Making Sense of Microarray Data: Development of an Integrated Bioinformatics Tool

Guoqing Lu

A thesis

in

The Department

Of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of Master of Computer Science at

Concordia University

Montreal, Quebec, Canada

April 2005

Your file Votre référence
ISBN: 0-494-05714-9
Our file Notre référence
ISBN: 0-494-05714-9

Abstract

Making Sense of Microarray Data: Development of an Integrated Bioinformatics Tool

Guoqing Lu

Microarray technology promises to monitor interactions among tens of thousands of genes simultaneously. Two types of microarrays, Oligonucleotide (oligo) and cDNA arrays, are in common use. Oligo arrays have the advantage of providing a platform that can be more readily compared between laboratories. With rapid evolution of hardware and lab protocols, the challenge becomes the analysis of a vast amount of data rather than the manufacture or the use of microarrays. Most software applications were developed dealing with cDNA arrays. There remains a lack of tools that can be used for oligo array analysis. The goal of this research project is to develop a bioinformatics tool dedicated to analyzing oligo array data. Our tool, AffyMiner, consists of three functional components: GeneFinding - finding significant genes in the experiment, GOTree - constructing a Gene Ontology (GO) tree, and interfaces – linking to third-party applications. AffyMiner effectively deals with multiple replicates in the experiment, provides users flexibility of choosing different data metrics for finding significant genes, and is capable of incorporating various gene annotations. In addition, AffyMiner maps genes of interest onto the GO spaces, providing assistance in the interpretation of findings in the context of biology. Furthermore, AffyMiner provides a portal to use Cluster and GenMAPP, two popular programs for microarray analysis. AffyMiner has been used by multiple users and was found to be an effective tool that has reduced plenty of time and efforts needed for data analysis.

Acknowledgement

I thank my supervisor, Dr. Clement Lam, for his various supports on this research work. Dr. Lam's many comments are critical to improve the quality of the thesis. I am particularly grateful to Dr. Greg Butler for initiating and stimulating my research interest in bioinformatics. Two bioinformatics courses lectured by Dr. Butler are very helpful in crystallizing ideas in this project. I also want to thank my committee member Dr. Volker Haarslev.

Special thanks go to Dr. Mike Hallett in the McGill Center for Bioinformatics at McGill University and Dr. David Thomas in the Department of Biochemistry at McGill University. I had lots of fun when working with them on the Human ER apercu (HERA) project.

I also want to thank my previous supervisors, Prof. Sifa Li at Shanghai Fisheries University and Dr. Louis Bernatchez at Laval University. The knowledge and skills obtained from their labs have been so helpful to this project in many ways.

Thanks should go to Dr. Yuannan Xia, the Director of Microarray Core Research Facility at the University of Nebraska Lincoln (UNL), and many end users at UNL. They provided me their requirements and reported bugs in the course of software development. Thanks also go to The Nguyen in the UNL Bioinformatics Core Facility for helpful discussions over software development and programming.

Finally, I thank my dear wife, Qiaomei, and our lovely kids, Eugene and Johnson, for all kinds of love, care, and support.

Table of Contents

List of Figures vii

List of Tables viii

Preface ix

Chapter 1 Introduction 1

1.1 DNA microarray technology 1

1.2 DNA microarray applications 3

1.3 Methods used for microarray analysis 6

1.4 Software tools for microarray analysis 9

1.5 The study problem 12

Chapter 2 Software Design 15

2.1 Data/information flow in a microarray assay 15

2.2 User requirements 17

2.3 Software architecture 19

2.4 Algorithms 20

2.4.1 GeneFinder 20

2.4.2 GOTree 24

2.4.3 Interfaces to Cluster and GenMAPP 27

2.4.4 Introduction to Cluster and GenMAPP 27

Chapter 3 AffyMiner – an Integrated Tool for Microarray Analysis 30

3.1 Introduction to AffyMiner 30

3.1.1 GeneFinder 31

3.1.2 GOTree 39

3.1.3 Interfaces to Cluster and GenMAPP 39

3.1.4 The use of AffyMiner 41

3.2 System requirements 45

4.1 Software comparison between AffyMiner and its counterparts 46

4.2 Limitations of AffyMiner 49

4.3 Future directions 49

4.4 Conclusion 51

Appendix A Arabidopsis up- and down-regulated genes in response to PI5P treatment . 64
Appendix B Glossary 84
Appendix C Index 87

List of Figures

Figure 1. Data/information flow in a microarray experiment 16

Figure 2. Assay steps and the outputs in a typical Affymetrix gene expression array assay (GeneChip® Operating Software Manual, Affymetrix, Inc.) 18

Figure 3. Software architecture 20

Figure 4. Multiple comparisons of experiment replicates (A1, A2, and A3) and control replicates (B1, B2, and B3). 22

Figure 5. Main interface of the Cluster program. 27

Figure 6. The Human biotin metabolism MAPP generated by the GenMAPP program.. 29

Figure 7. AffyMiner main window 31

Figure 8. GeneFinder parameter setting window 33

Figure 9. GeneFinder input setting window 34

Figure 10. Gene annotion popup window 35

Figure 11. GeneFinder output setting window 36

Figure 12. A Gene Ontology tree generated from the GOTree program in AffyMiner. .. 40

Figure 13. Distribution of genes with significantly altered expression levels after PI5P treatment according to the Gene Ontology tree generated in GOTree 44

Figure 14. Sample graphical view of the biological process hierarchy generated by the NetAffx Gene Ontology Mining Tool 47

List of Tables

Table 1 A comparison between cDNA and oligonucleotide arrays 2

Table 2 Applications of DNA microarray technology 5

Table 3 Inferential statistics methods used for microarray data analysis 7

Table 4 Some of microarray data analysis software and methods available therein 10

Table 5 The structure of GO terms generated by ChipInfo (abridged) 25

Table 6 Matches between GO IDs and Affymetrix probe set IDs 26

Table 7 Part of the output generated from GeneFinder 37

Preface

Microarray technology is a powerful approach for genomics research (Leung & Cavalieri 2003). The widespread use of this high-throughput data collection technique over years has produced a vast amount of heterogeneous data. The challenge faced by today's researchers is to develop effective ways to analyze genomic data that has been and will continue to be collected (Khatri et al. 2004; Lockhart & Winzeler 2000). My research project aims to tackle this challenging problem by developing a bioinformatics tool in order to facilitate microarray data analysis and assist researchers in discovering biological knowledge embedded within the massive gene expression data.

The thesis is organized in four chapters as detailed below. Chapter 1 is a general introduction, including a literature review on the microarray technology, its wide applications, and existing software tools, and a detailed description of the study problem. Chapter 2 is related to the software design, describing the flow of information in a typical microarray assay, user requirements, software architecture, and algorithms. Chapter 3 introduces AffyMiner, the bioinformatics tool that I developed for microarray data analysis and data mining. The functions of its components and the system requirement of AffyMiner will be detailed therein. Chapter 4 compares AffyMiner with several widely used counterpart software tools, and discusses the limitations and the future directions of AffyMiner. This chapter ends with a conclusion on the research work pointing out the significance of AffyMiner.

Chapter 1 Introduction

Microarray technology is a powerful tool for genomics research (Leung & Cavalieri 2003). In this chapter, I will briefly introduce the microarray technology, its broad applications, and the various methods and software tools used for the microarray analysis. At the end of this chapter, I will describe the study problem and the rationale for this master thesis project.

1.1 DNA microarray technology

Two types of microarrays are in common use, cDNA microarrays and oligonucleotide microarrays (Table 1) (Lockhart et al. 1996; Schena et al. 1995). In cDNA microarrays, each probe DNA corresponds to a unique gene and is prepared by high-throughput PCR on cDNA libraries or the genomic DNA. With synthetic oligonucleotide arrays, sequence information is needed and used for chemical synthesis of the probes, i.e., oligonucleotides. The difference between the two technologies is summarized in Table 1. In general, cDNA arrays are relatively flexible since they do not rely on genomics sequence information and the cost is comparatively low. However, the cDNA array has some disadvantages, such as the variable amount of DNA spotted in each spot, and the misidentification of clones. Oligonucleotide arrays provide a platform that can be more readily compared between laboratories since they are factory designed and synthesized (The Tumor Analysis Best Practices 2004). However, oligonucleotide arrays are relatively expensive. Minimizing nonspecific hybridization is another challenging issue in oligonucleotide array technology. Bioinformatics-based search of unique gene sequences is also a critical factor in the design of oligonucleotide probes (Rouillard et al. 2002; Rouillard et al. 2003).

Table 1 A comparison between cDNA and oligonucleotide arrays

	Spotted cDNA arrays	Oligonucleotide arrays
Major arraying approach	Spotting	Photolithography
Probe	cDNA, complete sequence of a gene	Oligos, a series of fragments of a gene
Measurement in one array	Relative expression of two samples	Expression of a single sample
Quality of data	Varied	Higher
Cost	Cheaper	Expensive
Need to know genome sequence	No	Yes
Invented by	Stanford University	Affymetrix Inc.

Both cDNA and oligonucleotide probes can be arrayed on glass slides using robotic pin spotting (Cheung et al. 1999) or ink-jet printing (Okamoto et al. 2000). The widely used Affymetrix GeneChip array uses the photolithographic method and phosphoramidite chemistry for in situ synthesis of high density (300,000 features on a 1.28 x 1.28 cm2 chip) (Lipshutz et al. 1999). A different method for in situ synthesis of oligonucleotide probes (60-mer) using ink-jet technology was described in Hughes et al.

(2001). The use of longer oligos is reported to address issues of lower hybridization specificity and sensitivity with shorter oligos (Barczak et al. 2003; Erle et al. 2002; Kane et al. 2000).

The basic principle behind microarray analysis is the hybridization of complementary DNA strands (Dharmadi & Gonzalez 2004; Southern 2001), one being immobilized DNA called “probes” and the other being labeled cDNA called “targets” (Zareparsi et al. 2004). The rationale for DNA microarrays is that the signal intensity of the hybridized DNA probes serves as an estimate of the abundance of each target species and hence a measure of the expression level of each specific gene. A typical microarray experiment comprises four steps below: 1) DNA complementary to genes of interest is generated and laid out in microscopic quantities on solid surfaces at defined positions; 2) DNA from samples is labeled and eluted over the surface, and complementary DNA binds; 3) presence of bound DNA is detected by florescence following laser excitation (imaging processing); 4) array data are preprocessed (i.e., normalized), analyzed, and interpreted (Dharmadi & Gonzalez 2004).

1.2 DNA microarray applications

DNA microarrays have many applications varying from gene expression, point mutation, SNPs, to pharmacogenomics (Heller 2002). Two entire supplemental issues of Nature Genetics, *Chipping Forecast I* and *Chipping Forecast II*, published in 1999 and 2002, respectively, were devoted to reviews not only on microarray technology itself but also on its various applications. Microarrays as a state of the art technology allow researchers to elucidate how the genome is organized and how developmental processes are orchestrated, and offer an important aid in the diagnosis and prognosis of cancer and in the selection of drug targets (Duyk 2002). Many researchers have extensively done reviews on microarray applications. For example, Epstein & Butow (2000) provides a comprehensive review on the uses of microarrays in deletion mapping, measurement of gene dosage, and transcript profiling that includes transcriptional analysis of cancer, organ- and disease-specific arrays, budding yeast mitosis and meiosis, stress response, and of aging. Recent reviews tend to focus on a specific subject, such as cell biology (Panda et al. 2003), or a specific area, such as bacterial systems (Dharmadi & Gonzalez 2004) and applied breeding (Walsh & Henderson 2004).

Various types of applications of microarray technology are summarized in Table 2. This table indicates several research trends. At first, most studies are related to gene expression, where investigators expected to identify differentially expressed genes in unlike physical or environmental conditions, such as diseased and healthy tissues, and/or to find co-regulated genes, i.e., those activated by a particular transcription factor. Secondly, DNA microarrays through measuring genome-wide gene expression patterns are becoming an important tool for pharmacogenomic applications, such as the identification of molecular targets for drugs, toxicological studies and molecular diagnostics. Thirdly, some new directions include the reconstruction of metabolic pathways and genetic networks. A combination of expression data with other experimental data plus effective computer algorithms seems essential to work in the new directions.

Table 2 Applications of DNA microarray technology

Application	References
Gene of interest and functional annotation	Bunney et al. 2003; Collins et al. 2004; Eisen et al. 1998; Reinke 2002; Tusher et al. 2001
Cancer	Chung et al. 2002; DeRisi et al. 1996; Grant et al. 2004; Lyons-Weiler et al. 2004; Rubin et al. 2004; Watson et al. 2004
Neuroscience	Luo & Geschwind 2001; Nisenbaum 2002; Parrish et al. 2004
Drug response and Discovery	Debouck & Goodfellow 1999; Gmuender 2002; Ivanov et al. 2000; Levy 2003; Ulrich & Friend 2002; Weeraratna et al. 2004
Disease prognosis and diagnosis	Dyrskjot 2003; Kononen et al. 1998; Robson & Garnier 2002; van 't Veer et al. 2002; Weeraratna et al. 2004; Yershov et al. 1996
Developmental biology	Chen & Corey 2002; Smith & Greenfield 2003
Toxicogenomics	Bottone et al. 2003; Medlin 2001; Ulrich & Friend 2002
Pathway mapping / networks	Abbott et al. 2003; Bremnes et al. 2002; Chung et al. 2004; Iglesias et al. 2004; Mircean et al. 2004; Reinke 2002; Scandurro et al. 2001; Shalev et al. 2002; Slonim 2002; Weldon et al. 2002
Ecology and evolution	Cherkasova et al. 2003; Eizirik et al. 2003; Gibson 2002; Held et al. 2004; Taroncher-Oldenburg et al. 2003

1.3 Methods used for microarray analysis

Microarray data analysis includes data preprocessing, inferential statistics computation, descriptive statistics estimation, and pathway/network analysis.

Data preprocessing, i.e., normalization, is a necessary step in microarray analysis, since it allows comparison of datasets generated from different array experiments by making sure the samples are equivalent in terms of technical and biological biases. For oligonucleotide arrays, there are at least three methods available for normalizing probe-level data. The first method is referred to as the scaling method, where intensities should be scaled so that each array has the same average value. The second method is called the non-linear method, which used non-linear smooth curves proposed by Li & Wong (2001). The third method is based on empirical results demonstrating the ability to reduce variance without increasing bias. A comparison of the above three methods can be found in Bolstad et al. (2003). In addition, Stoyanova et al. (2004) proposed to use principal component analysis to normalize the data generated from oligonucleotide arrays. Normalization methods for cDNA arrays are quite different. Their detailed description can be found in several good reviews (Fan et al. 2004; Leung & Cavalieri 2003; Quackenbush 2001; Smyth & Speed 2003).

Inferential statistics are frequently used to test the null hypothesis, for example, there is no difference in signal intensity for the gene expression measurements in normal and diseased samples. A test statistic is used to decide whether to accept or reject the null hypothesis. Table 3 listed inferential statistics methods in three different paradigms, i.e., comparisons of paired groups, unpaired groups, and three or more groups. Based on data distribution, either parametric or nor-parametric methods can be used for the hypothesis test. If a dataset follows normal distribution, parametric methods are used. Otherwise, non-parametric methods are used. Pan (2002) compared three methods: the t-test, a regression modeling approach and a mixture model approach. It concluded that the mixture model approach is superior with respect to plausible assumptions of data distribution, the resulting significance levels, and the number of genes detected. For many applications, the significance level was set to be 0.05 or 0.01. However, there is a multiple testing problem in the case of microarray data, where thousands of genes are compared simultaneously. The traditional p-value cutoffs should be made stricter to avoid an abundance of false positive results. Zhong et al (2004) discussed various strategies on the p-value adjustment procedures.

Table 3 Inferential statistics methods used for microarray data analysis

Paradigm	**Parametric test**	**Nonparametric test**
Compare two unpaired groups	Unpaired t-test	Mann-Whitney test
Compare two paired groups	Paired t-test	Wilcoxon test
Compare 3 or more groups	ANOVA	

Microarray data are highly dimensional. There are many thousands of data points made from a small number of samples. Descriptive (exploratory) statistical analysis is needed to help find meaningful patterns in the data. A first step is to arrange the data in a matrix, and the second is to use a distance method to define the relatedness of the different data points. Various methods are currently employed including classical methods such as hierarchical clustering and K-means clustering as well as methods originated from this group such as super-paramagnetic clustering (Blatt et al. 1996) and coupled two-way clustering (Getz et al. 2000). Many others, such as self-organizing tree algorithm, principal component analysis, discriminant analysis and clustering tools in different favors have recently been introduced. Methods of this type are reviewed in several publications (Quackenbush 2001).

There are at least three interesting approaches in analyzing microarray data in a pathway perspective (Leung & Cavalieri 2003). The first is an extension of the exploratory cluster analysis described above. The second is to identify the global regulatory network architecture from microarray data. The final approach is to study the expression data in a pathway perspective through mapping expression data onto metabolic pathways. Various methods have been proposed for constructing a network from microarray data, such as a Boolean network that simplifies gene expression as a binary logical value to infer the induction of a gene as a deterministic function of the state of a group of other genes (Akutsu et al. 2000; Liang et al. 1998; Maki et al. 2001) and a Bayesian network that models interactions among genes, evaluates different models and assigns them probability scores (Friedman et al. 2000; Hartemink et al. 2001). More recent modifications of the Bayesian network methods focus further on finding probabilistically supported gene interactions or on combining these into subnetworks, on modeling “latent” or hidden variables representing biological information unavailable to the model, and on incorporating prior biological knowledge or annotation (Slonim 2002).

1.4 Software tools for microarray analysis

Although at present there is no clear standard solution for microarray data analysis software, numerous tools of this type are available, both commercially and freely (Table 4). Most of the software available are still in the early phase of the software development process (Holloway et al. 2002). An exhaustive listing of microarray software can be found at Y.E. Leung's website (http://ihome.cuhk.edu.hk/~b400559/array.html). Some important open source and commercial software have been previously reviewed. Dudoit et al. (2003) provides a review on three open source software, Bioconductor, TM4, and BASE. What these tools have in common is the availability of their software source code, which allows users to modify the code as well as to expand the functionality. Dresen et al. (2003) compared three commercial software packages, arraySCOUT, GeneSpring and Spotfire DecisionSite, to evaluate their applicability for analysis of gene expression data (Dresen et al. 2003). GeneSpring seems to be an analysis tool more impressive than the rest. Liu et al. (2004) compared several software tools, listed a subset of tools most commonly used, and described the features that would constitute an ideal microarray analysis software suite.

Table 4 lists some software tools used by microarray investigators. It is demonstrated that most software packages, such as Affymetrix Data Mining Tools, BioDiscover software, and GeneData Expressionist ™, focus on the exploratory methods, including k-means clustering, hierarchical clustering, SOMs, and PCA. There are applications, e.g., GenMAPP and arraySCOUT ™, with an emphasis on integrative methods, including the pathway and Gene Ontology approaches. Despite an increasing number of software tools available for microarray analysis, there are still many aspects with poor or incomplete coverage (Herrero et al. 2004). For example, only a few of them were developed to deal with data generated from oligonucleotide arrays. Most existing software applications are not open source. Moreover, these software focus on unsupervised cluster methods that, in many cases, are used for inadequate purpose (Guffanti et al. 2002; Herrero et al. 2003; Simon et al. 2003; Simon & Dobbin 2003). Thus, there is a high demand for efficient software tools for analyzing oligo array data.

Table 4 Some of microarray data analysis software and methods available therein

Source	**Software**	**Methods/applications**
Affymetrix	Data Mining Tool $	Clustering and discriminatory gene analysis.
BioDiscovery	ArrayPack™ $	Integrated expression management system.
	GeneSight™ $	K-means and hierarchical clustering, SOMs and PCA. Discriminatory gene analysis.
GeneData	GeneData Expressionist™ $	K-means and hierarchical clustering and SOMs; Discriminatory gene analysis.
Gladstone Institutes, UCSF	GenMAPP	Biological pathways, GO trees
InforMax	Xpression NTI $	Hierarchical and nonhierarchical clustering

		methods.
Lion Biosciences	ArraySCOUT™ $	Connectivity to modules for analysis of molecular networks and biological pathways.
Lund University	BASE	Database, normalization, data analysis
Molmine	J-express	Hierarchical and K-means clustering. SOMs and PCA. Profile similarity search.
Silicon Genetics	GeneSpring™ $ GeNet™ $ Metamine™ $	Machine learning tools, clustering methods and PCA. Integrated platform for gene expression research.
Spotfire	DecisionSite™ $	Clustering and prediction tools. Integrated platform for functional genomics.
Stanford University	Cluster/Treeview SAM	Hierarchical clustering, SOMs, k-means clustering, PCA. Significant analysis of microarrays
The Institute for Genomic Research	TIGR Microarray 4 (TM4)	Hierarchical clustering, k-means clustering, self-organizing maps, principal components analysis, support vector machines, gene shaving, and relevance networks
Whitehead Institute	GeneCluster	K-nearest neighbors (KNN) Weighted voting (WV), SOM

MDS: Multidimensional scaling; PCA: Principal component analysis; SOM: Self-organizing map; $: commercial software

1.5 The study problem

We are interested in developing a bioinformatics tool for Affymetrix gene expression array data analysis and mining. Two reasons drove us to focus on the Affymetrix microarrays. First, Affymetrix oligo array have the advantage over other types of arrays since it provides a platform that can be more readily compared between laboratories. The increasing use of Affymetrix microarrays and the emerging use of this technology in clinical trials have led to the development of best practices for microarray data generation and interpretation in both the pharmaceutical and academic research communities (The Tumor Analysis Best Practices 2004). Affymetrix system provides a choice of several data metrics for the detection of the gene expression level and the significance of changes, as detailed in the next paragraph. Secondly, only a few software tools are currently available for GeneChip® array analysis (Table 4). These tools are often not integrated in a way that biologists can use them to analyze array data effectively and easily.

Gene expression analysis of a single Affymetrix array generates a number of data metrics such as a Detection p-value, a Detection call (i.e., Present, Marginal, or Absent), and a Signal value for each probe set (Affymetrix GeneChip® Expression Analysis: Data Analysis Fundamentals). These data metrics are used in the database of gene expression profiles, and facilitate sample classification and transcript clustering analysis. Comparison analysis between two arrays of the same type generates a Change p-value, an associated Change (i.e., Increase or Decrease), and a Signal Log Ratio. One important approach for determining genes that demonstrate robust changes in the experimental samples compared to the control samples involves the following three metrics: Detection, Change, and Signal Log Ratio. When looking for robust increases, for example, it is important to select for transcripts with "Present" in the experimental sample, "Increase" in the Change call, and a Signal Log Ratio exceeding a certain threshold, e.g., 1.0. Note that when the above guideline is applied in determining robust changes, conflicting information may occur for some genes, due to the fact that Detection, Change, and Signal Log Ratio are calculated separately using different algorithms. The benefit of this approach, however, is that genes can be assessed using independent data metrics.

Affymetrix GeneChip® Operating Software (GCOS) calculates the above quantitative and qualitative data independently. Finding genes with significant changes in the experimental samples currently requires considerable time and effort to manually process the gene expression data following the above guidelines. This tedious process becomes even more complicated when biological replicates are involved, since each experimental sample needs to be compared with each control sample and each of the Detection, Change, and Signal Log Ratio values in all the experimental samples must be compared. In addition, when replicates are introduced, statistical analysis such as the Student's t-test or the Mann-Whitney Nonparametric test may be applied. This adds a new variable into the data metrics that need to be considered in defining significant genes. We have been using Affymetrix software packages and other third-party programs such as GeneSpring for microarray analysis (http://www.silicongenetics.com/), but none of the existing software programs can be easily used to analyze Affymetrix gene expression array data with multiple replicates in the experiment.

Once significantly up-regulated and down-regulated genes are found, the subsequent challenge is how to interpret the gene expression analysis result. Towards this goal, several public resources such as GO (http://www.geneontology.org/ (Camon et al. 2004; Harris et al. 2004), KEGG (http://www.genome.jp/kegg/ (Kanehisa et al. 2004), and DAVID (http://david.niaid.nih.gov/david/), have made significant efforts in defining molecular functions and metabolic pathways. Some open source software packages such as GenMAPP and MAPPFinder (Dahlquist et al. 2002; Doniger et al. 2003) and several commercial software systems such as Ingenuity Pathways Knowledge Base (http://www.ingenuity.com/) have been made available. The NetAffx Analysis Center, one of the public resources developed by Affymetrix, is of particular interest to researchers since it correlates their GeneChip® array results to a catalog of array design and annotation information. However, NetAffx does not offer a flexible way to readily perform this operation. For example, when the user does a batch query, the output is sorted alphabetically based on the probe set IDs, which differs from the input order. Another disadvantage of NetAffx is that there is no way to incorporate quantitative data, e.g., signal log ratio, with its annotation information into a single table.

The goal of this research is to develop an integrated bioinformatics tool to address aforementioned issues. The tool is expected to have three functions: 1) finding the genes with significant changes in the experimental samples compared to the control samples using any of the four data metrics and their threshold values; 2) mapping the significant genes onto the GO spaces, i.e., molecular function, biological process, and cellular component; 3) providing interfaces of integrating other popular microarray analysis tools, such as Cluster and GenMAPP, into this application.

Chapter 2 Software Design

As mentioned in Chapter 1, microarray technology generates a vast amount of data. Without using appropriate bioinformatics tools, such data could not be transferred into useful knowledge. This chapter describes aspects related to software design, including data/information flow in a microarray assay, user requirements, software architecture, and some important algorithms. The design was based upon my understanding of the users' requirements and my analysis of current counterpart tools. Given the short time of graduate research, I have been focusing on the development of an integrated tool for high-level data analysis that does not include methods for low-level data analysis, such as image processing and data normalization.

2.1 Data/information flow in a microarray assay

The flow of information in a microarray assay is shown in Figure 1. The information flow in a microarray experiment begins with experimental design and followed by hybridization, data acquisition, data preprocessing, and data analysis. Taking together inferential analysis, exploratory analysis, and pathway analysis, plus known biology and validation, researchers expect to understand the biology, such as molecular functions and metabolic pathways.

Microarray experimental design defines biological questions and hypothesis, and the microarray experiment scheme, including biological replication in the experiment. Hybridization is the formation of double-stranded DNA, RNA, or DNA/RNA hybrids by complementary base pairing. Suitable protocols and procedures are used in the hybridization step. Data acquisition is the process where images are scanned at an appreciate resolution. Microarray data pre-processing, i.e., normalization, adjusts the average value of an experimental array so that it equals that of the baseline array, thus allowing datasets generated from different arrays to be compared. Normalized data are then analyzed at different levels, such as inferential analysis, exploratory analysis, and integrated analysis. This thesis focuses on inferential analysis and integrated analysis, as illustrated in the grey boxes in Figure 1. As for the exploratory analysis, we will take advantage of the open source software instead of reinventing the wheel.

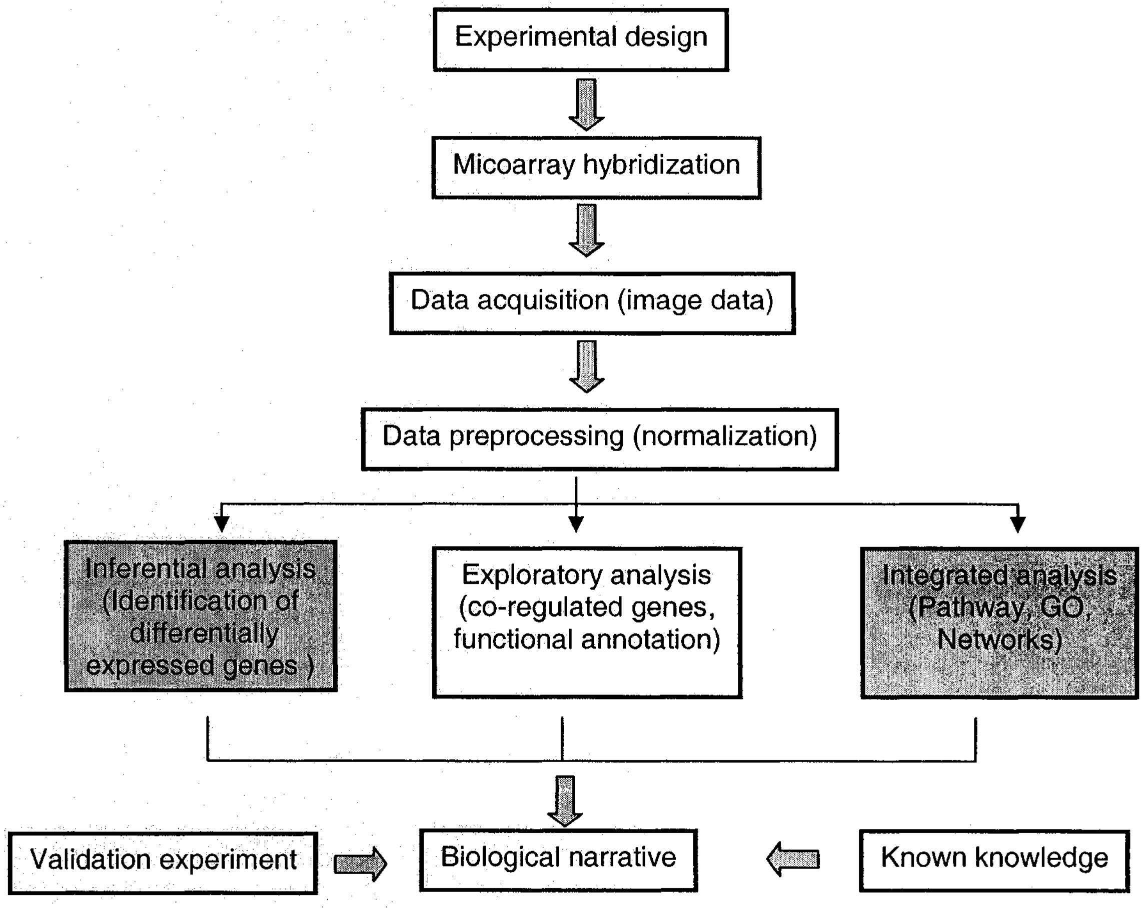

Experimental design
Micoarray hybridization
Data acquisition (image data)
Data preprocessing (normalization)
Inferential analysis (Identification of differentially expressed genes)
Exploratory analysis (co-regulated genes, functional annotation)
Integrated analysis (Pathway, GO, Networks)
Validation experiment
Biological narrative
Known knowledge

Figure 1. Data/information flow in a microarray experiment

In an Affymetrix GeneChip® microarray assay, there are five steps involved: 1) preparing the target; 2) setting up an experiment; 3) hybridizing, washing and staining the probe array; 4) scanning the probe array; 5) analyzing the hybridization data (Figure 2). The GCOS from Affymetrix generates several output files, *.exp, *.dat, *.cel, and *.chp. The analyzed array data can be published to the database. The *.exp file contains experimental information about the experiment performed and the array used. Array images are scanned with an appropriate scanner, and the image data are stored in the *.data file. A signal intensity value for each cell is saved in the *.cel file. The *.chp file contains data from signal chip analysis or pairwise comparison analysis, including signal value, signal detection (Present or Absent), signal change (Increase or Decrease), and signal log ratio. The data exported from the .chp files in the GCOS will be used by the tool to be developed.

2.2 User requirements

- Be able to deal with data exported from the GCOS. The data contain probe sets, signal detection, signal value, signal log ratio, signal change, et al.
- Be capable of sorting/filtering for significant genes. The user has the flexibility to choose different data metrics and set up different threshold values.
- Be able to perform inferential analysis, such as the t-test.
- Be able to do exploratory analysis. The user can use various clustering approaches and machine learning methods for exploratory analysis.
- Be able to perform data mining. The user is able to incorporate information of Gene Ontology and metabolism pathways.

- Have easy-to-use graphical interfaces. The program should be user friendly and easy-to-use and has attractive interfaces.
- Provide ready-to-publish charts and tables.

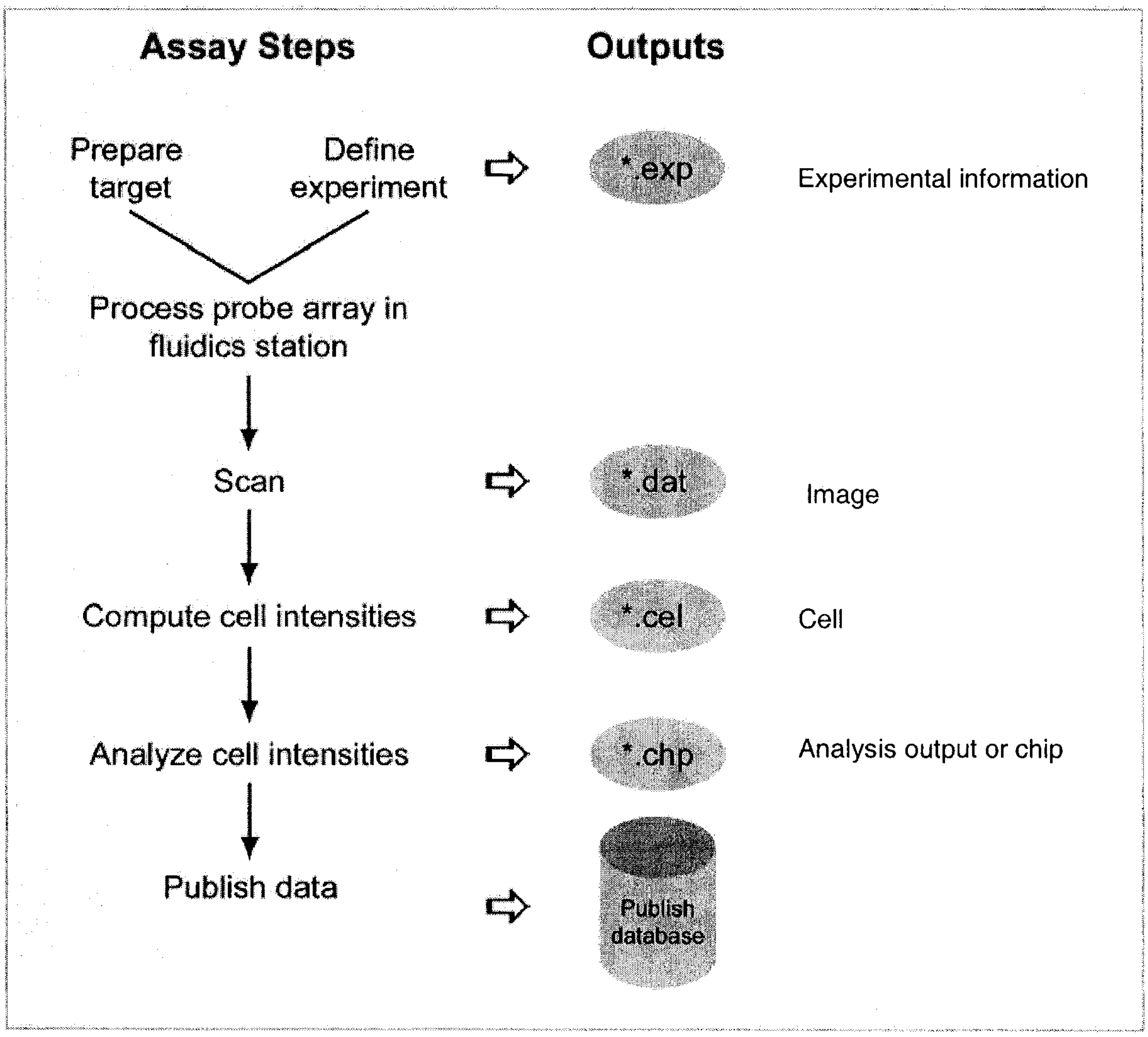


Figure 2. Assay steps and the outputs in a typical Affymetrix gene expression array assay (GeneChip® Operating Software Manual, Affymetrix, Inc.)

2.3 Software architecture

The tool to be developed is called AffyMiner, indicating mining Affymetrix microarray data for biological knowledge. AffyMiner comprises three functional components, GeneFinder, GOTree, and interfaces with Cluster and GenMAPP, as shown in the dash line box (Figure 3). These components are interconnected and all use as input the Affymetrix microarray data. Instead of reinventing the wheel, AffyMiner uses two well-known open source software tools, Cluster and GenMAPP, for exploratory and integrative analyses

GeneFinder finds differentially expressed genes based on the data metrics and thresholds values that the user chooses. GeneFinder can also incorporate gene annotation information. GOTree creates a hierarchical Gene Ontology tree illustrating the relationship of genes in the context of biological processes, gene functions, and cellular components. The input to GOTree can be the output generated from GeneFinder or the data defined by the user.

The interfaces to Cluster and GenMAPP provide a way of integration between AffyMiner and third party programs. AffyMiner transfers the data exported from GCOS into a format suitable for Cluster and GenMAPP. Cluster and GenMAPP will be introduced in 2.4.4.

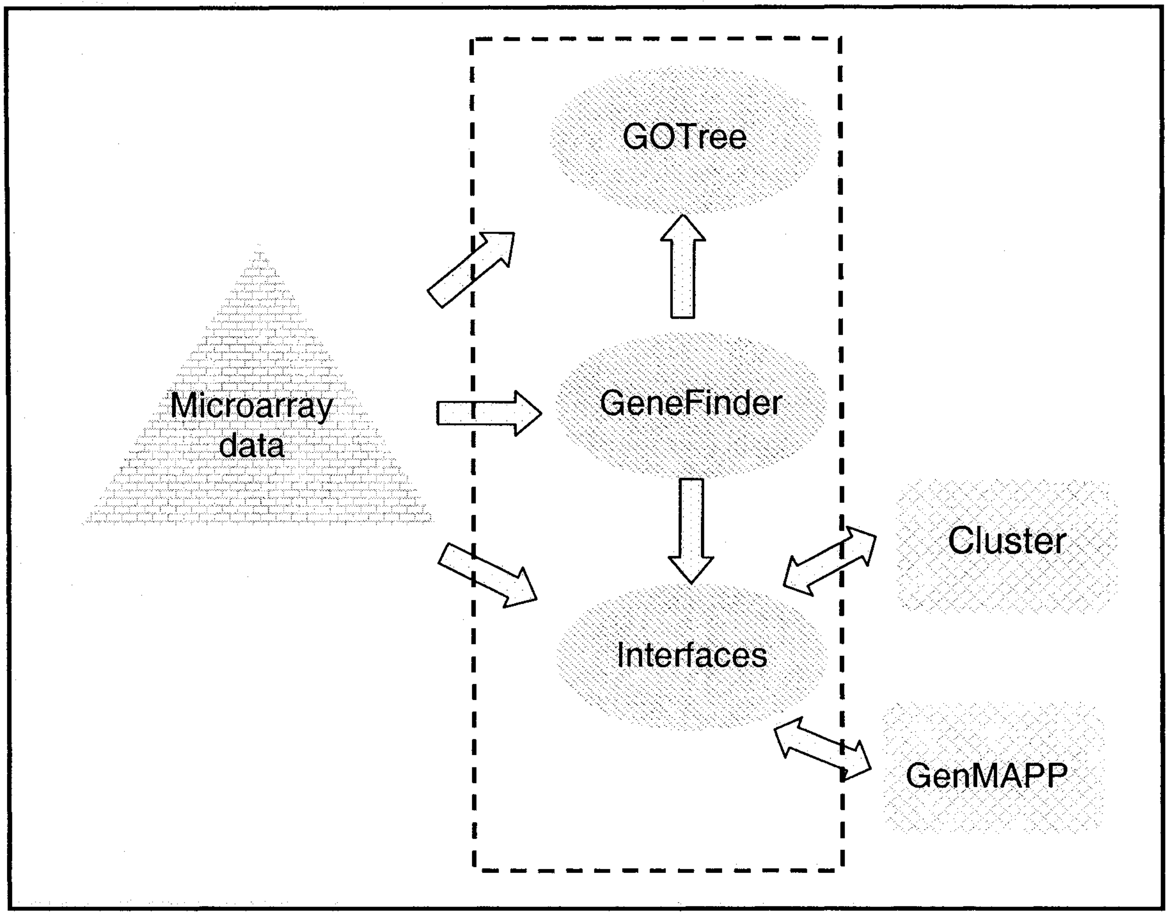


Figure 3. Software architecture

2.4 Algorithms

2.4.1 GeneFinder

Affymetrix GCOS (GeneChip Operating Software) provides users with both qualitative and quantitative measures of transcript performance, including Detection, Change, and Signal Log Ratio. Detection is the qualitative measure of presence or absence for a particular transcript. The Detection calls are a fundamental criterion for significance of the expression of a transcript between samples. For example, when looking for robust increases, it is important to select for transcripts that are called "Present" in the experimental sample, since it is uninformative when we see "Absent" to "Absent" changes, which need to be eliminated for further analysis.

"Change" is the qualitative measure of increase or decrease for a particular transcript. When looking for both significant increases and decreases, it is important to eliminate "No Change" calls. Signal Log Ratio is the quantitative measure of the relative change in transcript abundance. The Affymetrix Gene Expression Assay has been shown to identify Fold Changes greater than two 98% of the time by (Wodicka et al. 1997). Based on these observations, robust changes can be consistently identified by selecting transcripts with a Signal Log Ratio >1 for increases and < -1 for decreases. When performing a single comparison analysis, it is important that above three data metrics be applied. Note that some transcripts may provide conflicting information, for example, a transcript is called "Increase" but has a Signal Log Ratio of less than 1. Alternatively, a transcript is called "Absent" in both experimental and baseline files, but is also called "Increase." These contradictions arise due to the fact that Detection, Change, and Signal Log Ratio are calculated separately. The benefit of this approach is that transcripts can be assessed using three independent metrics. Thus, in order to determine the most robust changes, it is crucial to use all three metrics in conjunction.

Basic steps for determining robust increases include 1) eliminating probe sets in the experimental sample called "Absent"; 2) selecting for probe sets called "Increase"; 3) eliminating probe sets with a Signal Log Ratio below 1.0. Basic steps for determining robust decreases include 1) eliminating probe sets in the baseline sample called "Absent"; 2) selecting for probe sets called "Decrease"; 3) eliminating probe sets with a Signal Log Ratio above -1.0.

When biological replicates are introduced, multiple comparisons are needed. Figure 4 shows nine comparisons between three experimental sample replicates (A1, A2, and A3) and three control sample replicates (B1, B2, and B3). In this case, inferential statistics are not useful, since the sample size is too small (three replicates and two conditions). It is strongly suggested to do pairwise comparisons and to apply the rules described above to find genes with robust changes. When replicates are available in the experiment, it becomes possible to relax the sorting thresholds, for example, the number of Increase being 6 out of 9 comparisons and average signal log ratio being 0.8. The software tool needs to provide full flexibility for the user to decide all the thresholds based upon their experience.

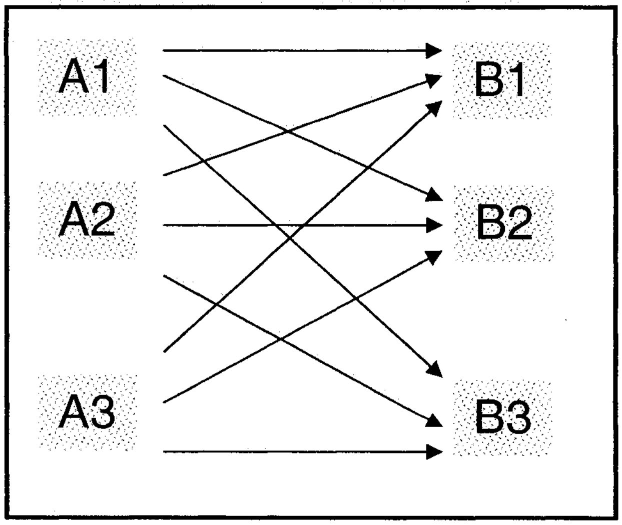


Figure 4. Multiple comparisons of experiment replicates (A1, A2, and A3) and control replicates (B1, B2, and B3).

The algorithm for finding genes with robust increase in expression is as follows:

```
double num_threshold, fold_threshold;
int num_present = 0;
while (!eof(inputFile)) { // read each line
        if (column = = “experiment” || detection = = “present”) num_present++;
        if (! beginWith(“AFFX”) &&
                num_present > num_threshold &&
                p < 0.05 &&
                signal_change = “I” &&
                fold_change > fold_threshold) {
                        print probe_set;
        } // end of if
}// end of while
```

The algorithm for finding genes with robust decrease in expression is as follows:

```
double num_threshold, fold_threshold;
int num_present = 0;
while (!eof(inputFile)) { // read each line
        if (column == “control” || detection ==”present”) num_present++;
        if (! beginWith(“AFFX”)  &&
                num_present > num_threshold  &&
                p < 0.05 &&
                signal_change = “D” &&
                fold_change < fold_threshold) {
                        print probe_set;
        } // end of if
}// end of while
```

2.4.2 GOTree

The Gene Ontology (GO) Consortium (Harris et al. 2004) produces structures of biological knowledge using a controlled vocabulary consisting of GO terms. GO terms are organized into three general categories: biological process, molecular function, and cellular component. The terms within each category are linked in defined parent-child relationships that reflect current biological knowledge. All genes from different organisms are systematically associated with GO terms, and these associations continue to grow in complexity and detail as sequence databases and experimental knowledge grow (Zhong et al. 2004). GO provides a useful tool to look for common features that are shared within a list of genes.

GOTree uses two files generated respectively from ChipInfo and GeneFinder. ChipInfo is designed for retrieving annotation information from online databases such as NetAffx and Gene Ontology and organizing such information into easily interpretable tabular format outputs (Zhong et al. 2003). ChipInfo has functions for computing related summary statistics of probe sets and Gene Ontology terms. A sample output produced from ChipInfo is shown in Table 5. The table is abridged. The actually number of GO terms is very large, for example, 7422 terms for molecular function. Table 5 includes three columns, GO_ID, Path, and GO_Term. GO IDs and GO terms are defined by the Gene Ontology Consortium. The path indicates the route passing from root to an internal node or a leaf node in the GO tree. For example, the path for adult behavior is from biological process (1), behavior (1.0), to adult behavior (1.0.0).

Table 5 The structure of GO terms generated by ChipInfo (abridged)

GO_ID	Path	GO_Term
8150	1	biological_process
7610	1,0	Behavior
30534	1,0,0	Adult behavior
8343	1,0,0,0	Adult feeding behavior
8344	1,0,0,1	Adult locomotory behavior
7628	1,0,9,0,0	Adult walking behavior
7630	1,0,9,0,2	Jump response
7636	1,0,9,0,2,0	chemosensory jump behavior
7636	1,0,4,0	chemosensory jump behavior
48148	1,5,11,2,3,1,2,4,0	behavioral response to cocaine

Another input required by GOTree is either the output generated from GeneFinder or the data defined by the user. The GeneFinder output will be described in 3.1.1. The input data should include the following four items: probe sets, Gene Ontology biological process, Gene Ontology cellular component, and Gene Ontology molecular function.

A high-level description of the algorithm to build the GO tree is as follows:

- Read the output file generated by GeneFinder
- Write in an array the GO Ids and their corresponding Affymetix probe set IDs as shown in Table 6

- Find GO Path IDs for each GO ID in the array, add the GO Path IDs to each element in the array
- Sort by the GO Path IDs, compute the sum of probe sets associated with each node
- Build the entire tree based on the GO Path IDs and write in each node GO term, GO ID, and the number of probe sets.

Table 6 Matches between GO IDs and Affymetrix probe set IDs

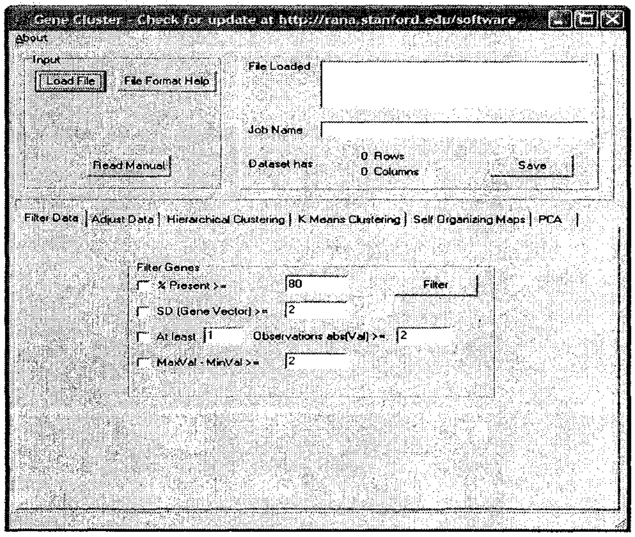
GO_ID	**Gene set**
4194	267580_at
4672	267481_at, 267461_at, 267624_at
4871	267477_at, 267516_at
6355	267477_at
6468	267481_at, 267461_at, 267624_at
7165	267477_at
9507	267592_at, 267624_at
9637	263669_at
9882	263669_at
12505	267481_at

2.4.3 Interfaces to Cluster and GenMAPP

The interfaces in AffyMiner provide functions for formatting data and for linking to Cluster and GenMAPP. Input data are formatted to satisfy the requirements of Cluster and GenMAPP. Systems calls will be used to launch the programs.

2.4.4 Introduction to Cluster and GenMAPP

Cluster is a well-known program developed for analyzing microarray data (Eisen et al. 1998). It performs a variety of types of cluster analysis and other types of processing such as filtering (Fig. 5). Cluster includes functions for hierarchical clustering, self-organizing maps (SOMs), k-means clustering, and principal component analysis. The software manual describes in detail how to use Cluster, which is available at Dr. Eisen's website, http://rana.lbl.gov/EisenSoftware.htm.



Dataset has 0 Rows 0 Columns

☐ % Present >= 80

☐ SD (Gene Vector) >= 2

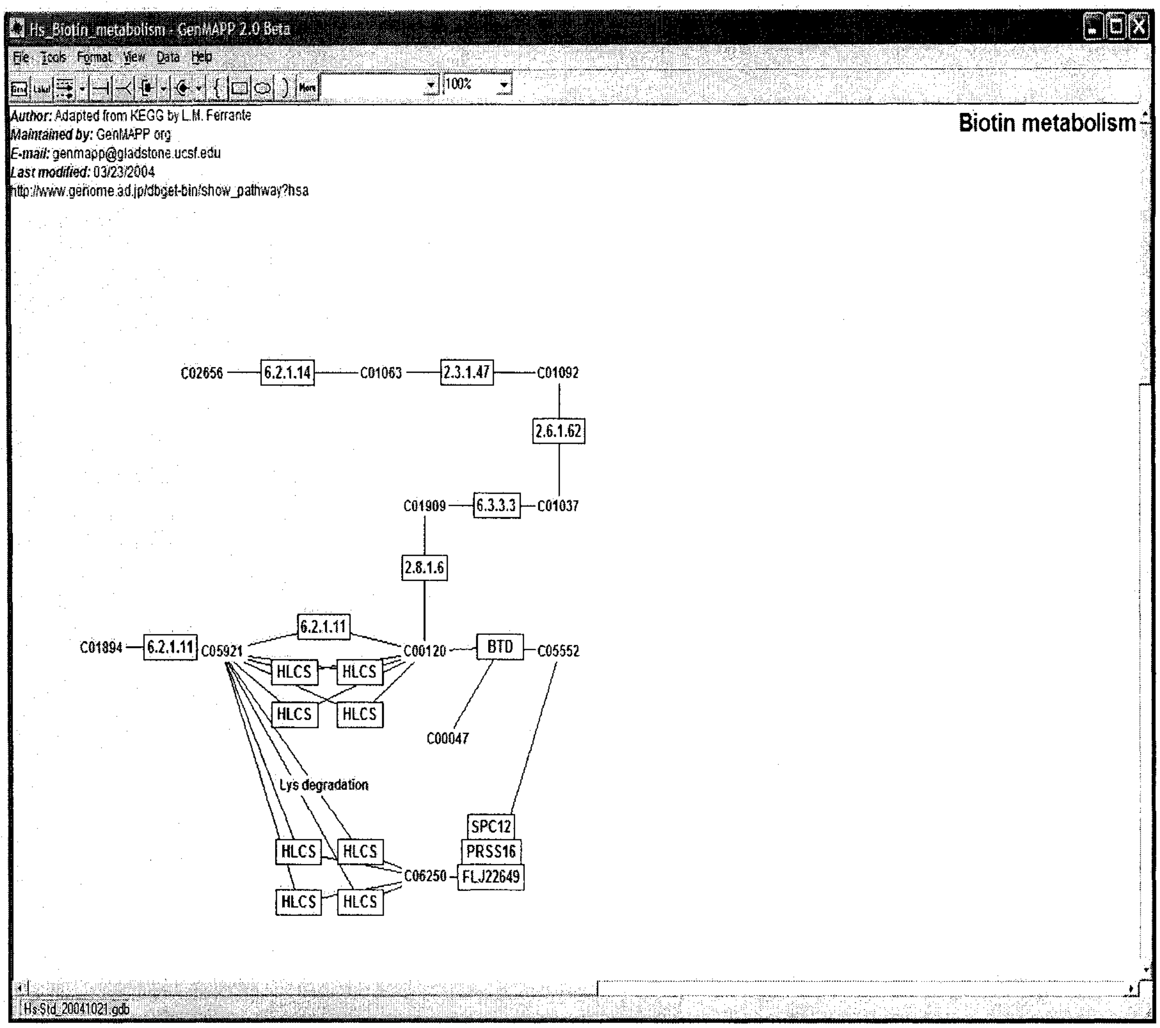
☐ At least 1 Observations abs(Val) >= 2

☐ MaxVal - MinVal >= 2

Figure 5. Main interface of the Cluster program.

GenMAPP (Gene MicroArray Pathway Profiler) is a package designed for visualizing gene expression data in a biological context with the graphical and more intuitive format of MAPPs (Dahlquist et al. 2002; Doniger et al. 2003; Eisen et al. 1998). A MAPP is a GenMAPP-produced file that graphically shows the biological relationship between genes or gene products. MAPPs can be used to group genes and view data by any organizing principle, such as metabolic pathways, signal transduction cascades, subcellular locations, gene families, or lists of genes associated with Gene Ontology categories. MAPPs for a standard set of biological pathways, as well as lists of functionally related genes from public sources such as the Gene Ontology Project, may be downloaded at www.GenMAPP.org. In addition, custom MAPPs for hypothesis testing may be drawn with the graphics tools provided by the GenMAPP program. It is a powerful tool for interpreting gene expression microarray data.

A sample MAPP of the biotin metabolism pathway is illustrated in Figure 6. MAPP information is shown at the top of the Drafting Board, including title, author, maintained by, email, last modified, remarks, copyright, and notes. The MAPP itself consists of objects (e.g., 2.3.1.47 and 6.2.1.14), links between objects (e.g., C01063), and labels (e.g., Lys degradation). The gene object represents a biological gene or gene product and is the link between the gene object on the MAPP and information in the Gene Database. The label is the text that the user wishes to appear on the MAPP.



Hs_Biotin_metabolism - GenMAPP 2.0 Beta

Author: Adapted from KEGG by L.M. Ferrante
Maintained by: GenMAPP org
E-mail: genmapp@gladstone.ucsf.edu
Last modified: 03/23/2004
http://www.geriome.ad.jp/dbget-bin/show_pathway?hsa

Hs-Std_20041021.gdb

Figure 6. The Human biotin metabolism MAPP generated by the GenMAPP program

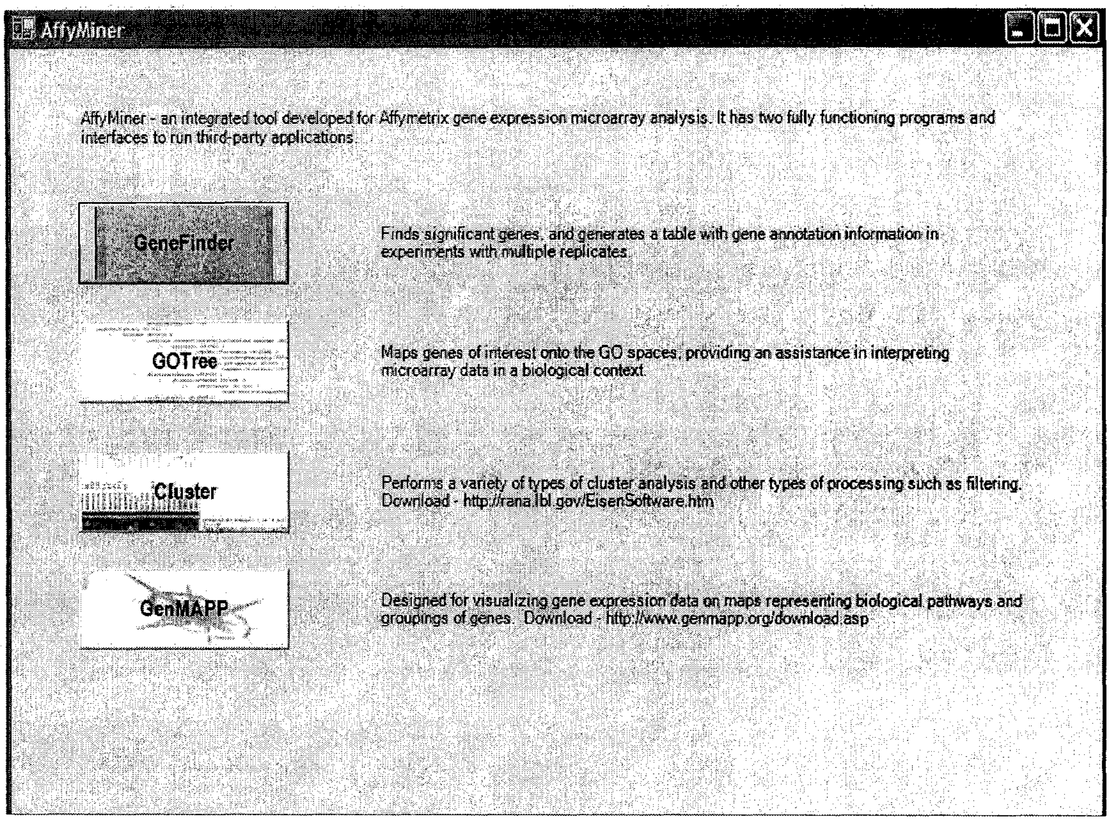
Chapter 3 AffyMiner – an Integrated Tool for Microarray Analysis

This chapter will introduce AffyMiner, the program that I developed for Affymetrix microarray analysis. The functionality for each component and the system requirements will be described.

AffyMiner was developed in the Microsoft .Net platform and programmed in Visual Basic .Net. VB .Net is the latest version of the Microsoft Visual Basic language. It has many attractive features, such as easy of use, fully object-oriented, and true visual development.

3.1 An introduction to AffyMiner

AffyMiner is an integrated bioinformatics tool developed for Affymetrix microarray data analysis. It includes two newly developed programs, GeneFinder and GOTree, and interfaces with two widely used packages, Cluster and GenMAPP. Figure 7 shows the AffyMiner main window. A brief description of AffyMiner and its components is available on this window. The user can run any of the above four programs by simply clicking an appropriate button.



AffyMiner - an integrated tool developed for Affymetrix gene expression microarray analysis. It has two fully functioning programs and interfaces to run third-party applications.

Finds significant genes, and generates a table with gene annotation information in experiments with multiple replicates.

Maps genes of interest onto the GO spaces, providing an assistance in interpreting microarray data in a biological context.

Performs a variety of types of cluster analysis and other types of processing such as filtering. Download - http://rana.lbl.gov/EisenSoftware.htm

Designed for visualizing gene expression data on maps representing biological pathways and groupings of genes. Download - http://www.genmapp.org/download.asp

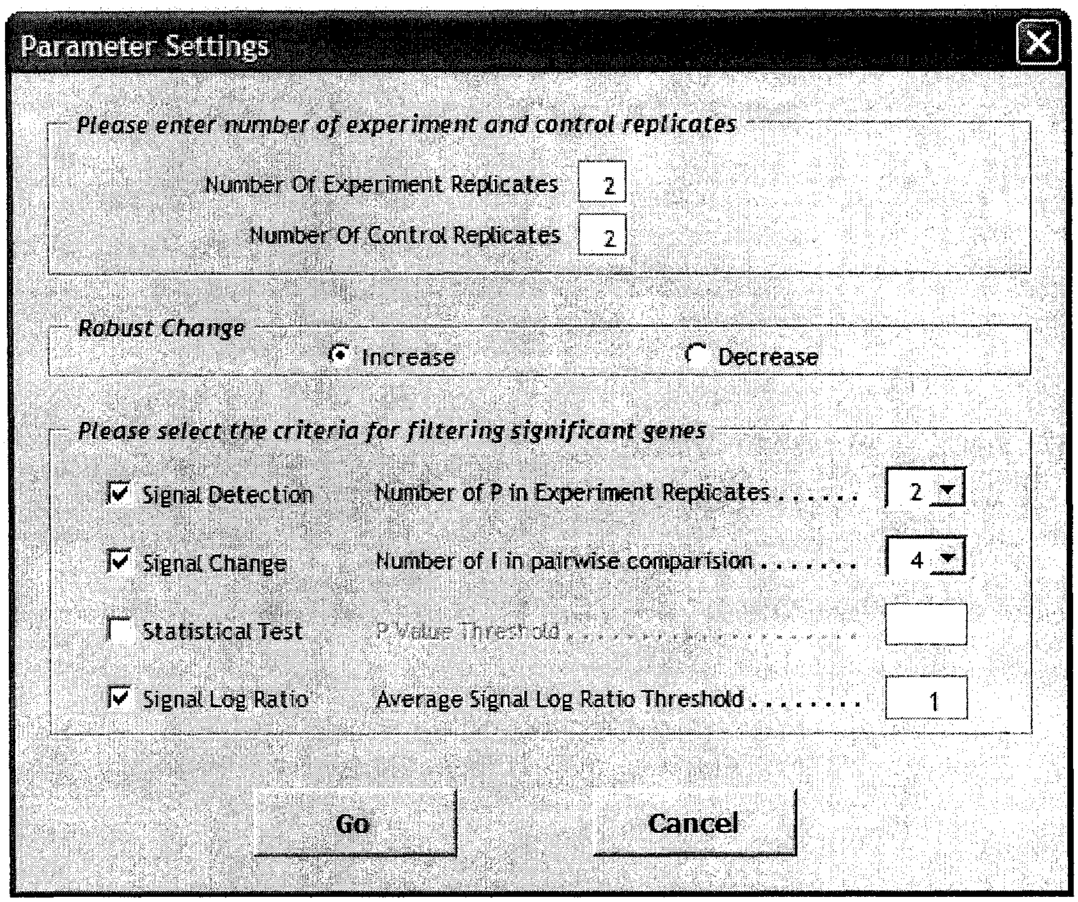
Figure 7. AffyMiner main window

3.1.1 GeneFinder

The input file to GeneFinder is a text file exported from Affymetrix GCOS software where single array analysis and pairwise comparison analysis are completed or from Affymetrix Data Mining Tools where statistics tests are performed. Data from different comparisons need to be combined into a single input file in a text format. The input file may contain qualitative data such as Detection, Change and quantitative data, e.g., Signal Log Ratio. Another input file needed is a NetAffx annotation file, which can be downloaded from the NetAffx Analysis Center (http://www.netaffx.com).

GeneFinder has three windows used to set up filtering parameters, upload input files, and define the output, respectively (Fig. 8, 9, 10). There are three frames in the parameter-setting window for setting the number of replicates, the direction of a robust change, and the data metrics to determine significant genes (Fig. 8). The data metrics consist of Signal Detection, Signal Change, Signal Log Ratio and Statistical Test. Note that the user has the flexibility of choosing which data matrices to use and setting threshold values. Once the parameters are set up in the first window, clicking the "Go" button will open another window for data input. To close the window, click the Cancel button.

Figure 8 demonstrates that two experiment replicates and two control replicates were used in the assay. The radio button Increase was checked for finding genes with robust increase. The data metrics used for finding significant genes included Signal Detection, Signal Change, and Signal Log Ratio. The signal detection was set to two, indicating two experiment replicates with signal detection as Present. The signal change was set to four, meaning four pairwise comparisons with signal change as Increase. The four comparisons are experiment 1 vs. control 1, experiment 1 vs. control 1, experiment 2 vs. control 1, experiment 2 vs. control 2. The average signal log ratio of one represents a two-fold change of signals in the experiment samples compared with the control samples. Note that the setting shown in Figure 8 is the most conservative one. Relaxing the conditions, for example, setting one for signal detection and three for signal change, would allow finding more genes.



Parameter Settings

Please enter number of experiment and control replicates

Number Of Experiment Replicates 2

Number Of Control Replicates 2

Robust Change

● Increase ○ Decrease

Please select the criteria for filtering significant genes

☑ Signal Detection Number of P in Experiment Replicates 2

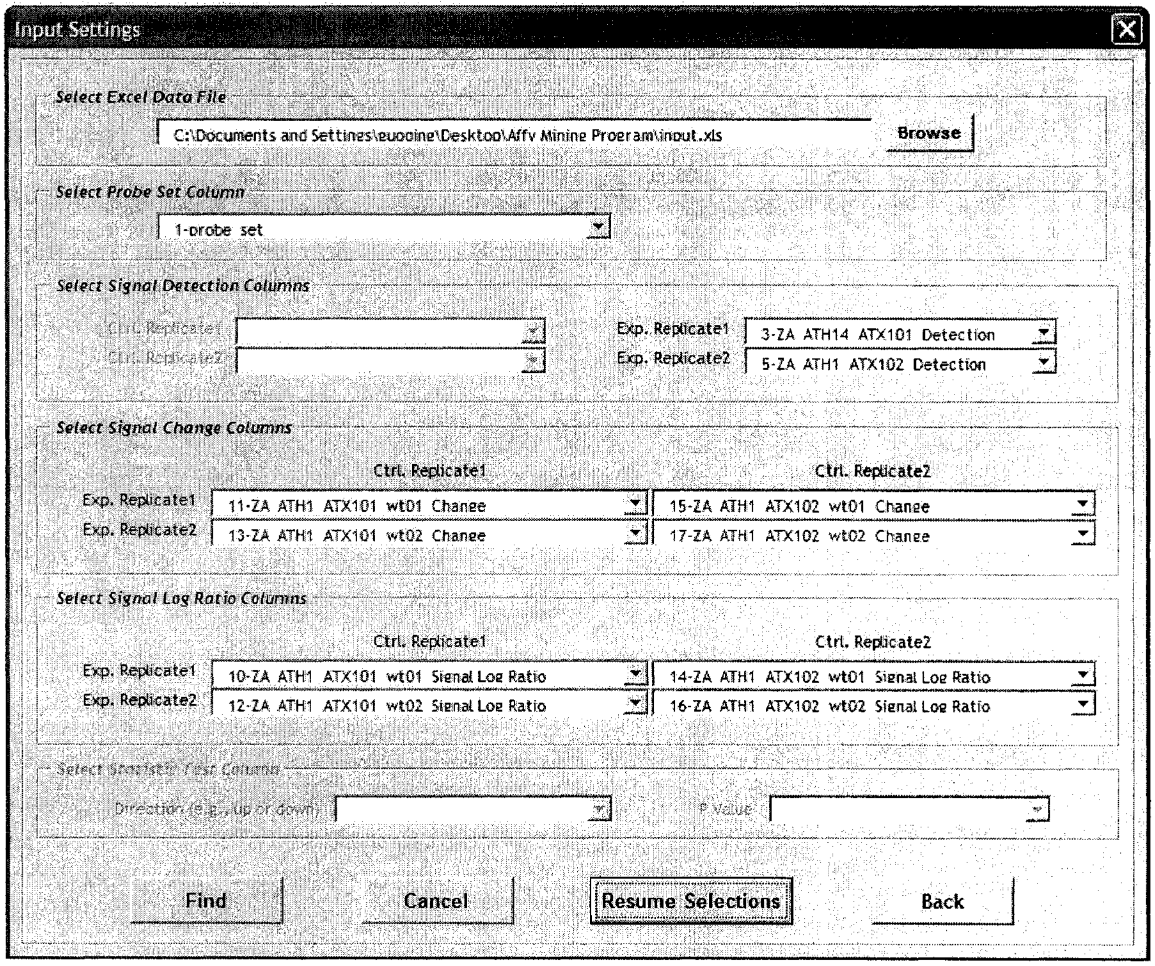
☑ Signal Change Number of I in pairwise comparision 4

☐ Statistical Test P Value Threshold . . . .

☑ Signal Log Ratio Average Signal Log Ratio Threshold 1

Figure 8. GeneFinder parameter setting window

In the data input window as shown in Figure 9, the user can choose the input file and select columns corresponding to specific samples and data metrics. The input data are exported from the Affymetrix GCOS software and saved in a text file. To change parameter settings, just click the "Back" button to the first window. If the program has been run once, clicking the "Resume Selections" button resumes the latest selections of the data file and columns. Clicking "Find" button starts the analysis process. As shown in Figure 9, column 1 of the input table contains the probe set, columns 3 and 5 have the signal detection values for two experiment replicates whereas columns 11, 13, 15, 17 comprise signal change values, and columns 10, 12, 14, 16 consist of the signal log ratios for the four pairwise comparisons.



Input Settings

Select Excel Data File

C:\Documents and Settines\euooine\Desktoo\Affv Minine Proeram\inout.xls

Select Probe Set Column

1-orobe set

Select Signal Detection Columns

Ctrl. Replicate1:

Ctrl. Replicate2:

Exp. Replicate1: 3-ZA ATH14 ATX101 Detection

Exp. Replicate2: 5-ZA ATH1 ATX102 Detection

Select Signal Change Columns

	Ctrl. Replicate1	Ctrl. Replicate2
Exp. Replicate1	11-ZA ATH1 ATX101 wt01 Chanee	15-ZA ATH1 ATX102 wt01 Chanee
Exp. Replicate2	13-ZA ATH1 ATX101 wt02 Chanee	17-ZA ATH1 ATX102 wt02 Chanee

Select Signal Log Ratio Columns

	Ctrl. Replicate1	Ctrl. Replicate2
Exp. Replicate1	10-ZA ATH1 ATX101 wt01 Sienal Loe Ratio	14-ZA ATH1 ATX102 wt01 Sienal Loe Ratio
Exp. Replicate2	12-ZA ATH1 ATX101 wt02 Sienal Loe Ratio	16-ZA ATH1 ATX102 wt02 Sienal Loe Ratio

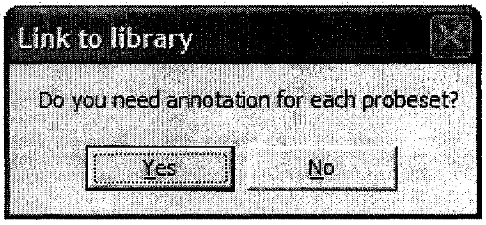
Select Statistic Test Column

Direction (e.g., up or down):

P Value:

Figure 9. GeneFinder input setting window

When the analysis is done, a window will be popupped, asking whether the user needs annotation for each probe set (Fig. 10). If "no" is clicked, a report on significant genes will be generated. If "yes" is clicked, the output-setting window will be displayed.

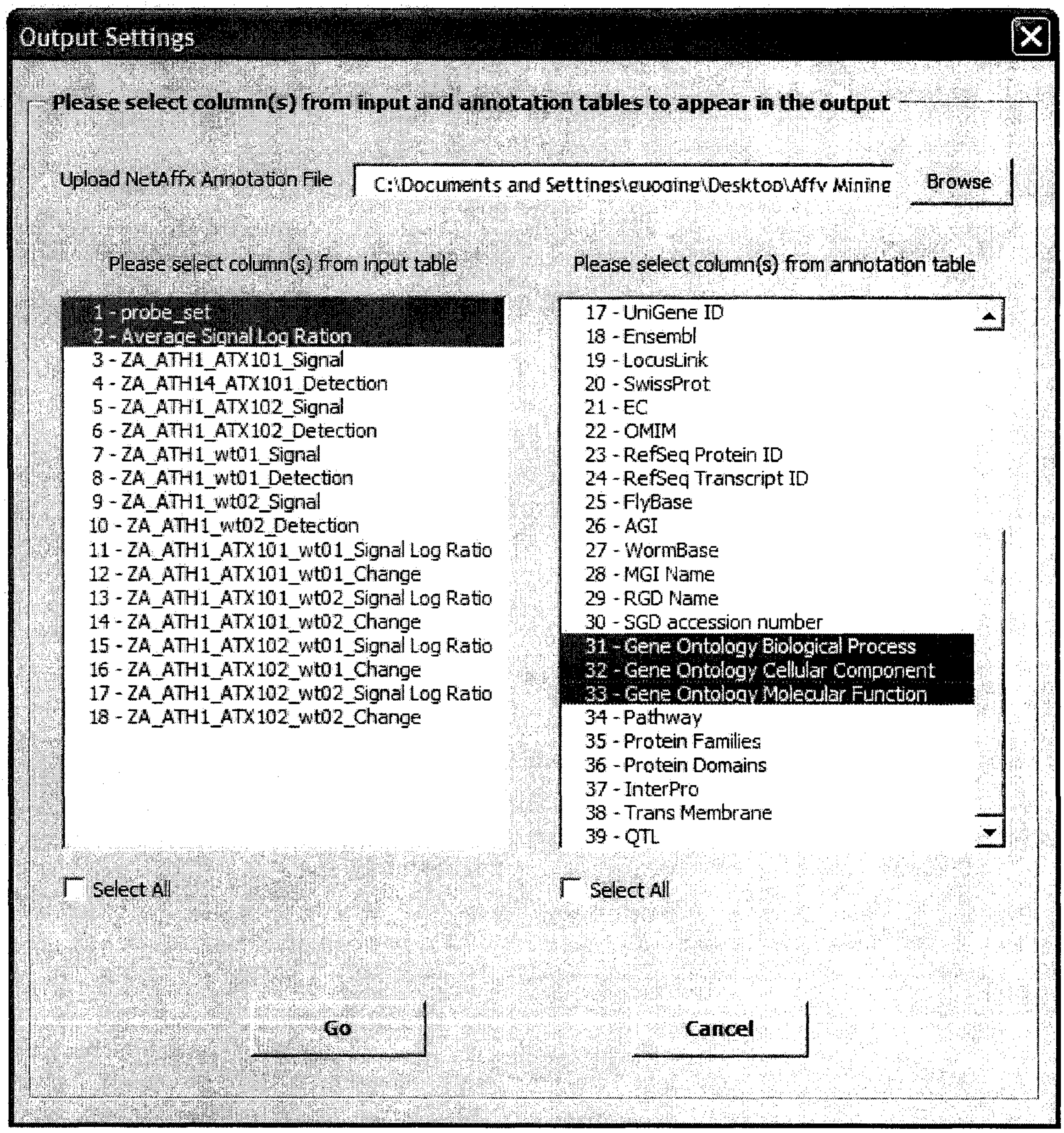

Link to library

Do you need annotation for each probeset?

Figure 10. Gene annotation popup window

The output-setting window allows the user uploading the annotation file and choosing columns to be included in the output table. The gene annotation file needs to be in the CSV (Comma Separated Value) format and can be downloaded from the NetAffx website (http://www.affymetrix.com/analysis/index.affx). In Figure 11, the user chose column names, Probe_set, Average Signal Log Ratio, Transcript ID, Target Description, Gene Ontology Biological Process, Gene Ontology Cellular Component, and Gene Ontology Molecular Function.

The output from GeneFinder is shown in Table 7. It comprises the probe set IDs, the average signal log ratio, and other columns selected by the user in the output-setting step. We did not show the whole table here and presented only six probe sets. The Gene Ontology Biological Process, Gene Ontology Cellular Component, and Gene Ontology Molecular Function were selected for GOTree to create a Gene Ontology tree for significant genes (see also 3.2.1).



Please select column(s) from input and annotation tables to appear in the output

Upload NetAffx Annotation File: C:\Documents and Settings\guoqing\Desktop\Affy Mining

Please select column(s) from input table

1 - probe_set
2 - Average Signal Log Ration
3 - ZA_ATH1_ATX101_Signal
4 - ZA_ATH14_ATX101_Detection
5 - ZA_ATH1_ATX102_Signal
6 - ZA_ATH1_ATX102_Detection
7 - ZA_ATH1_wt01_Signal
8 - ZA_ATH1_wt01_Detection
9 - ZA_ATH1_wt02_Signal
10 - ZA_ATH1_wt02_Detection
11 - ZA_ATH1_ATX101_wt01_Signal Log Ratio
12 - ZA_ATH1_ATX101_wt01_Change
13 - ZA_ATH1_ATX101_wt02_Signal Log Ratio
14 - ZA_ATH1_ATX101_wt02_Change
15 - ZA_ATH1_ATX102_wt01_Signal Log Ratio
16 - ZA_ATH1_ATX102_wt01_Change
17 - ZA_ATH1_ATX102_wt02_Signal Log Ratio
18 - ZA_ATH1_ATX102_wt02_Change

□ Select All

Please select column(s) from annotation table

17 - UniGene ID
18 - Ensembl
19 - LocusLink
20 - SwissProt
21 - EC
22 - OMIM
23 - RefSeq Protein ID
24 - RefSeq Transcript ID
25 - FlyBase
26 - AGI
27 - WormBase
28 - MGI Name
29 - RGD Name
30 - SGD accession number
31 - Gene Ontology Biological Process
32 - Gene Ontology Cellular Component
33 - Gene Ontology Molecular Function
34 - Pathway
35 - Protein Families
36 - Protein Domains
37 - InterPro
38 - Trans Membrane
39 - QTL

□ Select All

Figure 11. GeneFinder output setting window

Table 7 Part of the output generated from GeneFinder

Probe_set	Average Signal Log Ratio	Transcript ID	Target Description	Gene Ontology Biological Process	Gene Ontology Cellular Component	Gene Ontology Molecular Function
267580_at	1.0175	At2g41990	Unknown protein			
267524_at	1.405	At2g30600	Unknown protein			5515 // protein binding // inferred from electronic annotation
267523_at	1.2675	At2g30610	Unknown protein			5515 // protein binding // inferred from electronic annotation
267461_at	1.825	At2g33830	putative auxin-regulated protein; supported by full-length cDNA:	6468 // protein amino acid phosphorylation //		4672 // protein kinase activity // inferred from electronic annotation ///

			Ceres:1711.	inferred from electronic annotation		5524 // ATP binding // inferred from electronic annotation
267388_at	1.8825	At2g44450	putative beta-glucosidase	5975 // carbohydrate metabolism // inferred from electronic annotation	12505 // endomembrane system // inferred from electronic annotation	
267337_at	1.455	At2g39980	putative anthocyanin 5-aromatic acyltransferase; supported by cDNA: gi_13937225_gb_AF372968.1_AF372968			

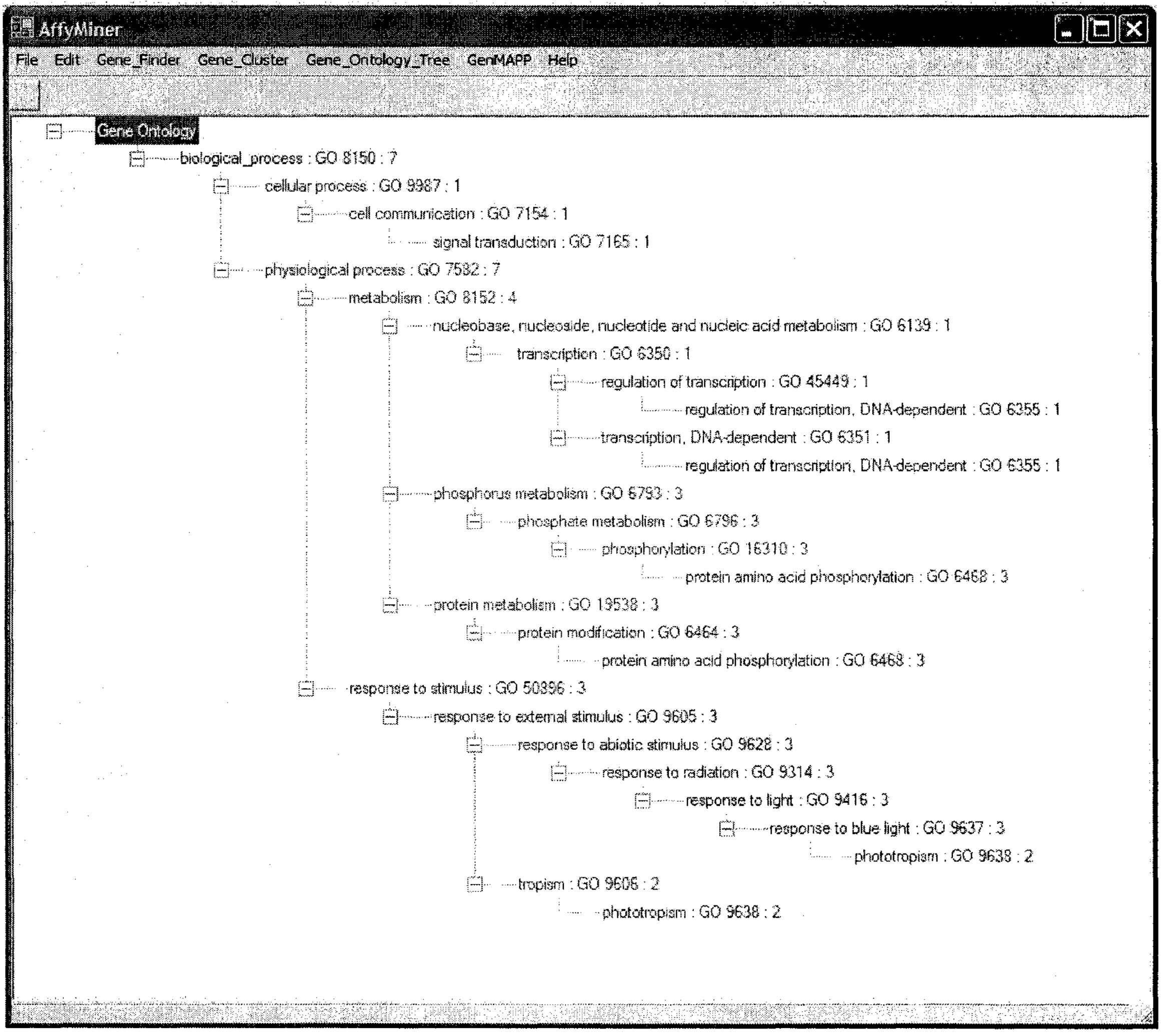
3.1.2 GOTree

GOTree takes as input two files. One file is called GOPath with information about hierarchical structure of GO terms whereas the other file has information regarding genes of interest and their GO term associations. Both files are described in 2.4.2. The GOPath file was generated from ChipInfo. The user can download it from the following website: http://www.biostat.harvard.edu/complab/chipinfo/. To run GeneChip, the user needs to provide the gene information file, which can be downloaded from the Affymetrix website.

The output GO tree can be presented at various levels. The user can click the small square box to expand or collapse the branches. Each node is labeled with the corresponding GO term, GO ID, and the number of genes associated. For example, line 3 of the Gene Ontology tree as shown in Figure 12 indicates the node represents cellular process with GO ID 9987 and 1 GO term associated. In addition, the GO tree can be saved in a GIF file, and be copied or pasted in a word document for publication.

3.1.3 Interfaces to Cluster and GenMAPP

The third component links AffyMiner with Cluster and GenMAPP. The user needs to download Cluster and GenMAPP Cluster respectively from the following websites: http://rana.lbl.gov/EisenSoftware.htm and http://www.genmapp.org/download.asp. Both Cluster and GenMAPP need to be installed on the local computer. The user can run them by merely clicking the buttons in the main window (Fig. 7). The data transformation function available in AffyMiner format Affymetrix microarray data or GeneFinder result data automatically for Cluster or GenMAPP. To learn more functions available in both programs, the user needs to refer to the user manuals or software tutorials available on the website.



```
Gene Ontology
    biological_process : GO 8150 : 7
        cellular process : GO 9987 : 1
            cell communication : GO 7154 : 1
                signal transduction : GO 7165 : 1
        physiological process : GO 7582 : 7
            metabolism : GO 8152 : 4
                nucleobase, nucleoside, nucleotide and nucleic acid metabolism : GO 6139 : 1
                    transcription : GO 6350 : 1
                        regulation of transcription : GO 45449 : 1
                            regulation of transcription, DNA-dependent : GO 6355 : 1
                        transcription, DNA-dependent : GO 6351 : 1
                            regulation of transcription, DNA-dependent : GO 6355 : 1
                phosphorus metabolism : GO 6793 : 3
                    phosphate metabolism : GO 6796 : 3
                        phosphorylation : GO 16310 : 3
                            protein amino acid phosphorylation : GO 6468 : 3
                protein metabolism : GO 19538 : 3
                    protein modification : GO 6464 : 3
                        protein amino acid phosphorylation : GO 6468 : 3
            response to stimulus : GO 50896 : 3
                response to external stimulus : GO 9605 : 3
                    response to abiotic stimulus : GO 9628 : 3
                        response to radiation : GO 9314 : 3
                            response to light : GO 9416 : 3
                                response to blue light : GO 9637 : 3
                                    phototropism : GO 9638 : 2
                    tropism : GO 9606 : 2
                        phototropism : GO 9638 : 2
```

Figure 12. A Gene Ontology tree generated from the GOTree program in AffyMiner.

3.1.4 The use of AffyMiner

Researchers at the University of Nebraska-Lincoln (UNL) have been using AffyMiner for microarray analysis. They found AffyMiner was an easy-to-use and effective tool and saved plenty of time and efforts needed for analyzing microarray data. In fact, AffyMiner has been cited in several manuscripts, which have been or will be submitted for publication (e.g., Z. Avramova, personal commun. 2004; M. Fromm, personal commun. 2005). In this section, I want to show a practical example in which I used AffyMiner to analyze Affymetrix microarray data.

Dr. Z. Avramova's group at the University of Nebraska-Lincoln was interested in knowing whether phosphoinositid 5-Phosphate and the yrithorax homolog, ATX1, define a novel signaling pathway in Arabidopsis. By doing so, two independent experiments were carried out. PI5P-treated, mock-treated wild type, and atx1 mutant plants were grown and handled under the same conditions. Whole plants, grown for 20 hours in the presence of exogenously added PI5P, were harvested and quickly transferred into liquid nitrogen. Total RNA was extracted from the frozen plants using TRIzol reagent following the manufacturer's instructions and further purified using Qiagen RNeasy column (Qiagen). Fifteen micrograms of total RNA was used to synthesize cDNA using Affymetrix One-Cycle cDNA Synthesis Kit according to the manufacturer's instructions (Affymetrix). All sample preparations followed prescribed protocols (Affymetrix Genechip Expression Analysis Technical manual). Hybridization was done on an Affymetrix Arabidopsis Genome ATH1 Array, stained with streptavidin-phycoerythrin conjugate on an Affymetrix Fluidics Station 450, followed by scanning with the GeneChip Scanner 3000 (Affymetrix). Affymetrix GCOS was used for washing, scanning, and basic data analysis.

For microarray data analysis, six Affymetrix GeneChip Arabidopsis ATH1 Genome Arrays were used for the analysis of gene expression in PI5P treated and ATX1 mutant Arabidopsis, by measuring florescence from gene-specific oligos. The Affymetrix microarray contains more than 22,500 probe sets, representing approximately 24,000 genes. Each probe set consists of 11 probe pairs with a perfect mach (PM) sequence corresponding to a specific region of a gene. For each PM sequence, there is also a corresponding mismatch (MM) oligo that differs by one base. In total, six microarray hybridizations were carried out and each experimental sample was analyzed versus each of the two wild type control sets. The numbers of total Arabidopsis genes detected by each individual hybridization were 60.4% and 57.4% for the wild type, 62.2% and 59.9% for the *atx1*, 57.7% and 58.5% for the PI5P-treated plants. Thereby, 60%, ~14,800 of all *Arabidopsis* genes have been detected by the analysis. For mining significant genes, the AffyMiner program was used and the criteria for defining genes with robust increase or decrease in expression.

The data were first analyzed with GCOS. For each array, overall intensity normalization for the entire probe sets was preformed using the scaling approach, which adjusts the average intensity or signal value of every array to a common value in order to make the arrays comparable. The target signal intensity 500 was set up for scaling. Single array analysis generated a detection p-value to determine the detection call, Present (P) or

Absent (A). Additionally, a signal value, a relative measure of abundance to the transcript, was calculated. For comparison analysis, the array for wild type (wt) is used as a baseline and the arrays for PI5P treated or ATX1 mutant were used as treatment. Instead of simply comparing signal values of each probe set, GCOS examines changes in the intensities of both PM and MM probes between the treatment and the baseline using a non-parametric Wilcoxon rank test. However, this method is available only for pairwise array comparison as discussed in 1.5. To compare two replicates for the treatment (e.g., PI5P1 and PI5P2) and two replicates for the control (i.e., wt01 and wt02), the data were exported from the GCOS and used by AffyMiner. The GeneFinder program in AffyMiner was used to identify genes significantly expressed in PI5P treated samples compared with the wild-type samples. The following criteria were used in calculation: i) detection call should be "present" in the 2 experiment replicates; ii) change calls from the pairwise comparisons should be all "I", i.e., increase, or ","D", decrease; iii) average fold change between the treatments and the controls should be no less that 1.5. The output table is shown in Appendix A.

To derive a GO tree, I used the GOPath file generated from the ChipInfo program and the output file from GeneFinder as input to the GOTree program. The tree is illustrated partially in Figure 12 and summarized in Figure 13. Of the genes with robust changes, 55.2% are related to metabolisms, 16.3% to cellular physical process, and 10% to response to stimulus.

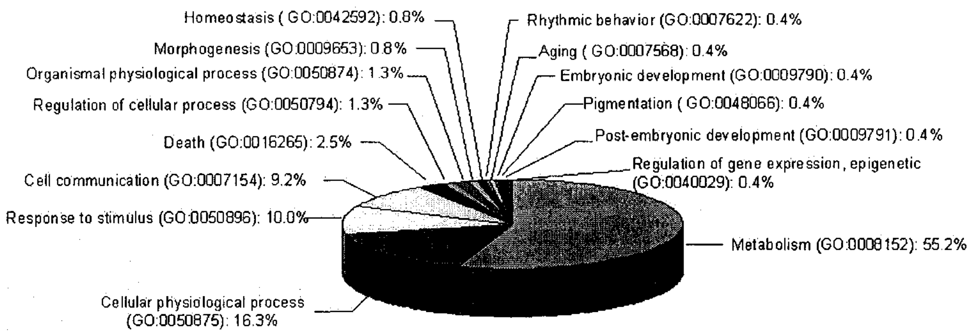


Figure 13. Distribution of genes with significantly altered expression levels after PI5P treatment according to the Gene Ontology tree generated in GOTree

The above results generated from AffyMiner were summarized in a manuscript, which has been submitted to the European Molecular Biology Organization (EMBO) journal for the consideration of publication.

The above exercise demonstrated that AffyMiner is not only an easy-to-use but also an effective tool. Without using GeneFinder, the user can sort for significant genes manually following the rules as described in 2.4.1. This manual approach is not only time consuming but could not easily deal with conflicts among different data metrics (for example, the conflict between signal log ratio and Signal Chang). AffyMiner provides the user with flexibility of setting up different threshold values for the data metrics, i.e., Signal Detection, Signal Change, Signal Log Ration, and Statistic Test. The user can play as many combinations as one likes and decide the final gene list. Finding significant genes with AffyMiner takes less than two minutes, however the manual process may take several hours or even longer to get the job done. Therefore, AffyMiner significantly reduces time and efforts needed for microarray data analysis.

3.2 System requirements

- Pentium III-class 600 MHz
- Microsoft Windows 2000 or above
- 256 MB RAM

Chapter 4 Discussion

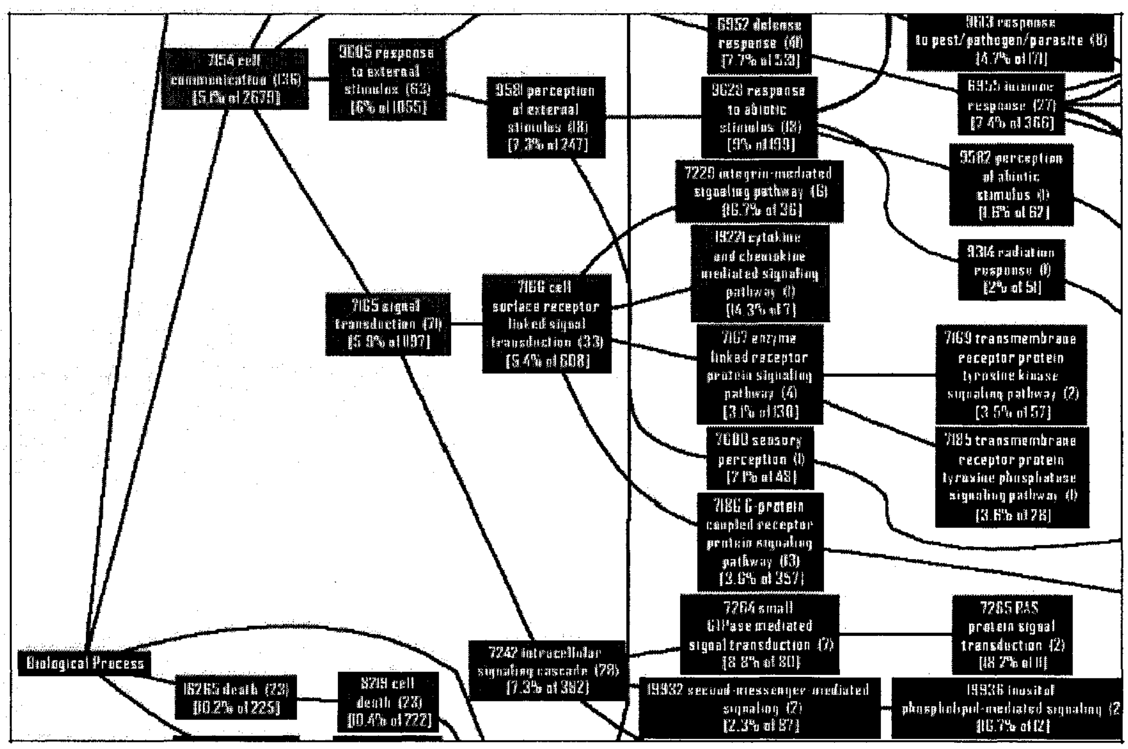
In this chapter, I will compare AffyMiner with its counterparts in order to demonstrate its unique features. Software limitations and future improvement of AffyMiner will be discussed as well. This chapter will end with a conclusion on the research work, i.e., the development of an integrated bioinformatics tool for microarray analysis.

4.1 Software comparison between AffyMiner and its counterparts

The challenge for microarray experiments is to analyze data and interpret the result in terms of biology rather than the generation of array data. In this project, I developed an integrated bioinformatics tool called AffyMiner to assist researchers in analyzing microarray data. AffyMiner has three interconnected components, GeneFinder, GOTree, and interfaces linking AffyMiner to other third-party programs.

Affymetrix, Inc. has developed the GeneChip® Operating Software (GCOS) and the Data Mining Tool (DMT) software (http://www.affymetrix.com/index.affx), with which AffyMiner has certain functions overlapped. For example, the GCOS software can perform single comparison analysis (a baseline versus a treatment) and sorting, and the DMT has functions used for array data filtering. However, both the GCOS and the DMT cannot be used for multi-comparison analysis of replicate samples (Fig. 4). AffyMiner provides users flexibility for choosing different data metrics (Signal Detection, Signal

Change, Signal Log Ratio, and Statistic Test) and setting threshold values in order to find significant genes. Moreover, both the GCOS and the DMT software do not have such a function that allows incorporating NetAffx gene annotation information into the final result. In this aspect, AffyMiner can include gene annotations in the analysis and result in a ready-to-publish user defined table (Table 7 and Appendix A). The NetAffx Gene Ontology Mining Too, made available by Affymetrix, Inc., provides a graphical view of probe set representation within the biological processes, molecular function, or cellular component hierarchies (Fig. 14). However, the graph is very difficult to read, which is the main reason driving us to develop GOTree. GOTree has flexibility of displaying the GO tree on different levels.



7154 cell communication (136) [5.1% of 2679]
9605 response to external stimulus (63) [6% of 1055]
9581 perception of external stimulus (18) [7.3% of 247]
6952 defense response (41) [7.7% of 533]
9613 response to pest/pathogen/parasite (8) [4.7% of 171]
9628 response to abiotic stimulus (18) [9% of 199]
6955 immune response (27) [7.4% of 366]
7229 integrin-mediated signaling pathway (6) [16.7% of 36]
9582 perception of abiotic stimulus (1) [1.6% of 62]
19221 cytokine and chemokine mediated signaling pathway (1) [14.3% of 7]
9314 radiation response (1) [2% of 51]
7165 signal transduction (71) [5.9% of 1197]
7166 cell surface receptor linked signal transduction (33) [5.4% of 608]
7167 enzyme linked receptor protein signaling pathway (4) [3.1% of 130]
7169 transmembrane receptor protein tyrosine kinase signaling pathway (2) [3.5% of 57]
7600 sensory perception (1) [7.1% of 48]
7185 transmembrane receptor protein tyrosine phosphatase signaling pathway (1) [3.6% of 28]
7186 G-protein coupled receptor protein signaling pathway (13) [3.6% of 357]
7264 small GTPase mediated signal transduction (7) [8.8% of 80]
7265 RAS protein signal transduction (2) [18.2% of 11]
Biological Process
7242 intracellular signaling cascade (28) [7.3% of 382]
16265 death (23) [10.2% of 225]
8219 cell death (23) [10.4% of 222]
19932 second-messenger-mediated signaling (2) [2.3% of 87]
19936 inositol phospholipid-mediated signaling (2 [16.7% of 12]

Figure 14. Sample graphical view of the biological process hierarchy generated by the NetAffx Gene Ontology Mining Tool

GenePicker (Finocchiaro et al. 2004) is a similar program to GeneFinder. It was developed for replicate analysis of Affymetrix gene expression microarrays. The analysis was done through definition of analysis schemes, data normalization, t-test/ANOVA, and Change-fold Chang-analysis, and the use of Change Call, Fold Change, and Signal mean ratios. GenePicker provides a comparison of noise and signal analysis scheme for determining a signal-to-noise in a given experiment, which is not available in GeneFinder. However, GeneFinder uses one more data matrix, i.e., Detection. Additionally, GeneFinder has the function incorporating gene annotation information with expression data in the result, which is not available in GenePicker.

The GoSurfer software was another tool developed for Affymetrix GeneChip data analysis (Li & Wong 2001; Zhong et al. 2003; Zhong et al. 2004). GoSurfer uses Gene Ontology information to analyze gene sets obtained from genome-wide microarray analysis. GoSurfer associates user input gene lists with GO terms and visualizes such GO terms as a hierarchical tree. GoSurfer compares two lists of genes in order to find which GO terms are enriched in one list of genes but relatively depleted in another. GoSurfer could not map genes from a single list onto the GO spaces. In this regard, GOTree and GoSurfer complement each other in the analysis of Gene Ontology.

As a whole, AffyMiner fills an important gap in finding significant genes from Affymetrix gene expression array data. AffyMiner filtering gene expression data for Affymetrix microarray users results in a list of genes showing significant changes in the experiment and generates a table in a format specified by the user that may include qualitative data (e.g., Detection, Change), quantitative data (e.g., Signal Log Ratio), and functional annotations (e.g., Gene Description, GO Molecular Function, and Pathway). AffyMiner has enhanced the capacity of existing Affymetrix software packages, i.e., the GCOS and the Data Mining Tool, and the NetAffx resource, and providing full flexibility for Affymetrix microarray data analysis and result interpretation.

AffyMiner has been tested by multiple users and their feedback was incorporated into its final implementation. Overall, AffyMiner greatly reduces the time and efforts needed to compare data from multiple arrays and provides the results in a flexible format dictated by the user.

4.2 Limitations of AffyMiner

AffyMiner is a Window application. It runs only in the Microsoft Window environments (MS Windows 2000 or above). Affymetrix has an inherent disadvantage, i.e., its dependence on the Affymetrix GCOS for the low-level analysis, including the single array analysis and pairwise comparison analysis, and on the NetAffx for gene annotation information. It seems not an easy task to adapt AffyMiner for analyzing array data generated from other platforms, such as cDNA microarrays.

4.3 Future directions

There is no a single approach suitable for all types of microarray analysis. The data sets, thus, need to be analyzed using a range of methods with increasing depth of inference (Leung & Cavalieri 2003; Lockhart & Winzeler 2000). It requires our future product be able to perform analysis at different levels. AffyMiner currently focuses on two important issues, finding differentially expressed genes and mapping genes of interest onto the GO tree. It is almost certain that there are functions missed in AffyMiner, for example, regulatory network inference. AffyMiner needs to get involved in the more ambitious realm of genetic network inference, where very promising results are reported in several recent papers (Altman & Raychaudhuri 2001; Friedman et al. 2000; Maki et al. 2001; Wu et al. 2004).

A database would be useful for managing gene expression data and storing analysis results. The database needs to compile with the MIAME standard (Brazma et al. 2001). A database would also allow effective data mining. In addition, the future version of AffyMiner should have a new function for the user to be able to integrate information from other resources, such as online databases, into the local database. It seems feasible since there have been efforts to combine expression data with other sources of information, which improves the range and quality of conclusions that can be drawn from microarray data analysis (Brazma & Vilo 2000).

Finally, in order to draw meaningful inferences from gene expression data, it is important to use an alternate technique to assay gene expression level. Researchers usually use the Real-Time PCR or the Western blot techniques to validate microarray findings. AffyMiner would become more powerful if it had the capability of dealing with validation data and the ability of integrating findings from microarray data and validation data.

4.4 Conclusion

In conclusion, we have developed an integrated bioinformatics tool, AffyMiner, for Affymetrix microarray data analysis and data mining. AffyMiner consists of two newly developed programs, GeneFinder and GOTree, and interfaces with third party programs. GeneFinder provides users flexibility of choosing different data metrics and effectively deals with multiple replicates to find significant genes in the experiment. Moreover, GeneFinder is capable of incorporating gene annotation information and generate a ready-to-publish table in a format specified by the user. GOTree maps genes of interest onto the GO spaces, which assists in the interpretation of findings in the context of biology. The interfaces provide a prototype of integrating open source software tools with AffyMiner, which is of great benefit to the user. AffyMiner has been tested and proved to be an effective tool that significantly reduced the time and efforts needed for microarray analysis. It is expected to implement more functions and a MIAME compliable database in the next version of AffyMiner.

References

Abbott, R. T., Tripp, S., Perkins, S. L., Elenitoba-Johnson, K. S. & Lim, M. S. 2003 Analysis of the PI-3-Kinase-PTEN-AKT pathway in human lymphoma and leukemia using a cell line microarray. *Mod Pathol* **16**, 607-12.

Akutsu, T., Miyano, S. & Kuhara, S. 2000 Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function. *J Comput Biol* **7**, 331-43.

Altman, R. B. & Raychaudhuri, S. 2001 Whole-genome expression analysis: challenges beyond clustering. *Curr Opin Struct Biol* **11**, 340-7.

Barczak, A., Rodriguez, M. W., Hanspers, K., Koth, L. L., Tai, Y. C., Bolstad, B. M., Speed, T. P. & Erle, D. J. 2003 Spotted long oligonucleotide arrays for human gene expression analysis. *Genome Res* **13**, 1775-85.

Blatt, M., Wiseman, S. & Domany, E. 1996 Superparamagnetic clustering of data. *Physical Review Letters* **76**, 3251-3254.

Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. 2003 A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-93.

Bottone, F. G., Jr., Martinez, J. M., Collins, J. B., Afshari, C. A. & Eling, T. E. 2003 Gene modulation by the cyclooxygenase inhibitor, sulindac sulfide, in human colorectal carcinoma cells: possible link to apoptosis. *J Biol Chem* **278**, 25790-801.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J. &

Vingron, M. 2001 Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* **29**, 365-71.

Brazma, A. & Vilo, J. 2000 Gene expression data analysis. *FEBS Lett* **480**, 17-24.

Bremnes, R. M., Veve, R., Gabrielson, E., Hirsch, F. R., Baron, A., Bemis, L., Gemmill, R. M., Drabkin, H. A. & Franklin, W. A. 2002 High-throughput tissue microarray analysis used to evaluate biology and prognostic significance of the E-cadherin pathway in non-small-cell lung cancer. *J Clin Oncol* **20**, 2417-28.

Bunney, W. E., Bunney, B. G., Vawter, M. P., Tomita, H., Li, J., Evans, S. J., Choudary, P. V., Myers, R. M., Jones, E. G., Watson, S. J. & Akil, H. 2003 Microarray technology: a review of new strategies to discover candidate vulnerability genes in psychiatric disorders. *Am J Psychiatry* **160**, 657-66.

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R. & Apweiler, R. 2004 The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res* **32 Database issue**, D262-6.

Chen, Z. Y. & Corey, D. P. 2002 Understanding inner ear development with gene expression profiling. *J Neurobiol* **53**, 276-85.

Cherkasova, E., Laassri, M., Chizhikov, V., Korotkova, E., Dragunsky, E., Agol, V. I. & Chumakov, K. 2003 Microarray analysis of evolution of RNA viruses: evidence of circulation of virulent highly divergent vaccine-derived polioviruses. *Proc Natl Acad Sci U S A* **100**, 9398-403.

Cheung, V. G., Morley, M., Aguilar, F., Massimi, A., Kucherlapati, R. & Childs, G. 1999 Making and reading microarrays. *Nat Genet* **21**, 15-9.

Chung, C. H., Bernard, P. S. & Perou, C. M. 2002 Molecular portraits and the family tree of cancer. *Nat Genet* **32 Suppl**, 533-40.

Chung, H. A., Hyodo-Miura, J., Kitayama, A., Terasaka, C., Nagamune, T. & Ueno, N. 2004 Screening of FGF target genes in Xenopus by microarray: temporal dissection of the signalling pathway using a chemical inhibitor. *Genes Cells* **9**, 749-61.

Collins, Y., Tan, D. F., Pejovic, T., Mor, G., Qian, F., Rutherford, T., Varma, R., McQuaid, D., Driscoll, D., Jiang, M., Deeb, G., Lele, S., Nowak, N. & Odunsi, K. 2004 Identification of differentially expressed genes in clinically distinct groups of serous ovarian carcinomas using cDNA microarray. *Int J Mol Med* **14**, 43-53.

Dahlquist, K. D., Salomonis, N., Vranizan, K., Lawlor, S. C. & Conklin, B. R. 2002 GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat Genet* **31**, 19-20.

Debouck, C. & Goodfellow, P. N. 1999 DNA microarrays in drug discovery and development. *Nat Genet* **21**, 48-50.

DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A. & Trent, J. M. 1996 Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* **14**, 457-60.

Dharmadi, Y. & Gonzalez, R. 2004 DNA microarrays: experimental issues, data analysis, and application to bacterial systems. *Biotechnol Prog* **20**, 1309-24.

Doniger, S. W., Salomonis, N., Dahlquist, K. D., Vranizan, K., Lawlor, S. C. & Conklin, B. R. 2003 MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol* **4**, R7.

Dresen, I. M., Husing, J., Kruse, E., Boes, T. & Jockel, K. H. 2003 Software packages for quantitative microarray-based gene expression analysis. *Curr Pharm Biotechnol* **4**, 417-37.

Dudoit, S., Gentleman, R. C. & Quackenbush, J. 2003 Open source software for the analysis of microarray data. *Biotechniques* **Suppl**, 45-51.

Duyk, G. M. 2002 Sharper tools and simpler methods. *Nat Genet* **32 Suppl**, 465-8.

Dyrskjot, L. 2003 Classification of bladder cancer by microarray expression profiling: towards a general clinical use of microarrays in cancer diagnostics. *Expert Rev Mol Diagn* **3**, 635-47.

Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. 1998 Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**, 14863-8.

Eizirik, D. L., Kutlu, B., Rasschaert, J., Darville, M. & Cardozo, A. K. 2003 Use of microarray analysis to unveil transcription factor and gene networks contributing to Beta cell dysfunction and apoptosis. *Ann N Y Acad Sci* **1005**, 55-74.

Epstein, C. B. & Butow, R. A. 2000 Microarray technology - enhanced versatility, persistent challenge. *Curr Opin Biotechnol* **11**, 36-41.

Erle, D. J., Koth, L., Abramson, O., Rodriguez, M. W. & Barczak, A. 2002 Use of spotted oligonucleotide arrays for large-scale analysis of mammalian gene expression. *Chest* **121**, 80S.

Fan, J., Tam, P., Woude, G. V. & Ren, Y. 2004 Normalization and analysis of cDNA microarrays using within-array replications applied to neuroblastoma cell response to a cytokine. *Proc Natl Acad Sci U S A* **101**, 1135-40.

Finocchiaro, G., Parise, P., Minardi, S. P., Alcalay, M. & Muller, H. 2004 GenePicker: replicate analysis of Affymetrix gene expression microarrays. *Bioinformatics* **20**, 3670-2.

Friedman, N., Linial, M., Nachman, I. & Pe'er, D. 2000 Using Bayesian networks to analyze expression data. *J Comput Biol* **7**, 601-20.

Getz, G., Levine, E. & Domany, E. 2000 Coupled two-way clustering analysis of gene microarray data. *Proc Natl Acad Sci U S A* **97**, 12079-84.

Gibson, G. 2002 Microarrays in ecology and evolution: a preview. *Mol Ecol* **11**, 17-24.

Gmuender, H. 2002 Perspectives and challenges for DNA microarrays in drug discovery and development. *Biotechniques* **32**, 152-4, 156, 158.

Grant, G. M., Fortney, A., Gorreta, F., Estep, M., Del Giacco, L., Van Meter, A., Christensen, A., Appalla, L., Naouar, C., Jamison, C., Al-Timimi, A., Donovan, J., Cooper, J., Garrett, C. & Chandhoke, V. 2004 Microarrays in cancer research. *Anticancer Res* **24**, 441-8.

Guffanti, A., Reid, J. F., Alcalay, M. & Simon, G. 2002 The meaning of it all: web-based resources for large-scale functional annotation and visualization of DNA microarray data. *Trends Genet* **18**, 589-92.

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J. M., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S., Fisk, D. G., Hirschman, J. E., Hong, E. L., Nash, R. S., Sethuraman, A., Theesfeld, C. L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S. Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E. M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T. & White, R. 2004 The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32 Database issue**, D258-61.

Hartemink, A. J., Gifford, D. K., Jaakkola, T. S. & Young, R. A. 2001 Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac Symp Biocomput*, 422-33.

Held, M., Gase, K. & Baldwin, I. T. 2004 Microarrays in ecological research: a case study of a cDNA microarray for plant-herbivore interactions. *BMC Ecol* **4**, 13.

Heller, M. J. 2002 DNA microarray technology: devices, systems, and applications. *Annu Rev Biomed Eng* **4**, 129-53.

Herrero, J., Al-Shahrour, F., Diaz-Uriarte, R., Mateos, A., Vaquerizas, J. M., Santoyo, J. & Dopazo, J. 2003 GEPAS: A web-based resource for microarray gene expression data analysis. *Nucleic Acids Res* **31**, 3461-7.

Herrero, J., Vaquerizas, J. M., Al-Shahrour, F., Conde, L., Mateos, A., Diaz-Uriarte, J. S. & Dopazo, J. 2004 New challenges in gene expression data analysis and the extended GEPAS. *Nucleic Acids Res* **32**, W485-91.

Holloway, A. J., van Laar, R. K., Tothill, R. W. & Bowtell, D. D. 2002 Options available--from start to finish--for obtaining data from DNA microarrays II. *Nat Genet* **32 Suppl**, 481-9.

Hughes, T. R., Mao, M., Jones, A. R., Burchard, J., Marton, M. J., Shannon, K. W., Lefkowitz, S. M., Ziman, M., Schelter, J. M., Meyer, M. R., Kobayashi, S., Davis, C., Dai, H., He, Y. D., Stephaniants, S. B., Cavet, G., Walker, W. L., West, A., Coffey, E., Shoemaker, D. D., Stoughton, R., Blanchard, A. P., Friend, S. H. & Linsley, P. S. 2001 Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol* **19**, 342-7.

Iglesias, A. H., Camelo, S., Hwang, D., Villanueva, R., Stephanopoulos, G. & Dangond, F. 2004 Microarray detection of E2F pathway activation and other targets in multiple sclerosis peripheral blood mononuclear cells. *J Neuroimmunol* **150**, 163-77.

Ivanov, I., Schaab, C., Planitzer, S., Teichmann, U., Machl, A., Theml, S., Meier-Ewert, S., Seizinger, B. & Loferer, H. 2000 DNA microarray technology and antimicrobial drug discovery. *Pharmacogenomics* **1**, 169-78.

Kane, M. D., Jatkoe, T. A., Stumpf, C. R., Lu, J., Thomas, J. D. & Madore, S. J. 2000 Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res* **28**, 4552-7.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. 2004 The KEGG resource for deciphering the genome. *Nucleic Acids Res* **32 Database issue**, D277-80.

Khatri, P., Bhavsar, P., Bawa, G. & Draghici, S. 2004 Onto-Tools: an ensemble of web-accessible, ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments. *Nucleic Acids Res* **32**, W449-56.

Kononen, J., Bubendorf, L., Kallioniemi, A., Barlund, M., Schraml, P., Leighton, S., Torhorst, J., Mihatsch, M. J., Sauter, G. & Kallioniemi, O. P. 1998 Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat Med* **4**, 844-7.

Leung, Y. F. & Cavalieri, D. 2003 Fundamentals of cDNA microarray data analysis. *Trends Genet* **19**, 649-59.

Levy, S. E. 2003 Microarray analysis in drug discovery: an uplifting view of depression. *Sci STKE* **2003**, pe46.

Li, C. & Wong, W. H. 2001 Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A* **98**, 31-6.

Liang, S., Fuhrman, S. & Somogyi, R. 1998 Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac Symp Biocomput*, 18-29.

Lipshutz, R. J., Fodor, S. P., Gingeras, T. R. & Lockhart, D. J. 1999 High density synthetic oligonucleotide arrays. *Nat Genet* **21**, 20-4.

Liu, D. K., Yao, B., Fayz, B., Womble, D. D. & Krawetz, S. A. 2004 Comparative evaluation of microarray analysis software. *Mol Biotechnol* **26**, 225-32.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. & Brown, E. L. 1996

Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* **14**, 1675-80.

Lockhart, D. J. & Winzeler, E. A. 2000 Genomics, gene expression and DNA arrays. *Nature* **405**, 827-36.

Luo, Z. & Geschwind, D. H. 2001 Microarray applications in neuroscience. *Neurobiol Dis* **8**, 183-93.

Lyons-Weiler, J., Patel, S., Becich, M. J. & Godfrey, T. E. 2004 Tests for finding complex patterns of differential expression in cancers: towards individualized medicine. *BMC Bioinformatics* **5**, 110.

Maki, Y., Tominaga, D., Okamoto, M., Watanabe, S. & Eguchi, Y. 2001 Development of a system for the inference of large scale genetic networks. *Pac Symp Biocomput*, 446-58.

Medlin, J. 2001 Array of hope for gene technology. *Environ Health Perspect* **109**, A34-7.

Mircean, C., Tabus, I., Kobayashi, T., Yamaguchi, M., Shiku, H., Shmulevich, I. & Zhang, W. 2004 Pathway analysis of informative genes from microarray data reveals that metabolism and signal transduction genes distinguish different subtypes of lymphomas. *Int J Oncol* **24**, 497-504.

Nisenbaum, L. K. 2002 The ultimate chip shot: can microarray technology deliver for neuroscience? *Genes Brain Behav* **1**, 27-34.

Okamoto, T., Suzuki, T. & Yamamoto, N. 2000 Microarray fabrication with covalent attachment of DNA using bubble jet technology. *Nat Biotechnol* **18**, 438-41.

Pan, W. 2002 A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* **18**, 546-54.

Panda, S., Sato, T. K., Hampton, G. M. & Hogenesch, J. B. 2003 An array of insights: application of DNA chip technology in the study of cell biology. *Trends Cell Biol* **13**, 151-6.

Parrish, M. L., Wei, N., Duenwald, S., Tokiwa, G. Y., Wang, Y., Holder, D., Dai, H., Zhang, X., Wright, C., Hodor, P., Cavet, G., Phillips, R. L., Sun, B. I. & Fare, T. L. 2004 A microarray platform comparison for neuroscience applications. *J Neurosci Methods* **132**, 57-68.

Quackenbush, J. 2001 Computational analysis of microarray data. *Nat Rev Genet* **2**, 418-27.

Reinke, V. 2002 Functional exploration of the C. elegans genome using DNA microarrays. *Nat Genet* **32 Suppl**, 541-6.

Robson, B. & Garnier, J. 2002 The future of highly personalized health care. *Stud Health Technol Inform* **80**, 163-74.

Rouillard, J. M., Herbert, C. J. & Zuker, M. 2002 OligoArray: genome-scale oligonucleotide design for microarrays. *Bioinformatics* **18**, 486-7.

Rouillard, J. M., Zuker, M. & Gulari, E. 2003 OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucleic Acids Res* **31**, 3057-62.

Rubin, M. A., Varambally, S., Beroukhim, R., Tomlins, S. A., Rhodes, D. R., Paris, P. L., Hofer, M. D., Storz-Schweizer, M., Kuefer, R., Fletcher, J. A., Hsi, B. L., Byrne, J. A., Pienta, K. J., Collins, C., Sellers, W. R. & Chinnaiyan, A. M. 2004 Overexpression, amplification, and androgen regulation of TPD52 in prostate cancer. *Cancer Res* **64**, 3814-22.

Scandurro, A. B., Weldon, C. W., Figueroa, Y. G., Alam, J. & Beckman, B. S. 2001 Gene microarray analysis reveals a novel hypoxia signal transduction pathway in human hepatocellular carcinoma cells. *Int J Oncol* **19**, 129-35.

Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. 1995 Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467-70.

Shalev, A., Pise-Masison, C. A., Radonovich, M., Hoffmann, S. C., Hirshberg, B., Brady, J. N. & Harlan, D. M. 2002 Oligonucleotide microarray analysis of intact human pancreatic islets: identification of glucose-responsive genes and a highly regulated TGFbeta signaling pathway. *Endocrinology* **143**, 3695-8.

Simon, R., Radmacher, M. D., Dobbin, K. & McShane, L. M. 2003 Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst* **95**, 14-8.

Simon, R. M. & Dobbin, K. 2003 Experimental design of DNA microarray experiments. *Biotechniques* **Suppl**, 16-21.

Slonim, D. K. 2002 From patterns to pathways: gene expression data analysis comes of age. *Nat Genet* **32 Suppl**, 502-8.

Smith, L. & Greenfield, A. 2003 DNA microarrays and development. *Hum Mol Genet* **12 Spec No 1**, R1-8.

Smyth, G. K. & Speed, T. 2003 Normalization of cDNA microarray data. *Methods* **31**, 265-73.

Southern, E. M. 2001 DNA microarrays. History and overview. *Methods Mol Biol* **170**, 1-15.

Taroncher-Oldenburg, G., Griner, E. M., Francis, C. A. & Ward, B. B. 2003 Oligonucleotide microarray for the study of functional gene diversity in the nitrogen cycle in the environment. *Appl Environ Microbiol* **69**, 1159-71.

The Tumor Analysis Best Practices. 2004 Expression profiling--best practices for data generation and interpretation in clinical trials. *Nat Rev Genet* **5**, 229-37.

Tusher, V. G., Tibshirani, R. & Chu, G. 2001 Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98**, 5116-21.

Ulrich, R. & Friend, S. H. 2002 Toxicogenomics and drug discovery: will new technologies help us produce better drugs? *Nat Rev Drug Discov* **1**, 84-8.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R. & Friend, S. H. 2002 Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530-6.

Walsh, B. & Henderson, D. 2004 Microarrays and beyond: what potential do current and future genomics tools have for breeders? *J Anim Sci* **82 E-Suppl**, E292-299.

Watson, J. E., Doggett, N. A., Albertson, D. G., Andaya, A., Chinnaiyan, A., van Dekken, H., Ginzinger, D., Haqq, C., James, K., Kamkar, S., Kowbel, D., Pinkel, D., Schmitt, L., Simko, J. P., Volik, S., Weinberg, V. K., Paris, P. L. & Collins, C. 2004 Integration of high-resolution array comparative genomic hybridization analysis of chromosome 16q with expression array data refines common regions of loss at 16q23-qter and identifies underlying candidate tumor suppressor genes in prostate cancer. *Oncogene* **23**, 3487-94.

Weeraratna, A. T., Nagel, J. E., de Mello-Coelho, V. & Taub, D. D. 2004 Gene expression profiling: from microarrays to medicine. *J Clin Immunol* **24**, 213-24.

Weldon, C. B., Scandurro, A. B., Rolfe, K. W., Clayton, J. L., Elliott, S., Butler, N. N., Melnik, L. I., Alam, J., McLachlan, J. A., Jaffe, B. M., Beckman, B. S. & Burow, M. E. 2002 Identification of mitogen-activated protein kinase kinase as a chemoresistant pathway in MCF-7 cells by using gene expression microarray. *Surgery* **132**, 293-301.

Wodicka, L., Dong, H., Mittmann, M., Ho, M. H. & Lockhart, D. J. 1997 Genome-wide expression monitoring in Saccharomyces cerevisiae. *Nat Biotechnol* **15**, 1359-67.

Wu, C. C., Huang, H. C., Juan, H. F. & Chen, S. T. 2004 GeneNetwork: an interactive tool for reconstruction of genetic networks using microarray data. *Bioinformatics* **20**, 3691-3.

Yershov, G., Barsky, V., Belgovskiy, A., Kirillov, E., Kreindlin, E., Ivanov, I., Parinov, S., Guschin, D., Drobishev, A., Dubiley, S. & Mirzabekov, A. 1996 DNA analysis and diagnostics on oligonucleotide microchips. *Proc Natl Acad Sci U S A* **93**, 4913-8.

Zareparsi, S., Hero, A., Zack, D. J., Williams, R. W. & Swaroop, A. 2004 Seeing the unseen: Microarray-based gene expression profiling in vision. *Invest Ophthalmol Vis Sci* **45**, 2457-62.

Zhong, S., Li, C. & Wong, W. H. 2003 ChipInfo: Software for extracting gene annotation and gene ontology information for microarray analysis. *Nucleic Acids Res* **31**, 3483-6.

Zhong, S., Tian, L., Li, C., Storch, F. & Wong, W. 2004 Comparative Analysis of Gene Sets in the Gene Ontology Space under the Multiple Hypothesis Testing Framework. *Proc. IEEE Comp Systems Bioinformatics* **2004**, 425-435.

Appendix A Arabidopsis up- and down-regulated genes in response to PI5P treatment

Probe Set ID	Fold Chang	Gene Title	AGI	Cellular Component (Gene Ontology ID)
Down-regulated genes				
267645_at	-1.79	glycosyl hydrolase family 1 protein	AT2G32860	endomembrane system (12505)
267518_at	-1.76	kinase interacting family protein	AT2G30500	mitochondrion (5739)
267063_at	-1.55	expressed protein	AT2G41120	chloroplast (9507)
266989_at	-1.63	jacalin lectin family protein	AT2G39330	
266922_s_at	-1.84	SKP1 family protein	AT2G45950	
266727_at	-2.1	ATP/GTP-binding protein family	AT2G03150	mitochondrion (5739)
266385_at	-2.03	pathogenesis-related protein 1 (PR-1)	AT2G14610	extracellular (5576) endomembrane system (12505)
266376_at	-3.25	xyloglucan:xyloglucosyl transferase, putative / xyloglucan endotransglycosylase, putative / endo-xyloglucan transferase, putative	AT2G14620	endomembrane system (12505)
266327_at	-1.76	homeobox-leucine zipper protein 7 (HB-7) / HD-ZIP transcription factor 7	AT2G46680	nucleus (5634)
266070_at	-2.21	expansin family protein (EXPR3)	AT2G18660	extracellular (5576) endomembrane system (12505)
265837_at	-1.59	expressed protein	AT2G14560	
265698_at	-1.62	expressed protein	AT2G32160	
265611_at	-1.69	expressed protein	AT2G25510	mitochondrion (5739)
265464_at	-1.6	myosin heavy chain-related	AT2G37080	chloroplast (9507)

265452_at	-2.16	basic helix-loop-helix (bHLH) family protein	AT2G46510	
265359_at	-1.73	myb family transcription factor	AT2G16720	nucleus (5634)
265339_at	-1.68	inorganic pyrophosphatase (soluble) (PPA) / pyrophosphate phospho-hydrolase / PPase	AT2G18230	membrane (16020)
265122_at	-1.68	flavin-containing monooxygenase family protein / FMO family protein	AT1G62540	endomembrane system (12505)
264968_at	-1.71	rubber elongation factor (REF) family protein	AT1G67360	endomembrane system (12505)
264931_at	-2.63	polygalacturonase, putative / pectinase, putative	AT1G60590	chloroplast (9507)
264790_at	-1.5	histidine kinase 1	AT2G17820	membrane (16020)
264400_at	-2.07	glucose-6-phosphate/phosphate translocator, putative	AT1G61800	chloroplast (9507) membrane (16020) integral to membrane (16021)
264318_at	-2.74	beta-ketoacyl-CoA synthase, putative	AT1G04220	
264217_at	-3.45	armadillo/beta-catenin repeat family protein / U-box domain-containing protein	AT1G60190	chloroplast (9507)
264102_at	-1.53	expressed protein	AT1G79270	
263852_at	-2.31	MutT/nudix family protein	AT2G04450	
263811_at	-1.55	long-chain-fatty-acid--CoA ligase family protein / long-chain acyl-CoA synthetase family protein (LACS8)	AT2G04350	
263802_at	-1.7	expressed protein	AT2G40430	
263739_at	-1.54	zinc finger (B-box type) family protein	AT2G21320	intracellular (5622) endomembrane system (12505)
263539_at	-1.89	aminotransferase, putative	AT2G24850	
263433_at	-1.63	inositol-3-phosphate synthase isozyme 2 / myo-inositol-1-phosphate synthase 2 / MI-1-P synthase 2 / IPS 2	AT2G22240	
263296_at	-1.81	calmodulin-binding protein-related	AT2G38800	
263252_at	-1.66	zinc finger (B-box type) family	AT2G31380	intracellular (5622)

		protein / salt tolerance-like protein (STH)		endomembrane system (12505)
263122_at	-1.69	solanesyl diphosphate synthase (SPS)	AT1G78510	
263112_at	-2.81	kinase interacting family protein	AT1G03080	
262958_at	-1.59	dehydrin family protein	AT1G54410	
262926_s_at	-2.55	S-receptor protein kinase, putative	AT1G65790	membrane (16020)
262871_at	-1.7	expressed protein	AT1G65010	chloroplast (9507)
262619_at	-1.65	enoyl-CoA hydratase/isomerase family protein	AT1G06550	
262526_at	-1.53	geranyl diphosphate synthase, putative / GPPS, putative / dimethylallyltransferase, putative / prenyl transferase, putative	AT1G17050	
262281_at	-1.62	proton-dependent oligopeptide transport (POT) family protein	AT1G68570	
262237_at	-1.62	thioesterase family protein	AT1G48320	
262128_at	-1.93	late embryogenesis abundant protein, putative / LEA protein, putative	AT1G52690	
261958_at	-2.09	glutaredoxin family protein	AT1G64500	
261892_at	-1.55	WRKY family transcription factor	AT1G80840	
261844_at	-2.12	expressed protein	AT1G15940	
261819_at	-1.6	S-locus protein kinase, putative	AT1G11410	endomembrane system (12505)
261804_at	-1.81	UDP-glucoronosyl/UDP-glucosyl transferase family protein	AT1G30530	endomembrane system (12505)
261333_at	-1.59	FF domain-containing protein / WW domain-containing protein	AT1G44910	nucleus (5634) chloroplast 9507
261240_at	-1.73	subtilase family protein	AT1G32940	extrachromosomal circular DNA (5727); endomembrane system (12505)
261177_at	-2.3	male sterility MS5 family protein	AT1G04770	mitochondrion (5739)
261150_at	-1.55	S-adenosyl-L-methionine:jasmonic acid carboxyl methyltransferase	AT1G19640	cytoplasm (5737)

		(JMT)		
261077_at	-3.2	protein phosphatase 2C, putative / PP2C, putative	AT1G07430	protein serine/threonine phosphatase complex (8287)
261031_at	-1.57	COP1-interacting protein-related	AT1G17360	
261016_at	-1.6	glycosyl hydrolase family 1 protein	AT1G26560	endomembrane system (12505)
260924_at	-1.91	protein kinase family protein	AT1G21590	
260916_at	-1.5	expressed protein	AT1G02475	chloroplast (9507)
260904_at	-2.1	NPR1/NIM1-interacting protein 1 (NIMIN-1)	AT1G02450	
260832_at	-1.55	glycosyl transferase family 8 protein	AT1G06780	endomembrane system (12505)
260769_at	-1.51	myb family transcription factor	AT1G49010	nucleus (5634) cytoplasm 5737
260727_at	-2.49	glycoside hydrolase family 28 protein / polygalacturonase (pectinase) family protein	AT1G48100	mitochondrion (5739)
260466_at	-1.73	phosphatidylinositol-4-phosphate 5-kinase family protein	AT1G10900	
260425_at	-2.13	CCAAT-box-binding transcription factor-related	AT1G72440	endomembrane system (12505)
260380_at	-2.22	zinc finger (B-box type) family protein	AT1G73870	intracellular (5622)
260203_at	-3.88	no apical meristem (NAM) family protein	AT1G52890	
260140_at	-2.41	myb family transcription factor, putative / production of anthocyanin pigment 2 protein (PAP2)	AT1G66390	nucleus (5634)
259794_at	-1.87	myosin heavy chain-related	AT1G64330	
259705_at	-1.75	no apical meristem (NAM) family protein	AT1G77450	
259561_at	-1.58	wall-associated kinase 1 (WAK1)	AT1G21250	extracellular matrix (5578); plasma membrane (5886)
259516_at	-1.5	dehydrin (ERD10)	AT1G20450	

259432_at	-2.06	myb family transcription factor	AT1G01520	chloroplast (9507)
259367_at	-1.5	expressed protein	AT1G69070	
259173_at	-2.14	glycosyl hydrolase family 1 protein	AT3G03640	endomembrane system (12505)
259058_at	-1.63	cytochrome P450, putative	AT3G03470	endomembrane system (12505)
259015_at	-2.11	expressed protein	AT3G07350	
258497_at	-1.57	zinc finger protein CONSTANS-LIKE 2 (COL2)	AT3G02380	
258362_at	-2.27	expressed protein	AT3G14280	mitochondrion (5739)
258333_at	-2.01	matrix-localized MAR DNA-binding protein-related	AT3G16000	thylakoid membrane (sensu Viridiplantae) (9535); plastid nucleoid (42646)
258321_at	-1.78	chlorophyll A-B binding family protein / early light-induced protein (ELIP)	AT3G22840	chloroplast (9507)
258158_at	-1.63	acid phosphatase type 5 (ACP5)	AT3G17790	cell surface (9986)
258119_at	-1.99	mitogen-activated protein kinase, putative / MAPK, putative (MPK19)	AT3G14720	
258017_at	-1.84	expressed protein	AT3G19370	chloroplast (9507)
257919_at	-1.84	myb family transcription factor (MYB15)	AT3G23250	nucleus (5634)
257855_at	-1.54	myb family transcription factor	AT3G13040	nucleus (5634)
257771_at	-1.51	CBL-interacting protein kinase 7 (CIPK7)	AT3G23000	chloroplast (9507)
257615_at	-2.02	octicosapeptide/Phox/Bem1p (PB1) domain-containing protein	AT3G26510	chloroplast (9507)
257262_at	-2.25	zinc finger (B-box type) family protein	AT3G21890	intracellular (5622)
257253_at	-1.66	ABC1 family protein	AT3G24190	chloroplast (9507)
256861_at	-1.66	beta-amylase, putative / 1,4-alpha-D-glucan maltohydrolase, putative	AT3G23920	chloroplast (9507)
256766_at	-2.04	expressed protein	AT3G22231	chloroplast (9507)
256596_at	-9.63	AAA-type ATPase family protein	AT3G28540	nucleus (5634) cytoplasm 5737

256497_at	-1.61	expressed protein	AT1G31580	cell wall (5618)
256431_s_at	-1.74	disease resistance family protein / LRR family protein	AT3G11010	
256300_at	-6.11	no apical meristem (NAM) family protein	AT1G69490	
256296_at	-2.57	EXS family protein / ERD1/XPR1/SYG1 family protein	AT1G69480	
256245_at	-1.68	heat shock protein 70, putative / HSP70, putative	AT3G12580	
255795_at	-2.13	calcium-binding RD20 protein (RD20)	AT2G33380	
255723_at	-1.73	expressed protein	AT3G29575	
255645_at	-1.76	auxin-responsive family protein	AT4G00880	
255588_at	-2.49	pentatricopeptide (PPR) repeat-containing protein	AT4G01570	
255566_s_at	-1.61	XH/XS domain-containing protein	AT4G01780	
255284_at	-1.56	5'-adenylylsulfate reductase (APR1) / PAPS reductase homolog (PRH19)	AT4G04610	chloroplast (9507) plastid 9536
255128_at	-2.27	expressed protein	AT4G08310	
254926_at	-1.78	1-aminocyclopropane-1-carboxylate synthase 6 / ACC synthase 6 (ACS6)	AT4G11280	
254869_at	-2.77	protein kinase family protein	AT4G11890	
254767_s_at	-1.68	cytochrome P450 71A19, putative (CYP71A19)	AT4G13290	endomembrane system (12505)
254764_at	-1.7	short-chain dehydrogenase/reductase (SDR) family protein	AT4G13250	
254680_at	-1.84	phytochrome E (PHYE)	AT4G18130	membrane (16020)
254390_at	-2.04	calcium-dependent protein kinase, putative / CDPK, putative	AT4G21940	chloroplast (9507)
254327_at	-1.65	protease inhibitor/seed storage/lipid transfer protein (LTP) family protein	AT4G22490	endomembrane system (12505)
254305_at	-2.08	potassium channel protein 2 (AKT2) (AKT3)	AT4G22200	membrane (16020)
254231_at	-1.95	WRKY family transcription factor	AT4G23810	
253994_at	-2.28	protein phosphatase 2C ABI1 / PP2C ABI1 / abscisic acid-insensitive 1	AT4G26080	

		(ABI1)		
253922_at	-1.8	expressed protein	AT4G26850	
253915_at	-1.72	calcium-binding EF hand family protein	AT4G27280	chloroplast (9507)
253872_at	-2.39	no apical meristem (NAM) family protein (RD26)	AT4G27410	
253834_at	-1.51	protein phosphatase 2C PPH1 / PP2C PPH1 (PPH1)	AT4G27800	mitochondrion (5739); protein serine/threonine phosphatase complex (8287)
253814_at	-1.75	expressed protein	AT4G28290	mitochondrial outer membrane translocase complex (5742); chloroplast (9507)
253305_at	-1.76			
253237_at	-1.5	aldehyde dehydrogenase (ALDH3)	AT4G34240	plastid (9536)
253228_at	-1.71	expressed protein	AT4G34630	
253061_at	-1.86	TAZ zinc finger family protein / BTB/POZ domain-containing protein	AT4G37610	nucleus (5634)
252958_at	-1.68	myb family transcription factor (MYB4)	AT4G38620	nucleus (5634)
252888_at	-1.51	glucose-1-phosphate adenylyltransferase large subunit 3 (APL3) / ADP-glucose pyrophosphorylase	AT4G39210	
252429_at	-1.54	Dof-type zinc finger domain-containing protein	AT3G47500	
252319_at	-1.87	expressed protein	AT3G48710	
252269_at	-2.95	expressed protein	AT3G49580	
251826_at	-1.6	ABC transporter family protein	AT3G55110	membrane (16020); inner membrane (19866)
251725_at	-7.12	expressed protein	AT3G56260	
251705_at	-3.08	WRKY family transcription factor	AT3G56400	
251400_at	-3	expressed protein	AT3G60420	

251309_at	-1.69	short-chain dehydrogenase/reductase (SDR) family protein	AT3G61220	
251272_at	-4.1	homeobox-leucine zipper protein 12 (HB-12) / HD-ZIP transcription factor 12	AT3G61890	nucleus (5634)
251247_at	-1.52	expressed protein	AT3G62140	
251060_at	-2.04	CBL-interacting protein kinase 14 (CIPK14)	AT5G01820	
250942_at	-2.94	legume lectin family protein	AT5G03350	endomembrane system (12505)
250735_at	-1.61	expressed protein	AT5G06280	chloroplast (9507)
250648_at	-1.98	late embryogenesis abundant group 1 domain-containing protein / LEA group 1 domain-containing protein	AT5G06760	
250598_at	-1.86	myb family transcription factor (MYB29)	AT5G07690	nucleus (5634)
250408_at	-1.78	CBL-interacting protein kinase 5 (CIPK5)	AT5G10930	
250296_at	-1.5	17.6 kDa class II heat shock protein (HSP17.6-CII)	AT5G12020	
250257_at	-1.6	pentatricopeptide (PPR) repeat-containing protein	AT5G13770	
249919_at	-1.68	expressed protein	AT5G19250	endomembrane system (12505)
249918_at	-1.69	expressed protein	AT5G19240	endomembrane system (12505)
249774_at	-1.56	squalene monooxygenase 1,1 / squalene epoxidase 1,1 (SQP1,1)	AT5G24150	endomembrane system (12505)
249769_at	-1.91	RNA polymerase sigma subunit SigE (sigE) / sigma-like factor (SIG5)	AT5G24120	chloroplast (9507)
249754_at	-1.92	oxidoreductase, 2OG-Fe(II) oxygenase family protein	AT5G24530	
249752_at	-1.89	expressed protein	AT5G24660	
249688_at	-1.55	aminotransferase-related	AT5G36160	endomembrane system (12505)
249614_at	-1.8	expressed protein	AT5G37300	
249271_at	-2.45	COP1-interactive protein 1 / CIP1	AT5G41790	cytoskeleton (5856)

249231_at	-1.58	expressed protein	AT5G42030	
249215_at	-1.61	dihydroflavonol 4-reductase (dihydrokaempferol 4-reductase) (DFR)	AT5G42800	
249200_at	-2.04	5'-3' exoribonuclease (XRN2)	AT5G42540	nucleus (5634)
249191_at	-1.92	O-methyltransferase N-terminus domain-containing protein	AT5G42760	
248764_at	-3.22			
248448_at	-1.71	AP2 domain-containing transcription factor, putative	AT5G51190	
248393_at	-1.62	BAG domain-containing protein	AT5G52060	
248344_at	-2.37	protein transport protein-related	AT5G52280	mitochondrion (5739)
248337_at	-2.16	low-temperature-responsive protein 78 (LTI78) / desiccation-responsive protein 29A (RD29A)	AT5G52310	
248311_at	-1.55	beta-carotene hydroxylase, putative	AT5G52570	chloroplast (9507)
248218_at	-2.15	expressed protein	AT5G53710	endomembrane system (12505)
248169_at	-2.22	ankyrin repeat family protein	AT5G54610	
248109_at	-1.73	DNA topoisomerase I, putative	AT5G55310	
248082_at	-2.2	fimbrin-like protein, putative	AT5G55400	chloroplast (9507)
248028_at	-1.94	expressed protein	AT5G55620	
247977_at	-1.85	expressed protein	AT5G56850	microtubule (5874); chloroplast 9507
247780_at	-2.17	dehydrodolichyl diphosphate synthase, putative / DEDOL-PP synthase, putative	AT5G58770	
247738_at	-1.67	myosin heavy chain-related	AT5G59210	mitochondrion (5739)
247723_at	-2.3	protein phosphatase 2C, putative / PP2C, putative	AT5G59220	chloroplast (9507)
247549_at	-1.58	myb family transcription factor (MYB28)	AT5G61420	nucleus (5634); mitochondrion 5739
247293_at	-2.68	expressed protein	AT5G64510	
247222_at	-1.77	ABC transporter family protein	AT5G64840	chloroplast (9507); membrane 16020;

				integral to membrane (16021); inner membrane (19866)
246968_at	-1.81	zinc finger (C3HC4-type RING finger) family protein	AT5G24870	
246901_at	-1.7	pentatricopeptide (PPR) repeat-containing protein	AT5G25630	
246490_at	-2.57	adenosylmethionine decarboxylase family protein	AT5G15950	
246476_at	-2.52	expressed protein	AT5G16730	extracellular (5576); chloroplast 9507
246468_at	-1.94	UDP-glucoronosyl/UDP-glucosyl transferase family protein	AT5G17050	chloroplast (9507)
246137_at	-2.31	expressed protein	AT5G28490	chloroplast (9507)
246069_at	-1.74	zinc knuckle (CCHC-type) family protein	AT5G20220	mitochondrion (5739)
245991_at	-1.73	24 kDa vacuolar protein, putative	AT5G20660	vacuole (5773)
245734_at	-1.69	hydrolase, alpha/beta fold family protein	AT1G73480	cytoplasm (5737); chloroplast (9507)
245628_at	-2.12	myb family transcription factor (MYB75)	AT1G56650	nucleus (5634)
245346_at	-1.78	beta-amylase (CT-BMY) / 1,4-alpha-D-glucan maltohydrolase	AT4G17090	chloroplast stroma (9570)
245319_at	-1.69	expressed protein	AT4G16146	
245302_at	-2.16	myb family transcription factor (KAN3)	AT4G17695	nucleus (5634)
245265_at	-1.54	ankyrin repeat family protein	AT4G14400	membrane (16020)
245092_at	-1.53	bZIP transcription factor family protein	AT2G40950	nucleus (5634)
244998_at	-1.54			

Up-regulated

263836_at	4.9	Bet v I allergen family protein	AT2G40330	mitochondrion (5739)
267624_at	1.59	protein kinase, putative	AT2G39660	chloroplast (9507)

267523_at	1.61	BTB/POZ domain-containing protein	AT2G30600	
267318_at	2.07	fatty acid hydroxylase (FAH1)	AT2G34770	
267260_at	2.29	arabinogalactan-protein (AGP17)	AT2G23130	endomembrane system (12505)
267209_at	2.71	expressed protein	AT2G30930	chloroplast (9507)
267169_at	1.85	short-chain dehydrogenase/reductase (SDR) family protein	AT2G37540	chloroplast (9507)
267034_at	1.69	expressed protein	AT2G38310	chloroplast (9507)
266956_at	2.95	expressed protein	AT2G34510	endomembrane system (12505)
266693_at	1.82	expressed protein	AT2G19800	
266658_at	2.31	expressed protein	AT2G25735	chloroplast (9507)
266545_at	3.11	expressed protein	AT2G35290	mitochondrion (5739)
266481_at	1.6	TCP family transcription factor, putative	AT2G31070	
266316_at	1.92			
266140_at	2.3	nodulin family protein	AT2G28120	endomembrane system (12505); membrane (16020)
266123_at	2.25	protease inhibitor/seed storage/lipid transfer protein (LTP) family protein	AT2G45180	endomembrane system (12505)
265648_at	2.06	glycosyl hydrolase family 17 protein	AT2G27500	endomembrane system (12505)
265561_s_at	1.7	glycine-rich protein	AT2G05510	endomembrane system (12505)
265481_at	1.81	expressed protein	AT2G15960	
265478_at	2.33	expressed protein	AT2G15890	chloroplast (9507)
265414_at	1.73	nodulin family protein	AT2G16660	endomembrane system (12505); membrane (16020)
265066_at	2.56	fasciclin-like arabinogalactan-protein (FLA9)	AT1G03870	endomembrane system (12505)
265005_at	2.08	expressed protein	AT1G61667	endomembrane system (12505)
264857_at	1.6	glycosyl transferase family 8 protein	AT1G24170	endomembrane system (12505)

264770_at	2.13	armadillo/beta-catenin repeat family protein / U-box domain-containing protein	AT1G23030	
264704_at	2.97	glycosyl transferase family 8 protein	AT1G70090	endomembrane system (12505)
264624_at	1.96	early-responsive to dehydration stress protein (ERD6) / sugar transporter family protein	AT1G08930	intracellular (5622); membrane (16020); integral to membrane (16021)
264433_at	2.35	glycosyl hydrolase family 1 protein	AT1G61810	endomembrane system (12505)
263598_at	2.07	xyloglucan:xyloglucosyl transferase / xyloglucan endotransglycosylase / endo-xyloglucan transferase (EXGT-A3)	AT2G01850	extracellular (5576); endomembrane system (12505)
263499_at	1.82	tetratricopeptide repeat (TPR)-containing protein	AT2G42580	mitochondrial outer membrane (5741)
263421_at	2.43	phosphate-responsive 1 family protein	AT2G17230	endomembrane system (12505)
263249_at	1.75	delta 9 desaturase (ADS2)	AT2G31360	endoplasmic reticulum (5783); membrane (16020)
263236_at	1.53	two-component responsive regulator / response regulator 4 (ARR4)	AT1G10470	nucleus (5634); cytoplasm (5737)
263207_at	4.59	xyloglucan:xyloglucosyl transferase, putative / xyloglucan endotransglycosylase, putative / endo-xyloglucan transferase, putative	AT1G10550	endomembrane system (12505)
263019_at	1.55	glycosyl transferase family 20 protein / trehalose-phosphatase family protein	AT1G23870	endomembrane system (12505)
262947_at	1.95	gibberellin-regulated protein 1 (GASA1) / gibberellin-responsive protein 1	AT1G75750	endomembrane system (12505)
262456_at	1.84	glucose transporter (STP1)	AT1G11260	membrane (16020); integral to

				membrane (16021)
261928_at	1.88	plastocyanin-like domain-containing protein	AT1G22480	endomembrane system (12505)
261926_at	1.74	SEC14 cytosolic factor family protein / phosphoglyceride transfer family protein	AT1G22530	intracellular (5622)
261727_at	2.2	S-adenosyl-methionine-sterol-C-methyltransferase	AT1G76090	endomembrane system (12505)
261453_at	1.88	O-methyltransferase, putative	AT1G21130	
261229_at	1.59	Rac-like GTP-binding protein (ARAC4) / Rho-like GTP-binding protein (ROP2)	AT1G20090	
260914_at	3.12	glycosyl hydrolase family 3 protein	AT1G02640	endomembrane system (12505)
260856_at	2.2	AP2 domain-containing transcription factor family protein	AT1G21910	
260668_at	2.61	expressed protein	AT1G19530	
260522_x_at	1.6	expressed protein	AT2G41730	mitochondrion (5739)
260451_at	1.52	ethylene-responsive element-binding protein, putative	AT1G72360	
260427_at	1.93	auxin-responsive protein-related	AT1G72430	chloroplast (9507)
260423_at	1.67	exocyst subunit EXO70 family protein	AT1G72470	exocyst (145)
260221_at	2.4	gibberellin-responsive protein, putative	AT1G74670	endomembrane system (12505)
259909_at	1.75	expressed protein	AT1G60870	
259879_at	2.11	calcium-binding EF hand family protein	AT1G76650	chloroplast (9507)
259875_s_at	1.6	12-oxophytodienoate reductase (OPR2)	AT1G76690	
259803_at	2	SEC14 cytosolic factor family protein / phosphoglyceride transfer family protein	AT1G72150	intracellular (5622)
259685_at	1.6	F-box family protein / SKP1 interacting partner 3-related	AT1G63090	
259681_at	1.67	nitrate reductase 1 (NR1)	AT1G77760	

259546_at	1.78	EXS family protein / ERD1/XPR1/SYG1 family protein	AT1G35350	
259466_at	1.97	two-component responsive regulator / response regulator 7 (ARR7)	AT1G19050	
259365_at	1.85	myb family transcription factor	AT1G13300	
259310_s_at	1.55	sugar transporter, putative	AT3G05165	membrane (16020); integral to membrane (16021)
259106_at	2.48	rapid alkalinization factor (RALF) family protein	AT3G05490	extracellular matrix (5578)
259072_at	1.63	beta-Ig-H3 domain-containing protein / fasciclin domain-containing protein	AT3G11700	endomembrane system (12505)
259020_at	2.44	expressed protein	AT3G07470	endomembrane system (12505)
258920_at	2.07	non-symbiotic hemoglobin 2 (HB2) (GLB2)	AT3G10520	collagen type I (5584)
258537_at	2.28	disease resistance protein (TIR-NBS class), putative	AT3G04210	endomembrane system (12505); membrane (16020)
258468_at	2.12	expressed protein	AT3G06070	
258432_at	1.79	rapid alkalinization factor (RALF) family protein	AT3G16570	extracellular matrix (5578)
258402_at	1.65	expressed protein	AT3G15450	
258156_at	1.58	expressed protein	AT3G18050	endomembrane system (12505)
258132_at	1.51	protein kinase family protein	AT3G24550	chloroplast (9507)
258100_at	1.62	MATE efflux family protein	AT3G23550	membrane (16020)
258060_at	1.65	serine/threonine protein phosphatase 2A (PP2A) regulatory subunit B', putative	AT3G26030	protein phosphatase type 2A complex (159)
257785_at	1.56	ubiquitin family protein	AT3G26980	
257315_at	2.31	proline oxidase, mitochondrial / osmotic stress-responsive proline dehydrogenase (POX) (PRO1) (ERD5)	AT3G30775	mitochondrion (5739)
257204_at	2.05	rapid alkalinization factor (RALF)	AT3G23805	extracellular matrix

		family protein		(5578)
257203_at	2.59	xyloglucan:xyloglucosyl transferase, putative / xyloglucan endotransglycosylase, putative / endo-xyloglucan transferase, putative	AT3G23730	endomembrane system (12505)
257076_at	1.65	expressed protein	AT3G19680	chloroplast (9507)
256940_at	2.29	expressed protein	AT3G30720	
256848_at	1.93	kinesin light chain-related	AT3G27960	
256578_at	1.66	peroxidase, putative	AT3G28200	endomembrane system (12505)
256525_at	1.55	aspartyl protease family protein	AT1G66180	endomembrane system (12505)
256522_at	1.97	U-box domain-containing protein	AT1G66160	mitochondrion (5739)
256516_at	1.6	leucine-rich repeat protein kinase, putative (TMK1)	AT1G66150	extracellular (5576)
256433_at	1.62	expressed protein	AT3G10980	
256396_at	1.55	expressed protein	AT3G06150	
256275_at	1.58	actin 11 (ACT11)	AT3G12110	cytoskeleton (5856)
256231_at	1.91	zinc finger (AN1-like) family protein	AT3G12630	
255818_at	1.57	expressed protein	AT2G33570	
255807_at	1.78			
255617_at	1.83	protein kinase family protein	AT4G01330	membrane (16020)
255506_at	1.82	glycosyl transferase family 8 protein	AT4G02130	endomembrane system (12505)
255479_at	1.99	late embryogenesis abundant 3 family protein / LEA3 family protein	AT4G02380	chloroplast (9507)
255433_at	1.7	xyloglucan:xyloglucosyl transferase, putative / xyloglucan endotransglycosylase, putative / endo-xyloglucan transferase, putative	AT4G03210	endomembrane system (12505)
255403_at	1.69	auxin-responsive GH3 family protein	AT4G03400	
255064_at	2.34	phosphate-responsive protein, putative (EXO)	AT4G08950	endomembrane system (12505)
254707_at	2.81	inositol polyphosphate 5-	AT4G18010	mitochondrion

		phosphatase II (IP5PII)		(5739)
254705_at	2.17	expressed protein	AT4G17870	
254573_at	2.31	pectinacetylesterase family protein	AT4G19420	endomembrane system (12505)
254553_at	2.14	disease resistance protein (TIR-NBS-LRR class), putative	AT4G19530	membrane (16020)
254492_at	2.25	DREPP plasma membrane polypeptide family protein	AT4G20260	
254396_at	3.89	proton-dependent oligopeptide transport (POT) family protein	AT4G21680	membrane (16020); integral to membrane (16021)
254384_at	4.81	26.5 kDa class P-related heat shock protein (HSP26.5-P)	AT4G21870	
254331_s_at	1.74	cytochrome P450 family protein	AT4G22710	endomembrane system (12505)
254098_at	1.66	superoxide dismutase (Fe), chloroplast (SODB) / iron superoxide dismutase (FSD1)	AT4G25100	
254042_at	2.5	xyloglucan:xyloglucosyl transferase, putative / xyloglucan endotransglycosylase, putative / endo-xyloglucan transferase, putative (XTR6)	AT4G25810	mitochondrial electron transport chain (5746); endomembrane system (12505); integral to membrane (16021)
253811_at	1.67	expressed protein	AT4G28190	
253722_at	1.9	zinc finger (CCCH-type) family protein	AT4G29190	
253667_at	1.59	peroxidase, putative	AT4G30170	endomembrane system (12505)
253666_at	2.26	MERI-5 protein (MERI-5) (MERI5B) / endo-xyloglucan transferase / xyloglucan endo-1,4-beta-D-glucanase (SEN4)	AT4G30270	endomembrane system (12505)
253631_at	1.79	NAD-dependent epimerase/dehydratase family protein	AT4G30440	

253485_at	1.57	WRKY family transcription factor	AT4G31800	
253104_at	1.6	pathogenesis-related thaumatin family protein	AT4G36010	endomembrane system (12505)
252997_at	2.47	expansin family protein (EXPL2)	AT4G38400	extracellular (5576); endomembrane system (12505)
252965_at	2.02	auxin-responsive protein, putative	AT4G38860	mitochondrion (5739)
252950_at	1.57	1-phosphatidylinositol phosphodiesterase-related	AT4G38690	
252563_at	3.66	expansin family protein (EXPL1)	AT3G45970	extracellular (5576); endomembrane system (12505)
252425_at	1.6	TCP family transcription factor, putative	AT3G47620	
252374_at	2.32	two-component responsive regulator / response regulator 5 (ARR5) / response reactor 2 (RR2)	AT3G48100	
251861_at	1.5	zinc finger (GATA type) family protein	AT3G54810	nucleus (5634)
251584_at	1.87	tetratricopeptide repeat (TPR)-containing protein	AT3G58620	mitochondrial outer membrane (5741)
251507_at	1.71	aspartyl protease family protein	AT3G59080	endomembrane system (12505)
251494_at	1.62	serine/threonine protein kinase, putative	AT3G59350	
251221_at	3.15	universal stress protein (USP) family protein	AT3G62550	
251192_at	1.85	galactosyl transferase GMA12/MNN10 family protein	AT3G62720	mitochondrion (5739)
251072_at	1.66	expressed protein	AT5G01740	
251059_at	1.76	CBL-interacting protein kinase 15 (CIPK15)	AT5G01810	mitochondrion (5739)
250936_at	2.96	expressed protein	AT5G03120	endomembrane system (12505)
250933_at	2.36	fasciclin-like arabinogalactan-protein (FLA11)	AT5G03170	endomembrane system (12505)

250777_at	3.09	expressed protein	AT5G05440	
250464_at	1.8	expressed protein	AT5G10040	
250398_at	1.71	expressed protein	AT5G11000	chloroplast (9507)
250217_at	2.43	nodulin family protein	AT5G14120	endomembrane system (12505)
250110_at	1.78	plastocyanin-like domain-containing protein	AT5G15350	endomembrane system (12505)
249996_at	2.33	glutaredoxin family protein	AT5G18600	
249955_at	1.53	sugar transporter, putative	AT5G18840	membrane (16020); integral to membrane (16021)
249922_at	1.67	auxin/aluminum-responsive protein, putative	AT5G19140	chloroplast (9507)
249862_at	2.39	zinc finger (C3HC4-type RING finger) family protein	AT5G22920	
249765_at	3.78	C4-dicarboxylate transporter/malic acid transport family protein	AT5G24030	integral to membrane (16021)
249234_at	1.58	zinc finger (C3HC4-type RING finger) family protein	AT5G42200	
249073_at	3.22	acid phosphatase class B family protein	AT5G44020	endomembrane system (12505)
249037_at	2.44	fasciclin-like arabinogalactan-protein, putative	AT5G44130	endomembrane system (12505)
249008_at	1.66	methyladenine glycosylase family protein	AT5G44680	chloroplast (9507)
248820_at	1.71	senescence-associated protein-related	AT5G47060	
248683_at	1.97	protease inhibitor/seed storage/lipid transfer protein (LTP) family protein	AT5G48490	endomembrane system (12505)
248622_at	3.16	glycosyl hydrolase family 3 protein	AT5G49360	endomembrane system (12505)
248606_at	1.6	bZIP family transcription factor	AT5G49450	nucleus (5634); chloroplast (9507)
248460_at	2.52	basic helix-loop-helix (bHLH) family protein	AT5G50915	
248419_at	2.19	phosphate-responsive 1 family protein	AT5G51550	endomembrane system (12505)

248252_at	3.33	arabinogalactan-protein, putative (AGP22)	AT5G53250	endomembrane system (12505)
248179_at	2.21	protein kinase family protein	AT5G54380	endomembrane system (12505)
247925_at	2.18	xyloglucan:xyloglucosyl transferase / xyloglucan endotransglycosylase / endo-xyloglucan transferase (TCH4)	AT5G57560	cell wall (5618)
247866_at	3.56	xyloglucan:xyloglucosyl transferase / xyloglucan endotransglycosylase / endo-xyloglucan transferase (XTR3)	AT5G57550	endomembrane system (12505)
247540_at	2.74	AP2 domain-containing transcription factor family protein	AT5G61590	
247533_at	2.71	protein kinase family protein	AT5G61570	endomembrane system (12505)
247462_at	1.79	protease inhibitor/seed storage/lipid transfer protein (LTP) family protein	AT5G62080	endomembrane system (12505)
247406_at	1.86	two-component responsive regulator / response regulator 6 (ARR6)	AT5G62920	
247297_at	1.6	peroxidase, putative	AT5G64100	endomembrane system (12505)
247280_at	1.73	phosphate-responsive protein, putative	AT5G64260	endomembrane system (12505)
247214_at	1.78	expressed protein	AT5G64850	
247188_at	1.7	14-3-3 protein GF14 kappa (GRF8)	AT5G65430	nucleus (5634); cytoplasm (5737); plasma membrane (5886); cell wall (sensu Magnoliophyta) 9505
247162_at	1.6	xyloglucan:xyloglucosyl transferase, putative / xyloglucan endotransglycosylase, putative / endo-xyloglucan transferase, putative	AT5G65730	endomembrane system (12505)
247132_at	1.56	armadillo/beta-catenin repeat family protein	AT5G66200	

247024_at	1.73	expressed protein	AT5G66985	
246781_at	2.04	sugar-porter family protein 1 (SFP1)	AT5G27350	membrane (16020); integral to membrane (16021)
246408_at	1.9	expressed protein	AT1G57680	endomembrane system (12505)
246114_at	2.6	raffinose synthase family protein / seed imbibition protein, putative (din10)	AT5G20250	chloroplast (9507)
246011_at	1.99	TCP family transcription factor, putative	AT5G08330	
245925_at	2.02	bZIP transcription factor family protein	AT5G28770	nucleus (5634)
245866_s_at	2.16	purine permease-related	AT1G57990	
245794_at	1.66	xyloglucan:xyloglucosyl transferase, putative / xyloglucan endotransglycosylase, putative / endo-xyloglucan transferase, putative (XTR4)	AT1G32170	endomembrane system (12505)
245757_at	4.5	phosphate-responsive protein, putative	AT1G35140	endomembrane system (12505)
245642_at	1.55	expressed protein	AT1G25275	endomembrane system (12505)
245334_at	2.01	rapid alkalinization factor (RALF) family protein	AT4G15800	extracellular matrix (5578)
245325_at	2.48	xyloglucan:xyloglucosyl transferase, putative / xyloglucan endotransglycosylase, putative / endo-xyloglucan transferase, putative (XTR7)	AT4G14130	endomembrane system (12505)
245262_at	3.57	aspartyl protease family protein	AT4G16563	endomembrane system (12505)
245176_at	1.72	DNAJ heat shock N-terminal domain-containing protein	AT2G47440	

Appendix B Glossary

Array: A collection of probes on glass encased in a plastic cartridge.

Baseline Array: An array used for normalization purposes during comparison analysis.

Bioinformatics: A multidisciplinary area, which applies computer science, statistics, and mathematics to solve biological and medical problems.

cDNA: Complementary DNA produced from an RNA template by the action of RNA-dependent DNA polymerase.

Change: A qualitative call indicating an Increase (I), Marginal Increase (MI), No

Change (NC), Marginal Decrease (MD), or Decrease (D) in transcript level between a baseline array and an experiment array.

Comparison Analysis: The analysis of an experimental array compared to a baseline array.

Detection: A qualitative measurement indicating if a given transcript is detected (Present), not detected (Absent), or marginally detected (Marginal).

Experimental Array: An array that is used in comparison analysis to be compared against a baseline array to detect changes in expression.

cDNA array: cDNA probes (500~5,000 bases long) are immobilized to a solid surface such as glass using robot spotting and exposed to a set of targets either separately or in a mixture.

Clustering analysis: a technique for grouping individuals or objects into unknown groups.

Data mining: extraction of useful information from data sets. Data mining serves to find information that is hidden within the available data.

Data preprocessing: any type of processing performed on raw data to prepare it for another processing procedure.

Exploratory data analysis: an approach for data analysis that employs a variety of techniques such as clustering to maximize insights into a data set and uncovers underlying structure.

GCOS: GeneChip® Operating Software. It manages GeneChip array data and automates the control of GeneChip Fluidics Stations and Scanners. GCOS provides workflow tracking of experiment, image and analysis data.

Gene annotation: A process that genes are annotated by cross-referencing to public databases like Gene Ontology and experimental data.

Gene expression: The process by which a gene's coded information is converted into the structures present and operating in the cell. Expressed genes include those that are transcribed into mRNA and then translated into protein and those that are transcribed into RNA but not translated into protein (e.g., transfer and ribosomal RNAs).

Gene Ontology: A set of controlled vocabularies used to describe biological features within a specified domain of biological knowledge.

Gene regulatory networks: The on-off switches and rheostats of a cell operating at the gene level.

Gene: The unit of heredity. A gene contains hereditary information encoded in the form of DNA and is located at a specific position on a chromosome in a cell's nucleus.

Hybridization: The formation of double-stranded DNA, RNA, or DNA/RNA hybrids by complementary base pairing.

Inferential statistics: Used to determine how likely it is that results that are obtained are not due to chance.

Interface: An interface among computer programs involves using agreed-upon commands and statements that let one computer program exchange information with the other in a way that the first program can integrate the second's.

K-means clustering: A clustering method. It initially takes the number of components of the population equal to the final required number of clusters.

Metabolic pathway: A series of chemical reactions occurring within a cell, catalyzed by enzymes , and resulting in either the formation of a metabolic product to be used or stored by the cell, or the initiation of another metabolic pathway.

Neural networks: Non-linear regression models that can be trained to learn with or without supervision.

Non -parametric statistics: A branch of statistics that are applied when data are not normally distribute.

Normalization: Adjusting an average value of an experimental array equal to that of the baseline array so that the arrays can be compared.

Oligo microarray: An array of oligonucleotide (20~80-mer oligos) probes is synthesized either in situ (on-chip) or by conventional synthesis followed by on-chip immobilization.

Parametric statistics: A group of statistical procedures that researchers use to test data that are normally distributed.

Pharmacogenomics: The study of the interaction of an individual's genetic makeup and response to a drug

Probe: A 25-mer oligonucleotide synthesized *in situ* on the surface of the array using photolithography and combinatorial chemistry.

Probe Set: A collection of probe pairs which interrogates the same sequence, or set of sequences. A probe set typically contains 11 probe pairs.

Self-organizing Map: A feed forward neural network that uses an unsupervised training algorithm, and through a process called self-organization, configures the output units into a topological representation of the original data.

Signal Log Ratio: The change in expression level for a transcript between a baseline and an experiment array.

SNPs: Single Nucleotide Polymorphism, differences (polymorphism) of individual bases within a genome from different individuals.

Visual Basic .NET (VB.NET or VB .NET): A version of Microsoft's Visual Basic that was designed, as part of the company's .NET product group, to make Web services applications easier to develop. VB.NET is the first fully object-oriented programming (OOP) version of Visual Basic, and as such, supports OOP concepts such as abstraction, inheritance, polymorphism, and aggregation.

Appendix C Index

A

Affymetrix GeneChip 2, 10, 11, 49

Affymetrix microarrays .... *See* Affymetrix GeneChip

AffyMiner 20, 31, 32, 38, 40, 47, 49, 50, 51

Algorithms 11, 16

Architecture 8

B

Bayesian 8, 14

C

cDNA microarrays 1

Cluster 13, 15, 20, 28, 31, 40, 42

Clustering 8, 11, 14, 15, 18, 28

D

Data analysis 6, 7, 9, 10, 14, 16, 49

Data flow 16

Data metrics 10, 11, 12, 13, 18, 20, 22, 33, 34, 49

Data Mining Tool 14, 32, 47, 50

Data preprocess 3, 6, 16

Descriptive statistics 6

DMT .... *See* Data Mining Tool

E

Experimental design 16

Exploratory statistics .... See Descriptive statistics

G

GCOS 11, 18, 21, 32, 47, 50

Gene expression 3, 4, 6, 8, 9, 10, 11, 12, 15, 29, 49, 51

Gene Ontology .... See GO

GeneChip .... *See* Affymetrix GeneChip

GeneFinder 20, 26, 32, 33, 47, 49

GeneSpring 9, 12, 15

GenMAPP 12, 13, 14, 20, 28, 29, 40, 42

GO 12, 13, 14, 18, 20, 25, 26, 27, 40, 49, 50, 51

GOTree 20, 25, 31, 40, 41, 47, 49

H

Hybridization 1, 3, 16, 18

I

Image processing 16

Inferential statistics 6

Information flow .... See Data flow

Ingenuity Pathways Knowledge Base 12

Interfaces 19, 20, 31, 40, 47

K

KEGG 12

K-means clustering 8, 15

M

MAPPFinder 12

MAPPs 29

Microarray experiment 3, 16, 17, 47

Microarrays 1, 3, 4, 10, 15, 49

- analysis methods 6
- applications 3, 5
- comparison between cDNA and oligo arrays 1, 2
- manufacture 2

Mixture model 7

Multiple comparisons 23

N

NetAffx 12, 32, 36, 48, 50

Network 6, 8, 14, 51

Normalization .... *See* data preprocessing

O

Oligonucleotide microarrays 1

P

Principal component analysis 6, 8, 28

Probe sets 22, 48

S

Self-organizing tree 8

Signal intensity 3, 6

Signal log ratio 13

Software 1, 9, 10, 12, 14, 15, 20, 21, 23, 32, 47, 49, 50

Super-paramagnetic clustering 8

System requirements 46

T

T-test 7, 49

Two-way clustering 8

U

User requirements 18