003
54. [Young 2005] A. Young, N. Whitehouse, J. Cho and C. Shaw, "OntologyTraverser: an R package for GO analysis," Bioinformatics, Vol 21(2) pp. 275-276, 2005
55. [Zehetner 2003] G. Zehetner, "OntoBlast function: From sequence similarities directly to potential functional annotations by ontology terms," Nucleic Acids Res, Vol 31(13), pp. 3799-3803, 2003
56. [Zhang 2004] B. Zhang, D. Schmoyer, S. Kirov and J. Snoddy, "GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies," BMC Bioinformatics, Vol 5(1), pp. 16, 2004
57. [Zhong 2004] S. Zhong, L. Tian, C. Li, F.K. Storch and W.H. Wong, "Comparative Analysis of Gene Sets in the Gene Ontology Space under the Multiple Hypothesis Testing Framework," Proc IEEE Comp Systems Bioinformatics, pp. 425-435, 2004
58. [Zhou 2004] M.Zhou and Y. Cui, "GeneInfoViz: constructing and visualizing gene relation networks," In Silico Boil, Vol 4(3), pp. 323-333, 2004

# VITA AUCTORIS

| | |
|---|---|
| NAME: | Tianbing Lin |
| PLACE OF BIRTH: | Guangxi, China |
| YEAR OF BIRTH: | 1975 |
| EDUCATION: | B.E. in Mechanical Engineering,<br>Northern Jiaotong University, Beijing, China<br>1992-1996 |
| | M.S. in Computer Science,<br>University of Windsor, Ontario<br>2003-2006 |