On the Effects of Large-scale Transcriptomics Datasets on Gene Functional Analyses

A Thesis
Submitted For the Degree Of

DOCTOR OF PHILOSOPHY

By

Prajwal Krishna Bhat

September 2011

School of Biological Sciences
Royal Holloway, University of London
Egham TW20 0EX
United Kingdom

Declaration of Authorship

I, Prajwal K. Bhat, hereby declare that this thesis and the work presented in it is entirely my own. Where I have consulted the work of others, this is always clearly stated.

Signed:

Date: 04th Feb, 2012

Acknowledgements

Guru brahma, guru vishnu, gurudevo maheshawaraha
Guru saakshaath parabrahma, thasmai shrigurave namaha

———

The teacher is the creator, he is the preserver, he is the destroyer
He is the source of the Absolute, I offer all my efforts to him

It is indeed a rare honour to be writing this section. I have benefitted from so many individuals who have unconditionally and wholeheartedly extended their support. I will never be able to adequately express my gratitude to them.

Firstly, I would like to thank my supervisors Dr. Alessandra Devoto, Prof. Dr. Laszlo Bogre and Prof. Dr. Peter Bramley for giving me the invaluable opportunity to pursue a PhD. I would like to thank Dr. Alessandra Devoto for her immense patience and unconditional support throughout the course of this project. Her words of encouragement especially made a world of difference to me. I would like to thank my advisor Dr. Alberto Paccanaro for being a mentor for which I will forever be in his debt. His energy and acumen will always be a benchmark for me. The lessons he taught have changed my science and life forever.

I would like to thank Dr. Haixuan Yang and Dr. Tamas Nepusz for their technical inputs and discussions which have enriched my work. They have been truly inspiring and I consider myself fortunate to have worked with them. I would like to specially thank Dr. Yang for his invaluable help with MATLAB throughout the project, and especially while implementing the experiment selection algorithm. I would also like to extend my thanks to Dr.

Alfonso Romero, Dr. Haixuan Yang and all the past and current members of the lab for making the lab an electrifying place to do science.

On a personal note, I would like to thank my parents for their invaluable support. I am grateful to them for stoking my scientific temper throughout my life. The long discussions with my father on science, technology and commerce were perhaps the seeds for this journey. Finally, I would like to thank my wife Akhila for her unconditional support and companionship without which this thesis would have been infinitely harder to write. I dedicate this thesis to my family.

Abstract

The Guilt-by-Association (GBA) principle, according to which genes with similar expression profiles are functionally associated, is widely applied for functional analyses using large heterogeneous collections of transcriptomics data. In this thesis we show that using such large collections could hamper GBA functional analysis for genes whose expression is condition specific. In these cases a smaller set of condition related experiments should instead be used, but identifying such functionally relevant experiments from large collections based on literature knowledge alone is an impractical task.

The study begins by discussing the basic principles underlying the definition of gene function and the use of large microarray collections for GBA based gene function analyses. We look at the effects of condition specific gene expression on GBA analyses and provide a mathematical and biological perspective. We show that using large microarray collections to calculate correlation can mask the effectiveness of the GBA principle. We suggest that using only those experiments that are relevant to the biological function under analysis can significantly improve GBA based gene functional analyses.

We then present a semi-supervised algorithm that can select functionally relevant experiments from large collections of transcriptomics experiments. The algorithm is able to select experiments relevant to a given GO term, MIPS FunCat term or even KEGG pathways. We extensively test our algorithm on large dataset collections for Yeast and Arabidopsis. We demonstrate that: (i) using the selected experiments there is a statistically significant improvement both in correlation between genes in the functional

category of interest and in GBA based function predictions; (ii) the effectiveness of the selected experiments increases with annotation specificity; (iii) our algorithm can be successfully applied to GBA based pathway reconstruction.

We conclude by discussing the potential applications of our technique. We outline several developments that could be implemented in the future to improve the efficiency of the experiment selection procedure.

Keywords

Guilt by Association, Condition specific gene expression, Experiment selection, Transcriptomics, Co-expression, Gene Ontology, Gene functional analyses

Abbreviations

GBA	Guilt By Association
GO	Gene Ontology
MIPS	The Munich Information Center for Protein Sequences
PPI	Protein-Protein Interaction
TAF	Transcription Associated Factors
SGD	Saccharomyces Genome Database
SNOMED	Systematized Nomenclature of Medicine
MIPSFunCat	MIPS Functional Catalogue
EC Nomenclature	Enzyme Commission Nomenclature
OBO	Open Biological Ontologies
DAG	Directed Acyclic Graph
BP	Biological Process
MF	Molecular Function
MJ	Methyl Jasmonate
CC	Cellular Component
TAIR	The Arabidopsis Information Resource
NCBI	National Center for Biotechnology Information
GEO	Gene Expression Omnibus
EBI	European Bioinformatics Institute
JA	Jasmonic Acid
SOM	Self-organizing maps
PCC	Pearson Correlation Coefficient
M3D	Many Microbes Database
MAS	MicroArray Suite
ROC	Receiver Operating Characteristic
LOO	Leave One Out
TPR	True Positive Rate
FPR	False Positive Rate
AUC	Area Under the Curve

List of Figures

List of Tables

Publication Based On This Thesis

Prajwal Bhat, Haixuan Yang, Laszlo Bogre, Alessandra Devoto and Alberto Paccanaro (2011) "Computational selection of Transcriptomics experiments improves Guilt-by-Association analyses", Under review.

Contents

1

Introduction

1.1 Context of the thesis

Proteins are widely recognised as the building blocks and functional components of the cell. Proteins are omnipresent; structural proteins form tissues of organisms, enzymes and other proteins regulate biochemical reactions and trans-membrane proteins act as transporters that maintain the cellular environment. Therefore, the knowledge of functions of proteins and the properties of genes which transcribe them are essential for understanding the mechanisms of biology. Gaining an insight into gene or protein function is essential for the development of new drugs, improving crop yield and development of essential biochemicals such as insulin and enzymes.

The early efforts to elucidate gene or protein function were mostly experimental and generally involved focussing on a small set of genes or a protein complex. These experiments were low-throughput in nature due to the enormous experimental resources, time and human effort required. However, the arrival of high-throughput experimental techniques such as rapid DNA Sequencing has changed the outlook of modern biology. In the subsequent phase now known as the post-genomic era, numerous high-throughput techniques have been developed, each offering a new perspective of the mechanisms by which a gene or protein executes its function. These high-throughput techniques provide us with an unprecedented opportunity to understand complex biological systems by providing insights into myriad facets of a gene's function and its dynamic interaction with other molecular components. These insights enable us to build increasingly complex models of regulatory interactions, helping us understand gene function.

1.2 Problem statement

Among the various high-throughput data available, microarrays for transcription profiling are currently the most abundant due to their accessibility, interpretability and decreasing experimental costs (Yeung, Medvedovic, & Bumgarner, 2004; Zien, Fluck, Zimmer, & Lengauer, 2003). The large quantities of data generated by microarray experiments have made manual analysis of the data impractical and a large number of computational tools and protocols have been developed to extract functional information from microarrays. Generally, most of the techniques developed to extract functional information from microarrays are based on the principle of Guilt-by-Association (GBA). The GBA principle suggests that genes that have a similar expression profile are suggested to share similar functions. This is largely based on the fact that genes which encode proteins that participate in a pathway are found to be co-regulated. GBA driven microarray analysis approaches include various clustering techniques such as hierarchical clustering, k-means clustering, biclustering and various co-expression based network analyses techniques. Generally, these techniques aim to group genes based on their expression profiles. The ways by which the genes relate to each other or the membership in a group of genes determine the function.

GBA-based microarray analyses approaches calculate similarity between gene expression profiles using a similarity metric such as Pearson correlation. Often, this has been done over a large heterogeneous collection of datasets (Manfield et al., 2006; Obayashi, Hayashi, Saeki, Ohta, & Kinoshita, 2009a; Zimmermann, Hirsch-Hoffmann, Hennig, & Gruissem, 2004). One reason behind this approach is that a large number of data points would result in a more robust correlation by combining weak expression signatures over many datasets. Formally, the significance of the correlation between the vectors is likely to increase with the size of the vector.

Generally, co-expression studies over large heterogeneous collections of datasets were aimed to reveal a global transcriptional response (Wu et al. 2002). Studies such as Zien et al. (2003) and Yeung et. al (2004) have shown that the size of the dataset has an upper limit of approx. 80 data points above which there is no discernable improvement in the quality of the analysis. We hypothesized that working with such large datasets may not always be beneficial to the functional analyses at hand.

For the principle of GBA to be effective for elucidating gene function or to perform any other functional analyses, genes which belong to the same functional category would be expected to have similar expression profiles. Consequently, the genes belonging to the same functional category are expected to be highly correlated. This property remains the cornerstone of GBA-based microarray data analyses. To verify this property, we looked at the distribution of correlation coefficients for genes which belong to the same functional category[1] such as genes annotated to the Gene Ontology Biological Process category "*Response to Jasmonic Acid Stimulus*". We prepared a microarray data collection containing 756 arrays from 44 individual experiments based on wild-type *Arabidopsis thaliana*. The data was sourced from NASCarrays (Craigon et al., 2004) and represented various experimental backgrounds such as stresses and developmental stages. Only Affymetrix ATH1 GeneChip data with MAS5.0 normalization was used. The processed data was normalized to a mean of zero and a standard deviation of 1. Further details regarding the experiments and the normalization pipeline can be found in Chapter 5 (Section 5.3.3) and Appendix I.

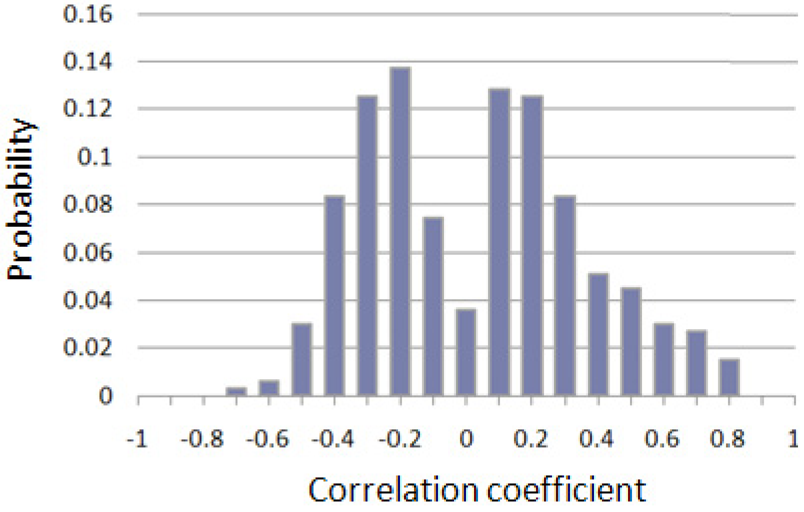


Figure 1: Distribution of correlation coefficients among genes in the GO category "Response to Jasmonic Acid Stimulus" obtained when a large collection of 44 microarray experiments (756 microarrays) were used to calculate the correlation. This result was found to be typical for most GO Biological Process categories.

[1] Only experimentally annotated genes were considered to ensure the reliability of the annotation. Details of the GO Evidence Codes used in this thesis are presented in Section 5.3.3.

Surprisingly, although the genes belonged to the same functional category we observed very poor correlation between the genes, with 81.4% of the correlation closer to zero (Fig.1). Such a distribution of correlation coefficients would limit the effectiveness of the GBA principle for functional analyses. If the correlation among genes in the same functional category is near zero, then putative genes that may belong to that functional category cannot be inferred based on correlation among genes already annotated to that category. We obtained similar results for most of the GO Biological Process categories and this was also replicated across functional classification systems such as MIPS and in other organisms such as Yeast.

We wanted to understand the causes of such poor correlation and investigate ways of limiting its effects on GBA-based analyses. We hypothesized that there could be two causes for observing such poor correlation. The first source of poor correlation could be an experimental artefact due to the noise inherent in the measurement. The second source of poor correlation could be biological phenomena such as cross-talk in the biological pathway of interest. We examine this in detail in Chapter 6. Gene function and gene expression are acknowledged to be condition-specific. Therefore, genes are also expected to be correlated based on the experimental conditions. In experiments where the pathways of interest are not activated, the genes in the pathway could show poor correlation. In cases where a large heterogeneous collection of microarrays is used for functional analyses, the poor correlations from functionally unrelated experiments could significantly dilute the high correlations observed in experiments where the pathways are sufficiently activated.

To limit the dilution of correlation, it is important that the sources of poor correlation are minimized. This would mean that in a functional analysis, the experiments where the genes of interest are poorly correlated are eliminated. However, identifying microarray experiments relevant to a functional category of interest is a non-trivial task as literature knowledge relevant to a biological process is seldom exhaustive. Further, the relevance of an experiment to a biological process may not be obvious and experiments that are deemed irrelevant by a researcher could in fact withhold significant information regarding the biological process of interest.

1.3 Objectives of the study

From the problem statement presented earlier in Section **1.2**, it is clear that the using large collections of microarrays in GBA-based analyses can result in poor correlation among genes that are deemed functionally related. This poor correlation among functionally related genes could undermine the effectiveness of GBA-based functional analyses because it would be difficult to determine the function of a gene based on its correlation with genes with known functions. In this thesis, we aim to present a comprehensive investigation of the problem of poor correlation among functionally related genes and also present a methodology for improving the correlation. The objectives of this study are summarized as follows:

1. We want to investigate the limitations of performing GBA-based functional analyses using large collections of microarrays. Specifically, we want to discuss some of the causes of poor correlation among genes involved in the same biological process.
2. We want to show that correlation among genes involved in the same process can be improved by using only functionally relevant sets of experiments. This is in contrast to the traditional approach where large numbers of experiments are used.

3. We want to show that selecting such functionally relevant experiments is a non-trivial task and we underline the need for an automated technique for identifying experiments that are relevant to a GBA-based functional analysis.
4. To address this need, we want to develop a method that is able to select from a large collection of experiments, a set of functionally relevant experiments.
5. We want to show that experiments selected using such a method would be able to improve GBA-based functional analyses. We would like to illustrate this by showing that,
   a. The selected experiments improve correlation among genes in the same functional category.
   b. The correlation resulting from the selected experiments is a better feature for classifying genes into a functional category.
   c. The effectiveness of the feature varies with the specificity of annotation.
   d. The selected experiments improve transcriptomics-based pathway reconstruction.

1.4 Overview of the thesis

Chapter 2: We discuss the concept of gene function and its interpretation in the context of post-genomic high-throughput biology. We briefly look at a few of the high-throughput experimental data types available for extracting functional information. We then briefly discuss the need for machine-friendly ontologies for organizing functional information. We outline the salient features of two widely used functional classification systems: Gene Ontology and MIPS FunCat.

Chapter 3: We investigate the principles employed for extracting functional information from microarrays. Primarily, we discuss the principle of Guilt-by-Association and its application in microarray-based functional analyses. We then highlight the importance of similarity metrics such as Pearson correlation, Mutual Information and Euclidean distance in GBA-based functional analyses. Lastly, we discuss the major functional analyses techniques that are based on the principle of GBA such as clustering, network-based approaches and basic co-expression based analytical tools.

Chapter 4: In this chapter, we investigate the reasons for observing poor correlation among genes which belong to the same functional category. We outline potential sources of noise in microarray data and suggest that these could have a far-reaching effect on any GBA-based functional analyses. We illustrate how in GBA analyses based on large heterogeneous datasets, poor correlation between genes in the same functional category can be limited by using only those experiments which are found to be relevant to the functional category of interest. We suggest that the identification of relevant datasets based on literature knowledge alone may not be efficient and propose the development of a computationally-driven method for the identification of functionally relevant experiments.

Chapter 5: In this chapter, we illustrate our experiment selection algorithm and prove its effectiveness for GBA analyses. We demonstrate that the algorithm is able to select experiments for a group of genes independent of their functional classification. We also show that the selection performance is replicable across various organisms. We demonstrate that the selected experiments lead to substantially improved correlation between genes in a functional category compared to using a large compendium of data. As a consequence, we show that using correlation obtained from the selected set of experiments leads to substantial improvements in GBA-based functional prediction.

Chapter 6: We outline the conclusions that can be drawn from the study and look at the future prospects and possible improvements of the experiment selection technique.

2

Understanding gene function

The concept of the "gene" as a discrete unit of heredity was first put forward by Gregor Mendel in 1866. The word "gene", the etymology of which can be traced to the Greek word "genos" (origin), was first used by Wilhem Johansson in the 1900s. The early concept of the gene considered it to be an abstract entity which was responsible for transmitting one or many phenotypes over generations. Subsequently a gene was seen as a physical molecule which is a blueprint for a protein. However, with the ever-increasing complexity of the insights into molecular biology, there is a need for a comprehensive framework to define a gene. Recently, Gerstein et al. (2007) defined a gene as "*a union of genomic sequences encoding a coherent set of potentially overlapping functional products*". Although the definition of the gene has continuously evolved, the intrinsic idea that a gene is responsible for a phenotype has remained constant. At the molecular level, this could mean that the DNA sequence of the gene determines the sequence and therefore the structure of the functional molecules that implement a specific phenotype. Regardless of the perspective, the identity of a gene is coupled to the corresponding functional products or phenotype.

Understanding the function of all known gene products has become the primary goal of modern molecular biology. It is widely recognized that elucidating the functions of various genes and gene products and the complex interactions between them is the key to understanding a biological system. The approaches for elucidating gene function have evolved dramatically over the past decades. Traditional approaches for elucidating gene function focussed on individual genes and were largely a single gene approach. These approaches were highly resource intensive and hence not very scalable. However, in the post-genomic era, there is a deluge of functional data from high-throughput experimentation techniques such as gene expression microarrays, protein-protein interaction experiments and genome-wide phenotype screens. The high-throughput data has necessitated novel ideas for functional analyses often adapted from computer sciences and statistics.

In this chapter, we look at gene function in the context of the changing experimental paradigms. We look at two state-of-the-art frameworks for organizing functional information, MIPS FunCat and Gene Ontology, and discuss their properties. Finally, we briefly discuss some of the few high-throughput data types available that can be exploited for linking a gene to its function.

2.1 What is Gene Function?

Traditionally, gene function elucidation techniques focussed on a single or a very small set of genes at a time. The notion of gene function was very conservative with every gene or protein having specific biological and molecular identities. For example, the molecular function of an enzyme would be its specific role in the catalysis of the reaction. Its biological function would be the pathway in which it participated. Therefore, the characterization of a gene or a protein would not be complete without elucidating the biological and molecular aspects. However, with the rapidly accumulating high-throughput experimental data, there is an unprecedented opportunity to automate gene function elucidation.

For instance, bioinformatic techniques have been successfully applied to annotate and characterize genes (Eddy, 1998; Mulder, 2003). For a gene whose functions cannot be predicted using sequence alone, a wide array of high-throughput data such as protein-protein interaction (PPI) data and co-expression data are available that can be exploited for functional analyses (Huynen, Snel, von Mering, & Bork, 2003; Vazquez, Flammini, Maritan, & Vespignani, 2003). The amount of PPI data available for the various model organisms is growing exponentially due to advances in techniques such yeast-2-hybrid (Y2H), Tandem Affinity Propagation (TAP) and Mass Spec Protein Complex Identification (HMS-PCI) (Gavin et al., 2002; Ito et al., 2001; H. Wang et al., 2007). However, data produced by these techniques suffer from high levels of false positives and false negatives. Typically, the false positive rates for Y2H are as high as 64% and TAP the false positive rate is as high as 77%. Similarly, the false negative rates could be as high as 71% and 50% respectively (Edwards et al., 2002).

Importantly, genes or proteins are no longer viewed in isolation but are recognized to be part of a complex network of interactions. A single gene can take part in multiple biological processes by having multiple collaborators and this has been recognized as the fundamental mechanism by which single genes control multiple traits. Thus the function of a gene is being defined by its interaction partners in a large network of interactions. Unlike the traditional view of gene function this new notion of gene function is fuzzy and encompasses a wide range of phenomena such as physical interaction, regulator-target interaction and co-expression. This has been termed the "*probabilistic view*" of gene function (Fraser & Marcotte, 2004; I. Lee, 2011) as opposed to the more deterministic traditional approach.

Although a rigorous definition of gene function has yet to emerge, it is appreciated that a general framework is necessary which is sufficiently robust to contain all the features of gene function (Fraser & Marcotte, 2004). One such feature is the promiscuous nature of gene function. Gene function is highly context sensitive as genes are known to be involved in multiple roles depending upon the biological process to be executed. For example, Transcription Associated Factors (TAF) have a role in DNA repair and transcriptional initiation; RAS protein regulates both mitogenesis and cytoskeletal rearrangement. Additionally, the framework also needs to keep up with the rapid pace of high-throughput analyses and accommodate new gene functions as they are discovered. With the development of computational approaches for elucidating gene function there has been a need for organizing gene functional information in a machine-friendly hierarchical ontology. Fraser and Marcotte (2004) outlined the two perspectives for organizing gene function, called the "*top down*" and the "*bottom up*" approaches. The top-down approach involves organizing all known functional information into a standardized vocabulary and organizing them into a hierarchical tree. This is similar to the current ontological projects such as the Gene Ontology (Ashburner et al., 2000a). In the top-down view, the function of a gene is defined by all the terms it is associated with in the ontology. However, in the bottom-up approach, genes or proteins are first organized into networks based on their interaction with each other. The interactions are determined using high-throughput data such as co-expression and PPI data. Here, the function of a gene is determined by its collaborators(C. v. Mering et al., 2003).Thus, the concept of gene function is fluid depending upon the nature of the investigation.

2.2 Organizing functional information: Ontologies

The need for building ontologies for describing gene function was realised at the beginning of the post-genomic era when functional analyses became increasingly driven by high-throughput data. Traditionally, functional information for genes in the various model organisms was maintained by the organism-specific databases such as Flybase (Gelbart et al. 1997) for Drosophila and the Saccharomyces Genome Database (SGD) (J. Cherry, 1998a) for Yeast. Although, such databases were very successful in their respective communities they were independent with no co-ordination between them. The potential of large scale functional analyses enabled by technologies such as microarrays revealed the potential for meta-analyses or interconnecting biological information on a global scale. A global functional ontology would enable easy access to functional information integrated in an unambiguous way.

The early approaches to describing function of a gene or a protein was based upon natural language where the functional label depended on the discretion of the investigator. Generally the functional annotation tended to be simple phrases, that are non-standard with no organizational structure

(Lan, Montelione, & Gerstein, 2003). Typical examples included gene names such as *redtape*, *roadblock*, *radish* and *turnip*. In addition to the obvious lack of organizational rule for naming, with thousands of new genes being discovered on a regular basis, not all genes could be named inventively. There is indeed a practical limitation to the number of gene names one can come up with. Evidently, due to the large variability, such a naming system was not amenable to analysis by a computer or even humans. The need for a machine-friendly, standardized, functional naming system was paramount. Some of the desirable characteristics of a potential functional naming scheme are listed below (Pandey, 2006):

1. **Wide coverage:** The functional scheme should cover the entire gamut of functions across as many organisms as possible. This is possibly the most desired characteristic.
2. **Standardized structure:** Adopting a standard data structure for the functional labels is imperative for minimizing variability between labels. This also results in easy readability and makes the label relatively computer friendly.
3. **Hierarchical structure:** Arranging the functional labels in a hierarchical arrangement starting from a general functional category leading to a specific function allows the researcher to select the relevant functional granularity for the analysis.
4. **Multiple functions:** As discussed earlier, it is well acknowledged that gene function is highly context-specific. To reflect the condition-specific nature of gene function, any functional labelling scheme would have to allow multiple labels for the same gene.

5. **Future proof:** The functional labelling scheme would be expected to be amenable to future additions. This would allow users to append new functional information as and when available.

The idea of a standard system for organizing biological knowledge was first proposed in the early 1990s with the introduction of the Enzyme Classification (E.C) numbers (Bairoch, 2000). Since then several functional classification schemes have been proposed such as EcoCyc (Keseler et al., 2005) and SNOMED (Spackman, 1997)with a majority of them being organism-specific systems. One of the early functional labelling schemes was the MIPSFunCat (Mewes et al., 2004). MIPS was one of the first classifications schemes to develop a machine-readable, standardized vocabulary for organizing functional information which was organism-independent. MIPS is widely used in bioinformatics-driven functional analyses due to its wide coverage and standardized hierarchical structure. However, one of the largest and the most comprehensive efforts in organizing functional information is the Gene Ontology Project (Ashburner et al., 2000b). The GO project was founded on strong ontological principles and is currently the most widely used functional labelling scheme with more than 7000 citations. In this thesis, generally, all the functional analyses and results are reported based on GO and MIPS functional classifications. In the following sections, we discuss the salient features of the two functional classification schemes.

2.2.1 MIPS FunCat

MIPS Functional Catalogue (H.W. Mewes et al. 2004; Andreas Ruepp et al. 2004) was one of the early attempts to generate a standardized, functional vocabulary. Unlike other functional schemes at the time such as EC (contd...)

Metabolism	
01	Metabolism
02	Energy
04	Storage Protein
Information pathways	
10	Cell cycle and DNA processing
11	Transcription
12	Protein synthesis
14	Protein fate (folding, modification and destination)
16	Protein with binding function or cofactor requirement (structural or catalytic)
18	Protein activity regulation
Transport	
20	Cellular transport, transport facilitation and transport routes
Perception and response to stimuli	
30	Cellular communication/signal transduction mechanism
32	Cell rescue, defence and virulence
34	Interaction with the cellular environment
36	Interaction with the environment (systemic)
38	Transposable elements, viral and plasmid proteins
Developmental processes	
40	Cell fate
41	Development (systemic)
42	Biogenesis of cellular components
43	Cell type differentiation
45	Tissue differentiation
47	Organ differentiation
Localization	
70	Subcellular localization
73	Cell type localization
75	Tissue localization
77	Organ localization
78	Ubiquitous expression
Experimentally uncharacterized proteins	
98	Classification not yet clear-cut
99	Unclassified proteins

Table 1: Major functional categories in MIPS FunCat and the corresponding category numbers. The numbers represent the main category identifiers.

nomenclature and SWISS-PROT (Boeckmann, 2003), MIPSFunCat focussed solely on associating gene products and functional information. Here, hierarchically structured keywords or a controlled vocabulary is used to describe gene function. The MIPSFunCat scheme was initially designed for *Saccharomyces cerevisiae* and was later extended to cover 11 other organisms including *Arabidopsis thaliana*, *Neurosporacrassa* and *Bacillus subtilis*. To account for the broad spectrum of biological processes found in the various organisms, the FunCat annotation scheme consists of 28 main functional categories that cover general functions such as cellular transport, metabolism and protein activity regulation (Andreas Ruepp et al., 2004). Each of the main functional categories is organized as a hierarchical tree-like structure. The main functional categories of the FunCat are listed in Table 1.

As discussed earlier, an important consideration for a new ontology or annotation scheme is machine readability and human usability. The FunCat scheme was aimed to find a balance between the two contrasting requirements. By design, the FunCat scheme is compact with a limited number of terms. The FunCat terms generally offer a broad classification of the gene of interest compared to similar vocabularies such as the Gene Ontology Project. Each of the functional categories is assigned a unique two digit number, as indicated in Table 1. The hierarchy between the functional categories is depicted by using a dot between the two digit category numbers e.g. **10** *Cell cycle and DNA processing*→**10.01** *DNA Processing* →**10.01.09** *DNA restriction and modification*→**10.01.09.05** *DNA conformation and modification.* The FunCat scheme allows for assigning a gene to multiple categories to accommodate for the condition-specific nature of gene functions.

2.2.2 The Gene Ontology Project

The Gene Ontology (GO)(Ashburner et al. 2000a) is the most widely used biological ontology for describing gene function and covers over 28 different organisms. The current version of the GO has over 30,000 terms and over 50,000 relationships (The Gene Ontology Consortium, 2010). The GO was developed in collaboration with a variety of biological databases such as SwissPROT (Boeckmann, 2003), GenBank (Benson, Karsch-Mizrachi, Lipman, Ostell, & Wheeler, 2008), MIPS (H W Mewes et al., 2004) and Pfam (Bateman et al., 2004). The GO consists of two discrete parts; firstly, the annotation of genes with GO vocabulary and secondly the vocabulary and the relationships between the terms in the vocabulary. The annotation of the genes to the terms is maintained by the individual organism-specific databases such as TAIR (Rhee et al., 2003) for Arabidopsis. The GO terms are created, organized and maintained exclusively by the GO consortium.

The GO is based on the Open Biological Ontologies (OBO) (Camon et al., 2004) framework and consists of three discrete classification systems called Biological Process, Cellular Component and Molecular Function. Each classification system addresses a different aspect of a gene's function. **Cellular Component** describes the location of the gene products in the cell. It is designed to describe the physical structure with which a gene or a gene product is associated e.g. *extra-cellular matrix, golgi apparatus*. **Molecular Function** is defined by GO as "the biochemical action characteristic of a gene product". The Molecular Function ontology describes the action without specifying the locality of action or the time of action e.g. *transporter*, *protein stabilization*. Biological Process is defined as "A phenomenon marked by changes that lead to a particular result, mediated by one or more gene products". Biological Process terms refer to a biological event to which a gene product contributes. The process may involve physical or chemical transformation i.e. the nature of the products before the event can be very different from the end product. Typical Biological Process terms include *response to stress, translation* and *cell cycle.*

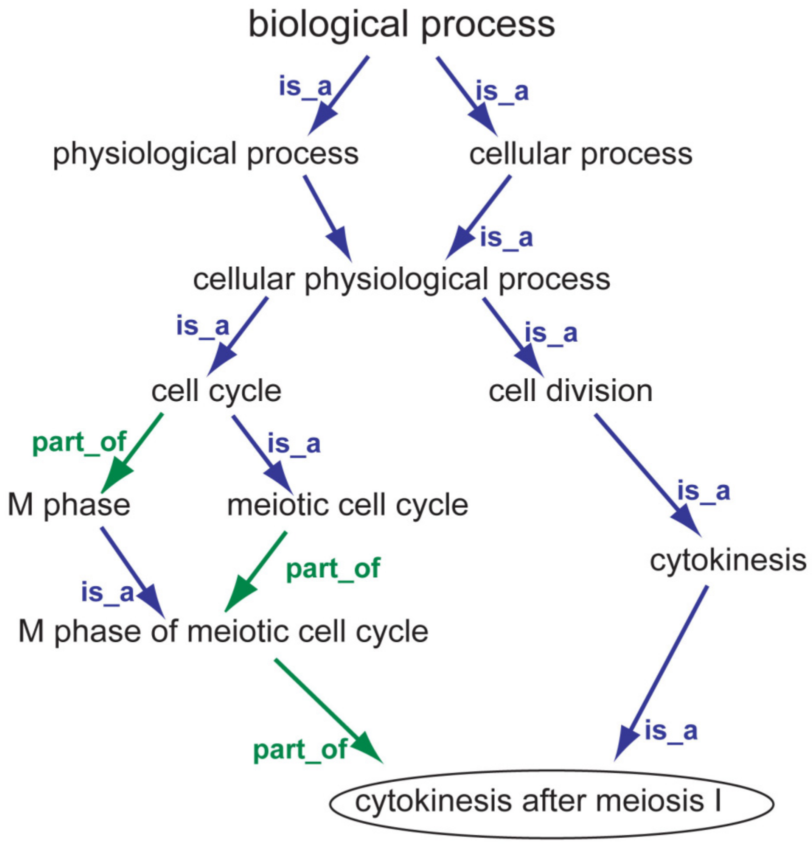


Figure 2: Structure of the Gene Ontology Biological Process Tree. The example shown above is the GO structure for the term "cytokinesis after meiosis I". Only two (is_a and part_of) of the 5 types of relationships are depicted in the figure.

The terms are arranged as a hierarchically arranged Directed Acyclic Graph (DAG) with increasing levels of granularity or specificity. The broadest term sits at the root of the DAG and the specificity increases with the distance from the root. The relation between the terms can be one of the 5 types: *is_a,* *part_of*, *regulates*, *negatively regulates* and *positively regulates*. Although in the earlier versions of GO, the Biological Process (BP), Molecular Function (MF) and Cellular Component (CC) trees were discrete with no links between them, the recent versions of the GO can accommodate links between the trees e.g. A gene can be annotated to a MF term and can have a "*regulates*" relationship to a BP term (The Gene Ontology Consortium, 2010).

The GO was designed to possess all the desirable properties for a biological ontology discussed earlier in the beginning of this chapter. The GO has a very **wide coverage** with nearly 28 databases such as SGD (J. Cherry, 1998b), EcoCyc (Keseler et al., 2005) and TAIR (Rhee et al., 2003). The terms in the GO follow a **standardized format** with each node having a unique GO id of the form GO:XXXXXXX e.g. GO:0009560. The terms are arranged in a DAG with well-defined relationship between them e.g. *is_a*. This allows the gene products to have multiple parents and multiple children; this design allows for denoting **multiple functions** for the same gene. The well-defined structure makes GO relatively better suited for computational applications and also for human readability. The three **disjoint** ontologies provide a multi-dimensional view of gene function. Finally, the GO is designed to be **dynamic**. As new research uncovers novel functions, the curators can easily incorporate the knowledge into the GO. These features have heavily contributed in making the GO the *de facto* standard for functional annotation.

3

Functional analyses using gene expression data

Functional analyses of genes and elucidation of gene function is central to the aims of Systems Biology. This includes understanding biological systems by uncovering novel gene interactions, reconstructing regulatory and metabolic pathways and characterizing novel genes or proteins that may be involved in these biological processes (Kitano, 2002).

Systems biology has adopted two main ideologies for functional analyses; "*top-down*" and "*bottom-up*" approaches[2](Bruggeman & Westerhoff, 2007; Noble, 2002). The bottom-up approach is largely based on prior knowledge. Prior knowledge about various biological components such as a gene regulatory pathway or reaction kinetics of enzymes is integrated and used to model the behaviour of the biological systems of interest. In contrast, the

[2] Not to be confused with the "top-down" and "bottom-up" principles of organizing gene function discussed in Chapter 2.

Top-down approach is based on sampling data concerning the various biological components in various experimental conditions. This is a generic approach and does not require any prior knowledge regarding the biological systems under investigation. Responses to experimental conditions can be sampled at various levels such as the transcriptome using microarrays, the proteome using the yeast 2-hybrid technique (Ito et al., 2001) or the metabolome using mass-spectrometry (Fiehn, 2002). The choice of data source depends entirely on the nature of the biological phenomena under investigation. Large quantities of data obtained from these techniques are mined using analytical techniques which are designed to look for patterns in the data that may support a prior formed hypothesis.

Gene expression data-based functional analyses are based on the top-down perspective of systems biology. The primary aim is to mine gene expression data to characterize novel genes and uncover novel relationships between genes. These interactions are summarized to form biological processes and pathways. The *Guilt-by-Association* (GBA) principle has been the most successful approach for mining functional information from gene expression data. In the following sections, we will take a closer look at the GBA principle and its general applicability in microarray data analyses. Subsequently, we investigate the role of similarity metrics in GBA-based analyses. These concepts are crucial for understanding the problem presented in this thesis.

3.1 Guilt-by-association in microarray data analyses

Microarray experiments focus on identifying patterns of gene expression in response to a treatment or stimulus such as chemical treatment or stress time course or simply a comparison between two or more tissue types such as wild-type and mutant. The gene expression profiles from these microarray experiments can be exploited for uncovering cellular pathways and the interaction between the myriad pathways in a biological system. Appendix I provides an overview of widely used microarray technologies and discusses the strategies used in extracting functional information from raw microarray data.

From the very beginning of the age of microarrays, it was evident that the biological functions of genes could be uncovered by applying the principle of guilt-by-association (GBA) (Quackenbush, 2003; Wolfe, Kohane, & Butte, 2005). In the context of gene expression data analysis, the GBA principle states that *genes with similar expression profiles may share similar functions*. This is based on the observation that genes encoding proteins that participate in a metabolic pathway are generally found to be co-regulated. Also clusters of genes with related functions often exhibit expression profiles that are correlated under several experimental conditions (Eisen, 1998; Ihmels, Bergmann, & Barkai, 2004; Stuart, Segal, Koller, & Kim, 2003). Therefore, the principle of GBA is based on the idea that a co-ordinated gene expression profile across several experimental conditions suggests the presence of a functional linkage.

Although the GBA approach, in the context of gene expression, holds great promise in the quest for uncovering gene function, its application is not without criticisms. It is important to note that the ideal end point for the description of a biological system would involve measuring protein levels and their respective activities rather than being limited to mRNA expression measurements alone. Studies by Gygi et al. (1999) have found that the correlation between mRNA and protein levels is insufficient to predict protein expression levels from quantitative mRNA measurements. In other words, the protein and mRNA abundance were found to be poorly correlated. Studies such as the ones by Clare & King (2002) clustered yeast microarrays and found poor correlation between the genes in the clusters and the functional annotations. Based on such reports, it was argued that GBA may be an unsatisfactory approach for understanding biological systems. Also, it can be observed that genes that are co-regulated may not necessarily be co-expressed and genes which are co-expressed are not necessarily functionally related. For example, it is well known that gene regulation is dependent on the presence of sequence motifs such as *cis* elements. However, cis-regulatory motifs have been shown to occur by chance in the genome leading to unexpected gene regulatory events. Events such as these would be hard to detect when analyzing gene expression data from single organisms.

Additionally, phenomena such as post-translational modifications of proteins heavily influence protein structure and functions. Such modifications cannot be detected by measuring gene expression alone. In such instances, GBA would not be effective for elucidating the functional roles of the genes.

Regardless of the criticisms, GBA has proven to be the most effective approach for the functional analyses of microarrays. Allocco, Kohane, & Butte (2004) show that there is a high degree of agreement between clusters of gene expression profiles and GO-based functional categories. This result is in contradiction to the results obtained by Clare & King (2002). Allocco et al. (2004) explain the discrepancy by suggesting that their more comprehensive approach is better suited for analyzing subtle similarities in gene expression profiles. They also attribute their superior results to the use of a larger and more comprehensive microarray collection. Regardless of the conflicts, a large corpus of microarray-based functional analyses techniques have been leveraged based on the idea of GBA. The techniques vary in their complexity ranging from simple correlation-based analyses (Manfield et al. 2006; Obayashi et al. 2009; Steinhauser et al. 2004; Usadel et al. 2009), gene expression profile clustering approaches (Andreopoulos, An, Wang, & Schroeder, 2009)to the recent network-based approaches (Babu, Luscombe, Aravind, Gerstein, & Teichmann, 2004; Long, Brady, & Benfey, 2008; Y. Wang, Joshi, Zhang, Xu, & Chen, 2006; Wolfe et al., 2005; Zhou et al., 2005). These techniques are briefly outlined later in section 4.2.

3.2 Similarity metrics for GBA analyses

At the core of the GBA principle lays the question of assessing the similarity between gene expression profiles found in microarray experiments. Microarray experiments take the form of a series of measurements taken at various points in time, response to treatments, as a comparison between biological samples or as a combination of the above. The vector containing the set of measurements is called the gene expression vector or gene expression profile. The primary step in any microarray-based functional analysis is to calculate the distance or the similarity between every pair of genes in the dataset. The resulting matrix of distance or similarity metrics is the starting point for nearly all data mining methods such as clustering and network construction.

The commonly used similarity measures can be divided into two main classes; Distance metrics and Similarity metrics. Distance metrics include Euclidean distance and City Block distance. Often used similarity metrics include Pearson correlation coefficient, Spearman's rank correlation coefficient and Mutual Information. Among the various similarity and distance metrics available, Pearson correlation, Euclidean distance is more widely used in microarray literature (Wen 1998; Khan et al. 2001; S K Kim et al. 2001; Lein et al. 2007; Eisen et al. 1998) although recently Mutual Information has also been applied in gene expression analysis (Margolin et al., 2006; Priness, Maimon, & Ben-Gal, 2007). Each of the similarity metrics has unique characteristics and the choice of the metric solely depends on the nature of the analysis. One commonly used approach in gene expression analyses is to select the distance measure which yields the best proportion of functionally related genes (Gibbons & Roth, 2002). In following sections, we present the properties of Pearson correlation and Euclidean distance.

3.2.1 Euclidean Distance

Euclidean distance is one of the most widely used distance measures in gene expression analyses. It is simply the straight line distance between two points (Fig.3).

In two dimensions, the distance is calculated using the Pythagorean Theorem (Eq. 1).

$$d(X,Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

(Eq.1)

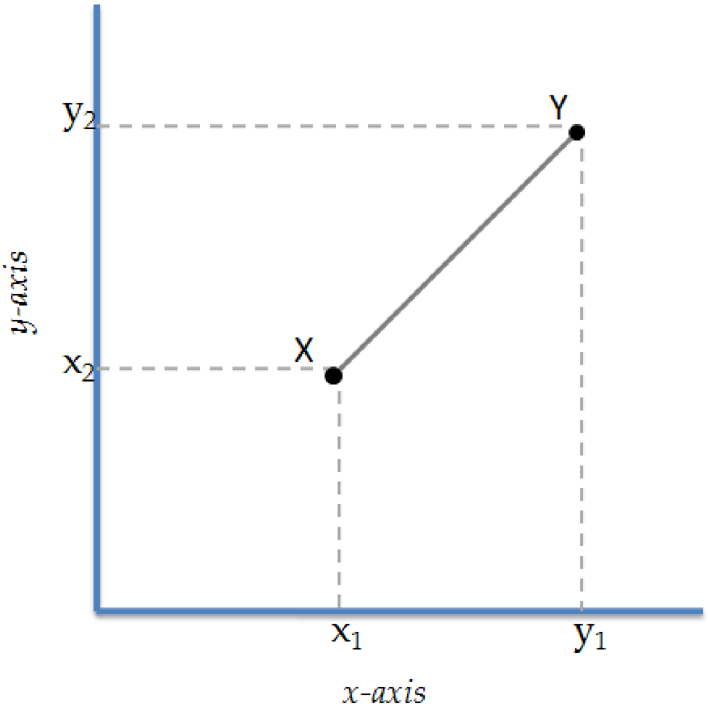


Figure 3: The Euclidean distance between two points *x* and *y* in a two dimensional space.

This concept is easily extended to higher dimensions found in gene expression profiles. Considering two gene expression profiles X and Y containing *n* measurements or dimensions, the distance between X and Y can be calculated using the formula illustrated in Equation 2.

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

(Eq.2)

Together with most distance measures, Euclidean distance follows a common set of ground rules (Stekel, 2003) which are as outlined below:

Given two vectors *x* and *y*,

1. Distance between *x* and itself must be zero.

$$d(x, x) = 0$$

2. Distance between x and y must be equal to the distance between y and x.

$$d(x, y) = d(y, x)$$

3. Distance between x and y cannot be negative.

$$d(x, y) > 0 \quad \forall\ (x, y)$$

4. Given another vector z, the distance between x and y must be less than the sum of the distances between x and z and y and z. This is also called the law of Triangle Inequality.

$$d(x, y) < d(x, z) + d(y, z)$$

Euclidean distance measures the similarity between two expression profiles based on the intensity of expression or the magnitude of the curve. In the context of gene functional analyses where the aim is to identify genes with similar expression profiles without considering the magnitude of expression, this property of Euclidean distance can be viewed as a limitation. The problem is illustrated in Fig.4, where we have three gene expression profiles measured over 5 data points. In this example, Euclidean distance would classify Profile B and Profile C as similar and Profile A would be considered as dissimilar. Although Profile A and Profile B have similar expression dynamics, they would be considered dissimilar.

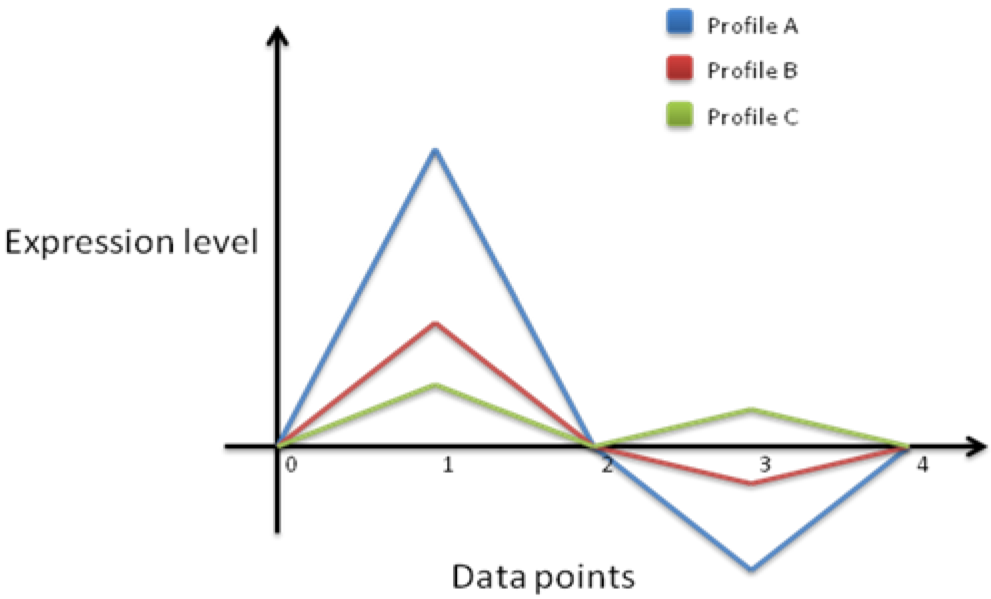


Figure 4: Hypothetical expression profiles of three genes over 5 data points. The choice of similarity metric dictates the similarity between the three expression profiles. Profile A and B would be considered similar according to Pearson correlation whereas Profiles B and C would be consider similar according to Euclidean distance.

3.2.2 Pearson correlation coefficient

Pearson correlation quantifies the similarity between two sets of gene expression measurements. If we denote the two sets of measurements with the notation (x_i) and (y_i), where $i$is an index from 1 to the total number of measurements (denoted by n), then the correlation coefficient r is given by the formula (Eq. 3):

$$r(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

(Eq. 3)

The Pearson correlation coefficient is bound between -1 and +1. A correlation coefficient of -1 between two gene expression profiles indicates that the shape of one expression profile is the exact mirror of the other i.e. in cases where one gene is activated and the other gene is correspondingly repressed. A coefficient of +1 indicates that the shapes of the two expression profiles are very similar to each other i.e. when two genes are similarly activated or repressed. A correlation coefficient of zero indicates that there is no similarity between the two expression profiles. Generally, in clustering applications where distance metrics are used, the absolute value of the Pearson correlation coefficient is subtracted from 1 to obtain a correlation distance (Eq.4).

$$d = 1 - |c|$$

where:

c is the correlation coefficient

(Eq.4)

In Pearson correlation, the significance of the correlation between two vectors or expression profiles can be calculated to obtain a *p*-value. The significance refers to the likelihood of having observed the correlation simply by chance alone. The *p*-values are obtained by transforming the correlation values to follow a *t* distribution using the formula in Eq. 5:

$$s = r\sqrt{\frac{n-2}{1-r^2}}$$

where:

r is the correlation coefficient

n is the number of pairs of data

source: (Neter, Wasserman, Kutner, & Li, 1996)

(Eq. 5)

Unlike Euclidean distance, Pearson correlation depends on the shape of the curve rather than the intensity or magnitude. In Fig.4, profile A and profile B would be classified as similar and profile C would be relatively dissimilar to profile A and B. This key property makes Pearson correlation well suited for clustering gene expression profiles. For this reason, we use Pearson correlation coefficient for our analyses.

It is reasonable to believe that genes that belong to a biological pathway or process can be positively or negatively correlated due to phenomena such as negative feedback loops in biological pathways. One such example would be the negative feedback loop found in the plant circadian clock machinery where (Alabadí et al., 2001) have shown that two MYB transcription factors called LHY and CCA1 repress the activation of their activator TOC1. For these core circadian genes, we retrieved microarray data from a time-series experiment conducted on wild-type Arabidopsis (NASCArray Database ID 137). Here, gene expression was measured at 30 min, 1h, 3h, 6h, 12h and 24h. We found that the LHY was positively correlated with CCA1 (r = 0.92) and TOC1 was negatively correlated with LHY (r = -0.85) and CCA1 (r = -0.78). For this reason, in this thesis, when considering the correlation between genes belonging to the same functional category we use the absolute value of the correlation coefficient ($r = |c|$).

4

Limitations of compendium-based correlation analyses

4.1 The emergence of microarray repositories

There has been a dramatic increase in the number of published microarrays for model organisms such as Yeast, Arabidopsis and *E.coli* in the recent times. A PubMed search reveals that over 20,000 papers have been published presenting microarray data in the last decade compared to just over a hundred at the end of the 90's. The earliest efforts to setup microarray repositories were led by consortiums that had earlier collaborated on sequencing projects. Some of the most comprehensive microarray repositories include NCBI's Gene Expression Omnibus (Edgar, Domrachev, & Lash, 2002) and EBI's Arrayexpress (Brazma, 2003) which are not limited to specific organisms. The number of microarrays currently available for a few model organisms has been listed in the Table 2. In addition to these, several organism-specific repositories have been set up such as NASCarrays for *A.thaliana*, SGD for Yeast and Flybase for *D.melanogaster*. Generally, these databases contain thousands of arrays from a wide spectrum of experimental backgrounds such as time courses, treatments, tissues and phenotype comparisons. Such organism-specific collections of microarrays are often termed as a "compendium", a term first used by T. R. Hughes et al.(2000).

S. no.	Organism	Number of Experiments
1	Arabidopsis thaliana	1349
2	Saccharomyces cerevisiae	1031
3	Mus musculus	6093
4	Escherichia coli	316
5	Homo sapiens	8913

Table 2: Number of microarray experiments available for model organisms in NCBI's GEO repository as of 2011.

4.2 Co-expression analyses tools for microarray compendia

Co-expression analysis simply involves identifying similarity in gene expression over a set of experimental conditions. The classical approach to microarray analyses involves identifying and comparing genes which have been expressed or repressed at a given experimental condition. However, instead of being limited to isolated experimental conditions, co-expression analysis takes a global view where the gene expression dynamics are considered over many experimental conditions.

In the post-genomics era, the high-throughput sequencing efforts have resulted in over 1200 organisms being sequenced till date (Szklarczyk et al., 2011). However, even in well studied organisms such as Yeast and Arabidopsis, the functional roles of only a fraction of the genes have been experimentally ascertained. In the previous chapter, we presented how GBA principle has been widely used in exploratory analyses of genes with unknown functions. Co-expression analysis is one of the simplest implementation of the GBA principle in microarray data analysis and can reveal the functional and the organizational relationship between genes. Although co-expression does not necessarily imply co-regulation of genes (Joshua M. Stuart et al., 2003), a large number of studies have been presented that support the applicability of co-expression analyses (e.g. Horan et al. 2008) used a co-expression based approach to identify 104 genes of previously unknown gene function as being involved in abiotic stress response in *A.thaliana*).

In this thesis, a microarray experiment is a set of microarrays measuring gene expression over a time course or a series of treatments (as shown in Fig. 5A). A microarray compendium is a collection of such microarray experiments (Fig. 5B). The primary step in a co-expression analysis is to calculate the similarity between gene expression profiles over a set of experimental conditions. This is generally done using various similarity metrics such as the ones discussed in Chapter 3. However, Pearson Correlation Coefficient is the most commonly used similarity metric in co-expression analysis.

At the outset, co-expression analysis was performed on single microarray datasets, where co-expression was measured over all the conditions (contd...)

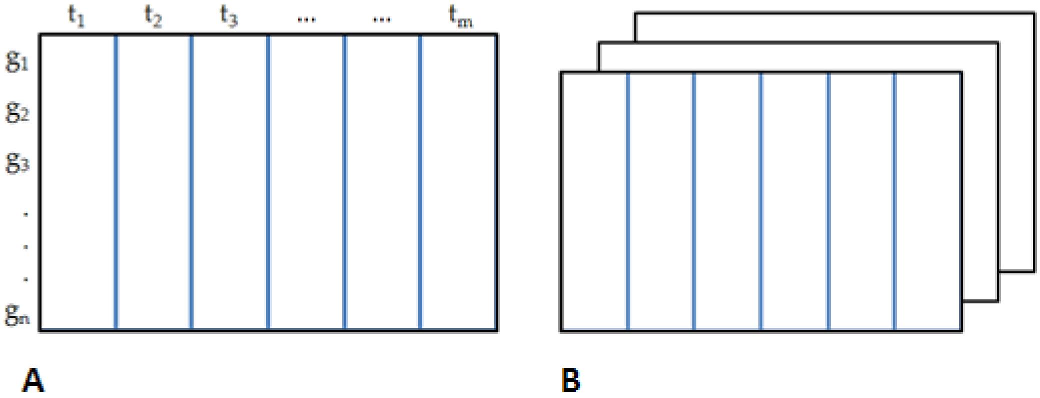


Figure 5:A. Microarray datasets are in matrix form, where the rows are genes and columns are experimental conditions. In the figure, *n* is the number of genes and *m* is the number of experimental conditions. B. Several such matrices, each representing a microarray experiment are appended together to form a microarray compendium.

in the experiment (Eisen et al. 1998). The aim was to uncover genes which showed co-ordinated responses in the given experiment. However, with the proliferation of microarray technology and easy availability of hundreds of datasets as a compendium, it was realized that co-expression could be measured over multiple experiments. Aggregating multiple microarray experiments allowed investigators to uncover a *global* co-expression pattern. The increased number of data points results in a more robust correlation as any weak correlation signatures are combined over many datasets. Importantly, the significance of the correlation between two gene expression profiles increases with the size of the experiment (denoted by *m* in Fig.5A).

Often, co-expression analysis does not require any programming expertise as a large number of microarray databases have integrated co-expression tools. Based on the data used, co-expression tools are of two types:

A. Condition-independent co-expression

In condition-independent co-expression analysis, the user inputs microarray datasets without any discrimination between tissue types and experimental conditions with the aim of using as many datasets as possible. This approach is suitable for illustrating a more general relationship between the genes and is for a general exploratory investigation of gene function. Most of the co-expression tools and resources available perform condition-independent co-expression analysis. Examples of some of the condition independent co-expression tools available are presented in the Table 3.

B. Condition-dependent co-expression

Condition dependent co-expression analysis uses user-selected sets of data for calculating co-expression. This approach is based on the condition-specific nature of gene expression and also gene function. The promiscuous nature of gene interaction and its effect on co-expression based function elucidation is discussed in detail in the later sections of this chapter. Generally, condition-dependent co-expression tools allow users to specify the type of experiments they would like to use in the analysis. Users can choose experiments with common biological themes such as specific plant organelles or cell lines allowing users to analyze co-expression in the context of the pre-defined biological backgrounds. Table 3 presents examples of condition-specific co-expression tools available online.

Regardless of the type of data available, a co-expression database can be queried either using a single gene or multiple genes or a gene list. Most co-expression databases allow both single and multiple genes as queries. In the single gene approach, the query gene is often called the "bait". The bait gene may be a transcription factor or a key member of a gene regulatory pathway of interest. The bait is used to retrieve a list of genes which are correlated with the bait. The size of the list is generally limited by a correlation threshold or simply by a limit on the length of the list such as top *N* correlated genes. In the multiple gene query, a list of genes of interest can be used to query the co-expression database. The aim of performing a multiple gene query is to retrieve co-expression information between the query genes themselves. This approach can be used to search for co-expressed gene combinations between two gene families (Usadel et al. 2009). Multiple gene querying technique can also be used to illustrate any co-expression between genes involved in a protein-protein interaction dataset. Ma, Gong, & Bohnert (2007) adapted multiple gene co-expression analysis to retrieve a list of *bait* genes that were then used to re-query the co-expression database. This approach was able to retrieve functional relationships between genes which were previously undetected using primary correlators alone.

Sl.No	Tool	Condition dependent	Author
1	ACT	No	(Manfield et al., 2006)
2	Genevestigator	No	(Zimmermann et al., 2004)
3	ATTED-II	Yes	(Obayashi, Hayashi, Saeki, Ohta, & Kinoshita, 2009b)
4	BAR Expression Angler	Yes	(Toufighi, Brady, Austin, Ly, & Provart, 2005)
5	Cress-Express	No	(Srinivasasainagendra, Page, Mehta, Coulibaly, & Loraine, 2008)
6	CSB.DB	Yes	(Steinhauser et al., 2004)
7	GeneCAT	No	(Mutwil, Obro, Willats, & Persson, 2008)
8	PED	No	(Horan et al., 2008)

Table 3: List of condition-dependent and independent tools available for co-expression analyses

In gene functional analysis, co-expression data is generally subjected to downstream analyses such as Gene Ontology term enrichment analysis (Al-Shahrour, D’iaz-Uriarte, & Dopazo, 2004; Eden, Navon, Steinfeld, Lipson,

&Yakhini, 2009; D. W. Huang, Sherman, & Lempicki, 2009; Maere, Heymans, & Kuiper, 2005; Q. Zheng & Wang, 2008) on the members of the correlated genes list. This can provide an abstracted view on the biological roles of the genes in the correlated genes list.

An emerging form of co-expression analysis is the construction of gene networks from co-expression data. In gene interaction networks, genes are represented as nodes in a fully connected graph where the edges are weighted by the co-expression scores. The central aim of this approach is to illustrate the organizational and functional relationships between genes of interest. Co-expression data can be visualized as networks using network drawing tools such as Pajek (Batagelj & Mrvar, 1998), Cytoscape (Shannon et al., 2003a), BioLayout (Enright & Ouzounis, 2001) and Gephi (Bastian, Heymann, & Jacomy, 2009). Co-expression networks, along with many other forms of biological networks, have been shown to exhibit a scale free architecture similar to the Internet or Social networks (A.-L. Barabási & Oltvai, 2004; Jeong et al., 2000). The scale-free nature of biological networks presents certain interesting characteristics. One such property is the presence of a large number of genes (hubs) with small number of interactions (edges) and similarly a small number of genes have a large number of connections. Also, biological networks are very robust and not susceptible to a breakdown when nodes are deleted randomly (Albert, Jeong, & Barabasi, 2000). Importantly, representing co-expression data as networks allows for the application of graph theoretical approaches for elucidating gene function e.g. GeneMANIA (Mostafavi, Ray, Warde-Farley, Grouios, & Morris, 2008).

4.3 Poor correlation between functionally related genes

Following the discussions in the preceding chapter (Section 3.1), it is evident that microarray data based functional analyses are predominantly based on the GBA principle. The majority of the techniques developed for gene function elucidation are aimed at identifying groups of genes that are co-expressed. This is based on the assumption that genes that are co-expressed could be co-regulated. This assumption is indirectly supported by studies such as (T Ideker et al., 2001; Tavazoie, Hughes, Campbell, Cho, & Church, 1999; Wolfsberg et al., 1999) that show that clusters of similar gene expression profiles often share common upstream sequence motifs. Therefore, high correlation is accepted as an important indicator of functional similarity.

Drawing from the principle of GBA, it is reasonable to expect that the distribution of correlation coefficients between genes belonging to the same functional category show high positive correlation or both positive and negative correlation. However this is not the case. As presented previously in Chapter 1, for *A.thaliana* genes belonging to the GO Biological Process category GO:009753 “Response to Jasmonic acid stimulus”, we retrieved microarray data from a compendium of 44 microarray experiments containing 756 arrays.

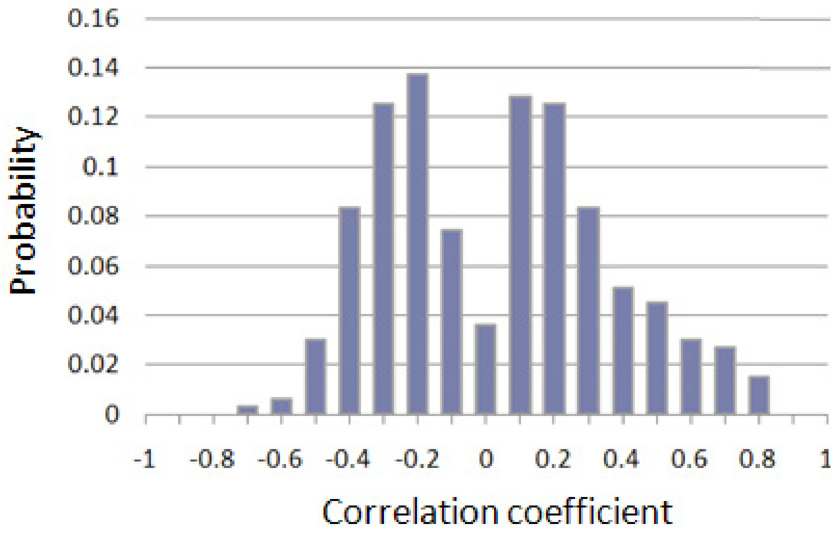


Figure 6: Distribution of correlation coefficients among genes in the GO Biological Process category GO:009753 "Response to Jasmonic acid stimulus" shows poor correlation overall. The correlation was calculated using data from 44 experiments (756 microarrays). The distribution is clearly enriched with low correlations.

The details of the experiments contained in the compendium are outlined in Chapter 5. Correlation coefficient between the expression profiles of every gene-gene pair were calculated over all the experiments in the compendium (the compendium contains 44 microarray experiments, containing 756 microarrays in total). The correlation coefficients were filtered for significant correlations ($p<0.05$) and were plotted as a histogram, shown in Figure 6. We found that only 18.6% of the gene pairs showed an absolute correlation above 0.5. In general, most of the gene pairs were found to share poor correlation despite being annotated to the same functional category. The poor correlation between the genes limits the effectiveness of GBA in gene expression-based functional analyses. In fact, if the correlation among genes in the same functional category is close to zero, the genes that may belong to that functional category may not be inferred based on correlation among genes already annotated to that functional category. Importantly, we found this effect across over 30 GO Biological Process categories we tested (data not shown); we also found it to be independent of the functional classification systems (e.g. GO Biological Process, MIPS) and organisms (e.g.Yeast, *A.thaliana*).

4.4 Causes of poor correlation among genes annotated to the same functional category

In this section, we explain how there could be poor correlation among genes which belong to the same functional category. We hypothesized two possible sources of poor correlation.

A. The extensive cross-talk present between various biological processes

Biological processes such as signalling, gene regulation and metabolic pathways involve cascading events and seldom occur in isolation. The cascade may contain a number of genes or proteins interacting with other genes or proteins downstream. Cascading events such as signalling do not necessarily occur linearly but rather through a complex web of interactions where genes are shared between pathways to bring about a non-linear response. These specific interactions between genes of multiple pathways are termed as *Crosstalk*. The phenomenon of signalling crosstalk has received much attention in biology as it has been observed that the specificity of biological responses to cues is largely due to the combinatorial integration of crosstalk. This phenomenon is salient in the plant defence response where in the absence of a dedicated immune system, plants activate a series of complex responses that lead to local and systemic induction of anti-pathogenic defences (Hammond-Kosack & Jones, 1996). Inherently, crosstalk is highly condition-dependent. Based on the condition, genes may interact or not interact with other genes in the same biological pathway. An example would be the crosstalk between phytochrome and cryptochrome signalling in *A.thaliana* where the proportion of blue, red and far red light in white light is interpreted by the organism is different ways (Casal & Mazzella, 1998). In experiments studying hypocotyl elongation, under short exposures of blue light in a red light background, the activity of the cry1 gene and phyB are related. However, during prolonged exposure to blue light the activity of cry1 and phyB are seen to be independent. Therefore, although cry1 and phyB participate in the same biological process, the correlation between them would be condition (i.e. blue light in a red light background) specific. Similar examples can be found in various other organisms such as the crosstalk in the glucose signal transduction pathways in Yeast (Kaniak, Xue, Macool, Kim, & Johnston, 2004).

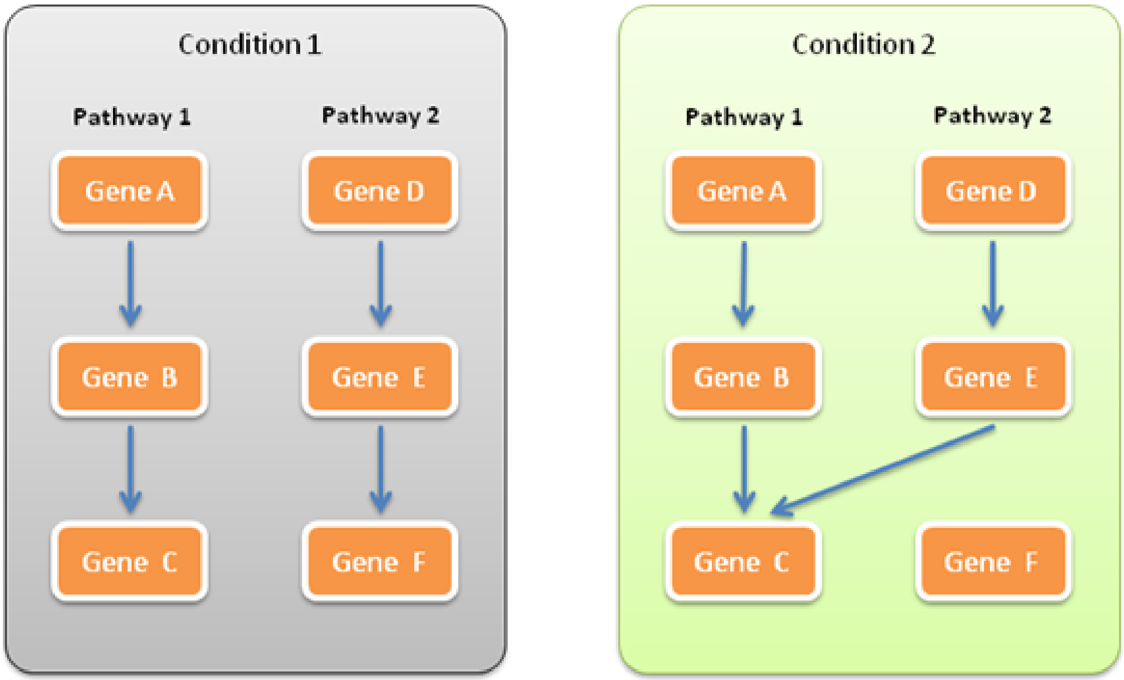
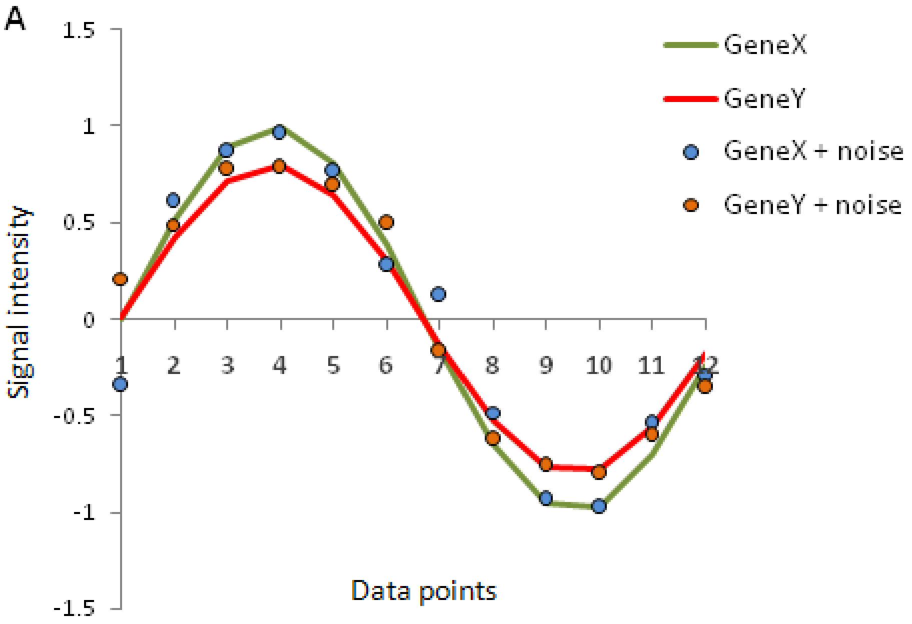


Figure 7: Phenomenon such as biological cross-talk can lead to poor correlation between genes in the same biological pathway. In condition 1,

genes in pathway1 are highly correlated and the same applies to genes in pathway2. However, in condition 2, as pathway2 is only partially activated the overall correlation could be lower.

To understand the effect of the condition dependent-nature of signalling crosstalk on correlation based functional analyses, consider two pathways (as shown in Fig. 7), pathway1 with genes A, B, C and pathway2 with genes D, E and F. Consider a condition, condition 1, where pathway1 and pathway2 are independent of each other. In this case, correlation between genes D, E and F can be expected to be high. Consider another condition, condition 2, where gene B from pathway2 interacts with gene C from pathway1. In such a scenario, the correlation between genes D, E and F would be lower as gene F would not have been activated.



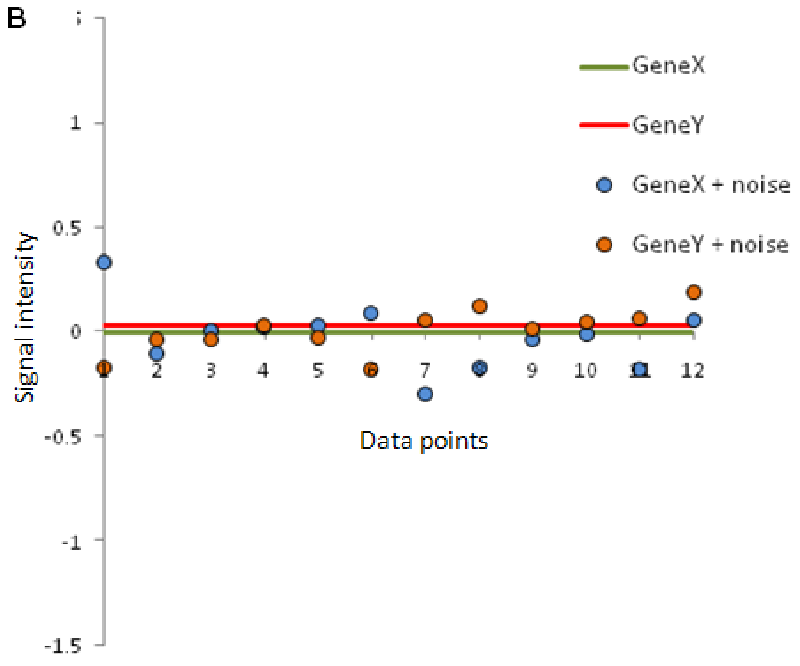


Figure 8: Simulated gene expression profiles for GeneX and GeneY over 12 data points. A. In the ideal case, GeneX and GeneY would be highly correlated even in the presence of noise. B. In the absence of a signal, only the inherent noise remains which is often very poorly correlated.

Therefore, it is possible that the compendium of microarray experiments used to calculate the correlation coefficients for the genes in Fig.6 contained a large number of experiments that are functionally unrelated to the functional category of interest.

B. Noise contributes to poor correlation

We hypothesized that another reason for the poor correlation between genes from the same functional category it could be experimental artefact due to noise in the amount of mRNA measured in the experiment. Consider a condition where two functionally similar genes are induced over a time course. The expression profiles of the two genes would be highly correlated regardless of the presence of a reasonable amount of noise. Consider a different condition, where the two genes have not been induced. Ideally, one would expect an absence of any signal, resulting in high correlation between the two genes. However, in the absence of any signal, only the noise is recorded which results in poor correlation between the two genes. To understand this better, we ran a simulation with artificial data where two genes, GeneX and GeneY, had been induced and their expression profile was been measured over 12 time points. Under ideal conditions, the two genes were found to be highly correlated with an *r* value of 0.99 (Fig. 8A). Noise was introduced by adding random values drawn from a Gaussian distribution to each of the measurements in the signal. However, even when the signals were noisy, the two genes were highly correlated with an r of 0.95 (illustrated by Eq.6).

$$c(s_1 + n_1,\ s_2 + n_2) = 0.95$$

(Eq. 6)

Where, s_1 and s_2 are the signal vectors corresponding to GeneX and GeneY. n_1 and n_2 are the corresponding noise vectors.

Consider an experimental condition where GeneX and GeneY have not been activated (Fig. 8B). In the ideal case, no mRNA would be recorded and hence there would be no signal. In this case, the two genes were found to be highly correlated with r = 1. However, in the real world, in the absence of any signal only noise is recorded resulting in a very low r value of 0.1(illustrated by Eq. 7).

$$c(\emptyset_1 + n_1,\ \emptyset_2 + n_2) = 0.1$$

(Eq. 7)

Where, $\emptyset_1$ and $\emptyset_2$ are the signal vectors corresponding to GeneX and GeneY. n_1 and n_2 are the corresponding noise vectors.

4.5 Poor correlation can rapidly dilute overall correlation in the compendium

The high level of technical noise due to the sensitivity of microarray technology has been widely discussed in literature (Klebanov & Yakovlev, 2007; E. Marshall, 2004). Several techniques such as various normalization techniques such as the ones discussed in Appendix I have been incorporated into microarray data analysis to limit the effect of noise on the experimental outcome. However, the effects of noise on the overall analysis when microarray data are pooled together such as in a compendium has not been very well investigated. We believe that for reasons discussed in section 5.4, genes annotated to the same functional category could be poorly correlated in certain experiments contained in the compendium. We believe that the instances of poor correlation can have a diluting effect on the overall correlation. We demonstrated the diluting effect using artificial data.

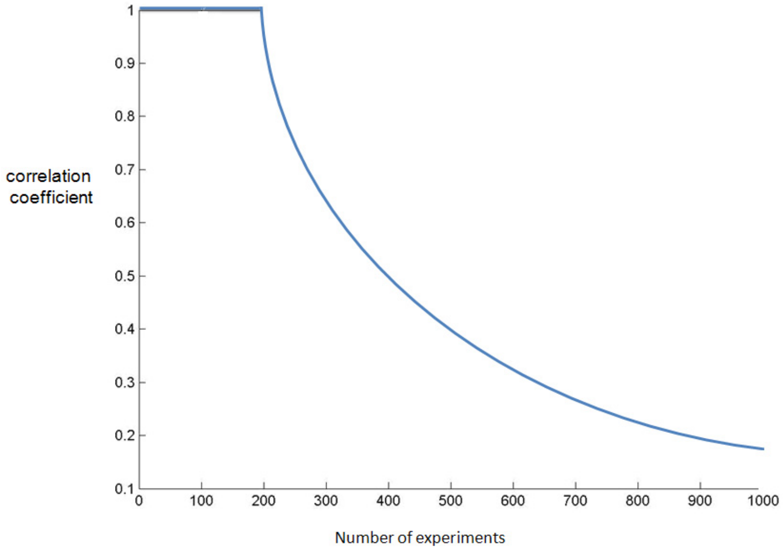


Figure 9: Plot showing the change in average correlation as the number of data points is increased. It can be clearly observed that there is high average correlation up till the 200th data point. As the noisy data points are added, the average correlation drops rapidly.

The artificial dataset contained 10 vectors representing genes, each with 1000 data points representing measurements from experiments. The first 200 data points were generated to produce a high correlation between the 10 genes and the next 800 data points were generated at random to represent noisy experiments. We measured the average correlation between the genes as each of the data points were added incrementally (Fig.9). As expected, the first 200 data points showed a high average correlation ($r = 1$). However, as the noisy data points were added, the average dropped dramatically to a very low value ($r = 0.2$). This result suggests that correlating genes over a large number of arbitrarily chosen experiments may not necessarily be optimal.

4.6 Disregarding functionally irrelevant experiments improves correlation

For the reasons outlined in the previous sections, we believe that correlation between genes annotated to the same functional category could be improved by limiting the sources of noise. In order to limit noise, we propose using only those experiments that hold some relevance to the functional category of interest and eliminating any irrelevant experiments from the analysis. We believe that using only the biologically relevant experiments in the analysis can have a significant impact on the correlation between the genes in the same functional category. To demonstrate the improvement in the correlation, we again consider the distribution of correlation, presented in

Fig.10. Here, for genes belonging to the GO category "Response to Jasmonic acid stimulus", we calculated the correlation coefficients using all the experiments in the compendium without any discrimination in the choice of the experiments. We observed that this resulted in very poor correlation between the genes, even though they belonged to the same functional category. We repeated this experiment; however, we calculated the correlation between the genes using data from only those experiments that are deemed functionally relevant. We selected the experiments based on literature knowledge. For the GO category "Response to JA stimulus" presented in the example, we selected the Wounding Times Series experiment and the MJ (Methyl Jasmonate) treatment Time Series experiment.

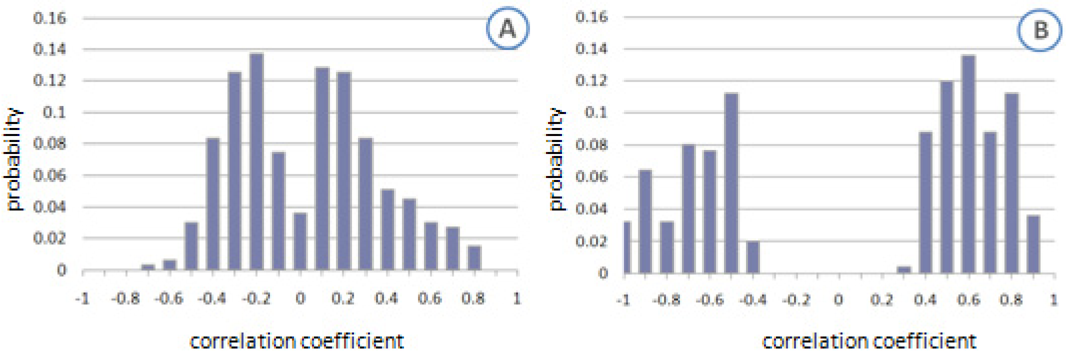


Figure 10: A. The distribution of correlation coefficients for genes in the GO category "Response to JA stimulus" was enriched with low correlations with much of the correlations around zero when a large functionally hetergeneous collection of microarrays was used. B. Using only functionally relevant datasets, there is an enrichment of higher correlation values. To account for the shorter vector length, all the correlation coefficients were filtered by a p-value threshold of 0.05.

From Fig.10B we can see that the distribution of the correlation coefficients is enriched with higher correlations. By visual inspection, we see that this distribution is significantly different from the one presented in Fig.10A where all experiments were used in the correlation. This was also confirmed by a t-test performed to test whether the two distributions were significantly different, which resulted in a very significant p-value of 9.8245e-15. It could reasonably be argued that the improvement in correlation observed when selected experiments are used is down to the smaller number of dimensions of data points to correlate. To account for this bias, we applied a p-value threshold ($p< 0.05$) and considered only those gene pairs that passed this threshold.

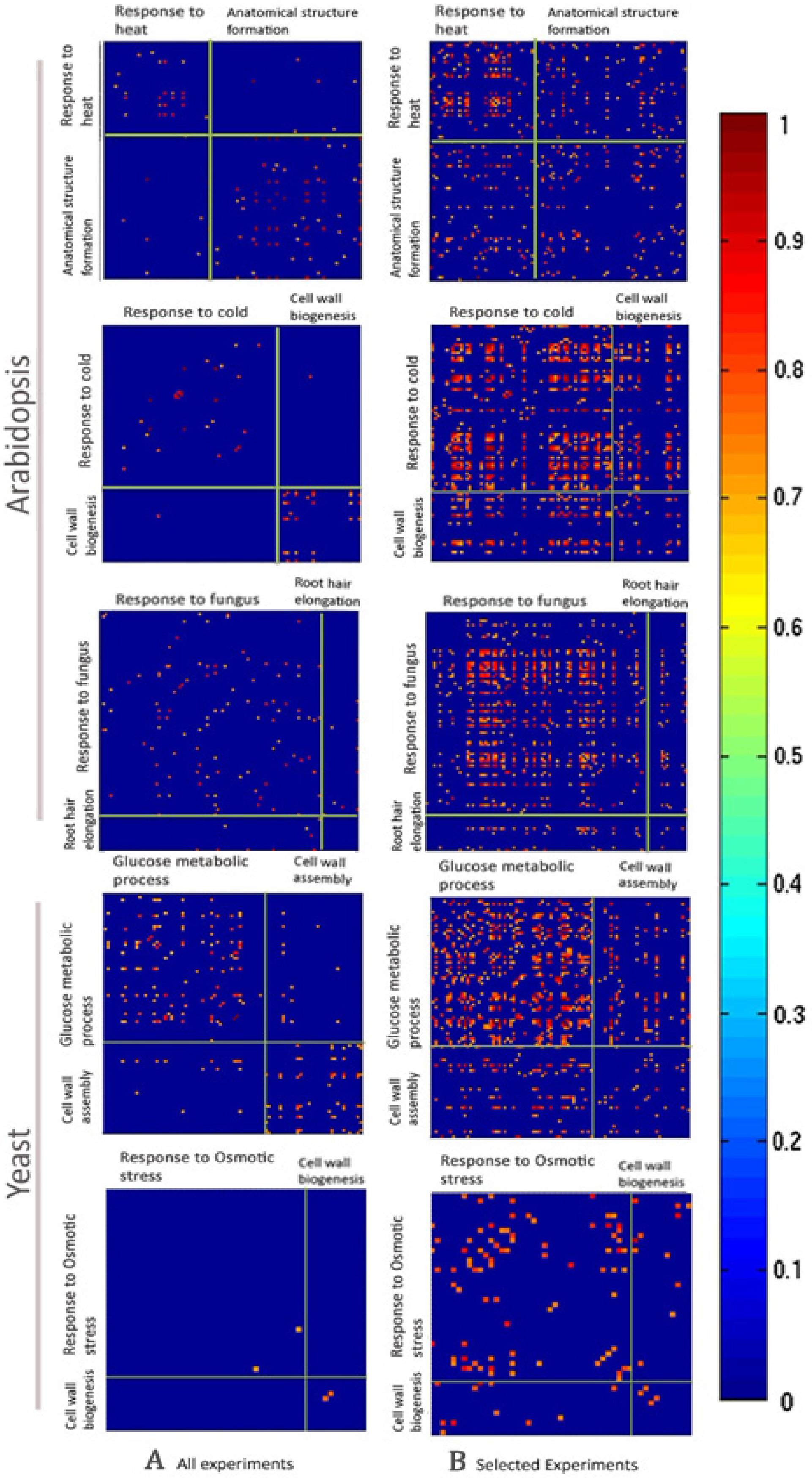
We believe that the improvement in correlation in the functional category of interest (as observed in Fig.10B) is due to the use of experiments functionally relevant to that category. To demonstrate that the improvement was not just due to shorter vector length in the chosen experiments compared to using all experiments, we paired GO categories that we deemed to be functionally dissimilar to each other e.g. Cell wall biogenesis - Response to osmotic stress, Root hair elongation- Response to fungus, and Cell wall assembly – Glucose metabolic process. By selecting experiments relevant to only one GO category in the pair, we can show that the improvement in correlation is relatively greater in the GO category for which the relevant experiments were selected. For each pair of functional categories, we retrieved genes belonging to the categories and calculated correlation using all the experiments in the compendium (The details of the experiments contained in the microarray compendium are reported in Section 5.3.1 of Chapter 5).

The correlation matrix for genes belonging to each GO category in the pair was visualized as a heatmap (Fig.11A). Subsequently, the correlation matrix was re-constructed using only those experiments that were functionally relevant to one of the GO categories in the pair (e.g. For the GO category pair "Cell wall biogenesis- Response to osmotic stress", we select experiments that are relevant only to Response to osmotic stress). The heatmap of the correlation matrix highlights the selective improvement in correlation among genes in the GO category pairs (Fig. 11B). The results from various GO category pairs selected from *A.thaliana* and *Yeast* are presented in Fig.11, where it can be observed that when all experiments were used for calculating the correlation, the correlation is generally lower among both the GO categories in the pair. However, in the second instance, where experiments relevant to one of the GO categories in the pair were used, although an increase in correlation was observed in both GO categories, there was a greater increase in the GO category of interest.

We quantified the change in correlation when biologically relevant experiments are used as compared to all the experiments in the dataset. The results are presented in Table 4. We computed the ratio of the averages of the absolute value of the correlations obtained between genes in the GO category of interest and the rest of the gene pairs in the test. In other words, considering the four quadrants in which the heat maps of the correlation matrices are divided (see Fig.11), we computed the ratio between the average of absolute values in the first quadrant and in the remaining three quadrants (termed 'background'). As expected, the ratio is much greater when biologically relevant experiments are used in the calculation.

Figure 11: Figure 11. Heatmap of the correlation matrix for pairs of Gene Ontology terms. A. Correlation calculated using all experiments in the microarray collection. B. Correlation calculated using experiments relevant to only one of the GO terms in the pair (marked in bold). In both Fig.11A

and Fig.11B, warmer colours indicate stronger correlation. The heatmaps are demarcated by green lines to indicate portions representing the two GO categories in the pair. Results for both Arabidopsis and Yeast are presented (Figure is presented in the following page)


Arabidopsis
Yeast
Response to heat
Anatomical structure formation
Response to cold
Cell wall biogenesis
Response to fungus
Root hair elongation
Glucose metabolic process
Cell wall assembly
Response to Osmotic stress
Cell wall biogenesis
A All experiments
B Selected Experiments
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

ARABIDOPSIS THALIANA						
	GO:ID	All Experiments		Selected Experiments		
		Ratio of correlation averages	t-test p-value	Ratio of correlation averages	t-test p-value	Experiments selected
1.	**Response to fungus** vs. Root hair elongation	4.147	0.5625	6.233	0.1443	Methyl Jasmonate treatment; Wounding; Psuedomonas infection
2.	**Response to Cold** vs. Cell wall biogenesis	2.2	0.1040	4.147	0.1270	Cold stress time series in shoots
3.	**Response to heat** vs. Anatomical structure formation	1.414	0.9986	2.639	6.62E-06	Heat stress time series in shoots
4.	**Response to salt stress** vs. Cell wall biogenesis	3.081	0.9938	3.797	0.2087	Osmotic stress in shoots; Salt stress in shoots;
YEAST						
1.	**Glucose metabolic process** vs. Cell wall assembly	2.057	0.0673	3.147	0.0481	Two glucose time series experiments
2.	**Response to osmotic stress** vs. Cell wall biogenesis	2.492	0.9091	5.143	0.7273	Hydrogen peroxide treatment
3.	**Aerobic Respiration** vs. Cell Death	3.459	0.7180	4.751	0.5042	Aerobic Phosphorus, Nitrogen, Sulphur time series; Aerobic anaerobic transition

Table 4: The ratio of average correlation among genes in the GO category of interest (in bold) and the genes in the other GO category in the pair is shown. A t-test was performed between the two sets of correlation and the

p-values are presented. Results are presented for both selected experiments and all experiments in the collection.

ARABIDOPSIS THALIANA							
	GO:ID	All Experiments			Selected Experiments		
	No. of positive correlations		No. of negative correlations	Average absolute correlation	No. of positive correlations	No. of negative correlations	Average absolute correlation
1.	Response to Abscisic acid stimulus	21	0	0.184	474	314	0.237
2.	Response to water deprivation	14	2	0.198	318	234	0.340
3.	Response to Jasmonic acid stimulus	14	0	0.214	57	22	0.256
4.	Response to Heat shock	12	0	0.200	73	24	0.244
5.	Response to Hyper-osmotic salt stress	0	0	0.177	6	3	0.227
YEAST							
1.	Antibiotic resistance	0	0	0.130	2	0	0.158
2.	Cellular glucose metabolic process	49	15	0.273	198	102	0.311
3.	Hexose metabolism	73	24	0.267	251	160	0.258
4.	Aerobic respiration	34	1	0.269	206	30	0.320

Table 5: Increase in the number of significant (|r| > 0.7) positive and negative correlations and the average correlation when selected relevant data are used.

As illustrated by Figure 11, the improvement in correlation is specific to the GO category for which the experiments are selected. To quantify this, we perform a one tailed *t*-test between the distribution of correlation obtained when selected experiments (experiments relevant to GO categories indicated in bold in Figure 11 and Table 4) are used and the distribution when all experiments are used. The *t*-test *p*-values (shown in Table 4) indicate that the distribution of correlation among genes in the GO category of interest is greater compared to the other GO category. Here a higher p-value means that when functionally diverse experiments are used the functional categories are less distinguishable from each other. It is interesting to note that in Table 4, although the p-values reported are not significant, there is a relative decrease in the p-values when the biologically relevant experiments are used.

We see that selecting experiments improves correlation generally in most GO Biological Process categories both in Yeast and Arabidopsis. To highlight the improvement in correlation, we compare it with the correlations obtained by using all experiments in the collection. We counted the number of gene pairs with a correlation of 0.7 and also for gene pairs with correlation below -0.7. Additionally, we also measured the average correlation among genes when selected experiments are used and when all experiments are used. The results are presented in Table. From the results it is clear that selecting experiments improves correlation among genes belonging to a functional category.

The difference in distributions found in Fig. 11A and Fig. 11B reflects the findings of (Adler et al., 2009) that acknowledged the pitfalls of using large microarray collections in co-expression analyses. Our results suggest that it is imperative to identify experiments which are relevant to the functional category of interest. However, manually identifying the relevant experiments may not be straight forward due to many factors. Firstly, in a typical compendium containing thousands of arrays it would be non-trivial to consider their literature individually. Secondly, literature knowledge is seldom exhaustive; the relevant experiment may not be obvious as biological process of interest may have been induced even in experiments where they are not the primary focus of investigation. These factors underline the need for a method to automate the selection of relevant experiments from a compendium of microarray experiments.

5

A novel method for selecting relevant experiments

5.1 State-of-the-art

In Section 4.4 of Chapter 4, we discussed how natural phenomena such as condition-specific gene expression can lead to poor correlation among genes, especially when large collections of microarray experiments are used. In correlation-based GBA analyses, several approaches were developed to account for the condition-specific nature of gene expression. Primarily, these approaches were based on the idea of biclustering. In the following section, we describe the concepts behind biclustering and outline some of the widely used biclustering algorithms.

5.1.1 Biclustering

In the functional analysis of gene expression data, the goal is to identify sets of genes or experimental conditions with similar expression profiles. Generally for this task, clustering techniques are applied that allow grouping of objects based on a selected feature. In case of gene expression data, genes are grouped based on correlation between gene expression profiles as the feature. Although classical single dimension clustering has been successfully applied for various functional analyses (Ben-Dor, Shamir, & Yakhini, 2004; D'haeseleer, 2005; Eisen, 1998), they suffer from two major drawbacks:

1. Single dimension clustering techniques such as *k*-means and hierarchical clustering (Tibshirani et al., 1999) and Self-organizing Maps (SOM) (Törönen, 1999) applied to gene expression data, find groups based on the *global* similarities between the expression profiles. When applied to large collections of microarray data, any similarity between the gene expression profiles in a subset of experimental conditions is lost (J. Wang, Delabie, Aasheim, Smeland, & Myklebost, 2002).
2. Single dimension clustering techniques do not allow for over-lapping clusters. However, as discussed in Chapter 2 and in Chapter 5, genes are co-expressed depending on the experimental condition. Therefore, it is reasonable to expect that genes can be in different clusters under different experimental conditions (Madeira & Oliveira, 2004).

Uncovering local patterns of expression similarities is considered vital to gene expression analyses, which led to the application of biclustering techniques in gene expression data analyses (Cheng & Church, 2000). In contrast to the single dimension clustering algorithms, biclustering algorithms are designed to cluster in two dimensions. Biclustering techniques when applied to gene expression data aims to cluster both genes and experimental conditions simultaneously. The concept of biclustering gene expression data was first proposed by Cheng and Church (2000).

The conceptual difference between clustering techniques and biclustering is illustrated in Fig.12. Single dimension clustering can be applied to group genes (Fig. 12A) or can be applied to group experimental conditions (Fig.12B). In both cases, the clusters are discrete with no overlap between the clusters.

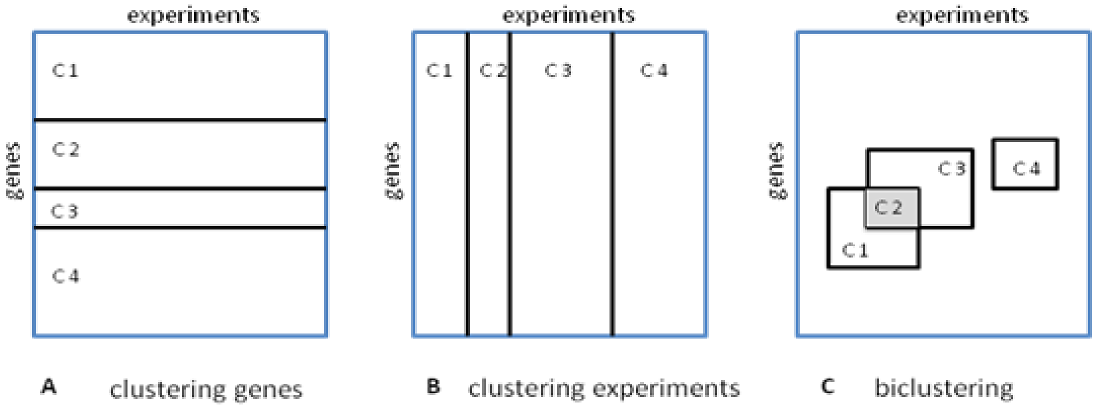


Figure 12: Illustration of the three basic approaches for clustering a matrix of gene expression profiles. A. Genes (rows) are grouped into clusters C1, C2, C3 and C4 based on similar expression in the experimental conditions. B. Experimental conditions (columns) are grouped into clusters C1, C2, C3 and C4 based on similar genes. C. Groups of genes and groups of experiments overlap to identify sub-groups where both genes and experimental conditions are similar. Clusters C1 and C3 overlap to form a sub-cluster represented by C2. C4 is a cluster with no overlaps, but is a discrete cluster of similar genes in similar experimental conditions.

Biclustering techniques aim to identify subgroups of experimental conditions where a subgroup of genes is co-expressed (Fig. 12C). Also, the clusters can overlap allowing for genes to associate with multiple clusters of conditions. Biclustering algorithms can find either one or multiple biclusters within a given gene expression dataset. Generally, biclustering techniques can handle multiple biclusters and allow for *apriori* definition of the number of clusters (A Califano, Stolovitzky, & Tu, 2000; Getz, Levine, & Domany, 2000; Ramanathan, 2001; Tanay, Sharan, & Shamir, 2002; Yu, 2003). Madeira & Oliveira(2004) provide an exhaustive review of biclustering techniques applicable to gene expression data.

5.2 Motivation for a novel experiment selection technique

In this work, the aim is to identify microarray experiments that are functionally relevant to a functional category of interest as genes in the functional category of interest are expected to be co-expressed in the experiments identified as relevant. Although the application of biclustering techniques in gene functional analyses was aimed at uncovering condition-specific co-expression, it is not designed to identify microarray experiments as a whole but only the experimental conditions that lead to a high clustering metric. Biclustering techniques when applied to microarray collections will identify clusters which may contain experimental conditions from different microarray experiments disregarding the integrity of the experiment. This is a limitation in biclustering experiments measuring time-courses where a bicluster containing a group of time points from different time-course experiments would be unreasonable. Further, biclustering is less effective on large microarray collections as the tendency to find local patterns due to noise also increases. Importantly, biclustering techniques are inapplicable in network-based analyses such as (Long et al., 2008; Mostafavi et al., 2008). It is important to note that although biclustering techniques have been applied with the aim of identifying condition-specific gene expression, it is distinct from our aim of identifying microarray experiments where genes belonging to a particular functional category are co-expressed.

In Section 4.3 of Chapter 4, we demonstrated that large functional heterogeneous collections of microarray datasets can lead to poor correlation among genes in the same functional category. In order to understand the causes of the poor correlation in the above mentioned scenario, we presented two hypotheses. Firstly, the poor correlation observed could be noise from the experimental technique. Secondly, we believe that the cause of the poor correlation lies in the nature of gene function itself. It is well known that gene expression and subsequently gene function are highly-condition dependent. Therefore, genes belonging to the same functional category may correlate only when the right biological conditions are considered.

In our investigations, we saw that when a large number of heterogeneous experiments are pooled together, the condition-specific nature of gene expression may have a significant influence on the overall correlation between the genes. We saw that, poor correlation between genes in functionally irrelevant conditions can rapidly dilute the overall correlation between the genes. Therefore, any functional relationship between the genes indicated by higher correlation would be masked. From our investigations, it was evident that choosing the right experiments was critical to the efficacy of GBA-based gene functional analyses.

However, identifying experiments relevant to a functional category is a non-trivial task. The obvious method would be to assess the relevance of a given microarray experiment using literature information. However, it is well known that literature knowledge is seldom exhaustive. Furthermore, the relevance of an experiment to a certain gene function or a biological process may not be immediately obvious. Experiments that are deemed irrelevant by an investigator could in fact withhold significant information regarding the biological process of interest as well as cross-talk between pathways. Also, the analysis of literature becomes increasingly impractical with the size of the microarray collection. These limitations provide a firm motivation for a computationally-driven method for identifying experiments relevant to a gene functional category or a pathway. In the following sections, we present a novel algorithm for systematically identifying experiments relevant to a functional category or a pathway. Importantly, this algorithm is able to identify relevant experiments not obvious by searching the literature on the experiments.

5.3 A novel algorithm for the selection of relevant experiments

5.3.1 Understanding functional relevance

Prior to the task of identifying experiments relevant to a functional category of interest, it is necessary to understand the concept of "relevance" in the context of correlation-based functional analyses. Let us consider the fundamental aim of a GBA analysis. Given a functional category of interest, the aim is to identify novel genes which may belong to that category. Therefore, it can be stated that relevant experiments are those which are "best" for identifying genes which may belong to that category. However, this definition serves as only an abstract description of "relevance".

For the task of developing an automated approach for identifying relevant experiments, one could simply consider relevant experiments as those where genes annotated to the functional category of interest are perturbed. It is reasonable to expect that in experiments where the functional category of interest has been perturbed, it is likely that the genes would show high correlation. However, simply searching for experiments where genes in the functional category of interest show high correlation is an inadequate approach. To understand this better, let us consider a correlation matrix constructed using data from an experiment of interest (Fig. 13).The area of the matrix denoted by A represents the correlation among genes annotated to the functional category of interest and B represents the correlation between genes annotated to all other functional categories and the genes annotated to the functional category of interest.

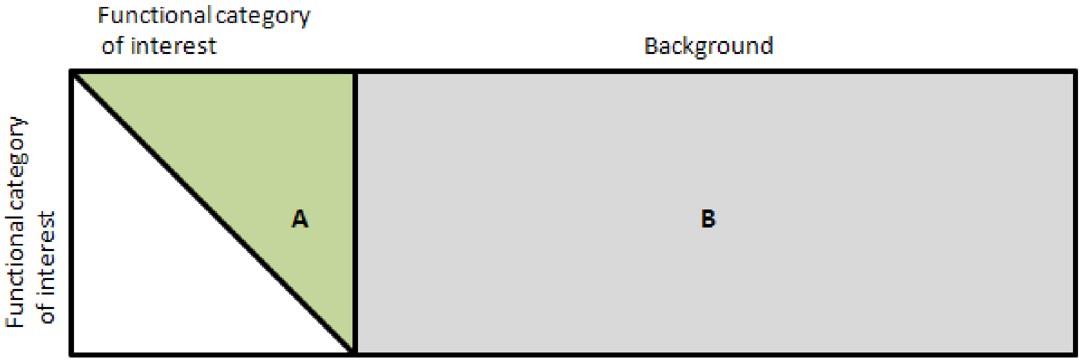

Functional category of interest
Background
Functional category of interest
A
B

Figure 13: Correlation matrix for functional category of interest and the background. The area *A* represents correlation among genes in category of interest and *B* represents correlation between functional category of interest and the background. Note: In the correlation matrix for the functional category of interest only the area above the diagonal (indicated in green) is considered as the correlation matrix symmetrical.

Let us consider an experiment where genes in the functional category of interest are perturbed, and A (Fig.14) contains highly correlated gene pairs.

However, in such an experiment it is highly likely several other biological functions were perturbed and as a result B would be populated by a significant number of highly correlated gene pairs. This could be due to cross-talk between various biological processes, processes that act in tandem with the functional category of interest or simply due to noise. In a GBA analysis where novel genes belonging to the functional category of interest are to be identified, the above outlined scenario would generate a large number of false positives i.e. many genes which belong to different functional categories may be falsely identified as belonging to the functional category of interest due to the existence of high correlation. Thus, selecting experiments simply based on the correlation between the genes in the functional category of interest would be unreliable. For this reason, prior to defining the "relevance" of an experiment to a functional category, we define a set of genes that we term as the *Background* set (as illustrated in Fig.14). For GO functional terms, the background set is defined as the genes annotated to all the GO terms except the GO term of interest and its children in the GO tree. The root of the GO tree is not considered part of the background set. Similar to the GO terms, for MIPS functional classification, all genes annotated to MIPS terms except to the MIPS term of interest is considered as the background. For KEGG pathways, the background set is simply all genes annotated to all the pathways described in the KEGG database except the pathway of interest. An experiment would be considered relevant if it can differentiate between the genes in the functional category of interest and the background set.

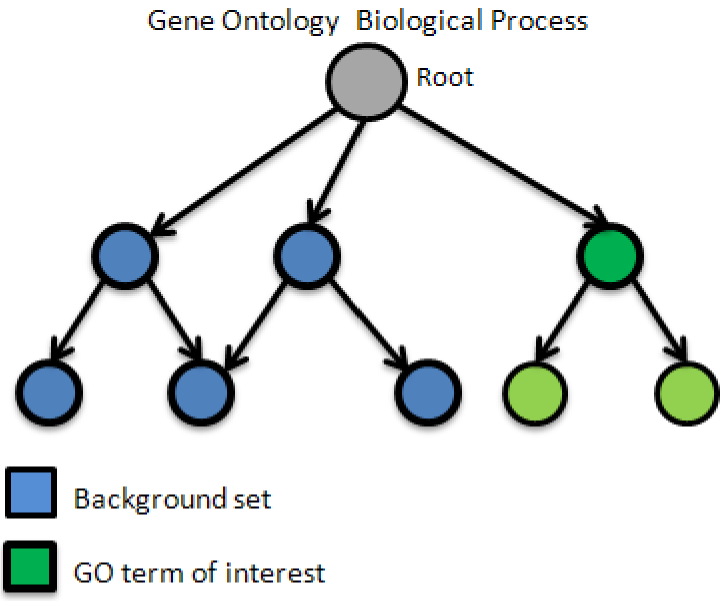
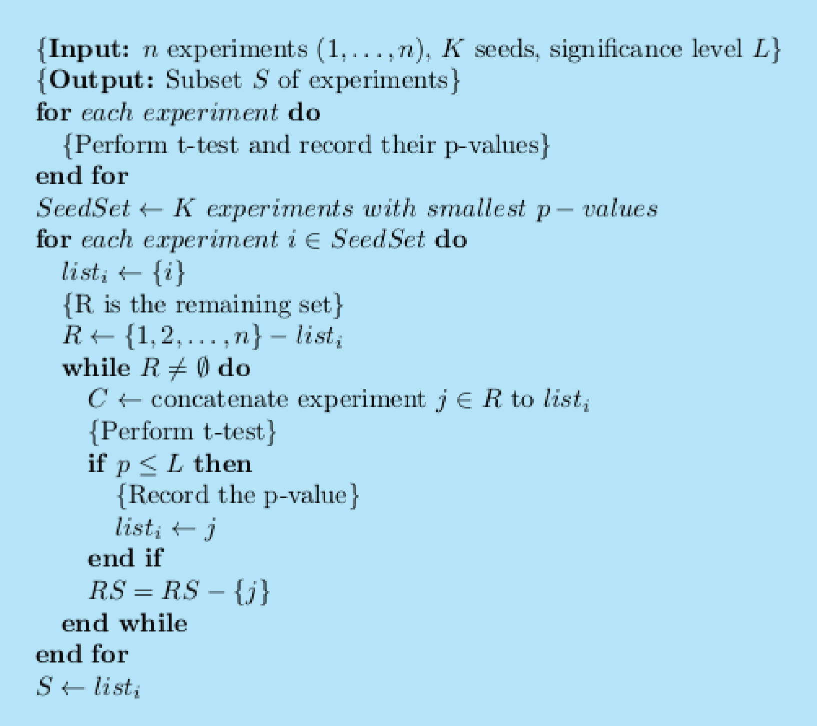


Figure 14: The contents of the background set with respect to the functional category of interest are illustrated in this figure. The dark green node depicts the chosen functional category of interest. The child nodes (coloured light green) are automatically considered to be part of the functional category of interest. The background set (in blue) is made of all the other nodes in the GO DAG that do not belong to the set containing the functional category of interest.

5.3.2 The experiment selection algorithm

Given a functional category of interest and a collection of microarray datasets, the task of the algorithm would be to select a subset of experiments that are optimal at differentiating between the genes in that functional category and those from the background. Our idea was to choose a feature that, if an experiment is relevant, would be able to differentiate the genes in the category of interest from those genes in the background set. The set of relevant experiments can then be found by maximizing the discriminative ability of the chosen feature. In other words, the set of relevant experiments will be the one for which the feature is best at discriminating the genes in the category of interests from the background ones. Since, we were primarily dealing with gene expression data, we chose the Pearson Correlation Coefficient (PCC) as the feature. We used the *t*-test to measure the discriminatory ability of the PCC in a given experiment. The *t*-test measures whether the distribution of correlation coefficient for the genes in the functional category of interest is higher than the correlation among genes in the background. The *t*-test is based on the null hypothesis that the two distributions of correlations are the same and it is assumed that the distributions are independent.

With typical microarray collections containing hundreds of experiments, clearly, an exhaustive search of the space of the possible subsets of the experiments is computationally intractable (the number of possible subsets of a set of *n* experiments is $2^{(n-1)}$). Thus a 'brute force' approach that analyzes every combination of experiments would not be feasible for typical microarray collections containing a large number of experiments – for example, for the set of 44 Arabidopsis microarray experiments which we present in the Results section (Section 5.4 – Section 5.8), this would require analysing over 8,000 billion combinations. Therefore we devised an efficient greedy heuristic which was able to select a set of experiments with high discriminatory ability while retaining a quadratic complexity. An informal description of the algorithm is presented below and the pseudo-code of the algorithm is presented in Fig.15. Our analysis assumes that we are given a certain functional category and a set of *n* microarray experiments, each comprising of several time-points or conditions. The procedure begins by performing a *t*-test for every experiment in the microarray collection assessing whether the distribution of the correlation among genes belonging to the label of interest (denoted by A in Fig.14) is significantly higher than the background (denoted by B in Fig. 14).



```
{Input: $n$ experiments $(1,\ldots,n)$, $K$ seeds, significance level $L$}
{Output: Subset $S$ of experiments}
for each experiment do
  {Perform t-test and record their p-values}
end for
$SeedSet \leftarrow K$ experiments with smallest $p-values$
for each experiment $i \in SeedSet$ do
  $list_i \leftarrow \{i\}$
  {R is the remaining set}
  $R \leftarrow \{1,2,\ldots,n\} - list_i$
  while $R \neq \emptyset$ do
    $C \leftarrow$ concatenate experiment $j \in R$ to $list_i$
    {Perform t-test}
    if $p \leq L$ then
      {Record the p-value}
      $list_i \leftarrow j$
    end if
    $RS = RS - \{j\}$
  end while
end for
$S \leftarrow list_i$
```

Figure 15: Pseudo-code describing the experiment selection algorithm. The *t*-tests are performed between the distributions of correlation among genes in the functional category of interest and the background set.

We then select a fixed number (K) of *seed* experiments with the best p-values from the *t*-tests. The algorithm builds experiment lists iteratively starting from these seed experiments. For a given list, at every iteration, an experiment is selected at random among those not already contained in the list and this experiment is tentatively added to the existing list. With the newly added experiment, the correlation matrix is reconstructed. A *t*-test is then performed to check whether this expanded list of experiments exhibits a distribution in the label of interest which is significantly higher than the background. The test is based on the assumption that the correlation among genes in the functional category of interest is independent of the correlation among genes in the background. If the p-value is smaller than a pre-defined threshold, the experiment is permanently added to the list; otherwise it is removed. This iterative procedure terminates when all experiments have been considered for every seed experiment for every list. Once the lists have all been created, the list with the overall final best p-value is kept as the optimal list of experiments that the algorithm returns.

Although this algorithm cannot guarantee that the selected set of experiments is optimal, in practice we found that this heuristic selected sets of experiments with high discriminatory ability while providing computational tractability. The number of t-tests our algorithm needs to consider at most is given by (Eq. 8):

$$n + K * [(n-1) + (n-2) + \cdots + 1] = K/2 * n * (n-1) + n = O(n^2) \qquad \text{(Eq. 8)}$$

This quadratic complexity allowed us to run all the experiments presented here in a few minutes on a regular desktop machine.

The algorithm has only two parameters: the significance level of the t-tests (denoted by L in the pseudo code) and the number of seed experiments (K). When testing our algorithm we set the significance level to the standard value of 0.05[3]. Importantly, we found that our algorithm is quite insensitive to the number of seed experiments – in the experiment presented below, in which we tested the procedure on different species and different sets of microarray experiments a value of K = 25 ± 15 gave similar results.

Compared to large collections of microarrays, smaller subsets of experiments may lead to higher correlation values purely because of the shorter length of the vectors. In all our analyses we account for this bias by filtering the

[3] We also tested the algorithm using a lower p-value threshold of 0.01 and obtained similar results.

correlation by a *p*-value threshold. This ensured that only statistically significant correlations are considered.

We tested our algorithm on publicly available microarray data collections. Here we present results obtained using 44 individual experiments in *Arabidopsis thaliana* from the NASCArrays collection and the M3D collection of 31 individual experiments in *Saccharomyces cerevisiae.* A full list and details of the microarray experiments can be found in Materials section in this chapter. Our experiments on both yeast and Arabidopsis prove that our procedure is also species-independent. To prove that our selection procedure is independent of the functional classification system adopted, we applied our algorithm for selecting experiments relevant to both GO Biological Process terms and MIPS FunCat terms.

In the following sections we will prove the effectiveness of the algorithm by showing that the selected set of experiments:

A. Result in higher correlations between genes in the same functional category (Section 5.4).
B. Improve the performance of a GBA-based classifier (Section 5.5).
C. Provide a discriminatory ability for a given functional category which increases with the specificity of the annotation (Section 5.6).
D. Lead to a better reconstruction of gene regulatory pathways (Section 5.7).

5.3.3 Materials

For *Arabidopsis thaliana*, our microarray data collection consisted of 756 Affymetrix ATH1-501 arrays from 44 experiments. The microarrays were sourced from NASCArrays (Craigon et al., 2004) Pathogen Series, Developmental Series, Stress series and Chemical and Hormone treatment series.

Serial. No	NASCARRAY ID	Experiment
1	152	Developmental series
2	137	Control (Shoot)
3	137	Control (Root)
4	138	Cold (Shoot)
5	138	Cold (Root)
6	139	Osmotic Stress (Shoot)
7	139	Osmotic Stress (Root)
8	140	Salt stress (Shoot)
9	140	Salt stress (Root)
10	141	Drought stress (Shoot)
11	141	Drought stress (Root)
12	142	Genotoxic stress (Shoot)
13	142	Genotoxic stress (Root)
14	143	Oxidative stress (Shoot)
15	143	Oxidative stress (root)
16	144	UV-B stress (Shoot)
17	144	UV-B stress (Root)
18	145	Wounding Stress (Shoot)
19	145	Wounding Stress (Root)
20	146	Heat Stress (Shoot)
21	146	Heat Stress (root)
22	120	Response to virulent, avirulent bacteria
23	122	Response to bacterial-(LPS, HrpZ, Flg22) and oomycete-(NPP1) derived elicitors
24	123	Response to Phytophthorainfestans
25	167	Response to Botrytis cinerea infection
26	168	Pseudomonas half leaf injection
27	169	Response to Erysipheorontii infection
28	172	ACC time course in wildtype seedlings
29	173	Zeatin time course in wildtype seedlings
30	174	Methyl Jasmonate time course in wildtype
31	175	IAA time course in wildtype seedlings
32	176	ABA time course in wildtype seedlings
33	179	Effect of brassinosteroids in seedlings
34	181	Cytokinin treatment of seedlings
35	183	Effect of ABA during seed imbibition
36	184	Basic hormone treatment of seeds
37	185	Effect of gibberellic acid inhibitors on seedlings
38	186	Effect of auxin inhibitors on seedlings
39	187	Effect of brassinosteroid inhibitors on seedlings

40	188	Effect of ethylene inhibitors on seedlings
41	189	Effect of cycloheximide on seedlings
42	190	Effect of proteasome inhibitor MG13 on seedlings
43	191	Effect of photosynthesis inhibitor PNO8 on seedlings
44	192	Effect of ibuprofen, salicylic acid and daminozide on seedlings

Table 6: The complete list of microarray experiments used in the Arabidopsis microarray compendium used for selecting functionally relevant experiments.

Raw data was downloaded, pre-processed and normalized by MAS 5.0 using R Bioconductor packages (Gentleman et al., 2004) as outlined in Appendix I. All the data used were from experiments based on Wild-type plants only. Experiments conducted on multiple organs such as Roots and Shoots were considered as separate experiments. The experiments used to construct the Arabidopsis microarray compendium are presented in Table 6 above.

For Yeast, the microarray collection consisted of 537 Affymetrix microarrays from 31 individual experiments. The data was downloaded from the Many Microbes Database (Faith et al., 2007) and consists of a mix of wild-type and mutant-based experiments under various stresses, growth, chemical and hormone treatments. The details of the individual experiments in the Yeast collection can be accessed at the web address: http://m3d.bu.edu/cgi-bin/web/array/index.pl?section=home.

Throughout our analysis, GO Biological Process annotations with only non-electronic evidence codes were considered; EXP, IDA, IPI, IMP, IGI, IEP, ISS, IC, ISO, ISA, ISM and IGC. Further, only 'is_a' and 'part_of' relationships were considered. This is because in the current framework of the algorithm, genes can only belong to a subset or a superset of a functional category. Other relationships such as "negatively_regulates" or "positively_regulates" do not necessarily satisfy this requirement. However, we hope to incorporate other such GO relationships in the future. While considering genes belonging to a GO term of interest, all genes belonging to the child terms were also included. Genes which belonged to all the other GO terms in the tree were considered as the *Background.* While selecting the GO terms for analyses, only terms with at least 25 genes were considered. This was done to afford sufficient number of genes for a statistically significant *t*-test and cross-validation. Similarly, for MIPS FunCat, terms of interest included all child terms and the Background included all the remaining terms. Pathways and genes annotated to the pathways were obtained from KEGG (Kanehisa, 2000).

5.4 Selected experiments improve overall correlation in the functional category

Earlier in Chapter 1 and Chapter 4, we discussed that, for effective GBA-based analyses, it is essential that genes belonging to the same functional category exhibit high correlation. However, we saw that this is not necessarily true when large microarray collections are used for calculating the correlation. Nevertheless, for a given functional category, we observed that the experiments selected by our algorithm uncover significantly higher correlation. For genes which belong to the same functional category, we compared the distribution of correlation coefficients obtained from experiments selected by the algorithm with the distribution obtained from using all experiments in the collection. Fig. 16 shows representative histograms of the distributions for both Arabidopsis and yeast GO terms. As expected, the distribution of correlation coefficients obtained from using all experiments is enriched with low correlations with a relatively low number of high correlation values. However, when experiments selected by the algorithm are used, the distribution is enriched with higher positive and negative correlations. We perform a *t*-test between the two distributions to test whether the distribution obtained by using selected experiments is greater than when all experiments are used. The *t*-test was performed using the absolute values in the two distributions. The low *t*-test *p*-values (indicated in Fig. 16) confirm that the distribution obtained with selected experiments is significantly higher than when all experiments are used.

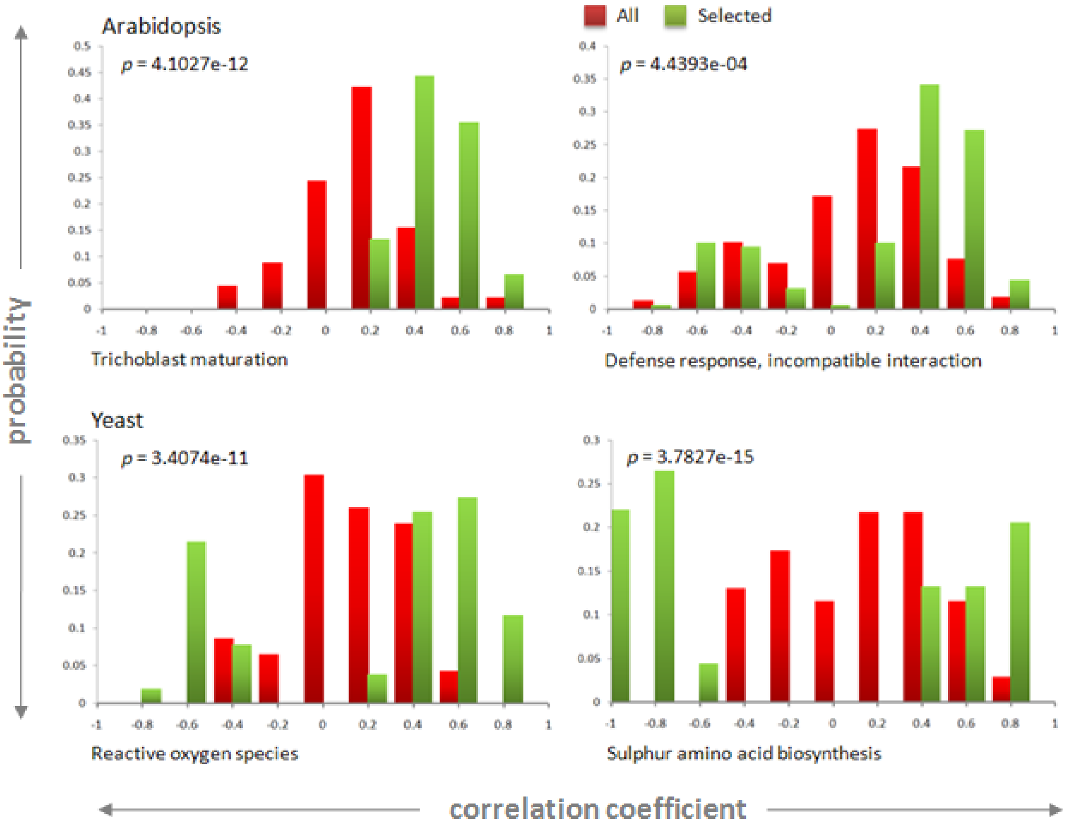


Figure 16: Comparison of distribution of correlation coefficients among genes in GO categories. It can be clearly observed that the distribution is enriched with higher correlations when experiments selected by our algorithm are used. The t-test p-values indicate that there is significant

difference between the distributions when selected and all experiments are used.

5.5 Quantifying the effectiveness of the selected experiments for function prediction

In order to evaluate the effectiveness of the experiments selected by the algorithm, we formulated a classification problem in Machine Learning. In this classification problem, genes are classified as belonging to a functional category of interest using Pearson Correlation between gene expression profiles as the feature[4]. In other words, a gene would be classified as belonging to the functional category of interest if it shares high correlation with the existing genes in the functional category. It is reasonable to state that the performance of a classifier depends on the quality of the input data i.e. if good quality data is provided the performance of the classifier should be better than if the input data is of poor quality. Therefore, evaluating the performance of the classifier allows us to evaluate the quality of the input data. In our case, we assess the performance of the classifier when the algorithm selected set of experiments is used as input data, compared to when all experiments in the microarray compendium are used as the input. Formulating the evaluation of the selected experiments as a classification problem allows us to use standard machine learning tools for evaluating classifier performance. The strategies adopted to measure and evaluate the performance of the machine learning classifier are detailed in the following sections. Two important tools that we have adopted for measuring and

[4] In Machine Learning, a classification is defined as an algorithmic procedure which assigns the input data into one or more categories (Bishop, 2007). The algorithm or the abstract *machine* which performs the classification is simply known as a Classifier.

evaluating the performance of our machine learning classifier are: Training and testing and Receiver Operating Characteristic (ROC) Curves. We have detailed these techniques in the following sections.

5.5.1 Training and Testing

A classifier learns the predictive relationships in the dataset, often called the *training set.* The performance of the classifier on the dataset where it "trained" is expected to be exaggerated. Hence, to evaluate the performance of a classifier, it is important to assess its predictive performance on new data that had no role in training the classifier. Such an independent dataset is called the *test set*. A central assumption is that both the training and the test set are representative of the data to be classified.

In the ideal case, where a large amount of data is available, a large sample is used for training. Another large sample is drawn which would then be used for testing. This would provide a reliable estimate of the performance of the classifier. However, in real cases such as the one we consider, the amount of data to be classified is small. In our case, for classifying a gene into a functional category, the total number of genes already annotated to that category (that can be used for training or testing) can be as low as 20. For this reason, a *holdout* procedure is preferred where a portion of the data available is held-out; the held-out portion of the dataset is used as the test set and the remaining portion is used as the training set. However, it is possible that the held-out portion or the data used as training set may not be representative of the dataset. Generally, it is difficult to verify whether every instance of the data is truly representative of the dataset. To counter this problem a *cross-validation* technique is adopted. In cross-validation, the user decides on a fixed number of portions or *folds* to divide the dataset. For example, if the user decides on a *ten-fold cross-validation*, the data is divided into ten approximately equal portions and each in turn is used as a testing set and the remaining portions are used for training. Therefore, in each turn 9/10th of the dataset is used for training and 1/10th is used for testing. The training and testing routine is repeated 10 times so that each portion of the dataset has served as a training set and a testing set. The performance of the classifier is recorded for each of the ten turns and it is averaged to obtain an overall performance figure. In our experiments, suppose a functional category of interest has 40 genes annotated to it, the 40 genes would be divided into 10 portions of 4 genes each. 36 genes would be used for training and 4 genes would be used for testing. This would be repeated 10 times and the average performance over the 10 repeats is considered as an estimate of the classifier performance.

Although the ten-fold cross validation procedure is often used for performance evaluation, the *leave-one-outcross-validation* (LOO cross validation) provides a more stringent estimate of the performance of the classifier. LOO cross-validation is simply an *n*-fold cross validation, where *n* is the number of instances in the dataset. In this approach, each instance present in the dataset is held out and used as the test set while the remaining instances are used for training. The performance of the classifier is averaged over the *n*-repetitions and an overall performance is obtained. LOO cross validation is stringent for two reasons (Bishop, 2007). Firstly, since every instance is considered, there is no random partitioning of the dataset. Secondly, this method uses the largest number of instances possible for training. It is presumed that this will increase the accuracy of the classifier. In our experiments, let us suppose the functional category of interest has 40 genes annotated to the category. 39 genes would be used for training and 1 gene for testing. The training and testing routine would be repeated 40 times.

5.5.2 Measuring the performance of the classifier using ROC curves

Let us consider a two-class classification problem such as the one in our experiment. Here, a gene can either be classified as belonging to the functional category of interest or as not belonging to the functional category of interest. Simply counting the number of wrong classifications will give us an *error rate* for the classifier. Lower the error rate, better the performance of the classifier.

Often, the "cost" of making a right classification is not equal to the cost of making a wrong classification. For example, let us consider a bank which has to make a decision on loan applications. The classifier employed by the bank has to classify the loan applications into two classes namely "defaulter" (a person unlikely to pay back the loan) and "non-defaulter" (a person very likely to pay back the loan). Here, the cost of paying out a loan to a defaulter could be far greater than losing business with a non-defaulter. However, evaluation of the classifier based on the error rate or the classification accuracy alone will deem both the cases as having equal costs. Therefore, a more comprehensive analysis of the cost of making a wrong classification is required.

Actual class	Predicted class		
		yes	No
	Yes	True positive	False negative
	No	False positive	True negative

Table 7: Confusion matrix for recording the performance of a classifier

In a classification exercise such as our experiment, a gene either belongs to a class or it does not. In such a two-class case, there are four possible outcomes in every classification. A gene can be correctly identified as belonging to the class of interest (true positive) or can be correctly identified as not belonging to the class (true negative). Also a gene can be wrongly classified as belonging to the class (false positive) or can be wrongly identified as not belonging to the class of interest (false negative). The possible outcomes are summarized in a table called the *confusion matrix* (Table 7). An ideal classification will have large numbers in the true positives and true negatives and smaller numbers in the false positive and false negative.

The confusion matrix summarizes the cost of a classification based on a single threshold. In our classification experiment, this would translate to classifying the genes based on a single threshold of correlation between the gene expression profiles. Often it is not trivial to decide on a single optimal threshold. Hence it could be useful to analyze the performance of the classifier as the threshold is varied.

Receiver Operating Characteristic (ROC) curves are a graphical way of representing the performance of a classifier and the costs involved as the threshold is varied. In our experiments, ROC curves illustrate the trade-off between a true positive and a false positive as the correlation threshold is varied in the classifier. Each point on the ROC curve is given by the true positive rate (TPR) and the false positive rate (FPR).

The TPR is given by the formula:

$$\boldsymbol{TPR = \frac{TP}{TP + FN}}$$

The FPR is given by the formula:

$$\boldsymbol{FPR = \frac{FP}{FP + TN}}$$

Where:

TP is true positive, TN is true negative, FP is false positive and FN is false negative

A sample ROC curve has been presented in Fig.17. The blue line represents the performance of the classifier if the instances in the data were classified randomly. The orange line represents the actual performance of the classifier. In the ROC curve for a given classifier, the top left corner of the plot represents the preferred case, where there are a large number of true positives and a minimal amount of false positives. The ROC curve can be quantified by calculating the area under the curve (AUC). The AUC serves as a quantification of the performance of the classifier under a range of thresholds. In our experiments, we calculate the area above the curve (1-AUC). Smaller the 1-AUC, better the performance of the classifier.

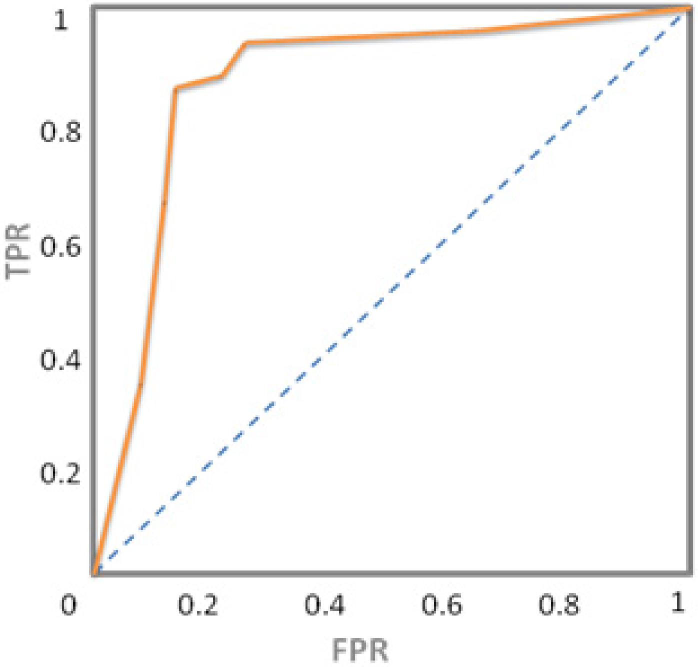


Figure 17: An example of a ROC curve (indicated in orange). The blue dotted line indicates the performance of the classifier if the input data is classified randomly.

5.5.3 Results from the evaluation and measurement of classifier performance

In our classification problem, the classifier's task is to use correlation to distinguish between pairs of genes which both belong to the category of interest and pairs of genes in which only one does. The rationale here is that a GBA analysis would be more effective if higher correlation values were obtained for gene pairs in which both genes belong to the same functional category.

For calculating the ROC curves, gene pairs in which both genes belong to the functional category of interest were considered the Positive Set; and gene pairs, in which only one gene belongs to the functional category of interest, were considered the Negative Set. To evaluate the performance of the selected experiments, we performed a ten-fold cross validation (as detailed in Section 5.5.1). Fig.18A shows four ROC curves which were calculated in this way. We can see that the ROC curves for selected experiments (shown in green) have a greater AUC compared to all experiments (shown in red). Following common practice, we also present the average (1-AUC) for both selected and non-selected datasets over the ten-folds (Fig. 18B). The average (1-AUC) is remarkably lower for the selected set of experiments. The superior performance of the selected set of experiments was observed for both Arabidopsis and Yeast GO Biological Process terms (Fig. 18B) and MIPS FunCat terms (Fig. 18C). For many examples of MIPS FunCat terms, we found that the difference in performance between the selected set and all experiments was lower compared to GO Biological Process terms. We believe that this could be due to the broad functional classification found in MIPS FunCat when compared to the more advanced GO (We discuss the relation between specificity of annotation and performance of the selected set of experiments in the next section.)

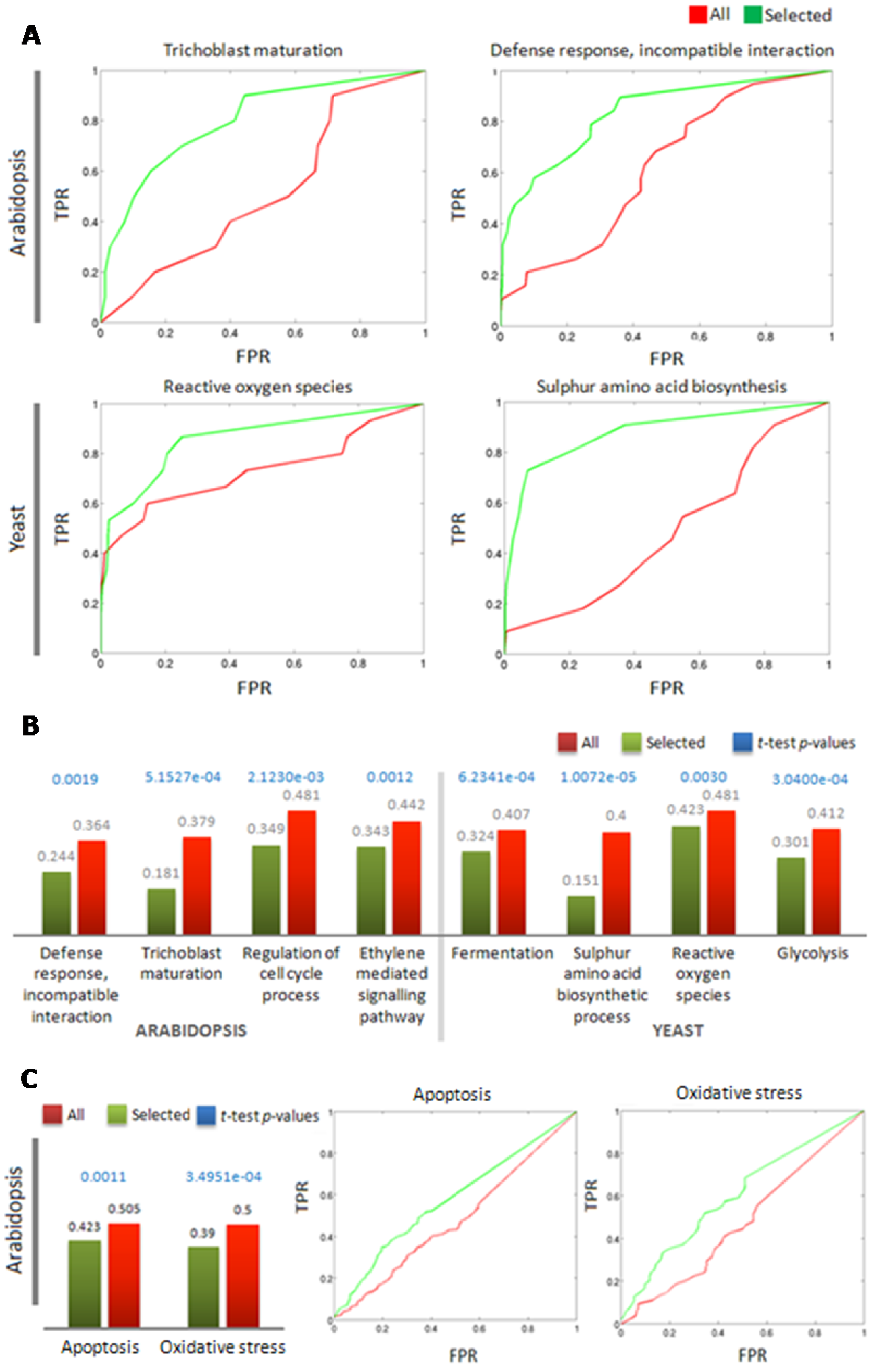


Figure 18: Evaluation of performance of the datasets selected by the algorithm compared to no selection (i.e. all experiments). Experiments were selected for both GO and MIPS FunCat terms. A. The ROC curves show a large improvement in classifier performance when the algorithm selected set of experiments is used. B. Average of the 1-AUC scores from each of the ten-fold cross validation show that the selected set of experiments are reliably better than using all experiments. The *p*-values from the t-test between the both the sets of 1-AUC show that the scores for the selected set are significantly better. C. Similar results can be observed for MIPS FunCat terms as well.

5.6 Effectiveness of the selected experiments increases with annotation specificity

The experiment selection algorithm is based on the idea that a functionally relevant set of experiments should be able to effectively differentiate the genes in the functional category of interest from all the other functional categories. Therefore, we expect the performance of the selected set of experiments to improve with the specificity of the functional annotation. Gene expression in a specific biological process could be expected to be relatively correlated compared to a broader category which includes several biological processes. As a result, it would be harder to differentiate a broad functional category from all the other processes. This gives us another way to prove the effectiveness of our experiment selection procedure where we show that the performance of the selected experiments at our classification task is higher as the functional category becomes more specific. For the GO BP term GO:0009861 "Jasmonic acid and ethylene dependent systemic resistance", at every level of the tree leading up to the root term, we selected experiments using our algorithm.

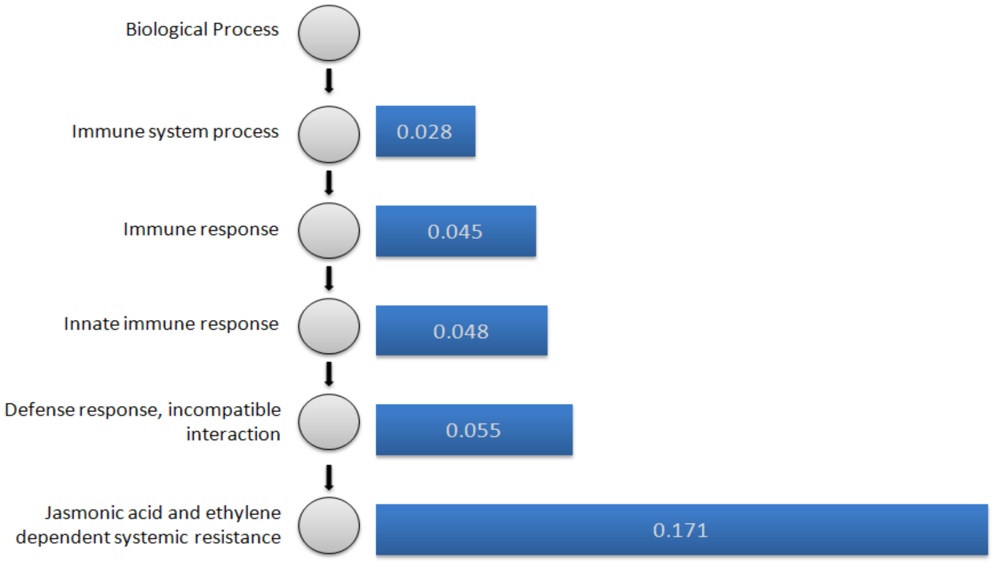


Figure 19: Increase in difference in 1-AUC between the selected set and all experiments when the specificity of annotation is increased.

To evaluate the performance of the selections, we constructed ROC curves for both selected and all experiments as described earlier (data not shown). If the performance of the selected experiments is no different to using all experiments, the difference in their 1-AUC was expected to be zero. Fig.19 reports the difference in 1-AUC between selected and all experiments. As expected, the specific annotation shows the largest gain in performance when selected experiments are used. The performance difference between the selected experiments and all experiments decreases as the functional classification gets broader. We have obtained similar results for several GO categories for both Arabidopsis and Yeast (data not shown).

5.7 Selecting relevant experiments: implications on pathway reconstruction

Reconstructing and modelling gene regulatory pathways using high-throughput data is a challenging problem in post-genomic biology. The GBA principle has been successfully applied to identify putative members of partially characterized pathways and gene networks using transcriptional data (Basso et al., 2005; Soinov, Krestyaninova, & Brazma, 2003). Identifying experiments relevant to the pathway of interest can be crucial for pathway reconstruction where the objective is to identify potential members of the pathway. The same idea we applied to select relevant experiments to functional categories can also be used to select experiments relevant to pathways. Here, all other pathways except the pathway of interest can be considered as the background and the set of relevant experiments are the ones which can best discriminate the pathway of interest from the background. We believe that the relevant experiments can uncover greater correlation between the genes in the pathway of interest and is a better predictor of potential membership of a gene in the pathway of interest. To demonstrate this, we hypothesized that a potential candidate would show greater correlation to the pathway of interest compared to all other pathways when only relevant experiments are used to calculate the correlation.

To verify this we obtained the “Alpha linolenic acid metabolic pathway” (ID: ath00592) from KEGG. Alpha linolenic acid is a precursor of a class of fatty acid derived regulators called Jasmonates. The main biosynthetic derivative of alpha linolenic acid is Jasmonic acid (JA). In plants, JA is known to be an important mediator of the defence response and other stress related signalling pathways (Avanci, Luche, Goldman, & Goldman, 2010; Balbi

&Devoto, 2008). The KEGG annotation of the pathway in *A.thaliana* consists of 30 genes of which 26 were found in our microarray collection. To demonstrate that the correlation obtained from the selected set of experiments is a better predictor of pathway membership, we framed this as a classification problem between two classes of genes: those in the pathway and those in the background. This classification is performed using very simple GBA-inspired classifiers that use only the Pearson correlation between the genes. The simplest possible classifier of this kind is one that classifies a gene using the sum of the correlations between that gene and the genes in the training set that belong to the category of interest: if this sum is above a certain threshold, it classifies the gene as belonging to the category of interest; otherwise it assigns it to the background.

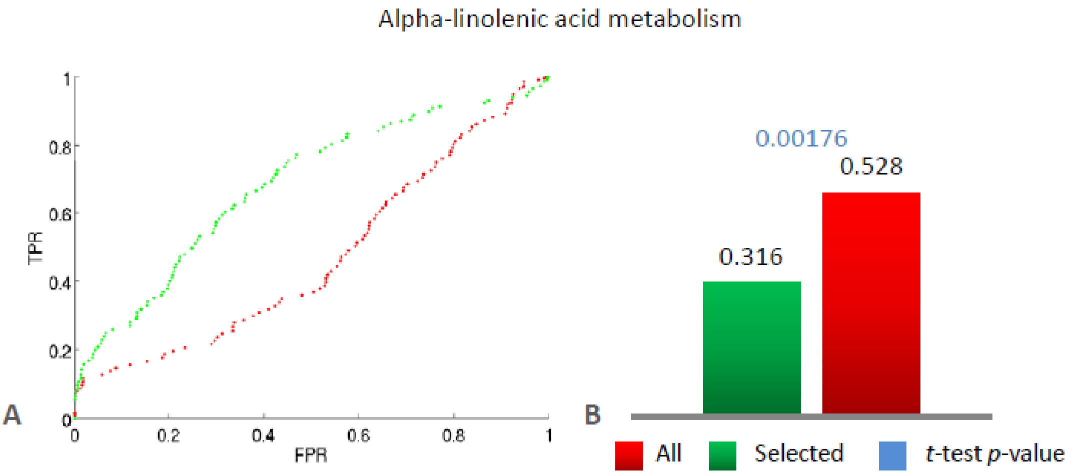


Figure 20: ROC curve analysis for genes in the "Alpha linolenic acid metabolic pathway" (KEGG ID: ath00592) from the KEGG Pathway Database. (A) Average ROC curves from the ten-fold cross-validation show the performance of the GBA-based classifier for predicting genes belonging to "Alpha-linolenic acid metabolism" pathway. (B) average (1-AUC) scores from 10-fold cross validation. The *p*-value for the t-test between the ten (1-AUC) values from the ten-fold cross-validation

obtained using the selected experiments and those obtained using all experiments is also reported (shown in blue).

As before, the performance of the classifier was evaluated by 10-fold cross-validation and average ROC curves were calculated over the ten folds. We compared the performance of the classifier when using correlations from the selected experiments and all experiments in the collection. The average ROC curves (Figure 20) and the average (1-AUC) bar plots (Figure 20) clearly show that the classifier using correlations from the selected set outperform the classifier using correlations from all experiments in the collection. This result clearly highlights the potential of the experiment selection algorithm in pathway modelling and reconstruction approaches. Similar results have been obtained for several pathways in Arabidopsis and Yeast (data not shown). This result demonstrates the potential of the experiment selection algorithm in pathway reconstruction.

5.8 Selected experiments generally agree with literature-based knowledge

Reassuringly, we found that the majority of experiments selected as relevant by our algorithm reflected the biological background of the functional category of interest. Some of the GO Biological Process terms and corresponding experiments selected as relevant by our algorithm are listed in (Table 8). For example, in the results obtained for Arabidopsis, the selection of experiments for "Ethylene mediated signalling pathway" seems relevant as ethylene is a well-studied mediator of osmotic stress and salt related responses (M. Fujita et al., 2006). Also, ethylene along with hormones such as abscisic acid has been shown to control many of the drought-related responses (Wilkinson & Davies, 2010). For growth-related terms such as

“Regulation of cell cycle process” and “Trichoblast maturation”, growth-related experiments such as the Weigel developmental stages experiments were selected. Similarly, experiments identified as relevant to plant defence were found to contain stress-related and pathogen infection-related experiments. For the GO term “Root epidermal cell differentiation” (data not shown), experiments related to abscisic acid treatment and ethylene treatment were selected. These selections are reasonable as studies such as (van Hengel, Barber, & Roberts, 2004) have demonstrated the role of abscisic acid, along with hormones such as ethylene in regulating epidermal cell-specific gene expression in *Arabidopsis thaliana* roots. Also it is interesting to note that, in general, the experiments were selected for a given GO term are plant organ-specific with no mixed sets of experiments such as from roots and shoots. This is particularly interesting as in a compendium of Arabidopsis abiotic stress experiments it has been observed that distinct clusters are formed for different organ types (Simon Barak, pers. comm.).

Similarly, yeast experiment selections also generally reflect the functional backgrounds of the GO terms such as in the case of “Response to reactive oxygen species”. Here, the experiment Hydrogen Peroxide treatment is relevant as hydrogen peroxide is widely used to mimic reactive oxygen species (Apel & Hirt, 2004).

A minority of experiments selected by the algorithm seemed to be unrelated to the GO terms of interest. We found this reasonable as the Biological Process of interest could also be activated in experiments originally designed to study a seemingly unrelated phenomena i.e. it could be a (contd...)

<table>
<tr><th colspan="4">Arabidopsis thaliana</th></tr>
<tr><th>Sl. No.</th><th>GO Identifier</th><th>Description</th><th>Selected experiments</th></tr>
<tr><td>1</td><td>GO009873</td><td>Ethylene mediated signalling pathway</td><td>Osmotic stress (shoot) TS
Cold stress (shoot) TS
Oxidative stress (shoot) TS
Salt stress (shoot) TS
Drought stress (shoot) TS</td></tr>
<tr><td>2</td><td>GO009817</td><td>Defense response to fungus, incompatible interaction</td><td>Developmental series (Flowers and Pollen)
Virulent and avirulent bacterial infection (leaf)
Gibberelic acid treatment (seed) TS
Hormone treatment (seed) TS
Cold stress (shoot) TS
Cytokinin treatment (seed) TS
ABA treatment (seed) TS</td></tr>
<tr><td>3</td><td>GO010564</td><td>Regulation of cell cycle process</td><td>Developmental series (Flowers and Pollen)
Bacterial, elicitor treatment (leaf)
Osmotic stress (shoot) TS
No treatment (shoot) TS
Methyl jasmonate (shoot) TS</td></tr>
<tr><td>4</td><td>GO048764</td><td>Trichoblast maturation</td><td>Ibuprofen, Salicylic acid treatment (seed) TS
Heat stress (shoot) TS
Brassinosteroid treatment (seed) TS
P.infestans infection (leaf) TS
Cold stress (shoot) TS
GA treatment (seed) TS
Osmotic stress (root) TS
UV-B stress (shoot) TS
B.cinerea infection (leaf) TS</td></tr>
<tr><th colspan="4">Yeast</th></tr>
<tr><td>5</td><td>GO000097</td><td>Sulphur amino acid biosynthesis</td><td>Carbon limitation TS
Antibiotic treatment TS
Aging TS</td></tr>
<tr><td>6</td><td>GO000302</td><td>Response to reactive oxygen species</td><td>Thiolutin s288c treatment
Antibiotic (Doxycycline) treatment
Hydrogen Peroxide treatment TS
Histone deacytelase mutant</td></tr>
<tr><td>7</td><td>GO006096</td><td>Glycolysis</td><td>UV and IR treatment TS
Thiolutin upf1
Mannose treatment TS
Hydrogen peroxide treatment TS
Antibiotic (doxycycline) TS
Antibiotic treatment TS
Histone deacytelase mutant
Carbon limitation TS
Dough fermentation TS
Aerobic respiration, Glucose P, N and S
Mannose</td></tr>
<tr><td>8</td><td>GO006113</td><td>Fermentation</td><td>Hydrogen peroxide treatment TS
Aerobic respiration, Glucose Phosphorus, Nitrogen and Sulphur utilization TS
Aging TS
Carbon limitation TS</td></tr>
</table>

Table 8: GO Biological Process terms and corresponding selected set of relevant experiments for *Arabidopsis thaliana* and Yeast. TS indicates time series.

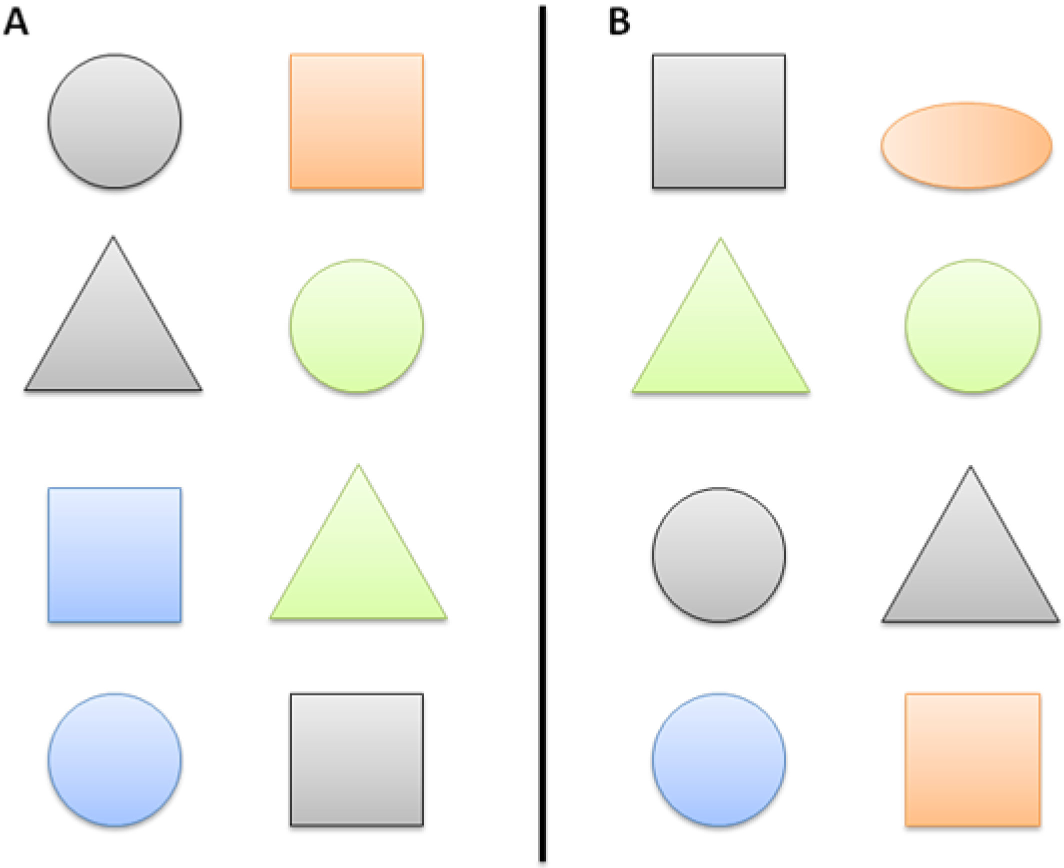


Figure 21: A collection of objects with multiple features such as colour and the number of edges. In a problem where circles have to be identified from the collection of objects, the feature that is useful to make the decision depends on the other members in the collection. A. Here, lack of edges is a useful feature for identifying the circles. B. Here, the lack of edges is no more a useful feature to identify the circles.

limitation of current literature. Importantly, viewing the experiment selection procedure as a classification problem provides an insight into the role of the seemingly irrelevant experiments in the selected set discussed in the previous section.

It is possible that such experiments may not have any biological relevance to the functional category of interest. However, they may be effective discriminators of the functional category of interest from the background. To understand this, let us consider a classification problem, where circles have to be identified from a collection of objects containing cubes, pyramids and circles (Fig.21). Several features exist that can effectively describe the circles in the collection e.g. colour, lack of edges etc. Let us consider the collection of objects presented in Fig. 21A. Here, a feature such as "colour", although a feature of the circle, will not be able to discriminate the circle from the other objects. Instead, the feature "lack of edges" is an effective discriminator of the circles in the collection. Now let us compare this with the collection of objects presented in Fig. 21B. Here, the feature "lack of edges" is no longer an effective feature to identify the circles in the collection as it also applies to the oval. Therefore, the choice of feature that can help in effectively classifying the circle depends on the other objects that are in the collection.

Similarly, the set of experiments selected as relevant to a functional category of interest may contain experiments which do not suggest any biological relevance but nonetheless be very relevant in GBA-based functional analyses. The choice of experiments selected as relevant depends on the other experiments in the microarray collection. We note that it would not be possible to identify such experiments based on literature knowledge alone.

5.9 Discussion

Previously, we have discussed the significance of using only relevant microarray datasets in functional analyses based on similarity metrics such as correlation. The idea of identifying relevant experiments follows from the discussion by Adler et al.(2009) who acknowledge the pitfalls of using large microarray collections in co-expression analyses and proposes manually selecting relevant datasets based on literature knowledge. However, this is increasingly impractical with the ever-increasing size of microarray databases. Additionally, experiments identified as relevant based on literature knowledge alone may not be sufficient to uncover co-expression between the genes of interest.

In this chapter we have presented an algorithm which is able to identify a set of experiments from a microarray collection which maximize the correlation between genes belonging to a process. The algorithm selected several experiments which have biological backgrounds relevant to the functional category of interest and in agreement with literature knowledge. For example, the experiments selected for defence-related terms were related to pathogen infection and stress signalling. Similarly, for growth-related terms such as regulation of cell cycle process and trichoblast maturation, growth-related experiment such as the Weigel developmental series experiment were selected.

Regardless of the biological background of the experiments in the selected set, the histograms of correlation coefficients show an enrichment of higher correlation coefficients. In the classification exercise, the significantly better 1-AUC scores in the ten-fold cross validation procedure show that the selections made by the algorithm are consistently superior to using experiments without selection. In Fig. 19, we observed that the performance of the selected set varied with the specificity of the functional annotation. As broader process annotations contain several smaller more specific processes, the overall correlation between the genes would be relatively lower than in specific processes. As a result, it is harder to differentiate the process from the background.

The conditional nature of co-expression has been the motivation for techniques such as Biclustering, first introduced by (Cheng & Church, 2000) and later developed by (Madeira & Oliveira, 2004). However, biclustering seeks to identify subsets with high correlation disregarding the biological background of the experiments. This is often unreasonable where the subsets are a part of larger experiments such as time courses. Further, biclustering is less effective on large microarray collections as the tendency to find local patterns due to noise also increases. Additionally, biclustering techniques are inapplicable in the graph-based functional analyses approaches such as (Long et al., 2008; Mostafavi et al., 2008).

The concept of a functional category is central to the idea of selecting relevant experiments. In this study we assume that genes assigned to a functional category are true members of the functional category. However, it is widely acknowledged that the overall error rates in Gene Ontology and MIPS is estimated to be 30% (C. E. Jones, Brown, & Baumann, 2007). The error rates can be even higher (up to 40%) in case of annotations assigned based on sequence and structure homology (C. E. Jones et al., 2007; Todd, Orengo, & Thornton, 2001).Therefore, it is reasonable that a given functional category is not functionally homogeneous because of wrongly assigned annotations. We believe that the performance of the algorithm would be lower as the error rate in annotation increases. The wrongly assigned genes are likely to lower the overall correlation in the functional category, thus making it difficult to differentiate the background set from the functional category of interest.

In this study we present experiment selections relevant to single functional categories only, the algorithm can be easily applied to select experiments relevant to multiple functional categories. This is based on the fact that fundamentally the algorithm is designed to select experiments relevant to a list of genes. Therefore by merging multiple functional categories, the algorithm can be easily applied to select experiments. The selected experiments would essentially be "best" datasets for discriminating the genes in the merged list from the background.

Our results for *A.thaliana* and Yeast show that the algorithm performs consistently independent of the type of organism. We also see that the selection performance is comparable regardless of GO or MIPSFunCat classification system. The algorithm is highly scalable and can be efficiently deployed to select experiments from large microarray collections. Currently, the run times of selection procedure increases exponentially with the number of microarray experiments in the selection.

One of the important applications of the experiment selection algorithm would be for selecting relevant datasets for modelling biochemical pathways. We believe that the experiments selected by the algorithm describe the co-relationships in a pathway better than without selection. We observe that with the selected set of experiments, the members of the pathway exhibit stronger edges among themselves compared to the edges in the background and the edges leading to the background genes. The selected set increases the likelihood of detecting true members of the pathway of interest. We believe, our semi-supervised experiment selection method can have a wide-reaching impact on the way datasets are selected for gene network construction, gene function prediction and biochemical pathway modelling.

6

Conclusions and Outlook

6.1 Conclusions

The context of this thesis is the elucidation of gene function using high-throughput transcriptomics data such as microarrays. An often used approach for characterising genes whose function is unknown is by using the principle of GBA. GBA-based approaches often use large collections of microarrays for calculating similarity between gene expression profiles. With this background, in this thesis we establish three facts:

1. Using large collections of microarrays in GBA-based functional analyses may not always be the optimal approach. We see that this approach may lead to poor correlation between genes in the same functional category.

2. Correlation between genes in the same functional category can be improved by limiting the dataset to functionally relevant experiments. However, selecting relevant experiments based on literature knowledge alone is a non-trivial task.
3. We have developed a greedy semi-supervised algorithm (Bhat et al. 2011, *under review*) that can select functionally relevant experiments for a given functional category.

In the first stage of the project we have closely looked at the idea of gene *function* and its organization in the post-genomic era. We discuss the principle of GBA and its application in analysing gene function. The motivation for using large collections of microarrays for calculating similarities between expression profiles is motivated by factors such as a longer gene expression profile will lead to a more robust correlation. However, previous studies have shown that this approach may not be suited to the functional analysis of genes whose expression is very condition-specific.

A pre-requisite of using gene expression data for functional analysis is that genes which belong to the same functional category should be highly correlated. Therefore, genes with unknown function could be characterized based on the similarity of their expression profile with the expression profile of genes with known functions. Hence, to set the stage for our work, we looked at the distribution of correlation coefficients among genes belonging to the same functional category. Here, the correlation between the genes was calculated with the often used approach of using a large collection of microarrays. We found that a majority of the functional categories we analyzed contained very low correlation between the genes.

We hypothesized that there could be at least two reasons for observing the poor correlations. Firstly, it could be an artefact due to the noisy nature of microarrays. The nature of technical noise in microarrays has been widely discussed in literature and numerous normalization pipelines have been proposed (B. M. Bolstad, Irizarry, Astrand, & Speed, 2003). A sophisticated normalization technique such as VSN (Qin et al., 2006) coupled with a customised quality monitoring such as M/A plots, RNA degradation curves and Normalized Unscaled Standard Error (NUSE) (B. Bolstad et al., 2005) is expected to limit the noise in the microarray datasets.

Secondly, the poor correlation could be due to cross-talk between the various biological processes. For these reasons, we hypothesized that using only functionally relevant experiments could limit noise and improve correlation between the genes. To verify this assumption, for functional categories of interest, from a large collection of microarrays, we selected experiments which we found relevant based on literature knowledge. The distribution of correlation between expression profiles improved significantly when only these experiments were used instead of a large microarray compendium. However, we realized that the distribution of correlation between genes in the same functional category is highly sensitive to the nature of experiments selected. To ensure the best distribution, selecting the right set of experiments was very important. However, identifying such a set of experiments was a non-trivial task as literature knowledge about a functional category was seldom exhaustive. Further, the relevance of an experiment to a given functional category may not be immediately obvious and experiments which are deemed irrelevant by a researcher could in fact withhold significant information.

In the second phase of the project, we developed a novel algorithm for systematically selecting from large collections those experiments which are relevant to a given functional category or pathway. A key concept behind the algorithm was the idea of "relevance" itself. We defined that an experiment would be relevant if it is able to clearly differentiate between the genes which belong to the category of interest and those which do not (called *background*), based on their gene expression profiles. The rationale behind this idea was that genes which belong to the same functional category would exhibit a high correlation between themselves compared to genes that belong to any other functional categories. The set of experiments selected by the algorithm maximize the differentiating ability between the functional category and background.

Importantly, the algorithm is able to identify relevant experiments not obvious by searching the literature on the experiment. Our results show that using experiments selected by the algorithm leads to substantially improved correlation between genes in the same functional category compared to using large heterogeneous collections of experiments. As a consequence, we also demonstrate that using correlation obtained with the selected experiments leads to substantial improvements in GBA-based function prediction. We are able to show that the improved performance of GBA-based analyses is independent of the species or functional classification systems. Finally, we have presented an example of how the algorithm can be applied for reconstructing biological pathways. Our algorithm is highly scalable and can be efficiently deployed to select experiments from large microarray collections. In conclusion, we believe that our semi-supervised experiments selection method can have a wide-reaching impact on the way datasets are selected for gene network reconstruction, gene function prediction and biochemical pathway modelling.

6.2 Future Perspectives

6.2.1 Improvements to the algorithm

In the current development of the algorithm, a limiting factor of the performance is the complexity of the search space. The nature of the complexity has been discussed in detail in Chapter 5. An important factor that is affecting the speed of the computation is calculation of the correlation matrices. In its current form, we use the full set of genes from the functional categories in the Background set in calculation. However, we believe that instead of the full set of genes, sampling only a few genes per functional category could also provide good results. As this will significantly trim the size of the correlation matrices to be computed, it can improve the run time by many magnitudes.

6.2.2 Selecting RNA-seq datasets for GBA analyses

In this thesis, our experiment selection algorithm was tested on microarray datasets. However, the application of the algorithm can be easily extended to other types of high-throughput transcriptomics datasets such as RNA-Seq data. This is feasible as the biological assumptions behind performing RNA-Seq experiments are similar to that of microarrays. In either of the cases, measured mRNA levels serve as a proxy for biological activity and correlation between gene expression profiles suggest a shared biological function. Additionally, the format of RNA-Seq data is also similar to processed microarray datasets, where rows of the data matrix represent genes and columns represent experimental conditions. With the increasing application of the RNA-seq method for Transcriptomics (Shendure, 2008), our algorithm could have a wide reaching impact on the way transcriptomics datasets are used in GBA-based functional analyses.

6.2.3 Mapping gene expression as a novel functional analysis technique

As discussed throughout this thesis, microarrays remain the single largest source of functional information. As discussed in Chapter 3, the core aim of performing microarray experiments is to assay the biological processes that are perturbed in response to a treatment. This has been achieved by following various levels of information extraction. These steps include identifying the differentially expressed genes in the dataset and clustering the differentially expressed genes to identify groups of genes with similar expression dynamics. Once the groups are identified , a typical analysis would be to perform a functional term enrichment analysis such as the ones presented in (D. W. Huang et al., 2009). Term enrichment methods look for functional categories that are over-represented in a list of genes such as a cluster. These methods provide a summarization of the various biological processes in a dataset.

We believe that the ideas developed during the course of this thesis could be applied to develop a novel functional analysis technique that is capable of illustrating the various biological processes in a microarray dataset. In this thesis, we have presented a method that is able to identify experiments that are relevant to a functional category of interest. Conversely, it is also possible to identify the relevant functional categories given a microarray experiment. Similar to the assumptions of a GBA analysis, genes in functional categories that have been perturbed by the treatment would be expected to have high overall correlation. However, a simple listing of functional categories with high overall correlation would lead to spurious results due to compounding factors such as variations in the number of genes in functional categories and redundancies in functional terms due to parent-child relationships. Based on this idea, we hope to develop an intelligent search technique that is able to identify a non-redundant list of functional categories that have been perturbed by the microarray experiment of interest. This technique can be used to map out the various biological processes that have been perturbed in response to a stimulus and can be visualized on a template of the GO tree. This approach would provide a completely novel perspective on the biological processes perturbed in a microarray.

6.2.4 Applications in modelling regulatory pathways

Systems approaches are increasingly becoming mainstream in various fields, particularly for modelling gene regulatory pathways and elucidation of gene function. Several techniques such as (Margolin et al., 2006) and (Zhou et al., 2005) have been developed for reconstructing gene regulatory networks from transcriptomics data. (W.P. Lee & Tzou 2009) is a recent review of the state-of-the-art in gene regulatory network reconstruction using transcriptomics data. These approaches generally employ large collections of microarrays to build an interaction networks. A common hurdle faced by these techniques is that genes which are known to be involved in the same pathway or process are not found to be co-expressed. This results in correlation-based interaction networks that do not reliably reflect regulatory relationships. We believe that by selecting datasets using our algorithm, the quality of the interaction networks can be significantly improved.

Appendix I

Microarray Data Analysis

Genes are transcribed into single stranded RNA molecules called the mRNA. Subsequently, the mRNA molecules are translated into sequences of amino acids which may undergo post-translational modifications to form proteins. Due to the central role played by mRNA in the above process, mRNA measurements are used as a proxy for protein production and subsequently biological function. Hence, microarray application is generally known as *Expression Analyses*. mRNA levels when measured over various time points or treatments are known as *Expression Profiles*. Although several techniques have been developed for measuring mRNA such as Northern Blot (Kevil et al., 1997), Quantitative Real-time Polymerase Chain Reaction (Higuchi, Fockler, Dollinger, & Watson, 1993), they remain low-throughput with a typical experiment measuring tens of genes. In contrast, DNA microarrays enable large-scale monitoring the activity of thousands of genes or the entire genome simultaneously, easily dwarfing the amount of data generated by any of the post-genomic experimental techniques.

The basic technology behind DNA microarrays is the process of *hybridization* (Stekel, 2003). Two DNA strands (and also RNA) will hybridize only if they are complementary to each other. The principles of hybridization were first used in the Northern and Southern Blotting to measure gene expression. In fact, DNA microarrays can be viewed as a high-throughput version of the Northern and Southern blotting. The general scheme of a microarray is a glass or a silicon slide on which thousands of DNA molecules are attached at specific locations called spots. Each spot may contain thousands of copies of DNA molecules and a typical microarray has thousands of spots where each spot is specific to a single gene. The spots are placed within micrometers of each other enabling entire genomes to be represented on a single chip. The mRNA sample to be tested is labelled with fluorescent dyes such as Cy3 and Cy5 and then incubated on the chip for hybridization. Any excess unhybridized sample is washed away. Subsequently, the fluorescent dyes attached to the sample are excited using lasers that scan the surface of the chip. Generally, the intensity of fluorescence is considered directly proportional to the amount of mRNA hybridized at a given spot. The intensity is recorded to obtain a quantification which serves as a proxy for the amount of the specific mRNA present in the sample.

Although several competing microarray technologies are available, they can be divided into two major categories; Single-colour arrays and Two-colour arrays. In the following sections, we discuss the general principles involved in a single colour array such as the Affymetrix GeneChip and a spotted two-colour array such the CATMA project (Crowe et al, 2003). We also outline the general protocols for data extraction and quality control for the two types of microarrays and apply the protocols in case studies.

3.1.1 Oligonucleotide arrays

Affymetrix GeneChip® is the most efficient and widely-used single colour, oligonucleotide microarrays (Lenoir & Giannella, 2006). The

GeneChip technology is based on principles perfected in the semiconductor industry where a light beam is used to control the deposition or removal of silicon. This process is known as Photolithography. Affymetrix uses lithographic masks to control the synthesis of nucleotides on a predetermined site on the array surface. The masks control the synthesis on several thousand squares (or spots) on the array, each containing several copies of the oligonucleotides. Each oligo is several nucleotides long and each gene is represented by up to 40 oligos. Affymetrix chooses 11 to 20 oligos to be perfect matches (PM) i.e. to be fully complementary to the incoming mRNA of the gene. In addition, 11 to 20 oligos, identical to the PM except for position 13 where one nucleotide has been changed from its complementary nucleotide to generate a mismatch (MM) are chosen. This system of PM and MMs was designed to deal with background and non-specific hybridization which make it harder to detect weakly expressed genes. GeneChip technology uses only a single fluorochrome. This necessitates the use of separate chips for the control and test in every control vs. test comparison. The intensity of fluorescence is considered proportional to the amount of mRNA bound to its complementary oligo. The intensity of fluorescence from each spot or square in the array is quantified by a sensor and used for subsequent analysis.

Before extracting a differentially expressed genes list from microarray data, the data is subjected to a data pre-processing pipeline. These steps are designed to prepare the raw microarray data for the subsequent statistical analyses. In the next section, we present some of the basic methods involved in the processing of raw data from Affymetrix GeneChip microarrays.

3.1.1.1 Data pre-processing and Summarization

Generally, the aim of a microarray analysis is to identify a set of genes that have been perturbed in response to the experimental conditions imposed on an organism or a sample of interest. The first step towards this goal is to ensure the integrity of the raw data obtained from the microarray as microarrays are very sensitive to technical variation and experimental noise which can lead to spurious results. It is important that the sources of variation in the experiments are corrected to achieve acceptable levels of accuracy of the data. The steps taken for correcting for technical variation depends on the microarray technology. The steps taken for Affymetrix GeneChips are relatively elaborate when compared to two-colour microarrays.

Step 1: Background correction

As soon as the fluorescence signal is obtained from the chip, the first step involves correcting the intensity reading in relation to any fluorescence in the background. An ideal case in the hybridization process on a microarray would be that the labelled sample binds specifically to the complementary oligonucleotides on the spots and nowhere else on the array surface. This would result in zero fluorescence from the background or the non-spot area. However, various factors such as non-specific binding of the labelled samples to the chip surface, inefficient washing after the hybridization or simply noise from the optical sensors can lead to varying levels of fluorescence from the background. For accurate quantification of signal intensity, it is necessary that microarray pre-processing algorithms make robust estimation of technical noise. Several pre-processing algorithms have been developed for GeneChip data such as MAS5.0, RMA and GC-

RMA. The salient features of each of the algorithms will be outlined later.

Step 2: Normalization

The normalization process is designed to account for any technical variation between the arrays used in an experiment. Technical variations between arrays occur due to subtle variations in the prevailing experimental conditions such as any mild variations in temperature, quantity of mRNA hybridized on the sample and even any slight difference in hybridization times. Such discrepancies lead to scaling differences in the fluorescence intensities between the various chips. This can potentially render the intensities from different chips incomparable. The normalization procedure ensures that gene expression levels recorded by the various chips are comparable.

Step 3: Perfect Match (PM) correction

As discussed earlier, the system of Perfect Match (PM) and Mis-match (MM) are unique to the Affymetrix GeneChip technology. This system was designed to measure both the relative abundance of the corresponding gene and the level of non-specific binding. Each MM probe reports the amount of non-specific binding for the gene represented by the PM probe. The PM corrections are handled differently by the various microarray pre-processing algorithms. The simplest procedure implemented by the early Affymetrix MAS algorithms involves simply subtracting the intensity of the MM probes from the intensity of the corresponding PM probes.

Step 4: Summarization

The GeneChip arrays have a unique design of 11 different PM probes, each targeting 11 separate sections of the target mRNA. Due to this unique design feature, it is necessary that the intensities for each of the 11 different PM probes are combined to obtain a single signal value for the corresponding gene. This process is called Summarization and is unique to GeneChip arrays.

3.1.1.2 Pre-processing pipelines developed for the Affymetrix GeneChip microarray

In the previous section, we introduced the basic concepts involved in the pre-processing of raw microarray data. Several pre-processing pipelines are available for processing GeneChip data. In the following, we present the three widely used systems.

Microarray Suite 5.0 (MAS5.0)

MAS 5.0 system was developed by Affymetrix when they first introduced the GeneChip technology in early 2000. The MAS 5.0 algorithm performs background correction on every PM and MM probes. As mentioned earlier, it is possible that the hybridization at the MM probes produces a greater intensity than the corresponding PM probe. For this reason, the MM intensity values are converted to ideal MM values that are always smaller than the values obtained from the corresponding PM probe. The robust mean (Tukey Bi-weight method) of the $\log_2$ transformed differences between the various PM and the MM values is recorded. The expression values obtained from each spot are then normalized by adjusting the mean of the signal value to a preset or a user-specified value. This normalized value is used as the processed microarray data.

Robust Multi-array Analysis (RMA)

The RMA algorithm (Rafael A Irizarry et al., 2003) is an open source effort for quantifying probe level signal intensities to gene expression data. In the hybridization step outlined earlier, it was assumed that the signal intensities at the PM probes would always be higher than at the MM probes. However, it was observed that in a significant number of cases, the signal intensities at the PM probes was found to be less than the corresponding MM probes. The developers behind the RMA algorithm argued that although the MM values are useful, it introduces a significant amount of noise. For this reason, the RMA algorithm completely ignores the MM values and considers only the PM values in the pre-processing. The algorithm works by adjusting for the background noise to ensure that the PM values are greater than the background intensities. The $\log_2$ transformed value of each background-corrected PM probe is obtained and these values are normalized using Quantile normalization (B.M. Bolstad et al., 2003). The RMA algorithm is then applied on the quantile normalized values.

GeneChip RMA (GC-RMA)

GC-RMA method is largely based on RMA. However, unlike RMA, in the background correction step the MM values are not discarded. Along with the MM values, Guanine and Cytosine content of the probe sequence are used to estimate the background (Z. Wu, Irizarry, Gentleman, Martinez-Murillo, & Spencer, 2004).

3.1.2 Two-colour arrays

Developed in the 1980's, the two-colour arrays are one of the earliest microarray technologies. Two-colour arrays were developed on open-source principles and as a result various approaches exist for its manufacture. Perhaps the most prolific microarray technique is the Spotted Array. Spotted Array technology is used by consortia such as CATMA (Complete Arabidopsis Transcriptome Microarray) for making both custom arrays and complete genome arrays. Primarily, manufacturing spotted arrays involves using automated (robot) spotters to place picoliter quantities of probes in solution onto a glass or a silicon surface. The probes used for spotting can be cDNA or oligonucleotides where each probe is complementary to a gene. The probes are attached to the surface by non-specific binding to poly-lysine coated glass or using processes such silanization. Primarily, each probe contains two channels where separate fluorochromes are used for two different biological samples such as *control* and *test*. Equal quantities of control and test samples are used. Generally, the fluorochromes Cy3 (green) and Cy5 (red) are incubated with one of the samples and spotted as a probe. The fluorescence intensity from each probe is quantified via a scanner. The ratio between the green and the red intensities is used as the quantification of gene expression. The greater the ratio, greater is the differential expression between the two biological samples. Visually, a red spot indicates up-regulation, green spot indicates down-regulation and a yellow spot indicates no change in gene expression between the test and the control samples.

Spotted arrays are well suited for manufacturing arrays with small number of probes (Stekel, 2003). Therefore, spotted arrays are widely used to make custom arrays such as partial genome arrays and pathway-specific arrays. Compared to *in-situ* synthesis in GeneChips, the accuracy of the spotted arrays decreases with the increase in the size of the array. This makes them more suitable for manufacturing smaller custom arrays rather than whole genome arrays.

3.1.2.1 Pre-processing two-colour microarray data

Similar to the Affymetrix GeneChip arrays, all two-colour microarray data require pre-processing to minimize technical noise and correct any bias in the signal measurements. However, due to the nature of the technology the pre-processing varies from single colour arrays. A brief outline of the data pre-processing pipeline has been presented below.

Step 1: Background correction

Following image analysis and quantification of the signal intensities from the two channels, the data needs to undergo a background correction step. Similar to the process in GeneChips, the aim of this step is to eliminate poor quality spots. Background fluorescence occurs due to non-specific binding of the labelled samples onto the chip surface in the non-spot area. The first step in background correction process is to eliminate any spot with intensity lower than the background plus two times the standard deviation (Leung & Cavalieri, 2003). The ratio of the intensities from the two channels is log transformed to make the control and the test signals comparable (Quackenbush, 2002).

Step 2: Normalization

Unlike GeneChip data, there are no Mismatch and Perfect Match values for each probe and hence no summarization step. The background corrected data are directly subjected to normalization routines to adjust for systematic biases in the data that might compromise the downstream analyses of the data. Major sources of such bias are the dyes used in the array. The dye bias can come from a variety of sources such as variation in dye and cDNA sample binding efficiencies, differences in heat and light sensitivities of Cy5 and Cy3 and also sensitivity of the scanner to the different wavelengths involved. Some of the commonly used normalization routines used in cDNA or two-colour microarray data analyses are the Loess normalization and the dye-swap normalization.

3.1.2.2 Information extraction: Case study

The data and analysis presented here are a part of a manuscript currently being prepared for publication.

> *Meristem outgrowth is repressed by a stress activated MAPK kinase pathway through regulation of auxin transport (2011), Hatzimasoura E, Doczi R, Ditengou F, Bhat P, Magyar Z, Helfer A., Menke F, Hirt H, , Lopez, E, Paccanaro A, Palme K, Bogre L.*

Introduction

Organ growth and morphogenesis in plants show extraordinary plasticity in response to environmental factors such as light, nutrients, temperature and biotic factors such as pathogen attack. Delaying growth is a common response in plants to environmental stresses. In Arabidopsis, meristem initiation is considered highly sensitive to environmental stresses. Most of the insights into environmental growth inhibition come from studies on various plant hormones, while little is known about the actual signalling interactions involved. Physiologically, hormones regulate systemic responses to signals perceived by cellular receptor and signalling mechanisms and conversely systemic hormonal signals are translated into cellular responses by perception and signal transduction. The mitogen-activated protein (MAP) kinase phosphorylation cascades are conserved signalling modules in all eukaryotes and known to have pivotal roles to regulate cell division, cell growth and stress responses in animals. The aim of the study is to adopt an experimental approach to uncover a negative regulatory function of MKK7/MKK9-MPK6 modules of the MAPK family. The study shows that meristem de-repression in response to exposure to light is accelerated in both the *mkk7* and *mpk6* mutant seedlings. Gene expression analysis was performed to reveal that MKK7 and MKK9 regulate a transcriptional reprogramming diverting the plant from photosynthetic growth to defence response.

Materials

A time course induction was performed on seedlings carrying the empty *pER8GW* vector, *pER8GW:myc:MKK7* and *pER8GW:myc:MKK9* constructs for 0, 2 and 8 hours using 1 μM β-estradiol, as well as 0.5 and 1 hour induction samples for empty vector and *myc:MKK7* seedlings. This arrangement was designed to exclude any changes due to circadian rhythm to be detected as differential gene expression. Three biological replicas were obtained for each sample.

The whole genome expression profiling was performed using seedlings at the 1.02 developmental stage. 6-day-old seedlings grown on 0.5x MS media were transferred to liquid 0.5x MS media and rested overnight. For transgene induction the media was drained and replaced by 0.5x MS supplemented with 1 μM β-estradiol. Each sample was obtained by pooling three biological replicas of approximately 50 seedlings. Total RNA was isolated by the RNeasy Plant Mini Kit (Qiagen), DNase treatment was performed by DNase away (Qiagen). The Cy3 and Cy5 dye labelled cDNA samples were hybridised to CATMA (Complete Arabidopsis Transcript MicroArrays) microarrays(Crowe et al., 2003), produced at the University of Utrecht. The microarrays were scanned using the Scanarray software (Perkin Elmer). The scanned arrays were then quantified using the Imagene software (Biodiscovery). Microarray preparation, scanning and image quantification were performed by Elizabeth Hatzimasoura.

Results

1. Data pre-processing

Microarray was pre-processed using the GeneSpring GX10 Suite (Agilent Technologies) and all the subsequent data analyses were performed in MATLAB (The Mathworks). Following quantification, the text files containing the signal intensities from each of the 48 individual chips was imported into GeneSpring. The imported CATMA data is arranged in the matrix format by GeneSpring where each row is a probe/gene and every column is a chip. The first step is a normalization procedure where the background signal intensity was subtracted from the foreground. Subsequently, the data was log-transformed and normalized using Loess-normalization for each print-tip individually (also known as Print-tip normalization). The data was then $\log_2$ transformed and averaged over the two dye-swaps. The normalization routine followed is similar to the one outlined in (Allemeersch et al., 2005).

To remove unreliable measurements, a filter on expression was applied where we assumed that the signal intensity of an expressed gene would be greater than the 20th percentile of all signal intensity values of the sample. A lower percentile cut-off of 20 and an upper percentile cut-off of 100 were applied. Genes with expression values out of this range in any of the time points were filtered out. Genes with Present and Marginal flags (assigned by GeneSpring) were retained and the rest were filtered out. This resulted in 21,871 probes out of a total of 27,649 probes. The data were then exported for downstream analysis.

Functional analysis of microarray data

Subsequent to the quality control measures, all downstream analyses was performed using MATLAB (The Mathworks). To detect the significantly differentially expressed genes, we performed a 1-way ANOVA with Time as the factor. The p-values of the ANOVA were adjusted for false discovery using the Benjamini-Hochberg method with a cut-off value of 0.05. This resulted in 6447 genes.

To identify the dominant patterns in the differentially expressed genes, firstly we applied *k*-means clustering with *k*= 20. The value of *k* was arbitrarily determined based on several trials ranging from *k*= 15 to *k*=30, where *k* =20 was found to be appropriate. Additionally, we also applied the Quality-Threshold clustering technique (Heyer, Kruglyak, & Yooseph, 1999). The QT clustering algorithm has several advantages compared to *k*-means approach. Importantly, the number of clusters is not decided *apriori* in contrast to *k*-means. We chose a maximum cluster diameter of 0.4, which resulted in 31 clusters (Fig.22). Compared to the clusters obtained by applying the *k*-means technique, the clusters obtained by QT algorithm were found to be qualitatively superior. This is based on the observation that in the downstream functional term enrichment analysis, clusters obtained from QT algorithm showed higher GO term enrichment scores. This suggested that clusters obtained from QT clustering could be functionally more homogenous compared to *k*-means. In addition to clustering, gene lists were also produced based on fold-change in gene expression between zero hour and the subsequent time points e.g. 0 hour vs. 1 hour and 0 hour vs. 8 hour. These lists were further sorted based on the gene expression dynamics e.g. two-fold late up-regulated and two-fold late down-regulated.

Subsequent to the clustering of the differentially expressed genes, in order to characterise the functional themes inherent in the clusters we performed GO term enrichment analysis using the BiNGO plug-in (Maere et al., 2005) of Cytoscape (Shannon, Markiel, Ozier, Baliga, Wang, Ramage, Amin, Schwikowski, & Ideker, 2003b). The term enrichment p-values were adjusted for multiple testing corrections using the Benjamini-Hochberg method. Gene lists prepared based on fold-change in expression were also tested for GO term enrichment. Genes which were found to be many-fold down-regulated at the 8 hour time point compared to 0 hour were found to be enriched in terms related to photosynthesis. Additionally, the list was also significantly enriched with cold-responsive genes. Genes that were found up-regulated by MKK7 were found to be enriched with genes related to defence response. This result was echoed by the clusters generated by QT clustering technique as well as *k*-means (result not shown): development and photosynthesis enrichment in down-regulated clusters (e.g. QT clusters 1, 2 and 4) while defence and catabolism was enriched in up-regulated clusters (e.g. QT clusters 3, 7 and 10). Selected overrepresented GO terms of the major co-regulated gene clusters by the QT algorithm are summarised in Table 9.

Downregulated clusters			**Upregulated clusters**		
cluster	**p value**	**GO term**	**cluster**	**p value**	**GO term**
1	4.15E-03	protein folding	3	6.83E-07	Phosphorylation
1	6.52E-03	Photosynthesis	3	1.79E-06	protein amino acid phosphorylation
1	8.03E-03	multicellular organismal development	3	2.76E-06	post-translational protein modification
1	8.79E-03	embryonic development	3	1.58E-05	response to stimulus
1	9.35E-03	chloroplast organization and biogenesis	3	1.92E-05	response to other organism
2	1.81E-04	regulation of biological quality	3	1.31E-04	response to stress
2	3.90E-04	cellular homeostasis	3	7.88E-04	signal transduction
2	1.79E-03	response to brassinosteroid stimulus	3	8.77E-04	immune response

2	2.91E-03	plastid organization and biogenesis
2	5.72E-03	lipid biosynthetic process
2	8.36E-03	chlorophyll biosynthetic process
2	9.85E-03	response to hormone stimulus
4	6.51E-07	pigment metabolic process
4	2.05E-05	pigment biosynthetic process
4	8.14E-05	cofactor metabolic process
4	3.03E-04	plastid organization and biogenesis
4	7.30E-04	lipid metabolic process
4	1.31E-03	steroid biosynthetic process
4	1.94E-03	chloroplast organization and biogenesis
4	4.77E-03	establishment and/or maintenance of chromatin architecture
5	2.42E-21	Translation
5	3.00E-20	gene expression
5	2.35E-19	macromolecule biosynthetic process
5	2.86E-18	cellular biosynthetic process
5	3.42E-10	organelle organization and biogenesis
5	6.92E-06	Photosynthesis
6	8.67E-05	cellular biosynthetic process
6	9.04E-05	nucleotide biosynthetic process
6	1.74E-04	biosynthetic process
8	1.95E-03	regulation of cellular protein metabolic process
8	5.03E-03	chromatin assembly
8	5.62E-03	DNA packaging
9	1.93E-04	defense response to fungus
9	2.04E-03	tissue development
9	6.75E-03	post-embryonic development
9	8.04E-03	system development
9	8.04E-03	organ development
9	8.83E-03	multicellular organismal process
9	9.40E-03	stomatal complex development
3	1.07E-03	defense response
3	1.70E-03	second-messenger-mediated signaling
3	3.22E-03	intracellular signaling cascade
3	3.34E-03	response to unfolded protein
3	3.54E-03	response to chemical stimulus
3	4.09E-03	establishment of localization
3	5.72E-03	Aging
3	6.57E-03	innate immune response
7	4.55E-06	vesicle-mediated transport
7	6.78E-06	catabolic process
7	3.22E-04	response to misfolded protein
7	9.96E-04	establishment of protein localization
7	8.90E-03	Glycolysis
10	1.89E-03	actin filament organization
10	5.11E-03	abscisic acid mediated signaling
10	7.76E-03	response to water deprivation
11	1.25E-04	regulation of apoptosis
11	3.99E-03	lipid catabolic process
12	7.02E-05	catabolic process
12	3.21E-04	vegetative to reproductive phase transition
12	8.86E-04	sexual reproduction

Table 9: Selected overrepresented GO terms characterizing the major down- and upregulated clusters generated by the QT algorithm.

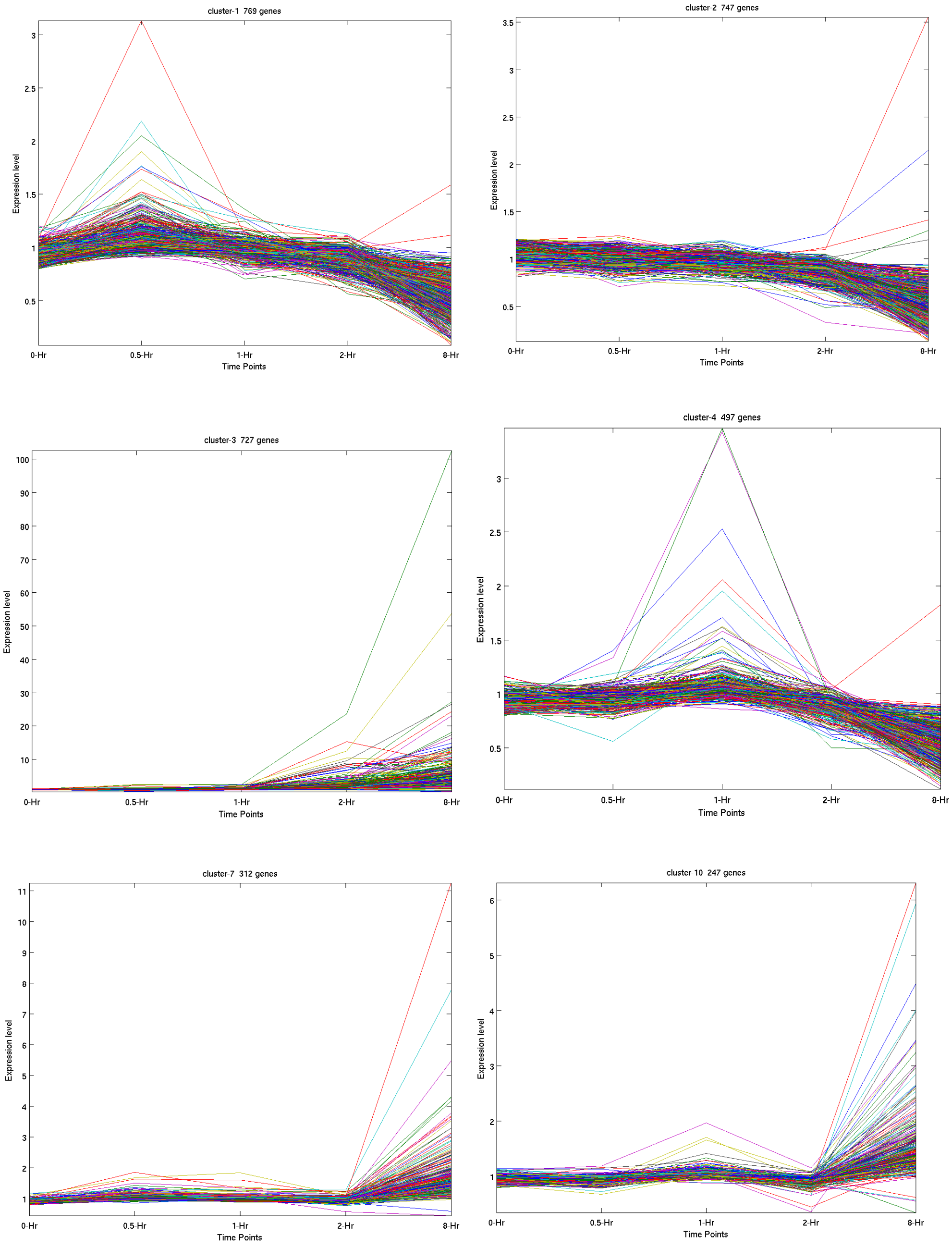


Figure 22: Gene expression profile plots of selected clusters from the QT method.

Discussion

Analysis of MKK7 and MKK9 function at the gene expression level revealed the prevailing tendency of down-regulation of growth and the induction of defences in good agreement with the regulatory functions implied by the observed phenotypes (data not shown). The large number of genes with altered expression levels, especially by MKK7, suggests that MKK7/MKK9 mainly act in modulating target gene expression rather than providing binary on/off inputs. The abundance and diversity of response genes also suggests that MKK7/MKK9 control complex regulatory machinery rather than a small number of executive genes. Thus, the main role of MKK7 and MKK9 is probably to fine tune various stress responses, which can explain the pleiotropic phenotypes caused by their altered expression.

In this study, although we also pre-processed microarray data showing the transcriptional response to induced MKK9 over-expression, we have not presented any results from the functional analyses. This is due to suspected quality issues in the MKK9 construct which resulted in leaky expression profiles in the early time points. Reflecting this possibility, a very low number of genes (133) passed the ANOVA performed for identifying differentially expressed genes. Poor agreement between the biological replicates resulted in very low significance levels for any hypothesis testing.

Appendix II

MATLAB implementation of the experiment selection algorithm

The MATLAB code for the experiment selection algorithm (presented in Chapter 6) and the datasets used in this thesis can be downloaded from: http://www.paccanarolab.org/papers/CorrGene/PAPER_CODE.zip

Instructions for running the code:

1. Extract the contents of the zip file into a folder named "PAPER_CODE".
2. Run MATLAB
3. Include "PAPER_CODE" and all its sub-directories in the path.
4. Run the script "Initialize.m" and specify organism of interest. Type "1" for Arabidopsis or "2" for Yeast. This reads in the microarray data, GO tree structure and gene annotations.
5. To run the algorithm, run the script "Run_and_display.m". IMPORTANT: Please ensure that "Initialize.m" has been run before executing this script
6. The user is prompted to enter the GO identifier of the functional category to select experiments for. For example, enter "51726" for the GO category "GO:0051726, Regulation of Cell cycle"
7. When prompted, enter the threshold for the t-test p-value. e.g. 0.05
8. When prompted, enter the number of experiments to be used as seed. e.g. K = 15.

9. The index numbers of the selected experiments will be output on the screen as well written into a text file named "selectedExperiments.txt".
10. The program will also output the ROC curves indicating the performance of the selected set vis-à-vis when all experiments in the collection are used. The ROC curves are saved as "*.png" files.
11. The 1-AUC values from the two ROC curves are recorded in a text file "AUCreport.txt". Here the first column indicates the GO identifier selected. Second column indicates the index number of selected set of experiments, the third column indicates the 1-AUC when all experiments are used and the fourth column indicates the 1-AUC when the selected experiments are used.

The MATLAB code for the experiment selection algorithm and all the functions developed for the algorithm are presented below:

Run_and_display.m

```
%This script runs the experiments selection algorithm and generates ROC
%curves for the selected datasets.
%IMPORTANT: Please run "Initialize.m" prior to running this script.

queries = input('Enter GO Identifier for category of interest: ');%Enter GO
identifier for the GO category of interest
TH = input('Enter threshold for t-test p-value: '); % set threshold for the t-test
p-value
seed_size = input('Enter experiment seed size: ');% specify the number of
experiments to be used as the seed set.

fornQ = 1:numel(queries)
qID = queries(nQ);
bgIDs = getallBGid(qID); %retrieves rest of the GOIDs from the GO, this forms the
background set
queryGOID = [qIDbgIDs];
disp('Retrieving data...')
    [alldata, labels, sizeMat] = getAllData(queryGOID); %extract data from ALL
microarrays, for the GOIDs in the queryGOID list
topExp = getTopExp(sizeMat, seed_size, TH); %gets the top N experiments specified
by "seed_size"
disp('Experiment selection...')
```

```
SelectSet = selectExperiments(topExp, sizeMat, maxExp, TH); %runs the experiment
selection algorithm

    %%%%%% Generate ROC curves for the selected and all experiments
disp('Plotting ROC curves...')
for S = 1:2
switch(S)
case(1) %Case 1 uses All experiments
exp_retr = [1:maxExp];
expLabel = 'All';
color = {'r'};
case(2) %Case 2 uses the Selected experiments
exp_retr = SelectSet;
expLabel = num2str(exp_retr);
color = {'g'};
end
            [corrM ,pval] = getCorrMat(exp_retr,sizeMat, TH);  %prepare
correlation matrix with data from chosen experiments
corrM = abs(corrM);
cmLength = length(corrM);
labelMat = zeros(sizeMat,cmLength);  % a zeros matrix of the same size as the cut
corr matrix above
labelMat(1:sizeMat, 1:sizeMat) = 1; % ones in the query GOID matrix , zeros for
the background matrix
corrVec = abs(corrM(:));
            [sortCorr, scidx] = sort(corrVec, 'descend');
labelVec = labelMat(:);
sortLabel = labelVec(scidx);
            [auc, prec_at] = get_precision_aucPB(sortCorr, sortLabel,color);
hold on
resultAUC(nQ, S) = auc;  % auc = 1-auc
end
title(labels(1))  %adds title to the plot drawn by getPrecisionAUC
        %Report generation
str = strcat(int2str(queryGOID(1)),':',int2str(sizeMat),':',char(labels(1)),' : ',
int2str(SelectSet));
strFig = strcat(int2str(queryGOID(1)));
dlmwrite('SelectedExperiments.txt', str,'delimiter','','-append')
print(gcf,'-dpng',strFig)
close
end
dlmwrite('AUCreport.txt', resultAUC, 'delimiter','\t','-append')
```

Initialize.m

```
% INITIALIZATION SCRIPT FOR YEAST AND ARABIDOPSIS THALIANA
%
% This script loads all the datasets required for running the experiment
% selection algorithm. Please set "PAPER_CODE" as the default directory
% before running this script and place the following files in the
% "DATA" sub-directory of "PAPER_CODE":
% arrayData: A compendium of microarray datasets in matrix form where every
% row represents a gene and every column is an experimental condition. The
% matrix should not contain any missing values.
% arrayInd: This is an index file containing identifiers for each
% experiment in the compendium.
% arrayGenes: This contains the list of gene identifiers for the data in
% arrayData.
% Annot: the file containing the gene annotation, can be downloaded
% from http://www.geneontology.org/GO.downloads.annotations.shtml
% GO: The gene ontology file in OBO format
%
% PrajwalBhat, July 27, 2011
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

clear all; close all; clc;
org=input('Which organism do you want to work on? [1 = ARABIDOPSIS, 2 = YEAST]:
','s');
fprintf('Please wait while the system is initialized. this might take a few
minutes... \n');
%input = 1;
switch(str2num(org))
case(1)
disp('Getting data...')

globalarrayData; %data matrix containing the microarray compendium
arrayData = importdata('../../DATA/ARABIDOPSIS_DATA/ATGE_data_Mod.txt');

globalarrayInd; %index file to identify experiments
arrayInd = importdata('../../DATA/ARABIDOPSIS_DATA/ATGE_exp_indices_Mod.txt');

globalarrayGenes; %file containing gene identifiers
arrayGenes = importdata('../../DATA/ARABIDOPSIS_DATA/ATGE_genelist.txt');
arrayGenes = lower(arrayGenes);

globalAnnot; %annotation file for the selected organism
Annot = goannotread('../../DATA/ARABIDOPSIS_DATA/gene_association.tair');  %
loading Arabidopsis annotation file

global GO; %load GO tree file
        GO = geneont('File','../../DATA/ARABIDOPSIS_DATA/GOobject');

globalalldata;

globalmaxExp;
maxExp = size(unique(arrayInd),1); %count the total number of
        % experiments available in the compendium

case(2)
disp('Getting data...')

globalarrayData; %data matrix containing the microarray compendium
arrayData =
importdata('//home/paccanaro/praj/Gene_corr_paper_yeast/Data/m3dyeast201010/m3ddat
a201010.txt');

globalarrayInd; %index file to identify experiments
arrayInd =
importdata('/home/paccanaro/praj/Gene_corr_paper_yeast/Data/m3dyeast201010/m3dexpI
ndex201010.txt');

globalarrayGenes; %file containing gene identifiers
```

```
arrayGenes =
importdata('/home/paccanaro/praj/Gene_corr_paper_yeast/Data/m3dyeast201010/m3dgene
s201010.txt');
arrayGenes = lower(arrayGenes);

globalAnnot; %annotation file for the selected organism
Annot =
goannotread('/rmt/csnewton/paccanarohome/praj/Gene_corr_paper_yeast/Data/gene_asso
ciation.sgd');

global GO; %load GO tree file
        GO =
geneont('File','/rmt/csnewton/paccanarohome/praj/Gene_corr_paper_yeast/Data/GOobje
ct');

globalexpLabels;
expLabels =
importdata('/home/paccanaro/praj/Gene_corr_paper_yeast/Data/m3dyeast201010/m3dexpL
ist.txt');

globalalldata;

globalmaxExp

maxExp = 31;
end
fprintf('INITIALIZATION COMPLETED \n');
```

SelectExperiments.m

```
%USAGE:
% SelectSet : selected experiments for the test GOID set
% top : top N experiments which were selected by the function "topExp"
% labelSize : number of genes in the test GOID
% maxExp : total number of experiments in the microarray dataset
% threshold : p-value cut-off limit for the t-tests
% Also requires "alldata" which can be declared globally

functionSelectSet = selectExperiments(top,labelSize, maxExp, threshold)
SelectSetCell = cell(maxExp,1);

fori = 1:length(top)
SelectSetCell{i} = top(i);
end

BestP = ones(maxExp,1);
Y = threshold;

fori = 1:length(top);
SelectSet = SelectSetCell{i};
pList = ones(maxExp,1);
first = 1;
while (first || min(pList)<threshold)
first = 0;
        REM = setdiff(1:maxExp, SelectSet);
```

```
pList = ones(maxExp,1);
tic
for j = REM;
            [corrMat, pMat] = getCorrMat([SelectSet,j], labelSize,threshold); %
corrMat = abs(corrMat);
            [hpval] = doTtest(corrMat, labelSize); %t-test
pList(j) =  pval;  %store pvalue from the t-test2
end
toc
        OLDY = Y;
        [Y,I] = min(pList);
if(Y<OLDY)
SelectSet=[SelectSet,I];
BestP(i)=Y;
            [i, BestP(i), SelectSet]
else
SelectSetCell{i}=SelectSet;
break;
end
end
end

[Y, I]=min(BestP);
SelectSet=SelectSetCell{I}
```

getTopExp.m

```
% This function performs t-test between the label of interest and the
% background for every experiment in the microarray collection

function [topexpID] = getTopExp(sizeMat, topN, TH)
globalmaxExp;
pvalList = ones(maxExp, 1);
fori = 1:maxExp
sprintf('Calculating TopN : Loop %d',i)
    [cMat, pMat] = getCorrMat2(i, sizeMat,TH,1);
    [hpval] = doTtest(cMat, sizeMat);
pvalList(i) = pval;
end
[~, Idx] =  sort(pvalList,'ascend');
topexpID = Idx(1:topN) %top N experiments from the sorted list
```

getCorrMat.m

```
%USAGE:
%expIdx : index number of the experiments to be selected from the
%microarray collection
%threshold : p-value cut-off limit for the correlation
%
% "getCorrMat" also requires "alldata" and "arrayInd" which can be declared
globally
% alldata : data from all the microarrays in the dataset
% arrayInd : file containing index numbers of the all the experiments
```

```
function [corrMat, pMat] = getCorrMat(expIdx, sizeMat, threshold)
globalarrayInd;
globalalldata;

Eidx = ismember(arrayInd, expIdx);
selectData = alldata(:,Eidx);  %retrieves data from selected experiments, for all
the genes
[corrMatpMat] = Fast_Corr([1:sizeMat], selectData);  %HX's fast correlation
function
corrMat(pMat>threshold) = 0;
```

getCorrMat2.m

```
% This function calculates the correlation matrix for specified experiments
% USAGE:
% expIdx : index number of the experiments to be selected from the
% microarray collection
% threshold : p-value cut-off limit for the correlation
%
% "getCorrMat" also requires "alldata" and "arrayInd" which can be declared
globally
% alldata : data from all the microarrays in the dataset
% arrayInd : file containing index numbers of the all the experiments

function [corrMat, pMat] = getCorrMat2(expIdx, sizeMat, threshold, type)
globalarrayInd;
globalalldata;
Eidx = ismember(arrayInd, expIdx);
selectData = alldata(:,Eidx);  %retrieves data from selected experiments, for all
the genes
if (type==2)
[corrMatpMat] = corrcoef(selectData');
elseif (type==1)
[corrMatpMat] = Fast_Corr([1:sizeMat], selectData);  %HX's fast correlation
function
end
threshold; %not used
```

getAllData.m

```
function [alldata, labels, sizeMat] = getAllData(queryGOID)

globalarrayData;
globalarrayInd;
globalarrayGenes;
globalAnnot;
maxExp = 44;   %total number of experiments in the dataset
global GO
arrayInd = arrayInd';
arrayGenes = lower(arrayGenes);

GOgenes = {Annot.DB_Object_Symbol};  %the full arabidopsis gene list from the
annotation file
GOID = [Annot.GOid];    %get associated GO terms
Aspect = {Annot.Aspect};  %get the 3 ontologies
```

```
Evidence = {Annot.Evidence};
BPMask = strcmp({Annot.Aspect}, 'P');  %Get only biological process terms
GOgenes = GOgenes(BPMask);  %get genes from biological process tree
GOID = GOID(BPMask);   %get GO terms from biological process treee
Evidence = Evidence(BPMask);    %get evidence codes from genes in biological
process tree
EvMask = ismember(Evidence, {'EXP' 'IDA' 'IPI' 'IMP' 'IGI' 'IEP' 'ISS' 'IC' 'ISO'
'ISA' 'ISM' 'IGC'}'); % all annotations except microarrays and electronic
GOgenes = GOgenes(:,EvMask);  %retrieve experimentally annotated genes
GOgenes = lower(GOgenes);
GOID = GOID(:, EvMask);   %retrieve experimentally annotated GO terms

geneproc = [];
dataproc = [];
labelproc = [];
sizeproc = [];
GOstruct = GO(queryGOID);
fori = 1:length(queryGOID)

queryLabel = cellstr(GOstruct.term(i).name);   %retrieves GO annotation for each
of the query GOIDs
desID = getdescendants(GO, queryGOID(i));   %gets all the descendant GO:IDs
Gidx = ismember(GOID, desID);
retGenes = unique(GOgenes(:, Gidx));   %retrieve genes which belong to the
descendants
retGenes = cellstr(lower(retGenes'));

ifi ==1
        N = 35;  % get only N number of genes from each category
else N = 1;
end
ifisempty(retGenes)
retGenes = ('');
elseif (length(retGenes)<N)
retGenes = retGenes(1:length(retGenes));
elseretGenes = retGenes(1:N);  %restrict the number of genes retrieved to N for
each GO category
end
    %%%% extracting gene expression data
Lidx = ismember(lower(arrayGenes), lower(retGenes));
allData = arrayData(Lidx,:);  %retrieves data from all experiments
getGenes = arrayGenes(Lidx,:);
getSize = length(getGenes);
getLabels = repmat(queryLabel, length(getGenes),1);
    %%% complete dataset retrieved
geneproc = [geneproc; getGenes];
labelproc = [labelproc; getLabels];
dataproc = [dataproc; allData];
sizeproc = [sizeproc; getSize];

end
sizeMat = sizeproc(1);
%%%%seperate background data
restGenes = geneproc(sizeproc(1)+1:end, :);
restLabels = labelproc(sizeproc(1)+1:end, :);
restData = dataproc(sizeproc(1)+1:end,:);
%%%%%%make unique genes dataset for the background data
```

```
[uniqGenes, m, n] = unique(restGenes);
uniqLabels = restLabels(m);
uniqData = restData(m,:);
%%%%%%%% prepare data for output: concatenate unique background gene data
%%%%%%%% with the query gene data
labels = [labelproc(1:sizeMat(1)); uniqLabels];
alldata = [dataproc(1:sizeMat(1),:); uniqData];
```

getallBGid.m

```
%function takes the foreground id as the input, gets all background GOID
%such that the list contains no descendants of the input id and the GOIDs
%haveatleast one of the annotated genes in the microarray data

function [bgLabels] = getallBGid(fgLabel)

globalAnnot
global genes
global GO

GOgenes = {Annot.DB_Object_Symbol};  %the full arabidopsis gene list from the
annotation file
GOID = [Annot.GOid];    %get associated GO terms
Aspect = {Annot.Aspect};  %get the 3 ontologies
Evidence = {Annot.Evidence};
BPMask = strcmp({Annot.Aspect}, 'P');  %Get only biological process terms
GOgenes = GOgenes(BPMask);  %get genes from biological process tree
GOID = GOID(BPMask);   %get GO terms from biological process treee
Evidence = Evidence(BPMask);    %get evidence codes from genes in biological
process tree
EvMask = ismember(Evidence, {'EXP' 'IDA' 'IPI' 'IMP' 'IGI' 'IEP' 'ISS' 'IC' 'ISO'
'ISA' 'ISM' 'IGC'}'); % all annotations except microarrays and electronic
GOgenes = GOgenes(:,EvMask);  %retrieve experimentally annotated genes
GOgenes = lower(GOgenes);
GOID = GOID(:, EvMask);   %retrieve experimentally annotated GO terms
GOIDlist = GOID;
GOgenesList = GOgenes;
inputLabels = getdescendants(GO, fgLabel);
goidx = ismember(GOIDlist, inputLabels);
GOIDlist(goidx) = [];  %filter out Foreground GO label and its descendents from
the GOID list
GOgenesList(goidx) = [];
geneidx = ismember(lower(GOgenesList), lower(genes)); %Remove genes in the GO list
which are not found in the microarray
GOIDlist(geneidx) = [];
%%%%%%% remove any obsolete IDs present in the list
obs = get(GO.terms,'obsolete');
mask = ismember(cell2mat(obs), 1);
obsTermsStruct = GO.terms(mask);
obsTerms = cell2mat(get(obsTermsStruct,'id'));  % make a list of obsolete IDs

terms = get(GO.terms,'id');
idx1 = ismember(GOIDlist, cell2mat(terms'));
GOIDlist(~idx1) = [];  %filter out any IDs not present in the GOtree
idx = ismember(GOIDlist, obsTerms');
GOIDlist(idx) = [];   % filter out obsolete terms from GOID list
```

```
bgLabels = unique(GOIDlist);
```

get_precision_aucPB.m

```
%This function produces ROC curves and AUC

function [auc,prec_at] =get_precision_aucPB(Pre,G, color)

recall_list=[0.01,0.1,0.5,0.8];
prec_at=zeros(4,1);

ind=find(G>0);

Th_List=unique(Pre(ind));
Th_List=[min(Pre);Th_List; max(Pre)];
Th_List=unique(Th_List);


AC_P=length(ind);
AC_N=length(G)-AC_P;

N=length(Th_List);

if(AC_P*AC_N==0 || N<3)
auc=0.5;
prec_at=0;%recall_list;
return;
end


TP=zeros(N,1);
TN=zeros(N,1);
FP=zeros(N,1);
FN=zeros(N,1);


for(i=1:N)

    TH=Th_List(i);
    I=find(Pre>=TH);
    J=find(Pre<TH);

    TP(i)=length(find(G(I)==1)); %TP should be 1, 2, 3..
    FP(i)=length(find(G(I)==0)); %TP should be 1, 2, 3..

TN(i)=length(find(G(J)==0));
FN(i)=length(find(G(J)==1));

end%TH


FPR=FP/AC_N;
TPR=TP/AC_P;

Recall=TP./(TP+FN); %check Recall==TPR
Precision=TP./(TP+FP);
```

```
%plot(Recall, Precision);

for(i=1:4)
    [Y,I]=min(abs(Recall-recall_list(i)));
prec_at(i)=Precision(I);
end

plot(FPR,TPR, char(color),'LineWidth',2);% for ROC
auc=0;
for(i=1:N-1)
    a=TPR(i);
    b=TPR(i+1);
    h=abs(FPR(i+1)-FPR(i));
auc=auc+0.5*(a+b)*h;
end
auc=1-auc;
end
```

fastCorr.m

```
%This function calculates a mxn matrix of correlation
function [Co, P]=Fast_Corr(IND, Matrix)
%IND is an index for the genes of interest
%Matrix is the micorarray matrix, rows for genes, and columns for
%experiments


if(size(IND,1) > size(IND,2))
    IND=IND';
end
n=length(IND);
Part=n*3;

N=size(Matrix,1);
REM=setdiff(1:N, IND);

Co=zeros(n,N);
P=zeros(n,N);

Iter=ceil((N-n)/Part);

%%get division
for(i=1:Iter)
if(i<Iter)
PartSet{i}=REM( (i-1)*Part+1: i*Part );
else
PartSet{i}=REM((i-1)*Part+1:end);
end
end

for(i=1:Iter)
callset=[IND, PartSet{i}];
    Ma=Matrix(callset, :);
    [c,p]=corrcoef(Ma');
```

```
Co(:,callset)=c(1:n, :);
P(:,callset)=p(1:n, :);
end


end
```

doTest.m

```
% Usage:
%
% corrMat : correlation matrix
% sizeMat: size vector containing sizes for genes in each GOID
% pval = p-value from the T-TEST2
%
% To use this function sizeMat must be GREATER than 1

function [h pval] = doTtest(corrMat, sizeMat)
%extract foreground matrix
pval = [];
corrMat = abs(corrMat);
bMat = zeros(length(corrMat), length(corrMat));
LCorr = corrMat(1:sizeMat(1), 1:sizeMat(1));   %corr matrix for the forground or
the label of interest
[m, n] = size(LCorr);
bMat(1:m,1:n)=LCorr(1:m,1:n);
bCorr = bMat - corrMat;  % corr matrix for the background
bCorr = abs(bCorr);
mask = repmat(2, m, n);  % mask for cutting corr matrix of label of interest
Trimask = triu(mask, 1);
LData = LCorr(Trimask==2);  % corr data vector for the foreground or label of
interest
[r, c] = size(bCorr);
maskB = repmat(2, r, c);
maskB(1:m, 1:n) = 0;  % mask for cutting out the background
TriMaskB = triu(maskB, 1);
BgData = bCorr(TriMaskB==2);
[h, pval, ci] = ttest2(LData, BgData,[],'right');
clearcorrMat
```

Bibliography

Adler, P., Kolde, R., Kull, M., Tkachenko, A., Peterson, H., Reimand, J., & Vilo, J. (2009). Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods. *Genome biology*, *10*(12), R139. doi:10.1186/gb-2009-10-12-r139

Al-Shahrour, F., D'iaz-Uriarte, R., & Dopazo, J. (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, *20*(4), 578. Oxford Univ Press.

Alabadí, D., Oyama, T., Yanovsky, M. J., Harmon, F. G., Más, P., & Kay, S. A. (2001). Reciprocal regulation between TOC1 and LHY/CCA1 within the Arabidopsis circadian clock. *Science (New York, N.Y.)*, *293*(5531), 880-3. doi:10.1126/science.1061320

Albert, R., Jeong, H., & Barabasi, A. (2000). Error and attack tolerance of complex networks. *Nature*, *406*(6794), 378-82. doi:10.1038/35019019

Allemeersch, J., Durinck, S., Vanderhaeghen, R., Alard, P., Maes, R., Seeuws, K., Bogaert, T., et al. (2005). Benchmarking the CATMA microarray. A novel tool for Arabidopsis transcriptome analysis. *Plant physiology*, *137*(2), 588-601. doi:10.1104/pp.104.051300

Allocco, D. J., Kohane, I. S., & Butte, A. J. (2004). Quantifying the relationship between co-expression, co-regulation and gene function. *BMC bioinformatics*, *5*, 18. doi:10.1186/1471-2105-5-18

Andreopoulos, B., An, A., Wang, X., & Schroeder, M. (2009). A roadmap of clustering algorithms: finding a match for a biomedical application. *Brief Bioinform*, *10*(3), 297-314. doi:10.1093/bib/bbn058

Apel, K., & Hirt, H. (2004). Reactive oxygen species: metabolism, oxidative stress, and signal transduction. *Annual review of plant biology*, *55*, 373-99. Annual Reviews. doi:10.1146/annurev.arplant.55.031903.141701

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000a). Gene ontology:

tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, *25*(1), 25-9. doi:10.1038/75556

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000b). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, *25*(1), 25-9. doi:10.1038/75556

Avanci, N. C., Luche, D. D., Goldman, G. H., & Goldman, M. H. S. (2010). Jasmonates are phytohormones with multiple functions, including plant defense and reproduction. *Genetics and molecular research🐦: GMR*, *9*(1), 484-505. doi:10.4238/vol9-1gmr754

Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M., & Teichmann, S. A. (2004). Structure and evolution of transcriptional regulatory networks. *Current opinion in structural biology*, *14*(3), 283-91. doi:10.1016/j.sbi.2004.05.004

Bairoch, A. (2000). The ENZYME database in 2000. *Nucleic Acids Research*, *28*(1), 304-305. doi:10.1093/nar/28.1.304

Balbi, V., & Devoto, A. (2008). Jasmonate signalling network in Arabidopsis thaliana: crucial regulatory nodes and new physiological scenarios. *The New phytologist*, *177*(2), 301-18. doi:10.1111/j.1469-8137.2007.02292.x

Barabási, A.-L., & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature reviews. Genetics*, *5*(2), 101-13. doi:10.1038/nrg1272

Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R., & Califano, A. (2005). Reverse engineering of regulatory networks in human B cells. *Nature genetics*, *37*(4), 382-90. doi:10.1038/ng1532

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media* (pp. 361–362).

Batagelj, V., & Mrvar, A. (1998). Pajek-program for large network analysis. *Connections*, *21*(2), 47–57.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., et al. (2004). The Pfam protein families database. *Nucleic acids research*, *32*(suppl 1), D138. Oxford Univ Press.

Ben-Dor, A., Shamir, R., & Yakhini, Z. (2004). Clustering gene expression patterns. *Journal of computational biology🐦: a journal of computational molecular cell biology*, *6*(3-4), 281-97. Mary Ann Liebert, Inc. doi:10.1089/106652799318274

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. (2008). GenBank. *Nucleic acids research*, *36*(Database issue), D25-30. doi:10.1093/nar/gkm929

Bhat, P., Yang, H., Bogre, L., Devoto, A., & Paccanaro, A. (2011). Computational selection of transcriptomics experiments improves guilt-by-association analyses. *submitted*.

Bishop, C. (2007). *Pattern Recognition and Machine Learning (Information Science and Statistics)* (1st ed. 20 ed.). Springer.

Boeckmann, B. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*, *31*(1), 365-370. doi:10.1093/nar/gkg095

Bolstad, B.M., Irizarry, R. ., Astrand, M., & Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, *19*(2), 185-193. doi:10.1093/bioinformatics/19.2.185

Bolstad, B., Collin, F., Brettschneider, J., Simpson, K., Cope, L., Irizarry, R., Speed, T. P., et al. (2005). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. (R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, & S. Dudoit, Eds.) (pp. 33-47). New York: Springer-Verlag. doi:10.1007/0-387-29362-0

Brazma, A. (2003). ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research*, *31*(1), 68-71. doi:10.1093/nar/gkg091

Bruggeman, F. J., & Westerhoff, H. V. (2007). The nature of systems biology. *Trends in microbiology*, *15*(1), 45-50. doi:10.1016/j.tim.2006.11.003

Califano, A, Stolovitzky, G., & Tu, Y. (2000). Analysis of gene expression microarrays for phenotype classification. *Proceedings / ... International*

Conference on Intelligent Systems for Molecular Biology🐦; ISMB. International Conference on Intelligent Systems for Molecular Biology, *8*, 75-85.

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., et al. (2004). The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic acids research*, *32*(Database issue), D262-6. doi:10.1093/nar/gkh021

Casal, J. J., & Mazzella, M. A. (1998). Conditional Synergism between Cryptochrome 1🌀and Phytochrome B Is Shown by the Analysis of phyA, phyB, and hy4 Simple, Double, and Triple Mutants in Arabidopsis. *Plant Physiol.*, *118*(1), 19-25. doi:10.1104/pp.118.1.19

Cheng, Y., & Church, G. M. (2000). Biclustering of expression data. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology🐦; ISMB. International Conference on Intelligent Systems for Molecular Biology*, *8*, 93-103.

Cherry, J. (1998a). SGD: Saccharomyces Genome Database. *Nucleic Acids Research*, *26*(1), 73-79. doi:10.1093/nar/26.1.73

Cherry, J. (1998b). SGD: Saccharomyces Genome Database. *Nucleic Acids Research*, *26*(1), 73-79. doi:10.1093/nar/26.1.73

Clare, A., & King, R. D. (2002). How well do we understand the clusters found in microarray data? *In Silico Biology*, *2*(4), 511-522.

Craigon, D. J., James, N., Okyere, J., Higgins, J., Jotham, J., & May, S. (2004). NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. *Nucleic acids research*, *32*(Database issue), D575-7. doi:10.1093/nar/gkh133

Crowe, M. L., Serizet, C., Thareau, V., Aubourg, S., Rouzé, P., Hilson, P., Beynon, J., et al. (2003). CATMA: a complete Arabidopsis GST database. *Nucleic acids research*, *31*(1), 156. Oxford Univ Press.

D'haeseleer, P. (2005). How does gene expression clustering work? *Nature biotechnology*, *23*(12), 1499-501. Nature Publishing Group. doi:10.1038/nbt1205-1499

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, *14*(9), 755-763. doi:10.1093/bioinformatics/14.9.755

Eden, E., Navon, R., Steinfeld, I., Lipson, D., & Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC bioinformatics*, *10*(1), 48. BioMed Central Ltd.

Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucl. Acids Res.*, *30*(1), 207-210. doi:10.1093/nar/30.1.207

Edwards, A. M., Kus, B., Jansen, R., Greenbaum, D., Greenblatt, J., & Gerstein, M. (2002). Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends in genetics🐦: TIG*, *18*(10), 529-36.

Eisen, M. B. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, *95*(25), 14863-14868. doi:10.1073/pnas.95.25.14863

Enright, A. J., & Ouzounis, C. A. (2001). BioLayout—an automatic graph layout algorithm for similarity visualization. *Bioinformatics*, *17*(9), 853. Oxford Univ Press.

Faith, J. J., Driscoll, M. E., Fusaro, V. A., Cosgrove, E. J., Hayete, B., Juhn, F. S., Schneider, S. J., et al. (2007). Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucl. Acids Res.*, gkm815. doi:10.1093/nar/gkm815

Fiehn, O. (2002). Metabolomics – the link between genotypes and phenotypes. *Plant Molecular Biology*, *48*(1), 155-171. Springer Netherlands. doi:10.1023/A:1013713905833

Fraser, A. G., & Marcotte, E. M. (2004). A probabilistic view of gene function. *Nature genetics*, *36*(6), 559-64. doi:10.1038/ng1370

Fujita, M., Fujita, Y., Noutoshi, Y., Takahashi, F., Narusaka, Y., Yamaguchi-Shinozaki, K., & Shinozaki, K. (2006). Crosstalk between abiotic and biotic stress responses: a current view from the points of convergence in the stress signaling networks. *Current opinion in plant biology*, *9*(4), 436-42. doi:10.1016/j.pbi.2006.05.014

Gavin, A. C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, *415*(6868), 141–147. Nature Publishing Group.

Gelbart, W. M., Crosby, M., Matthews, B., Rindone, W. P., Chillemi, J., Russo Twombly, S., Emmert, D., et al. (1997). FlyBase: a Drosophila database. The FlyBase consortium. *Nucleic acids research*, *25*(1), 63-6.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, *5*(10), R80. doi:10.1186/gb-2004-5-10-r80

Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbel, J. O., Emanuelsson, O., et al. (2007). What is a gene, post-ENCODE? History and updated definition. *Genome research*, *17*(6), 669-81. doi:10.1101/gr.6339607

Getz, G., Levine, E., & Domany, E. (2000). Coupled two-way clustering analysis of gene microarray data. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(22), 12079-84. doi:10.1073/pnas.210134797

Gibbons, F. D., & Roth, F. P. (2002). Judging the Quality of Gene Expression-Based Clustering Methods Using Gene Annotation. *Genome Research*, *12*(10), 1574-1581. doi:10.1101/gr.397002

Gygi, S. P., Rochon, Y., Franza, B. R., & Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. *Molecular and cellular biology*, *19*(3), 1720-30.

Hammond-Kosack, K. E., & Jones, J. D. (1996). Resistance gene-dependent plant defense responses. *The Plant cell*, *8*(10), 1773-91.

Heyer, L. J., Kruglyak, S., & Yooseph, S. (1999). Exploring expression data: identification and analysis of coexpressed genes. *Genome research*, *9*(11), 1106-15.

Higuchi, R., Fockler, C., Dollinger, G., & Watson, R. (1993). Kinetic PCR Analysis: Real-time Monitoring of DNA Amplification Reactions. *Bio/Technology*, *11*(9), 1026-1030. Nature Publishing Company. doi:10.1038/nbt0993-1026

Horan, K., Jang, C., Bailey-Serres, J., Mittler, R., Shelton, C., Harper, J. F., Zhu, J.-K., et al. (2008). Annotating Genes of Known and Unknown Function by Large-Scale Coexpression Analysis. *Plant Physiol.*, *147*(1), 41-57. doi:10.1104/pp.108.117366

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, *4*(1), 44-57. doi:10.1038/nprot.2008.211

Hughes, T R, Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., et al. (2000). Functional discovery via a compendium of expression profiles. *Cell*, *102*(1), 109-126. doi:10.1016/S0092-8674(00)00015-5

Huynen, M. A., Snel, B., von Mering, C., & Bork, P. (2003). Function prediction and protein networks. *Current opinion in cell biology*, *15*(2), 191-8.

Ideker, T, Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., et al. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science (New York, N.Y.)*, *292*(5518), 929-34. doi:10.1126/science.292.5518.929

Ihmels, J., Bergmann, S., & Barkai, N. (2004). Defining transcription modules using large-scale gene expression data. *Bioinformatics (Oxford, England)*, *20*(13), 1993-2003. doi:10.1093/bioinformatics/bth166

Irizarry, Rafael A, Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., & Speed, T. P. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic acids research*, *31*(4), e15.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., & Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(8), 4569-74. doi:10.1073/pnas.061034498

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., & Barabási, A. L. (2000). The large-scale organization of metabolic networks. *Nature*, *407*(6804), 651–654. Nature Publishing Group.

Jones, C. E., Brown, A. L., & Baumann, U. (2007). Estimating the annotation error rate of curated GO database sequence annotations. *BMC bioinformatics*, *8*(1), 170. BioMed Central Ltd. doi:10.1186/1471-2105-8-170

Kanehisa, M. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, *28*(1), 27-30. doi:10.1093/nar/28.1.27

Kaniak, A., Xue, Z., Macool, D., Kim, J.-H., & Johnston, M. (2004). Regulatory Network Connecting Two Glucose Signal Transduction Pathways in

Saccharomyces cerevisiae. *Eukaryotic Cell*, *3*(1), 221-231. doi:10.1128/EC.3.1.221-231.2004

Keseler, I. M., Collado-Vides, J., Gama-Castro, S., Ingraham, J., Paley, S., Paulsen, I. T., Peralta-Gil, M., et al. (2005). EcoCyc: a comprehensive database resource for Escherichia coli. *Nucleic acids research*, *33*(Database issue), D334-7. doi:10.1093/nar/gki108

Kevil, C. G., Walsh, L., Laroux, F. S., Kalogeris, T., Grisham, M. B., & Alexander, J. S. (1997). An improved, rapid Northern protocol. *Biochemical and biophysical research communications*, *238*(2), 277-9. doi:10.1006/bbrc.1997.7284

Khan, J., Wei, J. S., Ringnér, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., et al. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature medicine*, *7*(6), 673-9. doi:10.1038/89044

Kim, S K, Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J. M., Eizinger, A., et al. (2001). A gene expression map for Caenorhabditis elegans. *Science (New York, N.Y.)*, *293*(5537), 2087-92. doi:10.1126/science.1061603

Kitano, H. (2002). Systems biology: a brief overview. *Science (New York, N.Y.)*, *295*(5560), 1662-4. doi:10.1126/science.1069492

Klebanov, L., & Yakovlev, A. (2007). How high is the level of technical noise in microarray data? *Biology Direct*, *2*, 9-9. doi:10.1186/1745-6150-2-9

Lan, N., Montelione, G. T., & Gerstein, M. (2003). Ontologies for proteomics: towards a systematic definition of structure and function that scales to the genome level. *Current opinion in chemical biology*, *7*(1), 44-54.

Lee, I. (2011). Probabilistic functional gene societies. *Progress in biophysics and molecular biology*, *106*(2), 435-42. doi:10.1016/j.pbiomolbio.2011.01.003

Lee, W.-P., & Tzou, W.-S. (2009). Computational methods for discovering gene networks from expression data. *Briefings in bioinformatics*, *10*(4), 408-23. doi:10.1093/bib/bbp028

Lein, E. S., Hawrylycz, M. J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A. F., et al. (2007). Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, *445*(7124), 168-76. doi:10.1038/nature05453

Lenoir, T., & Giannella, E. (2006). The emergence and diffusion of DNA microarray technology. *Journal of biomedical discovery and collaboration*, *1*, 11. doi:10.1186/1747-5333-1-11

Leung, Y. F., & Cavalieri, D. (2003). Fundamentals of cDNA microarray data analysis. *Trends in Genetics*, *19*(11), 649-659. doi:10.1016/j.tig.2003.09.015

Long, T. A., Brady, S. M., & Benfey, P. N. (2008). Systems approaches to identifying gene regulatory networks in plants. *Annual review of cell and developmental biology*, *24*, 81-103. doi:10.1146/annurev.cellbio.24.110707.175408

Ma, S., Gong, Q., & Bohnert, H. J. (2007). An Arabidopsis gene network based on the graphical Gaussian model. *Genome Research*, *17*(11), 1614-1625. doi:10.1101/gr.6911207

Madeira, S. C., & Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics / IEEE, ACM*, *1*(1), 24-45. doi:10.1109/TCBB.2004.2

Maere, S., Heymans, K., & Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics (Oxford, England)*, *21*(16), 3448-9. doi:10.1093/bioinformatics/bti551

Manfield, I. W., Jen, C.-H., Pinney, J. W., Michalopoulos, I., Bradford, J. R., Gilmartin, P. M., & Westhead, D. R. (2006). Arabidopsis Co-expression Tool (ACT): web server tools for microarray-based gene expression analysis. *Nucleic acids research*, *34*(Web Server issue), W504-9. doi:10.1093/nar/gkl204

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., & Califano, A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, *7 Suppl 1*(Suppl 1), S7. doi:10.1186/1471-2105-7-S1-S7

Marshall, E. (2004). Getting the noise out of gene arrays. *Science (New York, N.Y.)*, *306*(5696), 630-1. doi:10.1126/science.306.5696.630

Mering, C. v., Martijn, H., Daniel, J., Steffen, S., Peer, B., & Berend, S. (2003). STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research*, *31*(1), 258-261. doi:10.1093/nar/gkg034

Mewes, H W, Amid, C., Arnold, R., Frishman, D., Güldener, U., Mannhaupt, G., Münsterkötter, M., et al. (2004). MIPS: analysis and annotation of proteins from whole genomes. *Nucleic acids research*, *32*(Database issue), D41-4. doi:10.1093/nar/gkh092

Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., & Morris, Q. (2008). GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome biology*, *9 Suppl 1*(Suppl 1), S4. doi:10.1186/gb-2008-9-s1-s4

Mulder, N. J. (2003). The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Research*, *31*(1), 315-318. doi:10.1093/nar/gkg046

Mutwil, M., Obro, J., Willats, W. G. T., & Persson, S. (2008). GeneCAT--novel webtools that combine BLAST and co-expression analyses. *Nucleic acids research*, *36*(Web Server issue), W320-6. doi:10.1093/nar/gkn292

Neter, J., Wasserman, W., Kutner, M. H., & Li, W. (1996). *Applied linear statistical models*. Irwin.

Noble, D. (2002). The rise of computational biology. *Nature reviews. Molecular cell biology*, *3*(6), 459-63. Nature Publishing Group. doi:10.1038/nrm810

Obayashi, T., Hayashi, S., Saeki, M., Ohta, H., & Kinoshita, K. (2009a). ATTED-II provides coexpressed gene networks for Arabidopsis. *Nucleic Acids Research*, *37*(Database issue), D987-991. doi:10.1093/nar/gkn807

Obayashi, T., Hayashi, S., Saeki, M., Ohta, H., & Kinoshita, K. (2009b). ATTED-II provides coexpressed gene networks for Arabidopsis. *Nucleic acids research*, *37*(Database issue), D987-91. doi:10.1093/nar/gkn807

Pandey, G. (2006). Computational Approaches for Protein Function Prediction: A Survey. *2006 [CITATION] Computational Approaches for Protein Function Prediction: A SurveyG Pandey,V Kumar… [CITATION] Computational Approaches for Protein Function Prediction: A SurveyG Pandey, V Kumar….*

Priness, I., Maimon, O., & Ben-Gal, I. (2007). Evaluation of gene-expression clustering via mutual information distance measure. *BMC bioinformatics*, *8*(1), 111. doi:10.1186/1471-2105-8-111

Qin, L.-X., Beyer, R. P., Hudson, F. N., Linford, N. J., Morris, D. E., & Kerr, K. F. (2006). Evaluation of methods for oligonucleotide array data via quantitative real-time PCR. *BMC bioinformatics*, *7*(1), 23. BioMed Central Ltd. doi:10.1186/1471-2105-7-23

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature genetics*, *32 Suppl*, 496-501. doi:10.1038/ng1032

Quackenbush, J. (2003). GENOMICS: Microarrays--Guilt by Association. *Science*, *302*(5643), 240-241. doi:10.1126/science.1090887

Ramanathan, M. (2001). *Interrelated two-way clustering: an unsupervised approach for gene expression data analysis*. *Proceedings 2nd Annual IEEE International Symposium on Bioinformatics and Bioengineering (BIBE 2001)* (pp. 41-48). IEEE Comput. Soc. doi:10.1109/BIBE.2001.974410

Rhee, S. Y., Beavis, W., Berardini, T. Z., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., et al. (2003). The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucl. Acids Res.*, *31*(1), 224-228. doi:10.1093/nar/gkg076

Ruepp, Andreas, Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., et al. (2004). The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic acids research*, *32*(18), 5539-45. doi:10.1093/nar/gkh894

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003a). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, *13*(11), 2498-504. doi:10.1101/gr.1239303

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003b). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, *13*(11), 2498-504. doi:10.1101/gr.1239303

Shendure, J. (2008). The beginning of the end for microarrays? *Nature methods*, *5*(7), 585-7. Nature Publishing Group. doi:10.1038/nmeth0708-585

Soinov, L. A., Krestyaninova, M. A., & Brazma, A. (2003). Towards reconstruction of gene networks from expression data by supervised learning. *Genome biology*, *4*(1), R6.

Spackman, K. (1997). SNOMED RT: a reference terminology for health care. *Proceedings of the AMIA* ….

Srinivasasainagendra, V., Page, G. P., Mehta, T., Coulibaly, I., & Loraine, A. E. (2008). CressExpress: a tool for large-scale mining of expression data from Arabidopsis. *Plant physiology*, *147*(3), 1004-16. doi:10.1104/pp.107.115535

Steinhauser, D., Usadel, B., Luedemann, A., Thimm, O., & Kopka, J. (2004). CSB.DB: a comprehensive systems-biology database. *Bioinformatics (Oxford, England)*, *20*(18), 3647-51. doi:10.1093/bioinformatics/bth398

Stekel, D. (2003). *Microarray bioinformatics*. Cambridge Univ Pr.

Stuart, Joshua M., Segal, E., Koller, D., & Kim, S. K. (2003). A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science*, *302*(5643), 249-255. doi:10.1126/science.1087447

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., et al. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, *39*(Database issue), D561-8. doi:10.1093/nar/gkq973

Tanay, A., Sharan, R., & Shamir, R. (2002). Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, *18*(suppl_1), S136-144. doi:10.1093/bioinformatics/18.suppl_1.S136

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., & Church, G. M. (1999). Systematic determination of genetic network architecture. *Nature Genetics*, *22*(3), 281-285. doi:10.1038/10343

The Gene Ontology Consortium. (2010). The Gene Ontology in 2010: extensions and refinements. *Nucleic acids research*, *38*(Database issue), D331-5. doi:10.1093/nar/gkp1018

Tibshirani, R., Hastie, T., Eisen, M., Ross, D., Botstein, D., & Brown, P. (1999). Clustering methods for the analysis of dna microarray data.

Todd, A. E., Orengo, C. A., & Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective. *Journal of molecular biology*, *307*(4), 1113-43. doi:10.1006/jmbi.2001.4513

Toufighi, K., Brady, S. M., Austin, R., Ly, E., & Provart, N. J. (2005). The Botany Array Resource: e-Northerns, Expression Angling, and promoter analyses. *The Plant Journal*, *43*(1), 153–163. Wiley Online Library.

Törönen, P. (1999). Analysis of gene expression data using self-organizing maps. *FEBS Letters*, *451*(2), 142-146. doi:10.1016/S0014-5793(99)00524-4

USADEL, B., OBAYASHI, T., MUTWIL, M., GIORGI, F. M., BASSEL, G. W., TANIMOTO, M., CHOW, A., et al. (2009). Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant, Cell & Environment*, *32*(12), 1633-1651. doi:10.1111/j.1365-3040.2009.02040.x

Vazquez, A., Flammini, A., Maritan, A., & Vespignani, A. (2003). Global protein function prediction from protein-protein interaction networks. *Nature biotechnology*, *21*(6), 697-700. Nature Publishing Group. doi:10.1038/nbt825

Wang, H., Collins, S., Krogan, N., & Koller, D. (2007). Identifying Protein Complexes in Saccharomyces cerevisiae.

Wang, J., Delabie, J., Aasheim, H., Smeland, E., & Myklebost, O. (2002). Clustering of the SOM easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study. *BMC bioinformatics*, *3*, 36.

Wang, Y., Joshi, T., Zhang, X.-S., Xu, D., & Chen, L. (2006). Inferring gene regulatory networks from multiple microarray datasets. *Bioinformatics (Oxford, England)*, *22*(19), 2413-20. doi:10.1093/bioinformatics/btl396

Wen, X. (1998). Large-scale temporal gene expression mapping of central nervous system development. *Proceedings of the National Academy of Sciences*, *95*(1), 334-339. doi:10.1073/pnas.95.1.334

Wilkinson, S., & Davies, W. J. (2010). Drought, ozone, ABA and ethylene: new insights from cell to plant to community. *Plant, cell & environment*, *33*(4), 510-25. doi:10.1111/j.1365-3040.2009.02052.x

Wolfe, C. J., Kohane, I. S., & Butte, A. J. (2005). Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC bioinformatics*, *6*, 227. doi:10.1186/1471-2105-6-227

Wolfsberg, T. G., Gabrielian, A. E., Campbell, M. J., Cho, R. J., Spouge, J. L., & Landsman, D. (1999). Candidate regulatory sequence elements for cell cycle-dependent transcription in Saccharomyces cerevisiae. *Genome research*, *9*(8), 775-92.

Wu, L. F., Hughes, T. R., Davierwala, A. P., Robinson, M. D., Stoughton, R., & Altschuler, S. J. (2002). Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters. *Nature Genetics*, *31*(3), 255-265. doi:10.1038/ng906

Wu, Z., Irizarry, R. A., Gentleman, R., Martinez-Murillo, F., & Spencer, F. (2004). A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association*, *99*(468), 909-917. American Statistical Association. doi:10.1198/016214504000000683

Yeung, K. Y., Medvedovic, M., & Bumgarner, R. E. (2004). From co-expression to co-regulation: how many microarray experiments do we need? *Genome biology*, *5*(7), R48. doi:10.1186/gb-2004-5-7-r48

Yu, P. (2003). *Enhanced biclustering on expression data. Third IEEE Symposium on Bioinformatics and Bioengineering, 2003. Proceedings.* (pp. 321-327). IEEE Comput. Soc. doi:10.1109/BIBE.2003.1188969

Zheng, Q., & Wang, X. J. (2008). GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic acids research*, *36*(suppl 2), W358. Oxford Univ Press.

Zhou, X. J., Kao, M.-C. J., Huang, H., Wong, A., Nunez-Iglesias, J., Primig, M., Aparicio, O. M., et al. (2005). Functional annotation and network reconstruction through cross-platform integration of microarray data. *Nature biotechnology*, *23*(2), 238-43. Nature Publishing Group. doi:10.1038/nbt1058

Zien, A., Fluck, J., Zimmer, R., & Lengauer, T. (2003). Microarrays: how many do you need? *Journal of computational biology : a journal of computational molecular cell biology*, *10*(3-4), 653-67. doi:10.1089/10665270360688246

Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L., & Gruissem, W. (2004). GENEVESTIGATOR. Arabidopsis Microarray Database and Analysis Toolbox. *Plant Physiol.*, *136*(1), 2621-2632. doi:10.1104/pp.104.046367

van Hengel, A. J., Barber, C., & Roberts, K. (2004). The expression patterns of arabinogalactan-protein AtAGP30 and GLABRA2 reveal a role for abscisic acid in the early stages of root epidermal patterning. *The Plant journal : for cell and molecular biology*, *39*(1), 70-83. doi:10.1111/j.1365-313X.2004.02104.x