

1-1-2009

A Method for Integrating Heterogeneous Datasets based on GO Term Similarity

Chamali Lankara Thanthiriwatte

Follow this and additional works at: <https://scholarsjunction.msstate.edu/td>

Recommended Citation

Thanthiriwatte, Chamali Lankara, "A Method for Integrating Heterogeneous Datasets based on GO Term Similarity" (2009). *Theses and Dissertations*. 176.
<https://scholarsjunction.msstate.edu/td/176>

This Graduate Thesis - Open Access is brought to you for free and open access by the Theses and Dissertations at Scholars Junction. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of Scholars Junction. For more information, please contact scholcomm@msstate.libanswers.com.

A METHOD FOR INTEGRATING HETEROGENEOUS DATASETS
BASED ON GO TERM SIMILARITY

By

Chamali Lankara Thanthiriwatte

A Thesis
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Master of Science
in Computer Science
in the Department of Computer Science and Engineering

Mississippi State, Mississippi

December 2009

Copyright by

Chamali Lankara Thanthiriwatte

2009

A METHOD FOR INTEGRATING HETEROGENEOUS DATASETS
BASED ON GO TERM SIMILARITY

By

Chamali Lankara Thanthiriwatte

Approved:

Susan M. Bridges
Professor of Computer Science
and Engineering
Department of Computer Science
and Engineering
(Major Professor)

W. Paul Williams
Supervisory Research Genetist,
USDA-ARS
Adjunct Professor of Plant
and Soil Sciences
Department of Plant and
Soil Sciences
(Committee Member)

Fiona M. McCarthy
Assistant Professor of
Basic Sciences
Department of Basic Sciences
(Committee Member)

Edward B. Allen
Associate Professor of Computer
Science and Engineering,
and Graduate Coordinator
Department of Computer Science
and Engineering
(Committee Member)

Sarah A. Rajala
Dean
of the Bagley College of Engineering

Name: Chamali Lankara Thanthiriwatte

Date of Degree: December 11, 2009

Institution: Mississippi State University

Major Field: Computer Science

Major Professor: Dr. Susan M. Bridges

Title of Study: A METHOD FOR INTEGRATING HETEROGENEOUS DATASETS
BASED ON GO TERM SIMILARITY

Pages in Study: 92

Candidate for Degree of Master of Science

This thesis presents a method for integrating heterogeneous gene/protein datasets at the functional level based on Gene Ontology term similarity.

Often biologists want to integrate heterogeneous data sets obtain from different biological samples. A major challenge in this process is how to link the heterogeneous datasets. Currently, the most common approach is to link them through common reference database identifiers which tend to result in small number of matching identifiers. This is due to lack of standard accession schemes. Due to this problem, biologists may not recognize the underlying biological phenomena revealed by a combination of the data but by each data set individually.

We discuss an approach for integrating heterogeneous datasets by computing the similarity among them based on the similarity of their GO annotations. Then we group the genes and/or proteins with similar annotations by applying a hierarchical clustering al-

gorithm. The results demonstrate a more comprehensive understanding of the biological processes involved.

Key words: Semantic Similarity, Similarity Matrix, Gene Ontology, Hierarchical Clustering, Functional Annotations, Gene Expression, Protein Expression, Proteomics, Transcriptomics

DEDICATION

To my beloved parents and husband who always provide me with a joyful surrounding
fulled with love and care.

ACKNOWLEDGMENTS

It is my pleasure to acknowledge many great individuals who have contributed to the success of this thesis.

First of all, I would like to thank my advisor Dr. Susan M. Bridges for her valuable guidance, support and encouragement through out the years of graduate school. She is an ideal advisor, provided ample freedom and flexibility to pursue my interests at my own pace. I am grateful for her constructive scientific input that helped me grow as a researcher and without which this work would not have been possible. Once again, my whole-hearted thanks go to her.

My very special thanks go to Dr. Paul Williams for generously providing me with financial support through out all the years in graduate school. Without his kind assistance, I would not have had a chance to pursue my dream of higher studies.

I would also like to extend my gratitude to my biology lecturers, Dr. Fiona McCarthy and Dr. Bindu Nanduri, for always patiently explaining the underlying biological phenomena of studies . With out their sincere support, bioinformatics research would have been more challenging. My special thanks also goes to our biology collaborators: Dr. Marilyn Warburton, a Research Geneticist with the USDA Corn Host Plant Resistance Laboratory, Dr. Rowena Kelley, a Postdoctoral Associate in the Department of Biochemistry and Molecular Biology, Seval Ozkan, a Research Associate in the Department of Plant and

Soil Sciences and the AgBase Biocurator for maize, and Dr. Leigh Hawkins, a Plant Geneticist with the USDA Corn Host Plant Resistance laboratory, and Dr. Zhaohua Peng, an Associate Professor in the Department of Biochemistry and Molecular Biology, who were generous with their expertise and precious time to analyze the results.

Words fail me to express my appreciation of beloved husband Sahan whose dedication, love and persistent confidence in me, which took the load off my shoulder. I am very fortunate to share my life with such an understanding partner who is always there for me. His unwavering support and steadfast belief were crucial for me to realize my dreams.

My very special thanks goes to my parents whom I owe everything, I am today my beloved mother Asoka and beloved late father Piyasena. Their consistent guidance, unwavering faith and confidence in my abilities are what have shaped me to be the person I am today. Many thanks go to my siblings Champika and Chaminda for their love and support. I cannot stop myself from thanking to my baby, Isitha for making me very happy all the times, even during the difficult periods. I owe him so much for his beautiful smile.

My sincere thanks also flow to our new faculty member, Dr. Andy Perkins for his support to explore graph theory and clustering algorithms.

I would also like to express my deep appreciation for the faculty and staff in Department of Computer Science and Engineering for their consistent guidance, help and support. Especially I express my sincere gratitude to Dr. Edward Allen, the graduate coordinator of the department for doing a very important service.

I would like to thank the staff in Mississippi State University Libraries. Their kindness and assistance will always be remembered.

The colleagues I have met while in graduate school have become my closest and dearest friends and counselors, and to all of you I give my love and thanks.

TABLE OF CONTENTS

| | |
|--|------|
| DEDICATION | ii |
| ACKNOWLEDGMENTS | iii |
| LIST OF TABLES | viii |
| LIST OF FIGURES | ix |
| LIST OF SYMBOLS | x |
| CHAPTER | |
| 1. INTRODUCTION | 1 |
| 2. LITERATURE REVIEW | 9 |
| 2.1 Transcriptome and proteome technology | 9 |
| 2.1.1 Linking heterogeneous datasets through identifiers | 13 |
| 2.1.2 Correlating protein and microarray data | 15 |
| 2.2 Gene Ontology (GO) | 17 |
| 2.3 Functional level mappings | 19 |
| 2.4 Computing the similarity of genes based on GO annotation | 20 |
| 3. APPROACH | 25 |
| 3.1 Background and hypothesis | 25 |
| 3.2 Steps in the approach | 27 |
| 4. RESULTS AND EVALUATION | 42 |
| 4.1 Experiments | 42 |
| 4.1.1 <i>Arabidopsis</i> Experiment | 43 |
| 4.1.2 Maize Experiment | 44 |
| 4.2 Results and Analysis | 46 |

| | | |
|-------|---|----|
| 4.2.1 | <i>Arabidopsis</i> Results Analysis | 46 |
| 4.2.2 | Corn Results Analysis | 73 |
| 5. | CONCLUSION AND FUTURE WORK | 84 |
| 5.1 | Summary of Results | 84 |
| 5.2 | Future Research | 86 |
| | REFERENCES | 88 |

LIST OF TABLES

| | | |
|-----|---|----|
| 3.1 | Gene similarity matrix for <i>Arabidopsis</i> data set 1 | 30 |
| 3.2 | Gene similarity matrix for <i>Arabidopsis</i> data set 2 | 31 |
| 3.3 | Gene similarity matrix for combined datasets | 32 |
| 3.4 | GO annotations for the two clusters in <i>Arabidopsis</i> dataset 1 | 38 |
| 3.5 | GO annotations for the three clusters in <i>Arabidopsis</i> dataset 2 | 39 |
| 3.6 | GO annotations for the two clusters in <i>Arabidopsis</i> combined datasets | 40 |
| 4.1 | Clusters for <i>Arabidopsis</i> gel dataset | 47 |
| 4.2 | Clusters for <i>Arabidopsis</i> shotgun dataset | 52 |
| 4.3 | Clusters for <i>Arabidopsis</i> combined dataset | 60 |
| 4.4 | Clusters for Maize combine dataset | 75 |

LIST OF FIGURES

| | | |
|-----|---|----|
| 1.1 | Measuring gene expression | 3 |
| 1.2 | Integration of proteomic and transcriptional data from [38] | 4 |
| 1.3 | The evolution of Crick's central dogma from 1950s to today [48] | 7 |
| 1.4 | Approach for integration of proteomic and transcriptional data (Adapted from [38]) | 8 |
| 2.1 | Example of an approximately 40,000 probe spotted oligo microarray with enlarged inset to show detail [39] | 10 |
| 2.2 | Protein identification methods from [38] | 12 |
| 3.1 | Two data sets consist of <i>Arabidopsis</i> gene identifiers | 29 |
| 3.2 | Cluster dendrogram for the <i>Arabidopsis</i> Data set 1 | 36 |
| 3.3 | Cluster dendrogram for the <i>Arabidopsis</i> Data set 2 | 37 |
| 3.4 | Cluster dendrogram for the combined <i>Arabidopsis</i> data set | 41 |

LIST OF SYMBOLS

| | |
|-------|---|
| 2D | Two Dimensional |
| 2DE | Two Dimensional Electrophoresis |
| BLAST | Basic Local Alignment Search Tool |
| DNA | Deoxyribonucleic acid |
| EST | Expressed Sequence Tag |
| GO | Gene Ontology |
| ID | Identifier |
| mRNA | messenger ribonucleic acid |
| MS | Mass Spectrometry |
| RNA | Ribonucleic acid |
| USDA | United States Department of Agriculture |

CHAPTER 1

INTRODUCTION

Computational biology is an interdisciplinary field that applies the techniques of mathematics, statistics and computer science to solve biological problems. A major focus of both biology and computational biology over the past decade has been the development of different methods for measuring changes in gene expression under different conditions. Data obtained from different methods often yield different, but complementary information. The goal of this thesis is to present a new approach for integrating information from different techniques and/or experiments about gene and protein expression in a meaningful way.

The central dogma of molecular biology explains the formation of major molecules in a living organisms: DNA, RNA and protein. DNA, the genetic information inherited from generation to generation, is a chain of nucleic acids from a four letter alphabet [16]. Small sections of the DNA strands (substrings from a computer science point of view) contain information for making particular proteins and are known as genes. Proteins are macromolecules consisting amino acids from a 20 letter alphabet. Proteins perform metabolic structural, defense and regulatory functions in and out of the cell. The central dogma describes how DNA is replicated and converted to messenger RNA (mRNA) and protein through transcription and translation. During replication, double stranded DNA forms

duplicate copies of itself. During transcription, DNA segments containing genes are transcribed into single stranded RNA (messenger RNA) which also has a four letter alphabet. RNA strands are then translated into amino acids and form the proteins.

All cells in the body of an organism contain the same set of genes, but not all of these genes are transcribed and translated into proteins in every cell. A gene is considered to be expressed when it is actively involved in transcription to produce mRNA, the first step of protein production. A protein is considered to be expressed when the mRNA is translated. Therefore, we can assay gene expression at either the mRNA level or the protein level as shown in the Figure 1.1. Gene expression microarrays are a popular platform for measuring mRNA levels across different biological samples [11]. Microarray technology allows scientists to have a view of the expression of thousands of genes simultaneously [4]. These types of studies help scientists identify differentially expressed genes under different conditions and pave the way for identification of response to stimuli, transcriptional pathways, cell differentiation, disease markers and drug targets in the long term [38].

Our goal is to integrate multiple datasets measuring gene and/or protein expression to gain an overall picture of the active biological processes under different conditions. The types of datasets that we want to integrate have several characteristics that makes this process challenging. Figure 1.2 shows the most common approach of integrating proteomic and transcriptional data.

A similar approach is used for integrating expression data from different technologies for the same data type (transcriptome or proteome). The two types of data are linked using a common reference database such as UniGene [38]. But the process of linking mRNA and

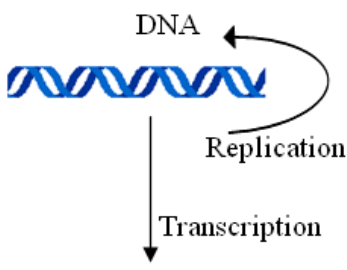
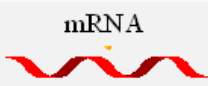

| Central Dogma | What is being measured? | How measurement is done? |
|--|-------------------------|--|
|  <p>DNA</p> <p>Replication</p> <p>Transcription</p> | | |
|  <p>mRNA</p> <p>Translation</p> | mRNA | Microarrays Real-time PCR High throughput sequencing |
|  <p>Protein</p> | Protein | 2 D-gel electrophoresis Shotgun proteomics |

Figure 1.1

Measuring gene expression

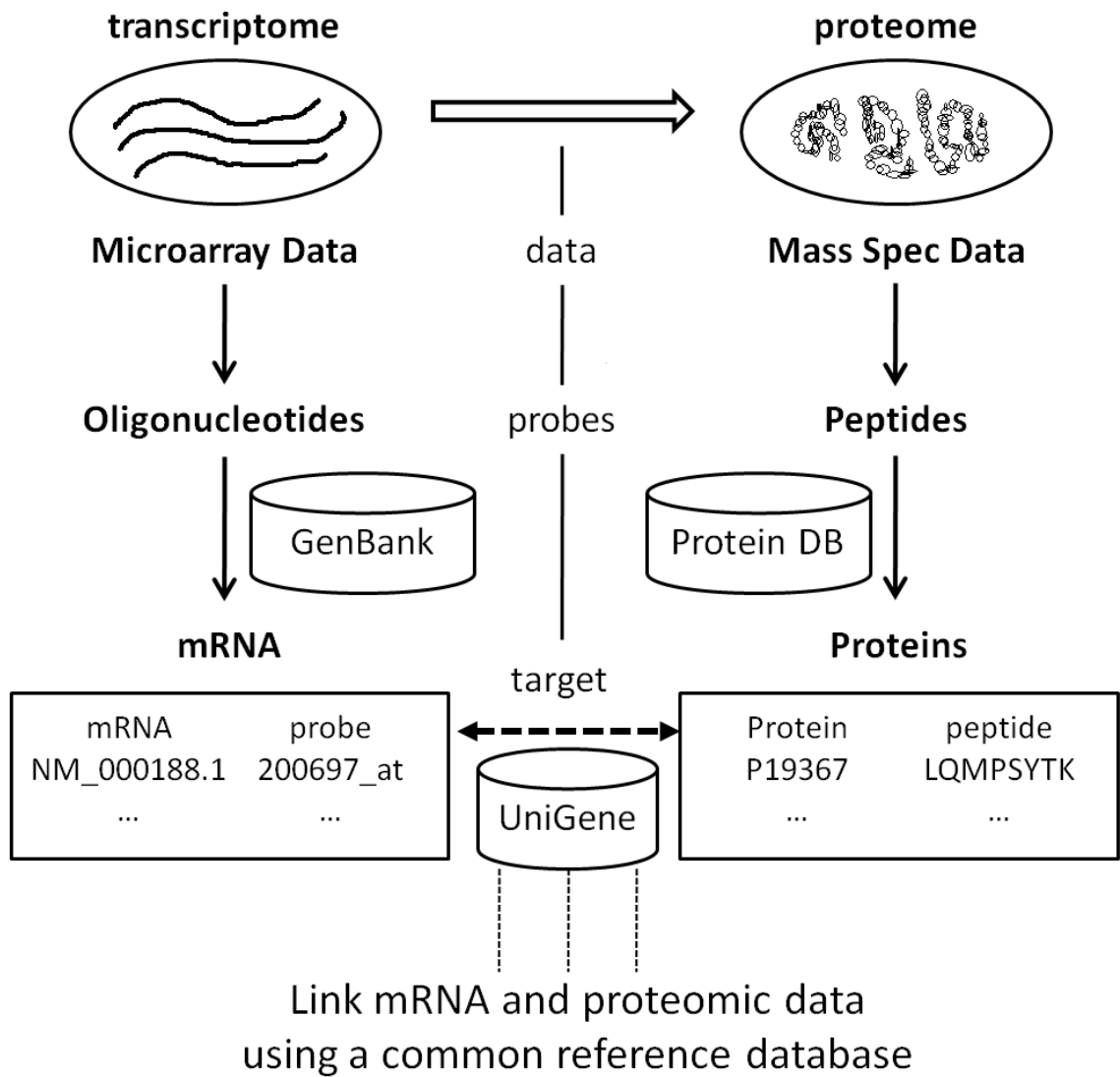


Figure 1.2

Integration of proteomic and transcriptional data from [38]

proteomic data through identifiers often results in a very small number of matches even in very controlled experiments [38]. There are many ambiguities involved in the process of connecting DNA probes to the target mRNA. First, the central dogma is not as simple as shown in Figure 1.1. Apart from replication, transcription and translation, there are many complex processes such as post transcription regulatory mechanisms and post translation mechanisms that take place as shown in Figure 1.3. Second, proteomics techniques and transcriptomic techniques are different and have different biases, sources of noise etc. There are complications that make the matching process difficult, even when dealing with a single type of data such as microarray data. When we measure gene expression using a microarray, there is a possibility of mapping multiple probes to the same gene or the same probe to different products of the same gene [38]. The situation is even worse for the heterogeneous datasets we are considering. First, we consider multiple genotypes of the same species (*Zea mays*) and there is substantial variation in the gene content of different genotypes in maize [53]. Second, plant genotypes often significantly differ in the genes activated in response to different conditions. Third, tissues from field grown samples where the environmental conditions are not controlled will exhibit a great deal of variation. Fourth, in some cases, the tissues were collected from different experiments conducted in different years. Fifth, in some cases we have measurements of expression from different technologies for the same tissue, and it has been demonstrated that there can be wide variations in the genes or proteins detected by the technologies. For example, two common methods of measuring protein expression are 2-d gel electrophoresis [14] and shotgun proteomics [37, 25]. A number of different studies have shown that the

overlap in the proteins identified by these two methods is quite low (20-30%) even when using exactly the same biological sample [41, 9]. Therefore, matching of identifiers across multiple data sets cannot be applied successfully in many of our experiments.

The main objective of this thesis is to develop a new method to obtain functional similarities among heterogeneous protein/gene data sets by constructing functional similarity matrices and applying a clustering algorithm. For each dataset, we will abstract the differentially regulated genes to the functional level, and analyze the data at this level as shown in Figure 1.4. During this process, first we assign functional annotations for heterogeneous gene/protein data sets using available online tools. We then compute the semantic similarities among these genes/proteins based on their functional annotations. Finally we adapt a hierarchical clustering algorithm to obtain functional clusters of genes/proteins. Resulting clusters consist of functionally similar groups of genes/proteins in heterogeneous data sets.

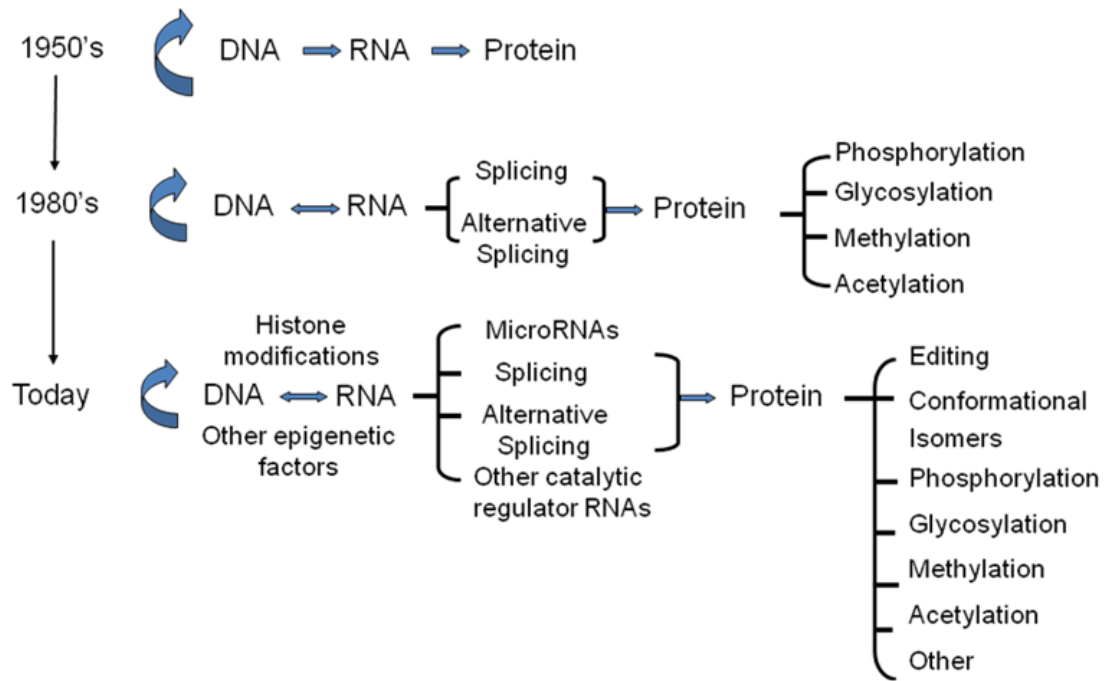


Figure 1.3

The evolution of Crick's central dogma from 1950s to today [48]

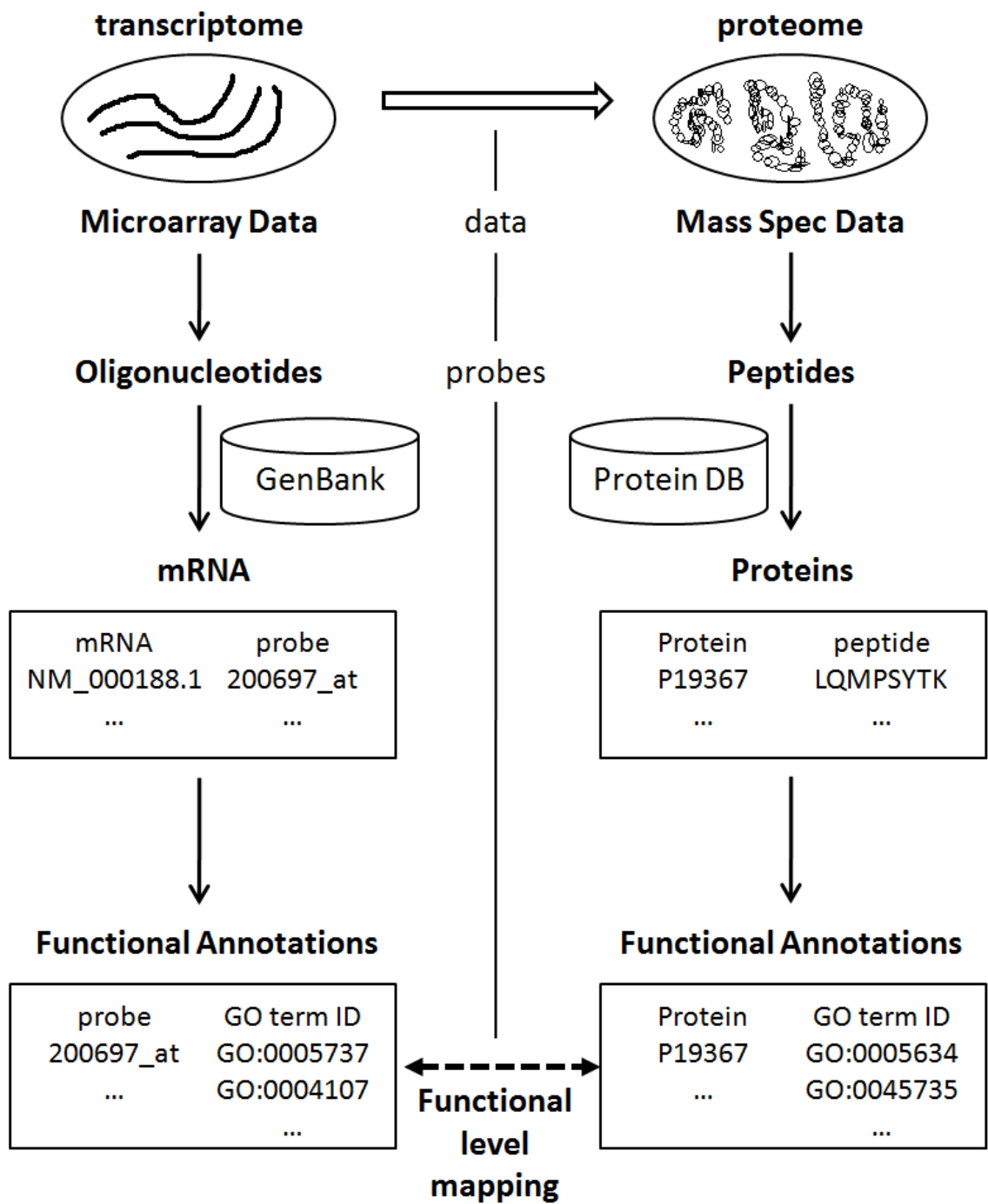


Figure 1.4

Approach for integration of proteomic and transcriptional data (Adapted from [38])

CHAPTER 2

LITERATURE REVIEW

The main objective of this thesis is to develop and implement an effective method for integrating heterogeneous gene/protein data sets at the functional level. In this chapter we review background information about proteomics and transcriptomics, current techniques used to integrate heterogeneous data, and the limitations of current techniques. Section 2.1 describes the most widely used technologies for measuring gene expression at the transcriptome and proteome levels. Section 2.1.1 discusses methods for linking heterogeneous datasets through identifiers and the strengths and weaknesses of these approaches. Section 2.1.2 describes methods used to correlate protein and microarray data. Because our method is based on integrating datasets at the functional level using the Gene Ontology (GO), Section 2.2 presents a description of the GO. The importance of functional level mapping and available computational tools that use this approach are discussed in section 2.3. Section 2.4 presents different semantic similarity measures which can be used to compute similarities among GO terms and genes.

2.1 Transcriptome and proteome technology

Proteomics and transcriptomics are relatively new research tools which help biologists understand how expressed proteins and genes change in complex biological systems.

Gene expression is currently most often analyzed using microarrays. A microarray is a chip of an arrayed series of thousands of microscopic spots of short segments of DNA or RNA called oligonucleotides. These oligonucleotides are designed to bind mRNA, and the bound oligos transmit a light signal which is detected. A series of needles controlled by robotic arms are used to deposit these oligonucleotides into the designated locations on the microarray chip. This resulting grid of oligonucleotides as in Figure 2.1 represents nucleic acid profiles and can be used to measure the gene expression in terms of messenger RNA (mRNA) or DNA. Gene microarrays can also be used to examine the global changes in mRNA throughout different biological settings [11, 27].

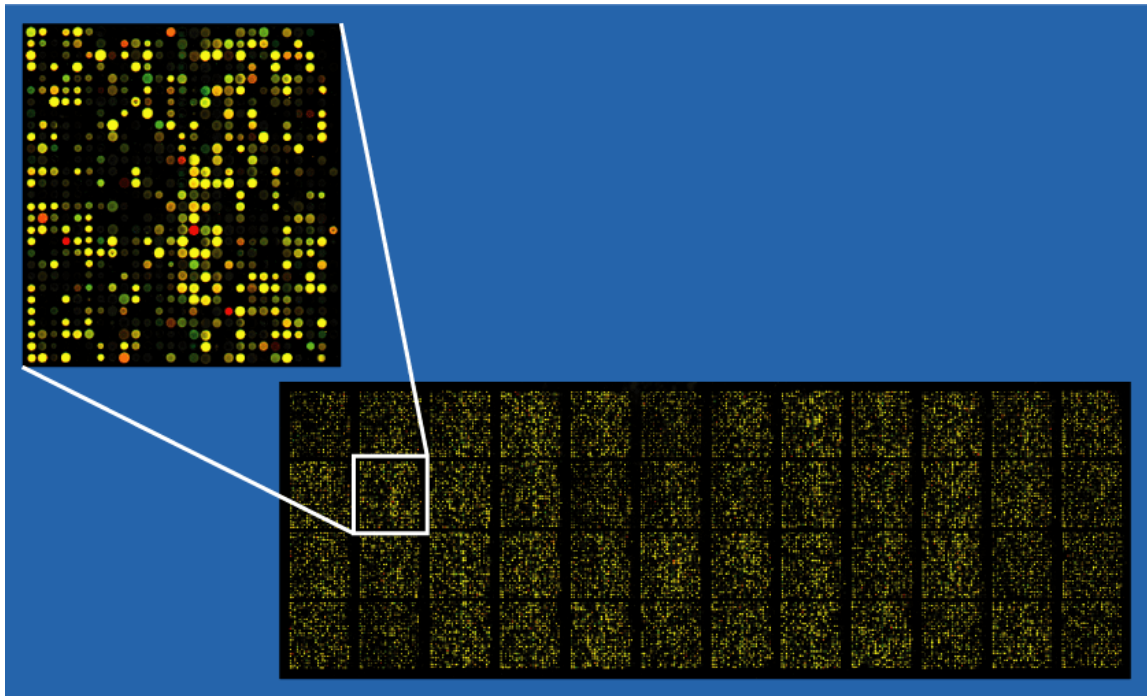


Figure 2.1

Example of an approximately 40,000 probe spotted oligo microarray with enlarged inset to show detail [39]

Two-dimensional gel-based electrophoresis (2D gel) and shotgun profiling methods followed by mass spectrometry are widely used to identify the relative abundance of proteins in complex biological samples [14]. Normally, there are two processes involved in each of these proteomic techniques: separation of proteins in a complex protein mixture and identification of the proteins. In a typical 2D gel-based approach, the proteins are separated, visualized and digested into peptides and then identified by mass spectrometry [38]. As Figure 2.2 shows, in both the 2D gel approach and shotgun approach, the protein mixture is digested into peptides and the resulting peptides are separated using liquid chromatography. When the peptides elute from the chromatography column, they are directly subjected to mass spectrometry (MS/MS) for sequencing. A database approach is used to identify the peptides based on tandem mass spectra assigned to each peptide and then used to identify the proteins. 2D gel methods can be used to identify different protein isoforms, and this cannot usually be done with shotgun proteomics [38]. Because of the large numbers of proteins that can be identified using the shotgun proteomics, this method is rapidly gaining in popularity over 2D gels. However, both the protein identification techniques provide complementary information about the biological samples.

It is important to be aware of the technical limitations associated with different platforms for profiling gene expression. For example, one major limitation of microarray experiments is that they can only detect genes with representative probes on the chip [11]. Mass spectrometry (MS) techniques for identifying proteins also have several limitations including incompleteness and redundancy of protein sequence databases used for searching MS spectra [38, 14]. In addition, the choice of the database and the search algorithm

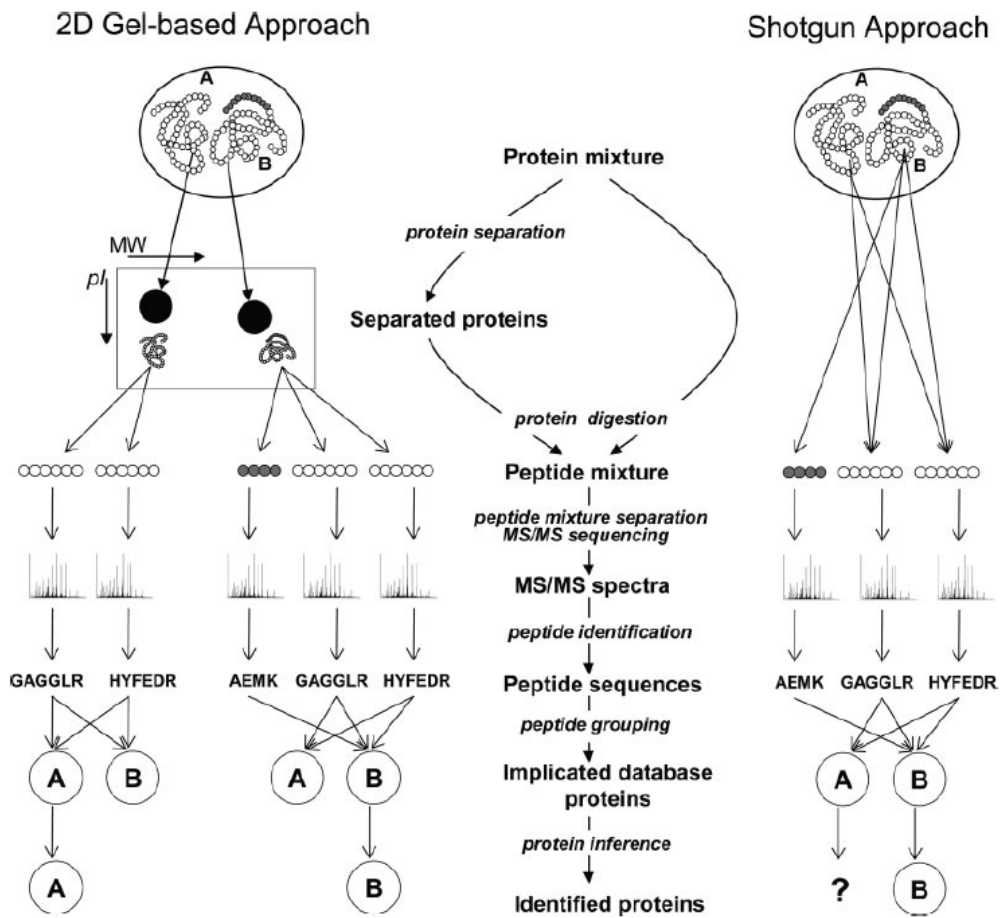


Figure 2.2

Protein identification methods from [38]

can be crucial to the success rates of protein identifications [51, 49, 34]. Extracting quantitative information for low density peptides is also a big challenge as high abundance proteins are preferably detected by liquid chromatography-mass spectrometry (LC-MS) [14]. Proper selection of samples is also equally important to generate accurate results. Because different techniques for measuring gene and protein expression have different strengths and limitations, researchers are interested in integrating complementary data sets to achieve a more complete picture of the complex biological systems they are investigating.

2.1.1 Linking heterogeneous datasets through identifiers

Once the microarray and proteomics experiments are completed, the next step is to match the genes represented on different microarrays or match the genes with the corresponding proteins identified in the proteomic datasets. Normally, commercial sources of microarrays such as Affymetrix chips provide a list of sequences spotted in the array along with GenBank accession number of the target RNA sequence, and brief functional annotation for each probe [38, 11]. In proteomic experiments, each MS/MS spectrum is assigned to a peptide, and the peptides are assembled to proteins using a variety of protein sequence databases [38, 14]. The process of integrating different protein and/or transcriptomic data sets is hindered by use of different accessioning schemes and lack of annotations. Regardless of the platform, biologists have to perform some cross referencing or indexing in order to know the corresponding protein sequence identifiers. There are several registered web sites available for cross-referenced annotations such as www.affymetrix.com for Affymetrix array users [38]. Most typical identifiers refer to databases such as Swis-

sProt / TrEMBL (SPTR), NCBI, ENSEMBL and UniGene. However, there are several drawbacks accompanied with the usage of most of these identifiers. For example, although SwissProt (SP) is a very popular choice for spectral database searches as it has highly curated data, generally it does not contain the complete set of proteins for many organisms [11]. TrEMBL (TR) is the companion database for SwissProt, which contains computer-annotated supplements for all the nucleotide translations which are not integrated into SwissProt. Although, TrEMBL provides more extensive coverage, the TR identifiers are frequently redundant, unannotated and continuously retired and replaced by SwissProt IDs as the proteins migrate to SP. Most of the gene and protein databases suffer from the similar kind of problems. Although NCBI has made an attempt to standardize and reduce the ID redundancy by creating RefSeq (protein) and NM (mRNA) accession systems, it still suffers from some of the above problems. UniprotKB is another database which tries to assign a unique ID for transcripts, which makes the situation worse, because sometimes they pick their own ID [1]. UniGene (www.ncbi.nlm.nih.gov) is a well annotated database which can be used as a common reference in correlating mRNA and protein data [42]. UniGene is generated from species-specific clusters created based on nucleotide sequence similarity [38, 11]. Recently, there are a number of tools developed which have the ability to link the probes from Affymetrix arrays to UniGene identifiers as well as to connect the RefSeq protein database sequences to UniGene [31, 19].

The drawback of using UniGene is whenever new members are added to the collection, all the clusters are recalculated. During this process, some members of previous clusters might move to new clusters and sometimes old cluster IDs are completely re-

moved [11]. This leads to a problem of having legacy data sets. Therefore it is important to make sure all UniGene clusters are built in the same date when linking data sets using Uniene. Ensembl (www.ensembl.org) is also an annotation database which assigns IDs in an effective manner. Ensembl IDs are assigned to genes/proteins if they can be associated with an assembled genome which makes them a more stable, non redundant set of identifiers [11]. For some instances, BLAST sequence alignment is the most suitable way to link databases. Species-specific sequences can be downloaded for the relevant sequence identifiers. Tools such as stand-alone BLAST or utilities like BioEdit [32] can be used to perform searches referring to one sequence as the query and the other one as the subject. BLAST results should be interpreted in terms of percent identity, sequence coverage and e value threshold.

2.1.2 Correlating protein and microarray data

Several methods have been developed to perform integration and comparison studies among functional proteomics and gene expression data. However, the most fundamental question is how these different patterns of gene expressions correspond to the protein abundance in the cellular level [11]. A significant number of correlation studies comparing gene expression and protein expression are reported in the literature. For example, the study of Gygi et al. [23] reveals the correspondence between gene expression and protein in yeast by using protein and mRNA quantitation by collecting complementary data for 156 genes. This experiment has shown a modest positive correlation of mRNA and protein levels. Another group of researchers, Mootha et al. [36], tried to correlate the ex-

pression patterns of mitochondrial proteins in mammalian tissues with public microarray data. They used a simple test for concordance assigning a positive score for similar expression patterns in tissue for corresponding protein and mRNA expression and found 426 of 569 detected genes were concordant. However, there were several criticisms raised for this experiment including the reliability of the scoring schema. On the other hand, there is a bias in the data since the average mRNA abundance of the detectable proteins was found to be nearly five-fold higher than for other mitochondrial genes. This suggests that only high abundance gene products strongly correlate [36]. Griffin et al. [22] tried to determine whether the changes in expression correlate at the protein and transcript levels between two yeast populations grown in two different carbon sources. They collected complementary protein and mRNA abundance data for 245 genes during the experiment. Although the genes linked to carbon metabolism showed some changes in abundance, there were no relative changes in the protein levels or mRNA levels in similar magnitude.

Researchers have identified a number of reasons for the lack of a direct correlation between gene expression patterns and corresponding protein levels. One problem is that gene expression patterns measured using mRNA do not take the influence of translational and post-translational mechanisms into account [38, 36, 23, 22, 8]. For an example, a recent study of protein abundance in yeast carried out by Ghaemmaghami et al. [20] reveals that many essential proteins and transcription factors are present at levels that are not readily predicted by mRNA levels. But still there are several important factors behind comparing transcriptome and proteome beyond the traditional correlation analysis which consider the relative levels of protein and mRNA detected for the same gene. For example, the stud-

ies of Greenbaum et al. [21] revealed that there is a considerable similarity between the transcriptome and proteome in terms of enrichment for specific structural and functional properties. This sort of comparative analysis is immensely helpful in filling the knowledge gap between proteomics and transcriptomics technologies. This type of knowledge will provide biologists with knowledge needed to link gene and protein expression patterns in different molecular pathways and to determine the suitability for using gene transcript levels as a substitute for measuring protein activities [11]. The research we present adopts the approach of integration at the functional level.

2.2 Gene Ontology (GO)

The most widely used method for specifying the function of gene products is the Gene Ontology, and we use GO annotation to link heterogeneous datasets. The GO was developed to facilitate integration of functional data into value-added databases. In 1998, the representatives of Saccharomyces genome database, Drosophila genome database and Mouse genome database founded the Gene Ontology (GO) Consortium and agreed jointly to apply the same vocabulary to describe gene functions for every gene in the respective databases [29]. This project was a novel functional classification system because it was implemented among cross-species for the first time. The members of GO consortium are responsible for the design, development and implementation of publicly available databases which consist of expertly-curated functional annotations using the GO. GO is a hierarchical structure which is implemented as a directed acyclic graph (DAG) and consists of well-defined terms and relationships. GO terms describe three attributes of genes

and gene products: molecular function, biological process and cellular component. Members of GO consortium ensure that the GO functional annotations consist of a controlled vocabulary. Each annotation is associated with some kind of evidence which provides the source of the annotation. The most common evidence code for annotations is IEA- inferred by electronic annotation, which means that GO annotations depend on automated recognition of functional motifs [6]. The GO annotations “Inferred from sequence or structural similarities”, or ISS is mostly assigned by running BLAST searches. For all the other evidence codes, annotations are assigned by curators using literature curation. Although manual curation provides high quality GO annotations, it is a very time consuming task and currently covers only a very small percentage of available annotations. An alternative approach to obtain GO annotations is to use computational tools for text mining. Besides the identification of annotations, these tools can locate their evidence in literature [10]. But these interactive text mining programs result in very high error rates [43] and assignment of GO annotations by human curators remains the “gold standard” [10, 13].

GO has become the standard method for describing function because it uses a common vocabulary to describe the same gene functions across different species. This helps biologists overcome the difficulty of biological interpretation of large gene lists derived from high throughput genomic and proteomic studies. Biologists can get their data annotated to varying levels depending on the completeness of available information in GO [7]. Another major use of GO is finding under-or over-represented GO terms associated with a dataset in microarray analysis [17, 5]. This use of the GO has led to many arguments in the literature because these analyses are not based on the quantitative values on the microarray,

but rather on counts of GO terms. However, ultimately, researchers use GO as a vital tool which enables turning data into knowledge. GO annotation has become the standard for functional annotation, and its usage is growing exponentially [7]. Computer scientists have made significant contributions to the development of computational tools that assign and analyze functional annotations and help to track related literature [17, 5, 10, 43, 13].

2.3 Functional level mappings

Many computational tools have been developed to facilitate interpretation of biological data in “batch” mode [4]. Most of these tools provide the user with functional annotations for each gene, summarize which genes are associated with specific biological processes, and rank these processes by over-representation analysis. Some of the tools which address this issue include, but are not limited to, GoMiner, DAVID, EasyGO, GOstat, GeneTools, AgBase [4, 55, 3, 35, 12, 26]. Although these tools are useful, they lack the ability to mine many-to-many gene-to-term relationships found in functional annotation databases, as well as the ability to condense redundant contents [12]. For example, individual genes can be associated with several biological terms, and those individual biological terms can be associated with several genes. Huang et al. [12] developed the tool DAVID, which uses a novel agglomeration algorithm that can extract this complex and redundant relationship by taking advantage of exploratory statistical methods. Their method identifies groups of genes sharing the same biological terms or groups of biological terms sharing similar genes and organizes them into biological modules. This is a powerful method to group functionally related genes and terms into biological modules

and has several advantages. First, it largely reduces redundant results into a manageable size while enhancing the understandability by visualizing gene-to-gene, term-to-term and gene-to-term relationships. Therefore investigators can quickly apply the information in a module to their study. Second, it is much easier to relate biological modules of interest to a study than it is to relate hundreds of individual terms. The database for annotation, visualization and integrated discovery (DAVID) has two implemented tools. One is gene functional classification tool, and the other one is functional annotation clustering tool, and both provide a module centric approach for functional analysis of large gene sets. DAVID is a user friendly, well-documented tool with an easily navigatable interface. DAVID accepts a range of different gene identifiers. After the user uploads the set of gene identifiers, DAVID converts those identifiers into its own DAVID identifiers before further processing. The drawback is sometimes DAVID does not have compatible identifiers for each of the identifiers uploaded by the user. Therefore the user cannot take maximum advantage of the functionalities implemented. DAVID displays results in a clear text and graphical formats. The unique fuzzy heat map visualization provides a clear global view of group-to-group relationships.

2.4 Computing the similarity of genes based on GO annotation

Researchers try to understand various aspects of relationship between gene function, gene expression and gene annotation. Most of the genomic studies are driven based on the assumption that functionally and biologically related genes would have similar expression

levels and gene ontology (GO) annotation [50]. This thesis focuses on how to explore gene similarity with respect to the semantic similarity of GO annotations.

Semantic similarity is a concept which describes the closeness of the relationship of GO terms in the GO hierarchical structure. The inverse of semantic similarity is semantic distance. There are a number of different methods available to calculate the semantic similarity among GO terms. One of the early techniques considers the path distances between GO terms [44]. Computation of the similarity merely considers the minimum number of edges that need to be traversed from one node to the other. The shorter the path between two GO terms, the more similar they are. However this edge-based method is implicitly based on the assumption that all the edges represent uniform distances and all nodes in the taxonomy are evenly distributed and have similar densities which is not necessarily true in the GO structure [46].

Instead of defining the similarity based on the structure of the GO, it is also possible to consider the information contained at the nodes based on the concepts in information science [2]. The information content of a node can be computed based on the known probability of each node within a lexical corpus. For example, the lexical corpus for a given organism is comprised of its GO annotations, and we can compute the probability of each term within the ontology [33]. When we traverse higher in the GO hierarchy, the probability increases and those top nodes are less informative. When we traverse deeper in to the GO hierarchy, the nodes have lower probabilities and therefore higher information content. This is very apparent because as we move up the GO taxonomy, the nodes are more general. Once the information content of the nodes are quantified, we can compute

node-based similarity measures. There are number of methods available to make use of information content of GO terms in order to compute the similarity between pairs of gene products including Resnik et al. [45], Jiang et al. [24] , Lin et al. [30].

The method developed by Resnik et al. relies on the notion of the shared information content of nodes as the basis for the semantic similarity measure. Information content $P(c)$ of particular node can be computed as the negated log of the likelihood as,

$$P(c) = -\log [p(c)] . \quad (2.1)$$

According to Resnik et al., semantic similarity between two nodes can be defined as information content of their minimum subsumer. Whenever there is more than one minimum subsumer, as often happens in the GO due to multiple inheritance, the most informative subsumer is choosen. Equation (2.2) defines the similarity between two GO terms,

$$sim(c_1, c_2) = -\log [p_{ms}(c_1, c_2)] , \quad (2.2)$$

where c_1 and c_2 are GO terms, and $p_{ms}(c_1, c_2)$, is the probability of minimum subsumer.

We focus on comparing two gene products rather than GO terms as explained above. Resnik et al., defines similarity between two genes, g_1 and g_2 , as the maximum similarity found between any two GO terms and the formula is given as,

$$sim(g_1, g_2) = \max [sim(c_1, c_2)] \quad (2.3)$$

where $c_1 \in A(g_1)$, $c_2 \in A(g_2)$, and $A(g_1)$ and $A(g_2)$ are the GO annotations of genes g_1 and g_2 respectively.

Jiang et al. proposes a similarity measure which is a mixed approach inherited from an edge-based method and is enhanced by the information content calculation methods. In addition to the information content, the other factors such as local density, node depth, and link type are also being considered. The overall edge weight wt for a child node c and its parent node a is defined as,

$$wt(c, a) = \left(\beta + (1 - \beta) \frac{\bar{E}}{E(a)} \right) \left(\frac{d(a) + 1}{d(a)} \right)^\alpha [\log(p(a)) - \log(p(c))] T(c, a) \quad (2.4)$$

where $d(a)$, denotes the depth of the node a , $E(a)$, the number of edges in the child links (local density), the average density in the whole hierarchy, $-\log(p(c))$ and $-\log(p(a))$ the information content of nodes c and a , and $T(c, a)$ the link relation/type factor. α and β are two weighting constants.

The overall distance between two nodes $dist(g_1, g_2)$ is defined as

$$dist(g_1, g_2) = \sum_{c \in \{path(c_1, c_2) - MS(c_1, c_2)\}} wt(c, a) \quad (2.5)$$

where $path(c_1, c_2)$, is the set that contains all the nodes in the shortest path from c_1 to c_2 . One of the elements in the set is $MS(c_1, c_2)$ which denotes the lowest subsumer of c_1 and c_2 [24].

Lin et al. also defines an information theoretic similarity measure which is applicable to different domains. When it is applied to GO, the similarity would be defined as:

$$sim(g_1, g_2) = \frac{2 \log(p_{ms}(c_1, c_2))}{\log(p(c_1)) + \log(p(c_2))} \quad (2.6)$$

where $c_1 \in A(g_1)$, $c_2 \in A(g_2)$.

There are number of studies available in the literature which investigate the utility of the above three measures to compare GO semantic similarity and its correlation to gene expression similarities and protein sequence similarities. Sevilla et al. [50] computed the similarities between genes based on the correlation between their expression profiles (calculating the Pearson correlation coefficient or its absolute value). Then they annotated the gene products to GO terms and computed semantic similarity using three similarity measures described above. Finally they analyze the correlation between the expression similarity of gene products and corresponding semantic similarity. They conclude that the Resnik semantic similarity clearly outperforms both Jiang's and Lin's semantic measures and suggests that Resnik's similarity measure is well suited for Gene Ontology.

Wang et al. [54] also evaluated above three different methods of semantic similarity measures and showed that Resnik's method is better than other methods in terms of the correlation with gene sequence similarities and gene expression profiles.

Another study carried out by Lord et al. [33] investigated the three measures to compare semantic similarities of GO and its correlation to protein sequences. They also reported that the Resnik measure may be the most discriminatory while Jiang distance shows the weakest correlation.

CHAPTER 3

APPROACH

3.1 Background and hypothesis

Biologists attempt to understand complex biological processes through the analysis of gene expression at either the mRNA level, protein level, or both. DNA microarray analysis is used to measure mRNA abundance, and quantitative MS/MS based proteomic analysis is used to measure protein abundance in biological samples. Since microarray technology is technically more advanced, it allows monitoring of RNA expression levels for a significantly larger number of genes than can be identified in a typical proteomics experiment [38]. Microarrays can also be effectively used for the analysis of alternative splicing and genome annotation. Often several different gene expression experiments are conducted over time and there is a need to integrate the data from multiple experiments. However, there may be changes in the arrays used for the experiments and in the experimental design and so there may not be a straightforward mapping from one dataset to the other. RNA expression levels alone are not sufficient to understand protein expression and function because the mRNA levels do not reflect post-transcriptional regulatory mechanisms such as protein translation, post translational modifications etc. Proteomics experiments can provide this sort of information. There are two commonly used technologies for studying protein expression—gel based proteomics and shotgun proteomics.

Shotgun proteomics experiments will typically detect many more proteins than gel-based experiments but shotgun proteomics cannot detect isoform differences or be able to distinguish proteins from large gene families. Therefore, there is often a need to combine data from multiple gene expression experiments, multiple proteomics experiments, or a combination.

Currently, the most popular approach to integrate these transcriptional and proteomic data sets is to cross-reference the data sets through a common ID such as SwissProt, Trembl, Ensembl etc. This approach is hindered by the lack of a standard accessioning scheme and lack of relevant annotations. Different protein sequence databases use unique accessioning schemes. The degrees of sequence annotations also usually do not allow an easy cross reference between either different protein sequence data bases or protein and genomic databases. Therefore it is very difficult to obtain a complete set of matching IDs during the process of linking transcriptomic and proteomic data sets. This problem particularly troublesome when the organism being studied is not sequenced or has only recently been sequenced and the structural annotation is quite immature. In addition, researchers have found only a weak correlation between gene expression measured at the mRNA level and protein level even under very highly controlled conditions in well-studied organisms [20].

This thesis presents a high level approach to solve the problem of dataset integration by obtaining a set of functional annotations for each of the datasets and mapping from items in one dataset to items in the other dataset based on GO annotations. The strength of the relationships between elements in the heterogeneous data sets is determined by the

gene similarity measured based on similarity of GO annotations. The groups of genes or proteins with similar functional annotations are obtained by applying a hierarchical clustering algorithm.

Hypothesis: Integration of heterogeneous gene expression datasets by mapping at the functional level using a hierarchical clustering algorithm can provide additional useful biological information that cannot be easily obtained by mapping at the identifier level.

3.2 Steps in the approach

Firstly, functional annotations for genes and/or proteins in the two datasets will be obtained and stored in a mapping file containing corresponding gene identifiers and GO terms along with their evidence codes. The GO Consortium reports associations between gene products and GO identifiers regularly, and this type of information is available through a number of websites including AgBase (www.agbase.msstate.edu), EMBL-EBI (www.ebi.ac.uk), and TAIR (www.arabidopsis.org). We used the GO annotations stored in a statistical package called GOSim.

Next the similarity between individual GO terms will be computed based on well known information theoretic similarity measures introduced by Resnik [45] using Equation (2.2).

This computation of GO term similarity requires the information content of each GO term for the three GO categories: molecular function, biological process and cellular component. The information content of GO terms is precomputed using Equation (2.1) and stored in data files in order to speed up computation of GO term similarity.

As the third step, the similarity among the genes in each individual data set and the similarity of genes among combined data set is computed based on the similarities of their GO annotations using the Equation (2.3). We are using GOSim (www.dkfz.de/mga2/gosim) for steps 2 and 3 [18].

Figure 3.1 shows an example of two sets of artificial *Arabidopsis* gene identifiers that were processed using the three steps above. Table 3.1, Table 3.2, and Table 3.3 display the gene similarity matrices obtained for data set 1, data set 2 and the combined data set respectively.

The final step of the implemented method is to apply a hierarchical clustering algorithm to group similar elements into clusters. Hierarchical clustering creates a hierarchy of clusters which may be represented in a tree structure called a dendrogram. The hierarchical clustering algorithm that we used is an agglomerative algorithm. It begins with each element as a separate cluster and merges them into successively larger clusters based on the distance measure. The distance measure determines the similarity of two cluster elements; in our case the similarity matrix is generated based on the similarity of GO annotations of each pair of gene products. Figure 3.2, Figure 3.3, and Figure 3.4 show the cluster dendrograms obtained by applying the hierarchical clustering algorithm to the similarity matrices given in Table 3.1, Table 3.2, and Table 3.3 respectively. These clusters provide the mappings between the data sets at the functional level. Data set 1 generates two clusters and Data set 2 generates 3 clusters. The GO annotations for the clusters in both datasets and combined data set are shown in Table 3.4, Table 3.5 and Table 3.6 respectively.

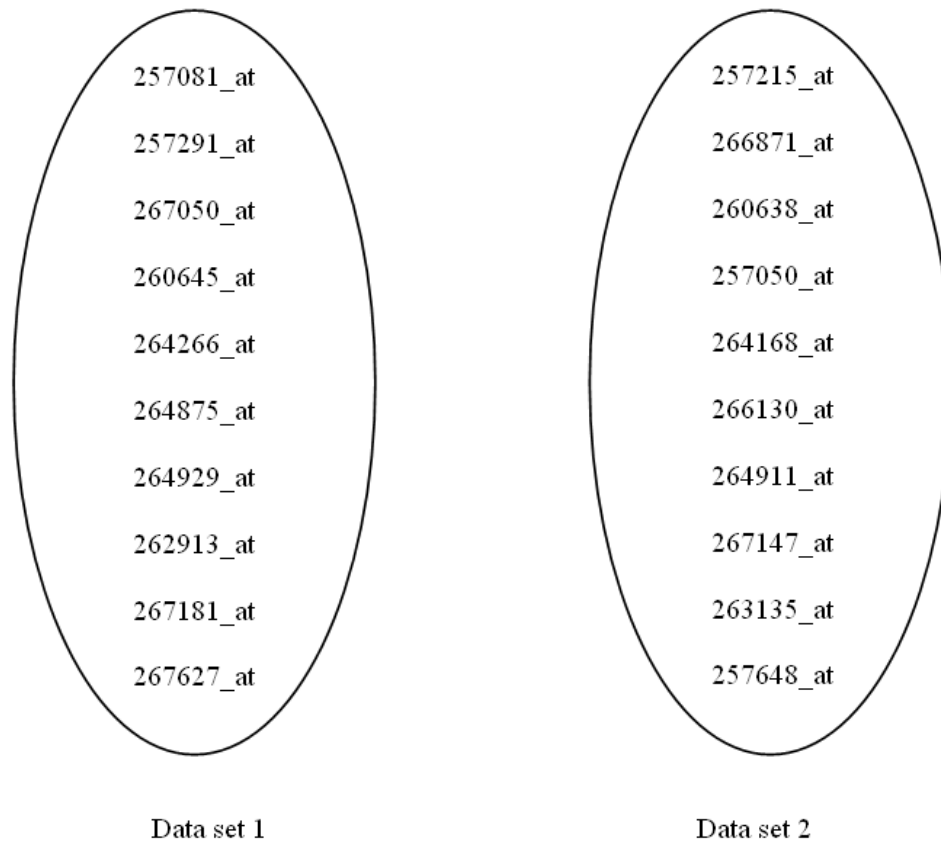


Figure 3.1

Two data sets consist of *Arabidopsis* gene identifiers

Table 3.1

Gene similarity matrix for *Arabidopsis* data set 1

| | 257081_at | 257291_at | 267050_at | 260645_at | 264266_at | 264875_at | 264929_at | 262913_at | 267181_at | 267627_at |
|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| 257081_at | 1.000 | 0.448 | 0.306 | 0.578 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 257291_at | 0.448 | 1.000 | 0.589 | 0.665 | 0.000 | 0.448 | 0.000 | 0.000 | 0.000 | 0.000 |
| 267050_at | 0.306 | 0.589 | 1.000 | 0.394 | 0.000 | 0.306 | 0.000 | 0.000 | 0.000 | 0.000 |
| 260645_at | 0.578 | 0.665 | 0.394 | 1.000 | 0.000 | 0.578 | 0.000 | 0.000 | 0.000 | 0.000 |
| 264266_at | 0.000 | 0.000 | 0.000 | 0.000 | 1.000 | 0.000 | 0.159 | 0.256 | 0.256 | 0.438 |
| 264875_at | 1.000 | 0.448 | 0.306 | 0.578 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 264929_at | 0.000 | 0.000 | 0.000 | 0.000 | 0.159 | 0.000 | 1.000 | 0.559 | 0.559 | 0.159 |
| 262913_at | 0.000 | 0.000 | 0.000 | 0.000 | 0.256 | 0.000 | 0.559 | 1.000 | 1.000 | 0.255 |
| 267181_at | 0.000 | 0.000 | 0.000 | 0.000 | 0.256 | 0.000 | 0.559 | 1.000 | 1.000 | 0.255 |
| 267627_at | 0.000 | 0.000 | 0.000 | 0.000 | 0.438 | 0.000 | 0.159 | 0.255 | 0.255 | 1.000 |

Table 3.2

Gene similarity matrix for *Arabidopsis* data set 2

| | 257215 | 266871 | 260638 | 257050 | 264168 | 266130 | 264911 | 267147 | 263135 | 257648 |
|-----------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 257215_at | 1.000 | 0.306 | 0.578 | 0.000 | 0.000 | 0.448 | 0.000 | 0.652 | 0.652 | 0.000 |
| 266871_at | 0.306 | 1.000 | 0.394 | 0.000 | 0.000 | 0.589 | 0.000 | 0.185 | 0.185 | 0.000 |
| 260638_at | 0.578 | 0.394 | 1.000 | 0.000 | 0.000 | 0.665 | 0.000 | 0.259 | 0.259 | 0.000 |
| 257050_at | 0.000 | 0.000 | 0.000 | 1.000 | 0.000 | 0.212 | 0.159 | 0.256 | 0.256 | 0.438 |
| 264168_at | 0.000 | 0.000 | 0.000 | 0.000 | 1.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 266130_at | 0.448 | 0.589 | 0.665 | 0.212 | 1.000 | 1.000 | 0.163 | 0.264 | 0.264 | 0.972 |
| 264911_at | 0.000 | 0.000 | 0.000 | 0.159 | 0.000 | 0.163 | 1.000 | 0.559 | 0.559 | 0.159 |
| 267147_at | 0.652 | 0.185 | 0.259 | 0.256 | 0.000 | 0.264 | 0.559 | 1.000 | 1.000 | 0.255 |
| 263135_at | 0.652 | 0.185 | 0.259 | 0.256 | 0.000 | 0.264 | 0.559 | 1.000 | 1.000 | 0.255 |
| 257648_at | 0.000 | 0.000 | 0.000 | 0.438 | 0.000 | 0.972 | 0.159 | 0.255 | 0.255 | 1.000 |

Table 3.3

Gene similarity matrix for combined datasets

| | 257081_at | 257291_at | 267050_at | 260645_at | 264266_at | 264875_at | 264929_at | 262913_at |
|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| 257081_at | 1.000 | 0.448 | 0.306 | 0.578 | 0.000 | 1.000 | 0.000 | 0.000 |
| 257291_at | 0.448 | 1.000 | 0.589 | 0.665 | 0.000 | 0.448 | 0.000 | 0.000 |
| 267050_at | 0.306 | 0.589 | 1.000 | 0.394 | 0.000 | 0.306 | 0.000 | 0.000 |
| 260645_at | 0.578 | 0.665 | 0.394 | 1.000 | 0.000 | 0.578 | 0.000 | 0.000 |
| 264266_at | 0.000 | 0.000 | 0.000 | 0.000 | 1.000 | 0.000 | 0.159 | 0.256 |
| 264875_at | 1.000 | 0.448 | 0.306 | 0.578 | 0.000 | 1.000 | 0.000 | 0.000 |
| 264929_at | 0.000 | 0.000 | 0.000 | 0.000 | 0.159 | 0.000 | 1.000 | 0.559 |
| 262913_at | 0.000 | 0.000 | 0.000 | 0.000 | 0.256 | 0.000 | 0.559 | 1.000 |
| 267181_at | 0.000 | 0.000 | 0.000 | 0.000 | 0.256 | 0.000 | 0.559 | 1.000 |
| 267627_at | 0.000 | 0.000 | 0.000 | 0.000 | 0.438 | 0.000 | 0.159 | 0.255 |
| 257215_at | 1.000 | 0.448 | 0.306 | 0.578 | 0.000 | 1.000 | 0.000 | 0.000 |
| 266871_at | 0.306 | 0.589 | 1.000 | 0.394 | 0.000 | 0.306 | 0.000 | 0.000 |
| 260638_at | 0.578 | 0.665 | 0.394 | 1.000 | 0.000 | 0.578 | 0.000 | 0.000 |
| 257050_at | 0.000 | 0.000 | 0.000 | 0.000 | 1.000 | 0.000 | 0.159 | 0.256 |
| 264168_at | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 |
| 266130_at | 0.448 | 1.000 | 0.589 | 0.665 | 0.212 | 1.000 | 0.163 | 0.264 |
| 264911_at | 0.000 | 0.000 | 0.000 | 0.000 | 0.159 | 0.000 | 1.000 | 0.559 |
| 267147_at | 0.652 | 0.229 | 0.185 | 0.259 | 0.256 | 0.218 | 0.559 | 1.000 |
| 263135_at | 0.652 | 0.229 | 0.185 | 0.259 | 0.256 | 0.218 | 0.559 | 1.000 |
| 257648_at | 0.000 | 0.000 | 0.000 | 0.000 | 0.438 | 0.000 | 0.159 | 0.255 |

Table 3.3

Gene similarity matrix for combined datasets (continued)

| | 267181_at | 267627_at | 257215_at | 266871_at | 260638_at | 257050_at | 264168_at | 266130_at |
|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| 257081_at | 0.000 | 0.000 | 1.000 | 0.306 | 0.578 | 0.000 | 0.000 | 0.448 |
| 257291_at | 0.000 | 0.000 | 0.448 | 0.589 | 0.665 | 0.000 | 0.000 | 1.000 |
| 267050_at | 0.000 | 0.000 | 0.306 | 1.000 | 0.394 | 0.000 | 0.000 | 0.589 |
| 260645_at | 0.000 | 0.000 | 0.578 | 0.394 | 1.000 | 0.000 | 0.000 | 0.665 |
| 264266_at | 0.256 | 0.438 | 0.000 | 0.000 | 0.000 | 1.000 | 0.000 | 0.212 |
| 264875_at | 0.000 | 0.000 | 1.000 | 0.306 | 0.578 | 0.000 | 1.000 | 1.000 |
| 264929_at | 0.559 | 0.159 | 0.000 | 0.000 | 0.000 | 0.159 | 0.000 | 0.163 |
| 262913_at | 1.000 | 0.255 | 0.000 | 0.000 | 0.000 | 0.256 | 0.000 | 0.264 |
| 267181_at | 1.000 | 0.255 | 0.000 | 0.000 | 0.000 | 0.256 | 0.000 | 0.264 |
| 267627_at | 0.255 | 1.000 | 0.000 | 0.000 | 0.000 | 0.438 | 0.000 | 0.972 |
| 257215_at | 0.000 | 0.000 | 1.000 | 0.306 | 0.578 | 0.000 | 0.000 | 0.448 |
| 266871_at | 0.000 | 0.000 | 0.306 | 1.000 | 0.394 | 0.000 | 0.000 | 0.589 |
| 260638_at | 0.000 | 0.000 | 0.578 | 0.394 | 1.000 | 0.000 | 0.000 | 0.665 |
| 257050_at | 0.256 | 0.438 | 0.000 | 0.000 | 0.000 | 1.000 | 0.000 | 0.212 |
| 264168_at | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 1.000 | 1.000 |
| 266130_at | 0.264 | 0.972 | 0.448 | 0.589 | 0.665 | 0.212 | 1.000 | 1.000 |
| 264911_at | 0.559 | 0.159 | 0.000 | 0.000 | 0.000 | 0.159 | 0.000 | 0.163 |
| 267147_at | 1.000 | 0.255 | 0.652 | 0.185 | 0.259 | 0.256 | 0.000 | 0.264 |
| 263135_at | 1.000 | 0.255 | 0.652 | 0.185 | 0.259 | 0.256 | 0.000 | 0.264 |
| 257648_at | 0.255 | 1.000 | 0.000 | 0.000 | 0.000 | 0.438 | 0.000 | 0.972 |

Table 3.3

Gene similarity matrix for combined datasets (continued)

| | 264911_at | 267147_at | 263135_at | 257648_at |
|-----------|-----------|-----------|-----------|-----------|
| 257081_at | 0.000 | 0.652 | 0.652 | 0.000 |
| 257291_at | 0.000 | 0.229 | 0.229 | 0.000 |
| 267050_at | 0.000 | 0.185 | 0.185 | 0.000 |
| 260645_at | 0.000 | 0.259 | 0.259 | 0.000 |
| 264266_at | 0.159 | 0.256 | 0.256 | 0.438 |
| 264875_at | 0.000 | 0.218 | 0.218 | 0.000 |
| 264929_at | 1.000 | 0.559 | 0.559 | 0.159 |
| 262913_at | 0.559 | 1.000 | 1.000 | 0.255 |
| 267181_at | 0.559 | 1.000 | 1.000 | 0.255 |
| 267627_at | 0.159 | 0.255 | 0.255 | 1.000 |
| 257215_at | 0.000 | 0.652 | 0.652 | 0.000 |
| 266871_at | 0.000 | 0.185 | 0.185 | 0.000 |
| 260638_at | 0.000 | 0.259 | 0.259 | 0.000 |
| 257050_at | 0.159 | 0.256 | 0.256 | 0.438 |
| 264168_at | 0.000 | 0.000 | 0.000 | 0.000 |
| 266130_at | 0.163 | 0.264 | 0.264 | 0.972 |
| 264911_at | 1.000 | 0.559 | 0.559 | 0.159 |
| 267147_at | 0.559 | 1.000 | 1.000 | 0.255 |
| 263135_at | 0.559 | 1.000 | 1.000 | 0.255 |
| 257648_at | 0.159 | 0.255 | 0.255 | 1.000 |

This small example demonstrates our method for constructing clusters from combined gene/protein expression data sets. Similarity measures between the proteins/genes in the two sets will be computed based on their functional annotations, and these will be used to establish similar clusters and thereby identify corresponding functional groups in the datasets.

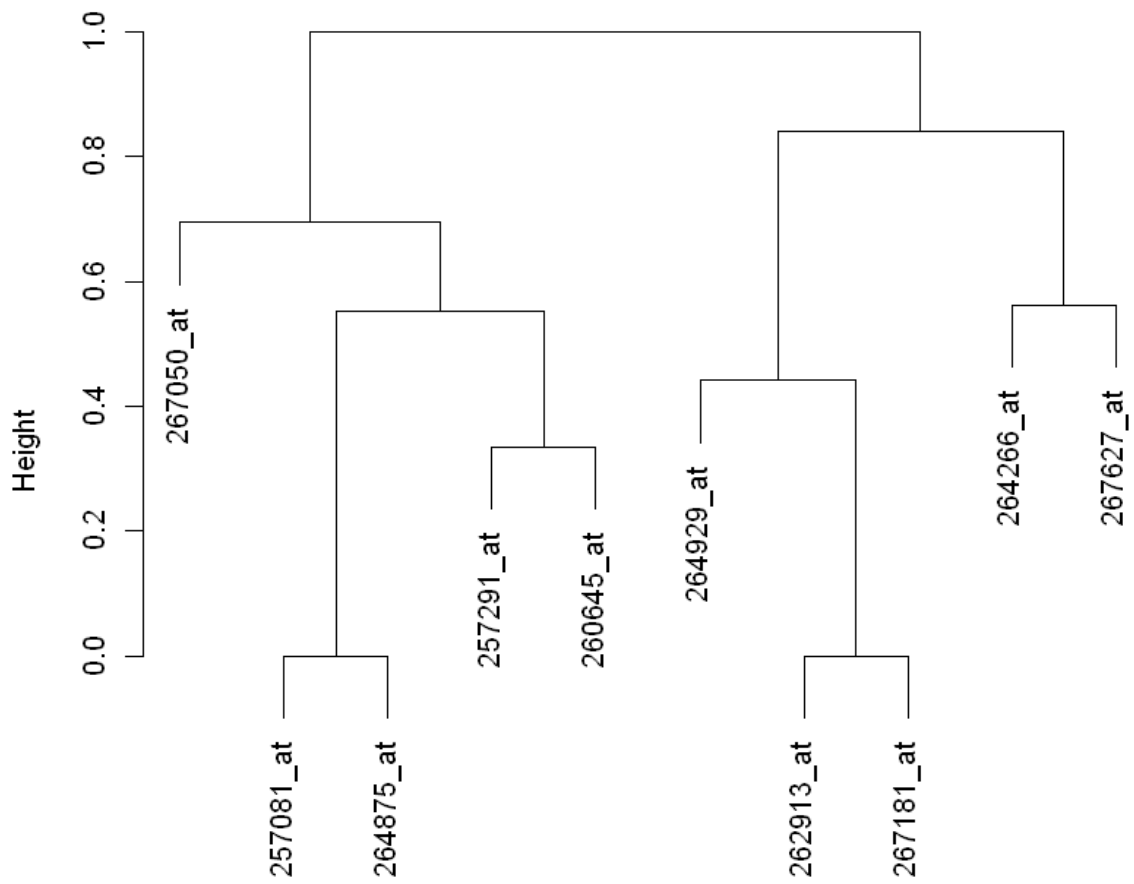


Figure 3.2

Cluster dendrogram for the *Arabidopsis* Data set 1

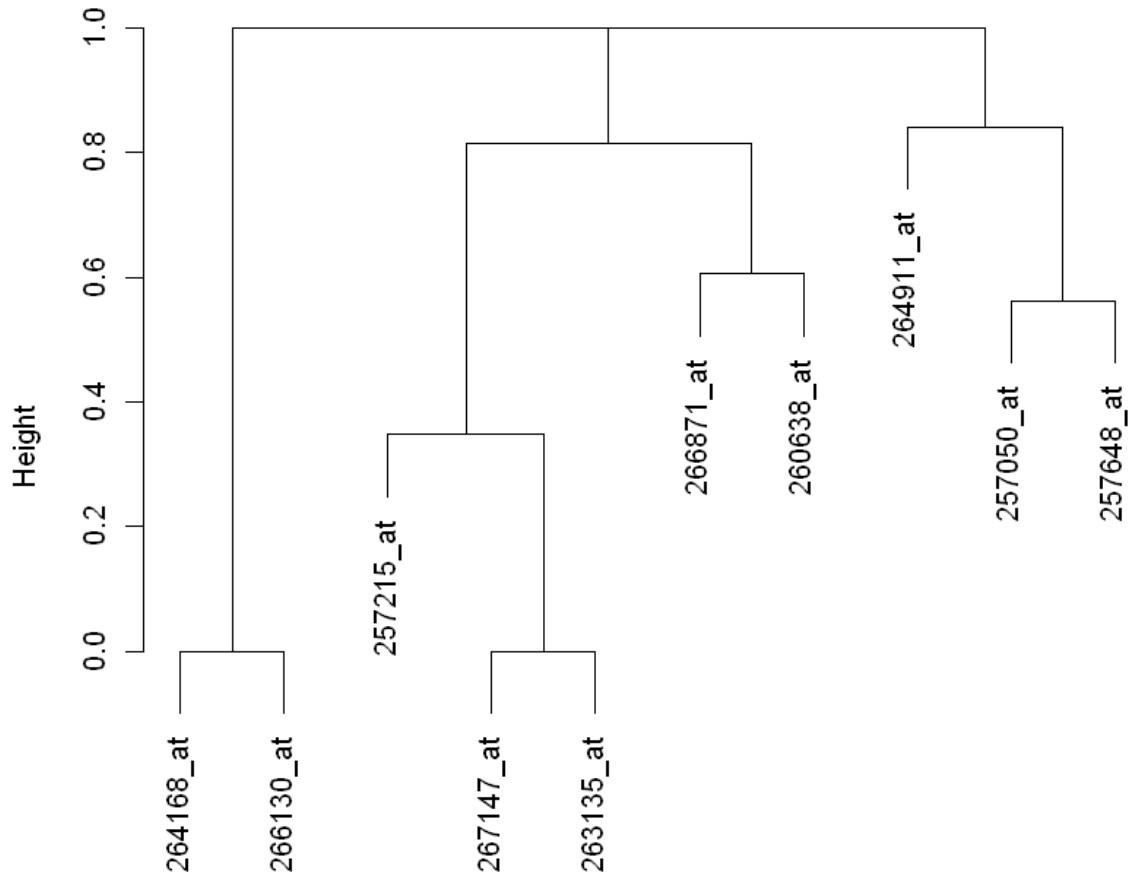


Figure 3.3

Cluster dendrogram for the *Arabidopsis* Data set 2

Table 3.4

GO annotations for the two clusters in *Arabidopsis* dataset 1

| Data set 1 | | | | | |
|------------|------------|-----------------|-----------|------------|---|
| Cluster 1 | | | Cluster 2 | | |
| ID | GO ID | GO term | ID | GO ID | GO term |
| 267050_at | GO:0003723 | RNA binding | 264929_at | GO:0004033 | aldo-keto reductase activity |
| 257081_at | GO:0005515 | Protein binding | 262913_at | GO:0016491 | oxido-reductase activity |
| 264875_at | GO:0005515 | Protein binding | 267181_at | GO:0016491 | oxido-reductase activity |
| 257291_at | GO:0003677 | DNA binding | 264266_at | GO:0004722 | protein serine/threonine phosphatase activity |
| 260645_at | GO:0005488 | binding | 267627_at | GO:0008026 | ATP-dependent helicase activity |

Table 3.5

GO annotations for the three clusters in *Arabidopsis* dataset 2

| Data set 2 | | | | | |
|------------|------------|---|-----------|------------|--------------------------|
| Cluster 1 | | | Cluster 2 | | |
| ID | GO ID | GO term | ID | GO ID | GO term |
| 264168_at | GO:0030528 | transcription regulator activity | 257215_at | GO:0005515 | Protein binding |
| 266130_at | GO:0003677 | DNA binding | 267147_at | GO:0016491 | oxido-reductase activity |
| | | | 263135_at | GO:0016491 | oxido-reductase activity |
| | | | 266871_at | GO:0003723 | RNA binding |
| | | | 260638_at | GO:0005488 | binding |
| Cluster 3 | | | | | |
| ID | GO ID | GO term | | | |
| 264911_at | GO:0004033 | aldo-keto reductase activity | | | |
| 257050_at | GO:0004722 | protein serine/threonine phosphatase activity | | | |
| 257648_at | GO:0008026 | ATP-dependent helicase activity | | | |

Table 3.6

GO annotations for the two clusters in *Arabidopsis* combined datasets

| Data set combined | | | | | |
|-------------------|------------|-----------------|-----------|------------|---|
| Cluster 1 | | | Cluster 2 | | |
| ID | GO ID | GO term | ID | GO ID | GO term |
| 267050_at | GO:0003723 | RNA binding | 264929_at | GO:0004033 | aldo-keto reductase activity |
| 266871_at | GO:0003723 | RNA binding | 264911_at | GO:0004033 | aldo-keto reductase activity |
| 257215_at | GO:0005515 | Protein binding | 263135_at | GO:0016491 | oxido-reductase activity |
| 257081_at | GO:0005515 | Protein binding | 267147_at | GO:0016491 | oxido-reductase activity |
| 264875_at | GO:0005515 | Protein binding | 262913_at | GO:0016491 | oxido-reductase activity |
| 257291_at | GO:0003677 | DNA binding | 267181_at | GO:0016491 | oxido-reductase activity |
| 266130_at | GO:0003677 | DNA binding | 264266_at | GO:0004722 | protein serine/threonine phosphatase activity |
| 260645_at | GO:0005488 | binding | 257050_at | GO:0004722 | protein serine/threonine phosphatase activity |
| 260638_at | GO:0005488 | binding | 267627_at | GO:0008026 | ATP-dependent helicase activity |
| | | | 257648_at | GO:0008026 | ATP-dependent helicase activity |

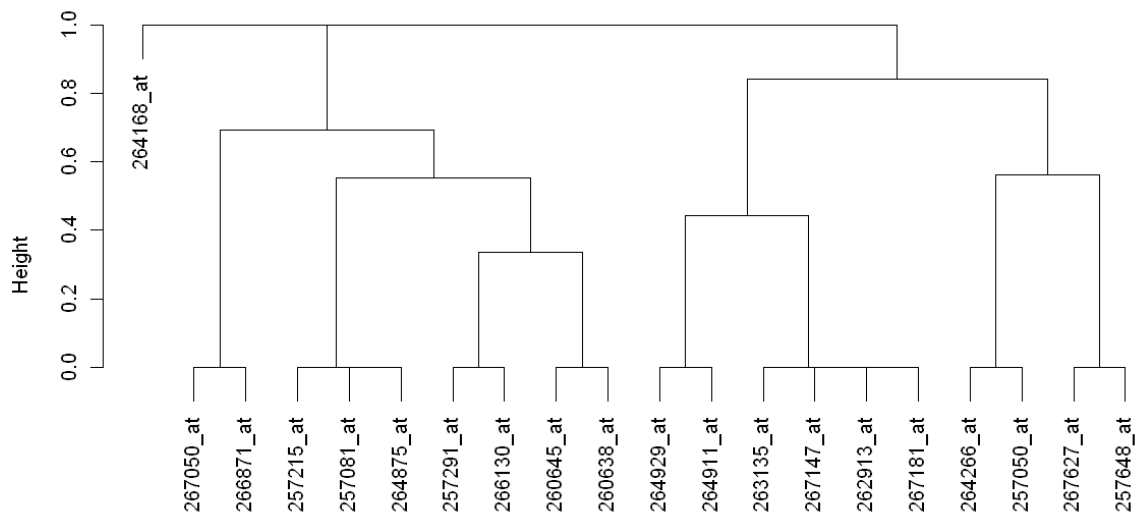


Figure 3.4

Cluster dendrogram for the combined *Arabidopsis* data set

CHAPTER 4

RESULTS AND EVALUATION

In Chapter 3 we described a new approach that we have developed for integrating multiple gene expression datasets. In this chapter we describe experiments we have designed and conducted to test the following hypothesis:

Integration of heterogeneous gene expression datasets by mapping at the functional level using a hierarchical clustering algorithm can provide additional useful biological information that cannot be easily obtained by mapping at the identifier level.

4.1 Experiments

We demonstrate our method by applying it to two different biological problems— protein expression during de-differentiation in *Arabidopsis* and gene expression in different corn lines upon infection by a fungus. In each case, we have two datasets available. We were unable to obtain gene expression and protein expression data for the same biological experiment. Instead, we use two proteomic data sets and two gene expression data sets. The same approach can also be applied to combine a gene expression data set with a protein expression data set.

4.1.1 *Arabidopsis* Experiment

Dr. Zhohua Peng provided us with two proteomics datasets from *Arabidopsis*. Analysis of these two datasets has been previously published [9]. The proteomics datasets represent up regulated proteins from a de-differentiation experiment in *Arabidopsis* where protein identification was done using two different technologies: shotgun proteomics and 2D gel electrophoresis. Cell de-differentiation is a process of switching the cell fate. During this process, cells undergo genome reprogramming to regain the competency of cell division and organ regeneration [9]. These proteins were chosen as an input to our experiment due to their availability and the author's familiarity with their data formats. Initially there were 193 *Arabidopsis* up regulated proteins identified by shotgun proteomics and 26 proteins up regulated identified by the 2DE gel approach. We mapped those proteins to *Arabidopsis* Affymetrix probe identifiers for input to the GOSim statistical package(www.dkfz.de/mga2/gosim). After the mapping, we obtained 95 differentially expressed proteins identified by shotgun proteomics and 20 differentially expressed proteins identified by 2DE gels. Of these proteins, only one protein was identified by both techniques. Therefore, little information for integration of the data sets is obtained by matching identifiers. From this particular experiment, biologists try to understand the reasons for recognizing different set of proteins using two different protein identification techniques in the same biological sample.

GO annotations for molecular function and biological process stored in GOSim were used to annotate these two different *Arabidopsis* sets of differentially expressed proteins.

We produced clusters based on the gene similarity of three different datasets: a set of pro-

teins identified by shotgun proteomics, and a set identified by 2DE gels, and the combined set. Protein similarity matrices were then computed using GOSim for each data set alone as well as for the combined data set.

Finally, each of the individual similarity matrices and the combined similarity matrix were used as input to the clustering algorithm. The hierarchical clustering algorithm we used is an agglomerative algorithm that builds the hierarchy from the individual elements by progressively merging clusters. We chose the complete linkage clustering method. Complete linkage computes the distance between two clusters as the maximum distance between any pair of elements in the clusters. The clustering dendrograms generated based on the similarity of GO molecular function annotation similarity are shown in Supp_Gel_up_mf_Arab.pdf, Supp_Shotgun_up_mf_Arab.pdf and Supp_Combined_up_mf_Arab.pdf (http://agbase.msstate.edu/Education/clt183_SuppFiles.pdf).

4.1.2 Maize Experiment

In our second experiment, we tested our hypothesis with two differentially expressed gene expression data sets from corn. In each set, the genes expressed in one maize line (Mp313E) were compared to the genes expressed in another maize line (Va35) when both were inoculated with the fungus *Aspergillus flavus*. Mp313E is considered to be resistant to infection by *Aspergillus* while Va35 is considered to be susceptible. Two *Maize* Unigene 1-1.05 arrays from the University of Arizona (www.maizearray.org) were used to evaluate differential expression. The first dataset using the MGDZ Zea Mays Unigene 1-1-05 maize microarray-GEO accession GPL6092 was conducted from a field experiment

in 2003 when samples were collected 2 days post infection. The second dataset using Maize Oligonucleotide Array version 4 was conducted from a field experiment in 2004 when samples were collected 4 days post infection. It is important to note that the first array has about 5000 probes while the second array contains about 32000 probes. The microarray for the 2-day post infection experiment contains a subset of the sequences on the array used for the 4-day post infection experiment. Analysis of the microarray for the 2-day post infection maize experiment resulted in 129 upregulated ESTs((Expressed Sequence Tag)) for Mp313E compared to VA35, and analysis of the microarray for the 4-day post infection corn experiment resulted in 234 upregulated ESTs. Then we obtained nucleotide sequences for those ESTs from www.ncbi.nlm.nih.gov and ran the BLAST algorithm on these EST sequences against *Arabidopsis* Affymetrix sequences in order to get the matching *Arabidopsis* probe identifiers to use as input to GOSim. BLAST resulted in 82 matching *Arabidopsis* probe IDs for the 2-day data and 203 *Arabidopsis* matching probe IDs for the 4-day data.

We then obtained biological process GO annotations for the Affymetrix probes and generated gene similarity matrices using GOSim for each of the individual data sets and for the combined data set created by combining the 2-day and 4-day data.

Then the gene similarity matrices were used as inputs for the clustering algorithm. Thus the resulting clusters are based on the functional similarity of genes. The dendrograms resulting from the clustering algorithm for the combined data set is as in [Supp_Combined_up_bp_Maize.pdf](http://agbase.msstate.edu/Education/clt183_SuppFiles.pdf) (http://agbase.msstate.edu/Education/clt183_SuppFiles.pdf).

4.2 Results and Analysis

In this section we discuss the results of applying our method to the *Arabidopsis* and Corn data sets. The clustering results were analyzed by our biology collaborators.

4.2.1 *Arabidopsis* Results Analysis

Dr. Zhaohua Peng from the Department of Biochemistry and Molecular Biology provided us with the *Arabidopsis* datasets and assisted us with the analysis. We produced the clusters based on highly expressed proteins identified using two different protein identification techniques, shotgun proteomics and 2D gel electrophoresis, in an *Arabidopsis* cell dedifferentiation experiment. We generated the clusters for the gel data set and proteomics data set alone as well as for the union of the two data sets. All clusters were generated based on the Gene Ontology Molecular Function.

The 2DE gel data had substantially fewer proteins. The resulting dendrogram for the gel data alone has four small, tight clusters as in Table 4.1. Mainly those clusters are formed based on the similarities of GO terms such as protein binding, nucleotide binding, and enzyme activities. The clusters in supplementary file Supp_gel_up_mf_Clusters_Arab.pdf (http://agbase.msstate.edu/Education/clt183_SuppFiles.pdf) has all set of gel clusters and those are labeled based on their position in the combined dendrogram. The gel dendrogram generated by the hierarchical algorithm is in Supp_Gel_up_mf_dendro_Arab.pdf (http://agbase.msstate.edu/Education/clt183_SuppFiles.pdf).

The dendrogram generated for the shotgun data alone has several clusters as in the Table 4.2. The largest cluster consist of 28 proteins (cluster 5) and formed based on the

Table 4.1

Clusters for *Arabidopsis* gel dataset

| Locus ID | Probe ID | Shotgun gel | GO ID | GO Term | Protein Identification |
|------------------|-------------|----------------|------------|---|--|
| Cluster 1 | | | | | |
| AT1G56330 | 256224_at | G | GO:0005525 | GTP binding | GTP-binding protein SAR1B |
| AT2G39730 | 245061_at | G | GO:0043531 | ADP binding | T5I7.3 (Hypothetical protein) |
| AT4G13850 | 254684_at | G | GO:0003723 | RNA binding | ATGRP2 (GLYCINE-RICH RNA-BINDING PROTEIN 2) |
| | | | GO:0003697 | single-stranded | DNA binding |
| | | | GO:0003690 | double-stranded | DNA binding |
| AT5G60390 | 247644_s_at | G | GO:0005524 | ATP binding | Putative translation elongation factor |
| | | | GO:0003746 | translation elongation factor activity | eEF-1 alpha chain (Gene A4) |

Table 4.1

Clusters for *Arabidopsis* gel dataset (continued)

| Locus ID | Probe ID | Shotgun gel | GO ID | GO Term | Protein Identification |
|------------------|-----------|----------------|------------|----------------------------|--|
| Cluster 2 | | | | | |
| AT5G42970 | 249175_at | G | GO:0005515 | protein binding | COP8 (Constitutive photomorphogenic) homolog (CSN complex subunit 4) |
| AT4G09000 | 255079_at | G | GO:0005515 | protein binding | F23J3_30 (14-3-3 protein GF14chi) (Grf1) |
| AT3G26650 | 257807_at | G | GO:0045309 | protein phosphorylated | |
| | | | GO:0008943 | amino acid binding | |
| | | | | glyceraldehyde-3-phosphate | Glyceraldehyde-3-phosphate dehydrogenase (NADP) (EC 1.2.1.13) |
| | | | | dehydrogenase activity | A precursor |
| AT5G10450 | 250439_at | G | GO:0005515 | protein binding | |
| | | | GO:0005515 | protein binding | 14-3-3 protein homolog RC12 |
| | | | GO:0045309 | protein phosphorylated | amino acid binding |
| AT1G32060 | 255720_at | G | GO:0005515 | protein binding | Phosphoribulokinase, chloroplast precursor |

Table 4.1

Clusters for *Arabidopsis* gel dataset (continued)

| Locus ID | Probe ID | Shotgun gel | GO ID | GO Term | Protein Identification |
|------------------|-----------|----------------|------------|---|---|
| | | | GO:0008974 | phosphoribulokinase | activity |
| AT2G18960 | 266939_at | G | GO:0005524 | ATP binding | Phosphoribulokinase, chloroplast precursor |
| | | | GO:0008553 | hydrogen-exporting ATPase activity, phosphorylative mechanism | V-type proton- ATPase |
| | | | GO:0016887 | ATPase activity | |
| | | | GO:0005515 | protein binding | |
| Cluster 3 | | | | | |
| AT1G21720 | 262497_at | G | GO:0008233 | peptidase activity | Proteasome subunit beta type 3-1 |
| AT5G42270 | 249244_at | G | GO:0016887 | ATPase activity | Cell division protein ftsH homolog 2, chloroplast precursor |
| Cluster 4 | | | | | |
| AT2G34590 | 266904_at | G | GO:0004739 | pyruvate dehydrogenase (acetyl-transferring) activity | Putative pyruvate dehydrogenase E1 beta subunit |
| | | | GO:0004802 | transketolase activity | |
| AT1G70580 | 260309_at | G | GO:0047958 | glycine transaminase activity | F26F24_4 |

Table 4.1

Clusters for *Arabidopsis* gel dataset (continued)

| Locus ID | Probe ID | Shotgun gel | GO ID | GO Term | Protein Identification |
|-----------|-------------|----------------|------------|--|---|
| | | | GO:0004021 | alanine transaminase activity | |
| AT1G74910 | 262174_at | G | GO:0016779 | nucleotidyltransferase activity | Putative GDP-mannose pyrophosphorylase (F9E10_24) |
| AT1G23820 | 265172_at | G | GO:0004766 | spermidine synthase activity | Spermidine synthase 1 |
| AT3G02230 | 259077_s_at | G | GO:0016760 | cellulose synthase (UDP-forming) activity | Reversibly glycosylated polypeptide-1 |

similarity of the GO term-structural constituent of ribosome. There are 3 little, very distinct clusters (cluster 1-3) formed based on the similarity of the GO terms such as nutrient reservoir activity, electron carrier activity and hydrogen ion transporting ATP synthase activity as shown in the Table 4.2. There are also some tight clusters from cluster 9-11 formed based on the GO term similarity of ATP binding, protein binding and calmodulin binding. All shotgun clusters generated based on Molecular Function GO annotation similarity are listed in the supplementary file Supp_shotgun_up_mf_Clusters_Arab.pdf (http://agbase.msstate.edu/Education/clt183_SuppFiles.pdf). Those clusters are labeled based on their position in the shotgun dendrogram in the supplementary file Supp_Shotgun_up_mf_dendro_Arab.pdf (http://agbase.msstate.edu/Education/clt183_SuppFiles.pdf).

The dendrogram for the combined data set maintains the same overall structure of clusters as in the dendrogram for the shotgun data set alone. This is probably due to the higher number of proteins identified by shotgun method compare to 2DE gel method. There are several clusters in the combined dendrogram which are exclusively formed of shotgun proteins as shown in supplementary file Supp_Combine_up_mf_Clusters_Arab.pdf (http://agbase.msstate.edu/Education/clt183_SuppFiles.pdf). For example, as in Table 4.3, clusters consist of tubulin proteins (cluster 4) and dehydrogenase family proteins (cluster 13) are not uniquely identified by 2DE gel. They all identified only by shotgun proteomics. One reason for having a small number of proteins identified by 2DE gel is due to a decision made by biologists during their dedifferentiation experiment. Although initially there were lots of differentially expressed proteins identified by 2DE gel, most of them were discarded because they are mixtures of multiple proteins. Therefore, the number of

Table 4.2

Clusters for *Arabidopsis* shotgun dataset

| Locus ID | Probe ID | Shotgun gel | GO ID | GO Term | Protein Identification |
|------------------|-------------|----------------|------------|--|---|
| Cluster 1 | | | | | |
| AT4G28520 | 253767_at | G,S | GO:0045735 | nutrient reservoir activity | CRU3 (CRUCIFERIN 3) |
| AT1G03880 | 265095_at | S | GO:0045735 | nutrient reservoir activity | CRU2 (CRUCIFERIN 2) |
| AT5G44120 | 249082_at | S | GO:0045735 | nutrient reservoir activity | CRA1 (CRUCIFERINA) |
| Cluster 2 | | | | | |
| AT2G27510 | 265649_at | S | GO:0009055 | electron carrier activity | ATFD3 (FERREDOXIN 3) |
| AT1G20340 | 255886_at | S | GO:0009055 | electron carrier activity | DRT112 (DNA-damage- repair/tolerance protein 112) |
| Cluster 3 | | | | | |
| AT1G76030 | 262684_s_at | S | GO:0046933 | hydrogen ion transporting ATP synthase activity, rotational mechanism | VACUOLAR ATP SYNTHASE SUBUNIT B1 |
| AT4G38510 | 252998_at | S | GO:0046933 | hydrogen ion transporting ATP synthase activity, rotational mechanism | VACUOLAR ATP SYNTHASE SUBUNIT B2 |

Table 4.2

Clusters for *Arabidopsis* shotgun dataset (continued)

| Locus ID | Probe ID | Shotgun gel | GO ID | GO Term | Protein Identification |
|------------------|-----------|----------------|------------|---------------------------------------|--|
| Cluster 5 | | | | | |
| AT3G49010 | 252294_at | S | GO:0003735 | structural constituent of ribosome | ATBBC1 (breast basic conserved 1) |
| AT4G09800 | 255000_at | S | GO:0003735 | structural constituent of ribosome | RPS18C (S18 RIBOSOMAL PROTEIN) |
| AT1G34030 | 255977_at | S | GO:0003735 | structural constituent of ribosome | 40S ribosomal protein S18 (RPS18B) |
| AT1G22780 | 264203_at | S | GO:0003735 | structural constituent of ribosome | PFL (POINTED FIRST LEAVES) |
| AT2G27710 | 266256_at | S | GO:0003735 | structural constituent of ribosome | 60S acidic ribosomal protein P2 (RPP2B) |
| AT1G04270 | 263667_at | S | GO:0003735 | structural constituent of ribosome | RPS15 (RIBOSOMAL PROTEIN S15) |
| AT5G09510 | 245886_at | S | GO:0003735 | structural constituent of ribosome | 40S ribosomal protein S15 (RPS15D) |
| AT5G09500 | 245883_at | S | GO:0003735 | structural constituent of ribosome | 40S ribosomal protein S15 (RPS15C) |
| AT1G78630 | 263131_at | S | GO:0003735 | structural constituent of ribosome | EMB1473 (EMBRYO DEFECTIVE 1473) |
| AT4G01310 | 255623_at | S | GO:0003735 | structural constituent of ribosome | ribosomal protein L5 family protein |
| AT3G53430 | 251938_at | S | GO:0003735 | structural constituent of ribosome | 60S ribosomal protein L12 (RPL12B) |

Table 4.2

Clusters for *Arabidopsis* shotgun dataset (continued)

| Locus ID | Probe ID | Shotgun gel | GO ID | GO Term | Protein Identification |
|-----------|-----------|----------------|------------|---------------------------------------|---|
| AT3G11510 | 259239_at | S | GO:0003735 | structural constituent of ribosome | 40S ribosomal protein S14 (RPS14B) |
| AT2G36160 | 263286_at | S | GO:0003735 | structural constituent of ribosome | 40S ribosomal protein S14 (RPS14A) |
| AT5G60670 | 247584_at | S | GO:0003735 | structural constituent of ribosome | 60S ribosomal protein L12 (RPL12C) |
| AT2G37190 | 265445_at | S | GO:0003735 | structural constituent of ribosome | 60S ribosomal protein L12 (RPL12A) |
| AT3G49910 | 252235_at | S | GO:0003735 | structural constituent of ribosome | 60S ribosomal protein L26 (RPL26A) |
| AT3G05560 | 259112_at | S | GO:0003735 | structural constituent of ribosome | 60S ribosomal protein L22-2 (RPL22B) |
| AT1G75350 | 261119_at | S | GO:0003735 | structural constituent of ribosome | EMB2184 (EMBRYO DEFECTIVE 2184) |
| AT4G00100 | 255706_at | S | GO:0003735 | structural constituent of ribosome | ATRPS13A (RIBOSOMAL PROTEIN S13A) |
| AT4G00100 | 255706_at | S | GO:0003735 | structural constituent of ribosome | ATRPS13A (RIBOSOMAL PROTEIN S13A) |
| AT5G10360 | 250440_at | S | GO:0003735 | structural constituent of ribosome | EMB3010 (EMBRYO x PROTEIN S13A) |
| AT4G10450 | 254980_at | S | GO:0003735 | structural constituent of ribosome | 60S ribosomal protein L9 (RPL90D) |

Table 4.2

Clusters for *Arabidopsis* shotgun dataset (continued)

| Locus ID | Probe ID | Shotgun gel | GO ID | GO Term | Protein Identification |
|-----------|-------------|----------------|------------|---------------------------------------|---------------------------------------|
| AT3G48960 | 252283_at | S | GO:0003735 | structural constituent of ribosome | 60S ribosomal protein L13 (RPL13C) |
| AT5G20290 | 246068_at | S | GO:0003735 | structural constituent of ribosome | 40S ribosomal protein S8 (RPS8A) |
| AT1G07320 | 261078_at | S | GO:0003735 | structural constituent of ribosome | RPL4 (ribosomal protein L4) |
| | | | GO:0008266 | poly(U) binding | |
| | | | GO:0003735 | structural constituent of ribosome | |
| AT3G58700 | 251552_at | S | GO:0003735 | structural constituent of ribosome | 60S ribosomal protein L11 (RPL11B) |
| AT4G18730 | 254617_s.at | S | GO:0003735 | structural constituent of ribosome | RPL16B (ribosomal protein L16B) |

Table 4.2

Clusters for *Arabidopsis* shotgun dataset (continued)

| Locus ID | Probe ID | Shotgun gel | GO ID | GO Term | Protein Identification |
|------------------|-------------|----------------|------------|---|--|
| Cluster 9 | | | | | |
| AT5G02500 | 250995_at | S | GO:0005524 | ATP binding | HSC70-1 (heat shock cognate 70 kDa protein |
| AT5G02490 | 250994_at | S | GO:0005524 | ATP binding | heat shock |
| AT5G28540 | 245956_s_at | S | GO:0005524 | ATP binding | cognate 70 kDa protein luminal binding protein 1 |
| AT3G12580 | 256245_at | S | GO:0005524 | ATP binding | HSP70 (heat shock protein 70 |
| AT1G16030 | 261838_at | S | GO:0005524 | ATP binding | HSP70B (heat shock protein 70B |
| AT4G09320 | 255089_at | S | GO:0004550 | nucleoside diphosphate kinase activity | NDPK1 (nucleoside diphosphate kinase 1) |
| AT5G56000 | 248043_s_at | S | GO:0005524 | ATP binding | heat shock protein 81-4 (HSP81-4) |
| AT4G22670 | 254275_at | S | GO:0005488 | binding | tetratricopeptide repeat (TPR)- containing protein |

Table 4.2

Clusters for *Arabidopsis* shotgun dataset (continued)

| Locus ID | Probe ID | Shotgun gel | GO ID | GO Term | Protein Identification |
|-------------------|-------------|----------------|------------|---|---|
| Cluster 10 | | | | | |
| AT3G57330 | 251649_at | S | GO:0005388 | calcium-transporting ATPase activity | calcium-transporting ATPase, plasma membrane -type, putative / Ca ²⁺ +ATPase |
| | | | GO:0005388 | calcium-transporting | ATPase activity |
| | | | GO:0005516 | calmodulin binding | |
| AT5G20010 | 246153_s_at | S | GO:0005515 | protein binding | RAN-1 (Ras-related GTP-binding nuclear protein 1) |
| | | | GO:0005525 | GTP binding | |
| | | | GO:0003924 | GTPase activity | |
| | | | GO:0005525 | GTP binding | |
| AT1G11740 | 262807_at | S | GO:0005515 | protein binding | ankyrin repeat family protein |
| AT3G15950 | 257798_at | S | GO:0005515 | protein binding | TSA1-LIKE |
| | | | GO:0005515 | protein binding | |
| Cluster 11 | | | | | |
| AT1G63940 | 260325_at | S | GO:0005524 | ATP binding | monodehydroascorbate reductase, putative |
| | | | GO:0005524 | ATP binding | |
| | | | GO:0005524 | ATP binding | |
| | | | GO:0005524 | ATP binding | |

Table 4.2

Clusters for *Arabidopsis* shotgun dataset (continued)

| Locus ID | Probe ID | Shotgun gel | GO ID | GO Term | Protein Identification |
|-----------|-------------|----------------|------------|--|---|
| AT5G02500 | 250995_at | S | GO:0005524 | ATP binding | HSC70-1 (heat shock cognate 70 kDa pro |
| AT5G02490 | 250994_at | S | GO:0005524 | ATP binding | heat shock cognate 70 kDa protein 2 |
| AT5G28540 | 245956_s_at | S | GO:0005524 | ATP binding | (HSC70-2) (HSP70-2) luminal binding |
| AT3G12580 | 256245_at | S | GO:0005524 | ATP binding | protein 1 (BiP-1) (BP1) HSP70 (heat shock protein 70) |
| AT1G16030 | 261838_at | S | GO:0005524 | ATP binding | HSP70B (heat shock protein 70B) |
| AT4G09320 | 255089_at | S | GO:0004550 | nucleoside diphosphate kinase activity | NDPK1 (nucleoside diphosphate kinase 1) |
| AT5G56000 | 248043_s_at | S | GO:0005524 | ATP binding | heat shock protein 81-4 (HSP81-4) |

differentially expressed proteins identified using the 2DE gel approach was small. Meanwhile, shotgun method identified a lot more proteins overall. Cluster 5 as in Table 4.3 is a large cluster which contains ribosomal proteins and all of them are identified by shotgun proteomics except one. The only gel protein in cluster 5 is an expressed protein, and it was not included in a cluster in the dendrogram generated only for gel data. But in the combined dendrogram, it was clustered with other ribosomal proteins. This is one good example of the advantage of combining the data sets. Once the data sets are combined, they form bigger, more meaningful clusters which reveal more useful biological information. The dendrogram generated for the combined data set is in the supplementary file Supp_Combined_up_mf_dendro_Arab.pdf (http://agbase.msstate.edu/Education/clt183_SuppFiles.pdf).

Most of the proteins found in small, highly associated clusters in 2DE gel dendrogram remained together in the combined dendrogram. Some of them are mixed with the proteins identified by shotgun proteomics in a reasonable way to form bigger, meaningful clusters in the combined dendrogram. For example, cluster 9 as in Table 4.3 in the combined dendrogram was formed based on GO terms such as protein binding and calmodulin binding. This is a mixture of both gel and shotgun proteins, but predominantly gel proteins. These types of binding proteins are highly abundant in the cell and have many close gene family members. High sequence similarity among these proteins makes the shotgun identification inaccurate due to common peptides of multiple proteins. Therefore 2DE gel had the advantage of identifying these type of binding proteins. Combined cluster 10 as in Table 4.3 is also an example for a mixture of proteins identified by each technique. It mostly

Table 4.3

Clusters for *Arabidopsis* combined dataset

| Locus ID | Probe ID | Shotgun gel | GO ID | GO Term | Protein Identification |
|------------------|-------------|----------------|------------|---|--------------------------------|
| Cluster 4 | | | | | |
| AT1G20010 | 261230_at | S | GO:0005200 | structural constituent of cytoskeleton | TUB5 (tubulin beta-5 chain) |
| AT5G44340 | 249049_at | S | GO:0005200 | structural constituent of cytoskeleton | TUB4 (tubulin beta-4 chain) |
| AT5G09810 | 250458_s.at | S | GO:0005200 | structural constituent of cytoskeleton | ACT7 (actin 7) |
| AT5G59370 | 247736_at | S | GO:0005200 | structural constituent of cytoskeleton | ACT4 (ACTIN 4) |
| AT2G29550 | 266295_at | S | GO:0005200 | structural constituent of cytoskeleton | TUB7 (tubulin beta-7 chain) |
| AT5G23860 | 249818_at | S | GO:0005200 | structural constituent of cytoskeleton | TUB8 (tubulin beta-8) |

Table 4.3

Clusters for *Arabidopsis* combined dataset (continued)

| Locus ID | Probe ID | Shotgun gel | GO ID | GO Term | Protein Identification |
|------------------|-----------|----------------|------------|---------------------------------------|--|
| Cluster 5 | | | | | |
| AT1G73850 | 260382_at | G | GO:0003735 | structural constituent of ribosome | Expressed protein |
| AT3G49010 | 252294_at | S | GO:0003735 | structural constituent of ribosome | ATBBC1 (breast basic conserved 1) |
| AT4G09800 | 255000_at | S | GO:0003735 | structural constituent of ribosome | RPS18C (S18 RIBOSOMAL PROTEIN) |
| AT1G34030 | 255977_at | S | GO:0003735 | structural constituent of ribosome | 40S ribosomal protein S18 (RPS18B) |
| AT1G22780 | 264203_at | S | GO:0003735 | structural constituent of ribosome | PFL (POINTED FIRST LEAVES) |
| AT2G27710 | 266256_at | S | GO:0003735 | structural constituent of ribosome | 60S acidic ribosomal protein P2 (RPP2B) |
| AT1G04270 | 263667_at | S | GO:0003735 | structural constituent of ribosome | RPS15 (RIBOSOMAL PROTEIN S15) |
| AT5G09510 | 245886_at | S | GO:0003735 | structural constituent of ribosome | 40S ribosomal protein S15 (RPS15D) |

Table 4.3

Clusters for *Arabidopsis* combined dataset (continued)

| Locus ID | Probe ID | Shotgun gel | GO ID | GO Term | Protein Identification |
|-----------|-----------|----------------|------------|---------------------------------------|---|
| AT5G09500 | 245883_at | S | GO:0003735 | structural constituent of ribosome | 40S ribosomal protein S15 (RPS15C) |
| AT1G78630 | 263131_at | S | GO:0003735 | structural constituent of ribosome | EMB1473 (EMBRYO DEFECTIVE 1473) |
| AT4G01310 | 255623_at | S | GO:0003735 | structural constituent of ribosome | ribosomal protein L5 family protein |
| AT3G53430 | 251938_at | S | GO:0003735 | structural constituent of ribosome | 60S ribosomal protein L12 (RPL12B) |
| AT3G11510 | 259239_at | S | GO:0003735 | structural constituent of ribosome | 40S ribosomal protein S14 (RPS14B) |
| AT2G36160 | 263286_at | S | GO:0003735 | structural constituent of ribosome | 40S ribosomal protein S14 (RPS14A) |
| AT5G60670 | 247584_at | S | GO:0003735 | structural constituent of ribosome | 60S ribosomal protein L12 (RPL12C) |
| AT2G37190 | 265445_at | S | GO:0003735 | structural constituent of ribosome | 60S ribosomal protein L12 (RPL12A) |
| AT3G49910 | 252235_at | S | GO:0003735 | structural constituent of ribosome | 60S ribosomal protein L26 (RPL26A) |
| AT3G05560 | 259112_at | S | GO:0003735 | structural constituent of ribosome | 60S ribosomal protein L22-2 (RPL22B) |
| AT1G75350 | 261119_at | S | GO:0003735 | structural constituent of ribosome | EMB2184 (EMBRYO DEFECTIVE 2184) |

Table 4.3

Clusters for *Arabidopsis* combined dataset (continued)

| Locus ID | Probe ID | Shotgun gel | GO ID | GO Term | Protein Identification |
|-----------------|-----------------|------------------------|--------------|------------------------------------|------------------------------------|
| AT4G00100 | 255706_at | S | GO:0003735 | structural constituent of ribosome | ATRPS13A (RIBOSOMAL PROTEIN S13A) |
| AT4G00100 | 255706_at | S | GO:0003735 | structural constituent of ribosome | ATRPS13A (RIBOSOMAL PROTEIN S13A) |
| AT5G10360 | 250440_at | S | GO:0003735 | structural constituent of ribosome | EMB3010 (EMBRYO DEFECTIVE 3010) |
| AT4G10450 | 254980_at | S | GO:0003735 | structural constituent of ribosome | 60S ribosomal protein L9 (RPL90D) |
| AT3G48960 | 252283_at | S | GO:0003735 | structural constituent of ribosome | 60S ribosomal protein L13 (RPL13C) |
| AT5G20290 | 246068_at | S | GO:0003735 | structural constituent of ribosome | 40S ribosomal protein S8 (RPS8A) |
| AT1G07320 | 261078_at | S | GO:0003735 | structural constituent of ribosome | RPL4 (ribosomal protein L4) |
| | | | GO:0008266 | poly(U) binding | |
| | | | GO:0003735 | structural constituent of ribosome | |
| AT3G58700 | 251552_at | S | GO:0003735 | structural constituent of ribosome | 60S ribosomal protein L11 (RPL11B) |
| AT4G18730 | 254617_s.at | S | GO:0003735 | structural constituent of ribosome | RPL16B (ribosomal protein L16B) |

Table 4.3

Clusters for *Arabidopsis* combined dataset (continued)

| Locus ID | Probe ID | Shotgun gel | GO ID | GO Term | Protein Identification |
|------------------|-------------|----------------|--|--|---|
| Cluster 9 | | | | | |
| AT5G60390 | 247644_s_at | G | GO:0003746 | translation elongation factor activity | Putative translation elongation factor eEF-1 alpha chain (Gene A4) |
| AT3G57330 | 251649_at | S | GO:0005516 GO:0005388 | calmodulin binding calcium-transporting ATPase activity | calcium-transporting ATPase, plasma membrane-type, putative / Ca2+ATPase |
| AT5G42970 | 249175_at | G | GO:0005388 GO:0005516 GO:0005515 | calcium-transporting ATPase activity calmodulin binding protein binding | COP8 (Constitutive photomorphogenic) homolog (CSN complex subunit 4) F23J3_30 (14-3-3 protein GF14chi) (Grf1) |
| AT4G09000 | 255079_at | G | GO:0005515 | protein binding | |
| AT3G26650 | 257807_at | G | GO:0045309 GO:0008943 | protein phosphorylated amino acid binding glyceraldehyde-3-phosphate dehydrogenase activity | Glyceraldehyde-3-phosphate (EC 1.2.1.13 dehydrogenase (NADP)) A precursor |
| | | | GO:0005515 | protein binding | |

Table 4.3

Clusters for *Arabidopsis* combined dataset (continued)

| Locus ID | Probe ID | Shotgun gel | GO ID | GO Term | Protein Identification |
|-----------|-------------|----------------|--------------------------|---|---|
| AT5G10450 | 250439_at | G | GO:0005515 GO:0045309 | protein binding protein phosphorylated amino acid binding | 14-3-3 protein homolog RCI2 |
| AT1G32060 | 255720_at | G | GO:0005515 | protein binding | Phosphoribulokinase, chloroplast precursor (EC 2.7.1.19) (Phosphopentokinase) |
| | | | GO:0008974 | phosphoribulokinase activity | |
| | | | GO:0005524 | ATP binding | Phosphoribulokinase, chloroplast precursor (EC 2.7.1.19) (Phosphopentokinase) |
| AT5G20010 | 246153_s_at | S | GO:0005515 | protein binding | RAN-1 (Ras-related GTP-binding nuclear protein 1) |
| | | | GO:0005525 GO:0003924 | GTP binding GTPase activity | |
| | | | GO:0005525 | GTP binding | |
| AT1G11740 | 262807_at | S | GO:0005515 | protein binding | ankyrin repeat family protein TSA1-LIKE |
| AT3G15950 | 257798_at | S | GO:0005515 GO:0005515 | protein binding protein binding | |

Table 4.3

Clusters for *Arabidopsis* combined dataset (continued)

| Locus ID | Probe ID | Shotgun gel | GO ID | GO Term | Protein Identification |
|-------------------|-----------|----------------|------------|---|---|
| Cluster 10 | | | | | |
| AT1G63940 | 260325_at | S | GO:0005524 | ATP binding | monodehydroascorbate reductase, putative |
| | | | GO:0005524 | ATP binding | |
| | | | GO:0005524 | ATP binding | |
| AT1G56330 | 256224_at | G | GO:0005525 | GTP binding | GTP-binding protein SAR1B |
| AT2G39730 | 245061_at | G | GO:0043531 | ADP binding | T5I7.3 (Hypothetical protein) |
| | | | GO:0046863 | ribulose-1,5-bisphosphate carboxylase/oxygenase activase activity | |
| | | | GO:0030234 | enzyme regulator activity | |
| AT5G02500 | 250995_at | S | GO:0005524 | ATP binding | HSC70-1 (heat shock cognate 70 kDa protein) |
| AT5G02490 | 250994_at | S | GO:0005524 | ATP binding | heat shock cognate 70 kDa protein 2 (HSC70-2) (HSP70-2) |

Table 4.3

Clusters for *Arabidopsis* combined dataset (continued)

| Locus ID | Probe ID | Shotgun gel | GO ID | GO Term | Protein Identification |
|-----------|-------------|----------------|------------|---|--|
| AT5G28540 | 245956_s_at | S | GO:0005524 | ATP binding | luminal binding protein 1 (BiP-1) (BPI) |
| AT3G12580 | 256245_at | S | GO:0005524 | ATP binding | HSP70 (heat shock protein) |
| AT1G16030 | 261838_at | S | GO:0005524 | ATP binding | HSP70B (heat shock protein 70B) |
| AT4G09320 | 255089_at | S | GO:0004550 | nucleoside diphosphate kinase activity | NDPK1 (nucleoside diphosphate kinase 1) |
| AT5G56000 | 248043_s_at | S | GO:0005524 | ATP binding | heat shock protein 81-4 (HSP81-4) |

Table 4.3

Clusters for *Arabidopsis* combined dataset (continued)

| Locus ID | Probe ID | Shotgun gel | GO ID | GO Term | Protein Identification |
|-------------------|-----------|----------------|------------|---|---|
| Cluster 11 | | | | | |
| AT1G09780 | 264668_at | S | GO:0046537 | 2,3-bisphospho- glycerate independent phosphoglycerate mutase activity | 2,3-biphospho- glycerate-independent phosphoglycerate mutase, putative |
| AT2G34590 | 266904_at | G | GO:0004739 | pyruvate dehydrogenase (acetyl-transferring) activity | Putative pyruvate dehydro- genase E1 beta subunit |
| AT5G19550 | 245951_at | S | GO:0004802 | transketolase activity | ASP2 (ASPARTATE AMINOTRANSFERASE 2) |
| AT1G70580 | 260309_at | G | GO:0004069 | aspartate transaminase activity | F26F24_4 |
| | | | GO:0047958 | glycine transaminase activity | |
| | | | GO:0004021 | alanine transaminase activity | |
| AT1G74910 | 262174_at | G | GO:0016779 | transaminase activity nucleotidyltransferase activity | Putative GDP-mannose pyrophosphorylase (F9E10_24) |
| AT3G52930 | 252022_at | S | GO:0003824 | catalytic activity | fructose-bisphosphate aldolase, putative |
| | | | GO:0004332 | fructose-bisphosphate | aldolase activity |

Table 4.3

Clusters for *Arabidopsis* combined dataset (continued)

| Locus ID | Probe ID | Shotgun gel | GO ID | GO Term | Protein Identification |
|-----------|-------------|----------------|------------|--|---|
| AT3G12290 | 256263_at | S | GO:0003824 | catalytic activity | tetrahydrofolate dehydrogenase/ cyclohydrolase, putative |
| AT4G34200 | 253274_at | S | GO:0005524 | nucleotide binding | EDA9 (embryo sac development arrest 9) |
| AT1G23820 | 265172_at | G | GO:0004766 | spermidine synthase activity | Spermidine synthase 1 (EC 2.5.1.16) |
| AT5G20980 | 246185_at | S | GO:0003871 | 5-methyltetrahydro- pteroyl/triglutamate- homocysteine S-methyl- transferase activity | ATMS3 (METHIONINE SYNTHASE 3) |
| | | | GO:0008705 | methionine synthase activity | |
| AT3G02230 | 259077_s_at | G | GO:0016760 | cellulose synthase (UDP-forming) activity | Reversibly glycosylated polypeptide-1 |
| AT4G23100 | 254270_at | S | GO:0004357 | glutamate-cysteine ligase activity | RML1 (PHYTOALEXIN DEFICIE- NT 2, ROOT MERISTEMLESS 1) |
| | | | GO:0004357 | glutamate-cysteine ligase activity | |
| | | | GO:0004357 | glutamate-cysteine ligase activity | |

Table 4.3

Clusters for *Arabidopsis* combined dataset (continued)

| Locus ID | Probe ID | Shotgun gel | GO ID | GO Term | Protein Identification |
|-----------|-------------|----------------|------------|---|---|
| AT3G12290 | 256263_at | S | GO:0003824 | catalytic activity | tetrahydrofolate dehydrogenase/ cyclohydrolase, putative |
| AT4G34200 | 253274_at | S | GO:0005524 | nucleotide binding | EDA9 (embryo sac development arrest 9) |
| AT1G23820 | 265172_at | G | GO:0004766 | spermidine synthase activity | Spermidine synthase 1 (EC 2.5.1.16) |
| AT5G20980 | 246185_at | S | GO:0003871 | 5-methyltetrahydro- pteroyltriglutamate- homocysteine S-methyl- transferase activity | ATMS3 (METHIONINE SYNTHASE 3) |
| | | | GO:0008705 | methionine synthase activity | |
| AT3G02230 | 259077_s_at | G | GO:0016760 | cellulose synthase (UDP-forming) activity | Reversibly glycosylated polypeptide-1 |
| AT4G23100 | 254270_at | S | GO:0004357 | glutamate-cysteine ligase activity | RML1 (PHYTOALEXIN DEFICIENT 2, ROOT MERISTEMLESS 1) |
| | | | GO:0004357 | glutamate-cysteine ligase activity | |
| | | | GO:0004357 | glutamate-cysteine ligase activity | |

Table 4.3

Clusters for *Arabidopsis* combined dataset (continued)

| Locus ID | Probe ID | Shotgun gel | GO ID | GO Term | Protein Identification |
|-------------------|-------------|----------------|------------|---|---|
| Cluster 13 | | | | | |
| AT1G59900 | 262908_at | S | GO:0004739 | pyruvate dehydrogenase (acetyl-transferring) activity | AT-E1 ALPHA (pyruvate dehydrogenase complex E1 alpha subunit) |
| AT1G19570 | 261149_s_at | S | GO:0045174 | glutathione dehydrogenase (ascorbate) activity | DHAR1 (DEHYDROASCORBATE REDUCTASE) |
| | | | GO:0005507 | copper ion binding | |
| | | | GO:0045174 | glutathione dehydrogenase (ascorbate) activity | |
| AT4G08390 | 255142_at | S | GO:0016688 | L-ascorbate peroxidase activity | SAPX |
| | | | GO:0016688 | L-ascorbate peroxidase activity | |

Table 4.3

Clusters for *Arabidopsis* combined dataset (continued)

| Locus ID | Probe ID | Shotgun gel | GO ID | GO Term | Protein Identification |
|-----------|-----------|----------------|------------|--|---|
| AT1G08830 | 264809_at | S | GO:0004784 | superoxide dismutase activity | CSD1 (copper/ zinc superoxide dismutase 1) |
| | | | GO:0004784 | superoxide dismutase activity | |
| | | | GO:0004784 | superoxide dismutase activity | |
| AT2G28190 | 266165_at | S | GO:0004784 | superoxide dismutase activity | CSD2 (COPPER/ ZINC SUPEROXIDE DISMUTASE 2) |
| AT5G41670 | 249266_at | S | GO:0004616 | phosphogluconate dehydrogenase (decarboxylating) activity | 6-phosphogluconate dehydrogenase family protein |
| AT5G43330 | 249147_at | S | GO:0016615 | malate dehydrogenase activity | malate dehydrogenase, cytosolic, putative |
| AT3G47520 | 252407_at | S | GO:0016615 | malate dehydrogenase activity | MDH (malate dehydrogenase) |

consists of heat shock proteins which are stress related proteins. The reason for having a lot of heat shock proteins could be the stress in the plant cells during the process of cell dedifferentiation as it was induced by high levels of hormones, which exceeded the growth inhibition concentration. At the same time, the tissues were excised from the plants to induce dedifferentiation, which was also a stress. Alternatively, a large number of proteins are synthesized during cell dedifferentiation, the heat shock proteins may be involved in protein folding. We can derive more biological information from the combined clusters rather than looking at the clusters in individual dendrograms for each data set.

Cluster 11 as in Table 4.3 is also another prominent cluster containing mixture of proteins. That cluster is formed based on the GO terms related to enzymic activity, and it also consists of proteins identified by both identification techniques.

There are few distinct clusters formed in the combined dendrogram which are not present in either of the individual dendrograms such as for the GO term chlorophyll binding. These new information help biologists to explore more aspects about the biological system.

4.2.2 Corn Results Analysis

The results of clustering the maize datasets have been analyzed by our collaborators. Both clusters based on the Gene Ontology Molecular Function and on Biological Process were generated. Those based on Biological Process proved to be most useful to the biologists for analysis. The USDA Corn Host Plant Resistance Laboratory developed the resistant maize line Mp313E and they also generated the gene expression data used in our study.

We produced clusters based on the genes that are significantly more highly expressed in the resistant line Mp313E compared to the susceptible line Va35 upon inoculation with *Aspergillus flavus* at a 2-day and a 4-day time point. Many genes known to be involved in response to stress were found in the up-regulated set.

The biologists found the combined clustering to be very informative in conveying the biological processes at work in the resistant line upon infection. We provide a short summary of their analysis of selected clusters. The clusters in supplementary file Supp_Combine_up_bp_Clusters_Maize.pdf (http://agbase.msstate.edu/Education/clt183_SuppFiles.pdf) has all the set of Maize clusters and those are labeled based on their position in the combined dendrogram. The combined dendrogram generated by the hierarchical algorithm is in Supp_Combined_up_bp_dendro_Maize.pdf (http://agbase.msstate.edu/Education/clt183_SuppFiles.pdf). Cluster 6 as in Table 4.4 contains four genes (one from the 2-day set and three from the 4-day set) that are involved in cell signaling. It is clear from this cluster and from several others that cells in the resistant infected plants are actively signaling other cells. Cluster 7 as in Table 4.4 contains only one gene, and we have typically ignored one-gene clusters. However, this gene was up-regulated in both the 2-day and 4-day datasets. The gene in this cluster is involved in autophagy, the process by which the cell breaks down its own components for reuse [32]. Autophagy is also known to play a protective role against infection by causing cell death at the infection site, preventing its spread into uninfected tissue [40]. Cluster 8 as in Table 4.4 has two genes from the 4-day dataset that both contribute to vacuole organization and acidification. An acidic pH in the vacuole is essential for protease activity (breaking down proteins) and protease activity is critical for

Table 4.4

Clusters for Maize combine dataset

| Locus ID | Maize ID | 2d/4dn | GO ID | GO Term | Protein Identification |
|------------------|------------|--------|------------|---|---|
| Cluster 6 | | | | | |
| At4g34920 | AW447878 | 2d | GO:0019432 | triglyceride | 1-phosphatidylinositol |
| | | | GO:0006629 | biosynthetic process lipid metabolic process | phosphodiesterase-related |
| | | | GO:0007242 | intracellular signaling cascade | |
| | | | GO:0008654 | phospholipid biosynthetic process | |
| At1g10210 | TC220557 | 4d | GO:0009734 | auxin mediated signaling pathway | Encodes ATPK1. |
| At4g03010 | BQ538143 | 4d | GO:0007165 | signal transduction | |
| At1g08340 | AZM4_91291 | 4d | GO:0007165 | signal transduction | Leucine-rich repeat family protein rac GTPase activating protein, putative |
| Cluster 7 | | | | | |
| At1g62040 | TC222043 | 4d,2d | GO:0006914 | autophagy | autophagy 8c (ATG8C) |

Table 4.4

Clusters for Maize combine dataset (continued)

| Locus ID | Maize ID | 2d/4dn | GO ID | GO Term | Protein Identification |
|------------------|----------|--------|------------|---------------------------------------|---|
| Cluster 8 | | | | | |
| At1g17260 | BG458764 | 4d | GO:0010023 | proanthocyanidin biosynthetic process | Belongs to H+-ATPase gene family involved in proanthocyanidin biosynthesis disturbs the vacuolar biogenesis and acidification process Homologous to yeast. VPS11. |
| | | | GO:0007035 | vacuolar acidification | Forms a complex with VCL1 and AtVPS33. Involved in vacuolar biogenesis |
| | | | GO:0007033 | vacuole organization | tubulin 3 process movement |
| At2g05170 | TC227930 | 4d | GO:0007033 | vacuole organization | Actin-depolymerizing factor (ADF) and cofilin define a family of actin-binding proteins essential for the rapid turnover of filamentous actin in vivo. |
| Cluster 9 | | | | | |
| At5g19770 | TC236810 | 4d | GO:0051258 | protein polymerization | |
| | | | GO:0007017 | microtubule-based | |
| | | | GO:0007018 | microtubule-based | |
| At3g46010 | BE012243 | 2d | GO:0007015 | actin filament organization | |

Table 4.4

Clusters for Maize combine dataset (continued)

| Locus ID | Maize ID | 2d/4dn | GO ID | GO Term | Protein Identification |
|-------------------|-----------------|---------------|--------------|--|--|
| At5g63800 | TC243912 | 4d | GO:0009827 | plant-type cell wall modification | Involved in mucilage formation |
| | | | GO:0048354 | mucilage biosynthetic process during seed coat development | |
| At3g46030 | CF628166 | 4d | GO:0006334 | nucleosome assembly | HTB11 |
| At5g54960 | TC223978 | 4d | GO:0001666 | response to hypoxia | pyruvate decarboxylase-2 |
| At2g33740 | AI855238 | 2d | GO:0010038 | response to metal ion | Copper binding protein that forms tetramers in vitro. |
| Cluster 10 | | | GO:0010038 | response to metal ion | |
| At5g65940 | AW787410 | 2d | GO:0009733 | response to auxin stimulus | hydrolyzes beta-hydroxyisobutyryl-CoA |
| | | | GO:0006635 | fatty acid beta-oxidation | |
| | | | GO:0006574 | valine catabolic process | |
| At3g23050 | AZM4_79559 | 2d | GO:0040008 | regulation of growth | Transcription regulator acting as repressor of auxin-inducible |

Table 4.4

Clusters for Maize combine dataset (continued)

| Locus ID | Maize ID | 2d/4dn | GO ID | GO Term | Protein Identification |
|-----------|------------|--------|------------|--|---|
| | | | GO:0009753 | response to jasmonic acid stimulus | gene expression. |
| | | | GO:0009611 | response to wounding | Plays role in the control of gravitropic growth and development in light-grown seedlings. |
| | | | GO:0009630 | gravitropism | |
| | | | GO:0009414 | response to water deprivation | |
| | | | GO:0009733 | response to auxin stimulus | |
| At2g04550 | AW400101 | 2d | GO:0009737 | response to abscisic acid stimulus | dual specificity protein phosphatase family protein |
| | | | GO:0009733 | response to auxin stimulus | |
| | | | GO:0007243 | protein kinase cascade | |
| | | | GO:0043407 | negative regulation of MAP kinase activity | |
| At5g09810 | AZM4_35410 | 4d | GO:0048364 | root development | Member of Actin gene family. |
| | | | GO:0009733 | response to auxin stimulus | |

Table 4.4

Clusters for Maize combine dataset (continued)

| Locus ID | Maize ID | 2d/4dn | GO ID | GO Term | Protein Identification |
|-----------------|-----------------|---------------|--------------|---|---------------------------------------|
| | | | GO:0009611 | response to wounding | |
| | | | GO:0010053 | root epidermal cell differentiation | |
| | | | GO:0048767 | root hair elongation | |
| | | | GO:0009845 | seed germination | |
| | | | GO:0007010 | cytoskeleton organization | |
| | | | GO:0048364 | root development | |
| | | | GO:0009416 | response to light stimulus | |
| | | | GO:0051301 | cell division | |
| At5g19140 | TC235519 | 4d | GO:0009733 | response to auxin stimulus | AILP1 |
| | | | GO:0010044 | response to aluminum ion | |
| At1g71230 | TC245291 | 4d | GO:0009640 | photomorphogenesis | Encodes a subunit of the COP9 complex |
| | | | GO:0010100 | negative regulation of photomorphogenesis | |
| | | | GO:0000338 | protein deneddylation | |
| | | | GO:0010387 | signalosome assembly | |
| | | | GO:0009733 | response to auxin stimulus | |

Table 4.4

Clusters for Maize combine dataset (continued)

| Locus ID | Maize ID | 2d/4dn | GO ID | GO Term | Protein Identification |
|-------------------|----------|--------|------------|------------------------------------|---|
| | | | GO:0000085 | G2 phase of mitotic cell cycle | |
| At2g28085 | TC243001 | 4d | GO:0009733 | response to auxin stimulus | auxin-responsive family protein |
| At1g15050 | TC223257 | 4d | GO:0009733 | response to auxin stimulus | Belongs to auxin inducible gene family. |
| Cluster 11 | | | | | |
| At3g02850 | TC226652 | 4d | GO:0006813 | potassium ion transport | member of Stellar K ⁺ outward rectifying channel (SKOR) family. |
| | | | GO:0009737 | response to abscisic acid stimulus | Mediates the delivery of K ⁺ from stelar cells to the xylem in the roots towards the shoot. mRNA accumulation is modulated by abscisic acid. K ⁺ gating activity is modulated by external and internal K ⁺ . |

Table 4.4

Clusters for Maize combine dataset (continued)

| Locus ID | Maize ID | 2d/4dn | GO ID | GO Term | Protein Identification |
|-------------------|----------|--------|------------|--|--|
| Cluster 12 | | | | | |
| At2g17290 | AW927389 | 2d | GO:0010119 | regulation of stomatal movement | Encodes calcium dependent protein kinase 6 (CPK6). |
| | | | GO:0006499 | N-terminal protein myristoylation | CDPKs protein belongs to auxin inducible gene family. |
| | | | GO:0009738 | abscisic acid | |
| | | | GO:0006468 | protein amino acid phosphorylation | mediated signaling |
| | | | GO:0010359 | regulation of anion channel activity | |
| At1g64060 | TC222718 | 4d | GO:0006800 | oxygen and reactive oxygen species metabolic process | Interacts with AtrbohD gene to fine tune the spatial control of ROI production and hypersensitive response to cell in and around infection site. |
| | | | GO:0002679 | respiratory burst during defense response | |
| | | | GO:0010119 | regulation of stomatal movement | |

Table 4.4

Clusters for Maize combine dataset (continued)

| Locus ID | Maize ID | 2d/4dn | GO ID | GO Term | Protein Identification |
|-----------------|-----------------|---------------|--------------|--|--|
| | | | GO:0050665 | hydrogen peroxide biosynthetic process | |
| | | | GO:0009738 | abscisic acid mediated signaling | |
| | | | GO:0006952 | defense response | |
| | | | GO:0043069 | negative regulation of programmed cell death | |
| | | | GO:0009873 | ethylene mediated signaling pathway | |
| | | | GO:0052542 | callose deposition during defense response | |
| | | | GO:0009723 | response to ethylene stimulus | |
| At3g57530 | TC238395 | 4d | GO:0006499 | N-terminal protein myristoylation | Calcium-dependent Protein Kinase. ABA signaling component that regulates the ABA-responsive gene expression via ABF4. AtCPK32 has autophosphorylation activity and can phosphorylate ABF4 in vitro |
| | | | GO:0009738 | abscisic acid mediated signaling | |
| | | | GO:0009651 | response to salt stress | |
| | | | GO:0006468 | protein amino acid phosphorylation | |

disease resistance [52]. Fungal infection leads to acidification of the vacuole and activation of protease enzyme activity [47]. Cluster 9 as in Table 4.4 is a mixture of genes from the 2-d and 4-d datasets that are involved in microtubule formation. Microtubules play key roles in intracellular transport, cell wall synthesis and in the adaptive response of plants to pathogen infection [28]. Cluster 10 as in Table 4.4 is a group of genes from both the 2-day and 4-day datasets involved in response to auxin stimulus. Auxin is a hormone produced by both plants and some fungi including *Aspergillus flavus* [15]. Therefore, it seems likely that these genes are activated in corn in response to auxin produced by the fungi. One of the genes specifically represses auxin-induced gene expression. Clusters 11 and 12 as in Table 4.4 are involved in regulation of stomatal movement. The openings on leaf surfaces used for gas exchange are called stomata. These provide an easy point of entry for an invading fungus and the maize plant may be reacting to the infection by closing the stomata. Many of the other clusters involve genes that have been implicated in previous research in providing defense mechanisms for plants. Thus, by combining the two datasets at the functional level, the biologists are able to gain a more comprehensive view of the biological processes that are activated in the resistant maize line upon inoculation with the fungus.

CHAPTER 5

CONCLUSION AND FUTURE WORK

This chapter summarizes the findings of the results obtained by integrating heterogeneous data sets at the functional level using a hierarchical algorithm. Directions of future research are also discussed in terms of possible enhancements and additional experiments that can be performed.

5.1 Summary of Results

We developed a method to integrate heterogeneous data sets by mapping at the functional level using a hierarchical clustering algorithm. In our method, Gene Ontology annotations are obtained for each dataset and the datasets are combined. The distance between all genes/proteins in the combined set is computed based on their GO similarity. GO similarity is computed using an information theoretic approach described by Resnik [45] and implemented in the GOSim package (www.dkfz.de/mga2/gosim). These similarity values are used to construct a distance matrix that is used as input for a hierarchical clustering algorithm. We have used complete link clustering. The resulting clusters represent groups of genes/proteins that are similar at the functional level.

We tested our method using two experiments: one experiment used two corn gene expression data sets and the other used two *Arabidopsis* proteomic data sets. Results

produced by both experiments confirm that our method of integrating heterogeneous data sets provides additional biological information which cannot be obtained by mapping at the identifier level.

In both the experiments, we generated the clusters for each individual data set as well as for the union of the data sets by merging each of the two individual data sets.

Most of the proteins or genes which did not belong to any of the clusters in clusters generated from individual datasets, grouped into meaningful clusters in the combined data set. This provides the biologists with additional information for exploring the biological systems they are studying. The biologists analyzing the results found clusters generated from the Biological Process hierarchy to be more useful than those generated from Molecular Function hierarchy.

The Arabidopsis dataset combined proteins from two types of proteomics experiments based on the same biological samples—2D gels and shotgun proteomics. According to the biologist's analysis, the combined clusters integrate information about the abundant proteins identified by 2D-gel electrophoresis with those identified by the more sensitive shotgun proteomics approach. The combined clusters provide a more comprehensive view of the processes that are up-regulated during cell dedifferentiation in *Arabidopsis*.

The maize experiment combined two gene expression datasets that used samples from different growing seasons and were based on two different arrays. The corn genitists also confirm that the combined clusters of two gene expression dataset reveal additional information than can be obtained by either individual dataset.

5.2 Future Research

One aspect of the proposal was not implemented in the current work: modeling of many to many correspondences between genes/proteins in the similarity matrix using a weighted bipartite graph. A bipartite graph is an undirected graph where the vertices are partitioned into two disjoint sets and edges only connect vertices from different sets. To integrate two datasets, the genes/proteins from each dataset becomes a vertex set and edges between the vertices are weighted by the gene similarity computed. Afterwards, functional co-clusters can be obtained by applying graph partitioning technique such as minimum cut algorithm to the bipartite graph. This will be an alternative method to the current one, which is supposed to result similar genes or proteins in groups at the functional level. We can compare the results of each method and use the best for biological analysis.

We used an R package GOSim to calculate semantic similarity among heterogeneous data sets. The only plant identifiers currently supported by GOSim are *Arabidopsis* Affymetrix probes. Therefore we had to map both our maize EST sequences and *Arabidopsis* protein identifiers to *Arabidopsis* probe ids to calculate similarities among *Arabidopsis* data sets. We plan to develop a custom interface to GOSim which enables the user to upload the GO annotations for any preferred species. This will make our method easier for biologists to use and will also provide more accurate results.

Finally, we would like to demonstrate that our method can be effectively used to integrate proteomic and transcriptomic data sets from the same or similar biological datasets. Dr.Olga Pechanova has protein expression data from cob tissue from the the same line of

corn (Mp313E) infected with *Aspergillus* and when this data becomes available, we will integrate it with the gene expression we already have in hand.

We plan to publish two papers from this work. The first will be submitted to a bioinformatics journal and will describe the new method. The second will be a detailed analysis of the clustering results for the Maize data and will be submitted to a biological journal.

REFERENCES

- [1] R. Apweiler, A. Bairoch, C. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. Martin, D. A. Natale, C. ODonovan, and L. Redasch, N.and Yeh, “UniProt: the Universal Protein knowledgebase,” *Nucleic Acids Res*, vol. 32, 2004, pp. D115–D119.
- [2] F. Azuaje and O. Bodenreider, “Incorporating ontology-driven similarity knowledge into functional genomics: an exploratory study,” *Bioinformatics and Bioengineering, 2004. BIBE 2004. Proceedings. Fourth IEEE Symposium on*, May 2004, pp. 317–324.
- [3] T. Beissbarth and T. Speed, “: GOstat: find statistically overrepresented Gene Ontologies within a group of genes,” *Bioinformatics*, vol. 20, no. 9, 2004, pp. 1464–1465.
- [4] V. Beisvag, F. Junge, H. Bergum, L. Jolsum, S. Lydersen, C.-C. Gunther, H. Rammampiaro, M. Langaas, A. Sandvik, and A. Laegreid, “GeneTools - application for functional annotation and statistical hypothesis testing,” *BMC Bioinformatics*, vol. 7, no. 1, 2006, p. 470.
- [5] G. F. Berriz, O. D. King, B. Bryant, C. Sander, and F. P. Roth, “Characterizing gene sets with FuncAssociate,” *Bioinformatics*, vol. 19, no. 18, 2003, pp. 2502–2504.
- [6] T. J. Buza, F. M. McCarthy, N. Wang, S. M. Bridges, and S. C. Burgess, “Gene Ontology annotation quality analysis in model eukaryotes,” *Nucl. Acids Res.*, vol. 36, no. 2, 2008, p. e12.
- [7] M. Cashburner, 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. The Gene Ontology Consortium,” *Nature Genetics*, vol. 25, 2000, pp. 25–29.
- [8] G. Chen, T. G. Gharib, C.-C. Huang, J. M. G. Taylor, D. E. Misek, S. L. R. Kardia, T. J. Giordano, M. D. Iannettoni, M. B. Orringer, S. M. Hanash, and D. G. Beer, “Discordant Protein and mRNA Expression in Lung Adenocarcinomas ,” *Mol Cell Proteomics*, vol. 1, no. 4, 2002, pp. 304–313.

- [9] B. R. Chitteti, F. Tan, H. Mujahid, S. M. Magee, Bryce G. and Bridges, and Z. Peng, “Comparative analysis of proteome differential regulation during cell dedifferentiation in Arabidopsis,” *PROTEOMICS*, vol. 8, no. 20, 2008, pp. 4303–4316.
- [10] F. Couto, M. Silva, and P. Coutinho, “Finding genomic ontology terms in text using evidence content,” *BMC Bioinformatics*, vol. 6, 2005, p. S21.
- [11] B. Cox, T. Kislinger, and A. Emili, “Integrating gene and protein expression data: pattern analysis and profile mining,” *Methods*, vol. 35, no. 3, 2005, pp. 303–314.
- [12] G. Dennis, B. Sherman, D. Hosack, J. Yang, W. Gao, H. Lane, and R. Lempicki, “DAVID: Database for Annotation, Visualization, and Integrated Discovery,” *Genome Biology*, vol. 4, no. 9, 2003, p. R60.
- [13] A. Doms and M. Schroeder, “GoPubMed: exploring PubMed with the Gene Ontology,” *Nucl. Acids Res.*, vol. 33, 2005, pp. W783–786.
- [14] J. A. Dowell, D. C. Frost, J. Zhang, and L. Li, “Comparison of Two-Dimensional Fractionation Techniques for Shotgun Proteomics,” *Analytical Chemistry*, vol. 80, no. 17, 2008, pp. 6715–6723.
- [15] T. P. Dvornikova, G. K. Skriabin, and N. N. Suvorov, “Enzymatic transformation of tryptamine by fungi,” *Mikrobiologiya*, vol. 39, no. 1, 1970, pp. 237–247.
- [16] A. Fagan, A. C. Culhane, and D. G. Higgins, “A multivariate analysis approach to the integration of proteomic and gene expression data,” *PROTEOMICS*, vol. 7, no. 13, 2007, pp. 2162–2171.
- [17] W. Feng, G. Wang, B. R. Zeeberg, K. Guo, A. Fojo, D. W. Kane, W. C. Reinhold, S. Lababidi, J. N. Weinstein, , and M. Wang, “Development of Gene Ontology Tool for Biological Interpretation of Genomic and Proteomic Data,” *Proceedings: American Medical Informatics Association Annual Symposium Proceedings*, 2003, p. 839.
- [18] H. Frohlich, N. Speer, A. Poustka, and T. BeiSZbarth, “GOSim: an R package for computation of information theoretic GO similarities between terms and gene products,” *BMC Bioinformatics*, vol. 8, no. 1, 2007, p. 166.
- [19] L. Gautier, M. Moller, L. Friis-Hansen, and S. Knudsen, “Alternative mapping of probes to genes for Affymetrix chips,” *BMC Bioinformatics*, vol. 5, no. 1, 2004, p. 111.
- [20] S. Ghaemmaghami, W.-K. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O’Shea, and J. S. Weissman, “Global analysis of protein expression in yeast,” *Nature*, vol. 425, 2003, pp. 737–741.

- [21] D. Greenbaum, R. Jansen, and M. Gerstein, “Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts,” *Bioinformatics*, vol. 18, no. 4, 2002, pp. 585–596.
- [22] T. J. Griffin, S. P. Gygi, T. Ideker, B. Rist, J. Eng, L. Hood, and R. Aebersold, “Complementary Profiling of Gene Expression at the Transcriptome and Proteome Levels in *Saccharomyces cerevisiae*,” *Mol Cell Proteomics*, vol. 1, no. 4, 2002, pp. 323–333.
- [23] S. P. Gygi, Y. Rochon, B. R. Franza, and R. Aebersold, “Correlation between Protein and mRNA Abundance in Yeast,” *Mol. Cell. Biol.*, vol. 19, no. 3, 1999, pp. 1720–1730.
- [24] J. J. Jiang and D. W. Conrath, “Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy,” *CoRR*, vol. cmp-1g/9709008, 1997.
- [25] R. S. Johnson, M. T. Davis, J. A. Taylor, and S. D. Patterson, “Informatics for protein identification by mass spectrometry,” *Methods*, vol. 35, no. 3, 2005, pp. 223 – 236.
- [26] P. Khatri and S. Draghici, “Ontological analysis of gene expression data: current tools, limitations, and open problems,” *Bioinformatics*, vol. 21, no. 18, 2005, pp. 3587–3595.
- [27] T. Kislinger, K. Rahman, D. Radulovic, B. Cox, J. Rossant, and A. Emili, “PRISM, a Generic Large Scale Proteomic Investigation Strategy for Mammals,” *Mol Cell Proteomics*, vol. 2, no. 2, 2003, pp. 96–106.
- [28] K. Kobayashi, Y. Kobayashi, and A. R. Hardham, “Dynamic reorganization of microtubules and microfilaments in flax cells during the resistance response to flax rust infection,” *Planta*, vol. 195, no. 2, 1994, pp. 237–247.
- [29] S. E. Lewis, “Gene Ontology: looking backwards and forwards,” *Genome Biology*, vol. 6, no. 1, 2004, p. 103.
- [30] D. Lin, “An Information-Theoretic Definition of Similarity,” *In Proceedings of the 15th International Conference on Machine Learning*. 1998, pp. 296–304, Morgan Kaufmann.
- [31] G. Liu, A. E. Loraine, R. Shigeta, M. Cline, J. Cheng, V. Valmeekam, S. Sun, D. Kulp, and M. A. Siani-Rose, “NetAffx: Affymetrix probesets and annotations,” *Nucl. Acids Res.*, vol. 31, no. 1, 2003, pp. 82–86.
- [32] Y. Liu, M. Schiff, K. Czymmek, B. Tallochy, Zsoltand Levine, and S. Dinesh-Kumar, “Autophagy Regulates Programmed Cell Death during the Plant Innate Immune Response,” *Cell*, vol. 121, no. 4, 2005, pp. 567–577.

- [33] 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, no. 10, 2003, pp. 1275–1283.
- [34] M. J. MacCoss, C. C. Wu, and J. R. Yates, "Probability-Based Validation of Protein Identifications Using a Modified SEQUEST Algorithm," *Analytical Chemistry*, vol. 74, no. 21, 2002, pp. 5593–5599.
- [35] F. McCarthy, N. Wang, G. B. Magee, B. Nanduri, M. Lawrence, E. Camon, D. Barrell, D. Hill, M. Dolan, W. P. Williams, D. Luthe, S. Bridges, and S. Burgess, "Ag-Base: a functional genomics resource for agriculture," *BMC Genomics*, vol. 7, no. 1, 2006, p. 229.
- [36] V. K. Mootha, J. Bunkenborg, J. V. Olsen, M. Hjerrild, J. R. Wisniewski, E. Stahl, M. S. Bolouri, H. N. Ray, S. Sihag, M. Kamal, N. Patterson, E. S. Lander, and M. Mann, "Integrated Analysis of Protein Composition, Tissue Diversity, and Gene Regulation in Mouse Mitochondria," *Cell*, vol. 115, no. 5, 2003, pp. 629 – 640.
- [37] A. I. Nesvizhskii, "Protein identification by tandem mass spectrometry and sequence database searching.," *Methods in molecular biology (Clifton, N.J.)*, vol. 367, 2007, pp. 87–119.
- [38] A. I. Nesvizhskii and R. Aebersold, "Interpretation of Shotgun Proteomic Data: The Protein Inference Problem," *Mol Cell Proteomics*, vol. 4, no. 10, 2005, pp. 1419–1440.
- [39] Paphrag, "DNA microarray," http://en.wikipedia.org/wiki/DNA_microarray (current 24 July. 2009).
- [40] S. Patel and S. P. Dinesh-Kumar, "Arabidopsis ATG6 is required to limit the pathogen-associated cell death response," *Landes Bioscience*, vol. 4, no. 1, 2008, pp. 20–27.
- [41] T. H. R. Paul A. Haynes, "Subcellular shotgun proteomics in plants: Looking beyond the usual suspects," *PROTEOMICS*, vol. 7, no. 16, 2007, pp. 2963–2975.
- [42] Pontius, J. U and Wagner, L. and Schuler, G.D., *UniGene: A Unified View of the Transcriptome*, The NCBI Handbook, National Center for Biotechnology Information, Bethesda, MD, 2003.
- [43] S. Pyysalo, F. Ginter, T. Pahikkala, J. Boberg, J. Jarvinen, and T. Salakoski, "Evaluation of two dependency parsers on biomedical corpus targeted at proteinprotein interactions," *International Journal of Medical Informatics*, vol. 75, 2006, pp. 430–442.
- [44] R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and application of a metric on semantic nets," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 19, no. 1, Jan/Feb 1989, pp. 17–30.

- [45] P. Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy," *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1995, pp. 448–453.
- [46] P. Resnik, "Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language," *Journal of Artificial Intelligence Research*, vol. 11, 1999, pp. 95–130.
- [47] I. Rodrigo, P. Vera, L. C. Van Loon, and V. Conejero, "Degradation of Tobacco Pathogenesis-Related Proteins : Evidence for Conserved Mechanisms of Degradation of Pathogenesis-Related Proteins in Plants," *Plant Physiol.*, vol. 95, no. 2, 1991, pp. 616–622.
- [48] B. Roe, "Key Note Address," *MidSouth Computational Biology and Bioinformatics Society (MCBIOS) Conference*, 2008.
- [49] R. G. Sadygov and J. R. Yates, "A Hypergeometric Probability Model for Protein Identification and Validation Using Tandem Mass Spectral Data and Protein Sequence Databases," *Analytical Chemistry*, vol. 75, no. 15, 2003, pp. 3792–3798.
- [50] J. L. Sevilla, V. Segura, A. Podhorski, E. Guruceaga, J. M. Mato, L. A. Martinez-Cruz, F. J. Corrales, and A. Rubio, "Correlation between Gene Expression and GO Semantic Similarity," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 2, no. 4, 2005, pp. 330–338.
- [51] D. L. Tabb, A. Saraf, and J. R. Yates, "GutenTag: High-Throughput Sequence Tagging via an Empirically Derived Fragmentation Model," *Analytical Chemistry*, vol. 75, no. 23, 2003, pp. 6415–6421.
- [52] M. Tian, B. Benedetti, and S. Kamoun, "A Second Kazal-Like Protease Inhibitor from *Phytophthora infestans* Inhibits and Interacts with the Apoplastic Pathogenesis-Related Protease P69B of Tomato," *Plant Physiol.*, vol. 138, no. 3, 2005, pp. 1785–1793.
- [53] Y. Vigouroux, J. C. Glaubitz, Y. Matsuoka, M. M. Goodman, J. Sanchez G., and J. Doebley, "Population structure and genetic diversity of New World maize races assessed by DNA microsatellites," *Am. J. Bot.*, vol. 95, no. 10, 2008, pp. 1240–1253.
- [54] H. Wang, F. Azuaje, O. Bodenreider, and J. Dopazo, "Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships," *Computational Intelligence in Bioinformatics and Computational Biology, 2004. CIBCB '04. Proceedings of the 2004 IEEE Symposium on*, Oct. 2004, pp. 25–31.
- [55] X. Zhou and Z. Su, "EasyGO: Gene Ontology-based annotation and functional enrichment analysis tool for agronomical species," *BMC Genomics*, vol. 8, no. 1, 2007, p. 246.