Pathway based microarray analysis based on multi-membership gene regulation

Stelios Pavlidis

A thesis submitted for the degree of
Doctor of Philosophy

Brunel University

February 2011

School of Information Systems, Computing and Mathematics

Abstract

Recent developments in automation and novel experimental techniques have led to the accumulation of vast amounts of biological data and the emergence of numerous databases to store the wealth of information. Consequentially, bioinformatics have drawn considerable attention, accompanied by the development of a plethora of tools for the analysis of biological data.

DNA microarrays constitute a prominent example of a high-throughput experimental technique that has required substantial contribution of bioinformatics tools. Following its popularity there is an on-going effort to integrate gene expression with other types of data in a common analytical approach. Pathway based microarray analysis seeks to facilitate microarray data in conjunction with biochemical pathway data and look for a coordinated change in the expression of genes constituting a pathway.

However, it has been observed that genes in a pathway may show variable expression, with some appearing activated while others repressed. This thesis aims to add some contribution to pathway based microarray analysis and assist the interpretation of such observations, based on the fact that in all organisms a substantial number of genes take part in more than one biochemical pathway. It explores the hypothesis that the expression of such genes represents a net effect of their contribution to all their constituent pathways, applying statistical and data mining approaches. A heuristic search methodology is proposed to manipulate the pathway contribution of genes to follow underlying trends and interpret microarray results centred on pathway behaviour. The methodology is further refined to account for distinct genes encoding enzymes that catalyse the same reaction, and applied to modules, shorter chains of reactions forming sub-networks within pathways. Results based on various datasets are discussed, showing that the methodology is promising and may assist a biologist to decipher the biochemical state of an organism, in experiments where pathways exhibit variable expression.

Declaration

I hereby declare that the research presented in this thesis is my own work except where otherwise stated, and has not been submitted for any other degree.

Stelios Pavlidis

Acknowledgements

Firstly I would like to thank my supervisors Annette Payne and Stephen Swift for their guidance, support, encouragement and patience throughout my PhD. Their contribution has been of great value for the completion of this work. I would also like to express my deep gratitude to my family for their extraordinary love and support all these years.

Importantly, I would like to thank my friends and most of all Evdoxia for their understanding and support.

Finally, I acknowledge the contribution of the Centre for Intelligent Data Analysis (CIDA) lab and its members and the financial support received from the studentship scheme of the School of Information Systems, Computing and Mathematics, Brunel University, UK.

Supporting Publications

The following publications have resulted from the research presented in this thesis:

Published:

1. Pavlidis, S., Payne, A. & Swift, S. (2011). Multi-membership gene regulation in pathway based microarray analysis. *BMC Algorithms for molecular biology*. 6:22.

2. Pavlidis, S., Swift, S. & Payne, A. (2011). A comparative analysis of single- and multi-membership gene expression, based on association rule mining and differential expression frequencies. In *Proceedings of the annual workshop on Intelligent Data Analysis in bioMedicine And Pharmacology (IDAMAP).*

3. Pavlidis, S., Swift, S. & Payne, A. (2010). Pathway based microarray analysis, facilitating enzyme compounds and cascade events. In *Proceedings of the annual workshop on Intelligent Data Analysis in bioMedicine And Pharmacology (IDAMAP).*

4. Pavlidis, S., Payne, A. & Swift, S. (2008). An Improved Methodology for Pathway Based Microarray Analysis Based on Identification of Individual Pathways Responsible for Gene Regulation. In *Proceedings of the annual workshop on Intelligent Data Analysis in bioMedicine And Pharmacology (IDAMAP).*

Under review:

5. An extended version of item 3 has been invited for publication in the Methods of Information in Medicine journal and is under review at the time of completion of this thesis.

Item 1 is based on the work presented in Chapters 4 and 5. Item 2 presents findings described in chapter 3 of the thesis. Item 3 results from the research presented in Chapter 6 while item 4 from preliminary work for this thesis and the research described in Chapter 4.

Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Overview

In the last couple of decades the face of biological research has undergone substantial transformation from both a qualitative and a quantitative perspective. The cumulative progress in a wide range of scientific fields including computer science, mathematics, physics and chemistry along with a rapid increase in automation have made it possible to develop novel, high throughput experimental techniques for the study of biological phenomena. It is now possible to gain insights into aspects of living structures and functions which have never before been accessible to us. At the same time we are now able to produce huge amounts of very diverse biological datasets. Bioinformatics sprung into existence and gained wide spread popularity in the 90s due to the need to find efficient ways to store, manipulate and analyse the newly acquired data.

DNA Microarrays have played a major role in the area of bioinformatics research. This experimental technique allows us to observe the expressional behaviour of entire genomes in a single experiment, by measuring the relative abundance of RNA molecules corresponding to individual genes, between conditions of interest. For over a decade there have been numerous publications dealing with one or other aspect of microarray data analysis and following the trend for data integration, there have been substantial efforts to incorporate different types of biological knowledge in the analytical process.

Pathway based microarray analysis is an attempt to exploit gene expression data to gain insight into the state of a cell or an organism from a biochemical point of view. Wet lab biological research has led to the identification of biochemical chains of reactions that take place in different organisms, to allow them to facilitate nutrients available in their environment and support their survival and development. This network of reactions is quite complicated and far from complete. Thus it has been organised into smaller units, the so called pathways, each one responsible for a defined gradual process of transformations of certain molecules into different chemical compounds required by the organism. Pathway based microarray analysis examines the expression of pre-defined sets of genes, which encode the proteins participating in each such chain of reactions, in order to identify the impact of different conditions on biochemical activity. That is, it tries to answer the question, which pathways need be activated and which de-activated in response to various stimuli, developmental stages and so on.

The central theme of this work is the analysis of biochemical pathway behaviour based on the expressional behaviour of genes forming them, facilitating microarray data. It aims to identify the true pathways to which gene members of more than one pathway contribute and subsequently identify the state of activity of pathways, centred on the behaviour of their constituent genes. This introductory chapter presents an overview of the motivation, content and contribution of the thesis.

1.2 Thesis outline

Chapter 2 provides an overview of the background behind the work in this thesis. It introduces the research area in some detail along with the relevant technological advances that have led to its emergence. It presents some basic biological concepts that are necessary for understanding the motivation behind this work and the analytical approaches applied. This includes a discussion of the basic features of biological systems and their main components, that is, genes and proteins along with their functional associations. A discussion of biochemical pathways, of different types and their role, as well as their importance in biological research from a theoretical and practical point of view, constitutes a major part.

Additionally, the field of bioinformatics and computational biology is explored, including a brief history and presentation of the main experimental technologies laying the foundations for its emergence and recent popularity. Given that heuristics are facilitated in the methodologies presented in this thesis, a section giving an overview of the area and some important relevant computational techniques is included.

There is an extensive part on the rationale, experimental procedure, data processing, applications and importance of DNA Microarrays technology. Biological databases are also discussed and the different types introduced with special emphasis on biochemical pathway and microarray data databases, which are relative to this research.

Naturally, this is followed by a discussion of pathway based microarray analysis the relevant experimental methodologies and computational tools it encompasses. Following the presentation of available software tools and how they approach the task in hand in terms of visualisation and analytical efforts the challenges faced by this approach are identified. The chapter concludes with a brief discussion of the importance of refining the methodology and the contribution that this work aims to achieve.

Chapter 3 deals with the main hypothesis on which the thesis is largely reliant and seeks to provide supportive evidence. It identifies the fact that a substantial number of genes in all organisms constitute members of a number of distinct pathways. There is an extensive comparative analysis of the behaviour of such genes as opposed to the expression of genes that participate in one unique pathway. A number of different approaches are implemented, including some statistical, as well as data mining analysis of microarray datasets, such as correlation analysis and association rule mining. Naturally, the methodologies and their implementation are presented where necessary.

Thus, this is an exploratory and descriptive chapter facilitating computational analysis to explore gene expression behaviour based on large microarray datasets. It concludes that there is some evidence that gene members of many pathways do exhibit different behaviour than genes that constitute members of one and only biochemical pathway.

Following these observations, chapter 4 proposes an analytical methodology that seeks to identify the state of activity of distinct pathways, centred on the behaviour of genes forming them. There is some further discussion of the motivation behind the specific computational approach, which facilitates a hill climbing algorithm to allocate genes to pathways. The allocation seeks to discover the true pathways whose regulation by the biological system requires genes to exhibit the expression observed in a microarray experiment. The algorithm is explained and the results of its application to some *Escherichia coli* and *Saccharomyces cerevisiae* are presented.

Given the nature of hill climbing and mainly its tendency to get stack in local optima, chapter 5 examines alternative heuristic search approaches. In particular it presents the implementation of a simulated annealing and a genetic algorithm approach, to search for the best gene to pathway allocation, as discussed above. Besides examining the fitness reached by each method, since it has no biological meaning, additional metrics of similarity are facilitated. Namely, a similarity measure based on the hamming distance metric and a method to extract the probability of observing two allocations of a given similarity or larger are presented. Additionally, there is an implementation of the fuzzy adjusted rand index measure. Together these measures allow the comparative analysis of results from different perspectives, which is discussed along with produced allocations.

In chapter 6 a slightly different methodology is discussed and implemented. The motivation behind it is mainly biologically driven. Since, proteins are the functional molecules responsible for enzymic reactions, rather than genes, the search approach in this chapter is altered in way that it becomes centred on enzymic positions in the chain of biochemical reactions. A main point here is that different genes often encode enzymes catalysing the same step in a pathway, while at the same time a particular gene may participate not only in different pathways but also in different steps of the same pathway. Hence, we examine the state of each such enzyme/step based on all the genes involved. Consequentially, we then examine the state of positions rather than just looking into lists of genes.

Chapter 7 summarises the research performed in this thesis and the obtained results, along with some critical discussion of its outcomes. It identifies shortcomings and possible improvements. The chapter concludes with some discussion of potential future work directions.

1.3 Thesis contributions

The key contributions of this thesis are below:

1. It identified substantial variability in the expression of genes forming biochemical pathways, in a small but considerable proportion (~15%) of examined microarray datasets (Chapter 3)
2. It identified differences in the expressional behaviour of multi- and single-membership genes, suggesting that the expression of the former group represents a net effect of their contribution to all their constituent pathways (Chapter 3)
3. It proposed a novel methodology to identify the state of activity of pathways, based on the expression of multi-membership genes, implementing a heuristic search approach (Chapter 4)
4. It showed that a hill climbing, a simulated annealing and a genetic algorithm search approach exhibit similar performance for the examined datasets. This is an interesting observation given the differences of their nature (Chapter 5)
5. It proposed and implemented a measure to estimate the similarity between genes to pathways allocations, based on discretised gene expression data and an approach to estimate the probability of observing a certain level of similarity or greater, purely by chance (Chapter 5)
6. It proposed a methodology to identify the activation state of pathways and modules centred on the activity of enzymes responsible for distinct steps in the process. The expression of all genes corresponding to each step is used as evidence for this activity (Chapter 6)

Chapter 2: Background

2.1 Introduction

The research presented in this thesis is in the field of systems biology and more precisely bioinformatics, facilitating heuristic search approaches for data analysis. These are relevantly novel areas of research that have drawn considerable interest, mostly in the last two decades, partly due to the large volume of accumulated biological knowledge and partly due to the availability of sophisticated technology and computer processing power that we now have at our disposal, in the effort to elucidate life processes and solve important biological issues.

This chapter provides an introduction to some basic biological concepts that are necessary for understanding the motivation behind this work. Additionally it deals with the main aspects of computational analysis of biological data and experimental techniques that have emerged in recent years and are relevant to this research.

2.2 Basic biological concepts

The following is a brief introduction to some basic biological concepts and mechanisms that govern living organisms. These form the conceptual basis of this work and are necessary for the reader to understand the biological issues this thesis deals with.

2.2.1 DNA, proteins and their role

All living organisms are formed by one or more cells, often described as the basic functional unit of life. Each cell carries genetic material, the DNA that can be seen as the hard drive, the storage facility that carries the information needed in order for a cell to maintain life processes and survive. DNA stores every single instruction that allows the cell and consequentially the organism to grow and multiply.

DNA molecules constitute large polymers build-up of nucleotides, which in turn consist of sugar residues with covalently attached nitrogenous base and a phosphate group. Nucleotides are linked together by phosphodiester bonds in a linear fashion and it is the particular sequence of bases in the linear DNA molecule that contains the information for life maintenance. More precisely there are four types of bases, the purines adenine (A) and guanine (G), and the pyrimidines cytosine (C) and thymine (T), that constitute the genetic ‘alphabet’ and the particular order in which they are placed encodes the genetic information of an organism (Strachan & Read 2004).

Table 2.1 The Genetic Code. The code is degenerate as a number of codons specify the same amino acid. In total 64 combinations of 3 bases specify all amino acids including the start and stop signals, for the process of transcription.

START	AUG	STOP	UAA, UGA, UAG
Alanine	GCU, GCC, GCA, GCG	Leucine	UUA, UUG, CUU, CUC, CUA, CUG
Arginine	CGU, CGC, CGA, CGG, AGA, AGG	Lysine	AAA, AAG
Asparagine	AAU, AAC	Methionine	AUG
Aspartic acid	GAU, GAC	Phenylalanine	UUU, UUC
Cysteine	UGU, UGC	Proline	CCU, CCC, CCA, CCG
Glutamine	CAA, CAG	Serine	UCU, UCC, UCA, UCG, AGU, AGC
Glutamic acid	GAA, GAG	Threonine	ACU, ACC, ACA, ACG
Glycine	GGU, GGC, GGA, GGG	Tryptophan	UGG
Histidine	CAU, CAC	Tyrosine	UAU, UAC
Isoleucine	AUU, AUC, AUA	Valine	GUU, GUC, GUA, GUG

Similarly to DNA, proteins are large polymers build-up of a linear sequence of repeating units, the so called amino acids. All proteins in all species are constructed from the same set of only 20 amino acids. However, these building blocks exhibit remarkable diversity in terms of their chemical properties, such as hydrophobicity, polarity and acidity/basicity, and the particular order in which they are placed is the main factor conferring a protein its structural and by extension functional properties (Stryer & Tymoczko 2006). It is this order of amino acids that is in fact stored in the order of nitrogenous bases of genes scattered within DNA molecules. In brief, a sequence of three bases, termed codon, encodes a particular amino acid, and thus provides the cell with the knowledge required to produce the proteins it needs by orderly arranging amino acid monomers. Table 2.1 presents the standard genetic code shared by most organisms.

This brings us to what is known as the central dogma of biology, describing the flow of genetic information, as portrayed on Figure 2.1. According to the classical view of the central dogma of biology genetic information hardwired in DNA molecules is transcribed into transposable RNA molecules, which in turn serve as templates translated by ribosomes in the cytoplasm to produce polypeptide chains that fold into active protein molecules (Figure 2.2). It should be noted that in RNA thymine is replaced by uracil (U), hence the absence of T from the codons on table 2.1.

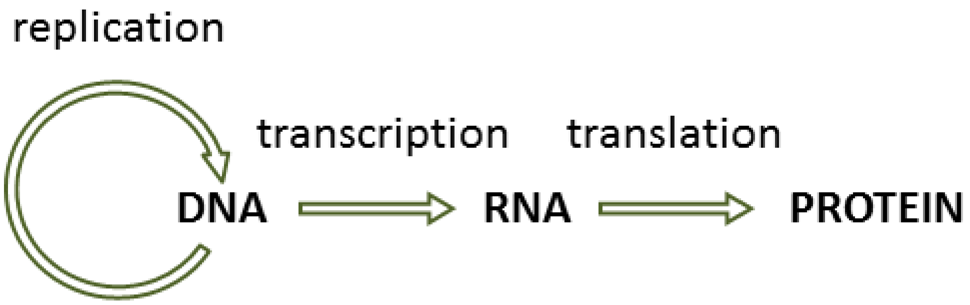


Figure 2.1 The central dogma of biology. Information stored in DNA can be copied onto another DNA molecule, or transferred to RNA which in turn can serve for protein synthesis.

Additionally, genetic information can be passed on to a new DNA molecule through the process of replication so that it can be passed on to a new cell. However, once genetic information has been turned into a protein it cannot be transferred back to DNA or another protein.

As will become apparent later on the process of DNA transcription into RNA molecules, which carry the message stored in a gene, so that it can be facilitated mainly for protein synthesis, is of great importance for microarray technology. In fact it is this sequence of events that constitutes gene expression. It is important to mention that while generally speaking microarray analysis is often termed gene expression analysis, to be precise we should note that it is the whole process starting from a gene and finishing with a functional protein that gene expression describes. This not only includes transcription but also the further processing of RNA and it's translation into a protein, as well as a number of post-translational modifications of the resulting molecule which are necessary for the production of a functional protein (Seo & Lee 2004).

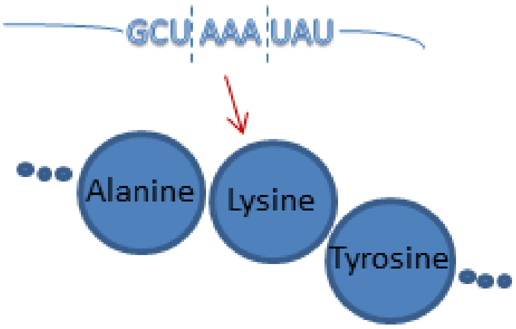


Figure 2.2 Translation. Each codon in the RNA molecule, resulting from the transcription of particular gene, specifies the amino acid that comes next in the respective protein. The codons are read one by one in the cells ribosomes and the transcription mechanism eventually produces a whole polypeptide chain.

Virtually every life process depends on proteins, which are the most abundant and functionally diverse molecules in any living organism. The vast majority of gene expression is dedicated to protein synthesis, which are the major functional end-point of DNA. Proteins can be composed of one or more polypeptides and account for the majority of the dry weight of a cell. Their name was derived from the Greek word *proteios*, meaning ‘of the first rank’ due to the wide range of important roles they have, including structural support, signalling, cell communication, transport and importantly catalysis which is essential for this work (Strachan & Read 2004).

2.2.2 Biochemical Pathways

All organisms are capable of carrying forward chemical transformations, facilitating nutrients available in their environment to make chemical building blocks, extracting and mediating the transformation of energy from one form to another, processes that are essential for their growth and the maintenance of life. Metabolism essentially refers to a linked series of chemical reactions that begins with a particular molecule and converts it into some other molecule or molecules in a strictly controlled fashion (Stryer & Tymoczko 2006).

Protein enzymes acting as catalysts are capable of specifically binding an extremely wide range of molecules, determining which one of a number of potential chemical reactions takes place. Proteins achieve this by accelerating the speed of reactions by factors of a million or more. A variety of such reactions are organised into multi step, synchronised sequences of events referred to as pathways (Harvey & Ferrier 2010).

Each pathway can be seen as a particular sequence of events, during which certain molecules are gradually modified to produce other molecules in order to accommodate the needs of the respective organism. At each step, the product of a reaction serves as the substrate for the next step of the process, until a final desired molecule product of a pathway is synthesised. This in turn may either be facilitated immediately or stored for future use. In other words, the product of a pathway may serve as a substrate for the initiation of another pathway. Naturally, the sum of such events constitutes a complicated network and can be seen as a flow of enzymatic activity, which has been categorised into separate units, the aforementioned pathways, to accommodate our intuitive needs in an effort to comprehend the biochemistry of living cells.

2.2.3 Metabolic Pathways

Metabolic pathways are responsible for two major cellular processes, the extraction of energy from the environment, and the synthesis of monomers, the building blocks of macromolecules and their subsequent utilisation for the synthesis of macromolecules themselves (Stryer & Tymoczko 2006). These processes constitute a highly integrated network of biochemical reactions taking place in a cell, the metabolic network.

Glycolysis is arguably the most studied and well characterised metabolic pathway, often used as an example in a variety of text books and on-line resources dealing with basic biochemistry. This is partly due to its universality across all living organisms, with enzymes involved in the catalysis of distinct reactions in the pathway being very similar in different species. It is one of the most ancient metabolic pathways and the first studied (Romano & Conway 1996). In brief, it is the process during which cells convert glucose, a very important carbohydrate, into pyruvate, producing energy. Figure 2.3 provides an overview of the main steps of glycolysis.

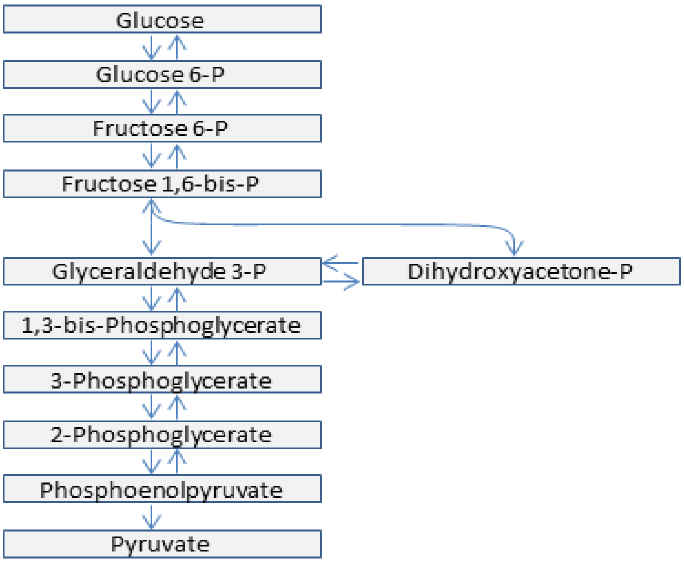


Figure 2.3 Diagrammatic representation of Glycolysis.

Each individual reaction requires catalysis by enzymes all of which constitute members of the glycolysis metabolic pathway. As discussed earlier, glycolysis also constitutes part of a wider network, with which it shares a number of interconnections. Figure 2.4 exemplifies that, using part of the Kyoto encyclopaedia of genes and genomes database (KEGG 2011) representation of the entirety of the biochemical network and interconnections between separate pathways (Kanehisa et al. 2008).

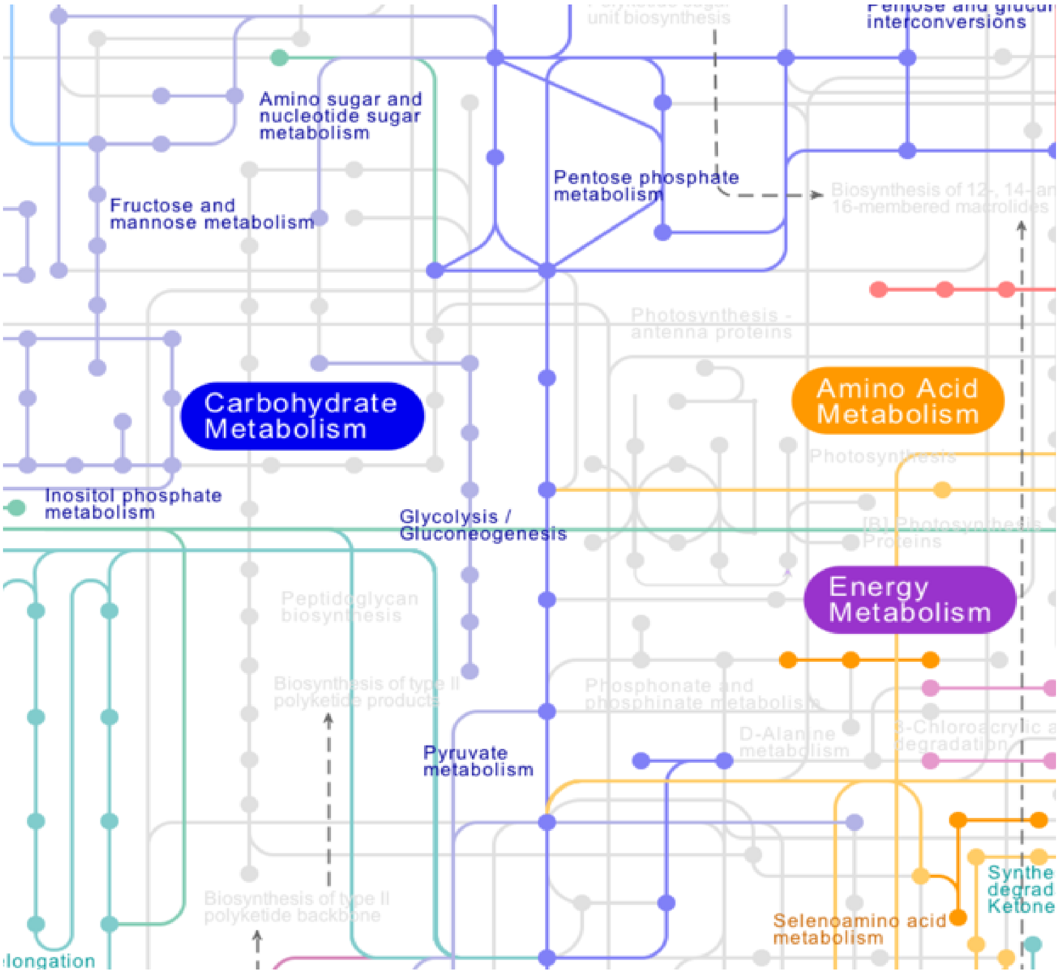

Pentose and glucur
interconversions
unit biosynthesis
Amino sugar and nucleotide sugar metabolism
Pentose phosphate metabolism
Biosynthesis of 12-, 14- an 16-membered macrolides
Fructose and mannose metabolism
Photosynthesis - antenna proteins
Carbohydrate Metabolism
Amino Acid Metabolism
Photosynthesis
Inositol phosphate metabolism
Glycolysis / Gluconeogenesis
[B] Photosynthesis - Proteins
Energy Metabolism
Peptidoglycan biosynthesis
Biosynthesis of type II polyketide products
Phosphonate and phosphinate metabolism
D-Alanine metabolism
3-Chloroacrylic a degradation
Pyruvate metabolism
Synthe degrad Ketone
Selenoamino acid metabolism
Biosynthesis of type II polyketide backbone
longation

Figure 2.4 The KEGG metabolic network. Part of the KEGG representation of the entire metabolic network and interconnections between individual pathways. This is a so called reference pathway map, a generalised view that can be individualised to distinct organisms which naturally share some similarities while at the same time differ in various aspects of their metabolism. Each circle represents a particular chemical compound such as glucose for example, while the lines represent the steps required to turn one compound into another, the so called substrate into a product. These steps are catalysed by one or more enzymes and take place in various cellular compartments. From (http://www.genome.jp/kegg-bin/show_pathway?map01100)

2.2.4 Signalling pathways

Another important category of pathways is the so called signalling pathways, which allow external signals to be passed through various cellular components, leading to a specific cellular response and allowing cell communication. The entire process consists of three stages, reception of the signal coming from outside the cell, transduction of the message, mostly through a sequence of changes in a number of different protein molecules forming the signalling pathway, and in conclusion an adequate cellular response (Campbell & Reece 2007).

Protein kinases, the enzymes that catalyse phosphorylation of other proteins as well as phosphatases responsible for the reverse reaction, hold a major role in signal transduction, as they are the main compounds of signalling pathways. Each molecule acts on another molecule in the pathway, in a sequenced manner. Each cell may contain hundreds of distinct protein kinases, each one acting on different proteins, regulating major cellular processes like reproduction, programmed cellular death, also known as apoptosis and so on.

2.2.5 Importance of biochemical pathways

The study and understanding of pathways constitutes a topic of intensive research as they are of upmost importance for proper cellular function. Metabolic regulation is quite complex due to the integrated nature of the metabolic network and aberrations of the genes involved may have serious impact on cellular state, which we aim to elucidate.

Abnormalities in the structure and activity of protein kinases have been implicated in the development of large variety of cancers (Campbell & Reece 2007). It has been established that kinases are involved in most ‘cancer pathways’, with the cell cycle attracting special attention due to the strong relationship between cell proliferation and tumour development (Nair 2005; Carter et al. 2006).There is plethora of evidence that cell cycle kinases aberrations may lead to uncontrolled proliferation and cell division, some of the main characteristics of cancer development (Malumbres & Barbacid 2007). For example the epidermal growth factor receptor (EGFR) signalling system, which plays a fundamental role in the morphogenesis of a diverse spectrum of organisms, has been implicated in a variety of human cancers. Just to mention a few, EGFR is overexpressed in 50% of epithelial cell malignancies, while HER-2 another receptor of the same family is highly over-expressed in 20-30% of breast cancers (Nair 2005).

Naturally, signalling pathways and their members are of special interest to therapeutics as plausible targets for cancer treatment drugs development. There is already substantial ‘proof of principle’ for the clinical use of kinase inhibitors in cancer treatment. To mention one success story, the development of the tyrosine kinase inhibitor imanitib, known as Gleevec®, has proved highly successful in the treatment of patients with chronic myeloid leukaemia and the majority of newly diagnosed patients have been shown to achieve complete remission (O’Brien, Guilhot & Larson 2003).

Moreover, the study of kinases has great potential in prognosis as it has been shown that gene expression data can be used for prognostic purposes. More precisely, a signature from specific genes, including several kinases, was found correlated to several cancer types, with overexpression of the particular set of genes proving predictive of poor clinical outcome (Carter et al. 2006). This constitutes a good example of the importance of microarrays in applied medical research.

2.2.6 MicroRNAs

MicroRNAs (miRNAs) are a family of short RNAs, approximately 21–25-nucleotide long which are not translated (He & Hannon 2004). However, they have been found to negatively regulate gene expression at the post-transcriptional level. Although microarray technology, which is central in this work, does not allow the study of microRNAs, this section provides a very brief overview of their role, given the recent discovery of their implication in regulating gene expression and pathway activity.

The founding member of the miRNA family, termed *lin-4*, was identified in the worm species *C. Elegans*. Studies demonstrated that the sequence of this 22 nucleotide long RNA is partially complementary to the RNA of *lin-14*, encoding a protein important for the regulation of the transition of the worm from one developmental stage to another. Through direct hybridisation between *lin-4* and *lin-*

14, lin-4 is involved in the control of LIN-14 expression. That is, it is able to block the process of protein translation and hence synthesis.

Almost at the same time with the discovery of the above phenomenon another kind of regulatory process was identified by other small RNAs termed siRNAs (for silence RNAs). In this case base pairing between a siRNA and an RNA transcript is followed by cleavage and degradation of the latter. It should however be noted, that questions have arisen regarding the distinction between the two types of small RNAs, as at least some examples have been identified where the two exchange mode of activity.

Recent studies have shown that miRNAs take part in important biological processes such us development, differentiation, apoptosis and proliferation. A number of groups have shown that changes in the expression levels of these molecules are associated with cancer development. It has been proposed that they can function as tumour suppressors or oncogenes (Callin & Croce 2006). Given these findings and the fact that the human genome may encode over 1000 microRNAs, which may target about 60% of genes, it is not surprising that their study has become quite popular and may contribute to our understanding of gene and protein regulation.

2.3 Bioinformatics

The term Bioinformatics, first proposed by Paulien Hogeweg (Hogeweg 1978), gained wide spread popularity in the 90s, initially used to describe the use of computers for the analysis of gene sequences (Claverie 2000). Today it is broadly accepted as defining an interdisciplinary scientific field that blends biology, computer science and mathematics.

There is a certain degree of confusion as to the distinction between bioinformatics and computational biology and the terms are often used interchangeably. According to the National Centre for Biotechnology Information (NCBI, 2011) computational biology is the actual process of analysing and interpreting biological data, while bioinformatics is the field of science in which biology, computer science, and information technology merge to form a single discipline. From another point of view bioinformatics is a sub-field within computational biology that is concerned with the development and application of algorithms and statistical analysis to interpret biological data. It can be seen as the process of creating the tools rather than the process of interpreting the results, admittedly a vague distinction.

According to the National Institute of Health of the U.S.A. (NIH) Biomedical Information Science and Technology Initiative Consortium (BISTI 2011) the following definitions apply:

> *Bioinformatics: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioural or health data, including those to acquire, store, organise, archive, analyse, or visualize such data.*
>
> *Computational Biology: The development and application of data-analytical and theoretical methods, mathematical modelling and computational simulation techniques to the study of biological, behavioural, and social systems.*

Nevertheless, the emergence of bioinformatics can be attributed to the advancement of genetics and genomics in the 80's, notably to the development of DNA sequencing, discussed in the following section (Moore 2007). The large amounts of sequence data produced became unmanageable without the use of computer power and storage. Indicatively, the human genome alone consists of over 3 billion DNA base pairs. Thus, one of the fundamental aspects of bioinformatics refers to the organisation of the newly acquired knowledge in databases, allowing us to store and manage the large volume of data (Moore 2007).

Naturally, the following step was to find useful and efficient ways to analyse and interpret this knowledge. Thus, today the scope of bioinformatics has expanded to encompass a variety of computationally intensive techniques such as data mining and machine learning methods applied to extract information from molecular biology experiments along with the process of interpreting the data (Moore 2007) . Popular approaches include gene mining, sequence analysis (Vinga & Almeida 2002), gene clustering (Hand & Heard 2005), protein structure (Zhang 2008) and interaction modelling (Skrabanek et al. 2008), just to mention a few.

For example, BLAST (for Basic Local Alignment Search Tool) and FASTA (for FAST-All, referring to the fact it works on any alphabet) algorithms identify regions of local similarity between nucleotide or protein sequences, measuring the statistical significance of the match, thus, allowing us to infer functional relationships between them (Altschul et al. 1990; Pearson 1990). Since their development more than a decade ago, a number of improvements and other algorithmic approaches have been developed. (Vinga & Almeida 2002) provide a useful review of alignment-free sequence comparison methods which attempt to overcome the limitations of the previous approaches, mainly due to the fact that they were largely based on text alignment methodologies that do not account for certain biological realities.

Importantly, microarray technology, essential in this work, is not only one of the major contributors to the accumulation of huge amounts of biological data, but also constitutes one of the major areas of bioinformatics research for more than a decade (Stoughton 2005). It has made it possible to study and compare entire genomes in a very short time span and its focus has already shifted from database management and search to gene discovery and characterisation, modelling gene networks, diagnostics and so on, as discussed in more detail in section 2.5.

2.3.1 DNA Sequencing

Bioinformatics in general and microarray technology itself would be unfeasible without the discovery of sequencing techniques which are in a way the cornerstone in the foundation of gene expression analysis. In fact genome sequencing is also the foundation of what is often referred to as ‘omics’ sciences, including not only transcriptomics (e.g. microarrays) but also genomics and proteomics, the study of DNA and protein structure and function respectively. Given that microarray technology is in the centre of this work, it is useful to provide a brief discussion of DNA sequencing.

In simple terms DNA sequencing refers to the process of reading the genetic code of an organism. While, until quite recently this used to be a laborious task, requiring years to establish the nucleotide sequence of a single gene, today the entire genome of many organisms is readily accessible in minutes, through search of public genome databases.

This became possible not so much due to ground breaking advances in molecular techniques, as due to the collaboration between scientists around the globe and the establishment of large sequencing centres that industrialised the Sanger sequencing method (Maxam & Gilbert 1977). In brief, DNA is synthesised on a single stranded template with random incorporation of modified bases that act as chain terminators. In this way we acquire a range of DNA fragments of varying size corresponding to each position of termination allowing us to read out chunks of nucleotide sequences.

This collaborative effort made the genomes of more than thousand organisms available to the world scientific community (1,550 according to KEGG, 18/02/11). As a result the concept of Systems biology has emerged, since we can now study biological processes in complete cellular systems.

2.4 Systems biology

Traditional biology has focused on the study of individual components of living organisms, such as cells, organelles, genes and proteins in an effort to establish their properties and specific functions. This static approach to biology has apparent limitations as it only provides us with sparse pieces of the puzzle, only examining few aspects of life processes at a time. Nevertheless, the continuous accumulation of more and more pieces of the puzzle in conjunction with technological advances has gradually brought us into a new era of biological research and the foundation of systems biology. This novel scientific field is an interdisciplinary approach to biological systems, integrating traditional biological research with computer science, mathematics, physics and engineering in a holistic analytical approach that seeks to elucidate the dynamics of biological systems. It is a field still in its infancy which is reflected on the fact that it has yet to attain a concise definition. Nevertheless, the popularity of the term has grown rapidly in recent literature, with systems biology institutes emerging around the globe (Ideker 2004; Hodgkinson & Webb 2007).

2.4.1 Overview

Systems biology is a holistic approach to the properties of living organisms. As opposed to the reductionist point of view that a system can be understood by reducing it to its individual parts, thus a biological system can be understood in terms of looking into the chemical and physical properties of the molecules of which it consists, the holistic view can be summarised by Aristotle’s point that ‘The whole is more than the sum of its parts’.

Systems biology’s roots can be traced in the work of the mathematician Robert Weiner’s and his book ‘Cybernetics, or Control and Communication in the Animal and the Machine’, who first introduced the theory of feedback systems, that is systems capable of self-regulation applicable to both living organisms and machines (Weiner 1948). Naturally, at the time we neither possessed the knowledge nor the processing power of computers to be able to apply this to practical research.

Today there has been considerable progress and our focus has drastically shifted towards understanding a system's structure and dynamics. Genes, proteins and their interconnections are merely a static roadmap, whereas we are truly interested in the traffic patterns and their properties (Kitano 2002b). Naturally, understanding the components of the system remains important. In fact, decryption of the genome, facilitated by advances in molecular technologies, along with the development of high throughput measurements, such as microarray technology, have been the driving force behind the emergence of systems biology.

To gain system-level understanding of a biological system we need to decipher four key properties. Firstly the system’s structures, that is genes and the nature of their interactions through biochemical networks. Second, we need to understand the behaviour of the system over time under various conditions. Third, decipher the control mechanisms employed by the cell to optimise its functions and avoid malfunctions and finally, elucidate the basic design principles that govern the properties of the biological system (Kitano 2002b). In this effort, gene expression analysis to identify co-expression of genes, and importantly identification of unknown genes that appear to interact with genes of known function, through clustering and correlation analysis plays a major role (Eisen et al. 1998).

While biology has always borrowed some scarce contribution from other sciences, notably mathematics, today this is more so than ever. Especially, in the last couple of decades the balance has shifted dramatically and biology has seized to be a single discipline, which is generally the case for most scientific disciplines. Today it is inconceivable to study biology in isolation, even at school level, let alone carry out advanced biological research without the use of computers. "Real" biology is increasingly carried out in front of a computer, be it so to simply retrieve a nucleotide sequence or produce a complex model of a gene network (Roos 2001). Physics, chemistry and engineering also have a wide range of contributions, for example in DNA sequencing, PCR technology, protein mass spectrometry, microarray analysis and so on. Mathematics, simple or advanced is indispensible part of research at every level.

The systems biology approach, has obvious practical advantages, notably in therapeutics as deeper understanding of biological functions allows us to develop new treatments. The combination of computational, experimental and observational enquiry is of great interest in drug discovery and individualisation of medical treatment regimes. While there is still a long way to go it is widely accepted that the future of medicine lies in the application of systems biology to medical practice (Kitano 2002a).

In a review of systems biology in drug discovery the authors identify three main principal approaches to the task in hand, namely the integration of distinct ‘omics’ data sets, the modelling of system physiology from cell and organ response level information in the literature, and the use of complex human cell systems in an effort to understand and predict the biological activities of drugs and gene targets (Butcher, Berg & Kunkel 2004). These are complementary approaches that need to be integrated if we are to gain deeper understanding of human disease. Nevertheless, they have already contributed to the process of drug discovery, by accelerating hypothesis-driven biology, providing useful models for target validation and increasing our ability to interpret organism responses to drugs. Perhaps, due to its infancy as a field systems biology has not yet produced a major ‘success story’.

Nevertheless, the intensive work and growing interest in the field is evident and we can be reasonably hopeful that breakthroughs will soon follow.

2.5 Microarrays technology

The fundamental goal of biological research is to improve our understanding of organisms and the underlying biological processes allowing them to facilitate the chemical compounds available in their environment in order to maintain life. As in other scientific fields biology relies heavily on subsequent rounds of hypothesis formulation and experimental design to test their credibility. It has been stressed that the development of novel highly automated, high throughput biological techniques have had a crucial impact on the rate at which data can be acquired. Microarray technology is arguably one of the best examples of highly sophisticated experimental methodology, which has revolutionised biological science in recent times.

This experimental technique was conceived by Mark Schena and Ron Davis in the early 90's while studying the function of transcription factors in the flowering plant *Arabidopsis thaliana*, and soon after published in science magazine (Schena et al. 1995). Microarrays allow us to observe transcription, the first step of gene expression subjected to extensive regulations by internal and external factors. In the past, more traditional methods to study gene expression were based on one gene per experiment principle. Microarray technology made it possible to study the expression levels of many thousands of genes from a particular cell in one single experiment. This not only increased the speed of experimental process but greatly reduced the cost of gene expression studies, making it possible to obtain genome-wide expression data and observe the effect of different physiological conditions by direct comparison between expression levels of genes or their products.

2.5.1 Underlying concept

Microarray technology is based upon the ability of a particular nucleic acid to hybridise specifically to the DNA template from which it originated, due to hydrogen bonds formation. It can be seen as an extension of Southern blot, the first DNA array, used to search for complementing sequences (Southern , Mir & Shchepinov 1999).

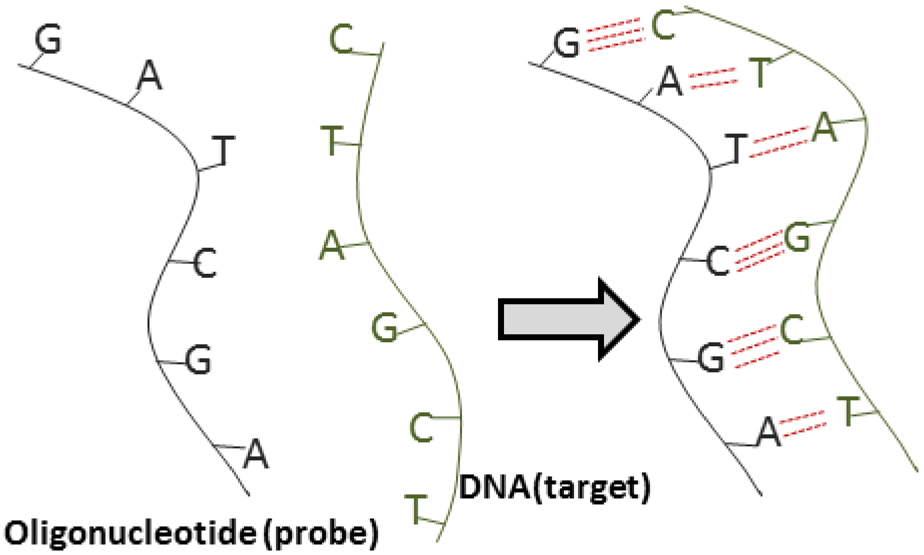


Figure 2.5 Hybridisation of nucleotide sequences. The target DNA sequence aligns with its complementary probe due to formation of hydrogen bonds (dashed lines) between the complementary base pairs, A-T and G-C.

Most commonly, RNA isolated from different types of cells or tissue, is converted to cDNA (for complementary DNA) and labelled with two distinct fluorescent tags (such as Cy5 and Cy3). The resulting mixture is added to a microarray chip that carries attached an orderly arrangement of nucleic acid sequences, each representing a specific gene. Upon exposure of the chip to the set of labelled samples hybridization takes place, due to the formation of base-pairs between complementary nucleic acids (Figure 2.5). In particular, two hydrogen bonds are formed between A-T base pairs and three between G-C base pairs, in the DNA duplex. Upon completion of that step we measure the amount of target bound to each sample. In particular we measure the intensity of a spot resulting from the amount of fluorophore present. The resulting image is used obtain a dataset consisting of raw intensity measurements for each individual spot representing a gene. The basic hypothesis is that the measured intensity level for each gene represents its relative expression level.

2.5.2 Raw data pre-processing

Importantly, before RNA levels can be compared appropriately, a number of transformations must be carried out on the data. A range of statistical treatments have been proposed for data normalisation with the aim to eliminate low-quality measurements, select genes that are significantly differentially expressed and to facilitate comparisons (Quackenbush 2002). The most basic one is the total intensity normalisation, where the ratio of expression change for each gene is divided by the summation of all intensities in both channels.

Most commonly, microarray experiments compare gene expression ratios between two conditions of interest, each one represented by a whole RNA, each labelled with a different dye, say green (*G*) for condition A and red (*R*) for B. Thus, each gene can be represented by a ratio, $T_i = \frac{R_i}{G_i}$ revealing the relative change of expression, as defined by the relative presence of each dye on the spot.

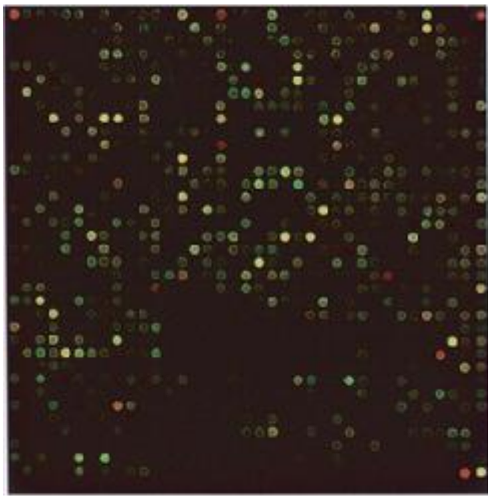


Figure 2.6 Microarray image from (Schulze & Downward 2001). Spots appearing red correspond to genes more actively transcribed under the condition labelled with red dye, while the opposite is true for green spots. Yellow spots correspond to genes of similar expression under both conditions.

This is due to the fact that red and green labelled RNA molecules, corresponding to different genes, compete between each other to bind the respective complementary oligonucleotide sequences, often termed probes, on each spot. Whichever one is present in the mixture in greater abundance will also win the competition and be present on the spot in larger quantities, upon completion of the hybridisation.

Scanning of the array, using two different wavelengths, corresponding to each dye, provides relative signal intensities and thus ratios of mRNA abundance for each individual gene. Figure 2.6 shows an example of a microarray image.

However, ratios have a disadvantage, especially from the point of view of graphical representation, since they treat up- and down-regulation differently. For example an increase of expression by a factor of 2 would be represented by a ratio of 2, while down- regulation by the same factor would be represented by a ratio of 0.5. To deal with that issue, most commonly, we use the base 2 logarithm, producing continuous spectrum of values. For the example discussed here, increased expression by a factor of two is represented by $\log_2(2) = 1$, while decrease by the same factor by $\log_2(1/2) = -1$. Figure 2.7 provides a brief overview of the main step of microarray technology preceding the final analytical process.

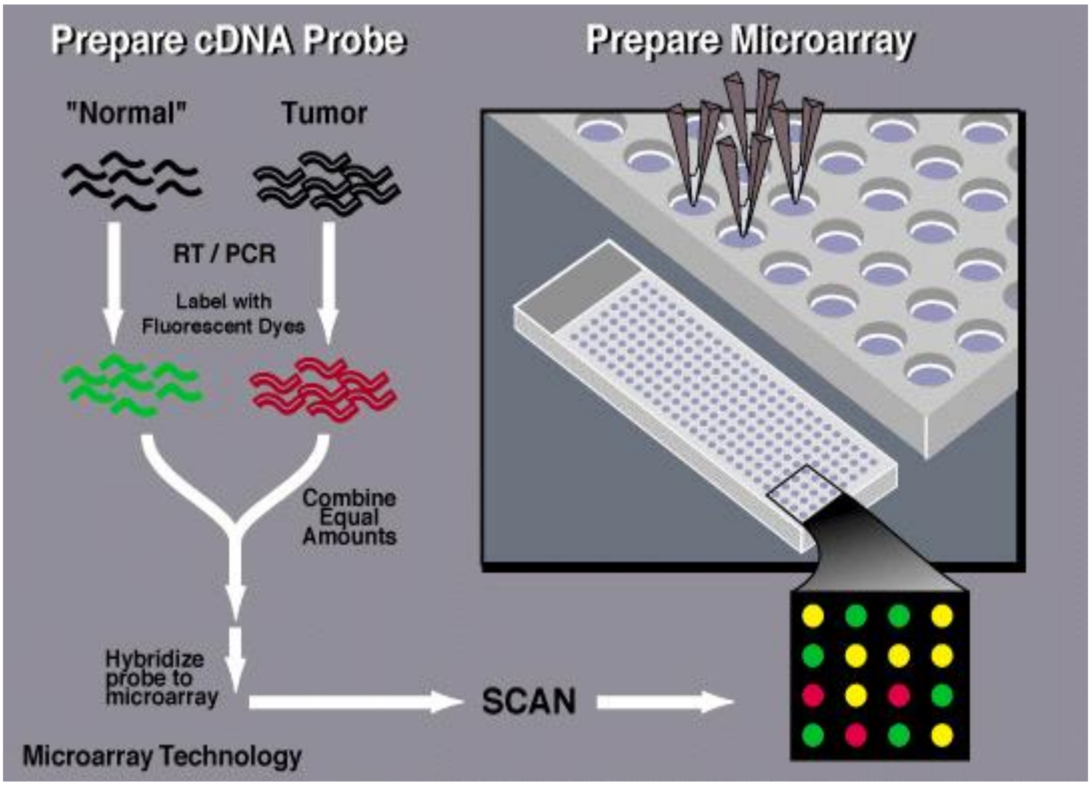


Figure 2.7 Overview of the major experimental steps of DNA Microarrays. From http://www.genome.gov/10000533,31/01/2011

2.5.3 Microarray platforms

There is a wide range of DNA microarray platforms including one- and two-channel formats, cDNA and oligonucleotide microarrays, in-house spotted microarrays, and commercially developed microarrays. The first widely used microarrays were based on the use of PCR-amplified cDNA fragments serving as probes, deposited in a matrix pattern of spots on a treated glass surface (Šášik, Woelk & Corbeil 2004).

The next microarray technology to emerge involved *in situ* synthesized oligonucleotide arrays using photolithographic technology pioneered by Affymetrix Company (Santa Clara, CA, USA) GeneChips (Lipshutz et al. 1999) which have become the industry standard and advances in the field are often measured against this technology. The popularity of the platform is due to the high density of the arrays, the facilitation of quality control and high reproducibility. The preparation process implements photolithography where oligonucleotides are synthesized in light directed manner, directly onto the chip, in 3’ to 5’ direction. In particular, a glass wafer is appropriately modified with photolabile protecting groups which prevent DNA base binding to its surface. In order to anchor a DNA base to the chip, with the use of a robot, a beam of light passing through a photolithographic mask eliminates the photolabile protecting groups at specific X, Y co-ordinates. At this point the surface of the chip is flooded with the appropriate mononucleotide, which is also photoprotected by a photolabile group at its 5’ end, resulting in anchorage of the nucleotide to the surface of the wafer at the exact positions where the light beams eliminated the protecting groups in the preceding step. Light passing through a second photomask de-protects selectively different positions on the substrate, so that a new 5’ proteced deoxynucleoside can be added. Affymetrix use 25-mers oligonucleotides multiple probes to estimate the abundance of each target transcript. Importantly, this approach is not based on competitive hybridisation, thus, it reveals the abundance of a particular transcript from a particular cell or tissue sample, using a single channel/dye format. To compare two samples, two separate microarrays must be produced.

Another approach is the *in situ* synthesis of probes, using inject printing. During the delivery process, each sample is loaded into a miniature nozzle controlled by a robotic system to ensure that it is spotted at an individual location, with the desired X, Y co-ordinates. Upon delivery of a sample the nozzle is washed and loaded with the next sample of interest. This technology is facilitated by Agilent (Palo Alto, CA, USA) and has the advantage of using longer 60-mer oligonucleotides as probes, allowing the use of only one probe per target (Hardiman 2004). Similarly to the original cDNA arrays Agilent implements competitive hybridisation using two distinct dyes.

Pre synthesis of oligonucleotides or cDNAs has the important advantage that the sequences eventually placed on the array can be exactly those desired. This strategy is implemented on the CodeLink^{TM} Bioarray platform from Amersham Biosciences (Piscataway, NJ, USA). The oligonucleotides are immobilized on the slide surface following their synthesis through covalent attachment. An advantage here is that the hydrophilic gel surface reduces non specific binding, hence minimizing background noise (Hardiman 2004). This methodology also uses a single channel approach.

Furthermore, Expression Array System from Applied Biosystems (Foster City, CA, USA) uses standard phosphoramidite chemistry to synthesise 60-mer oligonucleotides, which are validated by mass spectrometry prior to deposition on a nylon microarray substrate and subsequently mounted on a glass support (Hardiman 2004). Illumina chips also use standard oligonucleotide synthesis. However, Illumina facilitates beaded oligonucleotide arrays comprised of thousands of microwells (Šášik, Woelk & Corbeil 2004). Each microwell contains a single bead carrying more than 10^5 50-mer oligonucleotides probes targeting a unique gene. Importantly, beads that carry the same probe are scattered randomly across the microarray, to deal with variable signal across the chip.

It is important to note that regardless of the platform, the raw data comes in the same form, that is, a scanned image.

2.5.4 Applications

Microarray technology has become an indispensible tool for biological research, as evident in the huge number of scientific papers published in that area and the growing availability of microarray data. Gene expression analysis has been facilitated for a wide range of applications, notably for the identification of genes related to particular phenotypes, often referred to as biomarkers (Chu et al. 2005), drug discovery and development (Debouck & Goodfellow 1999), study of biochemical pathways (Wu et al. 2007), prognosis and diagnostics (Dai et al. 2005; Carter et al. 2006), therapeutics (Gerhold, Jensen & Gullans 2002) and so on. An informative review of the main features of microarrays technology and its applications in biology can be found in (Stoughton 2005).

A number of distinct strategies to analyse and exploit microarray data have been proposed. One of the most basic analytical methodologies emerging from the start, still popular to date, is the application of various clustering techniques to the expression profiles of genes (Kerr et al. 2008). In cluster analysis of microarray data, we wish to partition genes into groups, clusters, based on expression measurements. In this way we obtain classes of genes that show highly similar expression patterns, while being disjoint with genes in other classes. This is largely based on the ‘guilt by association’ notion, which assumes that genes with similar expression patterns are functionally related to each other (Brazhnik, de la Fuente & Mendes 2002).

Hierarchical and partitional clustering methods represent the two basic approaches to clustering (Jain, Murty & Flynn 1999). In the first case each gene is initially placed in a distinct cluster, followed by successive merging of clusters together until a stopping criterion is satisfied. This process results in a *dendrogram* representing the nested grouping of patterns and similarity levels at which groupings change. It is one of the most popular clustering techniques in microarray studies, introduced soon after the development of the methodology (Eisen et al. 1998). In contrast, partitional clustering produces a single partition of the data instead of a *dendrogram,* which optimises a chosen measure of clustering quality (Hand & Heard 2005). A popular example here is the k-means clustering algorithm, which iteratively moves observations between clusters in an effort to minimise the average squared distance between observations and their cluster centroid (MacQueen 1967). While the second is less computationally expensive than hierarchical clustering and more suitable for large datasets, it requires the choice of the number of desired output clusters, which in reality is usually unknown, and tends to produce distinct clusters at every application. An extensive review of clustering methodologies applied to gene expression data is provided in (Hand & Heard 2005).

As aforementioned, given that the function of a number of genes has been experimentally determined clustering can be useful in elucidating the role of unknown genes that appear in the same cluster and generally in providing the researchers with interesting targets for further analysis (Eisen et al. 1998). In the same context, (Yano et al. 2006) have proposed a methodology for identifying genes whose differential expression is related to a particular phenotype, using a set of microarray experiments associated with the trait of interest and a reference dataset.

Classification of microarray data is another strategy that has gained wide spread popularity. In contrast to clustering, an unsupervised learning technique, classification is a supervised learning method, as here we have a priori knowledge of the class that each experiment belongs to. This may be a certain disease state, environmental condition or developmental stage. The goal is to teach a classifier to distinguish between such states or phenotypes of interest, with obvious applications to medical research (Quackenbush 2006). A number of complicated artificial intelligence approaches have been applied in this area of research, such as neural networks and support vector machines with variable success (Furey et al. 2000; Ringnér & Peterson 2003).

2.6 Heuristics

The accumulation of biological data has presented us with new opportunities to gain insights into biological processes, but at the same time it has also presented us with an increasingly complicated range of problems, to which there are often no simple solutions. Making sense of the amount and complexity of biological data we have at our disposal constitutes a difficult task. This is where heuristics come in handy.

Heuristic approaches to analyse biological data are quite popular, in a number of areas, notably in the analysis of microarray data. Here heuristics are largely employed for clustering of genes, based on expression profiles and classification discussed in section 2.5.3. Another important area of interest is the construction of genetic network models from microarray data in order to reveal the regulation rules behind the gene expression profiles. While these networks are phenomenological and simplified, as they do not directly represent the proteins and metabolites involved in cell functions, they are a logical way of describing phenomena based on gene expression data (Brazhnik, de la Fuente & Mendes 2002)**.**

Given that the work presented here facilitates heuristic methods, to analyse gene expression data and propose solutions to biologically driven questions, a brief overview of such methods and some important, relevant concepts is included in this section.

2.6.1 Overview

The term heuristics, from the Greek word "heuriskein" meaning "to discover", refers to a wide range of problem solving approaches, which derive an approximate solution to a problem in a faster or more economical way than a mathematically strict algorithm. A heuristic can be seen as a rule of thumb that may help us to solve a problem. A good example is the left-hand rule for solving a maze, which states that by holding one hand in contact with one wall we are guaranteed to reach an exit. Heuristics apply to problems where exhaustive search for an exact solution is impractical, that is, problems that have such a wide range of possible solutions that even using the currently available computer processing power does not allow us to examine them all within a reasonable time frame.

Perhaps the most basic example is the Boolean satisfiability problem, where we wish to find the appropriate assignments to individual variables in a Boolean formula that will make it evaluate to TRUE (Michalewicz & Fogel 2004). A plausible solution for this problem can be represented as a binary string where ones correspond to TRUE and zeros to FALSE. Naturally, the number of potential solutions to such a problem depends on the length of the string. Given that we are presented with two choices of values for each variable, for a string of length n there are 2^n plausible solutions, corresponding to all combinations of assignment of values at each position of the string. Hence, as the size of the problem increases the number of plausible solutions follows suit at an exponential rate. For a binary string of 100 variables, there are 2^{100}

such solutions, a number so huge that it is impossible to examine each and everyone within the lifetime of the universe.

The purpose of Heuristics is to deal with such problems, by speeding up the search process and finding a solution, that may not be the exact solution to the problem in hand, but is a good approximation, at least good enough to satisfy our needs for the task in hand.

2.6.2 Basic concepts

In general, all algorithmic approaches to solving problems consist of three basic components, the representation, the objective and the evaluation function (Michalewicz & Fogel 2004). First we need to represent candidate solutions in a manner that is consistent with the problem and allows computational manipulation. In the satisfiability problem, discussed above, a fundamental representation corresponds to a binary string. Notably, the chosen representation for any problem implies the size of the search space, meaning the range of possible solutions, which is of upmost importance in algorithmic search.

Additionally, we need an objective, a definition of the goal we seek to achieve, which in the above discussed problem is to make the statement of Boolean variables evaluate to TRUE. Finally, we need an evaluation function, in order to be able to evaluate the worth of a solution and compare it to the worth of alternative solutions. This is often a mapping between a solution and its quality.

In the example here, the value of an approximate solution, FALSE, does not give us any indication of how close we are to reaching our goal. It does not allow us to compare alternative solutions and choose an appropriate direction for subsequent solutions in the search space. This is an important point, as in modelling a problem it is useful to design an evaluation function that will direct us towards better alternative solutions.

The traveling salesman problem, one of the most studied problems in computational mathematics, constitutes a useful example here (Michalewicz & Fogel 2004). In its most basic form, the goal is to find the shortest route visiting a number of cities and returning to the point of departure, without repetition. Figure 2.8 graphically represents a sample TSP problem of 4 cities. Here the most basic representation coming to mind is a string of natural numbers from 1 to *n*, where each number represents a city and the order of numbers a potential route. In essence, the size of the search space is equal to n!, corresponding to the number of all possible permutations. However, taking into account that in the basic form of the problem the distance between two cities is the same in either direction and that the circuit is the same regardless of the starting city, the search space is reduced to (n-1)!/2. Here we can evaluate each solution to the problem by summing the distance between each city and the following one in the route. Hence, we can compare a solution to a previous one and see if we are going in the right direction. This is an important point in algorithmic search approaches in the field of heuristics.

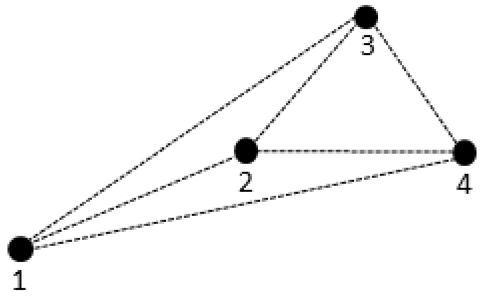


Figure 2.8 An example of 4 cities TSP problem.

2.6.3 Stochastic local search algorithms

There is a wide range of heuristic search algorithms whose purpose is to find an optimum solution to a problem in a given search space. Stochastic algorithms are a type of heuristic search methods that examine random solutions until reaching a time limit or criterion. This section is an overview of some basic choices relevant to this work. As aforementioned the idea is to find approximate solutions to problems where exhaustive search is infeasible. The underlying strategy in the case of local search, where we concentrate on a solution and its local neighbourhood is the following:

1. Choose a solution from the available search space and estimate its fitness.

2. Transform it in some way reaching a new solution and evaluate the new solutions fitness.
3. If the new solution turns out to be worse than the previous one it is discarded, if however it turns out to be better it is kept as the current solution.
4. The process at steps 2 and 3 is repeated until no improvement occurs or until chosen criteria are satisfied (e.g. a given number of iterations performed)

Hill climbing belongs to the category of local search algorithms, as it operates using a single current state and moves continually to neighbours of that state in a direction that increases the fitness of the corresponding a state (Russell & Norvig 2003). In the case of the traveling salesman problem, where a particular solution is represented as a string, each position corresponding to a city, the basic method to proceed is to swap a pair of cities in the string and evaluate the quality of the resulting route. Thus, we introduce a small change, moving to a neighbouring solution that differs in only two positions (i.e. cities) from the preceding string. If the new configuration is of better quality, that is, the overall length of the route is shorter, the route is kept to serve as the new current solution otherwise it is discarded. While this method needs not remember the route followed and is not computationally intensive, it tends to get stack in local optima. That is, local solutions which are better than other solutions in the surrounding neighbourhood, but do not constitute the best global solution.

Figure 2.9 serves as an example to demonstrate the concept of local optima and search space. It represents one dimensional state space landscape, where the aim is to find the highest point or global maximum. A hill climbing algorithm starting at the position indicated with a circle will evaluate the height of its immediate neighbours, and move uphill to the right until it reaches point C. From there on it will get stuck as no surrounding position is higher. This point is a local maximum, such as A, B and D, however, only B constitutes the global maximum, the highest point overall. This hill climbing example is analogous to a person trying to climb a hill in the fog. They feel around themselves until they find a point higher than they are at the moment and proceed until no such point is found. The height of the hills can be seen as the fitness when plotting the fitness space. Naturally, depending on the individual problem we may be looking for a global minimum, which follows a similar rationale.

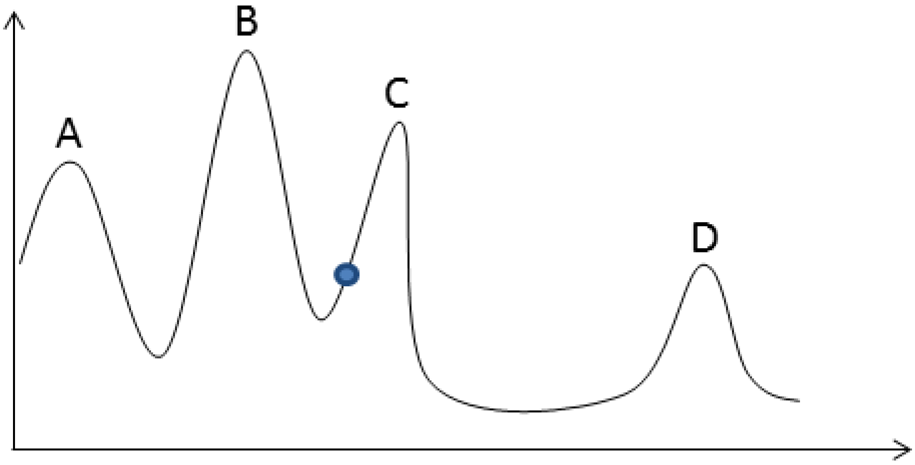


Figure 2.9 One-dimensional state space landscape. Points A, C, and D are only local maxima while B is the global maximum.

One common solution to the tendency of hill climbing to get stack in local maxima is to restart the search at various random positions and eventually choose the best of all solutions. Another way to deal with this is implementing a simulated annealing algorithm (Kirkpatrick , Gelatt & Vecchi 1983). It is quite similar to hill climbing, but here a solution of worse fitness may be accepted, in a controlled manner. It borrows its name from the process of annealing in metallurgy, where metals are initially heated at very high temperatures and then allowed to gradually cool down. At high temperatures the atoms wander around randomly, from time to time adopting states of higher energy. The chance of this occurring gradually decreases as temperature drops. This process allows misplaced atoms to adopt more thermodynamically favourable positions and hardens metals.

Simulated annealing incorporates the notion of temperature, through a parameter *T*, which defines the probability of accepting a solution of worse fitness. This parameter is initially set high and follows a scheduled gradual decrease as the algorithm progresses. Hence, after a certain number of iterations this probability becomes so small that only solutions of better fitness are accepted, as in the case of Hill climbing.

Genetic algorithms represent another case of local stochastic search methodology that follows the rationale of natural selection, first proposed by John Holland (Holland 1975). In nature individuals with traits that make them successful in their environment, have increased chances of survival and production of offspring. Consequentially, overtime the characteristics of such individuals show increased spread in the general population. In a fashion that mimics this process genetic algorithms start with a set of random solutions termed individuals which as a whole constitute a population. Individuals are usually represented as binary strings, or in some cases as strings of another finite alphabet (Russell & Norvig 2003), and their fitness can be evaluated through an appropriately chosen fitness function. Given a starting population, a change is introduced to randomly selected individuals forming it, leading to production of a new collection of individuals, increasing diversity. The individuals constituting the new population are selected and preserved based on their fitness. The process is repeated for a chosen number of iterations or until a desired fitness is reached.

Typically, the changes introduced to individuals at each iteration fall into two basic categories, crossover and mutation. In a crossover, mimicking biology, where two chromosomes may exchange parts during meiosis, two binary strings exchange parts at a particular positions giving birth to two new chromosomes. Each of the new chromosomes, typically referred to as daughter chromosomes, consists of parts of the initial ones. In the basic case of one point crossover, the new chromosomal parts come from either side of the position of exchange, as shown on Figure 2.10.

A mutation is most commonly implemented as a single change in a binary string at a particular position, as shown on Figure 2.11. This process mimics the biological process of a mutation, during which a change occurs in a single base in the nucleotide sequence of a genome.

In the simplest case once a new generation is introduced the fittest individuals are preserved. The number of such individuals, that is the size of the population is chosen based on the problem in hand. However, in order to maintain diversity, it is a common strategy for some individuals, even of low fitness, to be passed to the next generation at each step of the process. One very popular method to do this is the so called roulette-wheel selection.

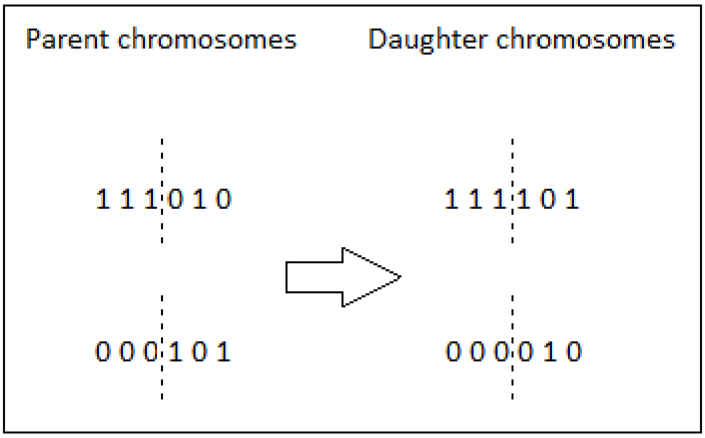


Figure 2.10 One point crossover. The figure exemplifies the exchange of parts between two parent chromosomes, at the position of the dashed line, to produce two daughter chromosomes, each one consisting of a part of each parent chromosome. This is a one point crossover, the simplest type. In this example the exchange takes place in the middle, but generally crossover can occur at any available position.

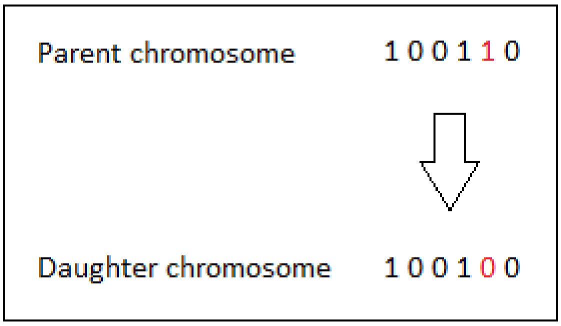


Figure 2.11 The genetic operator of mutation, giving birth to a new, daughter chromosome, from an existing parent chromosome.

In this approach, every individual is assigned a probability of being selected which is directly proportional to its fitness, thus the higher the fitness the most likely the individual is to be preserved. This is analogous to a roulette-wheel, in the sense that a proportion of the wheel is assigned to each individual based on its fitness. The higher the fitness the larger the proportion and thus the greater chance of randomly selecting the individual upon turning the wheel. Naturally, the method allows for individuals of lower fitness to be occasionally chosen.

Hill climbing, simulated annealing and genetic algorithms along with other search techniques are also referred to as metaheuristic methods, as they seek to optimise a problem through iterative search, facilitating an appropriate measure of quality, without prior expert knowledge regarding the problem in hand.

2.7 Biological Databases

The importance of public databases in spreading knowledge and providing raw material for data mining has already become apparent in the previous sections. What has often started as a simple collection of data regarding the research topic of a certain group has today become indispensible tool for biological research. The wealth and diversity of freely available information would be difficult to conceive only a decade ago. In fact, it is hard to imagine of any contemporary research effort that does not facilitate some type of database to a larger or lesser extent.

The importance of biological databases is reflected in their popularity and the rate at which established databases have grown and new ones have emerged in recent years. Indicatively, the 2005 release of the Nucleic Acids Research online Molecular Biology Database Collection (NAR, 2011) includes 719 databases, an increase of 171 over the previous year (Galperin 2005). In comparison, the 2010 release of the same collection contains 1230 carefully selected databases covering various aspects of molecular and cell biology, an increase of 5% over the last year (Cochrane & Galperin 2010).

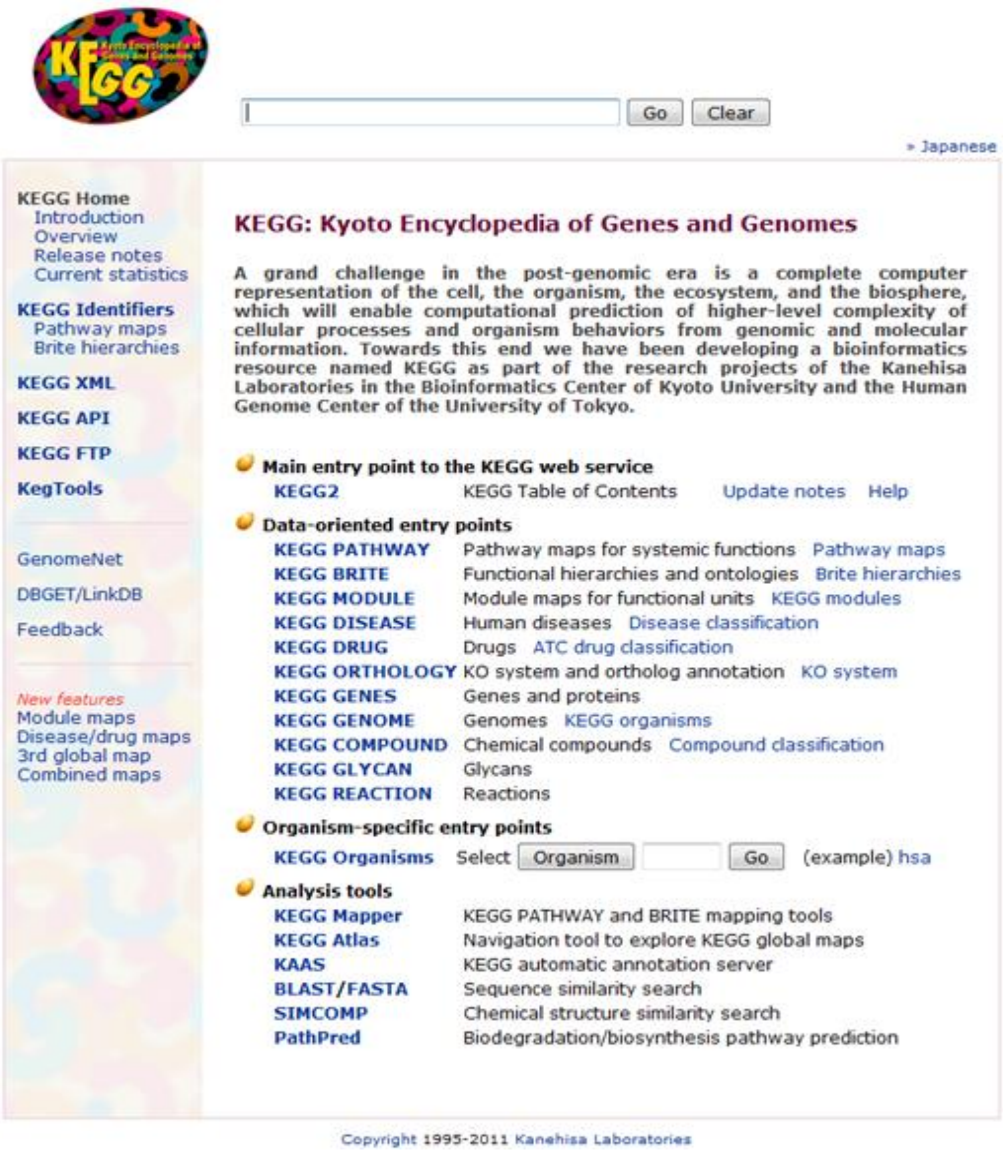
Databases can be roughly categorised according to the type of information stored in them. Table 2.2 provides an overview of the categories of biological databases according to the Nucleic Acids Research online Molecular Biology Database Collection. Importantly, this categorisation is only a rough guide as number of databases store data that can be assigned to more than one of these categories. For the scope of this work metabolic and signalling pathway databases along with microarray data databases are the most relevant and will be briefly discussed.

Table 2.2 2011 NAR Database Summary Paper Category List

Nucleotide Sequence Databases
RNA sequence databases
Protein sequence databases
Structure Databases
Genomics Databases (non-vertebrate)
Metabolic and Signaling Pathways
Human and other Vertebrate Genomes
Human Genes and Diseases
Microarray Data and other Gene Expression Databases
Proteomics Resources
Other Molecular Biology Databases
Organelle databases
Plant databases
Immunological databases

2.7.1 Metabolic and Signalling pathways databases

Pathway databases provide a collection of metabolic and regulatory pathways, including the genes and proteins involved along with chemical compounds participating in the respective reactions, which can be seen as the wiring diagrams of genes and molecules. The KEGG database (Kanehisa et al. 2008) which plays a prominent role in this research is a characteristic example of database that stores information which makes it assignable to more than one of the categories on Table 2.2. It is a general genomics database, storing information about individual genes and a number of completed genomes in its GENES section, while at the same time a pathway database, with graphical representation of cellular processes included in the PATHWAY section. Importantly these sections are linked providing information about the way genomic information is related with higher order functional information, that is, pathways. Figure 2.12 provides a snapshot of the KEGG database home page.



KEGG Home
Introduction
Overview
Release notes
Current statistics

KEGG Identifiers
Pathway maps
Brite hierarchies

KEGG XML

KEGG API

KEGG FTP

KegTools

GenomeNet

DBGET/LinkDB

Feedback

New features
Module maps
Disease/drug maps
3rd global map
Combined maps

KEGG: Kyoto Encyclopedia of Genes and Genomes

A grand challenge in the post-genomic era is a complete computer representation of the cell, the organism, the ecosystem, and the biosphere, which will enable computational prediction of higher-level complexity of cellular processes and organism behaviors from genomic and molecular information. Towards this end we have been developing a bioinformatics resource named KEGG as part of the research projects of the Kanehisa Laboratories in the Bioinformatics Center of Kyoto University and the Human Genome Center of the University of Tokyo.

Main entry point to the KEGG web service

KEGG2	KEGG Table of Contents	Update notes Help

Data-oriented entry points

KEGG PATHWAY	Pathway maps for systemic functions	Pathway maps
KEGG BRITE	Functional hierarchies and ontologies	Brite hierarchies
KEGG MODULE	Module maps for functional units	KEGG modules
KEGG DISEASE	Human diseases	Disease classification
KEGG DRUG	Drugs	ATC drug classification
KEGG ORTHOLOGY	KO system and ortholog annotation	KO system
KEGG GENES	Genes and proteins	
KEGG GENOME	Genomes	KEGG organisms
KEGG COMPOUND	Chemical compounds	Compound classification
KEGG GLYCAN	Glycans	
KEGG REACTION	Reactions	

Organism-specific entry points

KEGG Organisms Select Organism Go (example) hsa

Analysis tools

KEGG Mapper	KEGG PATHWAY and BRITE mapping tools
KEGG Atlas	Navigation tool to explore KEGG global maps
KAAS	KEGG automatic annotation server
BLAST/FASTA	Sequence similarity search
SIMCOMP	Chemical structure similarity search
PathPred	Biodegradation/biosynthesis pathway prediction

Copyright 1995-2011 Kanehisa Laboratories

Figure 2.12 KEGG homepage (KEGG 2011).

MetaCyc (Caspi et al. 2010) is another example of a popular metabolic-pathway database that describes more than 1000 pathways. An important characteristic of MetaCyc is that it only deals with pathways that have been determined experimentally through wet lab research. While this approach imposes some limitation on the amount of available data, on the positive side, it confers accuracy and reliability to the available information. Notably, MetaCyc provides a graphical user interface with a plethora of options and a number of applications including pathway analysis tool.

In fact, MetaCyc is part of a larger database, named BioCyc (Karp et al. 2005) which uses Pathway Tools software and MetaCyc as a reference to construct predicted metabolic networks. It holds a collection of 653 organism-specific Pathway/Genome Databases (14/07/10), each one containing the full genome and the predicted metabolic network of one organism.

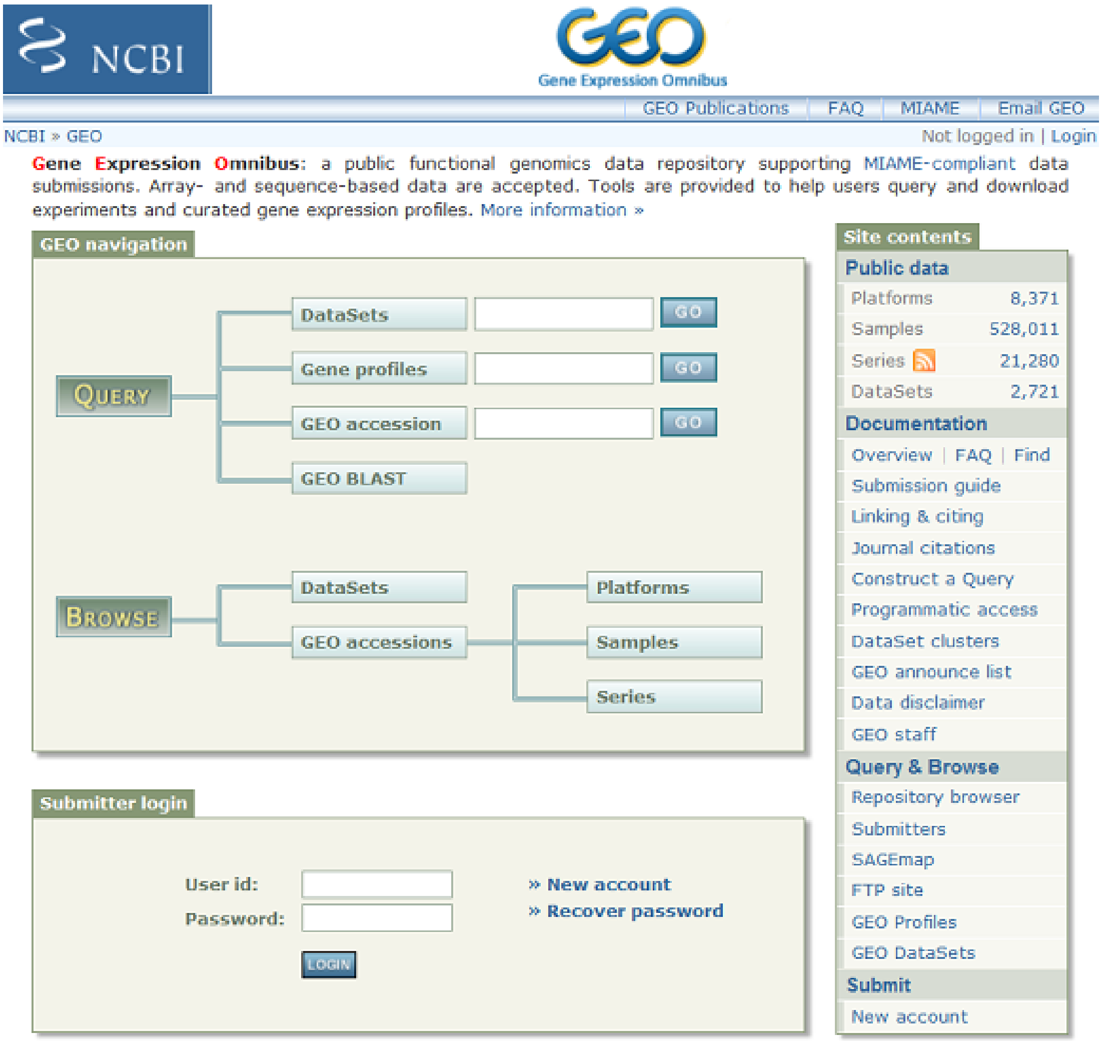
Finally, Reactome is another pathway database of importance relevant to this work, dealing with human pathways and processes (Croft et al. 2011). Importantly, the database is manually curated and peer-reviewed by an expert team of biologists and has gained widespread popularity. The core unit of the Reactome data model is the reaction, hence the name. Naturally, reactions are grouped into pathways representing a network of interconnecting processes. The data model generalizes the concept of a reaction to include the transport of a molecule from one compartment to another and the formation of complexes besides the classical biochemical transformations. Hence, pathways in Reactome include classic metabolism as well as signalling, transcriptional regulation, apoptosis and so on.

All pathways are cross-referenced to proteins, genes and small chemical compounds in relevant databases, primary research literature and GO controlled vocabularies. Besides an intuitive useful visualisation of such pathways, also allowing navigation and zooming in and out of processes, the database provides tools for pathway based analysis of microarray and other datasets. The user can supply a list of entities, such as genes and expression data to identify over expression in pathways.

2.7.2 Microarray Data Databases

Since the introduction of microarray technology it has become a widely used tool for the generation of gene expression data. This has been accompanied by an apparent growing demand for any publication to make the analysed dataset available to the wider research community. Naturally, a number of databases have been created to satisfy this need, with 69 listed in Nucleic Acids Research online Molecular Biology Database Collection (20/02/11). They include the National Centre for Biotechnology Information (NCBI) Gene Expression Omnibus database (Barrett et al. 2009) (Figure

2.13) and ArrayExpress (Parkinson et al. 2008), which have emerged as the main public repositories.



Gene Expression Omnibus: a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles. More information »

Public data	
Platforms	8,371
Samples	528,011
Series	21,280
DataSets	2,721

Figure 2.13 GEO homepage (GEO 2011).

Today, a variety of journals require that all authors using microarray data analysis in their research submit a complete dataset to a public repository in order to publish. As of 2011 GEO stores over half a million distinct microarray samples, meaning results of distinct gene expression experiments for a wide range of organisms from yeast to humans. It should be noted that the growing demand for publicly available microarray data has also stimulated the need to set some general standards regarding the format and the information accompanying each dataset, to allow subsequent analysis by different researchers. For example the Microarray Gene Expression Data Society (MGED 2011) advocates open access to genomic datasets and works towards developing standards for data quality, annotation and exchange. MIAME, which stands for the Minimal Information About a Microarray Experiment (Brazma et al. 2001), is designed to help authors, who submit microarray data, ensure that the data meets some minimum requirements, allowing other researchers to interpret the results of the experiment unambiguously and potentially to reproduce the experiment.

2.8 Pathway based microarray analysis

The huge wealth of information accumulated in recent times in distinct fields of biological knowledge along with the expanding influence of systems biology on contemporary research approaches have led to growing interest in data integration (Hwang et al. 2005; Bourguignon et al. 2010). Microarrays have for some time now been producing lists of differentially expressed genes, under experimental conditions of interest. Statistical analysis, clustering and classification, discussed in section 2.5.3, have been some of the basic choices of analytical approaches to facilitate expression data. There is a general notion that such approaches have not been able to leave up to the initial enthusiasm, often producing cryptic results (Werner 2008). It was soon realised that there is a need to move beyond, and come up with novel analytical methodologies to efficiently analyse global gene expression (Altman & Raychaudhuri 2001).

One such approach is to look at differentially expressed genes in terms of predefined lists, related to biological functions. For example gene-ontologies (GO), that is, an effort for clear and relatively simple gene classification according to the functional properties of their protein products (Sidhu et al. 2007), have been quite popular. A comprehensive review of GO based microarray analysis approaches can be found in (Ochs et al. 2007).

Pathway based microarray data analysis is a similar methodology, that aims to integrate microarray data analysis with biochemical pathway knowledge. Rather than concentrating on the often subtle change occurring in the expression of individual genes, gene expression analysis is facilitated to identify coordinated changes occurring in the expression of sets of genes, forming biochemical pathways (Cavalieri et al. 2007). This is a sensible choice given that an expression increase of 20% in genes encoding enzymes of a metabolic pathway may have a significant effect on the flux through the pathway, which may be more important than a huge increase in the expression of a single gene (Subramanian et al. 2005). Furthermore, as aforementioned deregulation of signalling cascades has major involvement in pathogenesis, notably in cancer development. Thus, differential expression analysis, in terms of pathways and regulatory networks has drawn considerable interest in biological and bioinformatics research (Keller et al. 2009). It holds promising potential in deciphering the functional state of a cell at the level of the underlying biochemistry.

The development of pathway databases, discussed in section 2.6.1, providing a collection of the components of metabolic and regulatory pathways, has been of upmost importance in this line of research. Due to the high rate of growth of relevant literature that needs to be constantly assimilated, academic efforts in the area of pathway based microarray analysis rely heavily on pathway databases (Werner 2008).

2.8.1 Available software

In (Kurhekar et al. 2002) the authors have proposed an interesting method for the analysis and visualisation of microarray data in metabolic and regulatory pathways in order to elucidate the effect of stimuli on these genetic networks. They combine gene expression data series with metabolic pathway data from KEGG for a number of organisms to score pathways according to three distinct criteria, namely activity, co-regulation and cascade effects. In brief, these refer to the proportion of differentially expressed genes in a pathway, the degree of correlation of gene expression per pathway and the degree of activation of genes along reaction chains per pathway, respectively.

Eu.Gene Analyzer (Figure 2.14) is a stand-alone application that allows microarray data analysis in the context of biological pathways and any other functional grouping of genes, such as gene-ontologies. This tool can be used to visualize expression data on metabolic pathways and to evaluate which metabolic pathways are most affected by transcriptional changes in whole-genome expression experiments (Cavalieri et al. 2007).

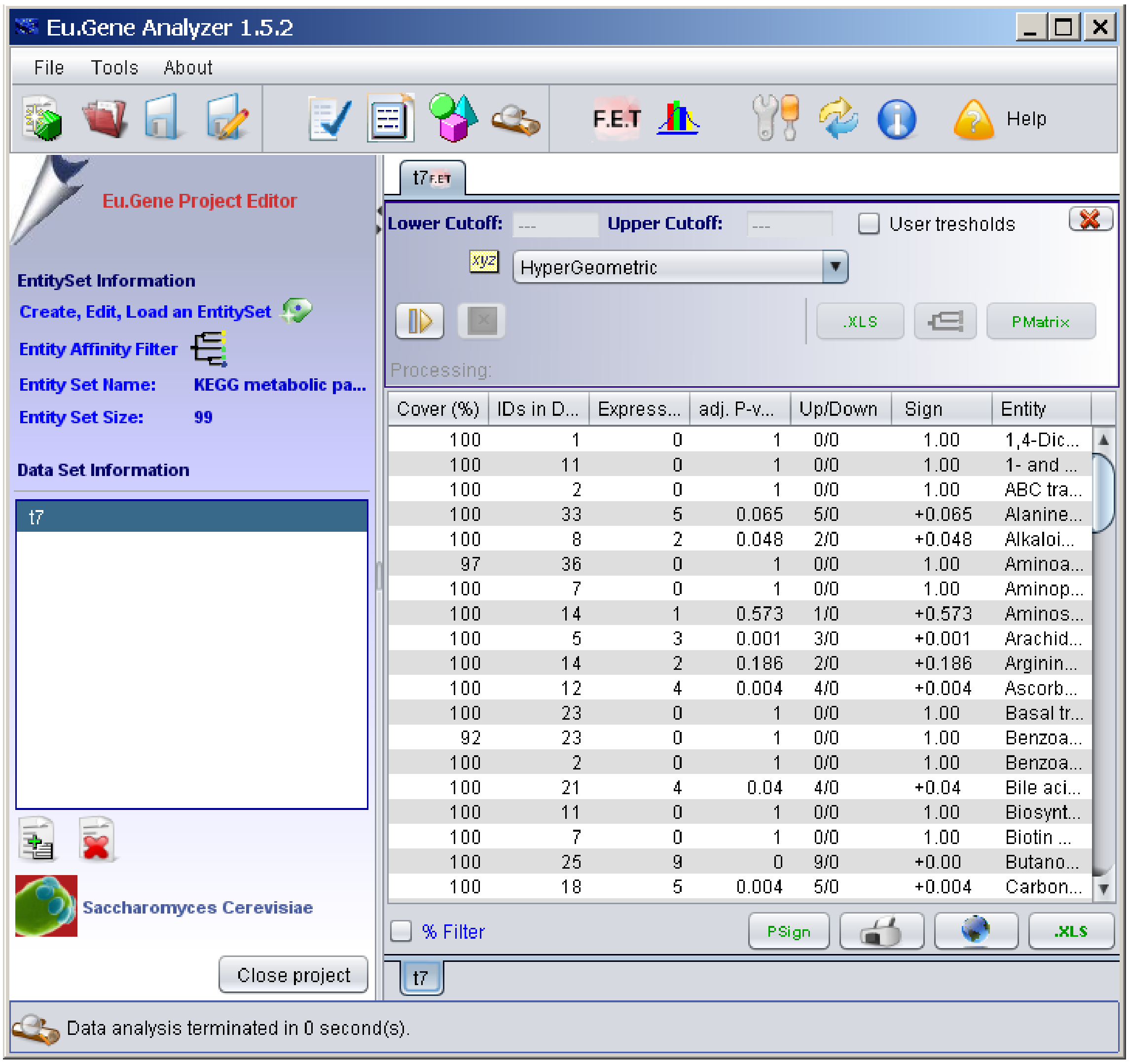
The scoring of pathways, in view of the effect of the experimental conditions on their activity, is based on the application of two different statistical approaches, the fisher exact test (FET) and gene set enrichment analysis. Regarding FET the software estimates the probability of observing as many or more differentially expressed genes in a pathway of given size, purely by chance, given the null hypothesis that the number of regulated genes in a pathway is random subset of regulated genes observed in the experiment as a whole.

Gene set enrichment analysis (GSEA, Figure 2.15) itself is a popular approach to analyse microarray data at the level of gene sets, be it sets of genes belonging to experimentally defined biochemical pathways, genes of the same chromosomal location or genes of similar ontology (Subramanian et al. 2005). The method uses a list *L*, where genes are ranked according to the degree of differential expression they show in a collection of microarray datasets corresponding to two different experimental conditions, and examines whether members of any gene set tend to occur toward the top or bottom of list *L*, thus showing up or down regulation.

Application of the method has produced promising results, for example revealing reduced expression of genes involved in oxidative phosphorylation in diabetics, which has been independently validated by in vivo functional studies (Petersen et al. 2004).

GSCope constitutes another example of popular software for pathway expression analysis (Toyoda , Mochizuki & Konagaya 2003). It provides a nice visualisation allowing the user to look at pathways from different levels, zooming in a more detailed view or zooming out to obtain an overview when appropriate. Other reference examples of similar software tools include GenMapp (Dahlquist et al.

2002), Cytoscape (Shannon et al. 2003), Pathfinder (Goesmann et al. 2002) and GeneNet (Ananko et al. 2002).



EntitySet Information

Create, Edit, Load an EntitySet

Entity Affinity Filter

Entity Set Name: KEGG metabolic pa...

Entity Set Size: 99

Data Set Information

t7

Saccharomyces Cerevisiae

Lower Cutoff: --- Upper Cutoff: --- □ User tresholds

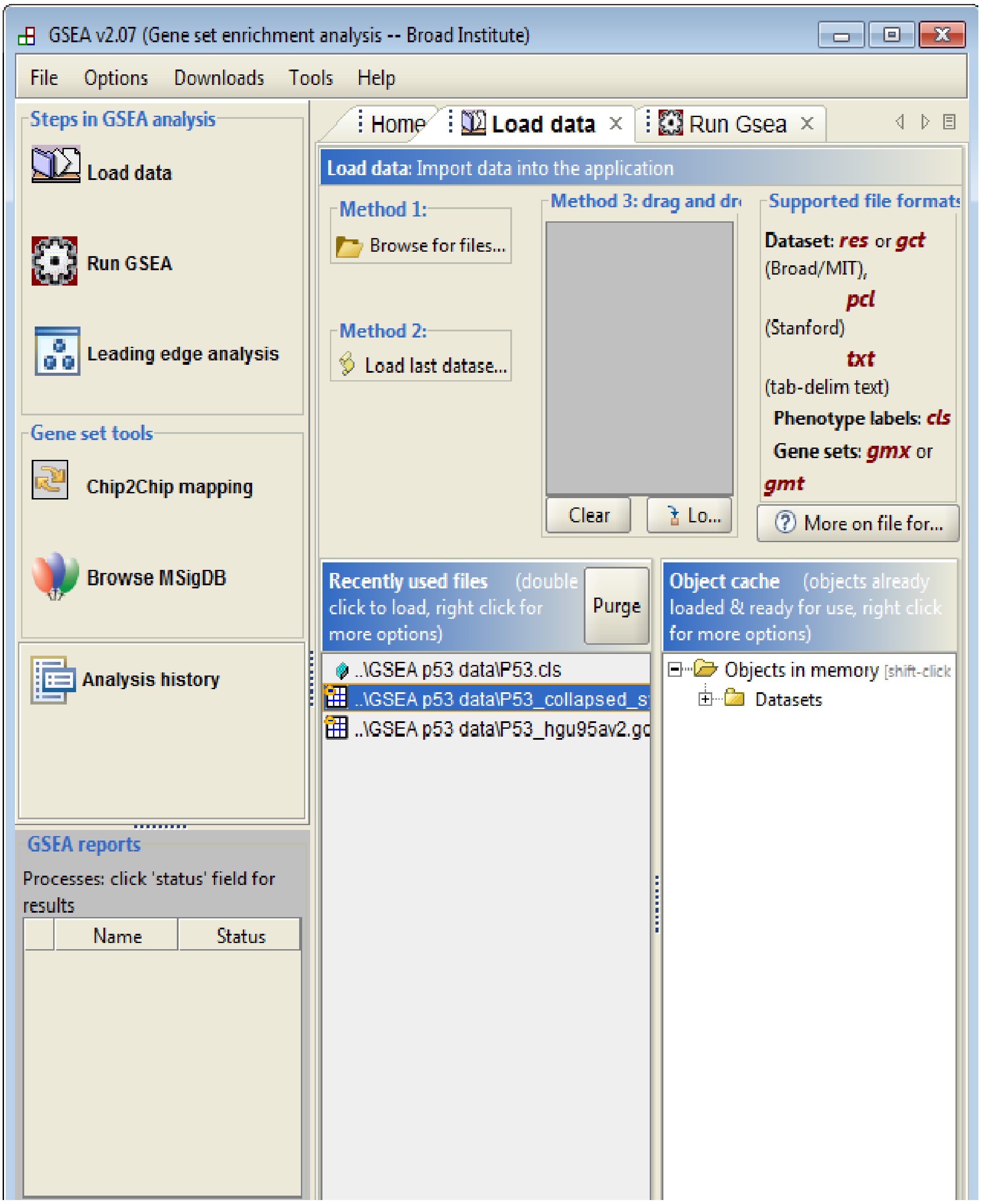
HyperGeometric

Cover (%)	IDs in D...	Express...	adj. P-v...	Up/Down	Sign	Entity
100	1	0	1	0/0	1.00	1,4-Dic...
100	11	0	1	0/0	1.00	1- and ...
100	2	0	1	0/0	1.00	ABC tra...
100	33	5	0.065	5/0	+0.065	Alanine...
100	8	2	0.048	2/0	+0.048	Alkaloi...
97	36	0	1	0/0	1.00	Aminoa...
100	7	0	1	0/0	1.00	Aminop...
100	14	1	0.573	1/0	+0.573	Aminos...
100	5	3	0.001	3/0	+0.001	Arachid...
100	14	2	0.186	2/0	+0.186	Arginin...
100	12	4	0.004	4/0	+0.004	Ascorb...
100	23	0	1	0/0	1.00	Basal tr...
92	23	0	1	0/0	1.00	Benzoa...
100	2	0	1	0/0	1.00	Benzoa...
100	21	4	0.04	4/0	+0.04	Bile aci...
100	11	0	1	0/0	1.00	Biosynt...
100	7	0	1	0/0	1.00	Biotin ...
100	25	9	0	9/0	+0.00	Butano...
100	18	5	0.004	5/0	+0.004	Carbon...

Data analysis terminated in 0 second(s).

Figure 2.14 A snapshot of Eu.Gene Analyser. A software tool developed for scoring gene sets, such as pathways and other functional groups. Each set is scored according to the proportion of differentially expressed genes it contains, in relation to other sets and the global expression of genes in the microarray dataset.

Overall, in most cases of such tools, a pathway database is superimposed onto a single microarray experiment in order to visualise the expression of individual genes forming the pathway in a collective view. Despite the large variety of methods they all take into account the expression of all the genes in a pathway, thus all methods are based on some type of averaging. However, it is important to note that genes in a biochemical pathway often show quite variable behaviour in terms of RNA production. Making sense of pathway activity based on expression data does not constitute a straightforward, simple task and a significant level of reluctance by the biological community to facilitate relevant tools has been reported (Saraiya, Chris North & Duca 2005).



GSEA v2.07 (Gene set enrichment analysis -- Broad Institute)

Load data: Import data into the application

Dataset: res or gct (Broad/MIT), pcl (Stanford) txt (tab-delim text) Phenotype labels: cls Gene sets: gmx or gmt

Recently used files (double click to load, right click for more options)

..\GSEA p53 data\P53.cls
..\GSEA p53 data\P53_collapsed_s
..\GSEA p53 data\P53_hgu95av2.gc

Object cache (objects already loaded & ready for use, right click for more options)

Processes: click 'status' field for results

Figure 2.15 GSEA software tool. Snapshot of GSEA software tool for the identification of gene sets, significantly enriched in differentially expressed genes.

Furthermore, it is not uncommon to observe simultaneous up- and down-regulation of closely related genes, in the same pathway, in the same experiment, as discussed in some detail in chapter 3. Naturally, identifying the state of activity of pathways exhibiting such behaviour is not a trivial task and requires further elaboration.

2.9 Conclusions

Visualisation is very useful, allowing the domain expert to gain insights into available data. It has been explored in some detail in the context of pathway based microarray analysis. However, biologists have expressed concerns regarding the advantages offered by available visualisation tools and a significant level of reluctance in facilitating them in their research, partly due to their complexity (Saraiya, Chris North & Duca 2005). Furthermore, visualisation alone does not interpret the data in hand. In that sense combining and intertwining visualisation and analytical methods can be very beneficial. In the analysis of biochemical pathway behaviour, centred on gene expression data, there is still considerable room left for speculation, and deciphering pathway behaviour remains a challenging task.

Additionally, it is often difficult for biologists to adopt to the current approaches to analysing expression data, which rely heavily on mathematics and complicated computational methodologies. Traditionally, biology has been an experimental science, mainly wet lab based, and it would be greatly beneficial to find ways to provide the researcher with simple and clear insight into the microarray data in hand.

This constitutes the main motivation behind the research presented in this thesis. It is an effort to add some contribution to pathway based microarray analysis and assist the biologist to interpret gene expression data with greater clarity and confidence. The analytical approach is centred on the collective expressional behaviour of all gene members of pathways in order to identify those truly associated with the observed pathway states, as discussed in the following chapter.

Chapter 3: Central hypothesis

3.1 Introduction

Gene expression analysis using microarray technology has been accompanied by the development of a large variety of analytical methodologies that attempt to exploit our ability to monitor the global response of gene activity to various experimental conditions. As previously discussed, following the popularity of the systems approach to biological research, gene expression analysis has not escaped the trend for data integration. There is an on-going effort to combine distinct experimental techniques in a holistic analytical approach, allowing us to arrange pieces of the puzzle that biological knowledge is and obtain a clearer picture of the behaviour of living systems.

Pathway based microarray analysis is an important effort to observe the behaviour of genes forming defined biochemical pathways and thus draw meaningful conclusions regarding the state of a cell or an organism. This chapter presents a discussion of the basic research hypothesis underlying this work and an attempt to validate it examining the behaviour of genes that constitute members of distinct biochemical pathways in large datasets of microarray experiments. The analysis, presented here, incorporates the notion of single- and multiple-membership, in respect to the participation of a gene in one or more biochemical pathways respectively. We facilitate different approaches to explore the behaviour of these types of genes, including expression frequency, correlation analysis and association rule mining.

3.1.1 Rationale and Motivation

A biochemical pathway can be seen as a collection of genes whose protein products collaborate in a highly organised fashion to produce a desired outcome. The entire collection of genes, members of distinct pathways, forms a complicated, highly integrated network. To cover our intuitive needs we break down this network into smaller parts, namely pathways, consisting of genes with close functional relationships that are responsible for a particular, well-defined cellular task. For example we can assume the following hypothetical pathway where a cell needs to produce proteins A, B, C, D and E, which in turn catalyse five steps of a biochemical process starting from an initial chemical substrate and leading to the production of a chemical compound required by the organism (Figure 3.1).

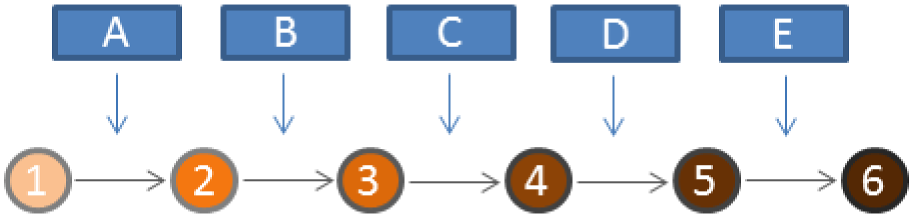


Figure 3.1 Hypothetical biochemical pathway. Enzyme A catalyses the conversion of compound 1 into compound 2, which is then transformed into compound 3 with the contribution of the enzymic activity of B and so on.

A parallel would be a production line in a car factory or a house building project where each molecule represents a certain step in the process. Different people involved in the project have different skills required to fulfil their tasks at different stages of the process. Some are needed to dig, others to lay the foundations of the house, others to paint or fit the electrical wiring, with the ultimate aim to build a functional house. In a similar fashion different proteins encoded by the genes in a pathway, have their individual tasks which they need to perform in a coordinated fashion for the pathway as a whole to be functional. In this analogy the organism can be seen as a city, where building, renovation and demolition projects constantly take place at different sites.

Ideally, in pathway based microarray analysis we would expect that activation of a pathway would allow us to observe an increase in the activity of genes forming the pathway in question and vice versa, following the above rationale. However, in practice we often observe quite variable behaviour in terms of expression, for gene members of the same pathway, with some showing up-regulation, others down-regulation and others stable expression at the same time, in the same experiment. Naturally, such observations raise issues and confer pathway based microarray analysis a tricky, non-trivial analytical methodology. It would be greatly beneficial for a biologist to find analytical approaches that aid him in better interpreting such contradictory results in the effort to elucidate pathway states using microarray data.

3.1.2 Common issues related to variable expression of genes in a pathway

There are a number of reasons discussed in available literature that provide some justification of the observed discrepancy in the expression of closely related genes forming a pathway. Above all, genes are characterised by large diversity, given that genes that are members of the same pathway may encode proteins of very different functionality, with some being transcription factors acting in the cell nucleus to facilitate the expression of other genes, while others transmembrane proteins (Stryer & Tymoczko 2006). Thus it is not surprising to observe that they respond differently in terms of how their corresponding RNA levels are affected in various conditions.

Previous work in the field has already targeted this issue and suggested a plausible explanation (Panteris et al. 2007). In brief, there is a flow of enzymatic activity taking place in an organism as a whole and in each of its pathways which, as already mentioned, consists of a network of biochemical reactions. In this collection of reactions there are rate determining steps, a common notion in chemistry (Zumdahl 2005), that describes the fact that the slowest step in a reaction is the one to determine its speed. Consequentially, it is likely that while certain genes in a pathway are important for the flow, probably encoding structural proteins, others are rate controllers, for example encoding enzymes and signalling molecules. The latter ones are likely to be more drastically affected by changes in the experimental conditions, in terms of RNA production rates. Hence their intensity values are more representative of the pathway's state.

Additionally, while the term gene expression is widely used as reference to DNA microarrays technology, due to the popularity of the technique, this use of the term is not entirely correct. To be more precise, in the case of protein encoding genes, gene expression describes the entire process starting with the transcription of a gene onto an mRNA molecule and finishing with a functional protein molecule. During this process there are a number of intermediate regulatory stages, including the translation of the mRNA molecule into a protein, while the rates of protein maturation and degradation also have a severe effect on the activity of a protein (Quadroni & James 1999).

One must also take into account the variety of post-translational modifications, such as phosphorylation, methylation and so on, which can regulate the function of proteins. Such modifications modulate the activity of most eukaryotic proteins and can turn a protein from an active into an inactive state and vice versa, affect their cellular location and their dynamic interactions (Mann & Jensen 2003; Seo & Lee 2004). Consequentially, gene expression alone may often be insufficient evidence of gene functionality, in terms of the abundance and, by extension, state of activity of its protein product (Greenbaum et al. 2003).

In addition, microarray technology itself is accompanied by a number of limitations, as it involves numerous error-prone experimental steps and requires the physical disruption of cells to gain access to their gene expression patterns (Russo , Zegar & Giordano 2003). The presence of noise may to distort any analytical approach facilitating microarray data. It should be noted, however, that popular beliefs regarding the extent to which noise is present in microarray data have been recently questioned (Klebanov & Yakovlev 2007).

3.2 Single- and Multi-membership genes

While considerations regarding the regulation of genes and microarrays technology discussed in the previous section may provide some plausible reasons for the observed discrepancy in the expression of genes belonging to the same biochemical pathway, we identify an additional issue of importance that to our knowledge has been overlooked and can provide deeper insight into the collective behaviour of pathways.

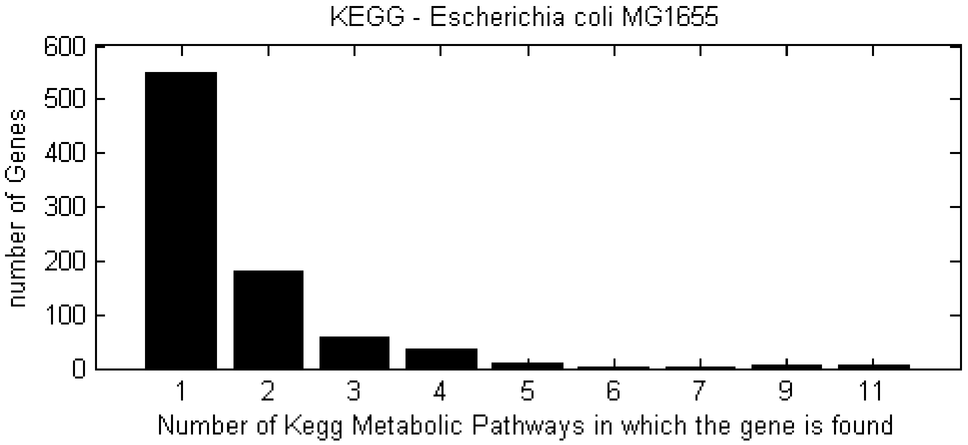


Figure 3.2 Number of *E.coli* genes of various membership degrees in KEGG metabolic pathways.

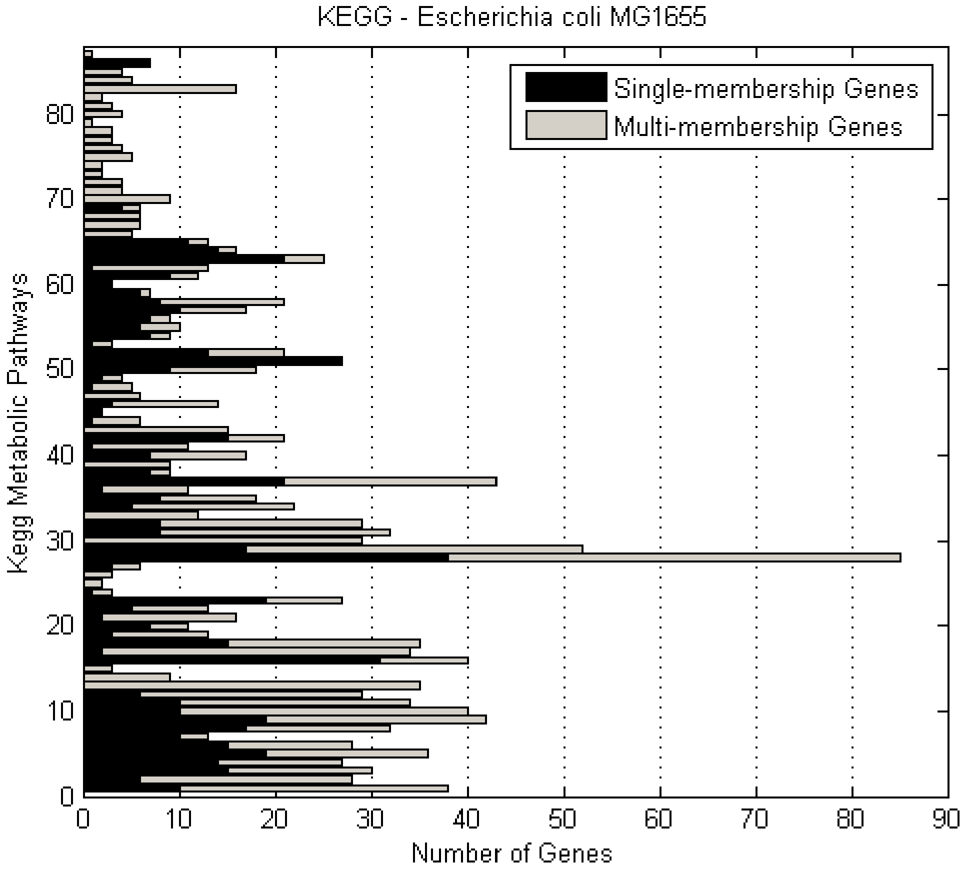


Figure 3.3 Number of *E.coli* single- and multi-membership genes per KEGG metabolic pathway.

In particular, a closer look at the Kyoto Encyclopaedia of Genes and Genomes Pathway database reveals that a number of genes in an organism constitute members of two or more biochemical pathways. Figure 3.2 graphically portrays the pathway membership of *Escherichia coli* metabolic genes, according KEGG (30/11/2010). Evidently, more than a third (35%) of the 849 unique *Escherichia coli* genes, present in KEGG metabolic pathways, are members of at least two biochemical pathways.

Figure 3.3 is also quite revealing, representing the number of genes present in each of the *Escherichia coli* KEGG metabolic pathways. Genes in black colour are unique members of the pathway in hand while genes in grey are members of two or more biochemical pathways.

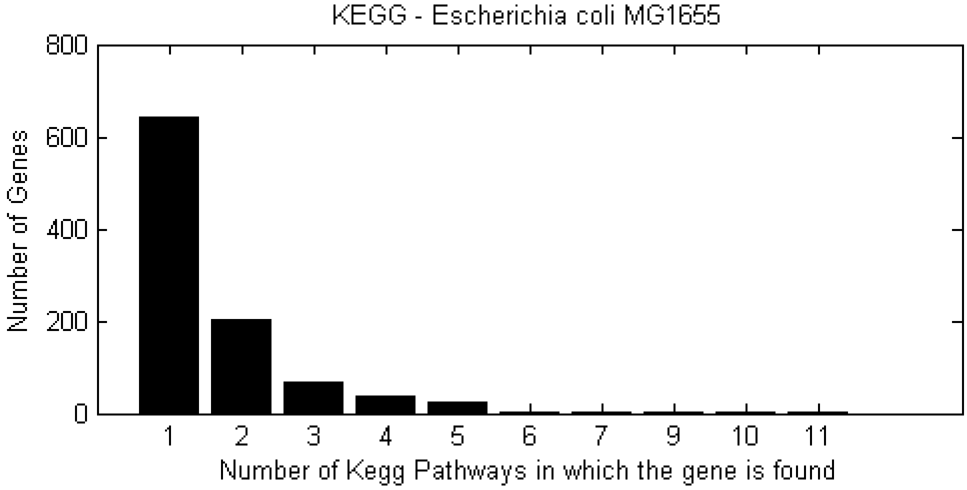


Figure 3.4 Number of *E.coli* genes of various membership degrees in all KEGG pathways.

On average around 40% of the genes in each path are unique members of that path, while more than 60% of the genes, constitute a part of at least one other pathway. In some extreme cases, pathways consist solely of genes of the latter category. Hereafter, we shall refer to such genes as multi-membership genes to distinct them from genes, unique members of one and only pathway, to which we shall refer as single-membership.

Notably, besides metabolic pathways, KEGG contains a number of other types of pathways. These include signalling pathways for environmental information processing, pathways for genetic information processing and various cellular processes. When the entire collection of pathways is taken into account the number of unique genes taking part in them as well as the number of genes shared by the totality of the KEGG pathways grows even further. This information is summarised on Figures 3.4 and 3.5.

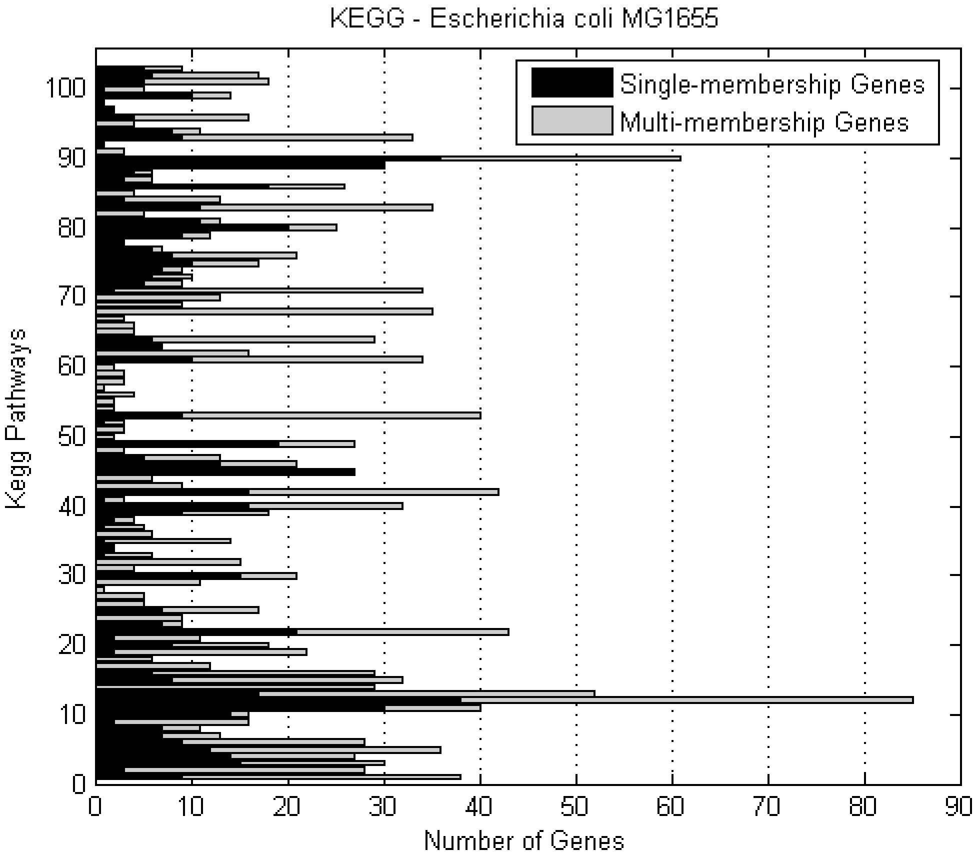


Figure 3.5 Number of *E.coli* single- and multi-membership genes per KEGG biochemical pathway.

Naturally, this phenomenon is not confined to *Escherichia coli*. In fact it applies to virtually all organisms whose genomes are currently stored in the KEGG database. To look into another characteristic example, *Saccharomyces cerevisiae*, the well-known budding yeast, which constitutes another popular experimental subject, shows similarly high presence of multi-membership genes in its pathways. On average 56% of the genes in each metabolic and 58% in the totality of KEGG pathways, respectively, are members of at least one more biochemical pathway. Figures 3.6 to 3.9 visually represent the data, similarly to *Escherichia coli* discussed above.

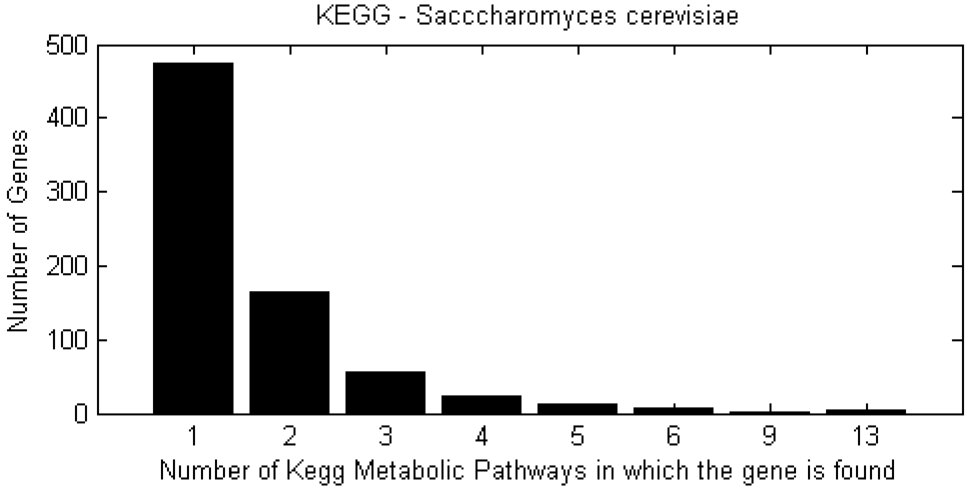


Figure 3.6 Number of *S.cerevisiae* genes of various membership degrees in KEGG metabolic pathways.

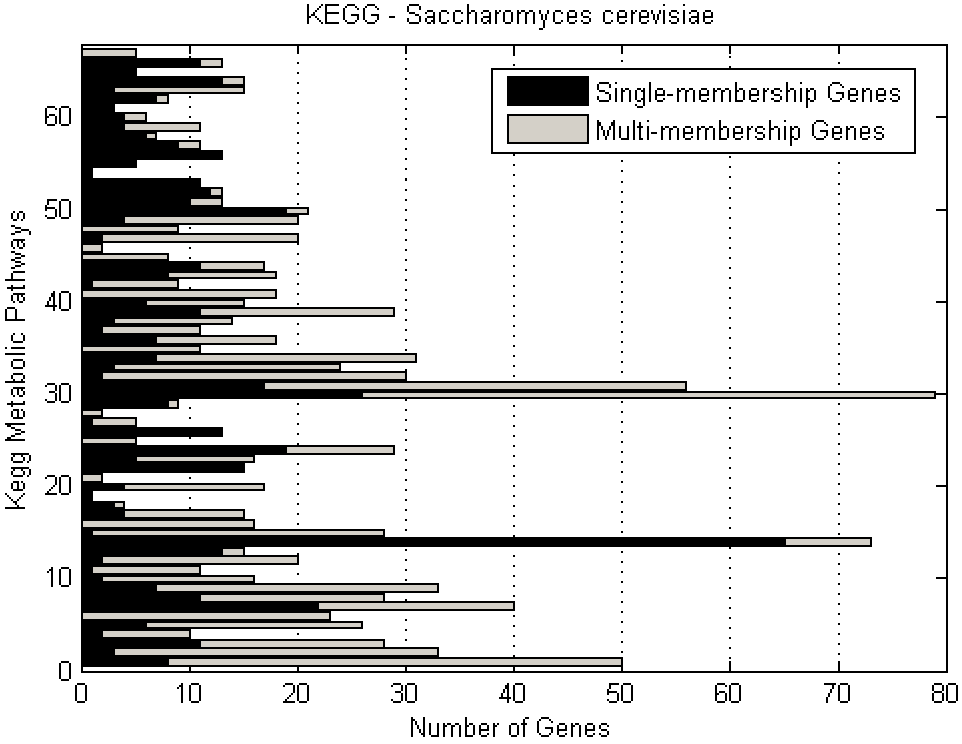


Figure 3.7 Number of *S.cerevisiae* single- and multi-membership genes per KEGG metabolic pathway.

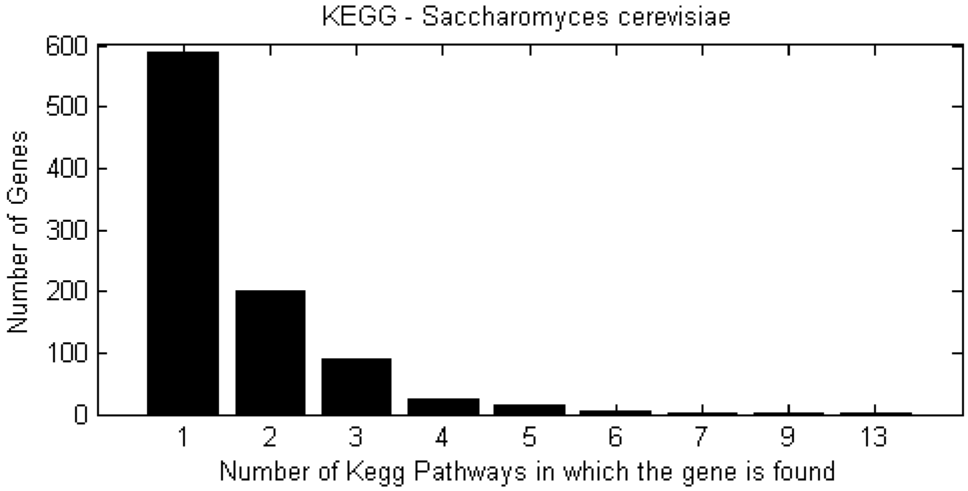


Figure 3.8 Number of *S.cerevisiae* genes of various membership degrees in all KEGG pathways.

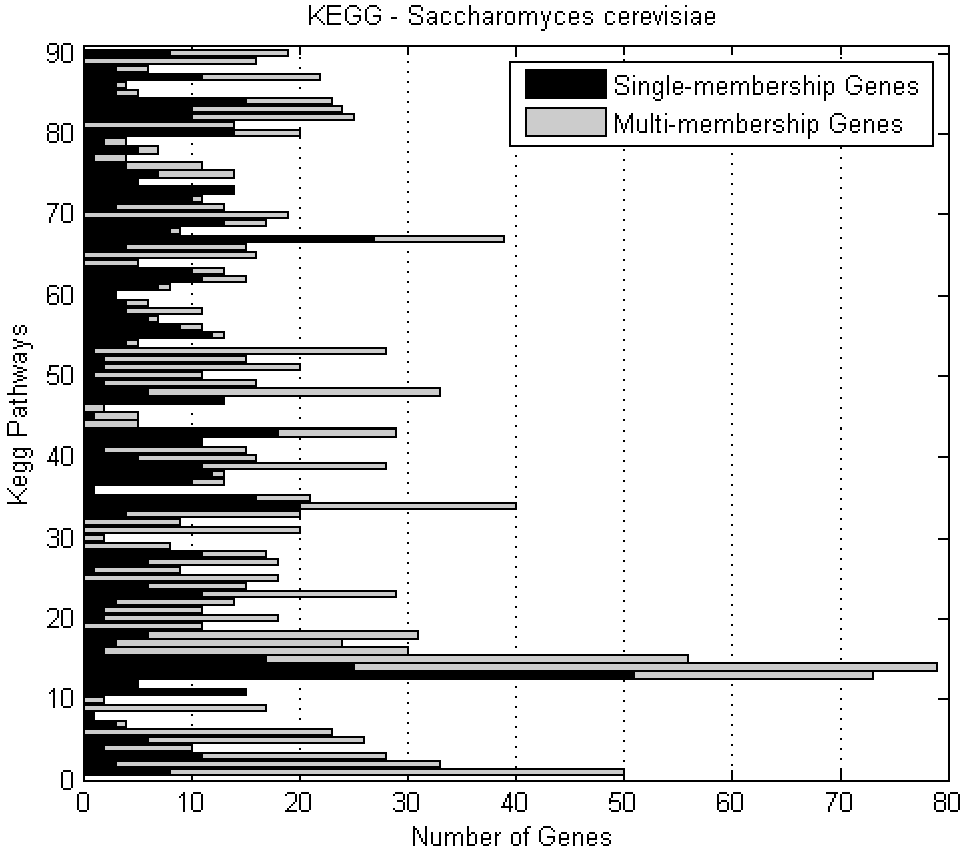


Figure 3.9 Number of *S.cerevisiae* single- and multi-membership genes per KEGG biochemical pathway.

For organisms higher in the evolutionary chain, with more complicated biochemical networks and larger genomes, the overlap of genes in distinct pathways grows significantly. Indicatively, KEGG contains 84 metabolic pathways of a total 216 pathways for *Homo sapiens* (30/11/2010). Figures 3.10 and 3.12 exhibit the various degrees of human gene membership, for metabolic and, biochemical pathways respectively.

The number of human genes allocated to pathways is 2169, significantly higher than in the previous examples, and the same is true for the 1458 unique genes allocated to *Homo sapiens* metabolic pathways. Figures 3.11 and 3.13 provide a visual representation of the number of single- and multi-membership genes present in each pathway, for metabolic pathways and the entire collection of KEGG pathways respectively. In the case of *Homo sapiens* each pathway contains a much higher proportion of genes shared by other pathways, than for the other discussed organisms.

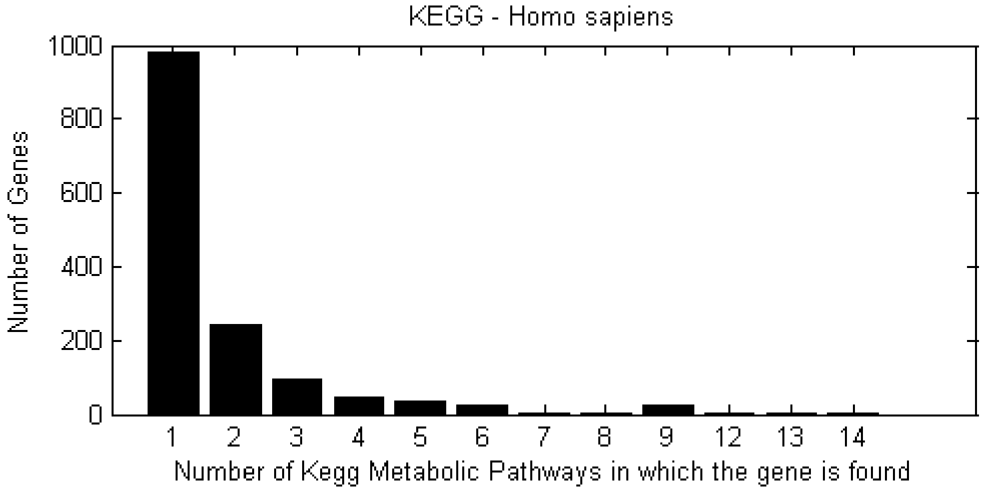


Figure 3.10 Number of *H.sapiens* genes of various membership degrees in KEGG metabolic pathways.

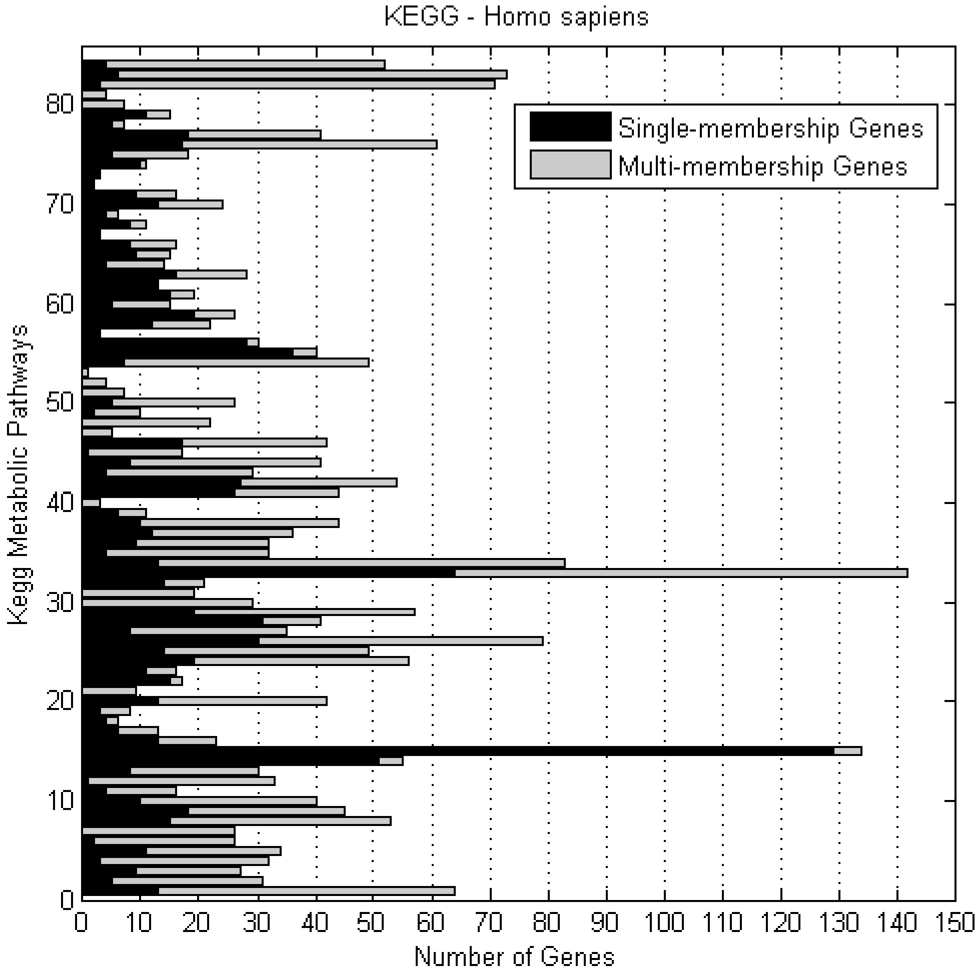


Figure 3.11 Number of single- and multi-membership *H.sapiens* genes per KEGG metabolic pathway.

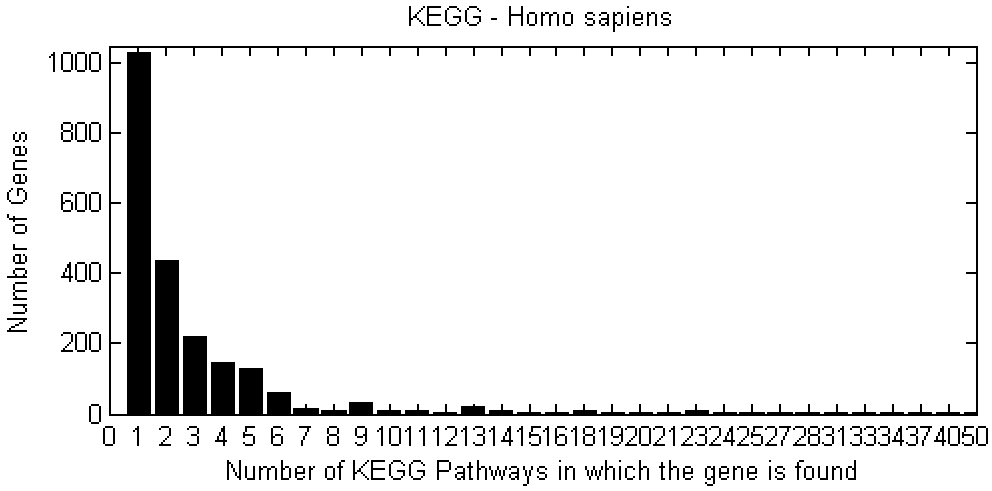


Figure 3.12 Number of *H.sapiens* genes of various membership degrees in all KEGG pathways.

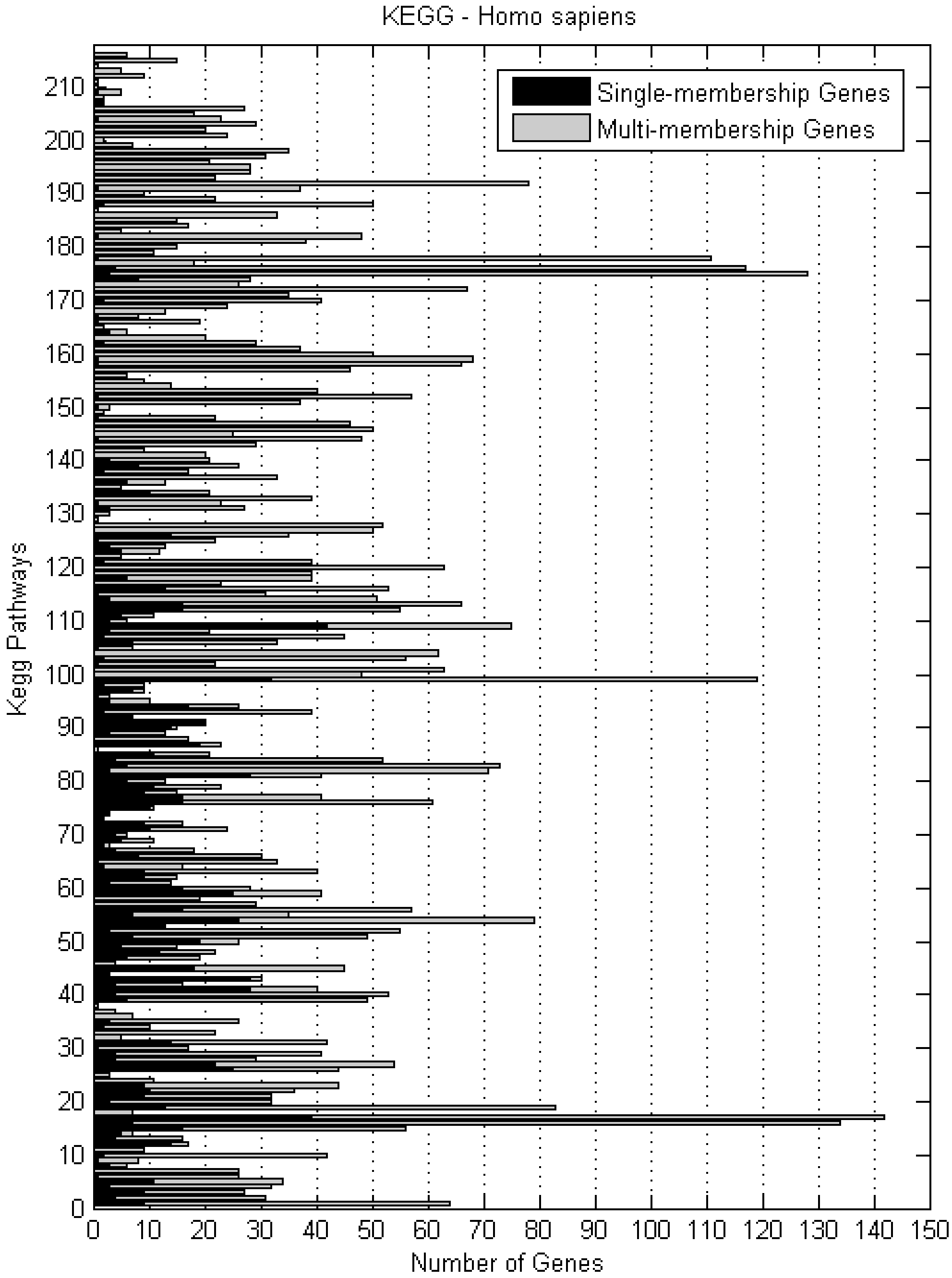


Figure 3.13 Number of single- and multi-membership *H.sapiens* genes per KEGG biochemical pathway.

In particular, in the case of the entire collection of *Homo sapiens* KEGG pathways the average percentage of multi-membership genes in each pathway is 80%.

3.3 Hypothesis

The preceding section has established that the lists of genes present in the biochemical pathways of sequenced organisms, stored in the KEGG pathway database, are characterised by significant overlaps, due to the presence of multi-membership genes. This observation has some interesting implications for pathway based microarray analysis. Namely, at any particular instance in time, the same gene may be potentially suppressed or stimulated to produce proteins that fulfil two or more distinct functions.

For example let’s assume for the case of the hypothetical pathway on Figure 3.1 discussed earlier, that proteins B, D and E are also members of one or more other pathways. In such a setting it may well be that under certain conditions the pathway is activated, while at the same time the other pathway or pathways in which the genes participate are severely repressed. It would then be quite likely to observe down regulation of some of the genes, even though the pathway we are looking at is in fact stimulated.

Similarly, in a situation where the hypothetical pathway is not affected by the experimental conditions, but the other pathways in which the genes participate are activated or repressed, we may observe differential expression of genes B, D and E that is not related to the activity of that particular pathway. If we were to examine it in isolation we might be driven towards the false conclusion that the pathway is activated or perhaps observe up-regulation of some and down-regulation of other genes, which would look contradictory and make it tricky to draw a meaningful conclusion regarding its state of activity.

In the analogy of the production line discussed earlier, it would be similar to observing a rapid increase in the activity of some workers but reduction in the activity of others. If this was our measure of establishing the state of productivity of a factory or to investigate if a new building project has started, as we do for biochemical pathways based on the activity of the genes forming them, we would probably reach the wrong or no conclusion at all.

Following the above rationale, we have established a plausible hypothesis that the expression exhibited by multi-membership genes represents the net effect of their contribution to each one of their constituent pathways. For example an up-regulated gene that constitutes a member of two pathways may be in such a state of expression due to its contribution to either or both of these pathways. The biological system regulates and controls the activity of its genes in a way that they contribute to the function of the biochemical network it governs in a manner that satisfies its needs in varying conditions.

In that sense, observing contradicting expression values for gene members of well-defined pathways is not surprising. In fact, if we were to directly observe and study the activity of the proteins encoded by the genes in a pathway we may not encounter any discrepancy. However, microarrays do not give us that information. Rather they can only tell us that a gene is more or less actively transcribed. While it has been shown that there is in some cases a correlation between the values of gene expression and the activity of the respective proteins, it is not always indicative on its own (Gygi et al. 1999). The proteins remain the functional molecules in a living cell. The expression of a gene that encodes a protein that can eventually be facilitated in various cellular processes does not reveal to us anything about this latter stage. To go back to the analogy of a house building project, microarrays serve as the information provided by an employment agency, telling as that on a certain day, twice as many painters where required in the area, without saying anything about the projects they are working on. It can be that work has started on a new house, or that work has started on two or more new houses, but it can also be that a house building project has been completed but at the same time a new bigger project has started.

3.4 Expression frequencies

In an effort to support our hypothesis we have performed comparative analysis of the frequency of expression of single- and multi-membership genes in large number of unrelated microarray experiments. For this purpose we facilitated data obtained from GEO, which consists of $\log_2$ transformed intensity ratios from more than 5000 microarray experiments on *Saccharomyces cerevisiae*. In cases of duplicate genes on the same chip intensity values have been averaged. For the normalised data values above 1 and below -1 standard deviation were considered as evidence of positive and negative regulation respectively.

We examined the number of experimental instances where each unique single- and multi-membership gene appears differentially expressed and the mean value corresponding to the totality of KEGG genes in each group. Following this approach we observed a mean value of 708 in about 5000 experiments, which corresponds to about 13.6% for genes that are members of one and only pathway. For genes that are members of two or more pathways the mean increases to 859, which corresponds to about 16.5%, thus, 2.9% higher than in the previous group. Table 3.1 below exemplifies the rationale.

Table 3.1 Average expression of hypothetical genes A and B, In 6 experiments. Ones represent up-regulation, minus ones down-regulation and zeros stable expression of the respective gene. Based on the table we can estimate the average expression of the group consisting of genes A and B.

Experiment	1	2	3	4	5	6	Average
Gene A	1	0	0	1	-1	0	3/6
Gene B	1	-1	1	1	0	0	4/6
Total Mean							0.583

Notably, a two sample t-test reveals that the proportions of expressed single- and multi- membership genes per experiment in the microarray dataset are different with a very low *p*-value of about 6.6×10^{-9}, which is highly significant.

Interestingly, the difference is more apparent when we only examine experiments with at least 5% of *Saccharomyces cerevisiae* pathway genes expressed, thus applying a threshold of expressed genes to incorporate an experiment in the analysis. This criterion is met by a subset of 3553 experiments in our dataset, where the average single-membership gene expression is 19.1% while the average multi-membership gene expression reaches 23.1% which corresponds to a difference of 4%. Again, single-membership gene expression appears significantly less frequent than that of multi-membership genes with a two sample t-test producing a *p*-value of 5.1×10^{-9}.

Moreover, the difference increases further when we consider genes with higher degree of membership, meaning genes constituting members of at least three biochemical pathways, gene members of at least four biochemical pathways and so on, as exemplified on Figure 3.14.

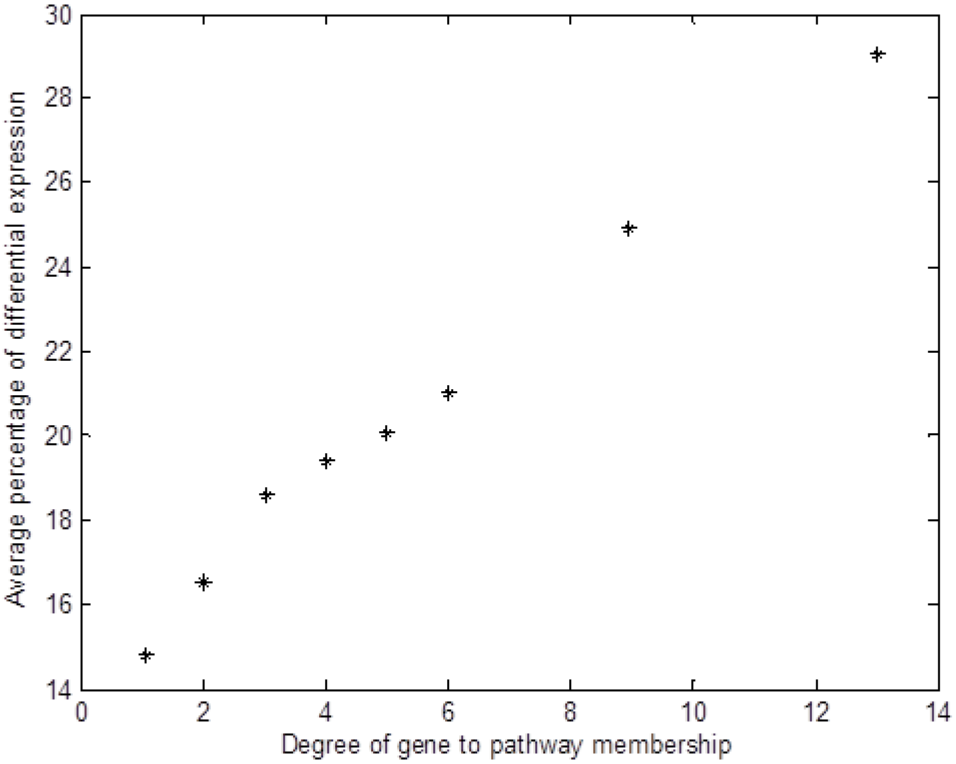


Figure 3.14 Average percentage of differential expression, for different membership degree thresholds. For each threshold the plot shows the average percentage of expression for genes of membership equal to the specific threshold or higher.

In agreement with our hypothesis, as the degree of minimum gene membership increases the average expression follows suit. Genes that are members of three or more pathways are clearly more frequently expressed than genes that are members of only two distinct pathways. Similarly genes that are members of four or more pathways are more frequently expressed than gene members of two or more pathways and so on.

If we examine the exact degree of membership in relation to the average expression of genes, we obtain the results on Figure 3.15. Apparently, for the case of genes with 6th and 9th degree of membership there is a deviation from the general pattern, however these only concerns 6 and 1 gene respectively and thus is not representative.

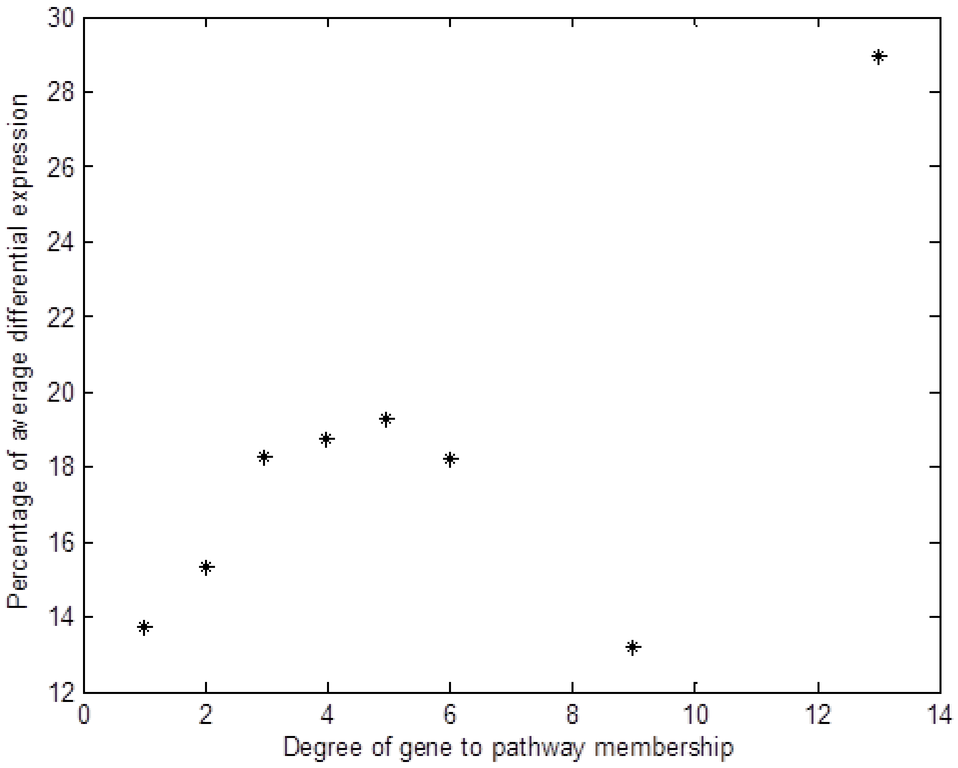


Figure 3.15 Percentage of average differential expression for genes of distinct membership degree bands.

Looking into these from another perspective we can examine the proportion of single-membership genes expressed in each microarray experiment in respect to the proportion of multi-membership genes expressed in the same experiment. Figure 3.16 shows the number of expressed single-membership genes in each experiment as proportion of the total number of single-membership genes in comparison to the number of regulated multi-membership genes as proportion of the total number of multi-membership genes on the platform. Again, we observe that in the majority of experiments the proportion of differentially expressed multi-membership genes is higher. In contrast, in just 20% of the experiments the proportion of differentially expressed single-membership genes is only marginally higher than the one of multi-membership genes.

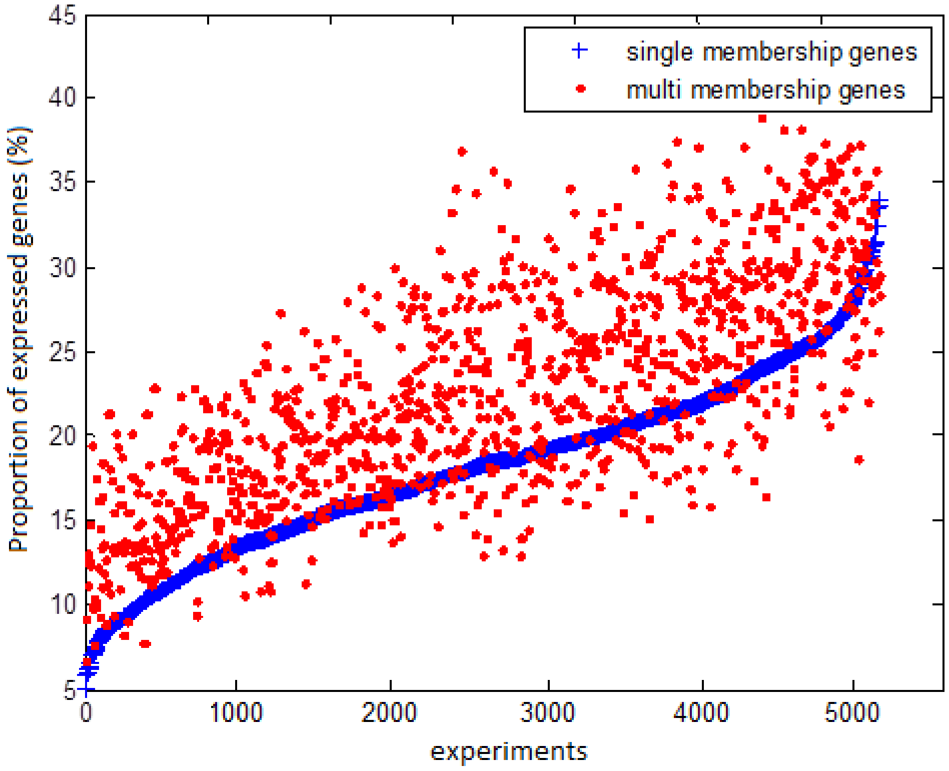


Figure 3.16 Percentage of regulated single- and multi-membership genes. Number of expressed single- and multi-membership genes, as proportion of the total number of single- and multi-membership genes respectively.

To further strengthen our hypothesis we also look into individual pathways and examine the average proportion of expressed single- and multi-membership genes per pathway, in the dataset. The results of this analysis are exhibited on Table 3.2.

Overall, the average expression of multi-membership genes per pathway seems to surpass the expression of single-membership genes both in numbers and instances. In a total of 46 pathways which contain both unique and multi-membership genes, 29 exhibit higher expression of the first as opposed to 17 of the latter group, while the average difference is 5.4 and 3.1 respectively. A two-sample t-test rejects the null hypothesis that the average expressions for the two groups come from normal distributions of equal means.

Table 3.2 Average proportion of expressed single- and multi-membership genes per *Saccharomyces cerevisiae* KEGG pathway.

KEGG Path ID	Singles %	Multi %	KEGG Path ID	Singles %	Multi %	KEGG Path ID	Singles %	Multi %
10	5.4	13.9	561	15.2	13.2	460	17.7	13.6
20	12.5	13.1	564	11.9	20.6	480	14.3	16.2
30	17.1	10.6	600	22.8	16.2	510	17.4	15.6
40	7.9	10.3	590	9.9	18.4	514	10.6	13.9
51	10.0	13.2	592	7.2	13.6	563	8.4	24.6
500	6.9	6.9	240	6.6	13.5	740	9.1	4.8
620	11.2	18.3	260	5.5	12.3	750	7.0	1.5
562	12.2	13.9	270	18.1	14.1	760	3.3	17.7
190	10.1	8.0	280	3.2	2.4	770	9.9	4.7
680	11.5	7.8	300	3.3	5.4	785	5.8	5.9
910	12.1	18.0	310	4.3	8.5	790	7.9	5.8
920	11.6	11.2	330	7.6	14.1	670	6.6	9.
61	8.8	12.9	340	9.8	14.9	900	17.2	15.2
71	13.3	16.9	350	4.9	5.0	980	9.6	14.0
72	8.8	11.8	380	8.0	5.3			
100	2.7	13.4	410	4.9	13.4			

As already mentioned, there is some apparent variability and in some cases a multiple membership gene may show less frequent expression than a single-membership one. This is not surprising given the nature of the data. In fact there are a number of reasons, discussed in section 3.1.2, that could explain why genes may show variable expression, including different levels of regulation, gene diversity and microarray limitations themselves. Additionally, one should also take into account that the analysis is based on randomly selected experiments and that some genes even though single-membership may be in pathways that are frequently regulated, while other multi-membership genes may be taking part in pathways less frequently regulated in the data in hand.

Importantly, the varying experimental approaches and research goals behind individual microarray experiments in the analysed dataset are likely to introduce some bias as far as the regulation of distinct biochemical pathways is concerned. For example, the glycolysis and gluconeogenesis pathway constitutes a popular choice of analysis. Consequentially, we would expect to observe relatively high differential expression of genes in these paths.

Indeed, ranking the KEGG pathways according to the average proportion of expressed genes per experiment reveals that the KEGG Glycolysis/Gluconeogenesis pathway is the 21st most highly ranked out of 69 *Saccharomyces cerevisiae* metabolic pathways. Figure 3.17 exhibits the *Saccharomyces cerevisiae* KEGG metabolic pathways, in descending order, from the one with the highest average proportion of expressed genes per pathway to the one with the lowest average proportion of expressed genes per pathway, in the analysed dataset. The calculation is based on estimating the proportion of expressed genes forming a pathway per experiment, and averaging over all experiments.

Alternatively, given that a threshold of even 20% of differentially expressed genes is reasonable evidence to consider a pathway affected by the experimental conditions (Subramanian et al. 2005), to apply a more stringent constrain we look at the average instances in our dataset where at least 30% of the genes in each path appear regulated (Figure 3.18).

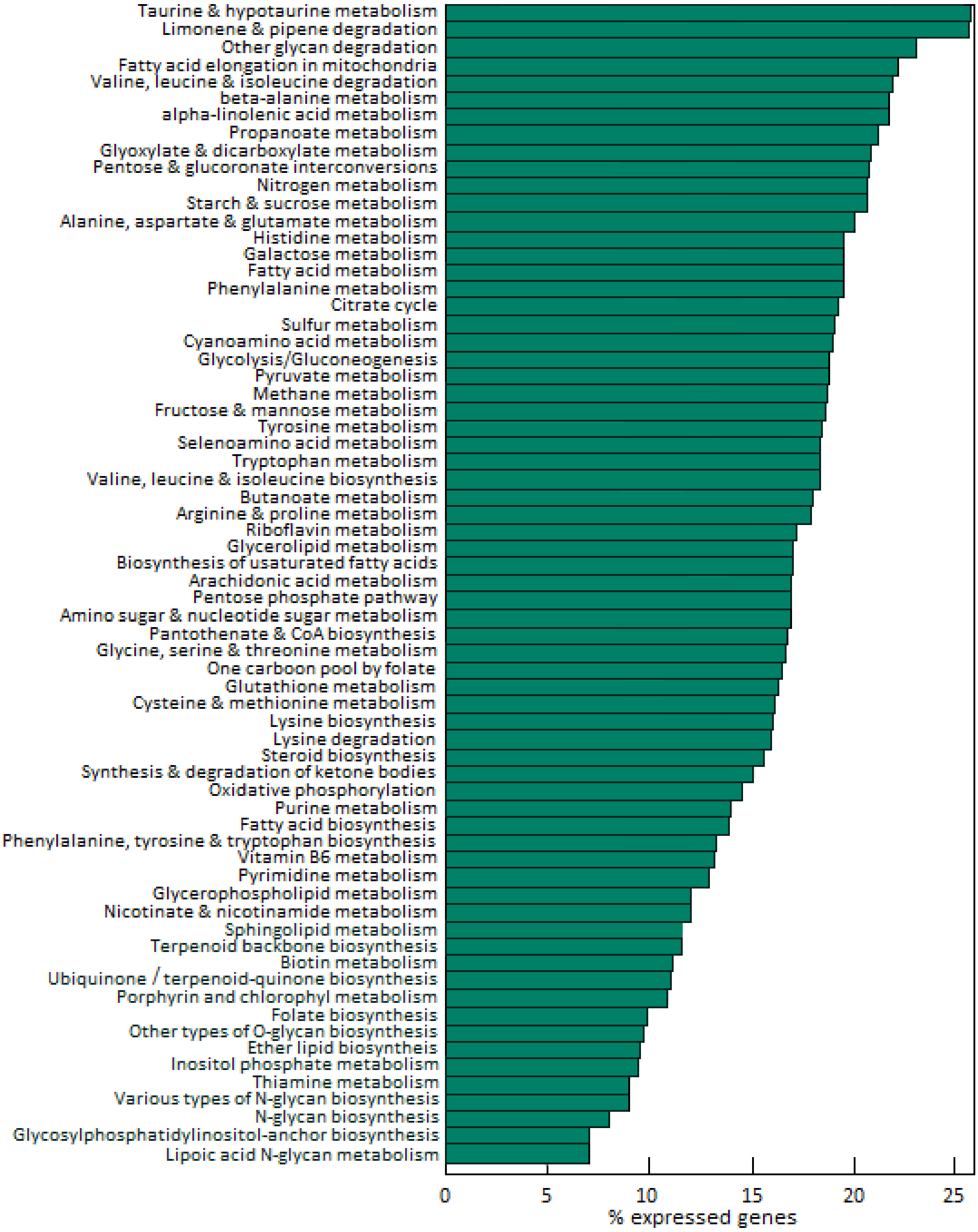


Figure 3.17 Average expression of *Saccharomyces cerevisiae* KEGG genes per pathway. Pathways are ranked from the one with highest proportion of expressed genes (top) to the one with lowest (bottom).

Again, as Figure 3.18 shows, pathways exhibit substantial diversity, with some cases like the fatty acid elongation in mitochondria appearing with more than 30% of genes expressed in only 0.002% of the experiments, while others like the Taurine and

Hypotaurine metabolism pathway showing above 30% expressed genes in 42.76% of the experiments.

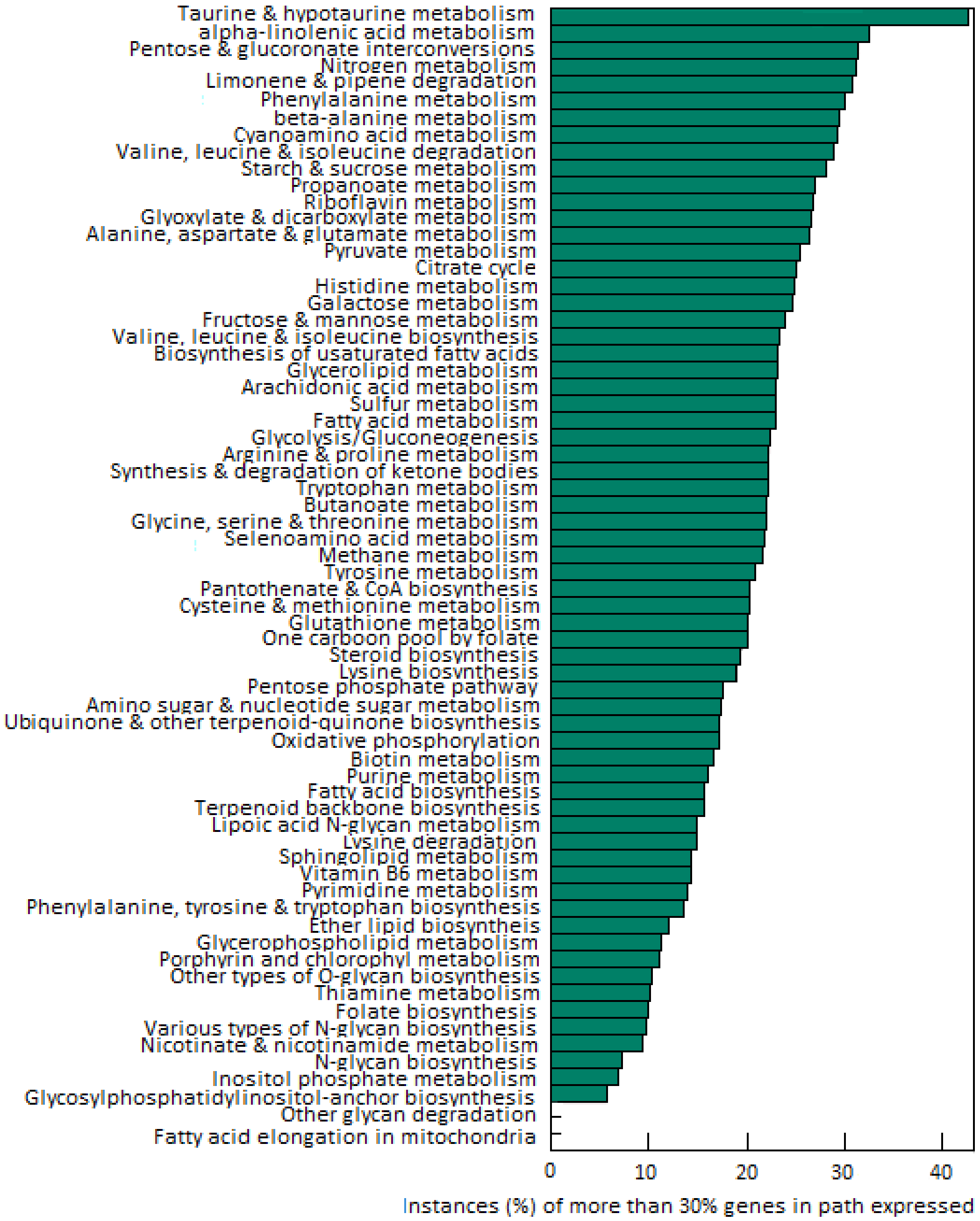


Figure 3.18 Instances of over 30% expressed *Saccharomyces cerevisiae* genes per pathway. Pathways are displayed in a sorted arrangement, from the one with most cases of 30% or more of its genes expressed (top) to the one with least such cases.

The KEGG Glycolysis/Gluconeogenesis pathway remains quite highly ranked, at position 26 out of 69 metabolic pathways. Moreover, if we consider the fact that the first highest ranking few pathways consist of only one or two genes which can explain why they frequently surpass the threshold of 1 expressed gene, the Glycolysis/ Gluconeogenesis pathway ranks even higher.

Overall, the result indicates that in spite of the above discussed issues, on average, multi-membership genes show more frequent differential expression than genes that are members of one and only pathway.

3.5 Contradicting expression values

According to our rationale, observing contradictions in the state of expression of multi-membership genes is largely due to the fact that the biological system regulates their expression to cover the needs of all their constituent pathways. Given that for genes working in more than one pathway, there will be instances where some of these paths may be activated while others supressed, we expect to observe the occurrence of contradicting gene expression at higher rate for multi-membership genes forming a pathway than in the case of their single-membership counterparts.

Indeed, we observed higher rate of contradicting gene behaviour in the groups of multi-membership genes than the ones of single-membership genes in each pathway, in the analysed *Saccharomyces cerevisiae* dataset. On average, for all pathways and experiments, single-membership genes show 7.7% rate of variable expression in the same pathway, in contrast to 13.9% for multi-membership genes. Figure 3.19 summarises the result per pathway, only for pathways that contain genes of both single- and multi-membership nature, allowing a comparison.

With only a few exceptions, cases of at least one gene contradicting the expression of the rest in the pathway are more frequent for genes that participate in more than one pathway. As Figure 3.19 reveals the difference in the number of such occurrences is often quite evident.

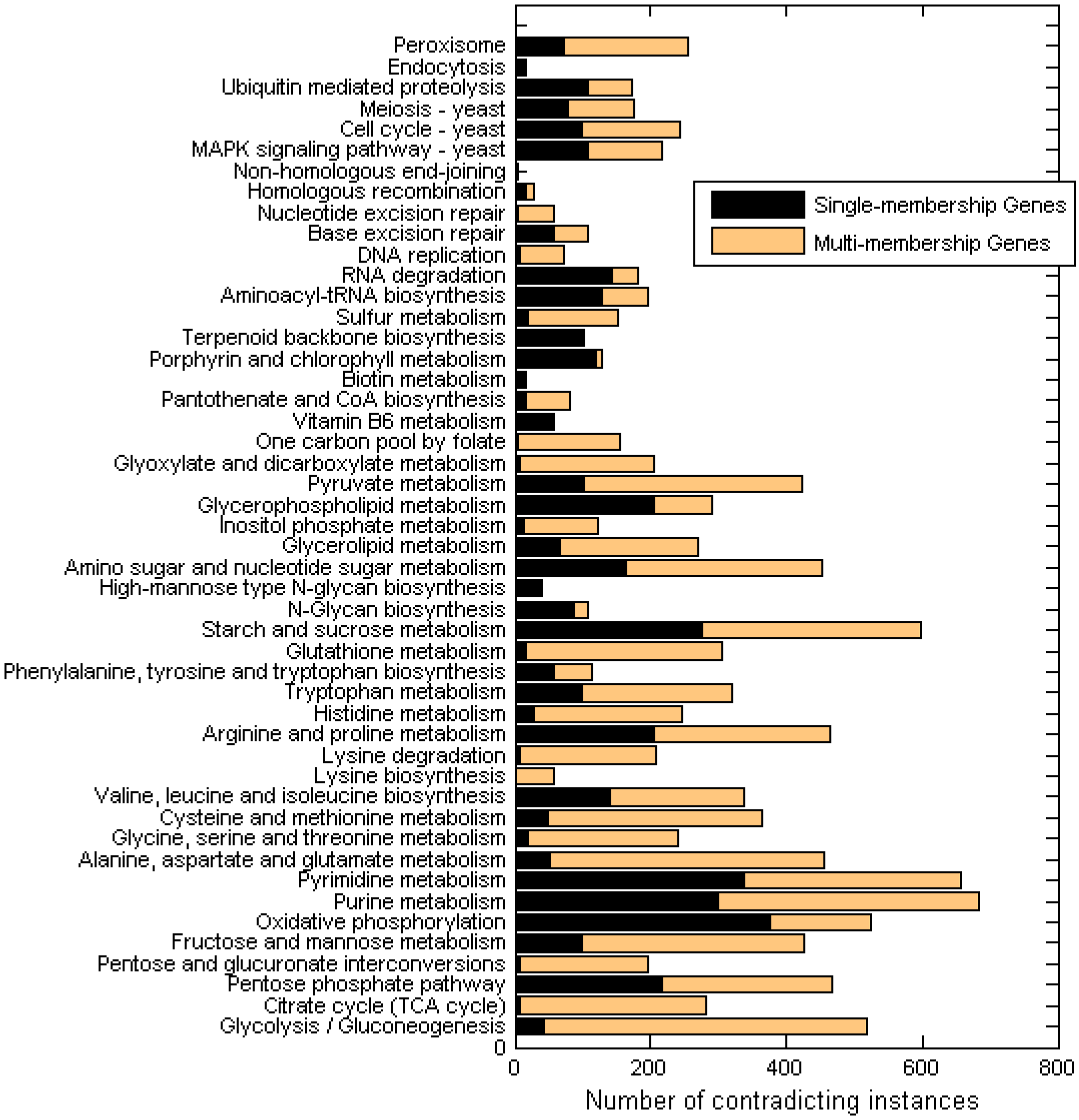


Figure 3.19 Contradicting *Saccharomyces cerevisiae* gene expressions per pathway. These are the instances where a gene in a pathway contradicts the behaviour of the rest of the genes forming the pathway in question.

3.6 Statistical analysis

Given the presence of multi-membership genes and the hypothesis established in the previous section, correlation analysis seems an appealing approach to further study the behaviour of such genes. Since each gene is likely to have a particular contribution to the activity of a pathway, that may be relatively steady, we can examine that contribution as a percentage of the total activity of the pathway.

We estimate the correlation between the expressions in terms of percentage for all single-membership genes in a pathway and for all multi-membership genes in the same pathway (Pavlidis, Payne & Swift 2008). Naturally, observing higher correlation for single-membership genes as they only contribute to the activity of the pathway in question and thus exhibit more consistent expression may provide evidence supporting our rationale. As previously discussed, unlike single-membership genes, multi-membership genes can participate in the functionality of any combination of the pathways they are members of, at any particular time. Thus, unlike single-membership genes, multi-membership genes' intensity values, as extracted from a microarray slide, represent a net effect. The biological system may require activation of certain pathways and regulate the production of a protein part of their network in a way that its quantity increases. At the same time it may require deactivation of other pathways in which the same protein participates. The resulting balance may affect the expression observed on the microarray leading to less consistent readings for groups of proteins part of a biochemical pathway, encoded by multi-membership genes, when each pathway is examined in isolation from the rest.

For example in a pathway consisting of genes A, B and C that say contribute 20%, 50% and 30% to the overall pathway activity, we can add the $\log_2$ ratios for genes A, B and C to get an estimate of that total activity of that path, and then examine the percentage of contribution for each gene in various microarray experiments. Ideally, we would want to obtain values close to the percentages above in each experiment where the pathway is activated. The obtained values should be more consistent in the case of single-membership genes than in the case of multi-membership genes.

To examine this we initially identified 19 experiments (GSM99081 to 83, GSM99108 to 112, and GSM99171 and GSM99172) on *Escherichia coli*, from microarray data available at Gene Expression omnibus (GEO), platform GPL3503 that contain a large number of expressed Urea Cycle genes (01/09/2008). The KEGG Urea Cycle pathway is a good candidate for our analytical approach as it consists of 16 single-membership and 12 multi-membership genes, reasonable numbers to allow meaningful comparison. We divide the intensities, separately for the group of single- and the group of multi-membership genes, per experiment by their sum, to obtain a measure of the contribution of each gene to the behaviour of the pathway. We then compare the correlation between the obtained contribution values of the 12 multi-membership genes and the 16 single-membership genes, throughout the 19 experiments. For both cases we acquire a set of 171 correlation values, and perform a two sample t-test which reveals that the values are significantly different with a *p*-value of 1.3×10^{-12}. Furthermore, in the case of single-membership genes the correlation values are higher with 86.5% of the values being above the level of significant correlation at *p*=1%. In contrast, for the multi-membership genes only 41.5% of the values exceed the threshold of significance at 1%. The assumption that multi-membership genes expression is the net effect of their contribution to their constituent pathways is in agreement with these findings. Single-membership genes apparently show more consistent behaviour as they only contribute to the functionality of the KEGG Urea Cycle pathway.

As KEGG is constantly updated it currently holds the Urea Cycle path in a larger pathway termed Arginine and Proline metabolism (01/02/2011) that contains more genes subsequently identified and added to the updated KEGG. The pathway consists of a total of 43 genes, 21 of which are unique members of the pathway in question, while 22 are multi-membership genes. We performed an analysis of the correlation of expression values for these new subsets of genes based on GEO platform GPL3503, consisting of 140 experiments on *Escherichia coli*.

We observed that the correlations of expression between each couple of the 140 experiments were higher in the case of single-membership genes than in the case of multi-membership genes. A two sample t-test revealed that the correlation values for these two subsets of genes in the KEGG Arginine and Proline metabolism pathway where significantly different, with a *p*-value of 2.4×10^{-12}.

Figure 3.20 graphically represents the correlation values for single- and multi-membership genes, for all the combinations of the 140 experiments by two. Interestingly, correlations for multi-membership genes have a tendency for significant negative values, which is in agreement with the observation that they show contradictive behaviour. We may ascribe this behaviour to their contribution to pathways that may have opposing activity under certain experimental conditions. In particular, the percentage of correlations for single-membership genes below -0.5 is only about 6% as opposed to close to 13% for multi-membership genes.

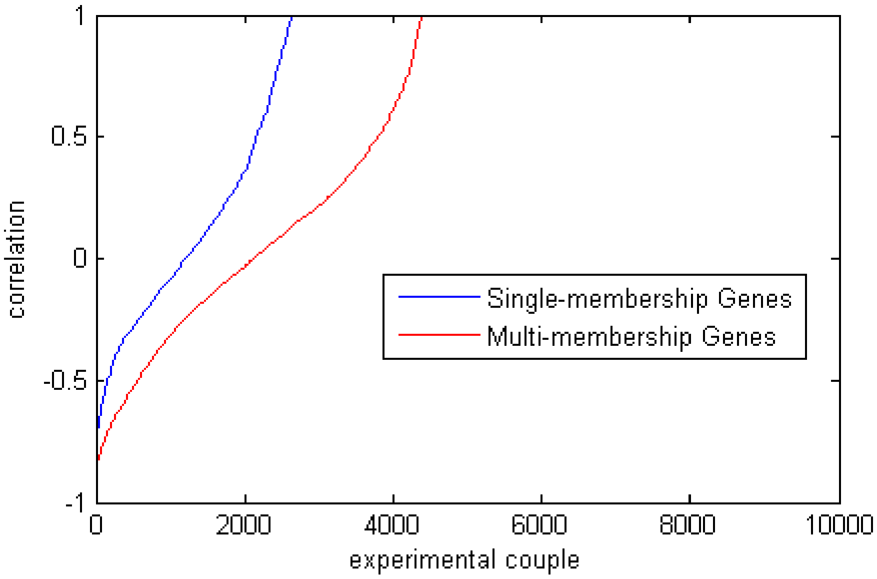


Figure 3.20 Correlation between expression values. The values correspond to *Escherichia coli* gene couples in the KEGG Arginine and Proline metabolism pathway in a subset of 140 experiments (GPL3503 from GEO).

We performed a similar analysis of the Oxidative phosphorylation KEGG pathway, based on experiments where at least 50% of the genes in the path show differential expression for a threshold of one standard deviation of intensity value. This criterion is satisfied by 28 experiments, allowing for 378 comparisons, where we observed a mean correlation of 0.30 for single-membership genes, as opposed to only 0.07 for multi-membership genes, thus more than 4 times lower value. A two sample t-test showed that the correlation values for all experiments were significantly different with a *p*-value of 1.3×10^{-5}. In about 12% of the cases for the first group of genes the correlation was significant at *p*-value of 0.01, while this was true for only about 4% of the latter group.

We performed the same comparative analysis for all 140 experiments without applying a threshold of expressed genes in the pathway. Naturally, we obtained lower correlation values, but the pattern was the same with single-membership genes showing an average correlation of 0.050 as opposed to only 0.001 for multi-membership genes. In this case a two sample t-test revealed that the 9730 correlation values for each group are significantly different with *p*-value of 1.6×10^{-8}.

3.7 Association rule mining

Association rule mining (ARM) is a very popular data mining methodology that was first proposed in the 90's for determining consumer purchasing patterns based on databases of consumer transactions (Agrawal 1993). These prove very useful to help managers identify items that are likely to be bought at the same store visit. In essence, an association rule reveals the probability that a customer that bought items X and Y will also buy item Z. Since its introduction it has been applied to discover useful information in many areas, including gene expression data.

In brief, an association rule is an expression of the form $\mathrm{LHS}\Rightarrow\mathrm{RHS}$, where LHS stands for left hand set and RHS for right handset of items (Hipp, Guntzer & Gholamreza 2000). The two sets are disjoint and the expression implies that given the occurrence of LHS, RHS is also likely to occur. Each association rule is characterised by two statistical measures, the support and confidence. For example for the rule $\mathrm{XY}\Rightarrow\mathrm{Z}$, support 30% and confidence 80% implies that whenever X and Y occur, Z also occurs in 80% of the cases, while all three occur in 30% of all cases. In this example we can infer that there is indeed some significant association between the co-occurrence of X and Y with Z.

3.7.1 Association rule mining in gene expression data

Given the popularity of association rule mining it was soon realised that the methodology can be facilitated for the analysis of gene expression data in various contexts. To mention a few examples, ARM can be used to determine how the expression of a certain gene may affect the expression of other genes. The underlying rationale is that identifying a gene whose expression determines the expression of other genes, with high probability, implies that the gene is likely to belong to the same functional group (Creighton & Hanash 2003). Additionally, ARM can be facilitated to identify genes that are expressed as a result of a particular cellular condition, for example genes that are expressed in certain disease condition, but remain silent in a healthy cell.

Due to the widespread application of ARM, currently there are a number of proposed variants of association rule mining approaches that can be roughly classified into categories according to the type of data they can handle. In brief the most popular types include ARM for the analysis of Boolean, nominal and quantitative data.

Given that we use microarray data, where a chosen threshold of standard deviation is used to discretise the state of a gene to stable, up- or down-regulated, we have facilitated an approach to extract association rules from ordinal data, described in (Chen & Weng 2008). Unlike nominal variables such as colours for example, where there are a number of unique possible states to which a variable can belong, in ordinal data the states are ordered. A multiple choice questionnaire where the possible answers include “good”, “very good”, “medium” and “bad” is a characteristic example. This is the so called Likert scale, initially introduced by Renis Likert in the field of psychology (Likert 1932). Clearly, “very good” is a better match to “good” than to “medium” while it is the worst match to “bad”. Similarly, in discretised gene expression data, up-regulated may not match stable, but is still a much worse match to down-regulated.

3.7.2 Definitions

Let $G = \{g_1, g_2, \ldots, g_n\}$ be a set of all unique genes in our dataset, and a value q that is a single ordinal value that can be equal to 1, -1 and 0, if the gene is up-, down-regulated or stable, respectively. We work with a similarity matrix as exhibited on table 3.3.

Assume that we have a single gene $a_i = (g_i, q_i)$ and a single gene $b_j = (h_j, r_j)$ and a similarity matrix $Sim_{i,j,}$ as the one on Table 3.3, to represent the similarity between q_i and r_i, where $sim(i,j)$ denotes the similarity between i and j. The degree to which a_i matches b_i is defined by equation (3.1).

$$\sup(a_i, b_j) = sim(a_i, b_j) \qquad \text{for } g_j = h_j \qquad (3.1)$$

For example if we have the expression of gene a_1 = (g_1,1) and of gene b_1=(h_j,1), the degree of $\sup(a_1, b_1) = \sup((g_1,1),(h_1,1)) = 1$, if g_1=h_1.

Table 3.3 Expression similarity matrix.

expression	-1	0	1
-1	1	0.5	0
0		1	0.5
1			1

Assume that we have a single gene expression set A = {(g_1,q_1),(g_2,q_2),…,(g_n,q_m)} and a gene expression set B = {(h_1,r_1),(h_2,r_2),…,(h_n,r_m)}, where we can find i_1, i_2,…, such that a_{ij} matches b_j for $1 \le j \le$ n. Let sup(A,B) denote the degree to which A matches B be defined as shown in equation (3.2).

$$\sup(A,B) = \prod_{j=1}^{n} \sup(a_{i_j}, b_j) \qquad (3.2)$$

For example, assume that A = {(g_1,1), (g_2,1), (g_3,0)} and B = {(h_1,1), (h_2,0), (h_3,1)} and the similarity matrix on Table 3.3. Then assume that g_1=h_1, g_2=h_2 and g_3=h_3, which simply means that gene g_1 is the same as gene h_1, with an expression value of 1 in both cases, gene g_2 the same as gene h_2 with expression value of 1 and 0 respectively and so on. Then the degree of sup(A,B) = $1 \times 0.5 \times 0.5$=0.25.

Assume that we have a large dataset D consisting of a number of microarray experimental results in the form of N rows by M columns matrix, where each row represents a gene and each column represents an experiment. Let A_i be the ith column in D, where A_{i} = {(g_1,q_1),(g_2,q_2),…,(g_n,q_n)}. Then if we have a single dataset of gene expression data B = {(h_1,r_1),(h_2,r_2),…,(h_m,r_m)}, where m$\le$n and b_j=(h_j,r_j), the support for B in D is defined as shown on equation (3.3).

$$\sup_D(B) = \frac{\sum_{sid=1}^{|D|} \sup(A_{sid}, B)}{|D|} \qquad (3.3)$$

where $|D|$ denotes the experiment number (column) in the dataset.

For example assume the hypothetical matrix on Table 3.4 representing gene expression data for 3 genes based on 4 experiments.

Table 3.4 Hypothetical example of expressional behaviour of 3 genes in 4 experiments. Ones represent up-regulation, minus ones down-regulation and zeros lack of differential expression.

Gene\experiment	1	2	3	4
1	0	0	1	0
2	1	0	-1	1
3	1	1	-1	1

Given this matrix and a set *B* = {(*g*1,0),(*g*2,1),(*g*3,1)}, the degree of support for *B*, based on *D*, is calculated as follows:

$$\sup_D(B) = (1\times1\times1 + 1\times0.5\times1 + 0.5\times0\times0 + 1\times1\times1)/4 = 2.5/4 = 0.625$$

Then we can calculate the confidence of a rule XY⇒Z based on the support of the item sets on the left and right hand side as follows:

$$conf\,(XY \Rightarrow Z) = \sup(XYZ)/\sup(XY) \qquad (3.4)$$

For example, for the matrix above, the confidence of a rule {(*g*1,0),(*g*2,1)} ⇒ {(*g*3,1)} is calculated as follows:

$$conf\,(\{(g1,0),(g2,1)\} \Rightarrow \{(g3,1)\}) = \frac{\sup(\{(g1,0),(g2,1),(g3,1)\})}{\sup(\{(g1,0),(g2,1)\})} =$$

$$= \frac{(1\times1\times1 + 1\times0.5\times1 + 0.5\times0\times0 + 1\times1\times1)/4}{(1\times1 + 1\times0.5 + 0.5\times0 + 1\times1)/4} = 0.625/0.625 = 1$$

Thus, in this example, when gene 1 is stable and gene 2 up-regulated, we expect gene 3 to be up-regulated with a confidence equal to 1.

3.7.3 Application

We apply association rule mining to sets of genes forming biochemical pathways. In particular we examine sets of single- and multi-membership genes, members of the same pathway, with the aim of observing differences in their behaviour, in an effort to validate our hypothesis.

In the case of using ARM to analyse microarray data we would prefer not to exclude rare occurrences of cases of gene expression values. For example, genes X and Y may occur up-regulated in only 1% of the instances but gene Y may only occur up-regulated when gene X is up-regulated. In this example, the rule stating that when X is up-regulated Y is also up-regulated would be characterised by very low support but at the same time confidence of 1. As we are interested in such observations, we do not follow the approach described in (Chen & Weng 2008), where the authors start with single item sets, apply a support threshold, and then use the item sets with sufficient support to add an additional item to build larger item sets and continue the same cycle until reaching a desired size for these item sets. Rather we exhaustively search and calculate the confidence for all possible item sets, regardless of their corresponding support values, that is, all existent combinations of expression values present on our dataset for genes of interest. However, we do not consider combinations of expression that never occur, that is, if gene X and Y are never up-regulated together we do not attempt to extract the confidence of the rule when X is up-regulated Y is also up regulated and vice versa.

Naturally, exhaustive search is computationally intensive and not applicable to large gene sets due to time constraints. It is however applicable to smaller sets of genes forming pathways, which were examined here.

First, we applied our association rule mining approach to obtain the rules produced by single- and multi-membership *Escherichia coli* genes, separately, in the Phenylalanine metabolism pathway based on GEO platform GPL3503. Naturally, not applying a threshold for the support of each rule, we end up with a very large number of such rules. It is interesting to observe this number for the group of single- and multi-membership genes, at different confidence threshold levels, as shown on Figure 3.21.

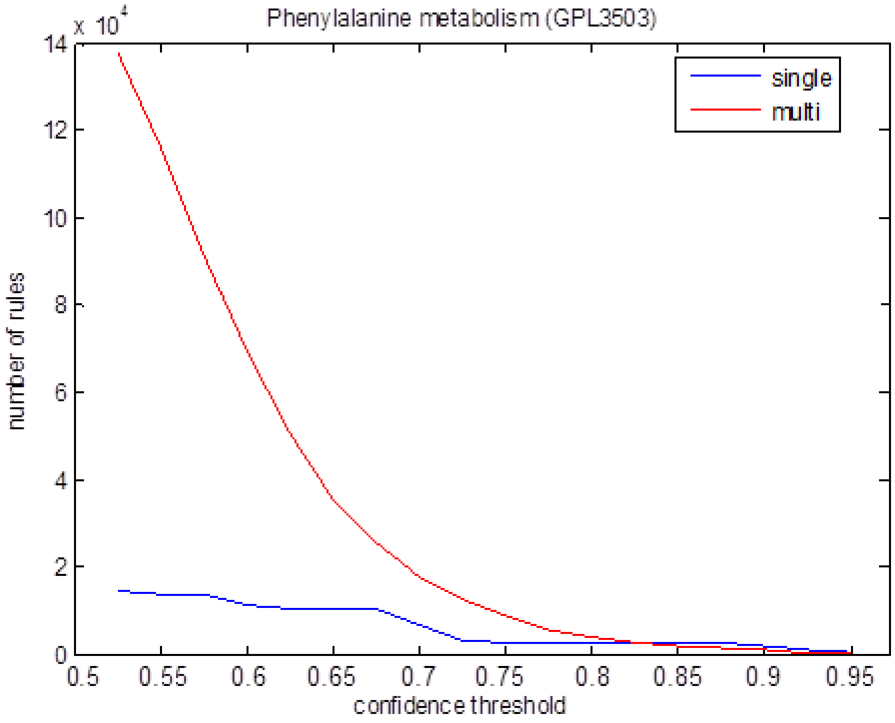


Figure 3.21 Number of association rules produced by single- and multi-membership *Escherichia coli* genes in the Phenylalanine metabolism pathway (GPL3503).

Given a confidence threshold between 0.50 and 0.95 (by 0.05), Figure 3.21 shows the number of rules where one gene implies all the rest. The much larger number of rules for multi-membership genes at low threshold levels is not surprising. We have shown that these genes are expressed more often on average as they are more likely to be functional due to their membership in many pathways. In other words, the biological system is more likely to facilitate a multi-membership gene at a given any time, as such genes have wider functionality. At high confidence threshold values the opposite is true as single-membership genes are more consistent in terms of expression and thus more likely to produce rules of high confidence.

If we examine all rules where one gene implies another one or more genes, for any given confidence above zero, we get 148243 rules for multi-membership genes with a mean of 0.27, and only 14962 for single-membership genes with a mean confidence of 0.46.

Figure 3.22 represents the mean confidence of rules produced at different confidence thresholds, revealing the higher values produced by single-membership genes.

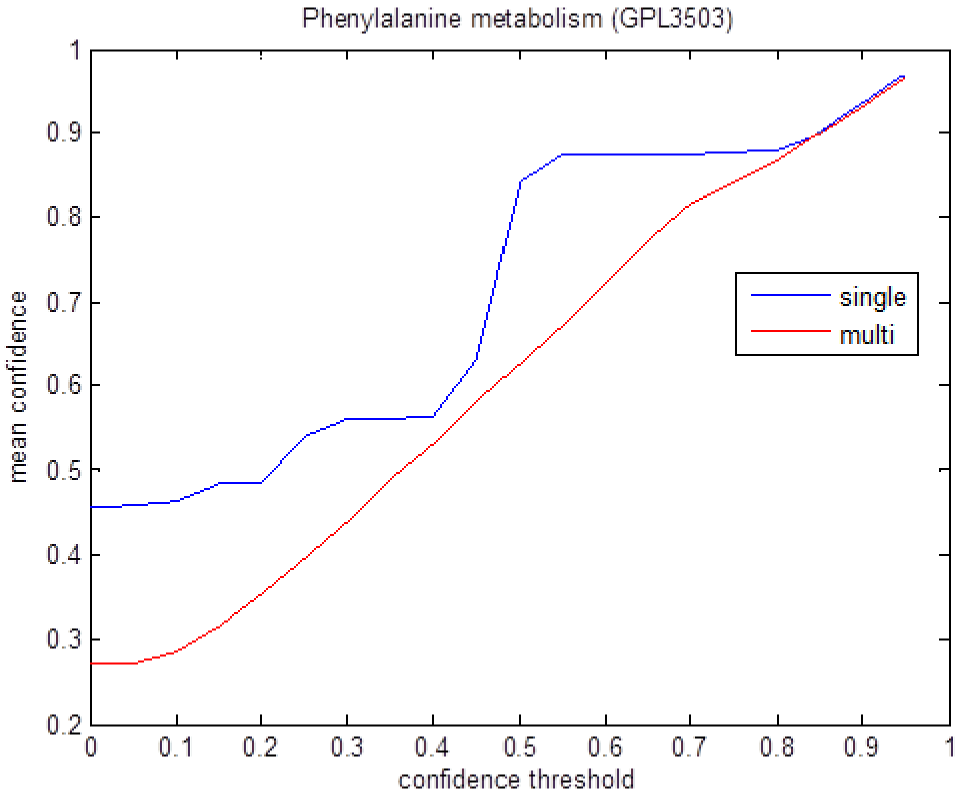


Figure 3.22 Mean *Escherichia coli* genes association rule confidence for rules of confidence above different thresholds (GPL3503).

One however might argue that the two extra multi-membership genes in the pnenylalanine pathway justify the hughely larger number of association rules produced by this group of genes, thus we performed the same analysis for the Glutathione metabolism pathway. This pathway is an ideal candidate for our analytical approach as it consists of exactly the same number of single- and multi-membership genes, nine in each case. This is also a number that allow us to exaustivily search for association rules correspondng to all possible combinations of genes and expression states in the dataset.

Our analysis produced 157896 in the case of single-membership genes and 562204, 350% more rules for multi-membership genes. In contrast, the mean confidence is higher in the first case, equal to 0.49 with standard deviation of 0.27, as opposed to mean of 0.38 and standard deviation of 0.22 in the latter case.

As Figure 3.23 reveals, increasing the threshold of confidence, the number of rules produced by multi-membership genes drops at a high rate, until for confidence above 0.8 it becomes single-membership genes that produce more rules. Additionally, for confidence of 1 the latter genes produce 256 rules as opposed to 0 for the former group.

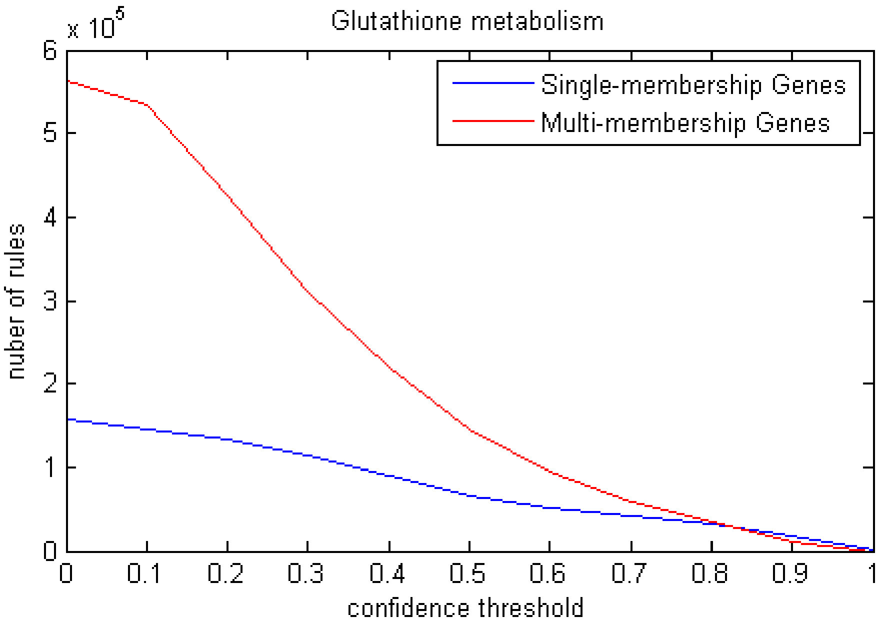


Figure 3.23 Number of association rules where one gene implies the behavior of the rest in the Glutathione metabolism pathway.

Multi-membership genes in the Glutathione metabolism KEGG pathway produce large number of rules of low confidence values, while at confidence threshold above about 0.8, single-membership genes produce more rules. As far as the mean confidence is concerned, single-membership genes produce rules of higher mean confidence at all thresholds, as exhibited on Figure 3.24.

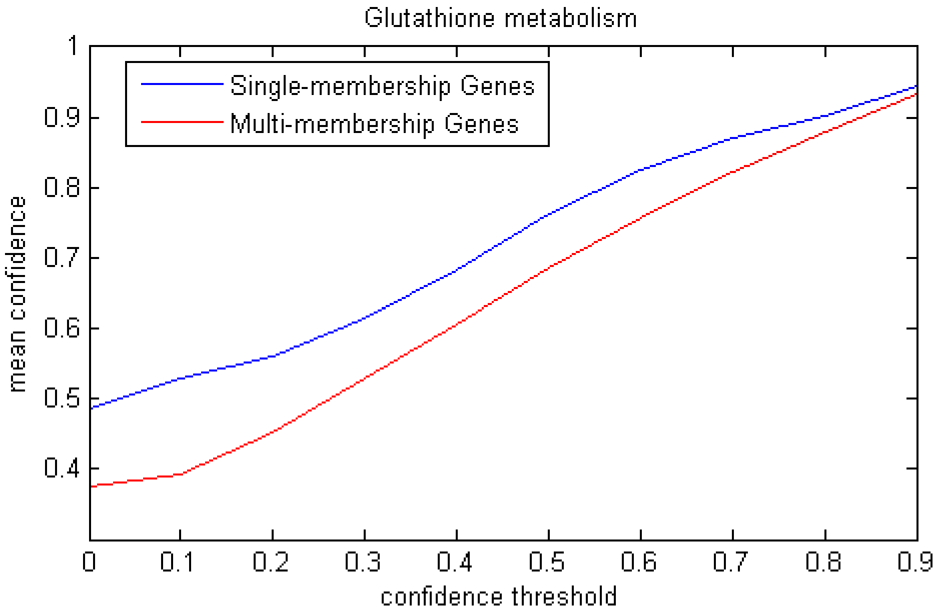


Figure 3.24 Mean confidence of rules produced by single- and multi-membership genes in the Glutathione metabolism pathway, for increasing minimum threshold.

To examine another case of a pathway with very small number of genes, we look into the D-alanine metabolism pathway. This path consists of only four genes according to KEGG, two of which are single- and two multi-membership genes. The data produces 12 and 10 association rules, with mean confidence of 0.58 and 0.54 respectively. Even though the number of genes is too small to allow as to draw clear conclusions it is interesting to observe that for the same very small number of only 2 genes, single-membership genes produce more rules, but still of higher average confidence.

3.8 Conclusions

Overall, four distinct approaches were facilitated to look into the relative behaviour of single- and multi-membership genes in terms of differential expression. The analysis is based on large datasets of randomly selected microarray experiments compiled from GEO.

Examining the frequency of differential expression we identify a clear tendency for genes that KEGG places in two or more pathways to be more frequently expressed on average than genes that constitute members of one and only pathway. This tendency shows a strong increase as we consider genes with higher degree of membership. Namely genes that are members of three or more pathways show higher frequency of activation and deactivation than genes that are members of one and genes that are members of two pathways. Moreover, we observe a positive correlation between the degree of membership and the average expression of genes belonging to each membership band.

This is a sensible result, since multi-membership genes are multitask genes, whose protein products are involved in more than one cellular process. As the biological system regulates the function of its genes, and subsequently its protein arsenal, to adapt to changing environmental conditions, it is more likely to require the contribution of multi-task genes than single-membership genes, at any given instance.

We also observed that multi-membership genes in a pathway appear to contradict each other’s behaviour more often than unique members of one and only pathway and a tendency for the latter group of genes to show higher correlation in terms of pathway contribution and expression values, than their multi-membership counterparts. We expect such behaviour as multi-membership genes are facilitated by the biological system in a number of different pathways, sometimes in opposing state of activation. Here the differences are less apparent but as discussed in the relevant section, we cannot expect perfect results for a number of reasons. In brief, microarrays can be noisy, gene activity is regulated at many levels, besides transcription, and different pathways are facilitated at varying degrees by the biological system. Importantly, the choice of experimental question by the various researchers supplying GEO with microarray data is likely to introduce some bias to pathway activation. For example, glycolysis and gluconeogenesis constitute a very popular choice of study, as these are well known pathways present in almost every organism. Thus it is not surprising to observe that genes in these paths are some of the most frequently differentially expressed genes in the dataset in hand.

Additionally, we observe that single–membership genes produce more consistent association rules, characterised by higher confidence values. Multi-membership genes on the other hand tend to produce more rules of lower confidence. In conclusion our analysis provides evidence which strengthens the hypothesis that the expression of genes that constitute members of various pathways represents a net effect and is regulated by the biological system in a way that meets the needs of all their constituent pathways. Hence, it would be greatly beneficial to develop a methodology to identify the pathways to whose activity multi-membership genes truly contribute which is the central aim of this work.

Chapter 4: Pathway based microarray analysis methodology, facilitating hill climbing search

4.1 Introduction

We have established that Pathway based microarray analysis constitutes an area of substantial interest in the effort to gain deeper insight into the behaviour of biochemical pathways and evaluate the state of an organism from a biochemical perspective based on gene expression data.

In chapter 2 we provided an overview of this analytical and visualisation approach and a list of currently available software tools for this type of analysis. Following that, in chapter 3, we discussed our basic hypothesis, that the expression of genes, which are members of more than one pathway, represents a net effect of their contribution to all their constituent pathways, as the biological system regulates their activity to fulfil its needs in a balanced manner.

Here we present our initial approach to analyse microarray data in terms of pathways, implementing a hill climbing algorithm and facilitating the idea of gene to pathway allocation, in an effort to add some contribution to the general methodology and assists a biologist to draw meaningful conclusions from available data. The motivation directing our approach is briefly exemplified followed by presentation of the rationale we follow in this work with some adequate examples. Then we proceed to discuss the methodology and algorithms in some detail.

The results section contains a detailed analysis of the results produced by the methodology on popular microarray datasets, in relation to the publications accompanying the data. We comment on the convergence of the algorithm and the consistency of the produced results, in separate runs of the scripts, along with a discussion of the effect of the processing on the probabilities of observing the expression of genes per pathway in hand. Finally, we discuss our conclusions regarding the results and the potential of the methodology.

4.2 Motivation

It has been shown that gene members of the same biochemical pathway do not always show consistent behaviour in terms of RNA production. While ideally we would expect to observe expression that reflects the state of a pathway, meaning up-regulation when the pathway is activated and down-regulation when the pathway is supressed, this is often not the case. In a number of microarray experiments there is an evident contradiction in the state of differential expression of genes belonging to the same biochemical pathway. More precisely, in these experiments we observe substantial number of up-regulated and down-regulated genes, in the same pathway. In a dataset consisting of 2135 randomly selected microarray experiments on *Saccharomyces cerevisiae*, on average 25 pathways exhibit at least one gene of expression contradicting the rest in each experiment. In some cases up to 81 pathways exhibit contradictions in the expression of their genes, a number almost equal to the totality of pathways in this organism.

To study this in more detail, we identified instances where a substantial number of the genes forming a biochemical pathway show differential expression. In particular we chose a 30% threshold, which as aforementioned is widely considered a valid indication of the pathway in hand being affected by the experimental conditions.

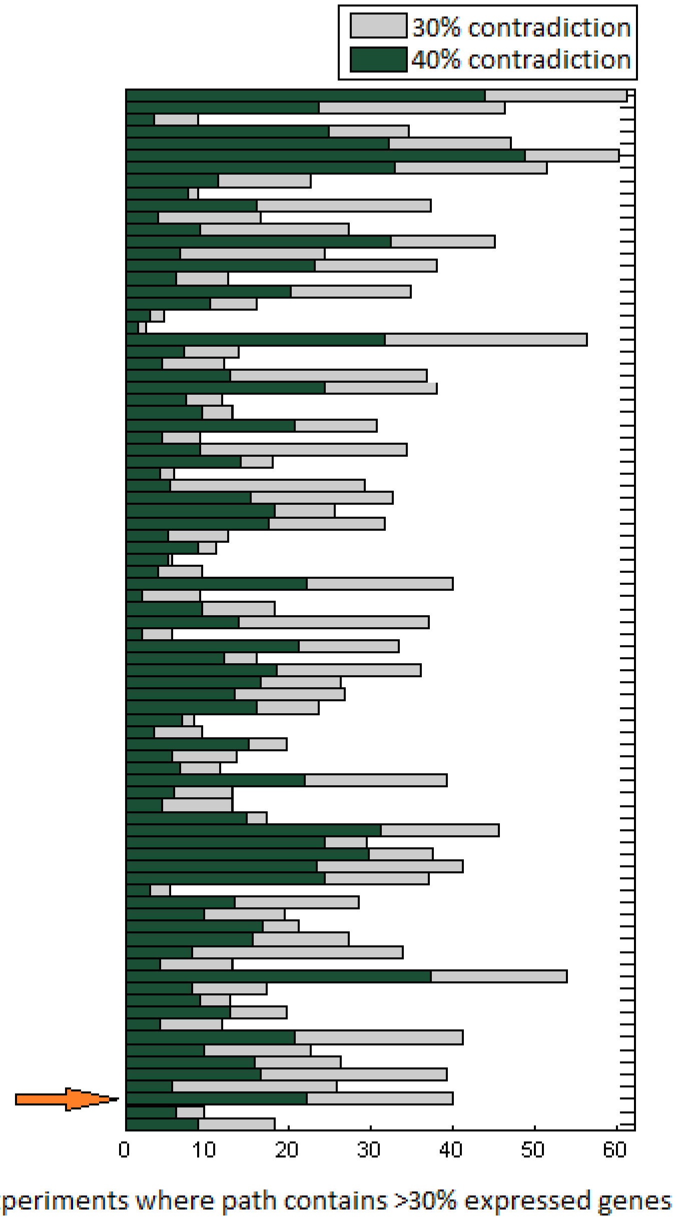


Figure 4.1 Proportion of genes of contradicting behaviour per pathway, in the subset of instances, where at least 30% of all genes in a path exhibit differential expression. The arrow indicates the pentose phosphate pathway, commented in the text.

Following this, we examined the instances, within this subset of experiments, where there is a considerable level of contradiction in the expression of genes, excluding 6 KEGG pathways that consist of 1 to 3 genes only. In particular we identified the proportion of cases within the subset, where at least 30% and 40% of the genes in the pathway exhibit differential expression in opposing direction to the rest of the expressed genes in the pathway.

As Figure 4.1 reveals this holds true in a considerable proportion of the microarray experiments where each pathway exhibits a large number of expressed genes, in the dataset in hand. For example, regarding the pentose phosphate pathway, indicated with the arrow on Figure 4.1, in more than 40% of the cases where significant number of genes appear differentially expressed, at least 30% of these genes show contradicting behaviour. Furthermore, in about 20% of these cases this is true for even higher proportion of genes, with at least 40% of them being in disagreement.

Evidently, in a number of instances, which represent a considerable proportion of microarray experiments in the dataset, the direction of differential expression of the majority of expressed genes does not allow as to draw a safe conclusion regarding the state of activation of a biochemical pathway. That is, both stimulated and suppressed genes are present in the same pathway in substantial and in some cases even the same numbers.

These are the cases that we are interested in, since elucidating the state of activation of these pathways is not trivial. Such microarray datasets constitute ideal candidates for our analytical approach that aims towards identifying the true state of activation of the totality of biochemical paths in a given experiment.

4.3 Rationale

We assume that given a certain pathway and a certain state, e.g. activation, proteins forming the pathway follow the trend of increased activation, which should be reflected on the expression of genes encoding them. That is, the respective genes produce more RNA for the synthesis of the protein they encode, which in turn contributes to the pathway function. Thus, we attempt to ascribe any observed down-regulation of genes in this pathway to decreased activity of other pathways of which these genes are also members, which in turn require less of the protein product of the genes in question. We assume that the net effect of the contribution of these genes to pathways of contradicting behaviour may be responsible for the contradicting intensity values extracted from the microarray chip.

In order to exemplify our line of thought we facilitated *Escherichia coli* data from (Khodursky et al. 2000) available as experiment GSM513 at Gene Expression Omnibus. The experiment examines the cell response in terms of global gene expression to addition of excess tryptophan in the growth medium.

Table 4.1 Log$_2$ ratios of tryptophan metabolism genes, for experiment GSM513.

Gene Symbol	Log$_2$ ratio	Gene Symbol	Log$_2$ ratio
'atoB'	1.12	'trpS'	5.85
'yqeF'	-1.81	'katE'	-0.44
'fadB'	2.63	'katG'	1.41
'sucA'	1.82	'tynA'	-0.79
'tnaA'	1.47		

Naturally, we expect the cell to intensify the process of the amino acid degradation. Indeed, in agreement with the observations of the authors, the addition of tryptophan is followed by up-regulation of the tryptophan metabolism pathway, as present in Kyoto Encyclopaedia of Genes and Genomes database. Most of the tryptophan metabolism genes show subtle to substantial up-regulation with the exception of gene yqeF which shows significant down-regulation, as highlighted in Table 4.1. However, according to KEGG gene yqeF is also member of other biochemical pathways, which may be responsible for its behaviour.

In another example, the Pentose Phosphate pathway in the diauxic shift experiments (DeRisi, Iyer & Brown 1997) discussed in the Results section (4.5) six genes included in the pathway show up-regulation while another six show down-regulation (Figure 4.2). Thus looking at the pathway in isolation is clearly not sufficient for us to be able to make an informative guess about its state of activity. However, an examination of the pathway membership of the genes reveals that most up-regulated genes are unique members of the Pentose Phosphate pathway, while all down-regulated genes are involved in other pathways, in most cases the Purine metabolism pathway. Given that Purine metabolism is severely down-regulated, as discussed in the publication accompanying the data, this pathway may well be responsible for the observed state of expression of the latter group of genes. Such observations offer strong evidence that taking the multi-membership nature of genes into account is a sensible choice that may be beneficial to pathway based microarray analysis.

4.4 Methods

To render data analysis more comprehensive, each microarray dataset is trimmed to only include genes contained in KEGG pathways. We apply discretisation, that is, the state of expression of each gene is defined as up-regulated, down-regulated or stable, based on a chosen set of thresholds.

4.4.1 Hill climbing

We facilitate a hill climbing algorithm (Michalewicz & Fogel 2004) that changes the possible multi-membership gene configuration. This is an optimization algorithm based on an iterative local search for a solution to a problem. A small change is introduced at each step of the process and the produced solution evaluated. If it is established that a change has led to a better solution to the problem in hand it is retained, otherwise it is discarded. The process is usually repeated until no better solution can be obtained.

In our implementation, we are essentially changing the allocation of multi-membership genes to their constituent pathways to identify the pathways that are more influential as far as the expression of each gene is concerned. Assigning a gene to a pathway, suggests that the biological system requires this gene’s involvement in the function of that pathway, and that the state of differential expression of that gene is due to its involvement in the activity of the pathway in hand. In contrast, removing a gene from a pathway suggests that the state of expression of that gene, be it up- or down-regulation, is not due to its involvement in the activity of that pathway, in the particular experiment.

For example, removing a down-regulated gene from a pathway implies that the reduced production of RNA by the gene is not related to the contribution of its protein product to that particular pathway. Consequentially it also suggests that some other of the genes' constituent pathways require less of its contribution and in that sense affects its expression in a negative manner.

Importantly, given that a gene member of one or more biochemical pathways exhibits differential expression, we assume that the gene is contributing to the activity of at least one these pathways. In agreement with this assumption, we do not consider as valid a configuration where a differentially expressed gene has not been assigned to at least one of its constituent pathways.

4.4.2 Algorithm

We will first define some notation that is used within our methods and algorithms. Let *P* be an *N* row by *M* column binary matrix, $P \in B^{NxM}$. Let p_{ij} (the element in the *i*th row and *j*th column of matrix *P*) = 1 if gene *i* is a member of pathway *j*, and P_{ij} = 0 if gene *i* is not a member of pathway *j*. Therefore *P* represents a snapshot of KEGG membership of genes to pathways for a given species and does not change.

Let $A \in B^{NxM}$ be a binary matrix such that $P\text{-}A \in B^{NxM}$. *A* represents a potential allocation of genes to pathways and will be used by our method, see algorithm 4.1. Here a_{ij} = 1 if gene *i* is allocated to pathway *j* and a_{ij} = 0 if gene *i* is not allocated to pathway *j*.

The restriction $P\text{-}A \in B^{NxM}$ means that *A* can define pathways to have less genes than originally in *P*, but can never have genes that contradict *P*, i.e. we do not allow allocations that would be contrary to that in KEGG.

Let us assume that we have a single set of gene expression data (one experiment) for the *N* genes called *G*. We score an allocation on how much each pathway is down or up regulated according to equations (4.1) to (4.3), note that the constant *c* is a threshold parameter.

$$X(i) = \begin{cases} +1, \text{if } G(i) > \text{c} \\ -1, \text{if } G(i) < \text{-c} \\ 0 \text{ ,otherwise} \end{cases} \tag{4.1}$$

$$F(A) = \sum_{j=1}^{M} \left| \sum_{i=1}^{N} H\left(a_{ij}\right) \right| \tag{4.2}$$

$$H(a_{ij}) = \begin{cases} X(i), \text{if } a_{\text{ij}} = 1 \\ 0 \quad \text{,otherwise} \end{cases} \tag{4.3}$$

X (*i*) has a value of +1, -1 or 0 if gene *i* is up-, down-regulated or stable respectively. *F*(*A*) is our fitness function, which we aim to maximise by changing the allocation of multi-membership genes to their corresponding pathways. We use equation (4.3) to define if gene *i* is a member of pathway *j*, which is true if a_{ij}=1, and if that is the case to define if the gene is up-, down-regulated or stable. The value of ∑$H(a_{ij})$ reveals the difference between the numbers of up- and down-regulated genes in pathway *j*. Thus, the more genes of similar expression are allocated to pathway *j* the greater the absolute value of ∑$H(a_{ij})$ becomes for that pathway.

We have explored the effect of three different starting genes to pathways allocations on the subsequent performance of the algorithm, each one characterised by different properties. Algorithm 4.1 presents the main body of the algorithm, in pseudocode, which performs the hill climbing search for the fittest genes to pathways allocation. Algorithm 4.2 represents the preliminary step of setting up the starting gene configuration.

In the case of single-membership starting allocation, only the single-membership genes are assigned to pathways. In the case of full membership, multi-membership genes are assigned to all possible pathways they belong to. In the case of directed membership, we allocate single-membership genes to their corresponding pathways, and then go through the pathways that are still empty, to check if they would contain more up- or down-regulated genes upon full allocation. If the full allocation contains more up-regulated genes, we randomly assign one of the up-regulated genes to the corresponding pathway in the starting allocation. If on the other hand the full allocation contains more down-regulated genes we randomly assign one of the down-regulated genes to the pathway, in the starting allocation.

ALGORITHM 4.1: SEARCH ALGORITHM

```
1)  Input: $a$ = list of gene IDs coupled with their pathway IDs
2)  Input: $b$ = Expression vector of $\log_2$ ratios (only KEGG
    pathways genes)
3)  Input: $c$ = threshold for up-/down-regulated genes
    Input: allocation_type = one of {single, multiple, directed}
4)  Remove all genes between $+c$ and $-c$
5)  If allocation_type = single then allocate single-membership
    genes to their pathways (thus create $A$)
6)  Elseif allocation_type = multiple then allocate all genes to
    all the pathways they are members of (thus create $A = P$)
7)  Elseif allocation_type = directed then Call Algorithm 4.2
8)  Get fitness $F(A)$, set F_old = $F(A)$
9)  For $j$ = 1: number of iterations
10)   Save gene configuration
11)   Use $P$ to randomly choose a gene ($i$) with multi-
      membership and randomly choose one of  the pathways($j$) it
      belongs to
12)   If according to $A$ gene ($i$) is already present in the
      pathway ($j$) then remove the gene, i.e. set $a_{ij} = 0$
13)   Else if not present, place it in the pathway, i.e. set
      $a_{ij} = 1$
14)   End if
15)   If upon completion of steps (10) to (14) the gene is not
      assigned to at least one pathway, randomly choose a
      pathway and assign the gene to it
16)   Estimate fitness $F(A)$
17)   If $F(A)$ > F_old set F_old = $F(A)$
18)   Else if $F(A)$ < F_old restore gene configuration (from step
      (10))
19) End for
20) Output: $A$ = Matrix representing genes to pathway allocation
```

ALGORITHM 4.2: SET DIRECTED STARTING ALLOCATION

```
Allocate single-membership genes to their pathways creating A
Set Q = a list of pathways that do not contain single-membership genes
For k = 1: length of Q
  If $\sum_{i=1}^{N} H\left(a_{iQ_k}\right)$ > 0 Then
    Let x = a random up-regulated gene from pathway $Q_k$
    Set $a_{xQ_k} = 1$
  End if
  If $\sum_{i=1}^{N} H\left(a_{iQ_k}\right)$ < 0 Then
    Let x = a random down-regulated gene from pathway $Q_k$
    Set $a_{xQ_k} = 1$
  End if
End for
```

It is important to note that each starting allocation has different properties. In the case of single-membership starting allocation, the presence of a single-membership gene in a pathway will direct the algorithm to fill that pathway with genes of similar behaviour. Hence, if a pathway initially contains a down-regulated single-membership gene, the algorithm will keep assigning more down-regulated genes to it. This is a sensible choice, because principally the behaviour of a single-membership gene can be only attributed to its involvement in that particular pathway. Thus it constitutes some strong evidence of the pathway's behaviour. It is worth noting that in chapter 3 we established that single-membership genes show more consistent behaviour and that observing contradicting expression in this group of genes is rarer than in the case of genes that are members of many pathways.

Starting from full membership allocation may also be beneficial in a different manner. Assigning all differentially expressed genes to a pathway will influence the hill climbing search to remove genes contradicting the expression of the majority from the path, leading towards a solution where the path is filled with up-regulated genes if they are present in greater numbers or down-regulated genes if the opposite is true. This also makes sense, as if for example, upon full allocation a certain pathway contains more up-regulated genes, this can be seen as some indication, although not a definitive one, that the pathway is likely to be up-regulated and vice versa.

However, there are cases where pathways do not contain single-membership genes or none of them show differential expression, to direct the subsequent filling of the pathway. To target such instances, we have implemented what we refer to as directed membership allocation. Here, the state of expression of single-membership genes directs the allocation of genes to the pathways that contain them, while the full allocation directs the filling up of pathways which do not contain single-membership genes. In essence this approach is a combination of the single and full membership starting allocations, where the behaviour of single-membership genes is taken into consideration first, while the expression of the majority is considered in pathways where this is not possible due to the lack of expressed single-membership genes. The choice of giving priority to single-membership genes is based on our hypothesis that these genes are generally more reliable indicators of pathway behaviour in agreement with the analysis presented in Chapter 3.

4.4.3 Comparison of Allocations

The allocation of a gene to its constituent pathways may be represented as a binary string. Here each position represents a pathway and 1 indicates allocation while 0 indicates that the gene is not allocated to the pathway. Consequentially, the Hamming Distance (*Hamm* below) measure (Hamming 1950) constitutes an ideal approach to identify the similarity between two allocations of the same gene. Given two binary strings of equal length, this metric considers the number of positions at which the strings differ. Table 4.2 exemplifies the gene representation we use and the hamming distance between two allocations of a gene.

In our implementation we divide the observed Hamming distance for two binary strings representing the allocations of a multi-membership gene to its constituent pathways by the length of the binary string. Then we add the results for all differentially expressed multi-membership genes and divide the sum by the number of such genes.

In mathematical terms, let $D,E \in B^{NxM}$ be binary matrices such that $P\text{-}D \in B^{NxM}$ and $P\text{-}E \in B^{NxM}$, i.e D and E are allocations of genes. Let the similarity between D and E be:

$$S(D,E) = \frac{1}{NM}\sum_{i=1}^{N}(M - Hamm(D_i, E_i)) \qquad (4.4)$$

where D_i is the ith row of D.

Table 4.2: Hamming distance between two gene allocations of gene YBR263W. The two allocations of gene YBR263W a member of 4 pathways differ at two positions, 3 and 4, for one carbon pool and methane metabolism respectively. Thus, the hamming distance between them is 2, or 2/4=0.5 (50%) as proportion of the length of the string.

	Glycine, metabolism	Cyanoamino acid metabolism	One carbon pool by folate	Methane metabolism
Allocation 1	0	1	0	0
Allocation 2	0	1	1	1

4.5 Results

This section provides a detailed discussion of the results produced by the application of the hill climbing search method to certain microarray datasets and an overview of the convergence of our algorithm. Following that there is a discussion of the consistency of the produced allocations. The section concludes with a comparison of our allocations to the original full membership allocations, using a standard statistical approach.

4.5.1 Data processing

We have applied our methodology to process data from diauxic shift experiments on *Saccharomyces cerevisiae* (DeRisi, Iyer & Brown 1997), using a threshold $\log_2$ ratio value of 1 and -1, to consider a gene up- and down-regulated respectively, as suggested by the authors.

This is a popular dataset, considered a golden standard for the analysis of the global gene expression response of *Saccharomyces cerevisiae* during diauxic shift, consisting of 7 time points for which the accompanying publication provides an informative analysis of the state of biochemical pathways. For time points 1 to 5 there is no substantial change in the expression of most genes. Following that, at time point 6 there is an evident change in the expression of a large number of genes present in *Saccharomyces cerevisiae* KEGG pathways, thus we have chosen to use this time point for analysis to demonstrate the utility of our algorithm. The results discussed here are based on twenty runs of the algorithm, from each starting allocation and a choice of the configuration exhibiting the best fitness.

In the experiment in question, yeast cells inoculated in glucose rich medium turn to aerobic utilisation of ethanol produced during fermentation, upon exhaustion of the available sugar. It is worth noting that KEGG includes both glycolysis and gluconeogenesis in one single pathway, as they share a number of common genes and a substantial part of each process is effectively a reversal of the other. Nevertheless, some genes are unique to glycolysis while others to gluconeogenesis and the two are never functional simultaneously, thus the two pathways have been separated to improve the efficiency of our analysis.

Figure 4.2 corresponds to changes occurring in the expression of genes following the diauxic shift and represents the pathway state observed when all multi-membership genes are considered active in all pathways they participate in, according to commonly used visualisation approaches. Evidently, most pathways contain both up- and down-regulated genes. Pathways including glycolysis, gluconeogenesis, the pentose phosphate pathway and pyruvate metabolism contain similar numbers of both up- and down-regulated genes, which makes it difficult to infer their state of activity.

As Figure 4.3 reveals, processing of the data with our hill climbing method changes the picture substantially. As expected, upon depletion of glucose the glycolysis pathway lacks fuel and is subsequently suppressed. Naturally, expression is now shifted in favor of the gluconeogenesis pathway.

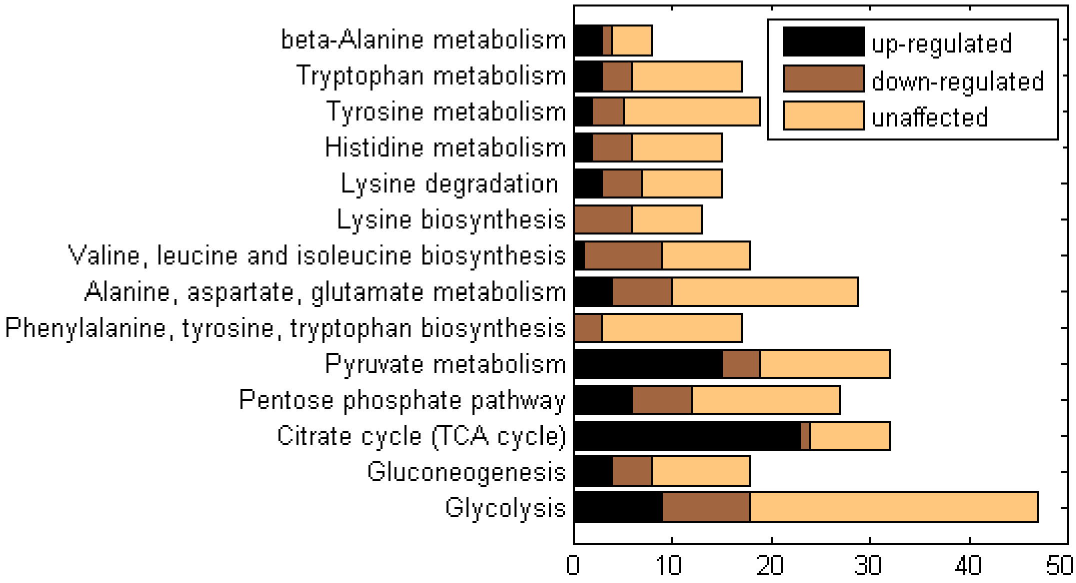


Figure 4.2 Pathway gene expression based on full allocation of genes to pathways. The figure shows the expression of all genes for a set of chosen pathways.

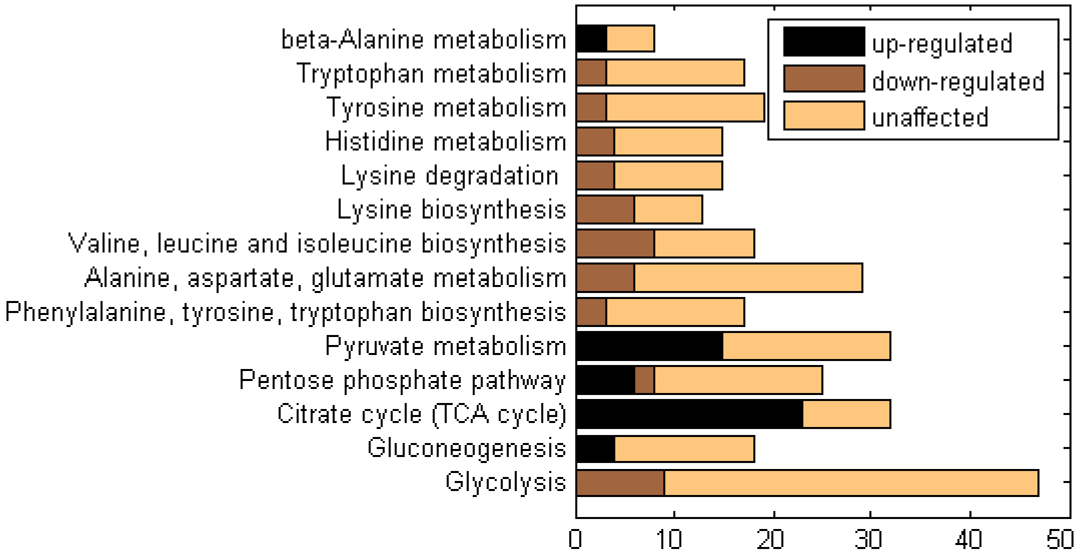


Figure 4.3 Pathway gene expression upon processing of the dataset. The figure reveals the allocation of genes for the same pathways, upon processing with the hill climbing method.

Thus, our method has correctly reallocated all down-regulated genes in the gluconeogenesis pathway to glycolysis where they also participate. Rather than towards pyruvate, reactions flow towards the biosynthetic precursor glucose-6-biphosphate which is channeled accordingly to supply the TCA cycle and gluconeogenesis.

The pyruvate metabolism pathway is now clearly activated, containing only up-regulated genes. At the same time amino acid metabolic pathways including the valine, leucine, isoleucine and methionine biosynthetic pathways are clearly repressed, in agreement with (Grosu et al. 2002). This is to be expected given the caloric restriction as the production of methionine is costly from a metabolic point of view, while valine, leucine and isoleucine are the most abundant amino acids in the cell.

The unique up-regulated gene in the valine, leucine and isoleucine biosynthetic pathways, LEU4, has been reallocated to the pyruvate metabolism KEGG pathway of which it is also a member, a pathway positively affected during the diauxic shift in agreement with our conclusion, as well as (DeRisi, Iyer & Brown 1997) and (Grosu et al. 2002).

For the unique down-regulated gene ALD6 in the beta-alanine pathway, which (Grosu et al. 2002) consider one of the 15 most positively affected pathways by the diauxic shift, our method implies that the observed down-regulation may well be due to involvement of the gene in other pathways. ALD 6 is a member of 13 distinct pathways, in fact a gene with the highest degree of membership in *Saccharomyces cerevisiae* according to KEGG, as shown on Table 4.3. Hence, the observed down-regulation may be due to its involvement in glycolysis pathway, or the metabolism of various amino acids which are suppressed.

Overall, here the algorithm has been able to allocate genes to pathways in a way that allows us to infer the state of individual pathways with increased certainty removing contradictions from the final results. Pathways are now mostly filled with genes of similar expression, which we consider to be the most indicative of a pathway’s state.

Table 4.3 *Saccharomyces cerevisiae* KEGG Biochemical Pathways containing gene ALD6
Glycolysis / Gluconeogenesis
Pentose and glucuronate interconversions
Fatty acid metabolism
Valine, leucine and isoleucine degradation
Lysine degradation
Arginine and proline metabolism
Histidine metabolism
Tryptophan metabolism
beta-Alanine metabolism
Glycerolipid metabolism
Pyruvate metabolism
Propanoate metabolism
Limonene and pinene degradation

To further investigate the results of data processing with our methodology we have applied it to *Escherichia coli* K-12 data from GEO available as experiment GSM513. *Escherichia coli* cells were grown in tryptophan enriched medium, leading to increased activity of the tryptophan metabolism pathway. Most tryptophan metabolism genes show subtle to substantial up-regulation except from yqeF which shows significant down-regulation, as noted in section 4.2 on Table 4.1.

These include the valine, leucine, isoleucine and lysine degradation pathways, that is, pathways responsible for the degradation of amino acids other than tryptophan. It is biologically meaningful to observe decline in the activity of such pathways given that the cell is presented with excess tryptophan to cover its nutritional needs. In agreement with this rationale our method has removed the down-regulated gene from the latter pathway, ascribing its behaviour to the activity of other amino acid degradation pathways.

In conclusion in the discussed experiments, our method produces results that are consistent with the findings of the publications accompanying the data, while reducing the number of genes per pathway showing contradicting expression and thus allowing us to infer the state of these pathways with higher degree of confidence.

The ability of this kind of approach to produce such consistent results and to substantially increase gene expression agreement per pathway seems interesting in itself. It adds some further evidence to the initial hypothesis that multi-membership gene expression represents a net effect, in the sense that the biological system regulates the expression of these genes to accommodate its need through the adequate function of the pathways they participate in.

4.5.2 Convergence

To examine the performance of the algorithm we looked into the convergence exhibited by it upon application to experiment GSM513, discussed in the preceding section. The result is graphically portrayed on Figure 4.4, where the solid line represents the convergence, starting from full membership, the dashed line starting from single membership and the dotted line starting from directed membership initial allocation. Each line represents the average performance of the search based on 20 separate runs of the algorithm.

Evidently, the full-membership allocation shows slightly faster convergence, however the directed membership allocation while slower seems capable of outperforming the other starting allocations in terms of fitness. Nevertheless, all three starting allocations show quite similar behaviour, with only small variability in the average final fitness and no significant effect on the overall picture produced by the processing of the data with our method.

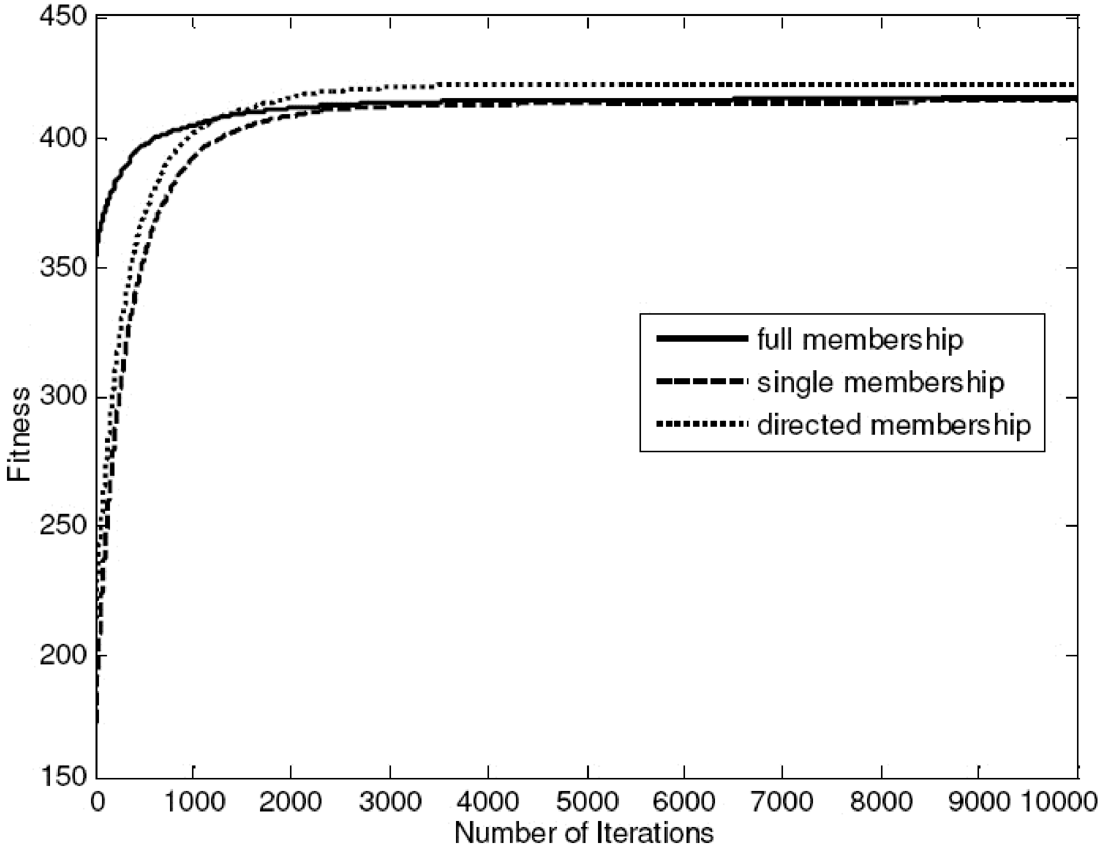


Figure 4.4 Algorithm Convergence. Each line shows the mean fitness for 20 runs of the algorithm on data from GSM513.

4.5.3 Consistency of Results

We have examined the consistency of the allocations produced by 20 separate runs of the algorithm, on the data in GSM513. The analysis was performed for all starting allocations, including the full, single and directed starting configurations, for 1000, 5000 and 10000 iterations.

This allows us to perform (*n*-1)×*n*/*2* comparisons, hence for *n* = 20 runs, we perform 190 comparisons for each configuration. Figures on Table 4.4 reveal that the algorithm produces sufficiently consistent results. Especially in the case of directed membership, for 5000 iterations, the two most distinct configurations produced by our method are still 94% similar.

Table 4.4. Comparison of results produced by 20 separate runs of the hill climbing algorithm for each separate starting point.

Membership		Full	Single	Directed
1000 Iterations	Max	97.4	92.6	94.9
	Min	86.7	72.8	88.1
	Mean	92.6	83.6	91.1
5000 Iterations	Max	98.2	93.8	99.9
	Min	87.82	78.5	94.0
	Mean	93.4	86.3	96.9
10000 Iterations	Max	97.6	97.7	99.7
	Min	88.7	86.9	92.8
	Mean	93.5	92.7	96.8

4.5.4 Comparison of Allocations

In pathway based microarray analysis, to validate data quality and establish the impact of the experimental conditions on the activity of pathways, it is common practice to estimate the probability per pathway of obtaining the results in hand by chance. For example in (Cavalieri et al. 2007) the authors describe Eu.Gene software, discussed in section 2.7.1, which calculates a *p* value for each pathway, revealing the probability of having obtained the results in hand by randomly picking genes and placing them in biochemical pathways. We have applied a similar approach to compare the results produced by our method to the standard full membership allocation. In order to achieve that, we have facilitated a microarray dataset, which can be downloaded from GEO and includes experiments for GEO platform 17. We run the algorithm, starting from directed membership, described in the Algorithm section 4.3.2, for the entire series of microarray experiments corresponding to GEO General Platform 17. We have excluded from our analysis microarray experiments that do not contain genes of fluctuating expression. This is often observed in cases of time series experiments, where at time point 1 (0 min)

RNA levels in the examined cell/tissue have not been yet affected by the experimental conditions.

As a measure for the comparison we have chosen the probability score, NBH (for Normal approximation to the Binomial distribution), described in (Swift et al. 2004), implemented to examine the quality of gene clustering results. The score is based on the hypothesis that, if a cluster of certain size contains the observed number of genes from a defined functional group of a certain size then the chance of this occurring randomly follows a binomial distribution. The *p-value* estimated based on this approach reveals the probability of observing a given number of genes or higher in a cluster, that belong to a particular functional group, purely by chance, given the overall number of genes and the overall number of genes in the cluster and the functional group.

In our implementation a cluster corresponds to a pathway, hence, given the overall number of genes and the overall number of differentially expressed genes in the experiment under consideration, the NBH statistic reveals the probability of obtaining the observed number of affected genes in a pathway or higher, purely by chance. Here the null hypothesis is that the relative changes in gene expressions in the pathway are a random subset of those observed in the experiment as a whole. For each experiment and each of the pathways containing regulated genes, we obtain this probability for full membership gene allocation and the allocation produced by our method using directed membership (see Algorithm section 4.3.2).

An issue that may arise here concerns what we define as the overall number of affected genes and overall number of examined genes, in any experimental data. One alternative that comes to mind straight away is to simply establish the sum of values above up-regulation and below down-regulation thresholds on the array of interest, and consider this to be the overall number of affected genes. In that case the overall number of genes is equal to the number of values/genes on the array that are also contained in KEGG biochemical pathways.

Figure 4.5 reveals the mean of NBH values per pathway and experiment, for the standard full allocation (solid line) and the directed membership (dotted line) allocation produced by our method, applying this rationale. The mean NBH values for the two allocations exhibit very significant correlation of 0.99 with a *p*-value of 1.376×10^{-036}.

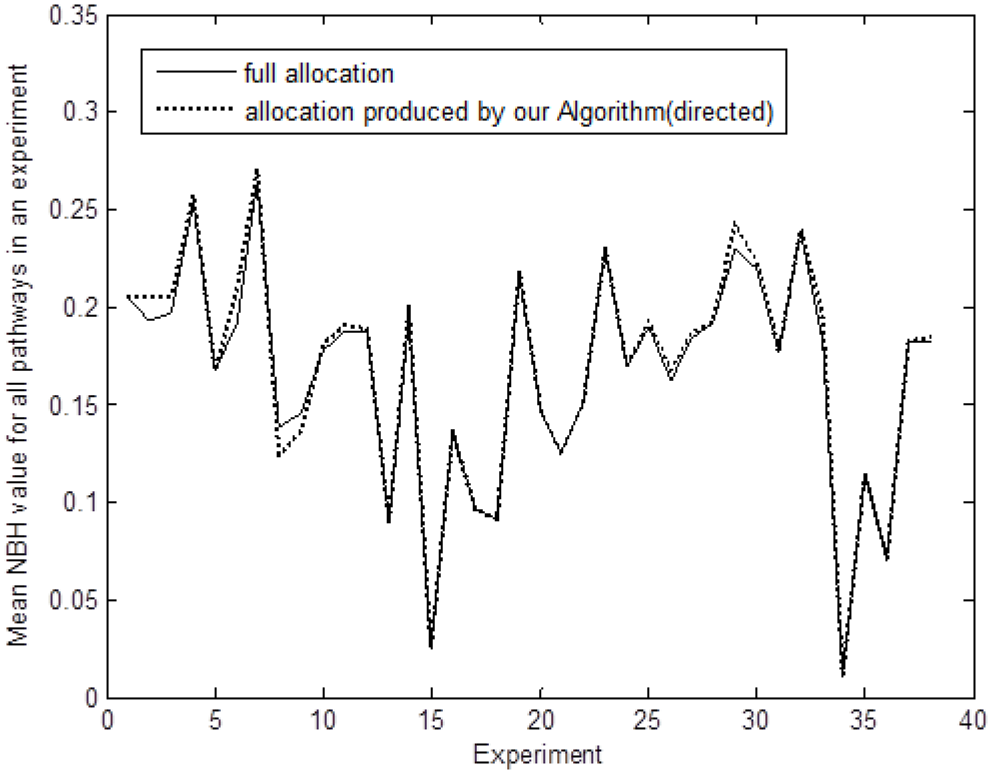


Figure 4.5 Mean NBH values per experiment (for all pathways containing genes of fluctuating expression), for standard allocation and the one produced by our methodology. As overall number of genes, we consider the intersection of genes between KEGG and each microarray. As affected we consider the number of genes contained in this intersection that show expression levels above or below the chosen thresholds.

However, given that as discussed previously, many genes are members of more than one biochemical pathway, if we simply add the number of regulated genes per pathway, we will come up with a greater number of overall affected genes per experiment, given that there are duplicates, since the same gene may appear in different pathways.

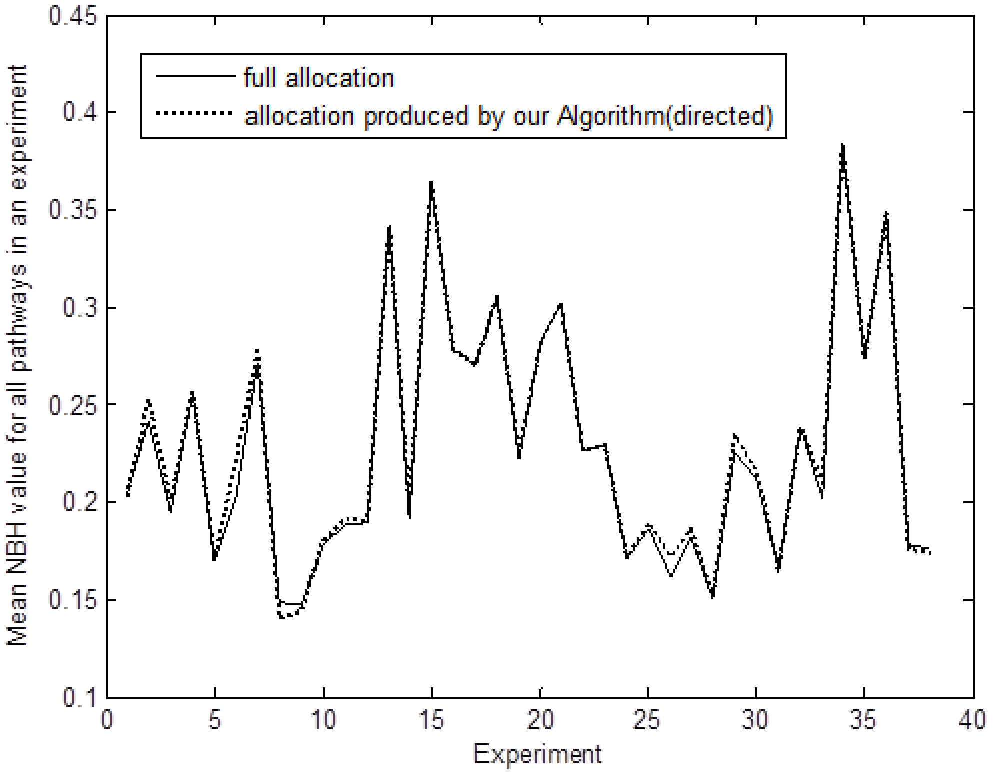


Figure 4.6 Mean NBH values per experiment (for all pathways containing genes of fluctuating expression), for standard allocation and the one produced by our methodology. The overall number of genes is equal to the sum of the sizes of all *Escherichia coli* biochemical pathways. The number of affected genes is equal to the sum of affected genes per pathway.

Similarly, if we add the sizes of all examined pathways, we will end up with a greater number of genes than the intersection between KEGG genes and the examined microarray. We have also applied this rationale to obtain the result on Figure 4.6. Again the mean NBH values exhibit highly significant correlation equal to 0.99, with a *p*-value of 6.4×10^{-58}.

In both cases it is evident that there isn't any substantial change as far as the NBH probability is concerned, when comparing standard full membership allocation to the allocation of genes produced by our algorithm. Hence, the proposed data processing approach does not affect the probabilities of obtaining the observed allocation in a positive or negative manner. However, while results seem equally valid, we have added an intuitional, biologically meaningful step to the data processing course.

4.6 Conclusions

The analytical method described here is used to identify the overriding behaviour of pathways given the up- and down-regulation of their constituent genes, observed in a microarray experiment. Given that many genes are members of more than one biochemical pathway, we have used *Escherichia coli* and *Saccharomyces cerevisiae* microarray data to allocate each of the affected genes to pathways, by maximising the number of genes that show similar behaviour, in each individual pathway. In doing this, we attempt to maximise pathway coverage, allocating as many genes as possible to a pathway, while at the same time minimise the number of contradictions, meaning the number of genes that show up- and down-regulation in the same biochemical pathway, in the same experiment.

We have shown that our method is able to effectively allocate multi-membership genes to their corresponding pathways in accordance with the underlying trend of gene regulation in that pathway and produce pathway categorised results that are biologically meaningful. By manipulating the pathway membership of the genes to follow underlying trends we can interpret microarray results centred on the behaviour of the biochemical pathways. We have also shown that the produced configurations are consistent, by comparing the results produced by subsequent runs of our algorithms. Additionally we have explored and compared different starting configurations, and discuss their advantages and disadvantages.

The methodology presented in this chapter is of potential interest, as it may assist a biologist to infer the state of individual biochemical pathways, based on microarray data. Given that the multi-membership pathway nature of genes has not been extensively considered in currently used tools for pathway based microarray analysis, this method suggests an interesting innovative approach.

Chapter 5: Pathway analysis using simulated annealing and a genetic algorithm

5.1 Introduction

We have proposed a methodology that takes into account the expression of all genes in a given organism, that are members of biochemical pathways, and the consensus of gene expression per pathway in order to identify the underlying pathway expression changes caused by the biological system through regulation of the expression of their constituent genes (Pavlidis, Payne & Swift 2008). Unlike other approaches where genes are treated as stable or differentially expressed (Cavalieri et al. 2007), our methodology considers the state of expression of individual genes in terms of up- or down-regulation and attempts to ascribe any observed inconsistencies in gene expression in a pathway, to the involvement of some of its genes in the activity of other pathways of which they are also members.

In the previous chapter, we implemented a hill climbing (HC) search approach which was able to produce consistent results, in agreement with the publications accompanying the data in question, presented and discussed in chapter 4. However, given the tendency of the hill climbing search to get trapped in local maxima, we proceeded further, applying a simulated annealing (SA) (Kirkpatrick , Gelatt & Vecchi 1983) and a genetic algorithm (GA) search technique in order to explore the performance of each one on the same and some additional microarray experiments.

Importantly, the differences in the final fitness reached by each of these methods do not have a straightforward biological meaning. Consequentially, we shifted our efforts towards exploring the similarity of the produced results, in conjunction with their corresponding fitness, by using two complimentary approaches. In particular, we developed a methodology for estimating the similarity of two gene allocations based on the hamming distance measure and the probability of observing any given hamming distance, or smaller, purely by chance. This approach directly reveals the similarity between two genes to pathways allocations. Additionally, we adopted the fuzzy adjusted rand index (FARI) metric (Brouwer 2009), widely used measure of agreement for categorical data. In this case we observe for each distinct pair of genes, if two allocations have placed them in the same pathways or in different ones. The greater the agreement, the greater the value we obtain.

In this chapter we present a more detailed discussion of the applied methodologies and proceed to present and discuss the performance of the three search algorithms. Interestingly, according to both implemented similarity metrics, results produced by all methods appeared highly consistent. While the simulated annealing search was able, in some cases, to reach slightly higher fitness values the difference was statistically insignificant.

5.2 Methods

Following the general methodology discussed in chapter 4, all microarray datasets were trimmed to only include genes present in KEGG pathways. For example, KEGG contains 1384 *Escherichia coli* pathway genes out of a total of 4288 protein-coding genes (Blattner et al. 1997), for the harmless laboratory strain K12 MG1655. As discussed, we applied discretisation of genes into three categories, namely up-, down-regulated and stable, based on an adequately chosen threshold and processed each experiment with a hill climbing, a simulated annealing and a genetic algorithm, to alter the possible allocation of multi-membership genes to their constituent pathways.

We assume that a differentially expressed gene is regulated by the biological system to contribute to the activity of at least one of the pathways it is a member of. Thus, any configuration that satisfies this criterion is considered valid, while an allocation where a differentially expressed multi-membership gene has not been assigned to any of its constituent pathways is rejected. Our goal was to identify, for each such gene, the pathways whose activity requires the observed behaviour of the gene in question. Allocation of a gene to one of its constituent pathways suggests that the biological system has adjusted the expression of that gene in the given manner to satisfy the activity of that pathway. Naturally, not allocating a gene to a pathway suggests its expression is not related to its involvement in the activity of this particular pathway.

Here we proceed to present the search methods we used to extend our analysis, namely simulated annealing and a genetic algorithm, followed by discussion of our adaption of the hamming distance metric and a methodology of calculating the probability of observing it purely by chance. Furthermore, we discuss our implementation of the fuzzy adjusted rand index metric, to analyse similarity between genes to pathways allocations produced by our methods.

5.2.1 Simulated Annealing

In contrast to hill climbing, simulated annealing may occasionally accept a solution of worse fitness in the initial stages of the process, depending on a probability which is defined by gradually decreasing parameter T, termed temperature. As T decreases the probability drops and it becomes less and less likely for a solution of lower fitness to be accepted. Allowing for worse configurations to be adopted in the context of the described process, allows simulated annealing to escape local maxima and minima.

In our application of the method, we have chosen a starting temperature $T = 1$ and a final temperature $T = 0.01$ as appropriate for 10000 iterations which have proven sufficient for the algorithm to converge. At step 17) of the hill climbing algorithm (Algorithm 4.1), described in chapter 4, section 4.3.1, the simulated annealing approach accepts an allocation of lower fitness with a probability which can be estimated based on equations (5.1) to (5.3).

$$P_t = e^{\frac{-\Delta F}{T_t}} \tag{5.1}$$

$$T_t = T_{t-1}\lambda \tag{5.2}$$

$$\lambda = e^{\frac{\log(T_R)-\log(T_0)}{R}} \tag{5.3}$$

Here P_t is the probability of accepting an allocation of lower fitness at the current iteration *t*, *-ΔF* is the difference between the current fitness and the one of the allocation at the previous iteration, T_t is the current temperature and T_R the temperature at the last iteration, *λ* is a constant, representing the cooling factor and *R* the number of iterations for the search to complete.

5.2.2 Genetic Algorithm

In a genetic algorithm, as discussed in section 2.6.3, candidate solutions are represented by the so called chromosomes, in analogy to chromosomes storing genetic information in living cells. Most commonly, from a computational point of view, a chromosome consists of a binary string of ones and zeros, where each bit constitutes a gene. At every iteration of the algorithm the totality of binary strings, which constitute the population, 'evolves' to give birth to a new population. The individuals constituting the new population are selected and preserved based on their fitness.

In our implementation of a genetic algorithm search approach to identify the pathways responsible for the behaviour of multi-membership genes, each chromosome represents an allocation of genes to pathways. As discussed in chapter 4, a binary string represents the allocation of each multi-membership gene to its constituent pathways. Each position for a particular gene corresponds to a pathway, with 1 suggesting that gene exhibits the observed differential expression due to its contribution to that path. On the other hand 0 suggests that the observed gene expression is not due to the contribution of the gene to the particular pathway.

```
ALGORITHM 5.1: GENETIC ALGORITHM

1)   INPUT: a = list of gene IDs coupled with their pathway
     IDs, b = expression vector of log2 ratios, c = threshold
     for up-/down-regulated genes
2)   Remove all genes between +c and -c
3)   Create S random Parent chromosomes
4)   Get fitness F of each Parent chromosome
5)   For i = 1:number of generations
6)     For j = 1:number of individuals in Parent
7)       Call mutation Algorithm with input Parent_j
8)     End for
9)     Create a random list List of (number of Mutated)
10)    For j = 1:(number of Mutated)/2
11)      Call crossover Algorithm with input Mutated(List(j)),
         Mutated(List(j+1))
12)    End for
13)    Get the fitness of each Mutated and Crossover chromosome
14)    Use roulette-wheel selection to select S chromosomes
15)    Set Parent =  selected S chromosomes
16)  End for
17)  OUTPUT: Best Individual and Fitness
```

At each generation, individuals are subjected to crossovers and mutations, changing the allocation of multi-membership genes to their constituent pathways. From the produced offspring we preserve the fittest individuals to serve as our current parent generation and repeat the process for a chosen number of iterations, in order to reach the best possible fitness.

Algorithm 5.1 represents the main body of the genetic algorithm, Algorithm 5.2 describes the crossover process, while algorithm 5.3 the mutation process, called at steps 7 and 11 of the main genetic Algorithm 5.1, respectively. Within Algorithm 5.2 the operator $C = [A_1,A_2,\ldots,A_x,B_{x+1},B_{x+2},\ldots B_N]$ concatenates the lists A and B preserving order and sets C to be the result. A generation consisting of a hundred individuals proved sufficient to reach the maximum possible fitness over about four hundred generations.

ALGORITHM 5.2: CROSSOVER ALGORITHM

```
1)  INPUT: Parent A and Parent B
2)  Choose a random number $x$ between 1 and length of Parent A
3)  Set Crossover A = $[A_1,A_2,\dots,A_x,B_{x+1},B_{x+2},\dots B_N]$
4)  Set Crossover B = $[B_1,B_2,\dots,B_x,A_{x+1},A_{x+2},\dots A_N]$
5)  OUTPUT: Crossover A, Crossover B
```

ALGORITHM 5.3: MUTATION ALGORITHM

```
1)   INPUT: Individual
2)   Create Mutated equal to Individual
3)   For $k$ = 1:length of Mutated
4)      Produce a random number $a$ between 0 and 1
5)      If $a$<1/length(Mutated$_k$) randomly choose a position $x$ in
        Mutated$_k$
6)      If according to Mutated gene ($k$) is already in path ($x$)
        remove it, i.e. set Mutated$_{kx}$=0
7)      Else if gene ($k$) is not allocated to path ($x$) place it
        in the path, i.e.  set Mutated$_{kx}$=1
8)      If upon completion of steps (4) to (7) the gene is not
        assigned to at least one pathway, repeat steps (4) to
        (7)
9)      End if
10)  End for
11)  OUTPUT: Mutated Individual
```

5.2.3 Hamming Distance and probabilities

To obtain a more meaningful interpretation of the observed hamming distances, we developed a methodology to estimate the probability of obtaining any hamming distance between two allocations produced by our methods, purely by chance. In particular, for any given multi-membership gene, we first estimate the probability of observing each possible hamming distance between pairs of allocations. The methodology is based on estimating the number of all possible valid binary strings representing the allocation of a multi-membership gene to the pathways it is a member of, according to the KEGG database.

Table 5.1 Hamming Distances between two allocations of a gene member of two pathways. The table reveals all possible combinations of two allocations for a multi-membership gene, participating in two distinct biochemical pathways, with the corresponding hamming distance between the binary strings representing these allocations. A string of zeros is considered invalid allocation, as we assume that a differentially expressed gene is contributing to the activity of at least one of its member pathways.

Allocation 1	Allocation 2		
	01	10	11
01	0	2	1
10	2	0	1
11	1	1	0

In the simplest case of a differentially expressed gene that is a member of only two pathways, its allocation is represented by a string of two binary digits. As already discussed in chapter 4, section 4.3.1, only solutions where the gene has been allocated to at least one of the pathways, of which it is a member, are considered valid. Therefore, we do not consider a string consisting solely of zeros as an acceptable, valid allocation. The square matrix on Table 5.1 represents all valid combinations of allocations, for a gene, member of two biochemical pathways, giving rise to all possible hamming distances.

The probability of observing any of the hamming distances on Table 5.1 is equal to the number of combinations giving birth to each of the possible hamming distances, namely 0, 1 and 2, divided by the overall number of possible combinations, equal to 9 in this example. Following this rationale, in the simplest case of a gene member of two pathways we obtain the probabilities shown on Table 5.2.

Table 5.2 Probability of obtaining any hamming distance between two allocations of a gene member of two pathways. Given the number of possible combinations (Table 5.1) of allocations for a gene member of two pathways, and the hamming distance between them, this table shows the probability of obtaining each possible hamming distance, purely by chance.

Hamming Dist.	0	1	2
Probability	0.333 (3/9)	0.444 (4/9)	0.222 (2/9)

For a gene that is a member of any possible number of pathways, the number of such combinations for any given hamming distance between 0 and *r* can be estimated according Table 5.3. Here *n* is the number of pathways the gene is a member of and *r* is the hamming distance between two allocations. As Equation (5.4) demonstrates, we can summate from 1 to *n* in order to get the number of possible combinations corresponding to all possible hamming distances.

Table 5.3: Number of combinations of pairs of allocations of hamming distance between 0 and *r*. Using the equations on the table we can estimate the number of all possible combinations of allocations, represented as binary strings, of hamming distance from 1 to *r*.

Hamming distance	Number of possible occurrences
0:	$2^n - 1$
1:	$(2^n - 1 - n)n + n(n-1) = (2^n - 2)n$
2:	$\left(2^n - 1 - \binom{n}{2}\right)\binom{n}{2} + \binom{n}{2}\left(\binom{n}{2} - 1\right) = \binom{n}{2}(2^n - 2)$
r:	$\left(2^n - 1 - \binom{n}{r}\right)\binom{n}{r} + \binom{n}{r}\left(\binom{n}{r} - 1\right) = \binom{n}{r}(2^n - 2)$

In the context of this text, we work with allocations of more than one expressed multi-membership genes to their pathways. This however, does not constitute a problem and the hamming distance probabilities can still be estimated following the above discussed rationale. In the simplest case of two genes, members of two pathways each, we can estimate the probability of obtaining all possible hamming distances using Table 5.2 and applying simple addition and multiplication of the values as shown on Table 5.4.

$$\begin{aligned}
&\underbrace{2^n-1}_{\text{Table 5.3,hamm=0}} + \underbrace{\sum_{r=1}^{n}\binom{n}{r}\left(2^n-2\right)}_{\text{Table 5.3,hamm=1,...,}n} = 2^n-1+\left(2^n-2\right)\underbrace{\sum_{r=1}^{n}\binom{n}{r}}_{=2^n-1} \\
&= 2^n-1+\left(2^n-2\right)\left(2^n-1\right) \qquad (5.4) \\
&= 2^n-1+\left(2^n-1\right)\left(2^n-1\right)-2^n+1 \\
&= \left(2^n-1\right)\left(2^n-1\right)
\end{aligned}$$

Table 5.4: Combined Hamming distance and probability for a pair of genes, members of two pathways. The table exemplifies how to estimate the combined hamming distance for two multi-membership genes, members of two distinct biochemical pathways each, along with the respective combined probability. Here again, we assume that any configuration, where each gene is allocated to at least one pathway is valid and that each one is equally likely to occur by chance.

Hamming/ Probability	0/0.333	1/0.444	2/0.222
0/0.333	0(0+0)/ 0.111(0.333x0.333)	1(1+0)/ 0.148(0.444x0.333)	2(2+0)/ 0.074(0.222x0.333)
1/0.444	1(0+1)/ 0.148(0.333x0.444)	2(1+1)/ 0.197(0.444x0.444)	3(2+1)/ 0.987(0.222x0.444)
2/0.222	2(0+2)/ 0.074(0.333x0.222)	3(1+2)/ 0.987(0.444x0.222)	4(2+2)/ 0.049(0.222x0.222)

Each pair of hamming distances is added to obtain the combined hamming distance, while each pairs' corresponding probability is multiplied to obtain the probability of observing the combined hamming distance in question. For any number of *N* genes we can obtain the corresponding values using an *N* dimensional matrix like the one table 5.4. As the number of genes grows this becomes computationally expensive, however the problem is circumvented, as each gene can be added at a sequential step, through a process of merging and expanding the matrix. For example merging the data for the two genes represented on Table 5.4, gives rise to the matrix on Table 5.5.

Finally, for any observed hamming distance between two subsequent runs of our search algorithms, for a given microarray experiment, we can calculate the probability of observing the hamming distance in hand or smaller through simple addition of probabilities. For example, using table 5.5, the probability of observing a hamming distance of 2 or smaller is equal to the sum of the probability of observing a hamming distance of 0, 1 and 2, that is, 0.111+0.296+0.345=0.752.

Table 5.5 Compact Hamming distance and probability for two genes, members of two pathways each. Table 5.5 is produced by merging Table 5.4, to only show each possible hamming distance and the corresponding probability of observing it by chance, for a set of two expressed multi-membership genes. Each gene is a member of two distinct biochemical pathways.

Hamming Distance	0	1	2	3	4
Probability	0.111	0.296	0.345	0.197	0.049

5.2.4 Fuzzy Adjusted Rand Index

The adjusted rand index (ARI) is a common quantitative measure of cluster similarity, widely accepted to possess the most desirable properties in the case of comparing crisp partitions. It has been recently extended to fuzzy clustering giving the fuzzy adjusted rand index (FARI) (Brouwer 2009). For each pair of elements FARI examines if both clustering arrangements have placed the pair in the same or different clusters. Unlike in the case of ARI where each element can only be placed in one cluster, here an element can be placed in a number of clusters, hence fuzzy ARI. We have adopted FARI, as defined in (Brouwer 2009), to compare allocations of multi-membership genes produced by separate runs of our algorithms on the same microarray dataset, given that each arrangement may place a gene in one or more pathways. For our purposes clusters correspond to pathways, assuming equal weights for the contribution of a gene to all its member pathways.

While the hamming distance between two multi-membership gene allocations reveals biological similarity, answering the question of how similar two allocations are the fuzzy adjusted rand index examines if each pair of genes is placed together or in different pathways by subsequent runs of our algorithms.

Interestingly, we have discovered that upon processing of microarray data with our methods we sometimes observe that two allocations exhibiting the same fitness differ in terms of hamming distance. This result can occur in cases where groups of genes are placed together but in different pathways by separate runs of our scripts. In particular for allocations of the same or very similar fitness, accompanied by significant hamming distance, high FARI value can reveal the occurrence of the above described phenomenon.

5.3 Results

In this section we present a comparison of the results produced by the three search algorithms, upon their application on set of microarray experiments. The comparison is first performed in terms of fitness values reached by each method and the number of iterations required for the algorithms to converge. Following that, we analyse the performance of the scripts based on application of the FARI metric, the hamming distance measure and the accompanying probability of observing it, along with some correlation analysis.

5.3.1 Methods' Performance

In order to examine and compare the performance of the three search algorithms the dataset consisting of 46 microarray experiments from GEO platform GPL17 was subjected to processing by each one of them. The resulting mean fitness reached by the implementation of each algorithm in twenty separate runs is shown on Figure 5.1.

Interestingly, all methods exhibit quite similar behaviour in terms of the fitness accompanying the produced allocations. In most cases the simulated annealing approach is able to reach only marginally higher fitness values. However, the difference is subtle with a two sample t-test revealing no significant difference between the values corresponding to each search methodology. This result is summarised on Table 5.6, which shows the mean of the minimum, maximum and mean fitness reached for the entire set of 46 experiments, upon twenty separate runs of each method.

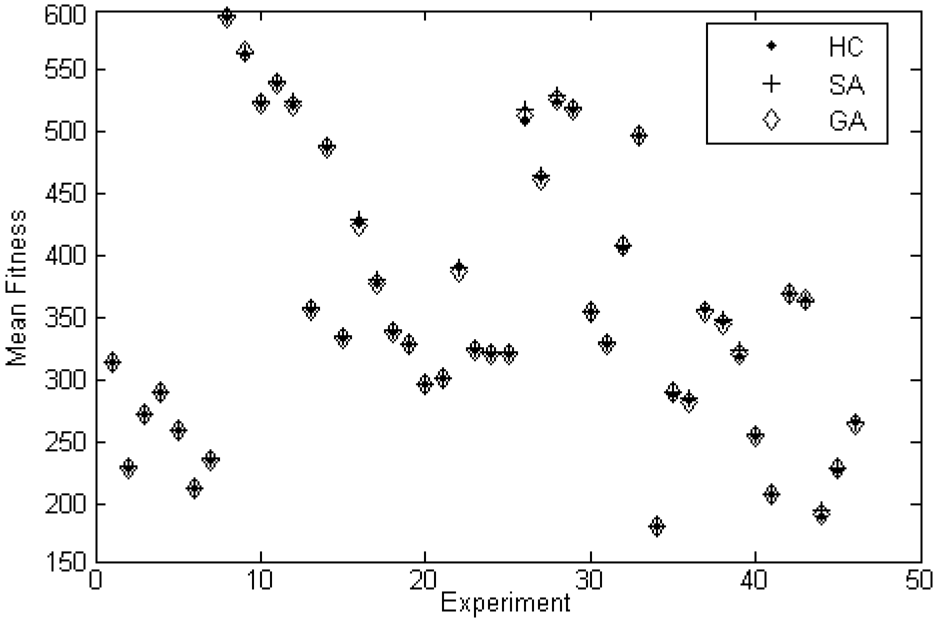


Figure 5.1 Mean fitness reached by each method, per experiment for GPL17. The figure reveals the mean fitness reached by method per experiment, for 46 experiments corresponding to platform GPL17 from GEO, in twenty separate runs of each method.

Table 5.6 Mean of the minimum, maximum and mean fitness reached by each method. The table summarises the fitness reached by the each method for GPL17.

Hill Climbing			Simulated Annealing			Genetic Algorithm		
Max.	Min.	Mean	Max.	Min.	Mean	Max.	Min.	Mean
357.5	354.0	356.1	359.0	355.0	357.4	357.8	353.9	356.1

The Convergence of each optimisation method for a subset of four microarray experiments is visually portrayed on Figure 5.2. The experiments were chosen based on the mean fitness reached by twenty separate runs of each search approach, in order to exemplify the entire range of fitness values reached for GPL17.

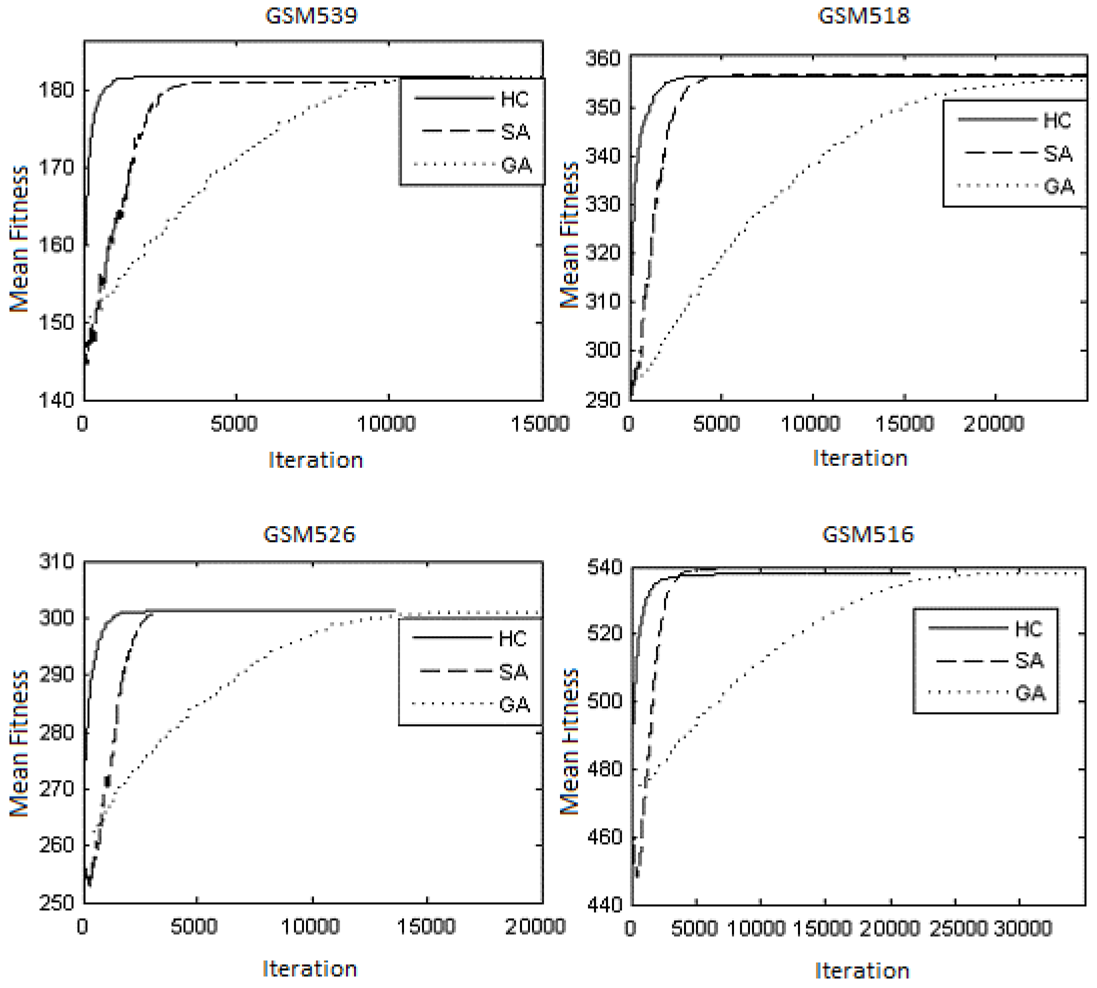


Figure 5.2 Convergence. The solid, dashed and dotted lines correspond to the mean hill climbing, simulated annealing and genetic algorithm fitness, upon 20 runs. Experiments roughly cover the range of fitness values reached in all experiments. GSM539 and GSM516 where the experiments with the least and most possible allocation positions, respectively, while GSM518 and GSM526 are equally distanced from the two extremes. The reached fitness follows suit.

In particular the mean fitness values reached for each experiment where sorted in ascending order. GSM539 corresponds to the lower mean fitness reached for an experiment in the dataset, GSM516 to the highest mean fitness and GSM526 and GSM518 to values equally distanced from these two extremes. Evidently, the genetic algorithm approach is slower than the other methods, requiring a significantly larger number of fitness calls to converge, a common issue associated with evolutionary algorithms (Davarynejad, Akbarzadeh-T & Pariz 2007).

The hill climbing and simulated annealing methods are roughly equally efficient, with the hill climbing being slightly faster, while the simulated annealing able to reach slightly higher fitness values, in experiments with large number of expressed multi-membership genes and thus larger search space. Naturally, as the search space grows larger, due to a larger number of expressed multi-membership genes and growing number of constituent pathways to which such genes can be assigned, the algorithms require more iterations to converge. Figure 5.3 graphically portrays the mean number of iterations required for the algorithms to converge for the experiments in the dataset.

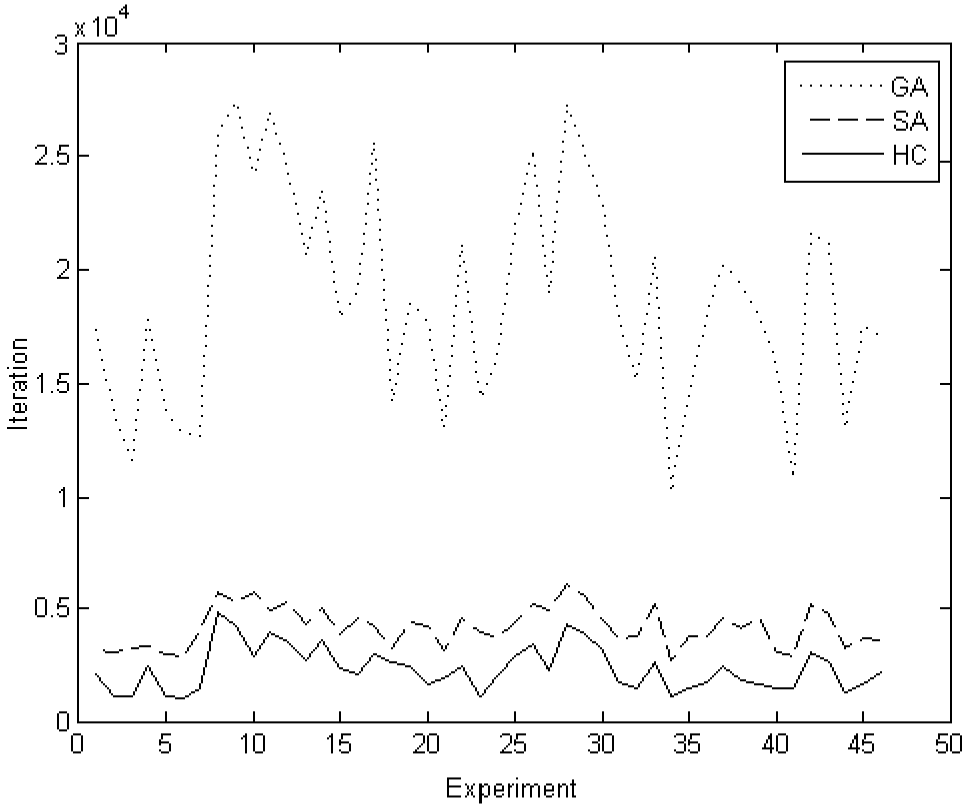


Figure 5.3 Mean convergence per experiment and method. The hill climbing method (solid line) is the fastest, closely followed by the simulated annealing (dashed line) approach, while the genetic algorithm (dotted line) proves significantly slower.

Figure 5.4 represents the same data, this time in an ordered fashion. In particular, experiments are sorted according to the number of expressed genes, from the experiment with the least number of differentially expressed multi-membership genes to the one with the largest number of such genes.

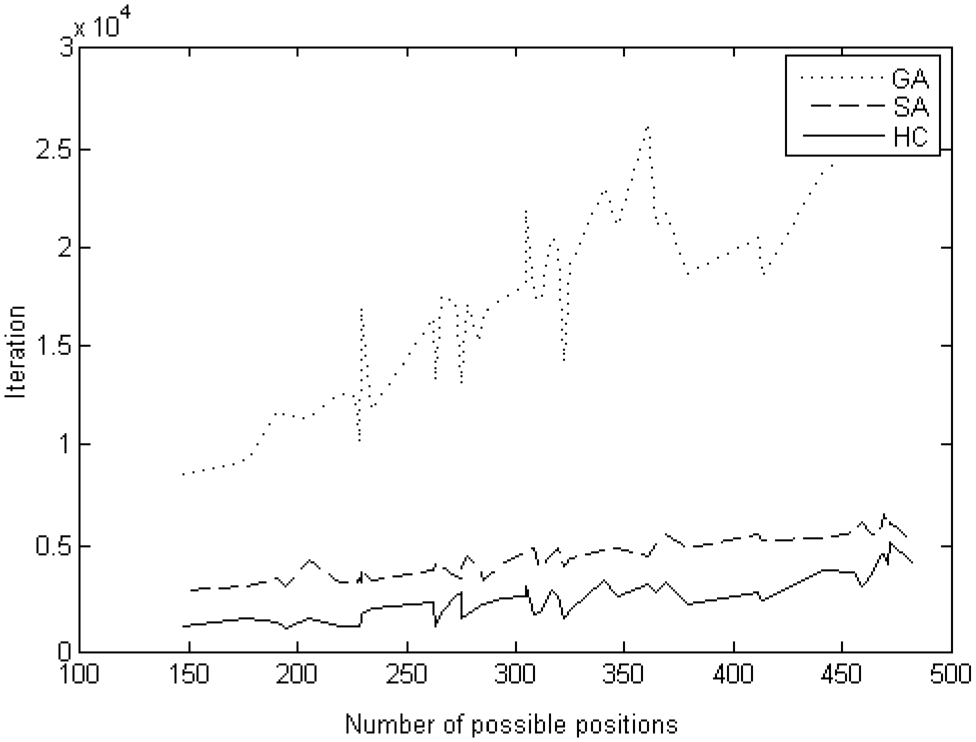


Figure 5.4 Mean convergence per experiment, according to search space size. Experiments are represented in an ordered fashion, from the one with least expressed multi-membership genes, and smallest search space to the one with most expressed multi-membership genes and largest search space.

As expected, the mean fitness value also shows an increase as the number of possible allocations of genes to pathways grows, as shown on Figure 5.5. The correlation values are highly significant, equal to 0.96, 0.97 and 0.97 for the hill climbing, simulated annealing and genetic algorithm respectively. The number of allocations of genes to pathways is determined by the number of expressed genes and the number of pathways in which the expressed multi-membership genes participate.

On the contrary, there is no significant correlation between the number of genes to pathways allocations and the mean hamming distance between allocations produced by subsequent runs of the three search algorithms, as exhibited on Figure 5.6.

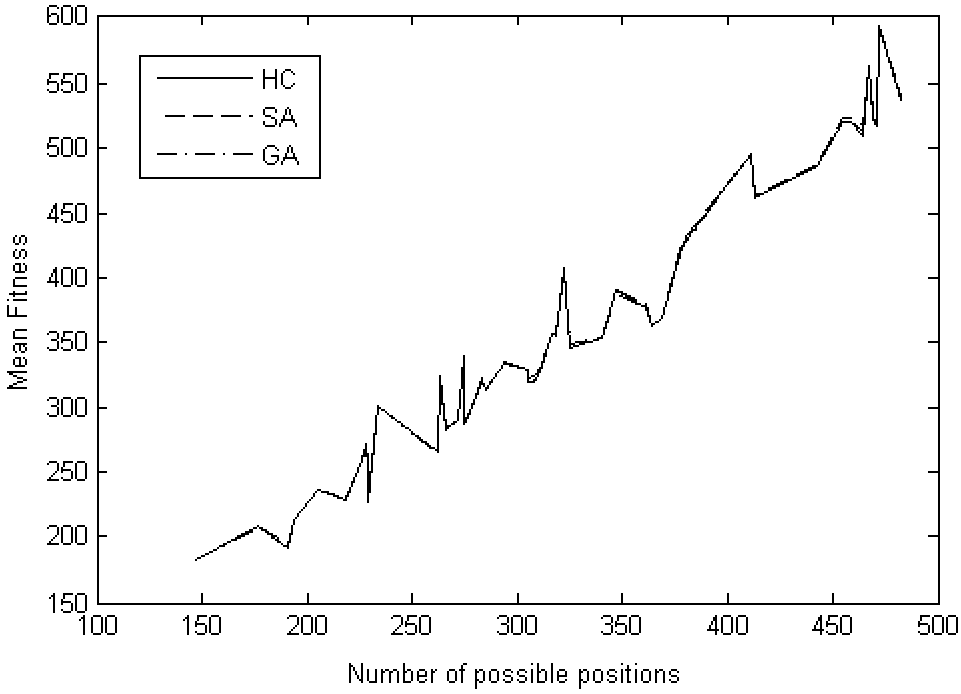


Figure 5.5 Mean fitness per experiment and method, according to search space. The hill climbing fitness values are represented with a solid line, the simulated annealing dashed and the genetic algorithm dotted line. As the size of the search space grows, following the number of possible genes to pathways allocations, all methods are able to reach higher fitness values and are virtually indistinguishable.

Here we look into the hamming distance between allocations of genes produced by each method separately. Interestingly, in this case we observe quite small correlation values of -0.27, -0.47 and -0.36 for the hill climbing, the simulated annealing and genetic algorithm respectively. Hence, allocations produced for experiments with greater number of expressed multi-membership genes, accompanied by a larger search space, do not appear less consistent and vice versa.

The same is true for the observed FARI's, where the size of the search space does not seem to exhibit influence on the consistency of the produced allocations, as shown on Figure 5.7.

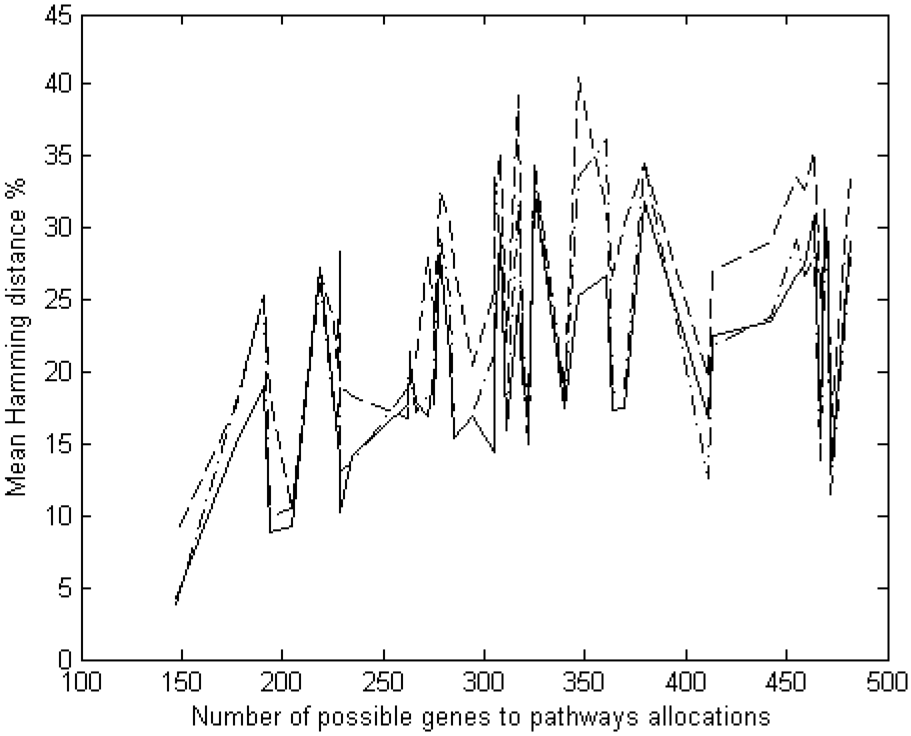


Figure 5.6 Mean hamming distance between allocations per experiment, according to search space. Experiments are ordered according to the number of possible genes to pathways allocations.

The correlation between the mean FARI value per experiment and the number of possible multi-membership gene to pathway allocations is -0.08, 0.05 and 0.03 for the hill climbing, simulated annealing and the genetic algorithm respectively. Nevertheless, FARI values themselves appear extremely high for allocations produced by separate runs of each of the search techniques, as summarised on Table 5.7.

The minimum FARI observed is 0.902, and the values remain high regardless of the observed variation in hamming distance. For example the mean FARI for pairs of allocations of hamming distance above 1 standard deviation is 0.964, 0.962 and 0.963 for the hill climbing, simulated annealing and genetic algorithm respectively.

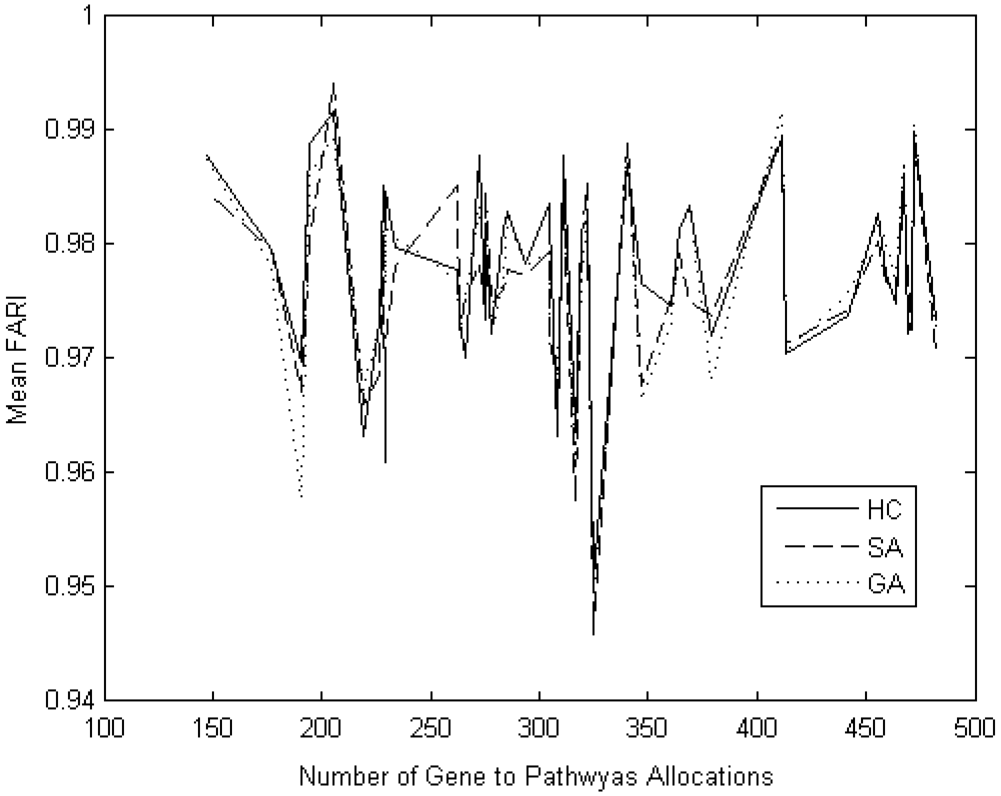


Figure 5.7 Mean FARI between allocations per experiment, according to number of possible genes to pathways allocations. For the FARI values between allocations produced by subsequent runs of each method there is no correlation whatsoever with the size of the search space (number of possible multi-membership genes' allocations).

Based on this observation we can assume with sufficient degree of confidence that in cases of pairs of allocations, exhibiting substantial hamming distance, and at the same time high FARI values, groups of genes have still been allocated together, in the same pathway, thus the FARI values are high. However, the pathways have been swapped in the two allocations, explaining the higher hamming distance values.

Nevertheless, facilitating the probability measure described in section 5.2.3 reveals that observing the hamming distances presented here, between subsequent allocations of genes to pathways by our methods, are of extremely low probability to have occurred by chance. Table 5.8 exhibits the minimum, maximum and mean probability of observing a certain hamming distance or smaller one for each method.

Table 5.7 FARI statistics between allocations produced separate runs of each search method. The table summarises the minimum, maximum and mean Fuzzy Adjusted Rand Indexes between allocations produced by twenty separate runs of the hill climbing, simulated annealing and genetic algorithm search approaches.

	Maximum	Minimum	Mean	Standard Deviation
Hill Climbing	1.000	0.928	0.978	0.012
Simulated Annealing	1.000	0.902	0.976	0.013
Genetic Algorithm	1.000	0.926	0.977	0.011

Table 5.8 Probability of observed hamming distance. The table summarises the minimum, maximum and mean probability between allocations produced by twenty separate runs of the algorithms on GPL17 data.

	Maximum	Minimum	Mean	Standard Deviation
Hill Climbing	1.92×10^{-12}	3.67×10^{-123}	9.31×10^{-16}	4.23×10^{-14}
Simulated Annealing	9.40×10^{-12}	2.25×10^{-114}	4.65×10^{-15}	2.07×10^{-13}
Genetic Algorithm	2.01×10^{-13}	4.08×10^{-116}	1.06×10^{-16}	4.42×10^{-15}

Once again, looking into the mean probability of observing any hamming distance or smaller, per experiment, and the size of the search space, as graphically portrayed on Figure 5.8, we establish an insignificant correlation value of -0.22.

Additionally, while the methods seem to reach similar values of fitness, we investigated how similar the allocations are in terms of FARIs and hamming distance, between the different search aproaches. Regarding FARIs we observed that the values remain equaly high, when we compare the results from each pair of methods, with a mean above 0.974 in all cases, as shown on table 5.9. Thus all methods seem quite consistent in placing groups of genes in the same pathways, as we previously observed for consequitive runs of each method separately.

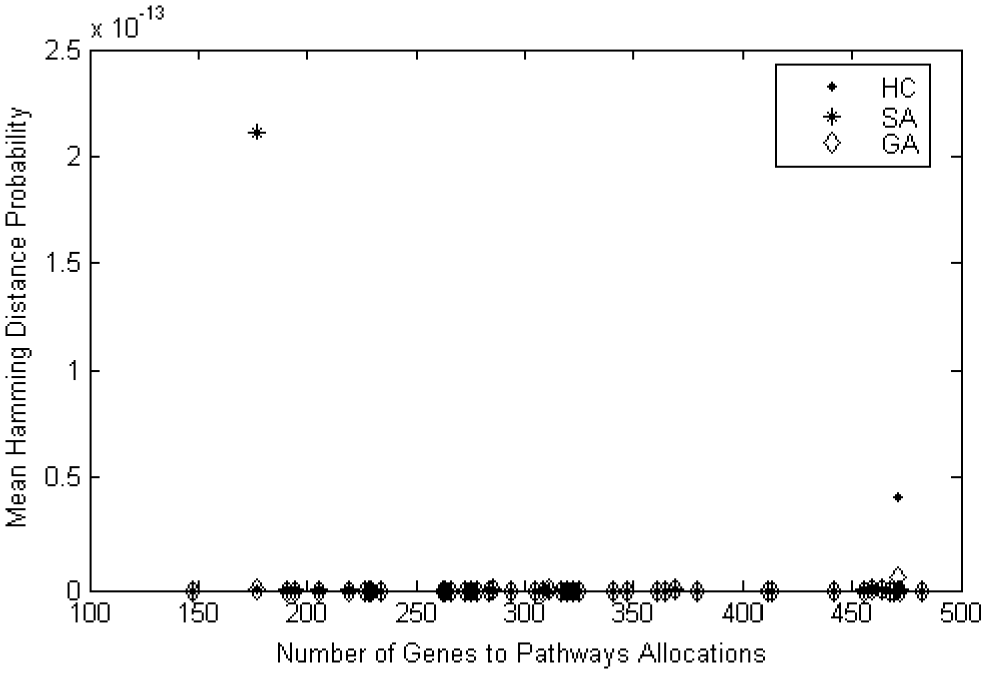


Figure 5.8 Mean Hamming distance probability according to search space size. The probability of observing a given hamming distance or smaller, per experiment is ploted against the possible number of gene to pathway allocations. The mean is based on comparisson of the results produced by 20 separate runs of each method.

Table 5.9 FARI statistics between allocations produced by the three search methods. The values are based on 20 separate runs of each script. However, this time we examine the similarity between results produced by each pair of methodologies, namely hill climbing and simulated annealing, hill climbing and the genetic algorithm, and simulated annealing and the genetic algorithm.

	Maximum	Minimum	Mean	Standard Deviation
HC versus SA	1.000	0.900	0.9742	0.012
HC versus GA	1.000	0.923	0.977	0.013
SA versus GA	1.000	0.895	0.975	0.011

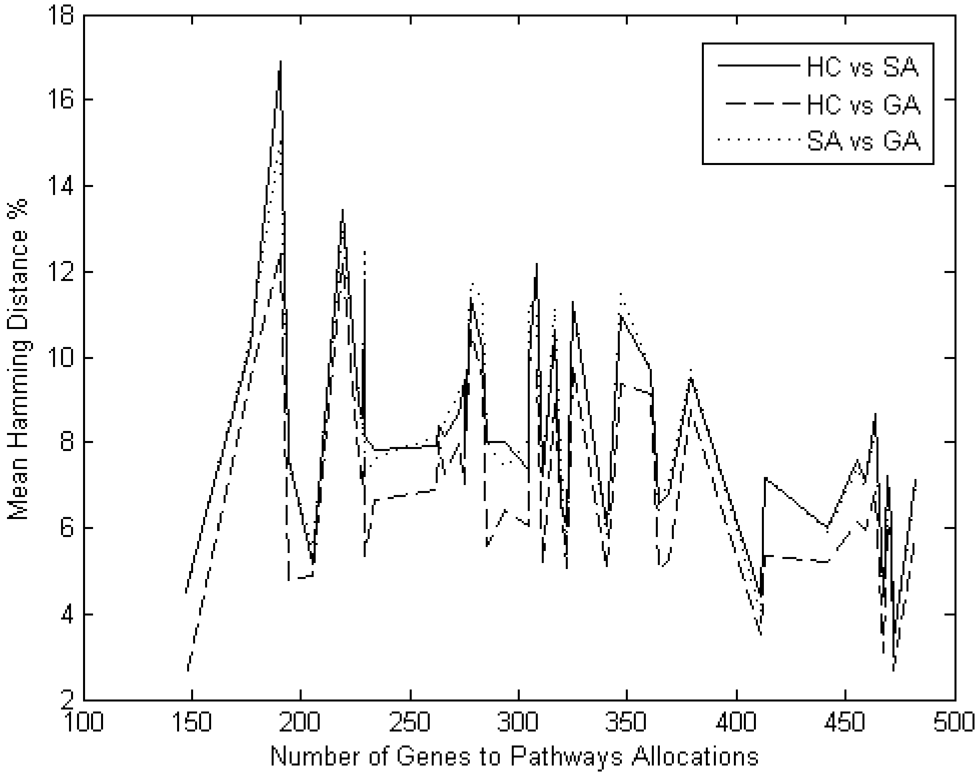


Figure 5.9 Mean Hamming distance between allocations for each pair of methods according to search space size. The probability of observing a given hamming distance or smaller, per experiment for each pair of search algorithms is ploted against the possible number of gene to pathway allocations, i.e. size of search space. The mean is based on comparisson of the results produced by 20 separate runs of each method.

The picture does not change when we examine the hamming distance measure and the acompanying probability of obtaining it by chance. Here, once again we observe that the values are small, with no coreltion to the size of the search space as revealed on Figure 5.9.

The accompanying probability of observing each hamming distance or smaller is similarly low, with a mean of 1.52×10^{-11}, 5.16×10^{-17} and 2.23×10^{-13}, for the hill climbing versus simulated annealing, hill climbing versus the genetic algorithm and simulated annealing versus the genetic algorithm produced gene configurations.

5.4 Conclusions

We have shown that our algorithms can effectively assign multi-membership genes to their constituent pathways, increasing the level of agreement, in terms of the direction of expression per pathway.

As discussed in the results section the allocations produced by separate runs of the search algorithms are highly consistent for all methods, in terms of FARIs and hamming distances. Moreover, the probabilities of obtaining two allocations of a given or smaller hamming distance are extremely low in all cases.

Interestingly, we have observed minimal variation in the performance of the three search approaches, namely the hill climbing, simulated annealing and genetic algorithm. All methods produce highly consistent results and reach roughly equal fitness values, although the simulated annealing approach does seem slightly superior. Furthermore, the consistency of the produced allocations, between methods, in terms of Hamming distance and FARI values does not show any correlation to the size of the search space, as defined by the number of possible genes to pathways allocations in each experiment.

A related issue that may be resolved following this approach is the observed swapping of piles of genes between pathways, by subsequent runs of the search algorithms. As discussed in the results section, in certain cases, allocations exhibiting the same fitness and extremely high FARI values exhibit relatively significant hamming distance. Given the nature of the metrics it appears that groups of genes allocated to different pathways, are still placed together by separate applications of the methods described here. There is room for further investigation in that respect.

Chapter 6: Pathway based microarray analysis centred on enzyme compounds

6.1 Introduction

We formulated the hypothesis that taking into account the multi-membership nature of genes and the collective expression data for the totality of genes involved in biochemical processes may help us improve the analysis of microarray data in the context of pathways (Pavlidis, Payne & Swift 2008). We applied heuristic search to acquire an educated guess regarding the state of individual pathways, by maximising the agreement in terms of differential expression of genes per pathway.

Here we proceed further in an effort to improve and refine this methodology, using the Glycolysis/Gluconeogenesis KEGG pathway as a model and proof of concept. This is one of the most ancient metabolic pathways, present in most organisms, and has been studied in great detail (Romano & Conway 1996).

The discussed analytical approach considers the fact that a number of genes may potentially act not only in distinct pathways as defined by the KEGG database but also in separate chains of events. Hence, we take into consideration the position in the chain of enzymic events where a gene can participate, as in a number of cases different genes encode the same protein responsible for the catalysis of a particular reaction. Additionally, we take into account isoenzymes, that is, distinct enzymes encoded by different genes capable of catalysing the same reaction.

Finally, we consider that pathways consist of smaller sub-networks or modules, which represent chains of events, leading to gradual alteration of a substrate into a desired product. In principal, we expect genes forming a module to agree in terms of expression, showing consistent up- or down-regulation in case of activation or repression of the module, respectively. We facilitate these observations to further develop the methodology discussed in chapters 4 and 5, in the effort to assist identification of the correct chain of events taking place in a pathway, allowing the biologist to infer the state of individual pathways.

6.2 Rationale

The rationale behind this work can be exemplified in Figure 6.1, showing a snapshot of the *Saccharomyces cerevisiae* KEGG glycolysis/gluconeogenesis pathway. Rectangles represent positions where genes and their enzyme products are required for the catalysis of a particular step in the chain of enzymic reactions. The number in each rectangle represents the so called enzyme commission (EC) number, a numerical classification of enzymes based on the type of reactions they catalyse (Webb 1992). Importantly, in cases where two or more enzymes catalyse the same reaction they are given the same EC number. Chemical compounds, that is, the substrates and products of these reactions, are represented by the circles on the diagram. Arrows indicate the direction of each reaction.

As it can be inferred from Figure 6.1, some enzymes can only catalyse a reaction towards one direction, such as EC 2.7.1.11 for 6-phosphofructokinase that catalyses the conversion of β-D-fructose-6P into β -D-fructose-1,6P$_2$. In contrast, many enzymes are able to catalyse a reaction in both directions, such as the case of EC 4.1.2.13. This commission number corresponds to fructose-bisphosphate aldolase, an enzyme capable of catalysing the interconversions between β -D-fructose-1,6P$_2$ and Glyceraldehyde-3P.

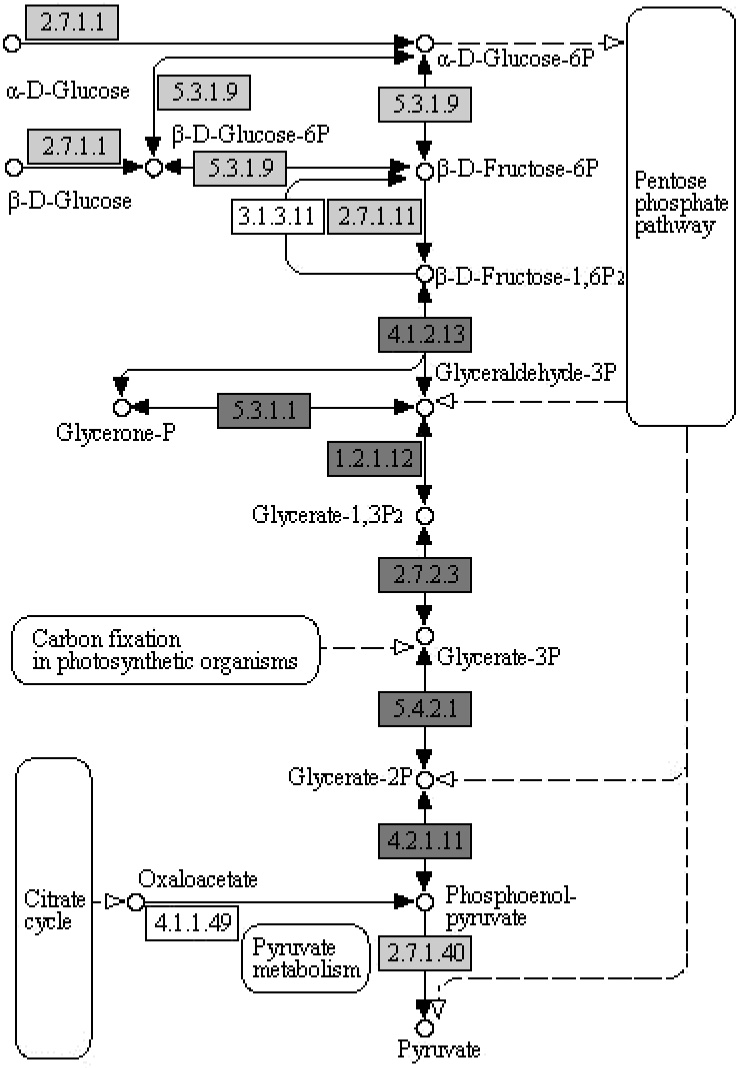


Figure 6.1 The KEGG Glycolysis and Gluconeogenesis modules and their interconnectivity (http://www.genome.jp/kegg-bin/show_pathway?sce00010). Rectangles represent positions where enzymes with the corresponding commission numbers are needed for catalysis. White rectangles represent enzymes unique to gluconeogenesis, light grey rectangles enzymes unique to glycolysis and dark grey rectangles enzymes involved in both modules.

Additionally, certain enzymes can catalyse more than one biochemical reaction, like for example EC 5.3.1.9, responsible for the interconversions between β-D-fructose-1,6P, α-D-Glucose-6P and β-D-Glucose-6P.

Importantly, we should note that there is not always a one to one relationship between genes and an enzyme compounds, corresponding to each rectangle on the diagram. This is exemplified on Table 6.1 which reveals all the experimentally identified enzyme encoding genes, whose products are involved in the catalysis of the respective reaction, for each commission number on Figure 6.1.

Table 6.1 Genes, encoding enzymes, corresponding to each commission number on Figure 6.1

EC number	Gene Symbol
1.2.1.12	TDH1 TDH2 TDH3
2.7.1.1	HXK1 HXK2 GLK1
2.7.1.40	CDK19 PYK2
2.7.2.3	PGK1
3.1.3.11	FBP1
4.1.1.49	PCK1
4.1.2.13	FBA1
4.2.1.11	ENO1 ENO2 ERR1 ERR3
5.3.1.1	TPI1
5.3.1.9	PGI1
5.4.2.1	GPM1 GPM2 YKR043C

Evidently, a number of distinct genes may encode proteins whose enzymatic activity can carry forward the same reaction. Such genes and the corresponding proteins may be active simultaneously or potentially become active or supressed under different environmental conditions and/or during various stages of the cell cycle and development of an organism.

To mention just a few characteristic examples, EC:1.2.1.12 represents genes TDH1 TDH2 and TDH3 which encode isoenzymes that show variable activity in different stages of the cell cycle (Delgado et al. 2001). In another example, EC 2.7.1.40

corresponds to CDK19 and PYK2, both encoding pyruvate kinase which catalyses the conversion of phosphoenolpyruvate to pyruvate (Boles et al. 1997).

Following these observations it becomes evident that the enzyme compounds taking place in the network of biochemical reactions in a pathway are the true indicators of the activity of that path, rather than the genes themselves. In fact, this is common knowledge and one of the main reasons for the widespread interest in studying protein activity directly whenever possible. In chapter 3 there was an extensive discussion of the fact that there are a number of regulatory stages during which a living system controls the activity of the protein arsenal it possesses, besides the process of transcription.

Nevertheless, since microarray technology only allows as to monitor transcription rates and we may often not have the luxury to observe protein function directly, it remains beneficial to find ways of exploiting gene expression data efficiently. Thus, we have opted for basing our search algorithms on enzyme compounds rather than only genes, while still facilitating microarray data in a more indirect manner. Instead of simply maximising the agreement of gene expression per pathway we have modified our methodology to maximise the agreement of enzymes, which in turn is based on the behaviour of the genes responsible for their synthesis.

Hence, we look for agreement between the positions represented by the rectangles on Figure 6.1, which in turn provide evidence for the state of the biochemical reactions, catalysed by the enzymes represented be these rectangles. This is a more biologically sensible choice as in the general effort to detect positive or negative regulation of a pathway we are more interested in establishing an increase or decrease in the rate of the reactions involved than just identifying up and down regulated genes in general. That is, the expression of a unique gene member of a group encoding the same enzyme or isoenzymes is not a clear indication of what is happening at that particular position of the pathway, without considering the expression of the rest of the genes.

To clarify this we can look at a few relevant examples. It has been suggested that yeast may switch between using HXK1, HXK2 and GLK1 glycolytic gene products depending on the carbon source used for growth (Herrero et al. 1995).

Consequentially, over expression of any of the three constitutes plausible indication of positive regulation, while parallel decreased expression of another may simply indicate a switch from using one gene or enzyme to another and does not necessarily imply decrease of the reaction rate. Hence, in a hypothetical instance where HXK1 and HXK2 are up-regulated, while GLK1 down-regulated, the position/enzyme may be activated, despite the state of differential expression of the latter gene.

In another example, EC:2.7.6.1 in the Pentose phosphate pathway corresponds to five homologous genes encoding phosphoribosyl diphosphate (PRPP) synthase. It has been shown that different combinations of the products of these genes, namely PRS1, PRS2, PRS3, PRS4 and PRS5, can result in formation of active PRPP synthase that can catalyse the interconversion between D-ribose 5-phosphate and 5-phospho-d-ribosyl α-1-diphosphate (Hove-Jensen 2004). Thus, a similar rationale can be applied as for the case of EC:2.7.1.1 discussed above.

Importantly, here we use KEGG modules rather than pathways to elucidate the state of the organism from a biochemical point of view. While KEGG includes glycolysis and gluconeogenesis in a single pathway due to the large number of genes shared by both, they are not simply the reverse of each other, but rather constitute two distinct modules. The Glycolysis KEGG module M00001 gradually breaks down glucose to pyruvate, producing energy during the process. In contrast, the gluconeogenesis module M00003 is responsible for the synthesis of glucose from precursors such as pyruvate.

Working with KEGG modules rather than pathways can be seen as zooming in the whole picture to work with shorter more compact chains of enzymic events, responsible for a specific biochemical outcome. In a way a module is a pathway within a pathway where we expect to observe consistent activity between its members which should be reflected on the genes encoding the respective enzymes.

On Figure 6.1, light grey rectangles represent enzymes unique to glycolysis, white rectangles enzymes unique to gluconeogenesis and dark grey rectangles enzymes shared by both modules. While, the cell keeps its regulatory networks functional, in a similar way to an engine when switched on, glycolysis and gluconeogenesis act against each other and are not activated together as this would lead to a futile cycle (Champe, Harvey & Ferrier 2004).

In the methodology discussed here, we account for the observations discussed above. Each gene is treated as a member of the corresponding catalytic position (EC) in a module and pathway, and a member of the module and pathway itself. An EC position appearing more than once is considered as a distinct step of the pathway. In that sense a gene like PGI1 for EC 5.3.1.9 is a multiple membership gene of the glycolysis module itself, capable of catalysing the inter conversion between both alpha-D-Glucose-6P and beta-D-Glucose-6P, and alpha-D-Glucose-6P and beta-D-Fructose-6P. Up-regulation of a gene like FBP1 in EC 3.1.3.11, only present in the gluconeogenesis module in this setting, can only be ascribed to its contribution to that module. In contrast differential expression of TDH1 in EC 1.2.1.12 may be due to its involvement in either the glycolysis or gluconeogenesis module.

6.3 Methods

Given that this work is currently confined to the Glycolysis/Gluconeogenesis KEGG pathway the membership of some genes in other pathways, which may be related to their expression state is not considered. To account for this omission, this work is mainly centred on experiments for which the accompanying literature provides clear analysis of the experimental conditions and the state of the glycolysis and gluconeogenesis pathways. Thus, we know that the conditions, such as for example addition of glucose, have a strong direct effect on this path and the accompanying publications confirm activation or repression of the pathway.

Additionally, a further constrain is applied, on principal facilitating experiments for which statistical analysis by both Eu.Gene (Cavalieri et al. 2007) and the method described by (Swift et al. 2004) and previously implemented in (Pavlidis, Payne & Swift 2008) reveal very significant enrichment of the pathway in hand in differentially expressed genes (*p*<0.01). Given that we work on some time series data where, particularly at the initial stages the cells have not yet responded to the environmental perturbations, we comment on the instances where the above restrictions do not apply.

For the datasets in question, we attempt to maximise the agreement of expression per module rather than the entire pathway. Furthermore, we base this maximisation on EC positions rather than the genes. Thus, a gene is first allocated to a catalytic position, each one corresponding to an EC rectangle for a particular module, as shown on Figure 6.1, which is then examined to infer the state of a pathway. The proportion of EC positions in a path to which at least one gene has been assigned constitutes the pathway coverage.

Given the size of the search space, an exhaustive search is not an adequate approach for large networks such as the entire metabolic network of an organism. Thus, we have once again opted for applying a hill climbing algorithm (Michalewicz & Fogel 2004) that changes the possible multi-membership gene allocation to EC positions and modules. In this way we attempt to elucidate which particular reaction in which module may require activation or repression of the gene in hand. Assigning a gene to an EC position, in a module, suggests that the state of differential expression of that particular gene is due to its involvement in that reaction. On the other hand, not assigning a gene to a certain EC position implies that the observed up- or down-regulation is not a consequence of the contribution of the enzyme product of the gene to the respective reaction.

6.3.1 Algorithm

The following mathematical notation is used within our methods and algorithm. Let N_1 represent the number of unique genes, let N_2 be the EC position identifier (e.g. 4.2.1.0 might be ID 7) where EC position identifiers that appear more than once have their own ID, and N_3 the number of modules.

We define a list of 5-tuples B where each 5-tuple B_i represents gene g_i, enzyme (EC) e_i encoded by the gene, module m_i in which the gene and the corresponding enzyme participate, the expression x_i of the gene and the state of allocation l_i of the gene to the particular enzyme and module. In Equation (6.1), b_i represents an instance of a 5-tuple and x_i has a value of +1, -1 or 0 if gene i is up-, down-regulated or stable respectively, based on a threshold parameter t (Equation (6.2)). l_i=1 if gene g_i is allocated to EC position e_i in module m_i, and l_i=0 otherwise.

$$b_i = (g_i, e_i, m_i, x_i, l_i) \qquad (6.1)$$

$$x_i = \begin{cases} +1, \text{if } G(i) > t \\ -1, \text{if } G(i) < \text{-}t \\ 0\ , \text{otherwise} \end{cases} \qquad (6.2)$$

$$E(i) = \begin{cases} -1, & |D_i| > |U_i| \\ 0, & |U_i| = 0 \wedge |D_i| = 0 \\ +1, & |U_i| > |D_i| \end{cases} \qquad (6.3)$$

$$V(i) = \begin{cases} 0, & \text{otherwise} \\ 0.5, & |U_i| = |D_i| \neq 0 \end{cases} \qquad (6.4)$$

$$U_i = \{b_j : l_j = 1 \wedge e_j = i \wedge x_j = 1\} \qquad (6.5)$$

$$D_i = \{b_j : l_j = 1 \wedge e_j = i \wedge x_j = -1\} \qquad (6.6)$$

Thus, we work with a binary string *A* of size equal to the length of the list of 5-tuples, representing the allocation of genes to EC positions/enzymes and modules. For each position we define a scoring function *E*(*i*) and *V(i)* as shown in equations (6.3) and (6.4), respectively. These functions use equations (6.5) and (6.6), which define the number of up and down regulated genes respectively.

$$M(i) = C(i)\left(\left| \sum_{\forall e \in K(i)} E(e) \right| + \sum_{\forall e \in K(i)} V(e) \right) \qquad (6.7)$$

$$C(i) = \frac{|\{e_j : l_j = 1 \wedge m_j = i\}|}{|\{e_j : m = i\}|} \qquad (6.8)$$

$$F(B) = \sum_{i=1}^{|B|} M(i) \qquad (6.9)$$

Therefore, if there are more up regulated than down regulated genes in a particular EC position, in a path, the position has a score of 1 while whenever the opposite is true the position has a score of -1. In cases where a position is not assigned any genes the respective score is 0.

Algorithm 6.1: Genes To Enzymes Allocation Algorithm

```
1)   Input 1: list of gene IDs coupled with their EC
     commission numbers
2)   Input 2: list of gene IDs coupled with their pathway IDs
3)   Input 3: = a list of EC commission numbers coupled with
     their pathway IDs
4)   Input 4: = Expression vector of $\log_2$ ratios for KEGG
     pathways genes
5)   Remove all genes between $+t$ and $-t$
6)   Allocate single-membership genes to their commission
     numbers and modules creating $A$
7)   Get fitness $F(A)$, set $F\_old = F(A)$
8)   For $p$ = 1:number of iterations
9)      Save gene configuration
10)     Use $A$ to randomly choose a gene ($i$) in EC position e($i$)
        and module m(i)
11)     If according to $A$ gene ($i$) is already present in the EC
        position ($i$) then remove the gene, i.e. set $a(i) = 0$
12)     Else if not present, place it in that EC position in
        the module, i.e. set $a_i = 1$
13)     If the gene is also allocated to the competing module,
        remove it
14)     End if
15)     If upon completion of steps (9) to (14) the gene is not
        assigned to at least one EC position in one module,
        randomly choose a position and assign it
16)     Estimate fitness $F(A)$
17)     If $F(A) > F\_old$ set $F\_old = F(A)$
18)     Else if $F(A) < F\_old$ restore gene configuration (from
        step (7))
19)  End for
20)  Output: $A$
```

Finally, for positions containing equal numbers of up and down regulated genes we assign a score of 0.5 (equation (6.4)). This is a biologically meaningful choice, as in terms of biochemistry a path may switch from facilitating one enzyme to another for the biological system to meet its needs, as exemplified in section 6.2. Thus cases of equally up- and down-regulated EC positions are not in disagreement with biological rationale. From this perspective, the contradicting expression of genes corresponding to the same EC number is not surprising.

We employ the fitness function in equation (6.9) to search for the best allocation of genes to EC positions and modules in order to infer the state of the modules and the overall state of the pathway. The more EC positions of similar expression in a module the higher the fitness we acquire for that module. Equation (6.7) calculates the fitness per module given the filling of positions with genes. Essentially, it subtracts the number of down-regulated positions from the number of up-regulated positions in the module. Then it adds 0.5 to the absolute value of the result for each neutral position, that is, for each position containing equal number of up- and down-regulated genes.

C(*i*) is the module coverage, acquired by equation (6.8). This is equal to the number of positions with allocated genes divided by all enzymes/positions in the module. Naturally, the better the allocation fits the module, with more positions filled with genes, the higher the value of *C*(*i*). Algorithm 6.1 presents the pseudocode for the hill climbing search implementation.

6.4 Results and Discussion

We applied the methodology presented here to a number of distinct microarray experiments on *Saccharomyces cerevisiae*. On principal we run the algorithm 20 times on each dataset mainly basing our discussion on the best allocation accompanied by the highest fitness value, whenever appropriate. We first discuss the performance of the algorithm from a biological perspective and proceed to examine the consistency of the produced results.

6.4.1 Pathway allocations

First, in (Cavalieri et al. 2007), initially discussed in chapter 4, the authors analyse the global expression of yeast during Diauxic shift, where cells inoculated in glucose rich medium turn to aerobic utilisation of ethanol produced during fermentation, upon exhaustion of the available sugar. According to their analysis, the KEGG glycolysis/gluconeogenesis pathway gradually becomes one of the most activated pathways in the experiment, in regard to all *Saccharomyces cerevisiae* KEGG pathways. In particular, following their statistical approach, the authors assign it a *p*-value of 0.01 at time point 7. This is the probability of observing as many or more differentially expressed genes in the pathway by chance, given the size of the pathway and the overall number of expressed genes in the entire dataset.

In this approach it is not directly clear what this activation actually means, increased synthesis or utilisation of glucose. The application of our method to the data from time point 7 reveals that the gluconeogenesis module is activated while glycolysis is actually repressed. This is in agreement with the more detailed analysis of the data, in the first publication accompanying the dataset (DeRisi, Iyer & Brown 1997). The authors identify repression of the glycolytic process and rechanneling of pyruvate through the gluconeogenesis path. Our method adequately identifies as fittest configuration the allocation of genes to the glycolysis module, so that 7 out of 13 EC positions appear repressed and one neutral, containing one up-regulated and one repressed gene. At the same time up-regulated genes are assigned to the gluconeogenesis module, covering 3 out of 8 EC positions in that path (Figure 6.2, t7).

The low coverage in the latter case is not unexpected and in agreement with our rationale. Both processes take place simultaneously, with cells rerouting the flow of metabolites. Thus genes along the reversible steps of the entire pathway, members of both the glycolysis and gluconeogenesis modules, need to be expressed in a way that balances two competing trends, repression in the first case but activation in the later. Consequentially, it is quite likely that the genes shared by both processes may appear repressed as a net effect of their regulation in order to satisfy both modules. It is likely that much less of their protein product is needed as glycolysis switches off, even though gluconeogenesis becomes active. Without applying the search process, the gluconeogenesis path contains 5 down-regulated, as compared to 3 up-regulated genes at time point 7, which could lead to the incorrect assumption that the module is repressed.

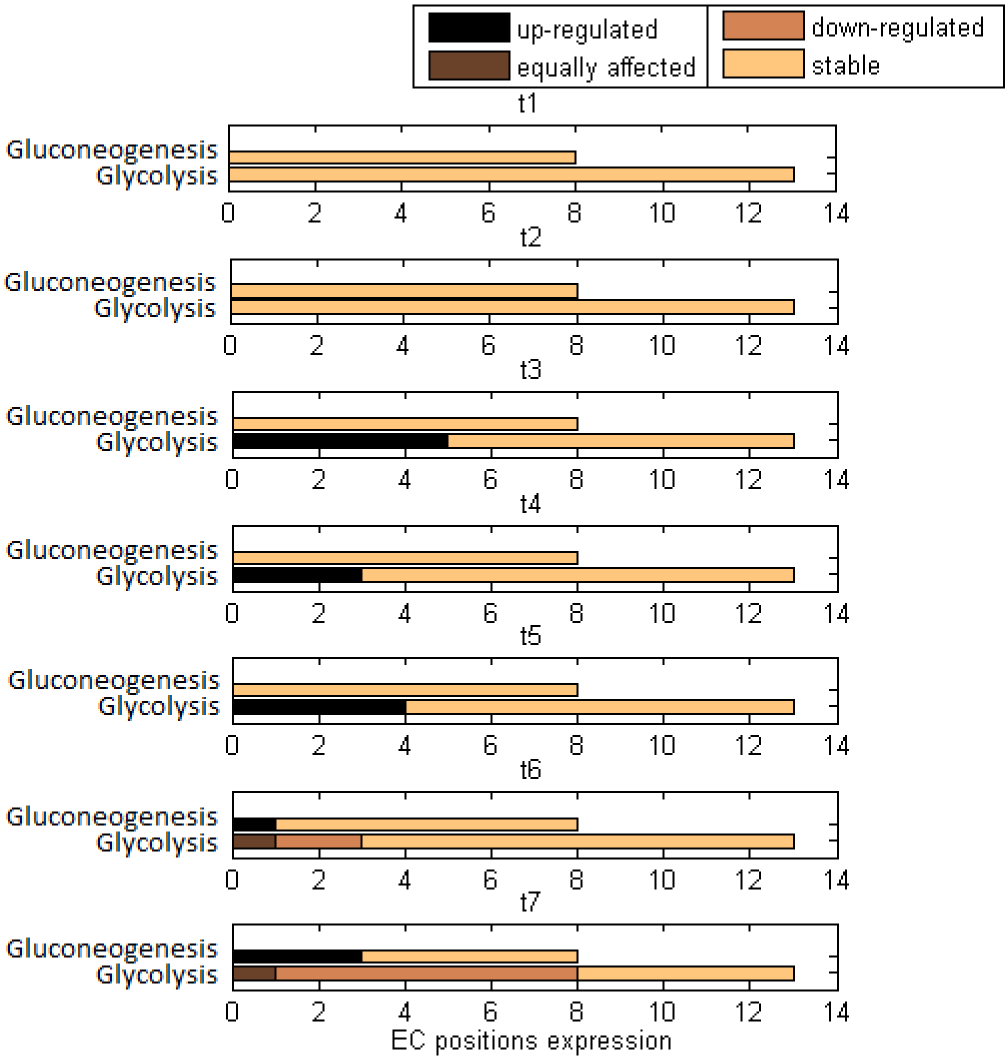


Figure 6.2 Glycolysis and Gluconeogenesis enzymes behaviour. The figure exhibits the activation and deactivation of various enzymic positions in the modules (EC positions) throughout 7 time points, upon processing with the proposed method.

Notably, before the exhaustion of glucose, the picture is quite different. In particular, as noted in (DeRisi, Iyer & Brown 1997), initially there is no substantial change in global gene activity, with most differential expression occurring towards the last two time points. At time points 3 and 4 there appears some glycolytic activity, while glucose is still available, with the balance starting to shift towards glucose synthesis at time point 6. In (Cavalieri et al. 2007) the authors calculate an insignificant *p*-value for the KEGG glycolysis/gluconeogenesis pathways at time points 1 and 2, which becomes significant at time point 3, to increase again afterwards and then regain its significance at the last time point.

We became interested in examining the allocation of genes to EC positions and modules in the time points preceding the exhaustion of available sugar and obtained a sensible result, as shown on the Figure 6.2. In particular, while initially there are no differentially expressed genes in the glycolysis and gluconeogenesis modules, at time point 3 to 5 all filled EC positions contain up-regulated genes, only for the glycolysis module. Then at time point 6 there is an evident shift with the appearance of an up-regulated position in the gluconeogenesis module, while the glycolysis module now exhibits 2 down-regulated EC positions, and one that contains an equal number of up- and down-regulated genes. At time point 7 the switch is complete, with the glycolysis module exhibiting 7 down-regulated positions, while the gluconeogenesis module 3 up-regulated ones.

To proceed to another dataset, in (Ronen & Botstein 2006) the authors describe a series of microarray experiments studying the response of steady-state yeast cultures to transient perturbations in carbon source. In GSM95012 in the analysed dataset, obtained from GEO, cells grown in steady conditions are subjected to a pulse of glucose (0.2 g/l) and microarray analysis performed on RNA extracted 20 minutes after the glucose addition. There are 12 differentially expressed genes in the KEGG Glycolysis/Gluconeogenesis pathway. Upon processing of the data the glycolysis module appears activated with 7 out of 13 positions covered, while the gluconeogenesis module repressed with 4 out of 8 positions covered, as expected given that cells are presented with excess glucose to cover their nutritional needs. If we were to base our analysis on the behaviour of genes only, for each module in isolation, we would observe 5 up-regulated and only 4 down-regulated genes in the gluconeogenesis pathway. Naturally, it is unclear if the module is activated or supressed and if we were to make a sensible guess activation would be more appealing, even though this is clearly not the case here.

The results are similar for GSM94996 from the same dataset, where RNA is extracted 120 minutes after the admission of a glucose pulse (2 g/l). The analysis identifies glycolysis as the activated module, with coverage of 6 out of 13 EC positions. In contrast, gluconeogenesis shows downward trend, given the increase of available glucose in its environment. The module coverage is 3 out of 8 EC positions in the latter case.

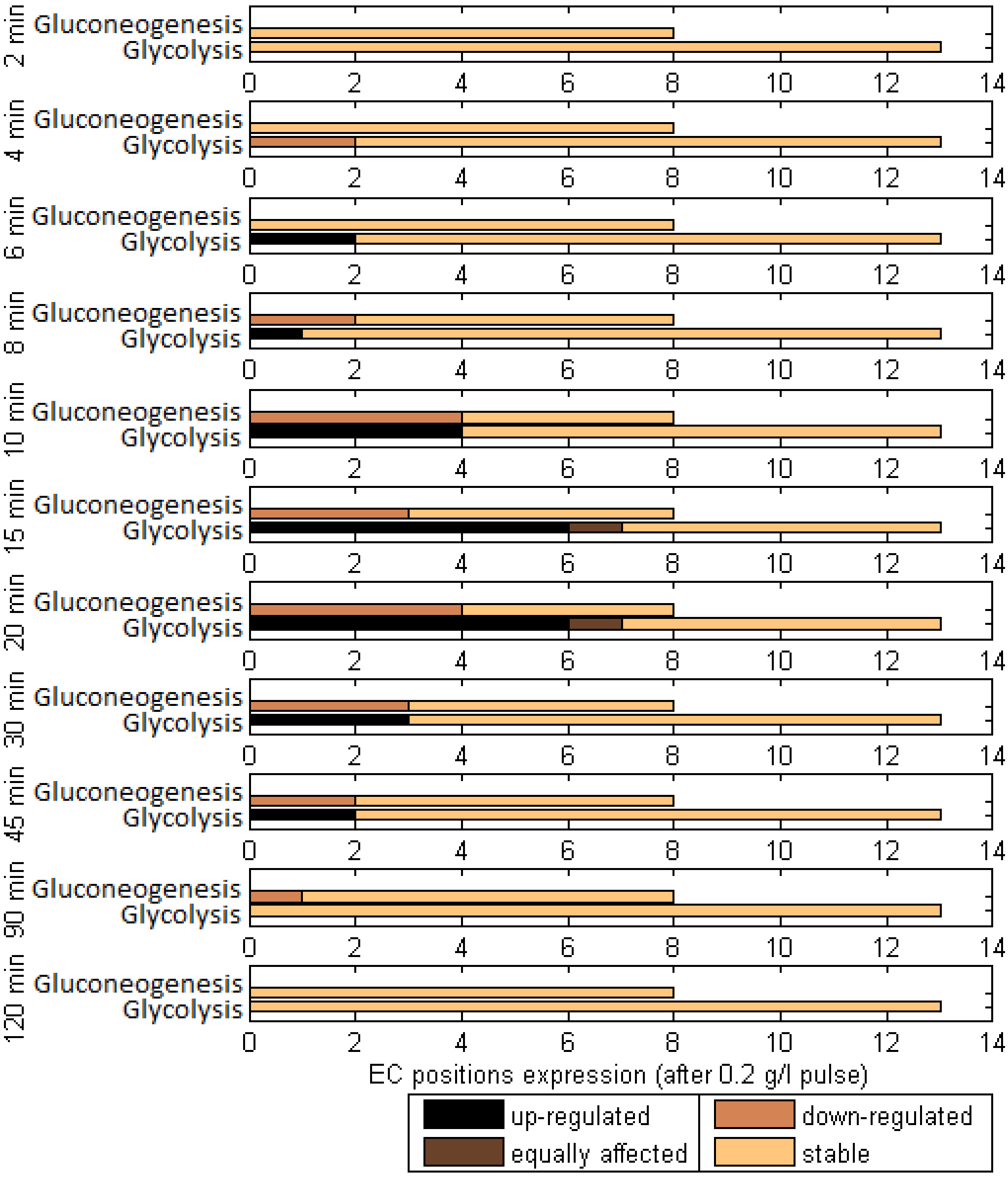


Figure 6.3 Response to excess glucose. Cells submitted to 0.2 g/l glucose pulse gradually exhibit increased expression of glycolytic enzymes, while gluconeogenesis subsides.

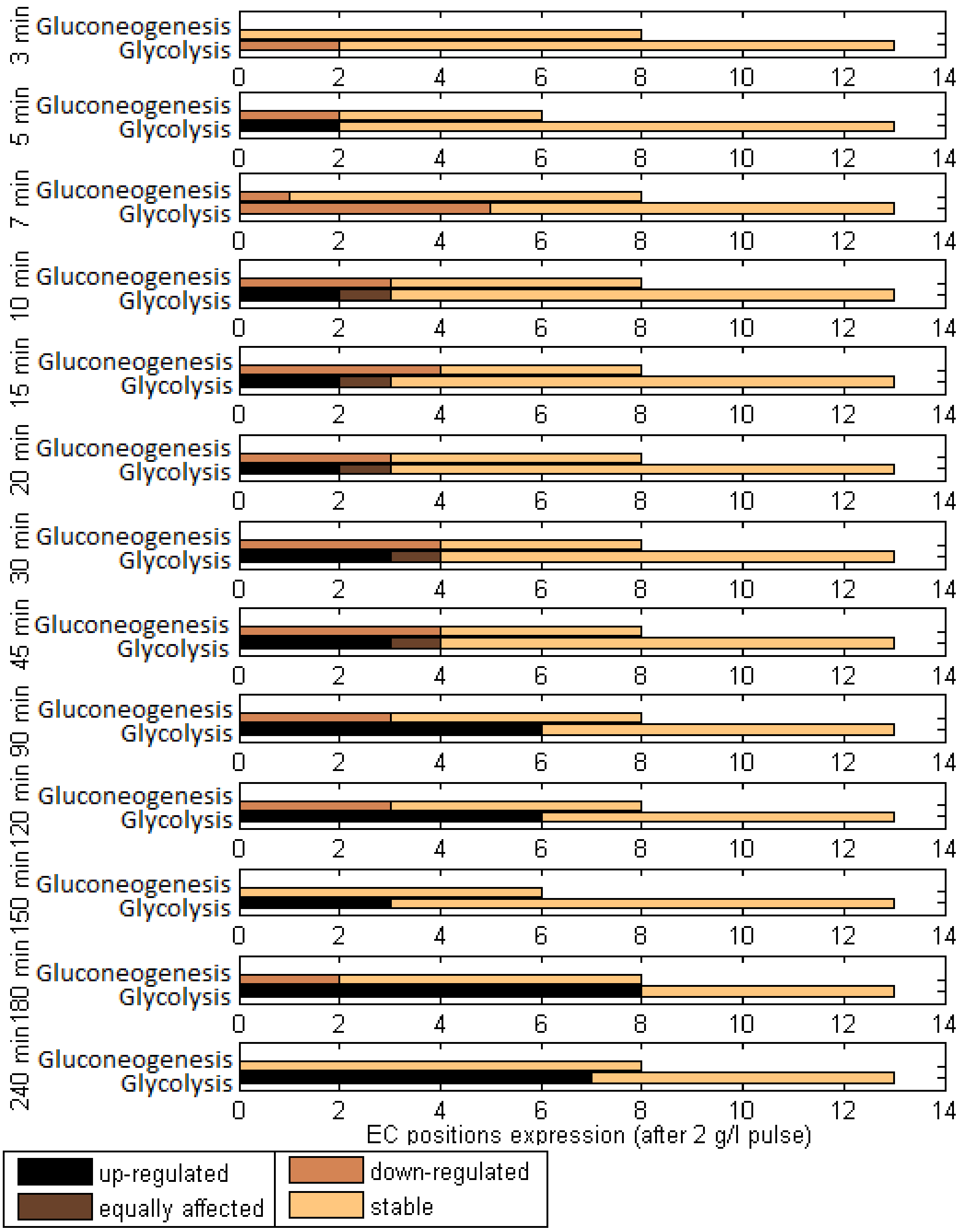


Figure 6.4 Response to excess glucose. Cells submitted to 2 g/l glucose pulse gradually exhibit increased expression of glycolytic enzymes, while gluconeogenesis subsides.

Notably, in this example the number of down-regulated is half the number of up-regulated genes in the gluconeogenesis group, 6 against 3 respectively. Consequentially, examining the expression of these genes in isolation, without considering their participation in the glycolytic process, would again suggest increased synthesis of glucose.

As these are time series experiments we applied the analysis to the entire dataset, even though as expected in some cases, at the initial and final stages, there are no expressed genes or their number is quite small. This includes both time series analysed in the publication, one following a 0.2 g/l glucose pulse and one following 2 g/l glucose pulse. Figure 6.3 exhibits the result of applying the search to consecutive time points after the 0.2 g/l pulse, with evident gradual increase in the number of up-regulated glycolysis and down-regulated gluconeogenesis EC positions, which then gradually subsides.

Not surprisingly, the effect of the 2 g/l pulse is stronger, as exemplified on Figure 6.4. After 10 min the number of expressed glycolysis enzymic positions gradually increases while gluconeogenesis exhibits an apparent downward trend. Here, unlike in the case of the 0.2 g/l pulse, the effect of glucose addition does not show signs of decrease until 240min following the pulse submission. Notably, in the first few time points the picture is somewhat confusing, which is not surprising given that the cells have not had time to respond to the extra glucose in their environment.

In another dataset analysed here, from (Gasch et al. 2000), the authors examine the global expression response of *Saccharomyces cerevisiae* to a number of environmental changes, in time series experiments. They identify that nitrogen depletion has a repressive effect on the cluster of glycolytic genes, throughout 9 consecutive time points (GSM874-882). Following their observations, we subjected the microarray data corresponding to each time point to processing with proposed methodology.

The algorithm was able to correctly assign down-regulated genes to the glycolysis module and up-regulated genes to the gluconeogenesis module, identifying suppression and activation in each case respectively, in all time points. As in the previous analysis, the obtained coverage shows an apparent gradual increase until it starts to subside at the last time point. There is an evident correlation to time (0.82, *p*-value=0.006), as shown on figure 6.5, with only time point 4 showing some divergence from the overall pattern. This however, could be due to the general limitations of microarrays technology which sometimes have been shown to lack accuracy.

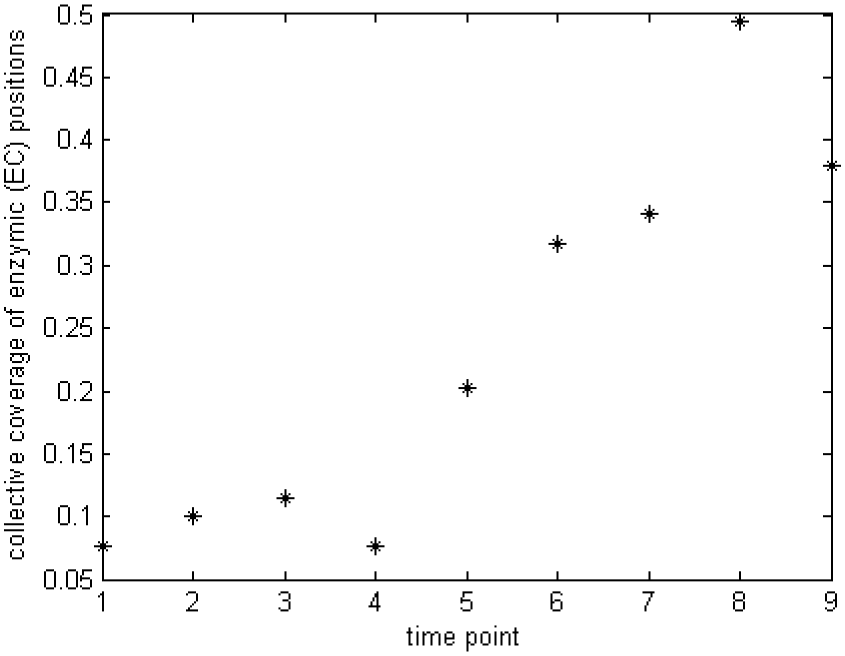


Figure 6.5 Glycolysis/Gluconeogenesis pathway coverage. Coverage represents the proportion of expressed EC position, for consecutive time point experiments discussed in (Gasch et al. 2000).

Additionally, we examined the performance of the algorithm on GSM290980, which deals with the response of yeast cells to glucose deprivation (Bradley et al. 2009). In this case we expected yeast cells to switch on the gluconeogenesis process and at the same time deactivate the glycolysis module. Indeed, the method identified activation of gluconeogenesis, in agreement with biological rationale, with coverage of 4 out of 8 EC positions. At the same time the glycolysis module appears severely repressed with coverage of all 13 EC positions, which appear down-regulated. This is expected given the lack of available glucose for degradation.

Here, there are 9 down-regulated and 5 up-regulated gluconeogenesis genes. Naturally, looking into the module in isolation without taking into account the participation of many of its genes in the glycolytic process would suggest suppression of glucose synthesis. However, the methodology produces a clearer setting allowing us to detect increase of glucose synthesis and decrease of glucose degradation.

For the data discussed in this section we examined the convergence of the algorithm. Figures 6.6 to 6.9 exhibit the convergence for 20 separate runs of the algorithm on the diauxic shift time point 7 data, GSM94996, GSM 95012 and GSM290980 respectively. Naturally, the larger the search space, as defined by the number of expressed multi-membership genes and the possible allocations, the larger the number of iterations required by the algorithm to converge. Processing of GSM290980 with the most expressed genes converges after 104 iterations on average, as opposed to the time point 7 diauxic shift data, converging after 40 iterations on average.

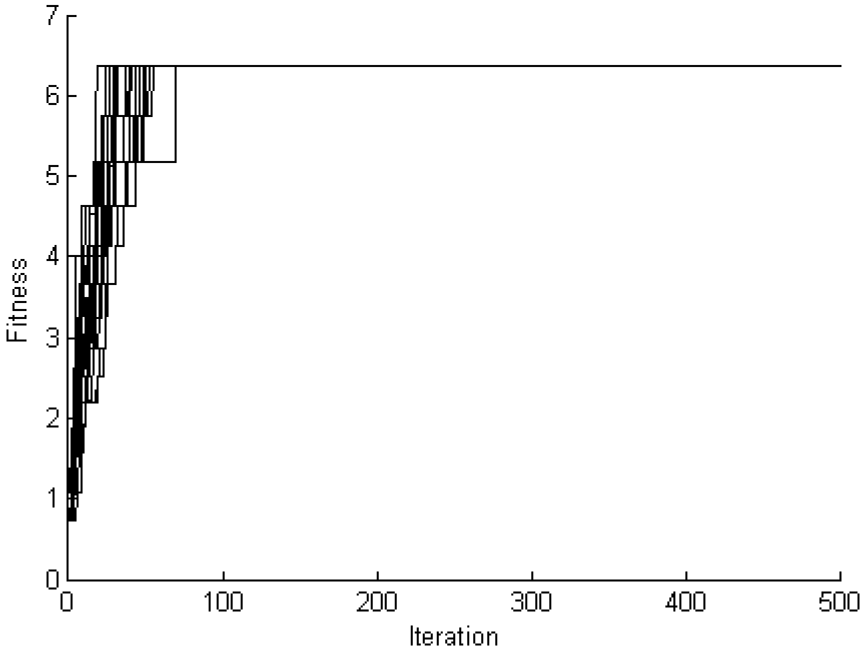


Figure 6.6 Convergence of 20 separate runs, on Diauxic Shift data, time point t7.

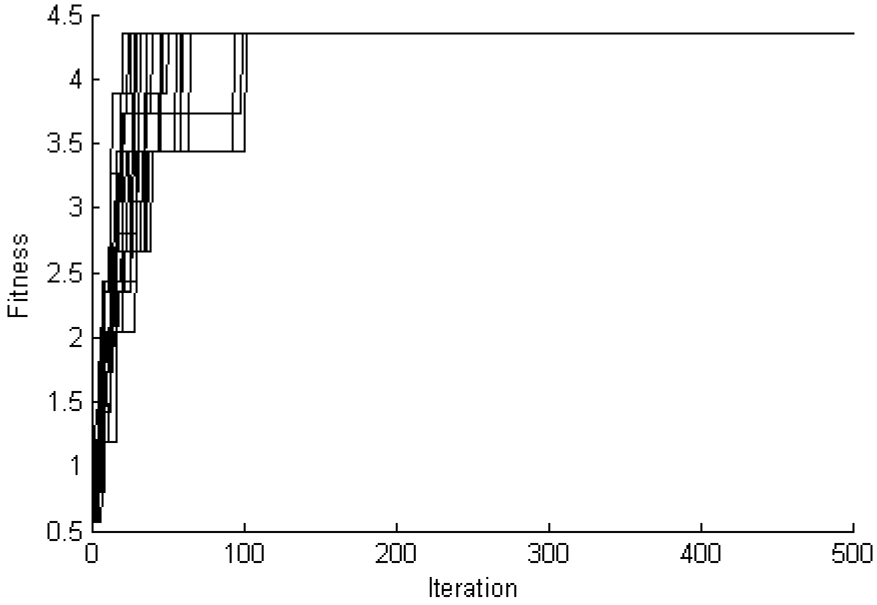


Figure 6.7 Convergence of 20 separate runs, on data from GSM94996.

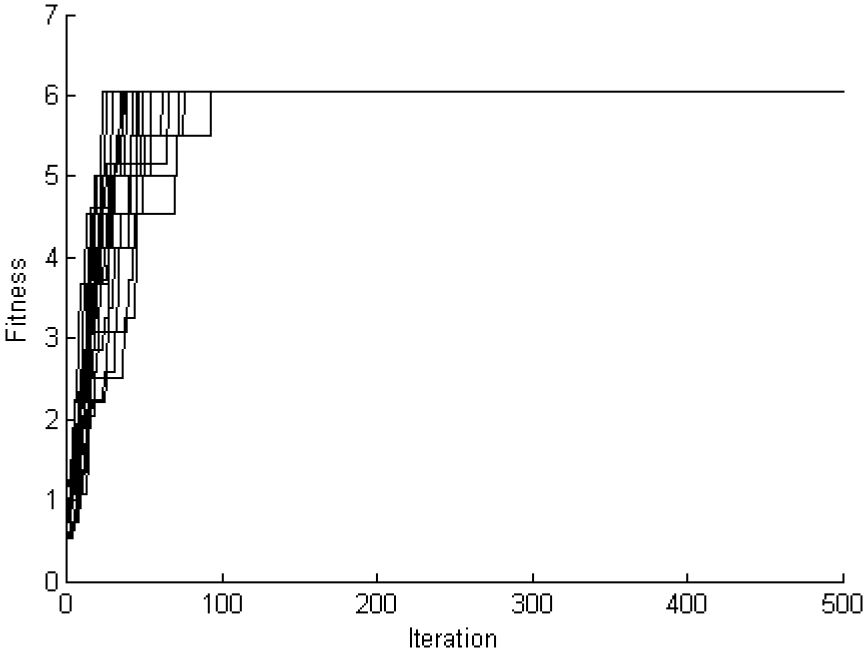


Figure 6.8 Convergence of 20 separate runs, on data from GSM95012.

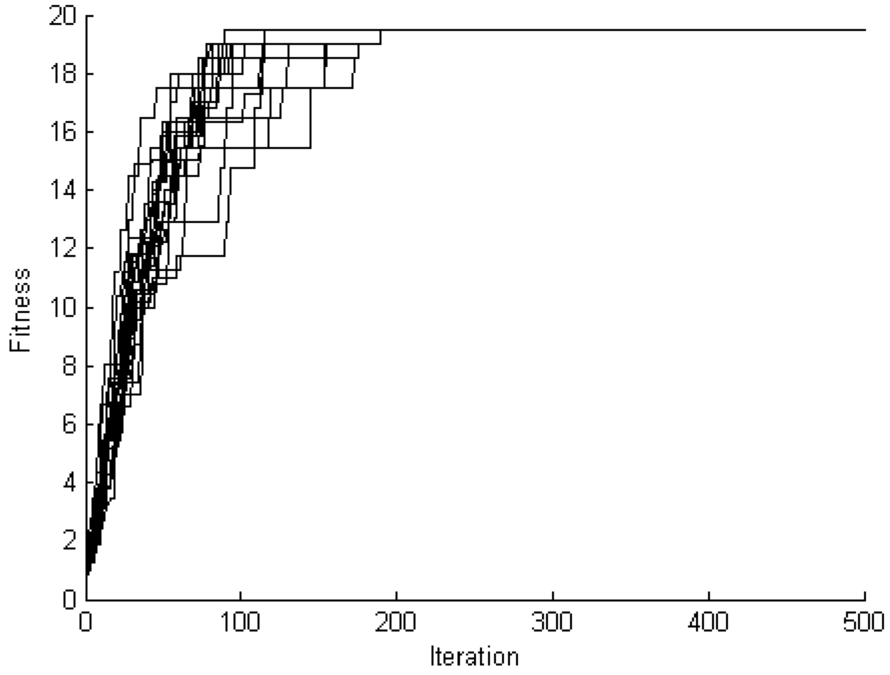


Figure 6.9 Convergence of 20 separate runs, on data from GSM290980.

6.4.2 Consistency of Allocations

Regarding the similarity of the produced allocations based on the hamming distance measure, discussed in chapter 4, section 4.3.3, the result were generally characterised by very high consistency. First for the Diauxic shift time series, the obtained allocations of genes were identical in all cases except from time point 6, where the average similarity was found to be 98.64% with a minimum of 95.12 and a maximum of a 100%, which was reached in 95 of the comparisons. The highest fitness values correspond to allocation of down-regulated genes to the glycolysis module and up-regulated to the gluconeogenesis module, in agreement with biological rationale.

For the dataset corresponding to admission of 0.2 g/l glucose pulse there is a significant proportion of differentially expressed genes in the experiments corresponding to 10, 15, 20, 30, and 45 minutes following the admission. All allocations were 100% identical, with the exception of the one corresponding to 15 minutes. In this case we observed an average similarity of 85.4172, which was revealed to be due to some cases of incorrect assignment of down-regulated genes to the glycolysis module. Nevertheless, the fitness of the latter allocation was significantly lower, half the value of the correct allocation's fitness.

The results exhibit some variability in the case of the 2 g/l glucose pulse in some of the initial stages, with a mean hamming distance of 82.93, 78.49 and 73.17 for 10 15 and 30 minutes respectively. However, in these instances the number of expressed genes is relatively small. The rest of the experiments produced consistent results.

There were no occurrences of variable allocations in the nitrogen depletion data, where the application of the search produces identical results for all experiments regardless of the proportion of differentially expressed genes. Similarly, consistent results were obtained for all runs on GSM290980.

6.5 Conclusions

Interestingly, the result produced by separate runs of the search on the same data where not only in agreement with biological rationale but also often identical for experiments with substantial proportion of expressed KEGG glycolysis and gluconeogenesis genes. While this is encouraging, suggesting that the approach is producing consistent allocations, it remains to be examined how the results may vary when a more complicated setting such as the entire metabolic network is subjected to such analysis.

Overall, the method seems capable of successfully differentiating between activation and repression of the glycolysis and gluconeogenesis modules. Importantly, the search algorithm takes into account the topology of the network, meaning the positions where genes interact through their protein products, and facilitates our knowledge of the competitive nature of the two modules. It considers the multi-membership nature of genes and all the reactions where a gene participates. Rather than simply providing us with a list of expressed genes, it gives as an indication of the state of activity at various steps and in the pathway as a whole, thus providing us with information regarding the direction of the reactions taking place.

Thus, the methodology is improved from a biological point of view. The representation of the setting and solution reflects the reality of pathway behaviour in a more accurate way, than the algorithmic implementation presented in chapter 4.

Here, the analysis is confined to the KEGG glycolysis/gluconeogenesis pathway, consisting of the two competing modules, using rigorous constraints to select microarray data suitable for applying the search. Namely, only experiments with sufficient accompanying information, subjected to statistical testing to be confident that the glycolysis/gluconeogenesis pathway is severely affected by the experimental conditions.

Naturally, the methodology needs to be extended to the entire metabolic network and it would be interesting to examine the results produced when applied to organisms with more sophisticated biochemistry than *Saccharomyces cerevisiae*. Organisms higher in the evolutionary chain provide more complicated networks with larger number of genes and pathways, as well as interconnections. The methodology seems an ideal candidate approach, for the identification of the most likely flow taking place in the metabolic network of a cell.

Interestingly, this can be extended to routes, going through the module, when there are alternative options. Given that we know the routes that reactions can follow within a module we can examine how well an allocation fits each possible route. In the case of the glycolysis and gluconeogenesis modules, a plausible route starts with α-D-glucose while another with β-D-glucose. While this is a simple example, more complicated diversions are present elsewhere in the metabolic network.

More sophisticated search approaches such as simulated annealing are worth exploring in cases of larger search space. This requires detailed preparation of the search setting, meaning identification of the modules constituting each pathway, their interconnectivity and nature, as well as the interconnections between entire pathways themselves. As pathway knowledge increases, through wet lab experimentation, we are able to construct this setting in greater detail.

Chapter 7: Conclusions and future work

7.1 Scope of work

Biology aims to decipher the mystery of living organisms and understand the processes that govern life. Through centuries of research, it has substantially increased our knowledge of living structures and their functionality. It was in the 19th century that Gregor Mendel first suggested the existence of a factor that carries genetic information from a parent to offspring, what we now term gene. It took almost a whole century before evidence emerged that DNA is the carrier molecule of genetic information and it was only about 60 years ago that the scientific community, reluctant at first, was finally convinced that it is indeed DNA rather than proteins, despite their divergence, that stores the data of living systems (Hershey & Chase 1952).

As a natural science biology seeks to achieve its goals through subsequent rounds of hypothesis formulation and experimentation. Traditionally, research has focused on different parts of organisms, including cells, organelles, tissue, and more recently genes and proteins, trying to decipher their role and mode of function. While acquiring the list of all genes and proteins in an organism is essential, as Hiraoki Kitano argues, it is by itself insufficient to understand the complexity of the organism and it function as a dynamic system (Kitano 2002b). He provides a revealing analogy of having an extensive list of all components of an airplane, which naturally by itself can not reveal to us the complexity and workings of the underlying object.

Thus, today with the development of automation and high-throughput experimental techniques, the scope of biological research is shifting as are the questions we are asking. Systems biology has emerged to build upon the immense knowledge we have acquired regarding the components of biological entities, in order to elucidate their collaboration and functional dynamics. This effort requires the contribution of a number of disciplines and clear understanding of biological processes. Computer science plays a central role allowing us to model and examine the dynamics of life processes and manipulate available data to draw meaningful conclusions.

Microarrays are largely dependent on computer science, especially at the analytical stage, where we strive to make sense of the behaviour of genes in different experimental conditions. In more than a decade microarrays have been widely used, but the initial enthusiasm has perhaps not been fully realised. There is an on-going effort to come up with ways to analyse the data in more efficient ways and find useful applications. One such approach is the integration of gene expression data with data produced by other experimental methodologies in a single holistic analytical approach.

Pathway based microarray analysis lies on the crossroad between gene expression data and our understanding of biochemical pathways. It seeks to analyse the behaviour of pre-defined sets of genes, forming biochemical pathways, largely identified by wet-lab biological research, in different experimental conditions. While the idea seemed promising it has struggled to produce clear results. It has been observed that genes forming a pathway often show quite contradictory expression not allowing us to identify the state of individual pathways with sufficient degree of certainty.

The somewhat unclear and erratic behaviour of such closely related genes, in terms of RNA production has been attributed to a number of reasons, discussed in section 3.1.2. In brief summary, genes store the sequence of amino-acids in protein polypeptide chains, but it is proteins that constitute the functional molecules in living organisms. The rate of protein translation, post-translational modifications and different half-life of distinct proteins affect their function, hence, the abundance of RNA corresponding to a gene is not always a good indicator of protein activity (Greenbaum et al. 2003). Additionally, genes in a particular pathway are often characterised by diversity and have distinct functionalities (Stryer & Tymoczko 2006). All these reasons may explain their varying response to changes in experimental conditions, in terms of RNA production.

Realising such obstacles, this work comes to add some contribution to pathway based microarray analysis. It identifies additional issues that increase the observed inconsistency in the expression of genes forming defined pathways, which unlike the obstacles discussed above can be targeted computationally. It further proposes an analytical approach to assist the biologist to identify the state of activity of biochemical pathways, wherever simple observation of gene expression levels is inconclusive.

Importantly, while direct observation of protein abundance is of great value, as it is more directly related to cell function than mRNA messages, it has proven technically more tricky and expensive process than microarray analysis (Stoughton 2005). Hence the widely available and easier to obtain gene expression datasets are likely to remain popular analytical subjects for some time, especially given the emergence of novel, direct methodologies for measuring transcript abundance with much greater accuracy (Morozova , Hirst & Marra 2009). In that sense devising analytical approaches that improve our ability to interpret gene expression remains a beneficial task.

7.2 Contribution

Visualisation is very useful in pathway based microarray analysis. It provides us with insights into the data and sometimes allows us to draw straightforward conclusions. However, we have seen that this is not always the case and to be able to interpret the data in hand we often need to combine visualisation with analytical methods.

Biologists often struggle to adapt to the current approaches to analysing expression data, which rely heavily on mathematics and complicated computational approaches.

They have exhibited a significant level of mistrust, from the early stages of bioinformatics development, to black box computational methodologies (Claverie 1999). Pathway based microarray analysis is often a challenging task, with considerable room left for speculation. This is not surprising, as traditionally, biology has been an experimental science, mostly wet lab based, and it would be greatly beneficial to find ways to provide the researcher with simple clear insight into the data in hand. Motivated by this reality, this work is an effort to gain clearer understanding of the behaviour of genes forming biochemical pathways, in terms of expression, by facilitating a straight-forward heuristic methodology.

7.2.1 Multi-membership genes

A number of analytical approaches applied to large datasets compiled from GEO, to analyse the relative behaviour of single- and multi-membership genes in terms of RNA production, were discussed in chapter 3.

Firstly, the frequency of differential expression of the two groups of genes was explored, revealing a clear tendency of genes that participate in more than one KEGG pathways to be more frequently expressed on average than genes that constitute members of one and only one pathway. Furthermore, an increase in the expression frequency of genes with higher degree of membership was identified. That is, genes that constitute members of three or more KEGG pathways appear differentially expressed more frequently, than genes that are members of two pathways. This pattern persists, showing an increase that follows the minimum degree of membership. A positive correlation between the degree of membership and the average expression of genes belonging to each membership band was evident. These results are in agreement with the underlying biological rationale. Multi-membership genes can be seen as multitask genes, encoding proteins of varying functionality. Consequentially, the organism is more likely to require the functional contribution of a multi-task protein, to adapt to a large number of random environmental conditions.

Additionally, single-membership genes in a pathway were shown to be less likely to exhibit differential expression of opposing direction, in a given experiment, than their multi-membership counterparts. At the same time the former group of genes exhibited higher correlation in terms of expression and pathway contribution.

Finally, through the application of association rule mining, single–membership genes were shown to produce more consistent rules, of smaller number and higher confidence values. In contrast, their multi-membership counterparts exhibited a tendency to produce more rules of lower confidence. Again, this is in agreement with the underlying hypothesis. The more functions a protein has, the more likely the system is to require its contribution at any given time. Thus, one would expect the gene encoding such a protein to appear differentially expressed more often, in a large number of random experiments, than a gene encoding a protein of a single function.

Admittedly, the distinction between multi- and single-membership genes is not very strong in some cases, in the context of this analysis. This is especially so in the case of the expression correlation analysis. To an extent this may be due to the aforementioned limitations of microarrays. However, we should also consider the fact that while the experiments in the datasets are compiled randomly, this does not override biases introduced by the choice of experimental questions by the researchers supplying the data. For example, the gene expression response of cells to addition or removal of popular nutrients from their environment is quite common choice of experimental approach. Hence, identifying frequent differential expression of gene members of the glycolysis pathway does not come as a surprise.

Even more importantly, as revealed in chapter 6, in some cases different genes encode a protein that can carry forward the same reaction. Thus, it is not necessary for such genes to show significant correlation in expression levels, since the organism can use them alternatively in different instances.

Nevertheless, overall, the results provide evidence that supports and strengthens the underlying hypothesis, that the expression of multi-membership genes represents a net effect of their contribution to any combination of their constituent pathways, in a given experiment. The biological system needs to regulate the activity of multi-task genes in a balanced manner to accommodate the needs of all pathways.

7.2.2 Hill climbing based gene to pathway allocation

Chapter 4 presented and explored the performance of an analytical methodology that is used to identify the behaviour of biochemical pathways, as shown in (Pavlidis, Payne & Swift 2008). The methodology is centred on the expression of multi-membership genes and attempts to identify the pathways responsible for any observed up or down-regulation of such genes.

The methodology facilitates a hill climbing search to maximise pathway coverage. It works by allocating as many genes of similar expression as possible to each pathway, while at the same time minimising the number of genes of opposing behaviour in each pathway.

Application of the method to *Escherichia coli* and *Saccharomyces cerevisiae* data showed that it is able to allocate multi-membership genes to their constituent pathways in configurations that make biological sense in accordance with underlying pathway activity. Importantly, the produced allocations exhibit consistency, revealed by simple observations of the results of subsequent runs of the algorithm, as well as through application of the hamming distance measure to examine their similarity. Notably, a number of distinct configurations were explored as starting points for the algorithm.

Analysis of the obtained results suggests that the methodology has a potential interest, especially in cases where gene members of the same pathway exhibit contradicting expression. The innovative characteristic of the proposed method lies within the fact that it considers the multi-membership nature of some genes. To our knowledge this has not been the subject of extensive research. At best, available software tools for pathway based microarray analysis employ various statistical approaches to look for pathways substantially enriched in differentially expressed genes. This is not to say that these approaches are of limited use or significance, rather that it may be beneficial to facilitate the proposed heuristic methodology in a complementary manner.

It should be noted that in some cases genes may be expected to show change in expression that contradicts the up- or down-regulated state of a pathway. As discussed later on, to a degree, this depends on what we define as a pathway, that is, the level of detail we zoom into to study a chain of reactions. Still, the method seems capable of producing satisfactory results, indicating an interesting direction for future work.

7.2.3 Heuristic search approaches comparison

Chapter 5 built on the methodology presented in chapter 4 discussing the application of simulated annealing and a genetic algorithm to search for the assignment of multi-membership genes to their constituent pathways, in a way that maximises the number of genes of similar expression per pathway.

Furthermore the methodologies were applied to a larger number of microarray experiments, each one run a number of times. The consistency of the produced allocations was examined, not only for the same method but also comparing the results of the different search implementations. Besides the obtained fitness which has no straightforward biological meaning, other measures of similarity were facilitated. Namely, an implementation of the fuzzy adjusted rand indexes, of the hamming distance measure and a methodology of estimating the probability of obtaining two allocations of a given hamming distance or smaller, purely by chance. In the context of all similarity metrics, all methods produced consistent results which were, interestingly, more or less similar.

The hill climbing, the simulated annealing and genetic algorithm were shown to reach roughly equal fitness values. The simulated annealing approach seemed capable of reaching only slightly higher fitness values without changing the overall pathway behaviour picture. Furthermore, no correlation was observed between the size of the search space, as defined by the number of possible genes to pathways allocations, and the consistency of the produced results, in terms of Hamming distance and FARI values.

However, in some instances there was a slight variation in the observed hamming distance, with FARI values remaining extremely high. This seems to reflect the swapping of some piles of genes between pathways, by subsequent runs of the algorithms. That is, a group of genes may be allocated to a different pathway, while still remaining together. This issue leaves room for further investigation.

7.2.4 Pathway analysis centred on enzyme compounds

Chapter 6 presented an effort to refine and improve the analytical approach, as presented in (Pavlidis, Swift & Payne 2010). The decision was mainly biologically driven, based on the fact that a number of genes may participate in the same step of a biochemical pathway, that is, the same reaction. Such are the cases of isoenzymes, which, while encoded by different genes, are responsible for the catalysis of the same reaction. At the same time a particular gene, through its protein product may appear at more than one distinct steps of the same pathway.

Given that it is proteins that fulfil enzymic functions rather than genes themselves, and that we are looking for evidence that a reaction takes place, regardless of the particular gene encoding the enzyme that catalyses it, it is a sensible choice to centre the analysis of pathway activity on enzymic positions rather than genes. In that sense the up-regulation of any of the genes involved in a particular step of a pathway serves as indication that the reaction takes place.

Furthermore, this analytical approach zooms in the biochemical network, working with KEGG modules, smaller sub-networks in the chain of enzymic reactions in KEGG pathways. While, when working on a pathway a certain route may become inactive, explaining possible disagreement of gene expression, here we are more confident that the activity of a module should be on principal reflected upon the expression of the genes involved.

Application of the method to the KEGG glycolysis and gluconeogenesis pathway and modules and showed that it can successfully differentiate between activation and repression of the glycolysis and gluconeogenesis modules in a number of unrelated datasets. Instead of working with lists of genes, the methodology gives us an indication of the state of activity at distinct steps of the pathway, providing us with information regarding the direction of the reactions taking place.

The initial analysis is confined to the KEGG glycolysis/gluconeogenesis pathway of *Saccharomyces cerevisiae*, and the two modules it contains. Thus, while it is encouraging that the produced allocations were in most cases identical or very similar, the performance of the method needs to be examined when applied to the entire metabolic network of an organism.

7.3 Future work

There is clearly a lot of room for future work on the ideas presented in here. Some of the possible directions have already become apparent in the previous sections of the chapter. First, the multi-membership nature of genes appears more complicated than initially conceived. As revealed in chapter 6, a gene may be a member of a particular pathway but participate in more than one different steps of the path. From that perspective, the gene becomes a multi-membership gene of the pathway itself. Additionally, distinct genes may correspond to the same enzymic position in a pathway. Hence any one of these genes may be producing an active protein, while the activity of the others is not required for the pathway in question. To an extent, the approach presented in chapter 6 accommodates for these observations, but there is room for further thought and analysis.

On another issue, *Saccharomyces cerevisiae* and *Escherichia coli***,** have relatively simple biochemical networks and it is worth applying the methodology to other organisms higher in the evolutionary chain. Overall, the methodology requires detailed knowledge of the biochemical network of an organism, the modules of which it consists, the functions of their genes and their interconnectivity. This knowledge is far from complete and remains an issue of intensive research, constantly updated and refined. As our understanding of biochemical networks increases, it becomes possible to prepare our search setting in more detail, with greater accuracy. This may allow us to obtain better results, through application of the proposed search methodology.

To a great extent, the limitations of microarray technology confer any analytical efforts, including the method proposed here, less reliable. Any inconsistences in the analysed data are directly reflected on the obtained genes to pathway configurations.

However, very recent advances in transcriptomics tools appear capable of surpassing or at least greatly reducing many of the problems faced by microarray technology. More precisely, the development of RNA-Seq, which uses deep sequencing technologies, is widely considered the next revolutionary step in biological research. This quantitative approach can be used to determine RNA expression with far greater accuracy and less noise than microarrays, which infer transcript abundance only indirectly from hybridization intensity (Wang, Gerstein & Snyder 2009). It has been shown to greatly increase the number of genes, identified to exhibit differential expression ('t Hoen et al. 2008). An informative review of this methodology can be found in (Morozova , Hirst & Marra 2009), as further discussion is beyond the scope of this text. The important point, however, is that it would be of great interest to examine the performance of the methodology proposed in this thesis, using more accurate expression data produced by such advanced techniques.

Finally, the proposed analytical methodology aims to reveal the state of activation of biochemical pathways in microarray experiments. Thus, applying it to do so, on a large number of experiments may in turn produce useful pathway data. This can be then subjected to association rule mining to search for functional relationships between pathways.

In conclusion the methods presented in this thesis suggest some promising directions for future work and can contribute to pathway based microarray analysis and help us elucidate individual pathway states, based on a collective view of gene expression and enzyme activity in a cell or tissue. Progress in the area of transcriptomics and our knowledge of biochemical networks can facilitate further improvement of the analytical framework and performance of the proposed methodology.

References

Agrawal , R 1993, 'Mining association rules between sets of items in large databases', *Proceedings of the ACM SIGMOD international conference on Management of data*, ACM.

Altman , RB & Raychaudhuri , S 2001, 'Whole-genome expression analysis: challenges beyond clustering', *Curr Opin Struct Biol.* , vol 11, p. 340:347.

Altschul, SF, Gish, W, Miller, W, Myers, EW & Lipman, DJ 1990, 'Basic local alignment search tool', *Journal of molecular biology*, vol 215, pp. 403-410.

Ananko , EA, Podkolodny , NL, Stepanenko , IL, Ignatieva , EV, Podkolodnaya , OA & Kolchanov , NA 2002, 'GeneNet: a database on structure and functional organisation of gene networks', *Nucleic Acids Res.*, vol 30, no. 1, pp. 398-401.

Barrett , T, Troup , DB, Wilhite , SE, Ledoux , P, Rudnev , D, Evangelista , C, Kim , IF, Soboleva , A, Tomashevsky , M, Marshall , KA, Phillippy , KH, Sherman , PM, Muertter , RN & Edgar , R 2009, 'NCBI GEO: archive for high-throughput functional genomic data', *Nucleic Acids Res.*, vol 37, no. suppl 1, pp. D1-D15.

BISTI (2011), *Biomedical Information Science and Technology Initiative*, viewed 20 February 2011, < http://www.bisti.nih.gov/>.

Blattner , FR, Plunkett , G, Bloch , CA, Perna , NT, Burland , V, Riley , M, Collado-Vides , J, Glasner , JD, Rode , CK, Mayhew , GF, Gregor , J, Davis , NW, Kirkpatrick , HA, Goeden , MA, Rose , DJ, Mau , B & Shao , Y 1997, 'The Complete Genome Sequence of Escherichia coli K-12', *Science* , vol 277, no. 5331, pp. 1453-1462.

Boles, E, Schulte, F, Miosga, T, Freidel, K, Schulter, E, Zimmermann, FK, Hollenberg, CP & Heinisch, JJ 1997, 'Characterization of a glucose-repressed pyruvate kinase (Pyk2p) in Saccharomyces cerevisiae that is catalytically insensitive to fructose-1,6-bisphosphate', *J Bacteriol.* , vol 179, no. 9, pp. 2987-2993.

Bourguignon, PY, Samal, A, Képès, F, Jost, J & Martin, OC 2010, 'Challenges in experimental data integration within genome-scale metabolic models', *Algorithms for Molecular Biology*, vol 5, pp. 1-4.

Bradley , PH, Brauer , MJ, Rabinowitz , JD & Troyanskaya , OG 2009, 'Coordinated concentration changes of transcripts and metabolites in Saccharomyces cerevisiae', *PLoS Computational Biology*, vol 5, no. 1, p. e1000270.

Brazhnik, P, de la Fuente, A & Mendes, P 2002, 'Gene networks: how to put the function in genomics', *Trends in Biotechnology*, vol 20, no. 11, pp. 467-472.

Brazma, A, Hingamp, P, Quackenbush, J, Sherlock, G, Spellman, P, Stoeckert, C, Aach, J, Ansorge, W, Ball, CA, Causton, HC, Gaasterland, T, Glenisson, P, Holstege, FC, Kim, IF, Markowitz, V, Matese, JC, Parkinson, H, Robinson, A, Sarkans, U, Schulze-Kremer, S, et al. 2001, 'Minimum information about a microarray experiment (MIAME)—toward standards for microarray data', *Nature genetics*, vol 29, no. 4, pp. 365-371.

Brouwer , RK 2009, 'Extending the rand, adjusted rand and jaccard indices to fuzzy partitions', *Journal of Intelligent Information Systems*, vol 32, no. 3, pp. 213-235.

Butcher, EC, Berg, EL & Kunkel, EJ 2004, 'Systems biology in drug discovery', *Nature biotechnology*, vol 22, pp. 1253 - 1259.

Callin, GA & Croce, CM 2006, 'MicroRNA signatures in human cancers', *Nature Reviews Cancer*, vol 6, pp. 857-866.

Campbell, AN & Reece, BJ 2007, *Biology*, 7th edn, Pearson Education.

Carter, SL, Eklund, AC, Kohane, IS, Harris, LN & Szallasi, ZA 2006, 'A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers', *Nature Genetics*, vol 38, no. 9, pp. 1043-1048.

Caspi, R, Altman, T, Dale, JM, Dreher, K, Fulcher, CA, Gilham, F, Kaipa, P, Karthikeyan, AS, Kothari, A, Popescu, L, Pujar, A, Shearer, AG, Zhang, P & Karp, PD 2010, 'The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases', *Nucleic Acids Res.*, vol 38, no. suppl 1, p. D473–D479.

Cavalieri, D, Castagnini, C, Toti, S, Maciag, K, Kelder, T, Gambineri, L, Angioli, S & Dolara, P 2007, 'Eu.Gene Analyzer a tool for integrating gene expression data with pathway databases', *Bioinformatics* , vol 23, no. 19, pp. 2631-2632.

Champe, PC, Harvey, RA & Ferrier, DR 2004, *Lippincotts Illustrated Reviews Biochemistry*, 3rd edn, Lippincott reverend & adventurer.

Chen, YL & Weng, CH 2008, 'Mining association rules from imprecise ordinal data', *Fuzzy sets and systems* , vol 159, no. 4, pp. 460-474.

Chu , W, Ghahramani , Z, Falciani , F & Wild , DL 2005, 'Biomarker discovery in microarray gene expression data with Gaussian processes', *Bioinformatics*, vol 21, no. 16, p. 3385–3393.

Claverie, J 1999, 'Computational methods for the identification of differential and coordinated gene expression', *Human Molecular Genetics*, vol 8, no. 10, p. 1821:1832.

Claverie, JM 2000, 'From Bioinformatics to Computational Biology', *Genome research*, vol 10, no. 1277-1279.

Cochrane , GR & Galperin , MY 2010, 'The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources', *Nucleic Acids Res.*, vol 38, no. suppl 1, pp. D1-D4.

Creighton, C & Hanash, S 2003, 'Mining gene expression databases for association rules', *Bioinformatics* , vol 19, no. 1, pp. 79-86.

Croft, D, O'Kelly, G, Wu, G, Haw, R, Gillespie, M, Matthews, L, Caudy, M, Garapati, P, Gopinath, G, Jassal, B, Jupe, S, Kalatskaya, I, Mahajan, S, May, B, Ndegwa, N, Schmidt, E, Shamovsky, V, Yung, C, Birney, E, Hermjakob, H, et al. 2011, 'Reactome: a database of reactions, pathways and biological processes', *Nucleic Acids Res*, vol 39, no. Database issue, pp. D691-D697.

Dahlquist , KD, Salomonis , N, Vranizan , K, Lawlor , SC & Conklin , BR 2002, 'GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways', *Nature genetics*, vol 31, no. 1, pp. 19-20.

Dai , H, van’t Veer , L, Lamb , J, He, YD, Mao , M, Fine , BM, Bernards , R, van de Vijver , M, Deutsch , P, Sachs , A, Stoughton , R & Friend, S 2005, 'A cell proliferation signature is a marker of extremely poor outcome in a subpopulation of breast cancer patients', *Cancer Research* , vol 65, no. 10, pp. 405940-405966.

Davarynejad, M, Akbarzadeh-T, M-R & Pariz, N 2007, 'A novel general framework for evolutionary optimization: Adaptive fuzzy fitness granulation', *Proceedings of the 2007 IEEE International Conference*.

Debouck , C & Goodfellow , PN 1999, 'DNA microarrays in drug discovery and development', *Nature Genetics*, vol 21, no. suppl 1, pp. 48-50.

Delgado , ML, O’Connor, JE, Azorín, I, Renau-Piqueras, J, Gil, ML & Gozalbo, D 2001, 'The glyceraldehyde-3-phosphate dehydrogenase polypeptides encoded by the Saccharomyces cerevisiae TDH1, TDH2 and TDH3 genes are also cell wall proteins', *Microbiology*, vol 147, no. 2, pp. 411-417.

DeRisi, JL, Iyer , VR & Brown , PO 1997, 'Exploring the metabolic and genetic control of gene expression on a genomic scale', *Science* , vol 278, no. 5338, p. 680–686.

Eisen, MB, Spellman, PT, Brown, PO & Botstein, D 1998, 'Cluster analysis and display of genome-wide expression patterns', *Proc. Natl. Acad. Sci. U.S.A.*, vol 95, no. 25, pp. 14863-14868.

Furey, TS, Cristianini, N, Duffy, N, Bednarski, DW, Schummer, M & Haussler, D 2000, 'Support vector machine classification and validation of cancer tissue samples using microarray expression data', *Bioinformatics*, vol 16, no. 10, pp. 906-914.

Galperin , MY 2005, 'The Molecular Biology Database Collection: 2005 update', *Nucleic Acids Res.*, vol 33, no. suppl 1, pp. D5-D24.

Gasch, AP, Spellman, PT, Kao, CM & Carmel-Harel, O 2000, 'Genomic expression programs in the response of yeast cells to environmental changes', *Mol Biol Cell* , vol 11, no. 12, pp. 4241-4257.

GEO (2011), Gene Expression Omnibus, viewed 20 February 2011, <http://www.ncbi.nlm.nih.gov/geo/>.

Gerhold, DL, Jensen , RV & Gullans, SR 2002, 'Better therapeutics through microarrays.', *Nature Genetics*, vol 32, pp. 547 - 552.

Goesmann , A, Haubrock , M, Meyer , F, Kalinowski , J & Giegerich , R 2002, 'PathFinder: reconstruction and dynamic visualization of metabolic pathways', *Bioinformatics* , vol 18, no. 1, p. 124–9.

Greenbaum , D, Colangelo , C, Williams , K & Gerstein , M 2003, 'Comparing protein abundance and mRNA expression levels on a genomic scale', *Genome Biology* , vol 4, no. 9, p. 117.

Grosu , P, Townsend , JP, Hartl , DL & Cavalieri , D 2002, 'Pathway processor: a tool for integrating whole-genome expression results into metabolic networks', *Genome Res.*, vol 12, no. 7, pp. 1121-1126.

Gygi, SP, Rochon, Y, Franza, BR & Aabersold, R 1999, 'Correlation between Protein and mRNA Abundance in Yeast', *Molecular and Cellular Biology*, vol 19, no. 3, pp. 1720-1730.

Hamming , R 1950, 'Error Detecting and Error Correcting Codes', *Bell System Technical Journal* , vol 26, no. 2, pp. 147-160.

Hand, DJ & Heard, NA 2005, 'Finding Groups in Gene Expression Data', *J Biomed Biotechnol*, vol 2, p. 15:25.

Hardiman, G 2004, 'Microarray platforms – comparisons and contrasts', *Pharmacogenomics*, vol 5, no. 5, pp. 487-502.

Harvey, AR & Ferrier, RD 2010, *Lippincott's illustrated Reviews: Biochemistry*, 5th edn, Lippincott publications.

He, L & Hannon, GJ 2004, 'MicroRNAs: small RNAs with a big role in gene regulation', *Nature Reviews Genetics*, vol 5, pp. 522-531.

Herrero , P, Galíndez , J, Ruiz , N, Martínez-Campa , C & Moreno , F 1995, 'Transcriptional regulation of the Saccharomyces cerevisiae HXK1, HXK2 and GLK1 genes', *Yeast*, vol 11, no. 2, pp. 137-144.

Hershey, AD & Chase, M 1952, 'Independent functions of viral protein and nucleic acid in growth of bacteriophage', *J Gen Physiol.*, vol 36, p. 39–56.

Hipp, J, Guntzer, U & Gholamreza, N 2000, 'Algorithms for association rule mining – a general survey and comparison', *SIGKDD Explorations*, vol 2, no. 1, pp. 58-54.

Hodgkinson, MJ & Webb, PA 2007, 'A system for success: BMC Systems Biology, a new open access journal', *BMC Systems Biology*, vol 1:41.

Hogeweg, P 1978, 'Simulating the growth of cellular forms', *Simulation*, vol 31, pp. 90-96.

Holland , JH 1975, *Adaptation in Natural and Artificial Systems*, Ann Arbor, MI: University of Michigan Press.

Hove-Jensen, B 2004, 'Heterooligomeric phosphoribosyl diphosphate synthase of Saccharomyces cerevisiae: combinatorial expression of the five PRS genes in Escherichia coli', *J Biol Chem.*, vol 279, no. 39, pp. 40345-40350.

Hwang, D, Rust, AG, Ramsey, S, Smith, JJ, Leslie, DM, Weston, AD, Atauri, P, Aitchison, JD, Hood, L, Siegel, AF & Bolouri, H 2005, ' A data integration methodology for systems biology', *Proc. Natl. Acad. Sci. U.S.A.* , vol 102, no. 48, p. 17296–17301.

Ideker, T 2004, 'Systems Biology 101-what you need to know', *Nature biotechnology*, vol 22, pp. 473-475.

Jain, AK, Murty, MN & Flynn, PJ 1999, 'Data Clustering: A Review', *ACM Computing Surveys*, vol 31, no. 3, pp. 264-323.

Kanehisa, M, Araki, M, Goto, S, Hattori, M, Hirakawa, M, Itoh, M, Katayama, T, Kawashima, S, Okuda, S, Tokimatsu, T & Yamanishi, Y 2008, 'KEGG for linking genomes to life and the environment', *Nucleic Acids Res.*, vol 36, no. suppl 1, pp. D480-D484.

Karp, PD, Ouzounis, CA, Moore-Kochlacs, C, Goldovsky, L, Kaipa, P, Ahren, D, Tsoka, S, Darzentas, N, Kunin, V & Lopez-Bigas, N 2005, 'Expansion of the BioCyc collection of pathway/genome databases to 160 genomes', *Nucleic Acids Res.*, vol 33, no. 19, pp. 6083-6089.

KEGG (2011), Kyoto Encyclopaedia of Genes and Genomes, viewed 20 February 2011,<www.genome.jp/kegg/>.

Keller, A, Backes, C, Gerasch, A, Kaufman, M, Kohlbacher, O, Meese, E & Lehnof, HP 2009, 'A novel algorithm for detecting differentially expressed paths based on gene set enrichment analysis', *Bioinformatics* , vol 25, no. 21, pp. 2787-2794.

Kerr, G, Ruskin, HJ, Crane, M & Doolan, P 2008, 'Techniques for clustering gene expression data', *Computers in Biology and Medicine* , vol 38, pp. 283-293.

Khodursky , AB, Peter , BJ, Cozzarelli , NR, Botstein , D, Brown , PO & Yanofsky , C 2000, 'DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptophan metabolism in Escherichia coli', *Proc. Natl. Acad. Sci. U.S.A.* , vol 97, no. 22, p. 12170–12175.

Kirkpatrick , S, Gelatt , CD & Vecchi , MP 1983, 'Optimization by simulated annealing', *Science* , vol 220, no. 4598, pp. 671-680.

Kitano, H 2002a, 'Systems Biology: A Brief Overview', *Science*, vol 295, no. 5560, pp. 1662-1664.

Kitano , H 2002b, 'Computational systems biology', *Nature*, vol 420, pp. 206-210.

Klebanov, L & Yakovlev, A 2007, 'How high is the level of technical noise in microarray data?', *Biol Direct.,* vol 2, no. 9.

Kurhekar , MP, Adak, S, Jhunjhunwala , S & Raghupathy , K 2002, 'Genome-wide pathway analysis and visualization using gene expression data', *Pacific Symposium on Biocomputing*, pp. 462-473.

Likert, R 1932, 'A Technique for the Measurement of Attitudes', *Archives of Psychology,* vol 140, pp. 1-55.

Lipshutz, RJ, Fodor, SP, Gingeras, TR & Lockhart, DJ 1999, 'High density synthetic oligonucleotide arrays', *Nature genetics supplement*, vol 21, no. 1 supplement, pp. 20-24.

MacQueen, J 1967, 'Some methods for classification and analysis of multivariate observations', *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press.

Malumbres, M & Barbacid, M 2007, 'Cell cycle kinases in cancer', *Curr Opin Genet Dev*, vol 17, no. 1, pp. 60-65.

Mann, M & Jensen, ON 2003, 'Proteomic analysis of post-translational modifications', *Nature Biotechnology* , vol 21, no. 3, pp. 255-261.

Maxam , AM & Gilbert , W 1977, ' A new method for sequencing DNA', *Proc. Natl. Acad. Sci. U.S.A.*, vol 74, no. 2, pp. 560-564.

MGED (2001), Microarray Gene Expression Data Society, viewed 20 February 2011,<http://www.mged.org>.

Michalewicz , Z & Fogel , DB 2004, *How to solve it: Modern heuristics,* 2nd edn, Springer.

Moore, JH 2007, 'Bioinformatics', *Journal of Cellular Physiology,* vol 213, no. 2, p. 365–369.

Morozova , O, Hirst, M & Marra , MA 2009, 'Applications of New Sequencing Technologies for Transcriptome Analysis', *Annual Rev Genomics Hum Genet.*, vol 10, pp. 135-151.

Nair, P 2005, 'Epidermal growth factor receptor family and its role in cancer progression', *Current Science,* vol 88, pp. 890-898.

NAR (2011), Nucleic Acids Research online Molecular Biology Database Collection, viewed 20 February 2011, <http://www.oxfordjournals.org/nar/database/a/>.

NCBI (2011), National Centre for Biotechnology Information, viewed 20 February 2011,<http://www.ncbi.nlm.nih.gov/>.

O’Brien, SG, Guilhot, F & Larson, RA 2003, 'Imatinib compared with interferon and low-dose cytarabine for newly diagnosed chronic-phase chronic myeloid leukemia', *New England Journal of Medicine*, vol 348, pp. 994-1004.

Ochs , MF, Peterson , AJ, Kossenkov , A & Bidaut , G 2007, 'Incorporation of gene ontology annotations to enhance microarray data analysis', *Methods Mol Biol.*, vol 377, no. 243-254.

Panteris , E, Swift, S, Payne, A & Liu, X 2007, 'Mining pathway signatures from microarray data and relevant biological knowledge', *Journal of Biomedical Informatics* , vol 40, no. 6, pp. 698-706.

Parkinson, H, Kapushesky, M, Kolesnikov, N, Rustici, G, Shojatalab, M, Abeygunawardena, N, Berube, H, Dylag, M, Emam, I, Farne, A, Holloway, E, Lukk, M, Malone, J, Mani, R, Pilicheva, E, Rayner, TF, Rezwan, F, Sharma, A, Williams, E, Bradley, XZ, et al. 2008, 'ArrayExpress update—from an archive of functional genomics experiments to the atlas of gene expression', *Nucleic Acids Res.*, vol 37, no. suppl1, pp. D868-D872.

Pavlidis, S, Payne, A & Swift, S 2008, 'An Improved Methodology for Pathway Based Microarray Analysis Based on Identification of Individual Pathways Responsible for Gene Regulation', *Proceedings of the annual workshop on Intelligent Data Analysis in bioMedicine And Pharmacology (IDAMAP)*.

Pavlidis, S, Swift, S & Payne, A 2010, 'Pathway based microarray analysis, facilitating enzyme compounds and cascade events', *Proceedings of the annual workshop on Intelligent Data Analysis in bioMedicine And Pharmacology (IDAMAP)*.

Pearson, WR 1990, 'Rapid and sensitive sequence comparison with FASTP and FASTA', *Methods Enzymol*, vol 183, pp. 63-98.

Petersen, KF, Dufour, S, Befroy, D, Garcia, R & Shulman, GI 2004, 'Impaired mitochondrial activity in the insulin-resistant offspring of patients with type 2 diabetes', *N. Engl. J. Med.*, vol 350, p. 664–671.

Quackenbush , J 2002, 'Microarray data normalisation and transformation', *Nature Genetics*, vol 32 supplement , pp. 496-501.

Quackenbush, J 2006, 'Microarray Analysis and Tumor Classification', *N Engl J Med*, vol 354, pp. 2463-2472.

Quadroni , M & James, P 1999, 'Proteomics and automation', *Electrophoresis* , vol 20, pp. 664-677.

Ringnér, M & Peterson, C 2003, 'Microarray-based cancer diagnosis with artificial neural networks', *Biotechniques*, vol Mar, no. suppl 1, pp. 30-35.

Romano, AH & Conway, T 1996, 'Evolution of carbohydrate metabolic pathways', *Res Microbiol*, vol 147, no. 6-7, pp. 448-455.

Ronen, M & Botstein, D 2006, 'Transcriptional response of steady-state yeast cultures to transient perturbations in carbon source', *Proc. Natl. Acad. Sci. U.S.A.* , vol 103, no. 2, pp. 389-394.

Roos, DS 2001, 'Computational biology. Bioinformatics--trying to swim in a sea of data', *Science*, vol 291, no. 5507, pp. 1260-1261.

Russell , S & Norvig, P 2003, *Artificial Intelligence: A Modern Approach*, 2nd edn, Pearson Education.

Russo , G, Zegar , C & Giordano , A 2003, 'Advantages and limitations of microarray technology in human cancer', *Oncogene* , vol 22, pp. 6497-6507.

Saraiya, P, Chris North, C & Duca, K 2005, 'Visualizing biological pathways: requirements analysis, systems evaluation and research agenda', *Information Visualization* , vol 4, p. 191:205.

Šášik, R, Woelk, CH & Corbeil, J 2004, 'Microarray truths and consequences', *Journal of Molecular Endocrinology*, vol 33, pp. 1-9.

Schena, M, Shalon , D, Davis , RW & Brown , PO 1995, 'Quantitative monitoring of gene expression patterns with a complementary DNA microarray', *Science*, vol 270, no. 5235, p. 467–470.

Schulze, A & Downward, J 2001, 'Navigating gene expression using microarrays — a technology review', *Nature Cell Biology*, vol 3, no. 8, pp. E190-E195.

Seo , J & Lee, KJ 2004, 'Post-translational Modifications and Their Biological Functions: Proteomic Analysis and Systematic Approaches', *Journal of Biochemistry and Molecular Biology*, vol 37, no. 1, pp. 35-44.

Shannon , P, Markiel , A, Ozier , O, Baliga , NS, Wang , JT, Ramage , D, Amin , N, Schwikowski , B & Ideker , T 2003, 'Cytoscape: a software environment for integrated models of biomolecular interaction networks', *Genome Research* , vol 13, no. 11, pp. 2498-2504.

Sidhu , AS, Dillon , TS, Chang , E & Chen , JY 2007, 'Ontologies for bioinformatics', *Int J Bioinform Res Appl.*, vol 3, no. 3, pp. 261-267.

Skrabanek, L, Saini, HK, Bader, GD & Enright, AJ 2008, 'Computational prediction of protein-protein interactions.', *Mol Biotechnol*, vol 38, no. 1, pp. 1-17.

Southern , E, Mir, K & Shchepinov , M 1999, 'Molecular interactions on microarrays', *Nature Genetics*, vol 21, pp. 5-9.

Stoughton , R 2005, 'Applications of DNA microarrays in biology', *Annual Review of Biochemistry* , vol 74, pp. 53-82.

Strachan, T & Read, AP 2004, *Human molecular genetics*, 3rd edn, Garland Publishing.

Stryer, L & Tymoczko, JL 2006, *Biochemistry*, 6th edn, WH Freeman.

Subramanian, A, Tamayo, P, Mootha, VK, Mukherjee, S, Ebert, BL, Gillette, MA, Paulovich, A, Pomeroy, SL, Golub, TR, Lander, ES & Mesirov, JP 2005, 'Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles', *Proc. Natl. Acad. Sci. U.S.A.*, vol 102, no. 43, p. 15545–15550.

Swift , S, Tucker , A, Vinciotti , V, Martin , N, Orengo , C, Liu , X & Kellam , P 2004, 'Consensus clustering and functional interpretation of gene-expression data', *Genome Biology*, vol 5, no. 11, pp. R94.1-R94.16.

't Hoen, PA, Ariyurek, Y, Thygesen, HH, Vreugdenhil, E, Vossen, RH, De Menezes, R, Boer, JM, Van Ommen, G-JB & Den Dunnen, JT 2008, 'Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms', *Nucleic Acids Res.*, vol 36, no. 21, p. e141.

Toyoda , T, Mochizuki , Y & Konagaya , A 2003, 'GSCope: a clipped fisheye viewer effective for highly complicated biomolecular network graphs', *Bioinformatics* , vol 19, no. 3, pp. 437-438.

Vinga, S & Almeida, J 2002, 'Alignment-free sequence comparison—a review', *Bioinformatics*, vol 19, no. 4, pp. 513-523.

Wang, Z, Gerstein, M & Snyder, M 2009, 'RNA-Seq: a revolutionary tool for transcriptomics', *Nat Rev Genet.*, vol 10, no. 1, pp. 57-63.

Webb, EC 1992, *Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes.*, International Union of Biochemistry and Molecular Biology, Academic Press, San Diego.

Weiner, N 1948, *Cybernetics or Control and Communication in the Animal and the Machine*, MIT Press.

Werner, T 2008, 'Bioinformatics applications for pathway analysis of microarray data', *Current Opinion in Biotechnology* , vol 19, no. 1, pp. 50-54.

Wu, X, Wang, J, Cui, XQ, Maianu, L, Rhees, B, Rosinski, J, So, WV, Willi, SM, Osier , MV, Hill, HS, Page, GP, Allison, DB, Martin, M & Garvey, WT 2007, 'The effect of insulin on expression of genes and biochemical pathways in human skeletal muscle', *Endocrine*, vol 31, no. 1, pp. 5-17.

Yano, K, Imai, K, Shimizu, A & Hanashita, T 2006, 'A new method for gene discovery in large-scale microarray data', *Nucleic Acids Res.*, vol 34, no. 5, p. 1532–1539.

Zhang, Y 2008, 'Progress and challenges in protein structure prediction', *Curr Opin Struct Biol*, vol 18, no. 3, pp. 342-348.

Zumdahl, SS 2005, *Chemical Principles*, 5th edn, Houghton Mifflin Company.