Computational Analysis of Gene Expression Data

Gráinne Kerr

B.A. (Mod.) Computer Science, M.Sc. Bioinformatics

A Dissertation submitted in fulfilment of the

requirements for the award of

Doctor of Philosophy (Ph.D.)

to the

Dublin City University

Faculty of Engineering and Computing, School of Computing

Supervisors: Prof. Heather J. Ruskin, Dr. Martin Crane

May, 2009

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

_____

Gráinne Kerr

_____

Student ID

_____

Date

CONTENTS

Abstract

Gene expression is central to the function of living cells. While advances in sequencing and expression measurement technology over the past decade has greatly facilitated the further understanding of the genome and its functions, the characterisation of functional groups of genes remains one of the most important problems in modern biology. Technological advancements have resulted in massive information output, with the priority objective shifting to development of data analysis methods. As such, a large number of clustering approaches have been proposed for the analysis of gene expression data obtained from microarray experiments, and consequently, confusion regarding the best approach to take. Common techniques applied are not necessarily the most applicable for the analysis of patterns in microarray data. This confusion is clarified through provision of a framework for the analysis of clustering technique and investigation of how well they apply to gene expression data. To this end, the properties of microarray data itself are examined, followed by an examination of the properties of clustering techniques and how well they apply to gene expression.

Clearly, each technique *will* find patterns even if the structures are not meaningful in a biological context and these structures are not usually the same for different algorithms. Also, these algorithms are inherently biased as properties of clusters reflect built in clustering criteria. From these considerations, it is clear that cluster validation is critical for algorithm development and verification of results, usually based on a manual, lengthy and subjective exploration process. Consequently, it is key to the interpretation of the gene expression data. We carry out a critical analysis of current methods used to evaluate clustering results. Clusters obtained from real and synthetic datasets are compared between algorithms.

To understand the properties of complex gene expression datasets, graphical representations can be used. Intuitively, the data can be represented in terms of a bi-partite graph, with weighted edges between gene-sample node couples corresponding to significant expression measurements of interest. In this research, this method of representation is extensively studied and methods are used, in combination with probabilistic models, to develop new clustering techniques for analysis of gene expression data in this mode of representation. Performance of these techniques can be influenced *both* by the search algorithm, and, by the graph weighting scheme and both merit vigorous investigation. A novel edge-weighting scheme, based on empirical evidence, is presented. The scheme is tested using several benchmark datasets at various levels of granularity, and comparisons are provided with current a popular data analysis method used in the Bioinformatics community. The analysis shows that the new empirical based scheme developed out-performs current edge-weighting methods by accounting for the subtleties in the data through a data-dependent threshold analysis, and selecting 'interesting' gene-sample couples based on relative values.

The graphical theme of gene expression analysis is further developed by construction of a *one-mode* gene expression network which specifically focuses on *local interactions* among genes. Classical network theory is used to identify and examine organisational properties in the resulting graphs. A new algorithm, *GraphCreate*, is presented which finds functional modules in the one-mode graph, i.e. sets of genes which are coherently expressed over subsets of samples, and a scoring scheme developed (using bi-partite graph properties as a basis) to weight these modules. Use of this representation is used to extensively study published gene expression datasets and to identify functional modules of genes with *GraphCreate*. This work is important as it advances research in the area of transcriptome analy-

sis, beyond simply finding groups of coherently expressed genes, by developing a general framework to understand how and when gene sets are interacting.

Acknowledgements

First and foremost I am indebted to my supervisors, Heather Ruskin and Martin Crane, for agreeing to guide me through the complicated landscape of academic research. They have helped me far beyond their duty. They have encouraged me to explore the region of gene expression analysis at all levels and expertly guided me through the development of ideas and methodologies. I would also like to recognize the help and funding support of the National Institute of Cellular Biotechnology in the early part of this work and the School Of Computing for some supplementary funding, and for giving me the opportunity of lecturing many groups at various levels.

Support from members of the Modelling and Scientific group in DCU, both past and present, has been considerable. Special mention to Noreen Quinn and Niall McMahon who early on in my work were always available to answer any questions I had. Special thanks to Dimitri Perrin for helpful, and long, discussions on graph edge-weighting schemes.

I would like to acknowledge past and present residents of 6 The Orchard for always maintaining a happy and joyful place to come home to after a long day of work. Especially, many thanks to Norman Davey for his constant encouragement, enthusiasm and insightful comments. And to Mairead, who I'm sure had better plans for her evening than reading chapters of a thesis.

I am eternally grateful to all my family for each individual pep-talk (I have quite a large family!). They have given me constant support and encouragement to continue with academic studies. In particular, Auntie Mena who provided me with lodgings towards the end of my thesis, and is possibly the most generous person you are likely to meet.

I can think of no-one more deserving to have this work dedicated to them than my parents, Hugh and Bridget. It is certain that this work would not have been completed if it were not for their support - financial and emotional. This thesis is as much yours as it is mine, I'm sorry it has taken so long. I will always endeavor to repay your patience and generosity - thank you.

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Approximately a century and a half ago an Augustinian monk, Gregor Mendel, hypothesized, through the study of *Pisum Sativum* (Pea plants), that there were units of heredity that controlled how traits were passed from one generation to another. Unknowingly he was preparing the core of genetic theory. In the intervening time, this discipline has changed substantially, although the analysis of genes in the nucleus of the cell has remained a fundamental concern in the study of biological organisms. Understanding how, why and when *genes are expressed*[1] is critical to the understanding of the functioning of the cell, and hence of biological organisms. Experimental work suggests that *biological networks* are *modular*, (Barabasi and Oltvai, 2004; Petti and Church, 2005) with modules defined over groups of genes and proteins, as well as other molecules that are involved with a common subcellular process. The underlying idea in *clustering* genes is that genes that are co-regulated will be grouped together, and if co-regulation indicates shared functionality, then clusters defined at gene level represent biological modules. Understanding how, why and when these groups operate is one of the most important questions in modern Biology.

[1] Genes are said to be expressed when the product they code for is realised, see Chapter 2 for more details.

1.1 Motivation for High Level Computational Analysis of Gene Expression Data

What is High Level Computational Analysis of Gene Expression Data?

Analysis of the transcriptome of an organism or group of cells to infer how gene expression affects it's function can take many forms. These include laboratory experiments, which might include gene "knock outs" to analyse the effect on phenotype, or time consuming, gene-by-gene investigations experiments. Computational analysis, involving simultaneous analysis of multiple gene expression can also be carried out. Computational algorithms and techniques include (i) extraction and estimation of expression levels - normally referred to as "low level" analysis, or (ii) identification and linkage of patterns of expression in the data - referred to as "high level" analysis. This categorisation, though crude, is useful and is expanded upon throughout the thesis.

Why is high level analysis of gene expression data important?

Single gene experiments can reveal only a limited amount of information. Genes do not work in isolation, but rather in modules in which the products of a number of genes come together. Furthermore, genes may be expressed coherently under one condition, but diverge under another. Identifying *functional groups of genes* can shed new light on the prognosis of a disease, (identifying for example targets for treatments), can elucidate functions of unknown genes, determine sets of genes involved in regulation of a particular process and so on. For example, in a study of the *Saccharomyces Cerrvisae* genome, hypoxic genes, which are transcriptionally repressed during aerobic growth (through recruitment of the *Ssn6-Tup1* repression complex by the DNA binding protein *Rox1*), where investigated, (Klinkenberg et al., 2005). In an oxygen deprived environment, cells are unable to main-

tain oxygen-dependent *heme* (a complex containing iron, among other elements, to which oxygen binds) biosynthesis and heme accumulates in the cell, which serves as an effector for the transcriptional activator *Hapl*, Fig. 1.1. A *heme-Hapl* complex activates transcription of the *ROX1* gene that encodes the repressor of one set of hypoxic genes. Under hypoxic conditions, heme levels fall, and a *heme-deficient Hapl* complex represses ROX1 expression. As a consequence, the hypoxic genes are derepressed (Klinkenberg et al., 2005). Put quite simply, if we examine the transcriptome of *S. Cerrvisae* during hypoxic and aerobic conditions, and we have prior information that the ROX1 gene encodes a repressor for hypoxic genes, we have the *potential* to evaluate which set of genes are affected by ROX1.

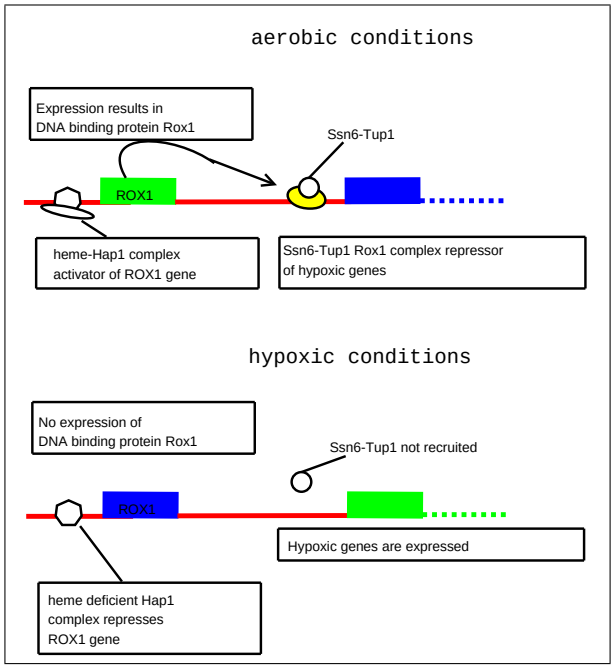


Figure 1.1: Through the repression or activation of ROX1 gene, hypoxic genes can be regulated. Identification of potential functional and/or regulatory groups is one of the aims of cluster analysis. Blue indicates repression, while green indicates activation.

What 'tools' are used for high level analysis of gene expression data?

High level analysis of gene expression data combines methods from Biology, Statistics and Computer Science to derive a picture of what genes are expressed in the nucleus of a living cell. Biological Science poses questions and hypotheses about active processes of the cell and gives us experimental tools to test these hypotheses. Statistical quantification of evidence is standard, not least because new automated tools mean that experimental techniques are rapidly becoming high-throughput in nature, resulting in a vast amount of data. Finally, Computer Science offers tools to organise, analyse and visualise the information generated in addition to the potential to simulate abstractions if biological theory and the theoretical implications. Statistical robustness is crucial, and results of analysis can provide a further basis for biological experiments. This cyclical process, with each disciplinary combination both directly and indirectly reinforcing investigations as a whole is a powerful combination. The focus here is computational although the context of the work is core to understanding the choices made.

1.2 Scope and Contribution

With the explosion of data following the typing of the human genome in 2000, much effort has focused on data generation, rather less on appropriate methods of analysis, with the initial assumption being that standard methods would apply. Systematic assessment of analytical techniques for high throughput technologies, such as microarray data has attracted limited interest, (Kerr and Churchill, 2001; van Bakel and Holstege, 2004; Zakharkin et al., 2005; Pham et al., 2006; Datta and Datta, 2006; Giancarlo et al., 2008). The view that clustering methods are universally applicable is a common mis-conception and recently, has provoked considerable controversy among practitioners, not least in the biological context, (Levsky

and Singer, 2003; Shendure, 2008). The main objectives of this thesis are thus: (i) An assessment of common unsupervised clustering methods and their applicability to gene expression data, (ii) an understanding of the purpose of such analysis, i.e. an extensive examination of the properties of the gene expression dataset, (iii) the development of a robust solution to computational analysis of these data, (iv) testing the solution proposed for diverse data.

1.3 Layout of Thesis

Chapter 2 introduces the main concepts and defines the terminology used throughout the thesis. Many of the definitions introduced in this Chapter can be found in the glossary for ease of reference.

Chapter 3 examines the theoretical state of the art in gene expression clustering. In particular, commonly-applied clustering techniques are examined for their appropriateness for gene expression data, and alternative, notably biclustering and graphical methods, are considered.

Evaluation of clustering results is non-trivial for gene expression data as very little may be known about the data before hand. In ***Chapter*** 4 we thus propose a practical approach for the evaluation of clustering techniques. We assess results obtained with selected clustering algorithms (identified in Chapter 3) for real benchmark and synthetic datasets. We demonstrate that recognition of valid clusters is problematic, and results frequently misleading in the context of gene expression data.

In ***Chapter*** 5, we adopt a framework for graphical modelling to carry out robust and extensive analyses of gene expression data behaviour. This allows us to link the theory of the gene expression data, identified in chapters 2 and 3, with the "realistic"

organisational properties and patterns found in large gene expression datasets.

In ***Chapter*** 6 we develop algorithms which draw on properties, identified in Chapter 5, to extract meaningful groups of genes and samples from the datasets. The ultimate goal of this analysis is to group subset of genes and samples, for which the genes show correlated behaviour, i.e. to extract bi-clusters or functional modules from the graph.

Chapter 7 discusses the challenges in analysing gene expression data and what can be achieved through appropriate computational analysis. Overall results are evaluated, and areas for future work highlight on the basis of conclusions drawn.

Due to the fact that this research deals with large datasets and each analysis is quite detailed and to maintain continuity of text, a lot of information has been enclosed into various appendices. ***Appendix A*** contains details of mathematical formulae. Details of the benchmark and synthetic datasets are given in ***Appendix B***. Additional graphs from various analysis can be found in ***Appendix C***.

CHAPTER 2

BACKGROUND

In this Chapter, we briefly introduce some key concepts concerning gene expression, its role in a biological cell and tools used for its measurement. For a more extensive overview, additional references are cited.

2.1 Introduction

Early observers using microscopes noted that living cells contained a light grey sap encapsulating a darker, denser globule of floating matter. In 1831, the botanist, Robert Brown, used the word *nucleus* to describe this dark, central globule, while the sap is the cytoplasm. Adding stains or dyes to thinly sliced tissue caused *chromatin* material in the nucleus to stand out. To early observers, chromatin appeared to be tiny granules or delicately intertwined threads scattered about inside the nucleus. These long entangled threads are what we now know as *chromosomes*.

Today we know that the chromosome structures, found in the nucleus of a cell, consist of linear *deoxyribonucleic acid (DNA)* polymers, in which the *monomeric subunits* are four chemically distinct *nucleotides* that can be linked together, in any order, into chains up to millions of units in length. Each nucleotide in a DNA polymer is made up of three components: a phosphate, a deoxyribose sugar and one of

four nitrogenous bases: *adenine*, *thymine*, *cytosine* or *guanine*, usually abbreviated to A, T, C and G respectively, (Brown, 2002d).

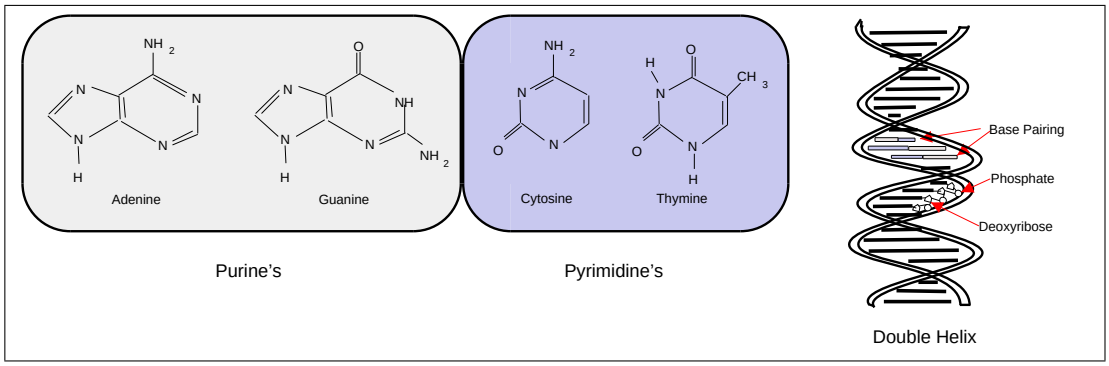


Figure 2.1: Left - Molecular structure of the four nitrogenous bases, which differentiate the DNA polymers and can be categorised as Purine's (double ring) or Pyrimidine's (single ring). Right - "The Double Helix"

DNA polymers assemble together in pairs within the nucleus of a cell to form a double stranded structure known as the *Double Helix*, (Figure 2.1). The double helix has structural flexibility due to *base-pairing* and *base-stacking*. Base-pairing between the two DNA polymer strands involves the formation of hydrogen bonds between an adenine on one strand and a thymine on the other strand, or between a cytosine and a guanine. *Only the A-T and C-G pairs are permissible*, partly because of the geometries of the nucleotide bases and the relative positions of the groups that are able to participate in hydrogen bonds, and partly pairing must be between a purine and a pyrimidine, (resulting in the distinctive helix structure). Base-stacking then involves hydrophobic interactions between adjacent base-pairs, adding stability to the double helix, (Walker and Rapley, 1997).

The limitation that only base pairs A-T and C-G are permissible has significant biological implications. It results in perfect copies of a parent molecule during DNA replication through the simple expedient of using the sequences of the pre-existing

strands to dictate the sequences of the new strands. This is template-dependent DNA synthesis and is the system used by all cellular *DNA polymerases*; (an *enzyme* fundamental to DNA replication). Further details of DNA polymerase function can be found in Brown (2002c) and Hartl and Jones (2002).

Genes are a DNA sub-segment of the genome which contain important biological information. The vast majority of genes *code* for *proteins* and a few for non-coding *Ribonucleic acid (RNA)*[1]. The so-called "expression" of protein-specifying genes involves intermediate *messenger RNA* (mRNA), or coding RNA. This is transported from the nucleus to the cytoplasm of the cell where it directs *synthesis* of the protein coded by the gene. The structure of RNA is similar to that of DNA in that it is also a polynucleotide. However, the sugar in the RNA nucleotide is a ribose sugar and, further, RNA contains the monomeric subunit *uracil* (U) instead of the base thymine found in DNA. Additionally, RNA is found in single strands in the cell, while DNA is usually double-stranded. RNA is *transcribed* from DNA by RNA polymerase, facilitating the biological encoding of DNA to be realised, (Brown (2002g,e) for more information)

Template-dependent RNA synthesis is used by RNA polymerases to make RNA copies of genes: these copies preserve the biological information contained in the sequence of the genomic DNA molecule, meaning that the sequence of nucleotides in a DNA template dictates the sequence of nucleotides in the RNA that is created. During transcription, ribonucleotides are added one after another to the RNA transcript, the identity of each nucleotide being specified by the base pairing rules: A-U, G-C[2]. RNA polymerase is the central component of the *transcription initiation com-*

[1]Non-coding in the sense that they are not translated into protein, but do carry out other essential functions in the cell

[2]Adenines in the DNA template do not specify thymines in the RNA copy as RNA does not contain thymine, instead adenine pairs with uracil in DNA-RNA hybrids

plex. Every time a gene is transcribed, a new complex is assembled immediately upstream of the gene. The initiation complexes are constructed at specific positions on the genome, marked by specific nucleotide sequences called *promoters*, which are only found upstream of the gene, (Brown, 2002b).

Genes are said to be expressed when the product they code for are expressed (see below for details). A large percentage of protein-coding genes are involved in expression, replication and maintenance of the genome. A smaller percentage specifies components of the signal transduction pathway that regulates genome expression and other cellular activities in response to signals received from outside the cell. Other genes code for enzymes, responsible for the general biochemical functions of the cell, while the remainder of the genes are involved in activities such as transport of compounds into and out of cells, the folding of proteins into their correct three dimensional structures, the immune response and synthesis of structural proteins, such as those found in the cytoskeleton and in muscles, (Petsko and Ringe, 2004b).

2.2 The Central Dogma of Protein Synthesis

The genome is a repository of biological information, where utilization of this requires the coordinated activity of enzymes and other proteins. These participate in a complex series of biochemical reactions collectively referred to as *genome expression*.

The initial product of genome expression is the *transcriptome*, a collection of RNA molecules derived from those protein-coding genes, having biological information required by the cell at a particular time. These RNA molecules direct the synthesis of the final product of genome expression, the *proteome*, (the cell's reper-

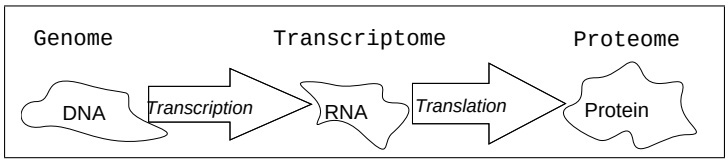


Figure 2.2: Central Dogma of protein synthesis

toire of proteins), which specifies the nature of the biochemical reactions that the cell is able to carry out. *Transcription*, (where individual genes are copied into RNA molecules, see above) constructs the transcriptome. Construction of the proteome involves *translation* of these RNA molecules into protein (Figure 2.2). Transcription does not result in synthesis of a new transcriptome but rather (i) its maintenance, (replacing mRNA that have been degraded), and (ii) changes to its composition, (by switching on and off different sets of genes).

It is an inadequate over-simplification to describe synthesis and maintenance of the transcriptome and proteome as the two-step process "DNA makes RNA makes protein", as the series of events involved is much more complex (Figure 2.3). Nevertheless, the *Central Dogma* has considerable acceptance for its simplicity. In reality, genome expression comprises the following steps, discussed by Brown (2002d,g,a,b,e,f,c).

1. Accessing the genome - Involves processes influencing chromatin structure in the parts of the genome that contain active genes, ensuring that these genes are accessible and are not buried deep within highly packaged parts of the chromosomes.

2. Assembly of the transcription initiation complex - this comprises a set of proteins that work together to copy genes into RNA. Assembly of initiation complexes is a highly targeted process because these complexes must be con-

structed at precise positions in the genome, adjacent to active genes.

3. Synthesis of RNA, during which the gene is transcribed into an RNA copy.

4. Processing of RNA molecule - Involves a series of alterations made to its sequence and chemical structure, and which must occur before the RNA molecule can be translated into protein or, in the case of non-coding RNA, before it can carry out other functions in the cell.

5. RNA degradation - The controlled turnover of RNA molecules, (plays an active role in determining the makeup of the transcriptome), (Lorkowski and Cullen, 2006).

6. Assembly of the translation initiation complex - Occurs at the termini of coding RNA molecules, and is a prerequisite for translation of these molecules.

7. Protein synthesis - The synthesis of a protein by translation of an RNA molecule, (Brown, 2002f).

8. Protein folding and protein processing. Folding results in the protein taking up it's correct three-dimensional structure. Processing involves modification of the protein by addition of chemical groups and, for some proteins, removal of one or more segments of the protein, (Branden and Tooze, 1999).

9. Protein degradation has an important influence on the composition of the proteome and, like RNA degradation, is an integral component of genome expression, (Petsko and Ringe, 2004a).

Control mechanisms exist for regulation of each step, (Figure 2.3), allowing the cell to 'adjust expression' in response to changes in its environment and to signals

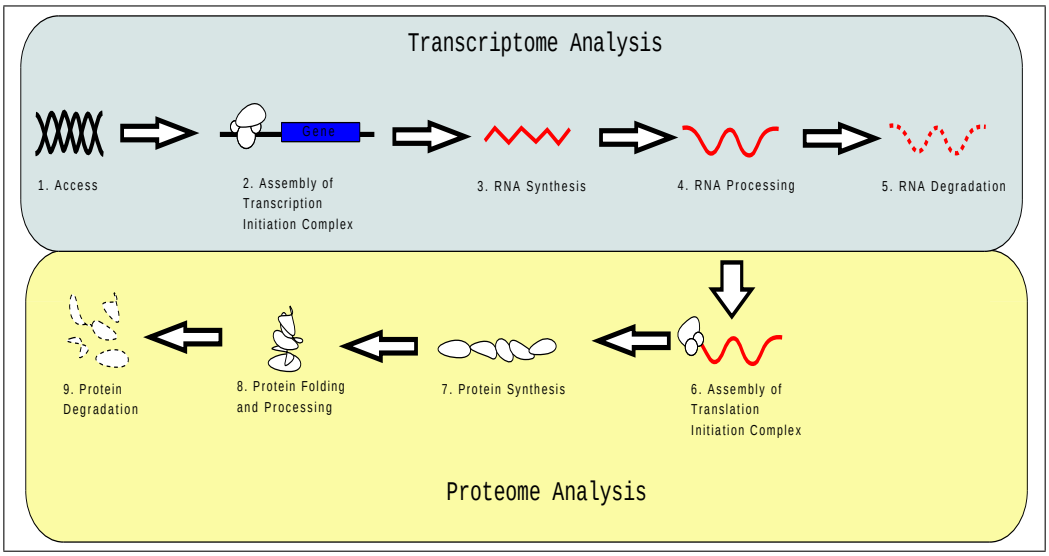


Figure 2.3: Detailed Central Dogma

received from other cells. These *regulatory events* determine not only function of individual cells but also the processes of differentiation and development.

Processing of mRNA has an important influence on the composition of the transcriptome. RNA editing, for example, can result in a single pre-mRNA being converted into two different mRNAs, coding for distinct proteins. Splicing, in which one pre-mRNA gives rise to two or more mRNAs by assembly of different combinations of exons[3], resulting in protein isoforms[4], is fairly widespread, (Modrek and Lee, 2002; Lee and Wang, 2005; Birzele et al., 2007). The mRNA resulting from both editing and alternative splicing often displays tissue specificity. These processing events increase the coding capabilities of the genome without the requirement for increased gene number. To some degree, this explains the "surprise" discovery of the Human Genome Project that an estimated $20,000 \sim 25,000$ protein-coding

[3]A segment of a gene that contains instructions for making a protein. In many genes the exons are separated by "intervening" segments of DNA, known as introns, which do not code for proteins; these introns are removed by splicing to produce messenger RNA.

[4]An alternative form of a protein resulting from differential transcription of the relevant gene either from alternative promoter or alternate splicing.

genes only are responsible for the synthesis of many thousands of functional proteins, (Pennisi, 2000; Lander et al., 2001; Consortium, 2004).

Many genes are active at any one time in a given cell. Transcriptomes are therefore *complex*, containing copies of hundred, if not thousands, of different mRNAs. Usually, each mRNA contributes a small fraction only of mRNA abundance, with the most common type rarely contributing more than 1% of the total. Overall, mRNA itself typically accounts for $\sim 4\%$ of total abundance of RNA in a cell, while non-coding RNA makes up the remainder, (Brown, 2002g)

2.3 Analysing Gene Expression

Identification of which genes are active, and under what circumstances, is a constant goal of the scientific community. The development of high-throughput technologies for the measurement of the entire transcriptome has evolved from gene-by-gene experimental methods, such as reverse transcriptome polymerase chain reactions (RT-PCR) (Erlich, 1989), reverse northern (Alwine et al., 1977) and Southern hybridisation (Southern, 1975). These relatively new high-throughput tools have broadened the size and scope of biological questions scientists can pose.

2.3.1 Microarray Technologies

One-at-a-time study of gene expression is time consuming and limited as multiple gene expression is typical for an organism. Microarray and DNA chip technology have facilitated determination of transcriptome composition, enabling comparisons between them.

Various manufacturers provide a large assortment of different microarray plat-

forms (Aligent and Affymatrix for instance). Fundamentally, all platforms take advantage of the specificity and affinity of complementary base-pairing of nucleic acid. Thousands of known discrete DNA sequences (known as *probes*) are attached, (via print, jet, photolithography), at known positions to a solid surface. To measure the quantity of transcripts of specific genes in a sample of interest, genetic material is extracted from the sample and labelled with a fluorescent dye. The labelled genetic material is referred to as the *target*. The target is then allowed to *hybridise*[5] to the probes on the slide. After hybridisation, a specialised scanner is used to measure the amount of fluorescence (i.e. amount of target) at each probe, which is reported as *intensity*. The "raw" or "probe-level" data are the intensities read for each of these components. As the address and sequence of each probe is known, probe intensities determine the abundance of the sequence of each specific gene in the target.

Microarray Platforms

Different platforms can be divided into two main classes that are differentiated by the type of data they produce.

(A) The **high density oligonucleotide array** platform contains probes of length 25 *base-pairs*(bp)[6] that are synthesised directly onto the array surface using photolithography. Rather than discrete spot association with a transcript of interest, a combinatorial probe is used for each gene, i.e. each gene on the array is represented by a series of smaller oligonucleotides that span different parts of the gene. Most probes are designed to represent the most common

[5]Base-pair to form double stranded structure

[6]Base-pairs are the units for measuring the length of a nucleic sequence. Each nucleotide in a sequence is one base-pair. 1000bp = 1kb (kilo base-pairs), 1000kb = 1mb (mega base-pairs) etc.

transcripts expressed from a gene. Probe design is thus constrained by needing to meeting the most consistent hybridisation parameters across the entire set - as the size and concentrations of each probe is predetermined, the most critical factors are the Guanine-Cytosine content[7], predicted secondary structure within the probe itself, as well as the uniqueness of the probe sequence itself, (Pham et al., 2006). There are typically mismatch probes and control probes incorporated in the design to measure the amount of non-specific binding and for use in normalisation procedures. Genetic material from one sample of interest is hybridised to the array resulting in one set of probe-level data per microarray. This platform is typically used for well-sequenced and annotated genomes. The data is, within reason, robust and reproducible between laboratories, with the design of each oligonucleotide being critical to the robustness of the array as a whole.

(B) For **two-colour spotted** (cDNA array) platforms, the probes may be as long as the gene product, ($\sim 2kb$). Probe sets are typically captured from gene products expressed in an organism of interest, so represent a set of likely gene expression patterns. The genetic material from two target samples are labelled with separate dyes (Cy3 (green) and Cy5 (red)), mixed together and hybridised to the same array, thus producing two sets of probe-level data per microarray (the red and green channels), (Gentleman et al., 2005a). This platform is typically used for incompletely sequenced genomes, but is often limited by poor probe annotation, redundancy in the array of gene products, cross-hybridisation of sequences common to different probes and sub-optimal

[7]Percentage of bases which are either guanine or cytosine. GC pairs are more thermostable compared to the alternative Adenine-Thymine pairs. Genes are often characterised by having a higher GC content in contrast to the background GC content for the entire genome. Evidence of GC ratio with that of length of the coding region of a gene have shown that the length of the coding sequence is directly proportional to higher GC content, (Oliver and Marn, 1996)

variation in hybridisation efficiencies[8] across the probe set. The platform is useful for organisms not well represented in the public sequencing databases, (Pham et al., 2006).

The choice of platform (A or B) determines the type of experimental design (two colour comparison, or single colour indirect comparison), as well as choice of normalisation and filtering strategies employed. Sources of variation need to be accounted for, and data must be heavily manipulated before the genomic level measurements used for analysis can be obtained.

2.4 Design of Gene Expression Analysis Experiments

The size and scope of the questions the data can answer is dependent on the experimental design. Important features are: (i) choice and collection of samples (tissue biopsies or cell lines exposed to different treatments); (ii) choice of probes and array platform to use; (iii) choice of controls (to measure non-specific binding, noise, and for normalisation procedures etc.); (iv) RNA extraction, amplification, labelling, and hybridisation procedures; (v) allocation of replicates; and (vi) scheduling. Unsurprisingly, the quality of experimental design to a large extent determines the utility of the data, and avoidance of confounding between biological factors and/or measurement artefacts is important. Examples of biological factors include tissue heterogeneity, genetic polymorphism, and changes in mRNA levels within cells and among individuals due to sex, age, race, genotype-environment interactions and so on. The Biological variation between experimental units (i.e. individual mice, rats, tissue samples etc.) is of intrinsic interest to investigators. However, technical variation inherent in preparation of samples, labelling, hybridisation and other steps

[8] determined by probe length and composition

of microarray experimentation, can significantly impact data quality, (Zakharkin et al., 2005) To minimise this quality control of RNA samples is required, (Gentleman et al., 2005a). For good reviews of microarray experiment design principles see Yang and Speed (2002), Churchill (2002) and Pham et al. (2006).

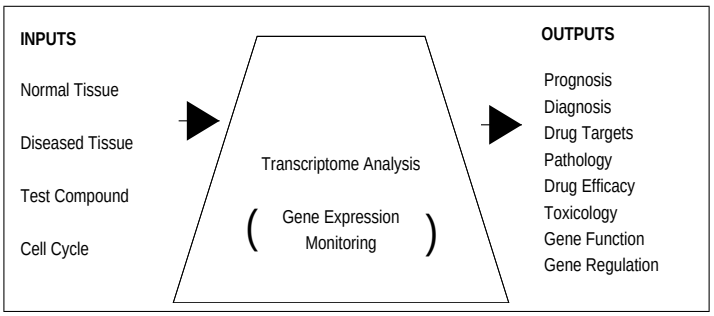


Figure 2.4: Microarrays have a large number of applications and expression data measurements can have an impact on a large spectrum of research areas.

Replicates: Biological replicates are samples extracted from independent[9] biological units (cell line, organism etc.). *Technical replicates* are genetic material extracted from the same biological unit and hybridised to two different arrays. Independent biological replication is very important in experimental design, to achieve adequate power and validity in statistical inference and testing. For technical replicates, conclusions are descriptive and limited to the samples used (i.e. descriptive as opposed to inferential), (Churchill, 2002).

Comparison: Two-colour cDNA arrays are inherently directly comparable between samples. With oligonucleotide arrays, however, only one sample can hybridise to an array, thus only *indirect* comparisons are possible. The main design issue with cDNA microarrays is to determine which RNA samples should be hybridised together on the same slide to achieve the desired precision. Popular de-

[9]Two measurements are considered independent if the experimental materials on which the measurements are based receive different treatments and if the materials were handled separately at all stages of the experiment.

signs and other practical considerations are discussed by Churchill (2002); Yang and Speed (2002) and references therein.

Reproduction and Analysis: The experimental design influences both data collected and the analysis performed. Comprehensive and meticulous details of the experimental procedures are vital for subsequent analysis of the data.

2.5 Microarray Data

2.5.1 Abundance Measures

High density oligonucleotide array and cDNA array platforms measure *overall* and *relative* abundance of a probe sequence in one and two target samples respectively, i.e. the former give **absolute** (log) intensities, while the latter give **ratio** (log) intensities. In many cases one of the samples in a cDNA array hybridization is a common reference used across multiple slides, with the sole purpose of providing a baseline for direct comparison of expression between arrays, (Gentleman et al., 2005b).

For oligonucleotide arrays, the direct comparison of expression measures within arrays is problematic, because fluorescent intensities are not the same across genes. The measured fluorescence intensities are roughly proportional to mRNA abundance but the proportionality factor, (p), is different for each gene. When using short oligonucleotide arrays, p is a function of the probes used and, in particular, of the frequencies of the different nucleotides in each. Specifically, the between-sample, within-gene comparisons are valid and sensible, but the within-sample, between-gene comparisons are not.

As an illustration of the difficulty, suppose that genes a and b, have estimated expression measures 100 and 200 respectively, in sample i. These observed data tell

us nothing about the real relative abundance of mRNA for these two genes. There could in fact be more copies of the mRNA for gene a. On the other hand, if in a second sample, j, gene a has an expression measure of 200, we could conclude that the abundance of mRNA for a in sample j is likely to be higher than that observed in sample i, (Gentleman et al., 2005b).

For cDNA arrays the measure of interest is a ratio of abundance, typically calculated relative to a standard reference. Consider a sample, i, with estimated relative abundance of 1 for gene a and 2 for gene b. It can be inferred that gene a is expressed at approximately the same level in sample i as in the reference sample, while gene b has approximately twice the abundance of mRNA as the reference sample. Note: relative values if a, b mRNA abundance are unknown as the value in the reference sample is not specified here. Certain designs are also less readily interpretable e.g. dye swaps (Churchill, 2002) (a gene ratio may be recorded a 2 for one slide and 0.5 for another, corresponding to the same abundance of target). These simple examples serve to show that data from transcriptome analysis experiments need to be carefully interpreted in the context of the experimental design.

2.5.2 Gene Expression Data Characteristics

Once raw gene expression data are collected and processed, these are typically presented as a real-valued matrix, with rows corresponding to gene expression measurements over a number of experiments, and columns corresponding to the pattern of expression of all genes for a *given* microarray experiment, Figure 2.5. Each entry, x_{ij}, is the measured expression of gene i in experiment j. *Dimensionality* of a gene or sample refers to the number of its expression or sample values recorded (number of matrix columns or rows respectively). A *gene/gene profile*, $\overrightarrow{g_i}$, is a single

data item (row) consisting of p measurements, $\overrightarrow{g_i} = (x_{i1}, x_{i2}, ..., x_{ip})$. An *experiment/sample* $\overrightarrow{s_j}$ is a single microarray experiment corresponding to a single column in the gene expression matrix, $\overrightarrow{s_j} = (x_{1j}, x_{2j}, ..., x_{nj})^T$, where n is the number of genes in the dataset. The notation adopted throughout this thesis is presented in Table 2.1.

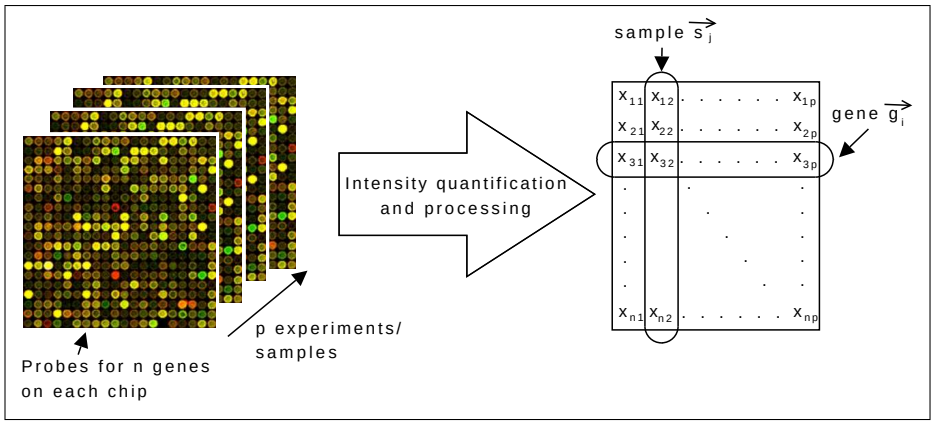


Figure 2.5: Gene Expression Matrix

X	Gene expression matrix
n	Number of genes
p	Number of samples
x_{ij}	A cell in the gene expression matrix
$\overrightarrow{g_i}$	Gene vector i
$\overrightarrow{s_j}$	Sample vector j

Table 2.1: Notation used throughout thesis

From the above discussion, we can summarise the properties of microarray data. *Accuracy:* The accuracy of gene expression data strongly depends on experimental design and minimisation of technical variation, whether due to instruments, observer or pre-processing[10], (Zakharkin et al., 2005). It also depends on the number

[10]Preprocessing is a processes applied to the raw data, such as standardisations, normalisations to remove noise etc. to produce data that can used as input to another program

of alterations of an mRNA molecule before it is measured by the array.

Incompleteness: Image corruption and/or slide impurities may lead to unusable or undetectable fluorescent intensities resulting in *incomplete* data, or missing values (Troyanskaya et al., 2001).

Noise: Due to the many uncontrollable factors in the experiments (biological variation, binding efficiencies, cross hybridisation etc.), gene expression data is intrinsically *noisy*, resulting in outliers, typically managed by: (i) robust statistical estimation/testing, (when extreme values are not of primary interest), (ii) identification, (when outlier information is of intrinsic importance), (iii) manual screening for defective slides, (Liu et al., 2002).

High Dimensional Data: The resulting dataset is of high dimension, with a few experiments, reporting on a large number of variables.

2.6 Beyond Microarrays

In addition to microarrays, other experimental methods for gene expression monitoring are continually under development. Serial Analysis of Gene Expression (SAGE), for example, is a technique used to produce a snapshot of the mRNA population in a sample, (Velculescu et al., 1995). The output of SAGE is a list of short sequence tags and the number of times each is observed. Using sequence databases a researcher can usually determine the original mRNA, (and therefore which gene), the tag was extracted from. Statistical methods can be applied to tag and count lists from different samples in order to determine which genes are more highly expressed between different samples. Several variants of SAGE have been developed, such as, LongSAGE (Saha et al., 2002), RL-SAGE (Gowda et al., 2004) and SuperSAGE (Matsumura et al., 2003). SuperSAGE advances its predecessors by using a

technique that expands the tag-size, (Matsumura et al., 2003). The longer tag-size allows more precise allocation of the tag to the corresponding transcript. By direct high-throughput sequencing techniques, thousands of tags can be analyzed in one run, producing precise gene expression profiles.

Many vendors provide an assortment of array platforms that are continually improved and upgraded. An alternative microarray, based on randomly arranged beads was developed by Illumina technologies. A specific 50bp oligonucleotide is assigned to each bead type, which is replicated approximately 30 times on an array, (Kuhn et al., 2004). The high degree of replication makes robust measurements for each bead type possible. Randomisation of the probe spots between arrays further avoids the potential systematic biases that could be introduced due to regular arrangements, (e.g. printing, scanning conditions etc.). Ideally, different arrays used in an experiment should have similar clones in different positions. Formerly, used for single nucleotide polymorphism detection amongst others. Illumina bead arrays can now be used to monitor the expression of genes in the entire genome.

Microarrays and their alterations offer a unique opportunity to analyse gene expression and regulation at a global cellular level. However, the generation of large datasets presents challenges in analysis and warehousing of the data, as well as its integration of that data with other high throughput platforms.

The "Central Dogma" of "DNA makes mRNA makes proteins" (that comprise the proteome) is overly simple. A single gene does not translate into one protein and protein abundance depends not only on transcription rates of genes but also on additional control mechanisms, such as mRNA stability , regulation of the translation of mRNA to proteins and protein degradation . Proteins can also be modified by post-translation activity (Brown, 2002(a)). Inevitably, the integration of transcriptome and proteome data will provide a more complete understanding of the

connection of gene expression to the physical chemistry of the cell. Integration and merger of proteomic[11] and transcription data sources across platforms is needed, together with development of automated high-throughput comparison methods if detailed understanding of cell mechanisms is to be achieved. To this end, as well as to successfully integrate cluster information from different datasets, standardisation of gene and protein annotation methods across databases is overdue, (Waters, 2006). Finally, recent developments in new ontologies and databases facilitate storage of expression and meta information, which assists enormously in validation of exploratory analyses of gene expression datasets.

2.7 Summary

Gene expression analysis represents only one parameter by which cells or tissues may be characterised. Depending on the experiment, epidemiological or molecular pathological data, genomic changes or sensitivity to drugs may be additional parameters that will influence the interpretation of microarray data. The ability to combine RNA and protein expression data to comprehensively profile both transcriptional and post-transcriptional changes in cells and tissues is particularly appealing, although the number of proteins that can be profiled at this stage is substantially less than the number of genes. Although it is more difficult to identify proteins that are differentially expressed, techniques for rapid and reproducible two-dimensional gel protein separation and mass spectrometry-based protein identification make high throughput proteomics, not only desirable, but feasible in the short term as an adjunct to microarray transcriptome analysis, (Bowtell, 1999). Consequently, accurate algorithms and computing techniques are needed to measure and understand

[11] Protein measurement methods include ICAT, MudPIT, 2-DE.

transcript data before integration with other levels.

Measurement of gene expression is critical for the understanding of cellular biological processes, and although it is not a complete link between sequence and cell function, it is an important element in the chain of events. While high-throughput technology has "evolved" from its ancestor technologies, there are "side effects" which need to be dealt with, not least the challenge of analysing abundant data. Even as this work was completed, new and more precise methods of measurement have emerged, but the need for robust analysis techniques remains.

CHAPTER 3

A REVIEW OF TECHNIQUES

This Chapter examines the strengths and weaknesses of algorithms used in the analysis of gene expression data. The properties of these data were introduced in Chapter 2 and we investigate here what is required from pattern-finding algorithms to meet the challenges of their analysis. We focus on unsupervised pattern recognition algorithms, reviewing key concepts; (further details are also referenced appropriately).

3.1 Introduction

Array data are used to determine which genes are expressed under which conditions, and for comparisons between transcriptomes of different biological samples. Searching for meaningful information patterns and dependencies in gene expression data, to provide a basis for hypothesis testing is non-trivial. An initial step is to cluster or "group" genes, with similar changes in expression. Lack of *a priori* knowledge means that *unsupervised* clustering techniques, where data are unlabelled (un-annotated), are common in gene expression work.

Many excellent reviews of gene expression analysis, using clustering techniques, are available. Asyali et al. (2006) provide a synopsis of class prediction and dis-

covery; (respectively, supervised pattern recognition and clustering), while Pham et al. (2006) provide a comprehensive literature review of the various stages of data analysis during a microarray experiment. In a landmark paper, Jain et al. (1999) provided a thorough introduction to clustering, and gave a taxonomy of clustering algorithms, (used in this work). Reviewing the state of the art in gene expression analysis is complicated by the high level of interest in exploratory analysis and the consequent proliferation of techniques, Figure 3.1. We restrict our assessment to a selection of those methods, which illustrate the properties of each group in the taxonomy according to Jain et al. (1999), and also to those which address shortcomings of conventional approaches, by introducing modifications to account for properties specific to gene expression data.

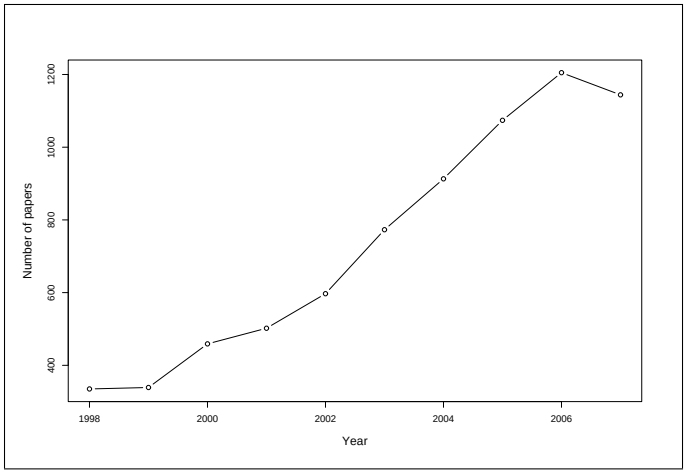


Figure 3.1: Data from ISI web of science on number of papers published in the area with key words “Clustering (Pattern Recognition)” and “Gene Expression”

3.2 Pattern Recognition

Pattern recognition is an exploratory technique and assumes that there is an unknown mapping that assigns a group “label” to each gene, where the goal is to

estimate this mapping. However, common clustering approaches do not always translate well to gene expression data, and may fail significantly to account for the data profile.

To recapitulate from Chapter 2: properties (which any clustering algorithm in this domain must account for) include: (i) Accuracy: this is not absolute and depends on a number of factors, not least the experimental design and the platform used; (ii) Missing values: common in gene expression datasets, mainly because of the complex and specific nature of the experiments. Frequently, the measured intensity is not convincing and deemed "absent" by experimentalists. Dust, scratches or image quality may also render a large proportion of values unusable. Consequently, any clustering algorithm which can not account for missing values must be supplemented by missing value estimation procedures prior to any exploratory analysis of the data; (iii) Incorporated noise: this is typically multiplicative or additive, may be caused by measurement or experimental error, and affects accuracy. A good clustering algorithm should not be greatly affected by spurious measurements, and should treat outliers with caution.

As cluster analysis is usually exploratory, lack of *a priori* knowledge on gene groupings or the number of these, K, is common. Arbitrary selection of K may undesirably bias the search, as pattern elements may be ill-defined unless signals are strong. Meta-data can guide choice of correct K, *e.g. genes with common promoter sequence are likely to be expressed together and thus are likely to be placed in the same group*. Methods for determining optimal number of groups, K, are discussed by Milligan and Cooper (1985) and Fridlyand and Dudoit (2001).

3.2.1 Cluster Types

In general, a cluster is defined as a group of objects, which exhibit similar properties according to some *objective* criterion. Clustering a gene expression matrix can be achieved in two ways:

1. genes can form a group which show similar expression across samples, (i.e. grouping *rows* of the gene expression matrix).

2. samples can form a group which show similar expression across all genes, (i.e. grouping *columns* of the gene expression matrix).

Both (i) and (ii) lead to *global clusters*, where a gene or sample is *grouped across all dimensions*. However, genes and samples can also be clustered simultaneously, with their inter-relationship represented by *bi-clusters*. These are defined over a subset of genes and a subset of samples thus focusing on a Section of the gene expression matrix and capturing *local structure* in the dataset. This is an important strength as **cellular processes are understood to rely on subsets of genes, which are co-regulated and co-expressed under certain conditions and behave independently under others**, (Ben-Dor et al., 2003).

Justifiably, this approach has been gaining much interest of late. For an excellent review on bi-clusters and bi-clustering techniques see Madeira and Oliveira (2004).

Additionally, clustering can be *complete* or *partial*, where the former assigns each gene to a cluster, and the latter does not. Partial clustering tends to be more suited to gene expression, as the dataset often contains irrelevant genes or samples. This allows: (i)“noisy genes” to be left out, with correspondingly less impact on the outcome and (ii) genes to belong to no cluster - omitting a large number of irrelevant contributions. This is important as microarrays measure expression for the entire

genome in one experiment, but genes may change expression independently of the experimental condition, (e.g. due to stage in the cell cycle). Forced inclusion, (as demanded by complete clustering), in well-defined but inappropriate groups may impact final structure found for the data. Partial clustering thus avoids the situation where an interesting sub-group in a cluster is obscured through forcing membership of unrelated genes.

Finally, clustering can be categorised as *exclusive (hard)*, or *overlapping*. Exclusive clustering requires each gene to belong to a single cluster, whereas overlapping clusters permit simultaneous membership of numerous clusters. This membership may qualify additionally as *crisp* or *fuzzy*. Crisp membership is boolean - either the gene does or does not belong to a group. In the case of fuzzy membership, each gene belongs to a cluster with a *membership weight* between 0, (definitely excluded), and 1, (definitely included). **Clustering algorithms, which permit genes to belong to more than one cluster are typically *more applicable* to gene expression since: (i)impact of "noise" is reduced - the assumption is that "noisy" genes are unlikely to belong exclusively to any one cluster but are equally likely to be members of several, (ii) this supports the underlying principle that genes, with similar change in expression for a set of samples, are involved in a similar biological function. Typically, gene products are involved in several such biological functions and groups need not be co-active under all conditions. Thus gene groups are fluid, so that constraining a gene to a single group (hard cluster) is counter-intuitive.**

3.2.2 Steps in Cluster Analysis

Cluster analysis includes several basic steps, (Jain et al., 1999). Initially, the data matrix is represented by number, type, dimension and scale of the gene expression profiles. Some features are set during the experiment, others are controllable, (e.g. scaling, imputation, normalisation etc.). An optional step of *feature selection* or *feature extraction* may also be carried out. The former refers to selecting, from the original features, a subset, which is most effective for clustering, while the latter refers to transformation of the input features to form a new set that may be more discriminatory in clustering, e.g. through *Principal Component Analysis*, (Yeung and Ruzzo, 2001).

Pattern proximity assessment is needed, usually provided by a "distance" measure between pairs of genes. (Alternatively, "conceptual" measures can be used to characterise similarity of gene profiles e.g. Mean Residue Score of Cheng and Church (2000), (see Section 3.3.1)). The next step is to apply a clustering algorithm to *determine structure* in the dataset. Methods can be broadly categorised according to taxonomy in Jain et al. (1999).

Structures are then described by *data abstraction*. For gene expression data, the context is usually direct interpretation by a human, so abstraction should ideally be straightforward, (for follow up analysis/experimentation). Required is usually a *compact description* of each cluster, through a prototype or representative selection of points, such as the centroid. Clusters are valid if they can not reasonably be achieved by chance or as an artefact of the clustering algorithm. *Validation* requires formal statistical testing, and can be categorised as Internal or External (see Chapter 4).

3.3 Clustering and Clustering Extensions

Analysis of large gene expression datasets is a relatively new task, although pattern recognition of complex data is well-established in a number of fields. Many common generic algorithms have, in consequence, been adopted for gene expression data, (e.g. Hierarchical (Eisen et al., 1998), SOM's (Kohonen, 1990), among others), but not all perform well. A good method must deal with noisy high dimensional data, be insensitive to the order of input, have moderate time and space complexity, (i.e. allow increased data load without breakdown or requirement of major changes), require few input parameters, incorporate meta-data knowledge (an extended range of attributes), and produce results, which are interpretable in the biological context.

3.3.1 Pattern Proximity Measures

The choice of proximity measure, needed to evaluate degree of expression coherence in a group of gene vectors, is as important as choice of clustering algorithm, and is based on data type and context of the clustering. Many clustering algorithms either employ a proximity matrix directly (e.g. hierarchical clustering), or use one to evaluate clusters during execution (e.g. K-Means). Proximity measures are calculated between pairs (e.g. Euclidean distance) or groups of genes (e.g. Mean Residue Error). For ease of reference, the mathematical formulae for each of the distances can be found in Appendix A.

Distances: Distance functions between two vectors include the so-called Minkowski measures, (Euclidean, Manhattan, Chebyshev, (Romesburg, 2004)), useful when searching for *exact* matches between two profiles in the dataset. These tend to find

globular structures and work well when these are compact and isolated. A drawback is that the largest feature dominates, so measures are sensitive to outliers, (Jain et al., 1999). However more sophisticated variants, such as *Mahalanobis distance*, also account for correlations in the dataset and are scale-invariant, (Romesburg, 2004). Different distance measures produce clusters of different shape, (e.g Euclidean are spherical, while Mahalanobis' are ellipsoidal). Alternatively, Kim et al. (2005) describe an *adaptive distance norm* (the Gaustafson-Kessel method). Here co-variances are estimated for the data in each cluster, (based on eigenvalue calculations), to obtain structure. Each cluster is then created using a unique distance measure.

Distances based on correlations reflect degree of similarity of *changes in expression across samples*, for two gene expression profiles, without regard to scale. For example, if, for a set of samples, gene X is up-regulated, and gene Y is down-regulated, i.e. are negatively correlated, then X and Y would form a cluster. This would clearly not be the case if Minkowski distances were used, since the average absolute distance between the points would be large. Correlation coefficients commonly used include both parametric (standard *Pearson* , *cosine*), and non-parametric (*Spearman's rank* and *Kendall's* τ), the latter used when outliers and noise are present, (Romesburg, 2004). In general, $distance = 1 - correlation^2$, if sign is unimportant.

Conceptual Measures: As an alternative to measures of distance, "conceptual" measures of similarity can be used. Models are based on *constant rows, columns* and *coherent values, (additive or multiplicative)*, (Madeira and Oliveira, 2004) (Fig. 3.2). A "good fit" indicates high correlation within a sub-matrix, (thus a possible cluster). These models are common to several clustering algorithms. For example, Cheng and Church (2000) and FLOC (Yang et al., 2003), use the additive model

(Fig. 3.2(C)), to evaluate biclusters obtained by determining the Mean Residue Score. Given a gene expression matrix, X, an element a_{ij} in a sub-matrix, $A = (I, J)$ is given by the constant additive model:

$$a_{ij} = \mu + \alpha_i + \beta_j + r_{ij} \tag{3.1}$$

Note that that the mean value of A corresponds to a_{IJ}, the offset of row i corresponds to $\alpha_i = a_{iJ} - a_{IJ}$, (the mean of row i minus the overall mean of A), the offset for column j corresponds to $\beta_j = a_{Ij} - a_{IJ}$, (the mean of column j minus the overall mean of A) and r_{ij} corresponds to unexplained error, which must be minimised. Simply rearranging equation 3.1 the residue of an element is calculated as:

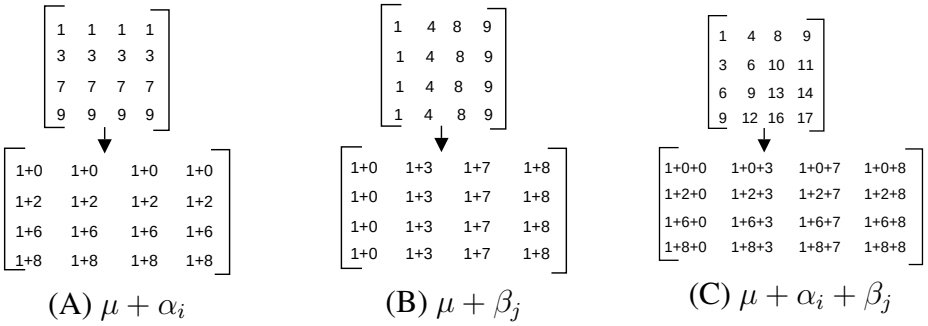
$$r_{ij} = (a_{ij} - a_{iJ} - a_{Ij} + a_{IJ}) \tag{3.2}$$

The *"H-score"* of the sub-matrix is then the sum of the squared residues, given by:

$$H(I, J) = \frac{1}{\mid I \mid\mid J \mid} \Sigma_{i \in I j \in J} (r_{ij})^2 \tag{3.3}$$

A *perfect bi-cluster* gives a H-score equal to *zero*, (corresponding to "ideal" gene expression data, with constant or additive matrix rows and columns).

The *Plaid Model* bi-cluster variant, (Lazzeroni and Owen, 2000), builds the gene expression matrix as a sum of layers, where each layer corresponds to a bi-cluster. Each value a_{ij} is modelled by $a_{ij} = \sum_{k=0}^{K} \theta_{ijk} \rho_{ik} \kappa_{jk}$ where K is the layer (bi-cluster) number, and ρ_{ik} and κ_{jk} are binary variables representing membership of row i and column j in layer k. Here, the value of an element in the gene expression matrix is a linear function of the contributions of the different bi-clusters to which



(A) $\mu + \alpha_i$ (B) $\mu + \beta_j$ (C) $\mu + \alpha_i + \beta_j$

Figure 3.2: (A) Bi-cluster with constant rows. Each row is obtained from a typical value μ and row offset α_i, (B) Constant columns. Each value is obtained from a typical value μ and column offset β_j, (C) Additive model. Each value is predicted from μ, and a row and column offset, $\alpha_i + \beta_j$. Similar model constructs apply for the multiplicative case with (A(i)) $\mu \times \alpha_i$, (B(i)) $\mu \times \beta_j$ and (C(i)) $\mu \times \alpha_i \times \beta_j$

the row i and the column j belong, (Fig. 3.3), (Lazzeroni and Owen, 2000). For layer k, expression level θ_{ijk} can be estimated using the general additive model, $\theta_{ijk} = \mu_k + \alpha_{ik} + \beta_{jk}$, *in layer k*, (Fig. 3.2 (C)).

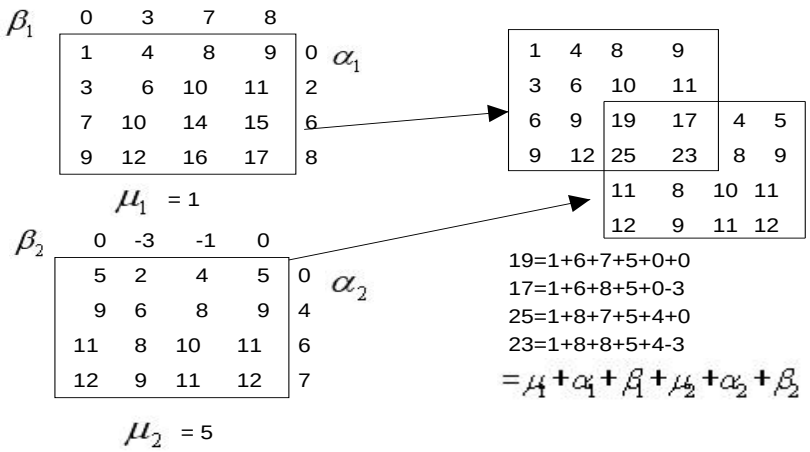


Figure 3.3: Values at overlaps are seen as a linear function of different bi-clusters.

For the *Coherent Evolutions model* the exact values of x_{ij} are not directly taken into account, but a cluster is evaluated to see if it shows coherent patterns of ex-

pression. In simplest model form, each gene expression value can have three states: up-regulation, down-regulation and no change. *Thresholds between states are crucial and additional complexity results from extending model definitions to include further states such as "slightly" up-regulated, "strongly" up-regulated and so on*, e.g. SAMBA, (Tanay et al., 2002).

Other Measures: Other measures used to evaluate coherency of a group of genes include *conditional entropy*: $H(C|X) = -\int \sum_{j=1}^{m} p(c_j|x) log p(c_j|x) p(x) dx$, (the average uncertainty of the random variable C (cluster category), when a random variable X (gene expression profile) is known). The optimal partition of the gene expression dataset is obtained when this entropy is minimised, (Li et al., 2004) i.e. a partition is achieved where each gene is assigned with a high probability to only one cluster. This requires the estimation of the *a posteriori* probabilities $p(c_j|x)$, usually by non-parametric methods, as this avoids assumptions on the distribution of the underlying gene expression data).

Note: Pattern proximity measures described so far make no distinction between time-series data and those obtained from expressions of two or more phenotypes. Applying similarity measures to time series data is not straightforward. Gene expression time series have non-uniform intervals and are usually very short, (4-20 samples while classically even 50 observations is low for statistical inference). Furthermore, data are not independently, identically distributed. Similarity in time series should be viewed only in terms of similar patterns in the direction of change across time points (i.e. trends in the data), while robust measures must allow for non-uniformity, in addition to scaling, shifting and shape (internal structure of clusters), (Moller-Levet et al., 2003).

3.3.2 Conventional Methods

In this Section we examine popular clustering methods and their more recent developments. Each algorithm described below relies, by definition, on some choice of proximity measure and inherits the limitations of that choice.

Agglomerative clustering

All agglomerative techniques naturally form a hierarchical cluster structure in which genes have crisp membership. Eisen et al. (1998) studied gene expression in the budding yeast, *Saccharmyces Cerevisiae*, using hierarchical methods, (which have been popularised due to ease of implementation, visualisation capability and availability). Methods vary with respect to choice of distance metric, decision on cluster merging, (linkage), as well as parameter selection affecting structure and relationship between clusters. Options include: *single linkage* (cluster separation as distance between two nearest objects), *complete linkage* (as previously, but between two furthest objects), *average linkage* (average distance between all pairs), *centroid* (distance between centroid's of each cluster) and *Ward's* method, (which minimises ANOVA Sum of Squared Errors between two clusters), (Sturn, 2001).

Distance and linkage determine level of sensitivity to noise: Ward's and the Complete method are particularly affected, (due to the ANOVA basis and outlier importance respectively, since clustering decisions depend on maximum distance between two genes). Single linkage forces cluster merger, based on minimum distance, regardless of other gene contributions to the cluster, so noisy or outlying values are among the last to be considered. Consequently, the *"chaining phenomenon"* may arise, (Romesburg, 2004). For commonly used Average and Centroid linkage

this problem is avoided, as no special consideration is given to outliers and clusters are based on highest density.

Results for agglomerative clustering may be intuitively presented by dendograms but there are 2^{n-1} different linear orderings consistent with tree structure, so care is needed in pruning. Dendrogram analysis, based on gene class information from specialised databases is presented by Toronen (2004), where optimal correlations are obtained between gene classes and used to form clusters from different branch lengths. Bar-Joseph et al. (2003) present an agglomerative technique for which each internal node has at most N children, allowing up to N genes (or clusters) to be directly connected, (extending traditional hierarchical concepts and reducing the effects of noise). Permutation is used to decide on the number of nodes (max N) to merge, based on a similarity threshold. Heuristically, algorithm complexity is comparable to traditional hierarchical clustering, although Bar-Joseph et al. (2003) also present a "divide and conquer" approach for optimal leaf ordering for small N, which has implications of increased time and space complexity.

Note: It should be stated that, such methods can not, in general, compensate for the greedy nature of the traditional algorithm, where mis-clustering at the beginning can not be corrected at a later stage and are magnified as the process continues. Further, Yeung et al. (2001) and Gibbons and Roth (2002) note that hierarchical clustering performance is close to random, despite its popularity and is poorer than other common techniques such as K-means and Self Organising maps (SOM).

Partitive Techniques

Partitive clustering divides data by similarity measure, where typical methods measure distance from a gene vector to a prototype vector representing the cluster, and

intra-cluster/inter-cluster distance are respectively maximised/minimised. A major drawback is the need to specify the number of clusters in advance. Table 3.2 summarises algorithms discussed here.

• *K-means* produces crisp clusters with no structural relationship between these, (MacQueen, 1967). It deals poorly with noise, since outliers *must* belong to a cluster and this distorts the means. Equally, cluster inclusion is dependent on the cumulative values of genes already present, so *order* matters. Results are dependent on initial cluster prototype (which varies between clustering attempts); this leads to instability and, frequently, to a local minimum solution. Incremental approaches to refine local minima solutions converging to a global solution, include the *Modified Global K-means (MGKM)* algorithm (Bagirov and Mardaneh, 2006), which computes k-partitions of the data using $k-1$ clusters from previous iterations. A tolerance threshold must be set which determines the number of clusters indirectly, and, as with regular K-means, returns spherical clusters. For the six datasets reported the MGKM algorithm showed slight improvement over K-means, but at higher computational time cost, (Bagirov and Mardaneh, 2006).

• The prevalence of local minima for K-means is linked to initial prototype selection. *Genetic algorithms* (GAs), as an evolutionary approach, work well for small datasets, (less than 1000 gene vectors and of low dimension), but have prohibitive time constraints for anything larger, so are less desirable for gene expression analysis. Although GA's find the global optimum, they are sensitive to user-defined input parameters and must be fine tuned for each specific problem. Studies which have combined K-means and GA include *Incremental Genetic K-Means Algorithm (IGKA)*, (Lu et al., 2004). This is a hybrid approach which converges to a global optimum faster than stand alone GA, and without the sensitivity to initialisation prototypes. The fitness function for the GA is based on *Total Within Cluster Vari-*

ance (TWCV), while the basis of the algorithm is to cluster centroids incrementally, using a standard similarity measure. The GA method requires the number of output clusters, K, to be specified, but is further complicated by inherent GA parameters (*mutation probability rate, number of generations, size of the chromosome populations* etc.), which influence time taken by the algorithm to converge to a global optimum.

• Fuzzy modifications of K-means include *Fuzzy C-Means (FCM)*, (Dembele and Kastner, 2003), and *Fuzzy clustering by Local Approximations of MEmberships (FLAME)*, (Fu and Medico, 2007). In both, genes are assigned a cluster membership degree indicating *percentage association* with that cluster, but the two algorithms differ in the weighting scheme used to determine gene contribution to the mean. For a given gene, FCM membership value of a set of clusters is proportional to its similarity to cluster mean. The contribution of each gene to the mean of a cluster is weighted, based on its membership grade. Membership values are adjusted iteratively until the variance of the system falls below a threshold. These calculations require the specification of a *degree of fuzziness* parameter which is problem specific, (Dembele and Kastner, 2003). As with K-Means, clusters are unstable, and considerably influenced by initial parameter values, while K, the number of clusters, must be specified *a priori*. In contrast FLAME requires membership of a cluster, i, to be determined by the weighted similarity of the gene to its K-nearest neighbours, and *their* membership of cluster i.

Note: This density-based approach (FLAME) further reduces noise impact, since genes with a density lower than a pre-defined threshold are categorised as outliers, and grouped with a dedicated 'outlier' cluster. FLAME produces stable clusters, but the size of the neighbourhood and the weighting scheme used affect K (as above) and hence clustering achieved.

For both FCM and FLAME, genes may have multiple and varied degrees of membership, but interpretation differs. FCM and FLAME use averaging, where each gene contributes to the calculation of a cluster centroid, and its overall membership value set sums to 1, (i.e. gene-cluster probability). Thus strong membership for a given gene does *not* indicate it to be more typical of the cluster, but rather the relative strength of its individual association, (Krishnapuram and Keller, 1993).

Table 3.1 illustrates three clusters of an FCM carried out on published yeast genomic expression data of Gasch and Eisen (2002), (available at `http://rana.lbl.gov/FuzzyK/data.html`). Membership values for genes B and D are very different for cluster 21, although both are approximately equidistant from the centroid of the cluster. Similarly genes C and D have comparable membership values for cluster 4, but gene C is more typical (closer to the centroid) than gene D. With similar centroid distances, membership value for gene B in cluster 21 is smaller than that for gene A in cluster 46. These anomalies arise from the membership sum constraint, which decreases gene membership in one cluster to increase it in another. *Listing genes in a cluster based on membership values is therefore counter-intuitive and does not reflect their compatibility with the cluster, but rather how they are shared between clusters*. Similarly for FLAME, as the memberships are weighted relative to the K-nearest neighbours, so a low membership value indicates a high degree of cluster sharing among these and not a more typical value of a given cluster.

This interpretative flaw was recognised by Cano et al. (2007), who developed the *possibilistic biclustering* algorithm, which removes the sum rule restriction. The authors used spectral clustering principles, (Kluger et al., 2003), to create, from the original gene expression matrix, a partition matrix, Z, to which possibilistic clustering is applied. The resulting clusters were evaluated using the H-Score,

GID	Cluster 4		Cluster 21		Cluster 46	
	Centroid Dist.	Mem.	Centroid Dist.	Mem.	Centroid Dist.	Mem.
A	10.691	0.002575	8.476	0.002002	3.864	0.482479
B	6.723	0.009766	3.855	0.009341	6.33	0.007381
C	6.719	0.007653	5.29	0.00515	8.024	0.005724
D	7.725	0.007609	3.869	0.01782	6.279	0.010249

Table 3.1: Fuzzy Membership Interpretation. Membership of a gene and distance to cluster centroid, as calculated by Euclidean distance.

(Eq.3.3), and improved on traditional techniques. The algorithm requires, *inter alia*, two specific parameters, namely *cutoff memberships* for (i) gene inclusion and (ii) sample inclusion in a cluster. In this case, these cutoffs are intuitively reasonable as membership does indicate how typical a gene/sample is to a defined cluster, and *not* the degree to which it is shared between clusters.

Neural Networks

Basic: Neural Networks (NN), loosely based on the biological parallel, can be modelled as a collection of nodes with weighted interconnections. Only numerical vectors are processed, so meta-information can not be included in the clustering procedure. Interconnection weights are *adaptively* learned i.e. features are selected by appropriate assignment of weights. In particular, *Self Organising Maps (SOMs)*, a type of NN, have proved popular for gene expression, (Kohonen, 1990; Tamayo et al., 1999; Golub et al., 1999). A kernel function, that defines the region of influence, (neighbourhood), for an input gene, distinguishes SOM from K-means. Updating the kernel function causes the output node *and its neighbours*, to track towards the gene vector. The network is trained, (adjusting strengths of interconnections), from a random sample of the dataset. Once training is complete, all genes

	Cluster Mem.	**Input**	**Proximity**	**Other**
K-Means	Hard	Starting Prototypes, Stopping Threshold, K	Pairwise Distance	Very Sensitive to input parameters and order of input.
MGKM	Hard	Tolerance Threshold	Pairwise Distance	Not as sensitive to starting prototypes. K specified through tolerance threshold.
IGKA	Hard	K, mutation prob., generation number, population size	TWCV	Time taken to converge to global influenced by parameters.
FCM	Fuzzy	Degree of fuzziness, Starting prototypes, Stop threshold, K	Pairwise Distance	Careful Interpretation of membership values. Sensitive to input parameters and order of input
FLAME	Fuzzy	K_{nn} - number of neighbours	Pairwise Distance to K_{nn} neighbours	careful interpretation of membership values. Output determined by K_{nn}.
Possibilistic biclustering	Fuzzy	Cut-off memberships, Max. residue, number of rows and number of columns	H-Score	Number of biclusters determined when quality function peaks by re-running for different numbers of eigenvalues.

Table 3.2: Summary of Partitive techniques. With the exception of FLAME and Possibilistic biclustering, all find complete global clusters.

in the dataset are then applied to the SOM. Cluster members, represented by output node i are the set of genes causing i to 'fire' (hard clustering).

SOMs are robust to noise and outliers, dependent on distance metric and neighbourhood function used. As for K-means, a SOM produces a sub-optimal solution if the initial weights for the interconnections are not chosen properly. Convergence is controlled by problem-specific parameters such as *learning rate* and *neighbourhood function*. A particular input pattern can fire different output nodes at different iterations; (while this can be overcome by gradually reducing the learning rate to zero during training, it can result in over-fitting, which leads to poor performance for new data). In specifying K, based on the number of output nodes, it should be noted that too few output nodes in the SOM gives large within-cluster distance, while too many results in meaningless diffusion across clusters.

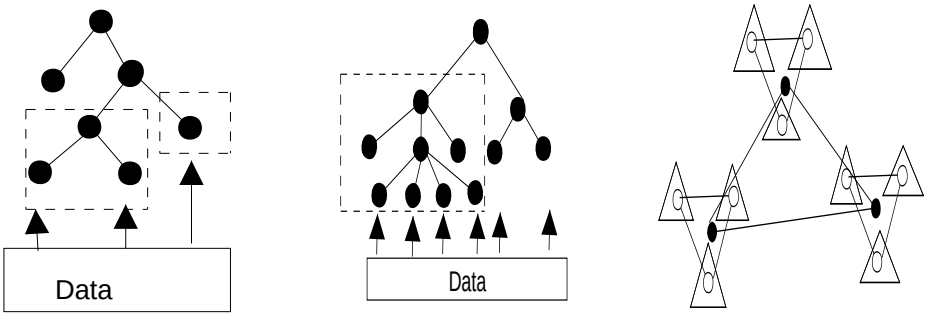
Extended: The *Self Organizing Tree Algorithm (SOTA)*, (Herrero et al., 2001), *Dynamically Growing Self Organizing Tree (DGSOT)* algorithm, (Luo et al., 2004) and, more recently, *Growing Hierarchical Tree SOM (GHTSOM)*, (Forti and Foresti, 2006) were developed to combine strengths of NN, (i.e. speed, robustness to noise) and hierarchical clustering, (i.e. tree structure output, minimum *a priori* requirement for number of clusters specification and training) to deal with properties of gene expression data. Here the SOM network is a tree structure, trained by comparing only leaf nodes to input gene expression profiles (each graph node representing a cluster). SOTA and DGSOT result in a binary and n-tree structure respectively, while in GHTSOM, each node is a triangular SOM (3 neurons, fully connected), each having 3 daughter nodes (also triangular SOMs), Fig. 3.4. Tree growth strategy determines K.

At each iteration of SOTA the leaf node with the highest degree of heterogeneity is split into two daughter cells. In the DGSOT case, the correct number of daugh-

ters, ($n_d \geq 2$), is determined dynamically by starting off with two and continually adding one until cluster validation criteria are satisfied. To determine n_d, a method was proposed, (Luo et al., 2004), based on geometric characteristics of the data (specifically, cluster separation in the minimum spanning tree of the cluster centroids). For this an empirical threshold, α, value must be specified; (the authors propose 0.8)). In SOTA and DGSOT, growth of the tree continues until overall heterogeneity crosses a threshold, β, or until all genes map onto a unique leaf node. The DGSOT method uses average leaf distortion to determine β for growth termination, while, for SOTA, this threshold is determined by re-sampling, (with system variability defined to be the maximum distance among genes mapped to the same leaf node). By comparing distances between randomized data and those of the real dataset, a confidence interval and distance cut-off are obtained. In GHTSOT, growth occurs if a neuron is activated when a sufficient number of inputs map to it, (i.e. at least 3 or a user defined number, β), which determines the resolution of the system. Growth continues as long as there is one neuron in the system which can grow. The advantage of these methods over most partitive techniques is that K is not pre-determined, but depends indirectly on the threshold, β, which is data dependant.

Key Features: SOTA, DGSOT and GHTSOM differ from typical hierarchical clustering algorithms in terms of adaptation. This occurs once a gene is mapped to a leaf node, but the neighbourhood of the adaptation is more restrictive than for SOM. DGSOT also overcomes the misclustering problem of traditional hierarchical algorithms, SOTA and GHTSOM, by specification of another input parameter, L - the immediate ancestor level in the tree of a given node which is growing. DGSOT then distributes all mapped values among the leaves of the subtree rooted at the L^{th} ancestor. In GHTSOM, new nodes (after growth) are trained using only those

inputs which caused the parent node to fire. Any neuron, which shows low activity, is deleted, and its parent is blocked from further growth. This has the advantage that inputs mapping to leaf neurons at the top of the hierarchy are usually noise, and clearly distinguishable from relevant biological patterns.



(A) SOTA (B) DGSOT (C) GHTSOM

Figure 3.4: (A) SOTA. A binary tree structure. Neighbourhood of adaptation indicated for (i) node with sibling, (ii) node with no sibling, (B) DGSOT. N-ary tree structure. Neighbourhood of adaptation indicated when $L = 2$, (C) GHTSOM. Each node represented by triangular SOM. Each layer indicated with line styles, (3 layers shown).

Search Based

Solutions for a criterion function are found by searching the solution space either deterministically or stochastically, (Jain et al., 1999). The former exhaustive search is of little use for high dimensional gene expression analysis and, typically, heuristics are used. *Simulated Annealing* is well-known and has been applied by Lukashin and Fuchs (2001) and Bryan et al. (2006), using $TWCV$ and *H-Score* (Eq. 3.3), respectively, as the fitness function, E, to be minimised. At each stage of the process, gene vectors are randomly chosen and moved to a new random cluster. E is evaluated for each move and the new assignment is accepted if E is improved or with a probability of $e^{-\frac{E^{new}-E^{old}}{T}}$ otherwise. The "temperature", T, controls readi-

	Structure	**Proximity**	**Input**	**Other**
SOM	None	Distance	Number of output neurons, Learning rate	Careful consideration of initialisation weights
SOTA	Binary Tree	Distance	Threshold β	
DGSOT	N-ary Tree	Distance	Thresholds and $L.\beta$, α	Corrects for misclusterings
GHTSOM	Each node triangular SOM, arranged in Tree structure	Distance	Minimal requirement - learning rate	

Table 3.3: Summary of Neural Network techniques presented.

ness of the system to accept the poorer situation by chance, enabling the algorithm to avoid local minima. As the search continues, T, is gradually reduced according to an *annealing schedule*, and ultimately achieves the global minimum, where the annealing schedule parameters dictate performance and speed of the search. Choice of initial temperature T_i governs convergence time and size of search space, (increased/decreased in the case of high/low T respectively). Similarly for search termination, (final effective T_F). The user must specify the rate at which T approaches T_F, which must be slow enough to guarantee a global minimum, as well as the number of swaps of gene vectors between clusters allowed in an iteration.

To determine K, a randomisation procedure is used, (Lukashin and Fuchs, 2001), to determine cut-off threshold for the distance, D, between two gene vectors in a single cluster. It is also necessary to determine P, the probability of accepting false positives, (e.g. $P = 0.05$). Simulated annealing is then applied for different numbers of clusters, until the weighted average fraction of incorrect gene vector pairs reaches the P-value.

3.3.3 Biclustering Methods

Biclustering methods are important for gene expression data analysis (Section 3.2.1), so we deviate slightly from the known taxonomy of clustering algorithms, (Jain et al., 1999) to consider algorithms which adopt biclustering strategy 'in isolation'.

'Cheng and Church' Algorithm and FLOC

This algorithm, (Cheng and Church, 2000, : adapted from Hartigan (1972)) obtains H-scores, (Eq. 3.3, Fig. 3.2, (Madeira and Oliveira, 2004)) of the sub-matrices of the gene expression matrix. This method is initialised for the entire gene expression matrix and considers a sub-matrix to be a bi-cluster if $H(I, J) < \delta$ for some $\delta \geq 0$, (user defined). Each row and column of the original matrix is thus tested for deletion. Once a sub-matrix is determined to be a bi-cluster, its values are "masked" with random numbers in the initial gene expression matrix. Masking bi-clusters prevents the algorithm from repeatedly finding the same sub-matrices, but there is a substantial risk that this replacement will interfere with the discovery of future bi-clusters. To overcome this problem of random interference, *Flexible Overlapped biClustering (FLOC)* was developed - a generalised model of Cheng and Church incorporating null values, (Yang et al., 2003). FLOC constrains the clusters to both a low mean residue score *and* a minimum occupancy threshold of α, $0 \leq \alpha \leq 1$ (user defined).

Note: FLOC does not require pre-processing for imputation of missing values. Both these bi-clustering algorithms find coherent groups (Section 3.3.1) and permit overlapping.

Coupled Two Way Clustering(CTWC)

Getz et al. (2000) adopted an iterative approach to find biclusters in the data. Firstly, all samples are clustered (using any clustering algorithm) using all genes, and vice versa to identify stable clusters of genes and of samples. Then, each gene cluster is used to cluster all stable sample clusters, and vice versa. Whenever a clustering operation generates a new stable subcluster, it is recorded and its members are used in the next iterative step. The process stops when no new stable clusters (that exceed a minimum size) are generated. This method is, of course, reliant on that algorithm used to cluster the data in the first place, and inherits the limitations of that method.

Inputs are determined by the clustering algorithm used. Unlike many others, the method adopted in Getz et al. (2000) (Superparamagnetic Clustering, (Blatt et al., 1996)) does not require the number of clusters to be specified before hand. However, if another clustering approach was used, for e.g. K-means, then the number of clusters would have to be specified. The results are also dependent on initial clusters found and the order used to cluster the samples etc. in the latter stages of the algorithm.

The Plaid Model

The Plaid Model, (Lazzeroni and Owen, 2000), (Section 3.3.1), assumes that bi-clusters can be generated using a statistical model and aims to identify the parameter distribution that best fits the available data, by minimising the error sum of squares for the k^{th} bi-cluster assuming that $k-1$ bi-clusters have already been identified.

Explicitly, it seeks to minimise for the whole matrix:

$$Q = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}(Z_{ij} - \theta_{ijk}\rho_{ik}\kappa_{jk})^2 \tag{3.4}$$

where Z_{ij} is the residual after deducting $k-1$ previous layers,

$$Z_{ij} = a_{ij} - \sum_{k=0}^{K-1}\theta_{ijk}\rho_{ik}\kappa_{jk} \tag{3.5}$$

Parameters (θ_{ijk}, ρ_{ik} and κ_{jk}, defined previously, Section 3.3.1) are estimated for each layer and for each value in the matrix, and are updated iteratively, providing refined estimates of μ_k, α_{ik} and β_{jk}, (Fig: 3.2(C)) and ρ_{ik} and κ_{jk} to minimise Q, (Lazzeroni and Owen, 2000).

The importance of a layer is defined by:

$$\delta_k^2 = \sum_{i=1}^{n}\sum_{j=1}^{p}\rho_{ik}\kappa_{jk}\theta_{ijk}^2 \tag{3.6}$$

To evaluate the significance of the residual matrix, Z is randomly permuted and tested for importance. If δ_k^2 is significantly better than δ_{random}^2, k is reported as a bi-cluster. The algorithm stops when the residual matrix Z retains only noise, again *with the advantage that the user does not need to specify the number of clusters beforehand*. Another important property of this method is that statistical evaluation is intrinsic in the results.

3.3.4 Graph Theoretic Methods

Methods selected and detailed below also span agglomerative/partitive, global/local structures, crisp/fuzzy cluster membership etc. and, as with all others, have an

	Proximity	**Deterministic Stochastic**	**Clusters**	**Other**
SA	Depends on application	Stochastic	Depends on application	Specification of Annealing Schedule
CC	Additive Model	Deterministic	Overlapping, partial bi-clusters	δ, random interference
FLOC	Additive Model	Deterministic	Overlapping, partial	bi-clusters α and δ to specify. Overcomes random interference, allows missing values.
CTWC	Depends on application	Deterministic	Crisp, partial bi-clusters	Results dependent on underlying algorithm used and on initial clusters found.
Plaid	Additive Model	Deterministic	Overlapping, partial, bi-clusters	Values seen as sum of contributions to bi-clusters

Table 3.4: Summary of biclustering techniques presented.

objective function to evaluate the clusters found. We isolate graph theoretic approaches for special consideration, due to their intuitive nature, interpretation and visualisation of gene expression matrices but also because the modelling paradigm is well established for analysis of other complex systems which exhibit similar properties to those found for gene expression.

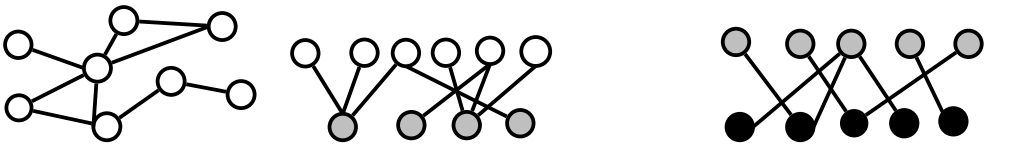
Data Organisation in Gene Expression Context

At a basic level, a graph, $G = (V, E)$, consists of two parts: a set of nodes (or nodes), V, and a set of edges (or links, connections), $(v_i, v_j) \in E$, which captures the concept of a 'relationship' between nodes v_i and v_j, with $v_i, v_j \in V$.

A graph can be undirected (where edge (a,b) is considered to be the same as edge (b,a) and is only recorded once) or directed, (distinction is made between an edge (a,b) and an edge (b,a), i.e. both edges are recorded). A graph can be *complete*, with an edge between every node in a graph, or *incomplete*, with an edge between a subset of nodes in a graph - also referred to as *strongly connected, connected or weakly connected*.

Definition/Representation of graph edges and nodes in terms of gene expression data is a first consideration. Nodes may be viewed as either genes or samples, linked by edges. In the gene expression context an edge can have different meanings - depending on how the graph is designed. For example, it could indicate a similarity above some threshold of a similarity measure, or adherence to some cohesion model. We will see further examples of these in this Section, and later in Chapter 5 when we describe a new method for extraction of a graph from a gene expression matrix. A graph may be constructed from 'gene' nodes alone, with an edge which represents similarity of expression. Alternatively, the graph could be

constructed from both 'gene' and 'sample' nodes. Further, (Tanay et al., 2002) used additional *'condition'* nodes and sample nodes, where *condition* nodes represented moderate and strong changes in gene expression, involvement in a protein complex etc.



(a) Gene Expression Graph containing only gene nodes

(b) Gene Expression Graph with both gene and sample nodes

(c) Gene Expression Graph according to Tanay et al. (2002) with both condition and sample nodes

Figure 3.5: Gene Expression Graphs. (a) is an example of an *one-mode graph* (one type of node). (b) and (c) are examples of *bi-partite graphs* (two types a node).

While benefits are associated with each representation, it is important to note that the final clusters/identification achieved follows from this choice. In graphical context, a cluster is defined to be *connected components*, i.e. a group of nodes that are connected to one another, but that have no connection to nodes outside the group. Examples include, **contiguity-based clusters** (connecting objects within a specified distance on one another), or a **clique** (a set of nodes in a graph that are completely connected to each other).

Given a dataset X, we construct an adjacency matrix, A, where $a_{ij} \in A$, $a_{ij} = f(i, j)$. The adjacency matrix can be a square $n \times n$ matrix (a *One Mode Representation*, Fig. 3.5a) or a $n \times p$ matrix (a *Bipartite Representation*, Fig. 3.5b and 3.5c), $n =$ cardinality of one set of nodes (e.g. number of genes), $p =$ cardinality of alternative set of nodes (e.g. number of samples). For some clustering schemes

each pair of nodes is connected by an edge with an assigned weight, $f(i,j)$, and thus the adjacency matrix records *edge weights* (and the structure is referred to as a weighted graph). In other instances, $f(i,j) \in \{0,1\}$: edges are constrained to exist between objects i and j only if $f(i,j) = 1$, and thus the adjacency matrix records whether an edge exists of not[1]. Additionally, each node may be assigned a weight[2], $g(v_i)$ and this information can be used in any clustering process. The clustering problem is thus explicitly presented in terms of graph theoretical properties.

One Mode Representation

A graph, $G = (V, E)$, is considered to be in one mode if an edge can exist between any two nodes, $v \in V$. A gene expression dataset can be modelled in this way where nodes in the graph represent genes and an edge exists between two gene nodes if they show common expression (e.g. measured as a distance function).

Bipartite Representation

A graph, $G = (\top, \bot, E)$, is considered to be bipartite if there are two disjoint subsets of nodes, $\top$, $\bot$, and there is no edge between two nodes in the same subset. A gene expression dataset can be modelled in this way, where $\top$ nodes represent genes and $\bot$ nodes represent samples. An edge $w_{ij} \in W$ is the weight matrix, were $w_{ij} \neq 0$ if there is an edge between $i \in \top$ and $j \in \bot$.

We consider applying these modelling ideas more specifically to gene expression data in Chapter 5.

[1] In computer science there are, of course, alternative methods of recording edge lists such as linked lists etc. However to introduce the concept we use only the idea of an adjacency matrix.

[2] the weight of a node can represent such things as importance in hub, e.g. internet networks, telephone networks etc.

Graph Theoretic Clustering Options

Graph theoretic approaches generally have gained ground recently in analysing large complex datasets and we consider in brief the principle options for the gene expression case.

• The *Cluster Affinity Search Technique (CAST)*, (Ben-Dor et al., 1999), models data as an undirected graph, $G = (V, E)$, where $\{V, E\}$ is the set representing $\{genes, \textit{similar expression}\}$. The model assumes that there is an ideal *clique* graph, (see 3.3.4), $H = (U, E)$, which represents the ideal input gene expression dataset, while data to be clustered is a "contamination" of the ideal graph H by random errors. In a clique graph each clique represents a cluster. For a pair of genes in G, the model assumes that an edge/non-edge was assigned incorrectly, with a probability of α. The true clustering of G is assumed to be that which requires fewer edge changes to generate H.

CAST uses an *affinity* (similarity) measure, either binary or real valued, to assign a node to a cluster. This must be above a threshold, t (user-defined, determining size and number of clusters). The affinity of a node v to a cluster, is the sum of affinities over all objects currently in the cluster, so v has high affinity with i if $affinity(x) > t|i|$, and low affinity otherwise. The CAST algorithm alternates between adding high affinity elements and removing low affinity elements, finding clusters one at a time. A disadvantage of this approach is that the result is dependent on the order of input, as once initial cluster structure is obtained, a node v is moved to that cluster for which it has a higher affinity value.

• *CLICK*, (Sharan and Shamir, 2000), builds on the work of Hartuv et al. (2000), introducing a probabilistic model for edge weighting. Pairwise similarity measures between genes are assumed to be normally distributed: between 'mates'

($N(\mu_T, \sigma_T^2)$), and between 'non-mates' ((μ_F, σ_F^2)), where $\mu_T > \mu_F$. These parameters can be estimated via Expectation Maximisation methods, (Dempster et al., 1977). The weight of an edge is derived from the similarity measure between the two gene vectors, and reflects the probability that $i(\in V)$ and $j(\in V)$ are mates, specifically that:

$$w_{ij} = log\frac{p_{mates}\sigma_F}{(1-p_{mates})\sigma_T} + \frac{(S_{ij}-\mu_F)^2}{2\sigma_F^2} - \frac{(S_{ij}-\mu_T)^2}{2\sigma_T^2} \tag{3.7}$$

Edges with weights below a user defined non-negative threshold are omitted from the graph. The graph is thus partitioned using a *minimum weight cut* algorithm, (Hartuv et al., 2000), and the weight of a partition is the average of these edge weights.

• The *Statistical Algorithmic Method for Bi-cluster Analysis (SAMBA)* method finds bi-clusters based on the coherent evolution model (Section 3.3.1), (Tanay et al., 2002). Firstly, the gene expression matrix is modelled as a bipartite graph, $G = (U, V, E)$, where U is the set of sample nodes, $U \cap V = \varnothing$ and an edge (u, v) only exists between $v \in V$ and $u \in U$ *iff* there is a significant change in expression level of gene v, w.r.t. to its normal level, in sample u. Key to SAMBA is the scoring scheme for a bi-cluster, corresponding to its statistical significance, where a weight is assigned to a given edge, (u, v), based on the log-likelihood of getting that weight by chance, (Tanay et al., 2002):

$$log\frac{P_c}{P_{(u,v)}} > 0 \text{ for edges and, } log\frac{(1-P_c)}{(1-P_{(u,v)})} < 0 \text{ for non-edges.} \tag{3.8}$$

The probability $P_{(u,v)}$ is the fraction of random bipartite graphs, with degree sequence identical to G, that contain edge (u, v) (and can be estimated using Monte-

Carlo methods). P_c is a constant probability $> max_{(u,v)\in UxV} P_{(u,v)}$. Assigning these weights to the edges and non-edges in the graph, the statistical significance of a subgraph H can be calculated:

$$logL(H) = \sum_{(u,v)\in E} log\frac{P_c}{P_{(u,v)}} + \sum_{(u,v)\in \overline{E}} \log\frac{1-P_c}{1-P_{(u,v)}} \tag{3.9}$$

and, given that the expected degree for each node is $\hat{d}_u = \sum_{v\in V} \phi(u,v)$, where $\phi(u,v)$ is the probability that u has an edge to v, the expected log-likelihood score of a subgraph is thus:

$$E(log(L(H))) = \sum_{u,v}(\phi(u,v) \times log\frac{P_c}{P_{u,v}} + (1-\phi(u,v)) \times log\frac{1-P_c}{1-P_{u,v}}) \tag{3.10}$$

The K heaviest (largest weight) sub-graphs for each node in G is found. Tanay et al. (2002) present two ways to calculate the weight of the resulting sub-graph.

(i) In the simpler model, bi-clusters, which reflect changes relative to normal expression level, without considering direction of change are sought.

(ii) The second model, focuses on *consistent bi-cliques*, targeting those samples which have the same or opposite effect on each of the genes.

	Mode	Proximity	Search	Other
CAST	One Mode	Similarity	Clique Graph	Parameters α and t. Finds global, complete, crisp clusters.
CLICK	One Mode	Distribution based on distance	Minimum weight cut	Stat. Sig. of clusters. EM to estimate parameters. Finds global, partial, crisp clusters.
SAMBA	Bi-Partite	Probability	Heuristic search of neighbours	Stat. sig. of clusters. Input P_c difficult to define. Finds partial overlapping biclusters.

Table 3.5: Summary of Graph theoretic methods presented.

Gene Expression Graphs and Random Graphs

Random graphs[3] are proposed as the simplest and most straight forward realisation of a complex network with no apparent design principles. The theory of random graphs lies at the intersection between graph theory and probability theory and are used to explore the existence of properties of real world graphs. First studied by Erdös and Rényi (1959), the model they propose starts with N nodes and connects every pair of nodes with probability p, creating a random graph with approximately $pN(N-1)/2$ edges distributed randomly. Growing interest in complex systems has prompted revision of this as a suitable model. Questions have been raised as to whether the real networks behind diverse complex systems, such as gene co-expression in the cell, are fundamentally random? Intuitively, the random network model is insufficient for gene expression, since we expect these systems to display some organizing principles, which at some level are encoded in their topology. If

[3]A random graph is a graph in which properties such as the number of nodes, edges and/or connections between are determined randomly.

the topology does deviate from that of random graphs, we need to develop tools and measurements to capture, in quantitative terms, the underlying organisational principles. In Chapter 5 we investigate the potential of using random graph models to compare to real gene expression graphs in order to highlight interesting organisational properties in real gene expression graphs.

Techniques for generating random graphs

There are a number of random graph models, here we list three.

- *Erdös-Rényi Model* There are two closely related variants of the Erdös-Rényi random model. In the first model, a graph, $G(n, M)$, is chosen uniformly at random from the collection of all graphs which have n nodes and M edges. In the second model, a graph $G(n, p)$ is constructed by connecting nodes randomly. Each edge is included in the graph with probability p, with the presence or absence of any two distinct edges in the graph being independent. An Erdös-Rènyi graph is not scale-free, (Erdös and Rényi, 1959), see below. Both of the two major assumptions of the G(n, p) model (that edges are independent and that each edge is equally likely) are unrealistic in modeling gene expression data.

- *Barabasi Albert Model* This technique generates scale-free random graphs. In scale-free networks, some nodes act as "highly-connected hubs", although most nodes are of small degree (number of edges incoming and outgoing from a node). Scale free network structure and dynamics are independent of, N, the number of nodes. Their defining characteristic is that the degree distribution follows a power law relationship defined by $P(k) \sim k^{-\gamma}$, where

the probability $P(k)$ that a node in the network connects to k other nodes is roughly proportional to $k^{-\gamma}$. The coefficient γ varies from 2 to 3 for most real networks, or, in some cases, between 1 and 2, (Barabasi and Albert, 1999).

- *Watts Strogatz Small World Model* This model produces graphs with small-world properties, i.e. where most nodes are not neighbours of one another, but most nodes can be reached from every other by a small number of hops or steps. Watts and Strogatz (1998) noted that graphs could be classified according to their clustering coefficient (the average proportion of edges shared by nodes neighbours in a graph) and mean shortest path length. Watts and Strogatz proposed a simple model of random graphs with (i) a small average shortest path and (ii) a large clustering coefficient .

Using graphical techniques to extract meaningful information from biological data is an intuitive and popular method. Bi-partite graphs representation, in particular, capture essential properties of the gene expression dataset, allowing for the extraction of biclusters, however, the alternative one-mode gene expression graph can also be been considered. This approach was used in Yip and Horvath (2007), where an edge existed between two genes nodes if they show *similar* expression across *all* samples, (measured e.g by a distance function). Yip and Horvath (2007) did not explicitly to investigate graphs with weighted edges. Carlson et al. (2006) and Zhang and Horvath (2005) studied weighted one-mode gene expression graphs (creating the graphs using a threshold on a similarity function) using classical network analysis techniques. Again, an edge existed between two genes nodes if they show *similar* expression across *all* samples. They found that, for the datasets tested, the gene networks generated by this technique were scale free, following a power-law distribution or an exponential. These results, however, are dependent the global

method used to extract a graph from the gene expression data. Analysis of complex weighted networks was also considered in Saramaki et al. (2007). All the aforementioned authors did not expand the investigation to the weighting scheme itself. It is important to recognise *the essential role a weighting scheme itself plays on cluster determination* from gene expression graphs, and thus it is important to investigate the intrinsic nature of these schemes in isolation.

3.4 Discussion

Despite shortcomings, application of clustering methods to gene expression data has proven to be of immense value, providing insight on cell regulation, as well as on disease characterisation. Nevertheless, not all clustering methods are equally valuable for high dimensional gene expression data. Recognition that well-known, simple clustering techniques, such as K-Means and Hierarchical clustering, do not capture complex local structure, has led to investigation of other options. In particular, bi-clustering has gained considerable recent popularity. Indications to date are that these methods provide increased sensitivity at local structure level in discovery of meaningful biological patterns.

An inherent problem with exploratory clustering is *ab initio* knowledge of K, the number of clusters. Consequently, those methods for gene expression analysis which do not need K specified *ab initio* have an advantage. Most algorithms seek empirically to determine this at run time, but derive complicated thresholds that may not make sense in the context of gene expression data. There is a risk that determination of these thresholds is not a one step process but requires testing and validation of clusters produced. A comprehensive survey of robust cluster validation and evaluation methods is given (Handl et al., 2005) but it seems clear that a

requirement for *information-driven* clustering is emerging, which integrates cluster and meta-information, (Choi et al., 2004; Liu et al., 2004; Kasturi and Acharya, 2005; Gamberoni et al., 2006; Kustra and Zagdanski, 2006). This provides a basis for validation, independent of the current problem, as well as interpretation of clustering results.

3.5 Summary

Cluster analysis, applied to gene expression data, aims to highlight meaningful patterns for gene co-regulation. The evidence suggests that, while commonly applied, agglomerative and partitive techniques are *insufficiently powerful* given the high dimensionality and nature of the data. While further testing on non-standard and diverse data sets is required, comparative assessment and numerical evidence, to date, supports the view that bi-clustering methods, although computationally expensive, offer better interpretation in terms of data features and local structure. While the limitations of commonly-used algorithms are well documented in the literature, adoption by the bioinformatics community of new (and hybrid) techniques, developed specifically for gene expression analysis has been slow, mainly due to the increased algorithmic complexity required. This would be catalysed by more transparent guidelines and increased availability in specialised software and public dataset repositories.

CHAPTER 4

CLUSTER ANALYSIS: A PRACTICAL EVALUATION

In the ***Assessment*** process, a clustering achieved is tested for specific properties. Assessment measures are rarely a fixed set but together form a diagnostic toolkit targeted at improving the clustering process. In general, clustering techniques optimise some form of this measure as a criterion function. ***Evaluation*** of clustering thus involves the synthesis of a number of assessment measures used to gauge final cluster quality in order to form an objective final judgement on the most suitable technique for the dataset involved. In this chapter we use these basic principals to investigate the applicability of clustering algorithms to gene expression data. The approach is to consider a series of measures which assess cluster quality on the basis of biological realism amongst other criteria. It also involves comparison of these measures between clustering algorithms and for different datasets. We evaluate clusterings obtained with selected algorithms[1] identified in Chapter 3. Clusters obtained from real and synthetic datasets are compared between algorithms. We demonstrate the fact that, with so many classification criteria for clustering, no one

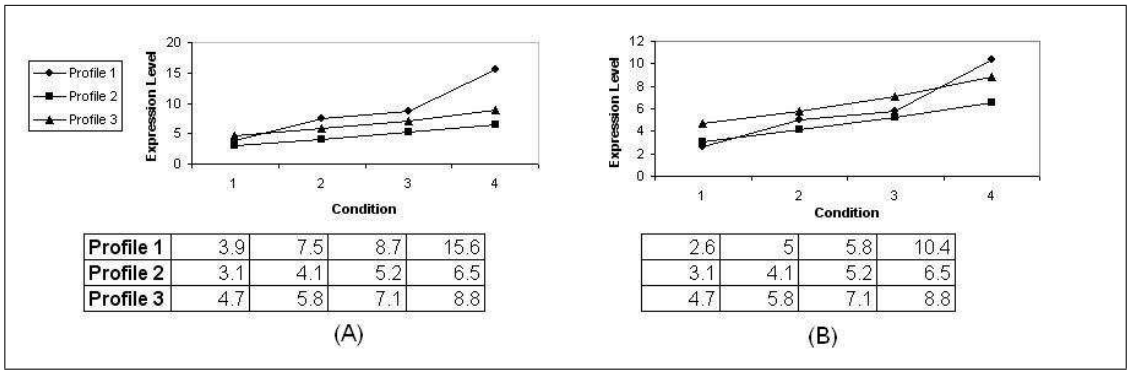
[1]Reporting for all algorithms is prohibitively detailed. Our aim is to give a 'flavour' of techniques and their validation, by applying selected algorithms from each group in the Jain et al. (1999) taxonomy.

algorithm is good for all datasets, so that a preliminary review of the most appropriate methods is essential. Further, it is also strongly advocated that no single method or interpretation is sufficient and that recognition of valid clusters is frequently indefinite or misleading.

4.1 Introduction

Evaluation of clustering requires both internal/external assessments of clusters obtained, and comparison between algorithms. This is a complicated area for gene expression data due to its unique properties, due to the fact that little may be known about the data before hand. Many clustering algorithms are designed to be exploratory; so that clusters (dependent on given criteria) found *will* discover "a structure" which, while meaningful in the context of these, may yet fail to be optimal or even biologically realistic. Algorithms are inherently biased, as properties of clusters reflect built-in clustering criteria, while structures found are not usually the same for different algorithms. For example, with regard to the K-Means criterion the "best" structure is one that minimises the sum of squared errors (MacQueen, 1967), while for the Cheng and Church biclustering algorithm (Cheng and Church, 2000), it is that which minimises the Mean Residue Score (MRS, Eq. 3.3). The two assessments are generally not directly comparable, as the former highlights *global patterns* in the data and the latter *local patterns*, (Section 3.2). Also, large deviations from the mean may correspond to large residue scores, but this is not always the case. For example, Fig. 4.1(a), and the corresponding table, highlight a simple case of three genes in a cluster across four samples. According to the K-Means criterion, the cluster (Euclidean and centroid) distance is approximately 11.02, while MRS $= 0$. In the second case (Figure 4.1(b)) the scale of profile 1 was reduced by

one third. In this case the cluster distance is decreased to 7.91 (indicating a better cluster), while the MRS is increased to 0.0168 (indicating an inferior cluster). Obviously, interpretation of cluster results relies at some level on subjective choice with regard to the assessment criterion to use. The greater the need, therefore, for independent validation by integration of findings with metadata. Subjective evaluation, (based on experience, background knowledge, expected results), even for low dimensional data, is non-trivial at best, but becomes increasingly difficult for high dimensional gene expression data. From these considerations, it is clear that cluster validation is critical for algorithm development and verification of results, with the latter usually based on a manual, lengthy and subjective exploration process.



Profile 1	3.9	7.5	8.7	15.6
Profile 2	3.1	4.1	5.2	6.5
Profile 3	4.7	5.8	7.1	8.8

(A)

2.6	5	5.8	10.4
3.1	4.1	5.2	6.5
4.7	5.8	7.1	8.8

(B)

Figure 4.1: Cluster (B) has profile 1 scaled down by one third.

4.2 Assessment Methods

Once a non-random structure is distinguished in the data, a technique that finds the "best" structure is desirable - a vague expression - since 'best' may refer to novelty, stability, size, suitability, and is again dependent on the nature of the analysis and experimental purpose, and so on. Some non-trivial considerations might include: whether the method is exploratory or predictive (with results used as the

foundation for further investigations), whether *all* samples and genes should be grouped, whether a novel structure is to be assessed, and so on. Objective assessment measures of clustering quality fall into two categories - external and internal - summarized in Table 4.1.

Internal Measures:

These measures use only information from the clustering result and the dataset itself to assess cluster quality. Properties considered include:

- *Compactness* - intra-cluster homogeneity e.g. assessment of average or maximum pairwise intra-cluster distances, average or maximum centroid-based similarities.

- *Separation* - inter-cluster distance, e.g. average weighted distance where the distance between clusters can be computed as the distance between their centroids, or as the minimum distance between data items of each, e.g. the minimum distance between any two clusters.

- *Connectedness* - to what degree data items are grouped with their nearest neighbours in the data space, i.e. also known as connectivity (Handl et al., 2005).

- *Combinations* - as compactness usually improves with the number of clusters and separation usually deteriorates, linear/non-linear combination measures to assess both can be used, e.g. the SD-validity index (Halkidi et al., 2000), Dunn Indices (Dunn, 1974), Davies-Bouldin Index (Davies and Bouldin, 1979), Silhouette Index (Rousseeuw, 1987), C-Index (Hubert and Schultz, 1976).

An equivalent measure for fuzzy clusterings includes the *Xie-Beni* index (Xie and Beni, 1991).

- *Fuzziness* - applicable only to fuzzy partitions, these measures assesses sharing of membership between clusters, included are the *partition coefficient* and *partition entropy* (Bezdek, 1973, 1974).

- *Stability/Predictive Power* - based on repeatedly resampling the original data, this measures the consistency of the results, which in turn provides an estimate of the significance of the clusters obtained from the original dataset e.g. Ben-Hur et al. (2002) and Levine and Domany (2001). The jackknife approach, Yeung et al. (2001), forms clusters based on $p - 1$ (with $p =$ number of samples) and uses the remaining sample to assess predictive power of the algorithm i.e. Figure Of Merit (FOM). Stability can also be assessed by perturbing data and comparing the different clusters found with the original partition, using external indices, (Bittner et al., 2000; Kerr and Churchill, 2001; Li and Wong, 2001).

- *Preservation of distance information* - the degree to which the distance information in the original data is preserved in a clustering and typically used for hierarchical clustering. Here, a *cophenetic distance matrix* is an $N \times N$ matrix where each entry (i, j) records the level at which the data items i and j are grouped in the same cluster for the first time. The preservation is usually assessed using the *cophenetic correlation coefficient* i.e. the correlation between the entries in the cophenetic distance matrix and the original distance matrix, (Sokal and Rohlf, 1962).

External Measures (Supervised):

These refer to assessments which reference external information (e.g. class labels or clusterings from alternative algorithms). The comparison of clusterings with external class labels is of critical importance as it provides a great deal of information to the user. For instance, genes that show similar pattern across clusters do not necessarily indicate the same pathway or similar function but *could do*. Examples of the properties key to external measurements include:

- *Agreement with metadata* - For biological function information included, in the gene list for each cluster, a more complete picture is inevitably provided of the dataset and the success of the technique. A number of functional annotation databases are available. The Gene Ontology database, (Ashburner et al., 2000), for example, provides a structured vocabulary that describes the role of genes and proteins in all organisms. The database is organised into three hierarchical ontologies: biological process, molecular function and cellular component. Several tools have been developed for batch retrieval of GO annotations for a list of genes (e.g. tools DAVID, (Dennis et al., 2003), Babelomics, (Al-Shahrour et al., 2005) or Machaon CVE (Bolshakova et al., 2006)). Statistically relevant GO terms can be used to investigate the properties shared by a set of genes. These tools typically use comprehensive measures, like the F-measure (introduced by Rijsbergen (1975)), or hypergeometric tests, (Falcon and Gentleman, 2007), to test the significance of cluster *purity*, (the fraction of the cluster taken up by the predominant class label) and *completeness*, (fraction of items in a class grouped in the current cluster). This assessment can be adapted for partially annotated datasets, by only including that fraction of genes that are annotated in the calculation of

the measure. This facilitates the transition from data collection to biological meaning by providing a template of relevant biological patterns in gene lists.

- *Agreement between clusterings (cluster runs)* -In a simulation dataset, the true partition is known, and the performance of a technique can be assessed in terms of its clustering similarity to the true partition. There are several such indices to measure this in the literature, (Fridlyand and Dudoit, 2001). Most popular is the *Rand Index* (RI), (Rand, 1971) and a number of variations of this exist, including the *adjusted RI*, (Hubert and Arabie, 1985) and the *weighted RI* (Thalamuthu et al., 2006). In general, these determine the similarity between two partitions as a function of positive and negative agreement in pairwise cluster assignments. The *Jaccard coefficient*, (Jaccard, 1908), looks at similarity as a function of only the positive agreements in pairwise cluster assignments. Most of these can also only be used where a single class label is unequivocally assigned to a data item, thus are *inappropriate for fuzzy clusterings or overlapping clusters*, although a fuzzy extension has been proposed recently for the Rand Index, (Campello, 2007) Note: these measures can also be used where the gold standard is not known, to assess relative similarity of two clusterings obtained.

Table 4.1: Summary of Evaluation Measures, categorised into Internal and External. *Assess'* refers to which property the measure it assessing, *Measure*, refers to the popular name for the measure in literature, *G, L, C,* $C_{>1}$, *F* indicates if the measure is suitable for assessing Global, Local structures, Crisp, Crisp in more than one cluster, and Fuzzy Membership respectively, *Max/Min* indicates whether the measure should be maximised or minimised, *Bounds* refers to the [maximum, minimum] possible value of the result.

Assessment Measures									
Category	**Assess'**	**Measure**	**G**	**L**	**C**	**C$_{>1}$**	**F**	**Max /Min**	**Bounds**
Internal	Connectedness	Conn	✓		✓			Min	$[0, \infty)$
	Compactness	Intra-cluster Distance	✓	✓	✓	✓		Min	$[0, \infty)$
	Separation	Inter-cluster Distance	✓	✓	✓			Max	$[0, \infty)$
	Combination	SD-validity Index	✓		✓			Min	$[0, \infty)$
		Dunn Indices	✓		✓			Max	$[0, \infty)$
		Davies-Bouldin Index	✓		✓			Min	$[0, \infty)$
		Silhouette Index	✓		✓			Max	$[-1, 1]$
		C-Index	✓		✓			Min	$[0, 1]$
		Xie-Bien Index	✓				✓	Min	$[0, \infty)$
	Fuzziness	Partition Coefficient	✓		✓		✓	Max	$[0, \infty)$
		Partition Entropy	✓		✓		✓	Min	$[0, \infty)$
	Continued on Next Page...								

	Table 4.1 – *Continued*								
Category	**Assess'**	**Measure**	**G**	**L**	**C**	**$C_{>1}$**	**F**	**Max /Min**	**Bounds**
	Stability	Cluster Overlaps (Average Proportion Non-overlap, Average Distance, Average Distance between Means)	✓	✓	✓			Min	$[0, 1]$
		Figure Of Merit	✓	✓	✓	✓		Min	$[0, \infty)$
	Distance Preservation	Cophentic Correlation	✓		✓			Max	$[-1, 1]$
External	Purity	Biological Homogeneity Index	✓	✓	✓	✓		Max	$[0, 1]$
	Completeness	Biological Stability Index	✓		✓	✓		Max.	$[0, 1]$
	Reconstruction of structure	Adjusted Rand Index	✓	✓	✓	✓		Max	$[0, 1]$
		Fuzzy Rand Index	✓	✓	✓		✓	Max	$[0, 1]$
		Jaccard Coefficient	✓	✓	✓	✓		Max	$[0, 1]$
		Hubert Γ Statistic	✓	✓	✓	✓		Max	$[-1, 1]$

These metrics are usually highly dependent on the number of clusters as an input parameter, (discussed in Chapter 3). The 'natural' number of clusters in the data depends on which clustering criterion are used in the algorithm and is not fixed between algorithms. For example, according to the K-Means criterion the optimal

number of clusters for a particular dataset may be 5, while for CLICK it may be 10 for the same dataset. The 'optimal' number of clusters depends on the dataset and algorithm so that absolute choice is difficult. Assessment measures are also biased. The *compactness* index, e.g. is biased towards a large number of clusters, while the *separation* index is biased towards a small number of clusters. Formulae for each of the assessments can be found in Appendix A.

4.3 Tools and Packages

R is a powerful statistical computing language and environment, available for download from CRAN (`http://CRAN.R-project.org/`). This, and associated contributed packages, were used extensively for this analysis of clustering algorithms and evaluation methods. Primarily, esoteric packages, provided as part of the *Bioconductor* open source and open development project, (Gentleman et al., 2004), were employed. As a result of the Bioconductor project hundreds of packages for the analysis of gene expression data are publicaly available and, as the project is open development, updated regularly. Navigation and understanding of this abundance of contributed packages presents a serious challenge even for the experienced user.

For development of software for this analysis we used a number of contributed packages that perform cluster validation and which are available from CRAN or Bioconductor (`http://www.bioconductor.org`). These include packages *clv*, (Nieweglowski., 2008), *clValid*, (Brock et al., 2008), *e1071*, (Dimitriadou et al., 2008), *clusterSim*, (Dudek, 2008) and *biclust*, (Kaiser et al., 2007). Not all assessment measures that we wanted to use were available in any package on CRAN or Bioconductor, and for those we implemented our own functions. These included

the C-Index, (Hubert and Schultz, 1976), and a form of the Xie-Beni index, (Xie and Beni, 1991).

4.4 A Framework for Evaluation

Each of the above measures assesses different properties of a clustering, but does not tell us which 'weigh' more heavily. The evaluation framework here provides a *road-map* to guide final judgement of quality and applicability. Steps in the evaluation framework, (Figure 4.2) are:

1. To understand the dataset properties (Chapter 2).
2. To understand the properties, (biases and cluster types found preferentially) of the chosen clustering algorithm (Chapter 3).
3. To make an educated guess for initial input parameters
4. To apply clustering algorithm.
5. For a range of input parameters to the clustering algorithm, to analyse internal assessment measures. Determine which type of internal measures are appropriate for the algorithm e.g. does it take into account fuzzy memberships, or overlapping clusters etc.?
6. To use optimal input parameters based on internal validation.
7. To apply external assessment measures.

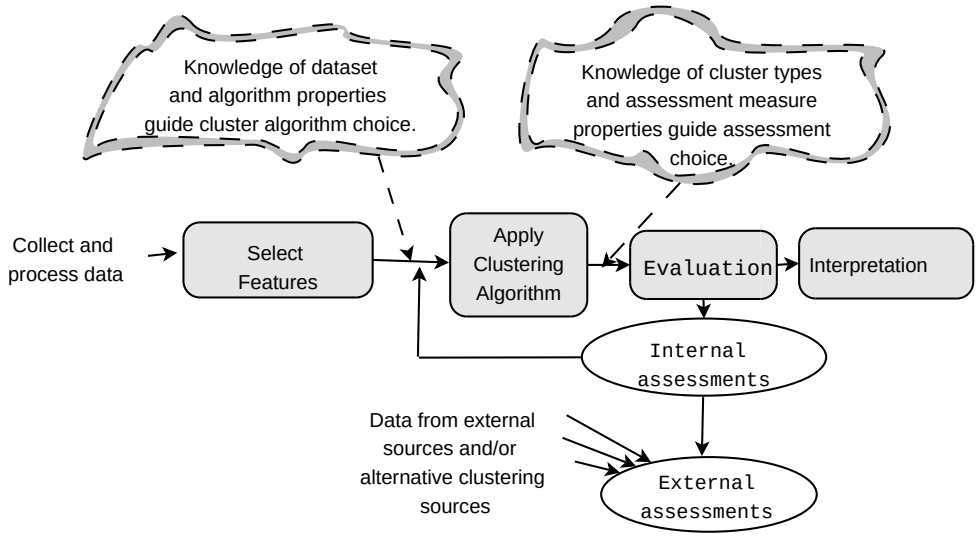


Figure 4.2: Evaluation Work-flow

4.5 Datasets

Table 4.2 details the range of benchmark datasets used to evaluate clustering techniques. The range of datasets was chosen to reflect different platforms, experimental design, number and type of samples and genes. Details of datasets used can be found in Appendix B.

4.5.1 Creation of synthetic datasets

Synthetic datasets allow us to test assessment measures with a known partition. A two-step process was used to create two synthetic datasets, where the number, size and type of clusters could be controlled. The two-step process involved:

1. Data generated according to artificial patterns, such that the true class of each gene is known for 11 classes of various sizes, for which all genes in a class

Author	Experiment	Type	Number of genes	Number of Samples
Alizadeth	Lymphoma	Spotted	4026 (3019)	96
Alon	Colon Cancer	Oligonucleotide	2000 (2000)	62
Cho	Yeast Cycle	Oligonucleotide	6601 (3000)	17
Gash	Yeast Stress	Spotted	6152 (3120)	173
Golub	Leukemia	Oligonucleotide	7129 (3571)	72
Hsiao	Human tissue	Oligonucleotide	7070 (2115)	59
Spellman	Yeast Cycle	Spotted	6178 (3049)	82
Stegmaier	Leukemia	Oligonucleotide	22283 (3145)	22
West	Breast Cancer	Oligonucleotide	7129 (3332)	49
Synthetic 1	-	-	3000	60
Synthetic 2	-	-	3000	90

Table 4.2: Test datasets used for analysis. The value in brackets represents the number of genes after filtering

have identical patterns before error is added. Genes not contributing to the programmed pattern, so technically 'irrelevant' for cluster formation, were also included.

2. Error was added to the synthetic patterns, aimed to control the level of biological noise cluster each class (and hence, the signal-to-noise ratio), such that classes were less separable when affected by a higher error. Errors were added randomly, (uniform distribution), so that the signal-to-noise ratio was 20 at maximum.

Artificial patterns were created to reflect patterns of real gene expression datasets, based on the analysis in Chapter 2. The artificial patterns were encoded as follows, (refer to Figure 4.3 and for exact details on functions for artificial patterns see Appendix B):

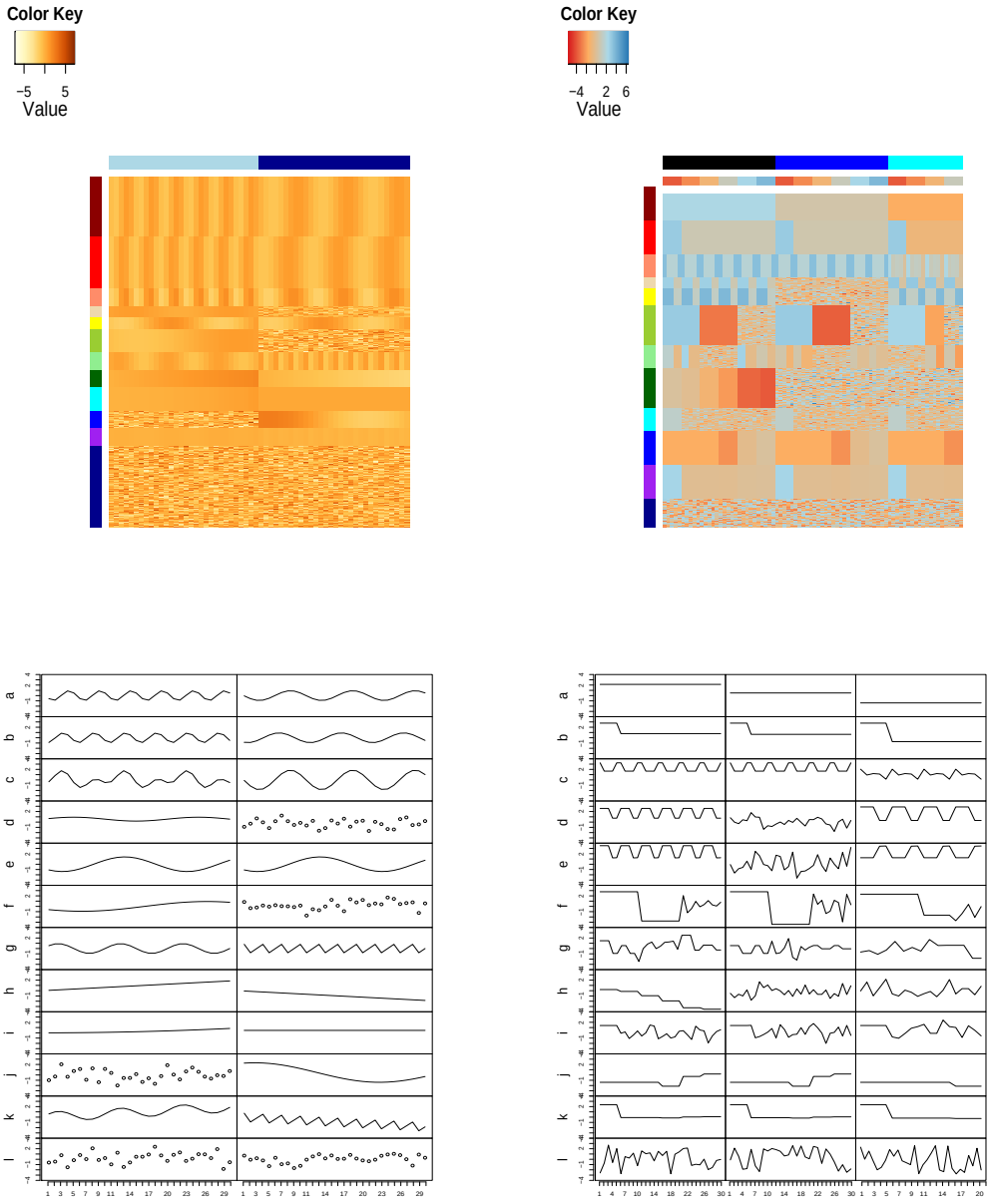
- Synthetic dataset A was designed to represent data derived from two time-course experiments, over 30 time points, creating a 60×3000 dataset, (see

Figure 4.3 (a)). The 11×2 groups were primarily modelled using $Asin(2\pi f + \theta)$, for various values of A = amplitude, f = frequency and θ = phase. Note that apart from pattern 'e', $A_{\text{time series 1}} \neq A_{\text{time series 2}}$, $f_{\text{time series 1}} \neq f_{\text{time series 2}}$ and $\theta_{\text{time series 1}} \neq \theta_{\text{time series 2}}$, and that, in some groups (d and f, Fig. 4.3,(a)) time series 1 shows definite structure, where gene values are random in time series 2 and vice verse (group j Fig. 4.3(a)). For exact details on functions see Appendix B.

- Synthetic dataset B was designed to represent data derived from an experiment on phenotypic samples, with 3 'treatment groups'. In the first two groups samples relate to 6 individuals, with 5 tests carried out on each sample ($2 \times (6 \times 5) = 60$ experiments). In the third group, samples relate to 4 individuals, again with 5 tests carried out on each sample ($4 \times 5 = 20$ experiments). This created a 80×3000 dataset. Patterns were created to either (i) affect all samples between treatment groups similarly, (ii) affect a subset of treatment groups similarly, (iii) affect a subset of individuals in a treatment group, or (iv) have a different effect on individuals between treatment groups. Note also that intentional overlap between groups was added to the patterns, (see Figure 4.3 (b)).

4.5.2 Random Datasets

Random datasets were created by randomly sampling expression values from the original dataset without replacement, thus destroying any cluster structure in the data while retaining all other properties.



(a) (b)

Figure 4.3: (a) Synthetic time series data, contained two time series over 30 time points. (b) Synthetic data from phenotypic samples. Contains data from 3 treatment groups, samples derived from 6 individuals in group 1 and 2, and 4 individuals in group 3, 5 tests carried out on each sample

4.5.3 Dataset pre-processing

An initial pre-processing step was applied to the 'real' datasets for testing, (Table 4.2) depending on whether the pre-processed data,due to the original authors, was available. Hence, the West, Golub, Hsiao, Stegmaier datasets were preprocessed as proposed by Speed (2000). This involves an initial step of applying a threshold on expression values of a floor of 100 and ceiling of 16000. The data was then log-transformed and finally, standardized to have zero mean and unit variance. The Alon and Cho datasets were standardized to have zero mean and unit variance.

A filter based on variation in gene expression was applied to all datasets to focus computations on informative genes across the samples. Each dataset was also filtered based on the percentage of missing values (if $> 25\%$ missing values occur in the gene vector it was excluded analysis).

4.6 Evaluation

In the following sections, we assess a number of clustering applications using a variety of assessment measures. These are intended to give a 'flavour' of techniques and assessments, and the reader should be aware that there are a number of permutations of input parameters etc. that could affect the final clustering. The intention is to assess the metrics and the performance of the metrics when applied to specific clustering algorithms.

4.6.1 Hierarchical application

The *agnes* implementation of hierarchical clustering (available in the *cluster* package of Bioconductor (Maechler et al., 2005)) was used to carry out this analysis.

As stated previously, the *cophenetic correlation* is often used to decide between various linkage methods when applying hierarchical clustering. Table 4.3 provides the *cophenetic correlation* of this clustering on the test datasets. In all cases, the distance matrix used for the clustering correlates most with the cophentic distance derived when average linkage was used. However, tests on random datasets suggest that the cophenetic correlation measure is, in fact, biased towards this linkage method. There is a higher degree of correlation for the synthetic datasets compared to the real datasets, most likely because groups in these datasets are less complex than in real cases.

The dendogram produced by the hierarchical clustering can, of course, be 'cut' at various levels to produce clusterings with different cluster numbers, K. The trend of the internal assessments (*SD-Validity, Dunn, Davies-Bouldin, Silhouette Index and C-Index*) across various values of K are given in Figures C.1 - C.9, Appendix C. For a *majority* of these internal assessments, the values remain roughly constant for the entire range of K, (for all linkage methods), indicating that, according to these measures, hierarchical clustering techniques do not identify a compact and well separated structure in the data.

For a minority of datasets, however, the internal indices did indicate an optimal number of clusters. Table 4.4 summarises optimal K, suggested by internal measures for selected datasets. The *SD-Validity* index for **single linkage**, was minimised at 6 and 4 clusters for the Cho and Stegmaier datasets respectively. This value was corroborated by the *Davies-Bouldin* index for the Cho dataset for single linkage, Fig. 4.4. Again, the *Dunn* index indicates 4 clusters for the Cho dataset using **average** and **ward linkage**, Fig. 4.5.

For Synthetic B dataset, (were 11 groups is optimal, excluding outliers) the *Dunn* index suggested 15 and 10 clusters optimal for **complete** and **ward** linkage

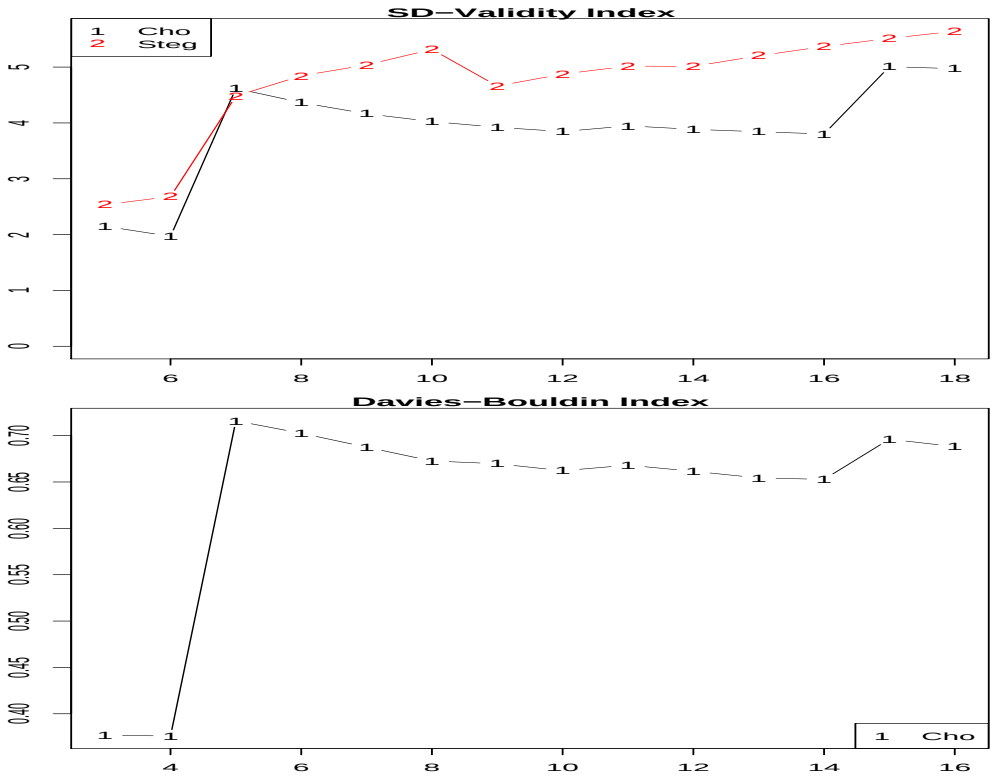


Figure 4.4: Hierarchical Assessment (Single) of Cho and Stegmaier datasets using Internal measures. The x-axis indicates the cluster number while the y-axis indicates the score obtained.

	Single	Complete	Average	Ward's
Aliz	0.363	0.211	0.485	0.178
Alon	0.245	0.304	0.544	0.428
Cho	0.296	0.332	0.558	0.335
Gasch	0.669	0.297	0.762	0.324
Golub	0.280	0.184	0.439	0.169
Hsiao	0.258	0.362	0.620	0.298
Spell	0.246	0.196	0.405	0.183
Steg	0.589	0.288	0.703	0.418
West	0.323	0.199	0.451	0.081
Synth. A	0.872	0.857	0.943	0.772
Synth. B	0.914	0.934	0.952	0.615
Random Datasets				
Aliz	0.005	0.104	0.129	0.057
Alon	0.014	0.126	0.146	0.091
Cho	0.011	0.147	0.198	0.145
Gasch	0.006	0.1	0.12	0.057
Golub	0.008	0.1	0.116	0.066
Hsiao	0.014	0.129	0.150	0.092
Spell	0.004	0.106	0.142	0.048
Steg	0.006	0.121	0.170	0.104
West	0.006	0.092	0.11	0.063
Synth. A	0.009	0.106	0.122	0.07
Synth. B	0.009	0.110	0.128	0.076

Table 4.3: Cophenetic correlation coefficient. Agglomerative Clustering Techniques, internal assessment using the Cophenetic Correlation Coefficient measure.

respectively, Fig. 4.6. For the same dataset, the *Silhouette* index is maximised at 15 and 9 clusters for **average** and **ward** linkage respectively, Fig. 4.8. For Synthetic A dataset, the majority of indices do not change across clusters, with the exception of the *Silhouette* index, which indicates that the optimal number of clusters is 8 and 12 for **complete** and **ward** linkage methods respectively. Indeed, the negative *Silhouette* index for all the *real* datasets indicates that the clustering is very poor and that on average genes are placed in the wrong cluster, for all K, Fig. 4.8. With the exception of **single** linkage, the trend of the *SD-Validity* index is to increase

		Dunn	Davies-Bouldin	SD-Validity	Silhouette	C-Index
Cho	Single	-	4	4	-	4
	Average	4	-	-	-	-
	Complete	-	-	-	-	-
	Ward	4	-	-	-	-
Steigmaier	Single	-	-	4	-	-
	Average	-	-	-	-	-
	Complete	-	-	-	-	-
	Ward	-	-	-	-	-
Synthetic A	Single	-	-	-	-	16†
	Average	-	-	-	-	-
	Complete	-	-	-	8	-
	Ward	-	-	-	12	7
Synthetic B	Single	-	-	-	-	16†
	Average	-	-	-	-	-
	Complete	15	-	-	15	11
	Ward	10	-	-	9	11

Table 4.4: K for selected datasets using HC Analysis. † - The value of C-index up to this point gradually declined.

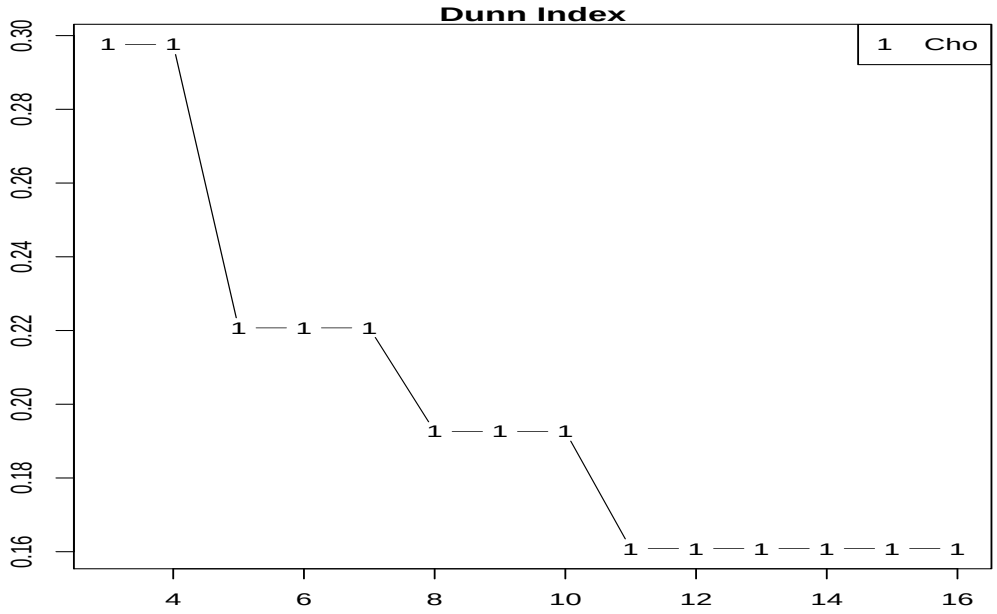


Figure 4.5: Hierarchical Assessment (Average) of Cho dataset using Internal measures. The x-axis indicates the cluster number while the y-axis indicates the score obtained.

with the number of clusters in all linkage methods, i.e. the cluster variance to. distance ratio deteriorates, Fig. 4.7. The major trend of the *C-Index* is to decrease with the number of clusters, Fig. C.5, Appendix C. For **ward** and **complete** linkage methods, the *C-Index* indicates that the optimal number of clusters is 11 (optimal number) for Sythentic dataset B, Fig. 4.9.

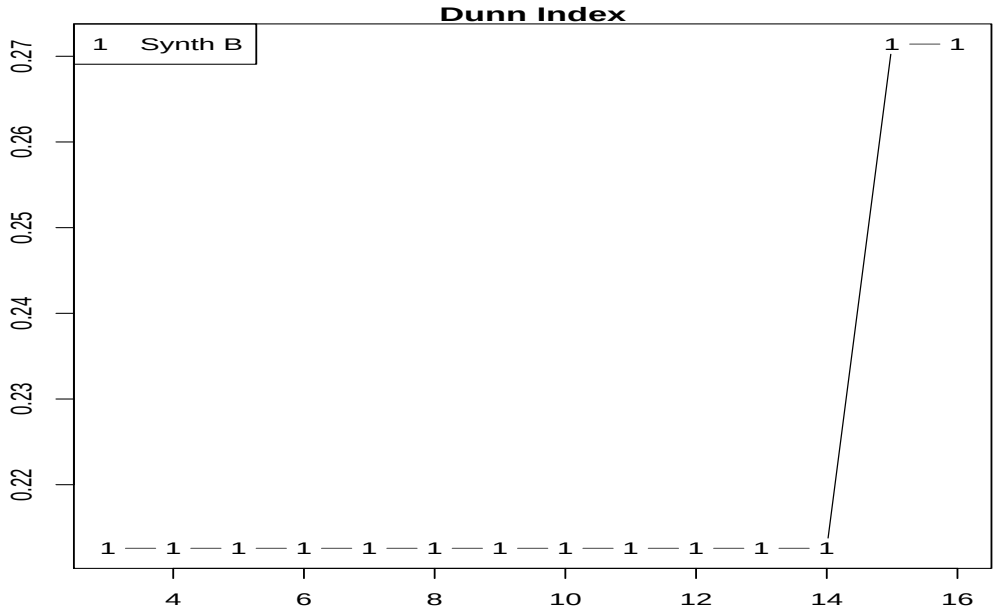


Figure 4.6: Hierarchical Assessment (Complete) of Synthetic dataset B using Dunn internal measures. The x-axis indicates the cluster number while the y-axis indicates the score obtained.

For a stability analysis[2] of the hierarchical clustering methods, shows that the behaviour of *ADM* measure across datasets is shown in Fig. 4.10. This measure generally increases with K and it highlights larger effects of K on the final clustering of Synthetic datasets A and B. For the **single** and **average** linkage methods this relationship is more gradual (range of values $0 - 1.2$) compared to **complete** and

[2]Note: these stability measures are extremely time consuming and memory intensive measures to calculate. This is in addition to the fact that they are already assessing a very memory intensive algorithm. For example, to assess the Synthetic dataset A took 344.24 hours of CPU execution time, while analysis of the Golub dataset took 744.56 hours of CPU time. This is a practical consideration of computation time and memory and is obviously a drawback of these stability measures.

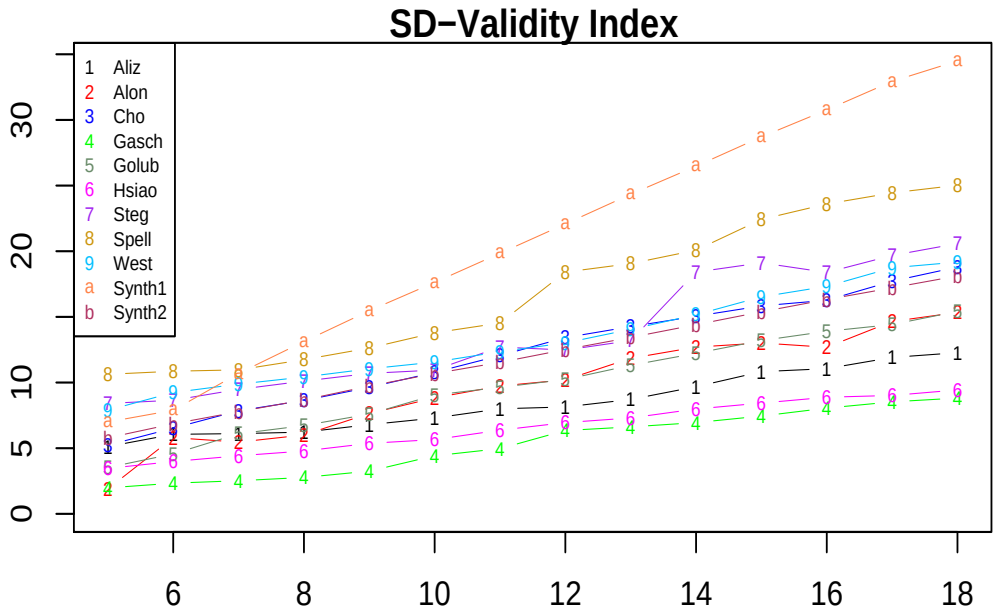


Figure 4.7: Hierarchical Assessment (Average) of datasets using SD-Validity Index. The x-axis indicates the cluster number while the y-axis indicates the score obtained.

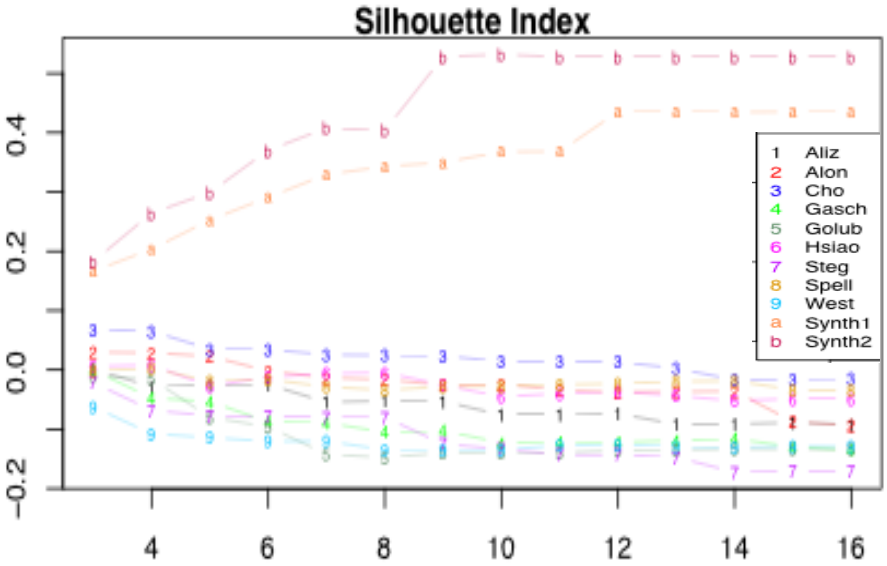


Figure 4.8: Hierarchical Assessment (Ward) of datasets using Silhouette Index. The x-axis indicates the cluster number while the y-axis indicates the score obtained. For the majority of 'real' datasets the silhouette is negative, indicating a bad clustering. Similar results were obtained for all linkage methods.

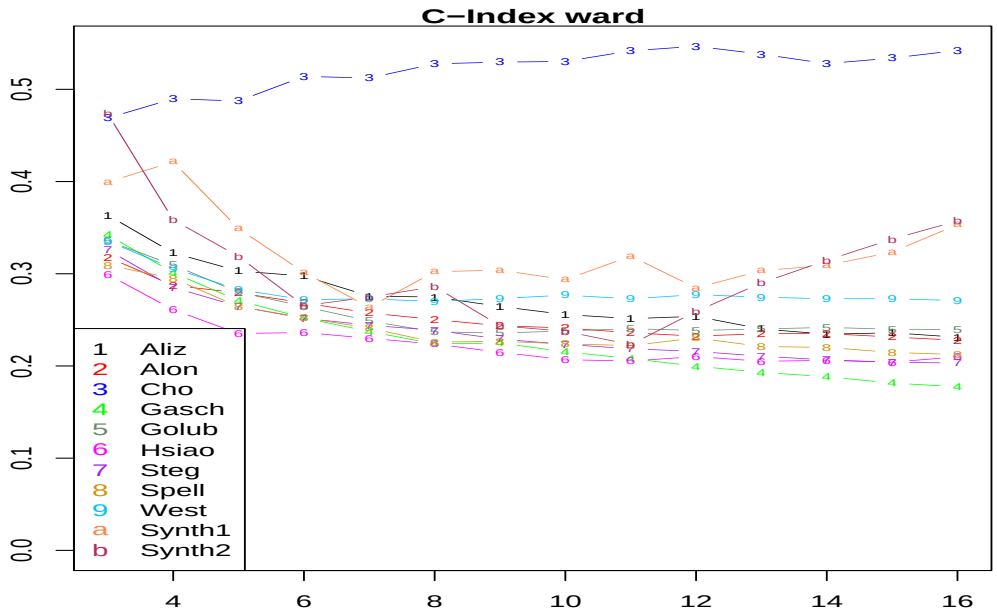


Figure 4.9: Hierarchical Assessment (Ward) of datasets using C-Index. The x-axis indicates the cluster number while the y-axis indicates the score obtained. The trend of the index is to decrease as the number of clusters increases. Similar results were found for all linkage methods.

ward (range $0 - 5.3$), Fig. 4.10.

Similar behaviour of *APN* across linkage methods is observed, whereby the value of this index increases with K. Like the *ADM* measure, this increase is more abrupt in the **complete** and **ward** linkage methods, Fig. 4.11. This indicates that as the number of clusters increases the consistency of results deteriorates. Exceptions include assessments of Synthetic datasets A and B, for **complete** and **ward** linkage.

Although there is a bias towards small K with these measures, some information can be obtained. For Synthetic dataset A the *ADM* index indicates 13 clusters for **ward** linkage. For the same dataset, the *ADM* and *APN* index indicate 9 clusters for **complete** linkage. For Synthetic B dataset the optimal number of clusters of 8 as indicated by the *ADM* and *APN* indices for **complete** linkage, while 17 clusters is indicated by the *AD* and *FOM* indices for the same linkage method (not shown). For **ward** linkage, all indices indicate 10 clusters for this dataset. For the Cho dataset, *ADM* and *APN* indicate 3 clusters for **average** linkage, while the same measures indicate 5 clusters for the Alon dataset for **ward** linkage. For other datasets, these assessments give little information. Table 4.5 summarises results of the stability measurements.

External assessment is crucial in gene expression analysis, because it lets the experimenter link data to knowledge. External assessment results, using the Biological Homogeneity (BHI) and Biological Stability (BSI) measures, is given in Figures 4.12 - 4.13. It is evident that the biological stability of the hierarchical clusters is maximised when K is small. The *BSI* index inspects the cluster consistency of genes grouped by similar biological function, for $K = 1$ stability $= 1$ and gradually decreases as K increases, i.e. the BSI index is biased towards a smaller number of clusters, Fig. 4.12. **Single** and **Average** linkage methods decrease linearly with K while **Complete** and **Ward** deterioration is more pronounced. The

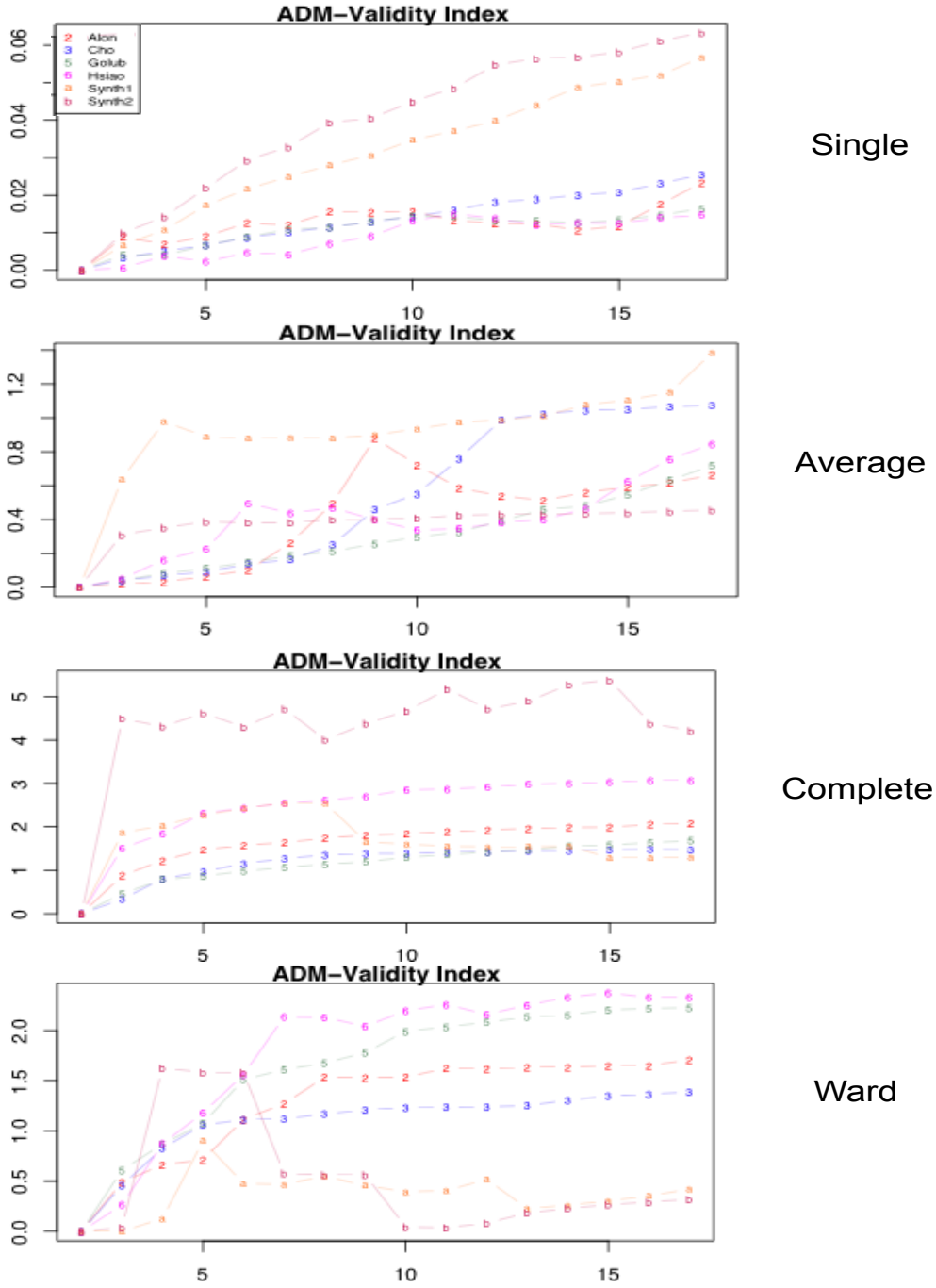


Figure 4.10: ADM measure across linkage methods. The x-axis indicates the cluster number while the y-axis indicates the score obtained. This measure highlights larger effects of K on the clustering of Synthetic datasets A and B (smaller values preferred).

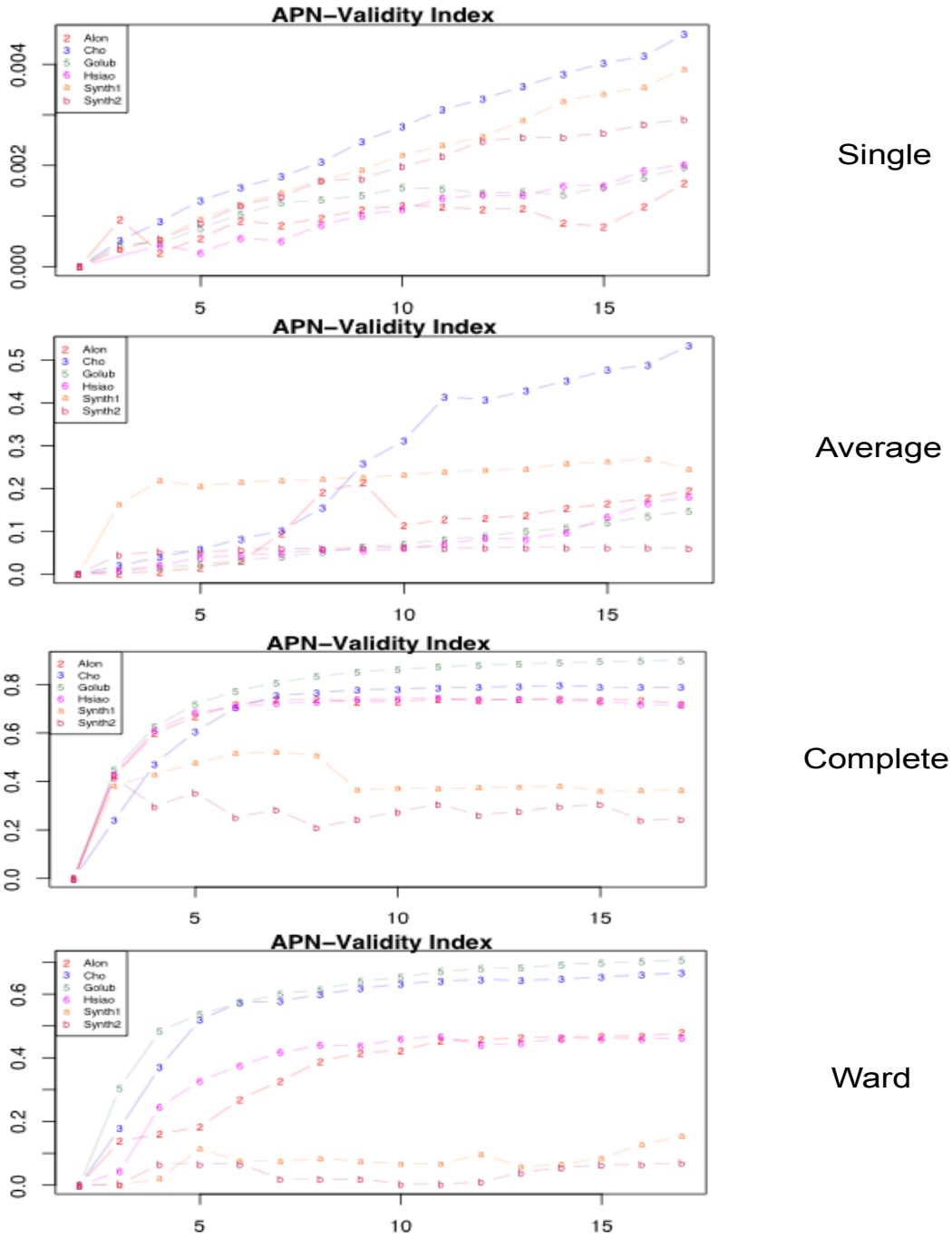


Figure 4.11: APM measure across linkage methods. The x-axis indicates the cluster number while the y-axis indicates the score obtained. For Ward and Complete linkage, similar to ADM behaviour, Synthetic datasets A and B show deviant behaviour compared to real datasets.

		ADM	AD	APN	FOM
Synthetic A	Single	4	-	-	-
	Average	-	-	-	-
	Complete	9	-	9	-
	Ward	13	12	-	-
Synthetic B	Single	-	-	-	-
	Average	-	-	-	-
	Complete	8	16	8	16
	Ward	10	10	10	10
Cho	Single	-	-	-	-
	Average	3	-	3	-
	Complete	-	-	-	-
	Ward	-	-	-	-
Alon	Single	-	-	-	-
	Average	5	-	5	5
	Complete	-	-	-	-
	Ward	-	-	-	-

Table 4.5: K selected by stability measures

BHI index measures whether genes placed in the same cluster belong to the same functional classes, through interrogation of GO ontologies. A value of unity indicates a biologically homogenous cluster. The maximum value for any technique is ~ 0.35, indicating that the clusters are not particularly homogenous. (This value was obtained for the Golub dataset, when $K = 2$, Fig. 4.13.)

Hierarchical Application Summary

Although there are some consistent predictions for K within the stability analysis, these are not consistent with internal assessment results. For example, stability indices indicate 10 clusters is optimal for the Synthetic dataset B dataset, while internal indices indications range from $9-16$. Again, the external assessments are not consistent with either the internal or stability assessment results, and suggest that

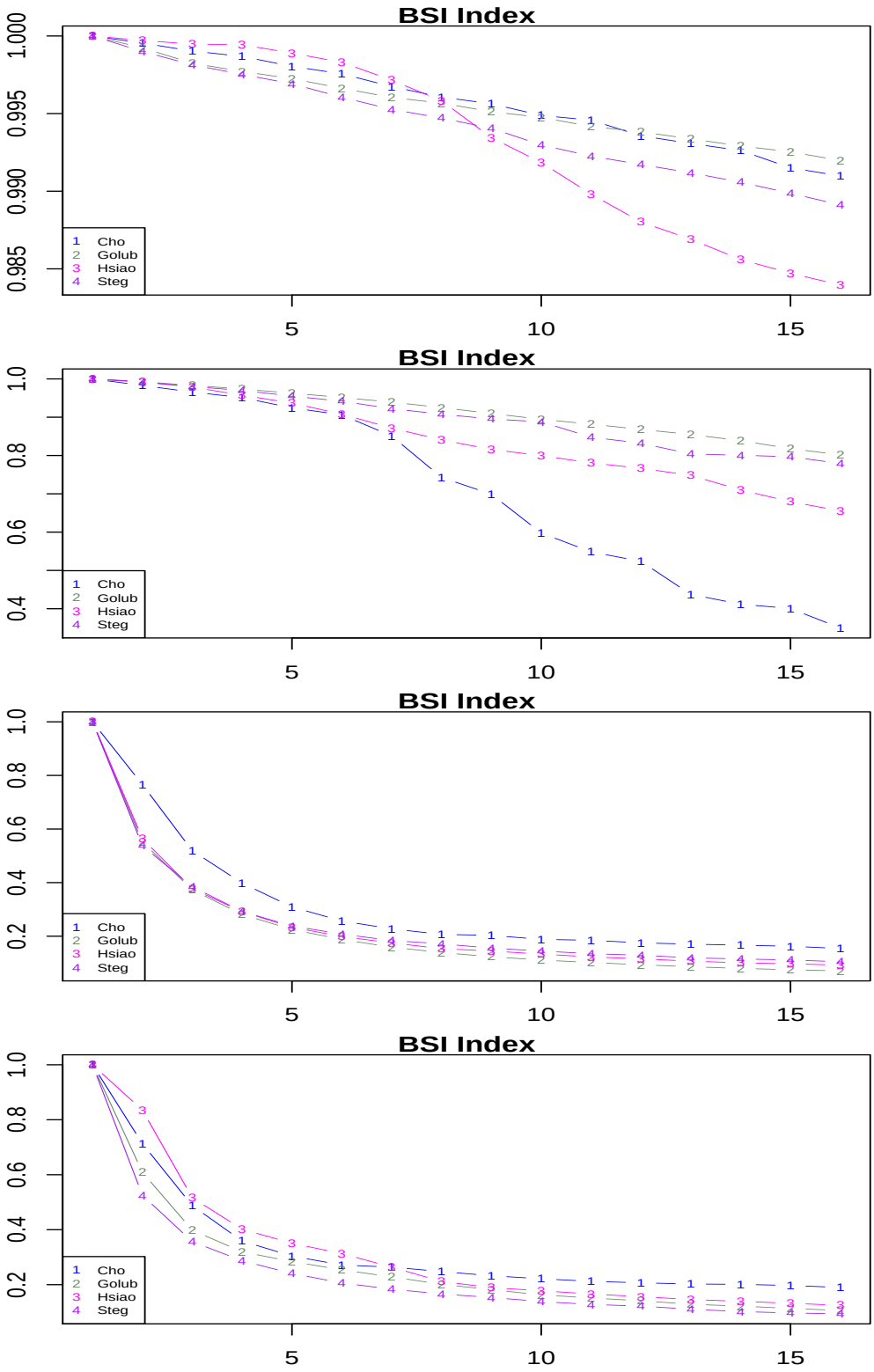


Figure 4.12: Hierarchical Assessment using BSI. Top - bottom: Single, Average, Complete, Ward. The x-axis indicates the cluster number while the y-axis indicates the score obtained. It is clear there is a bias towards small values of K.

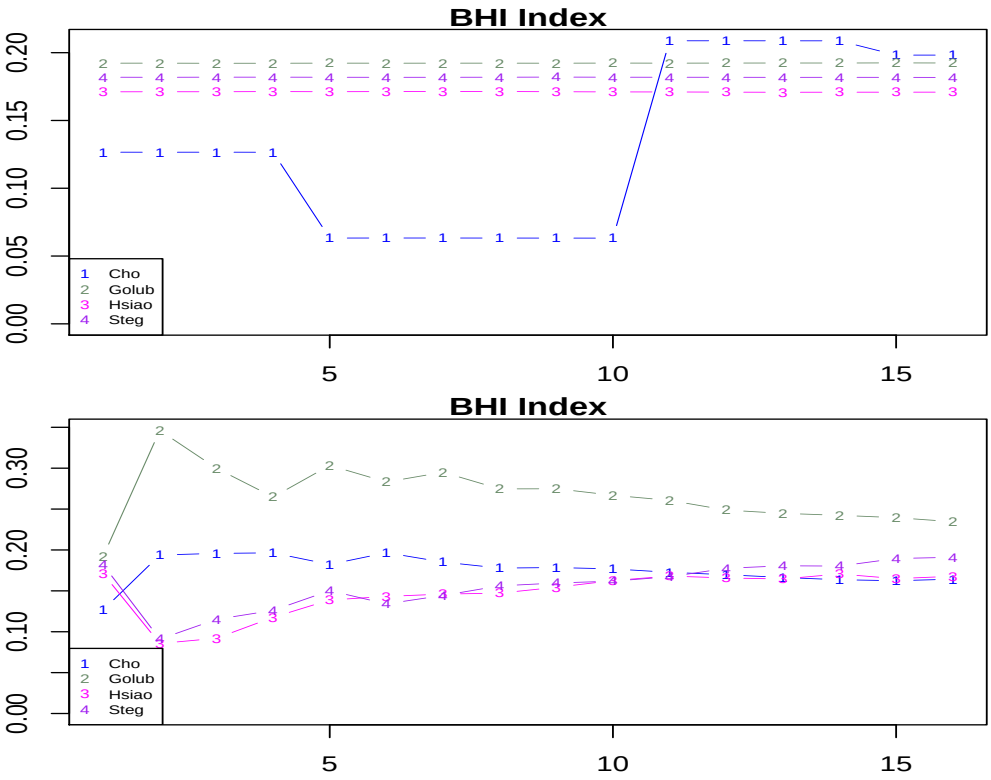


Figure 4.13: Hierarchical Assessment using BHI. Top - bottom: Single, Average. The x-axis indicates the cluster number while the y-axis indicates the score obtained.

clusters found by hierarchical clustering are not particularly homogenous. There are also obvious biases of indices (e.g. *SD-Validity, C-Index, APN, BSI*) towards a small number of clusters or linkage method (e.g. *APN* and *ADM* towards single linkage).

4.6.2 Partitive Application

K-Means and SOTA were selected as representative techniques for the partitive analysis. Internal assessments (or classification criteria) have a more obvious role here as these are typically used to provide the required parameter, K, to the algorithm.

K-Means Clustering

As discussed in Section 3.3.2, K-Means partitions the data by associating each gene vector with its nearest centroid and re-computing the cluster centroids. Thus, the K-Means technique is very sensitive to the random start positions and different executions will result in different clusterings. The K-Means algorithm available from the *cluster* package of Bioconductor (Maechler et al., 2005), was used for this analysis and starting positions were initalised using a technique which approximates the centres, based on the SPSS QuickCluster function, available in the *clusterSim* package (Dudek, 2008) of Bioconductor.

Figures C.12 - C.13 (Appendix C) summarises the internal assessments of clusterings obtained for various values of K. The *Dunn* index ranges from from ~ 0.03 to ~ 0.1 for $K = 3$ to 16, for all 'real' datasets, increasing slightly for Synthetic datasets A and B. This implies the datasets do *not* have compact and well separated

clusters when grouped according to the K-Means cost function. This assessment measure also indicates that the Synthetic B dataset (designed to have $K = 11$) has optimal compactness to separation ratio at $K = 4$. The *Davies-Bouldin* index indicates 6 for the same dataset, Fig. 4.14, (which measures the average error of each cluster group, rather than the maximum used by *Dunn*, thus incorrect grouping has less of an impact, see Appendix A for function details, (Davies and Bouldin, 1979)).

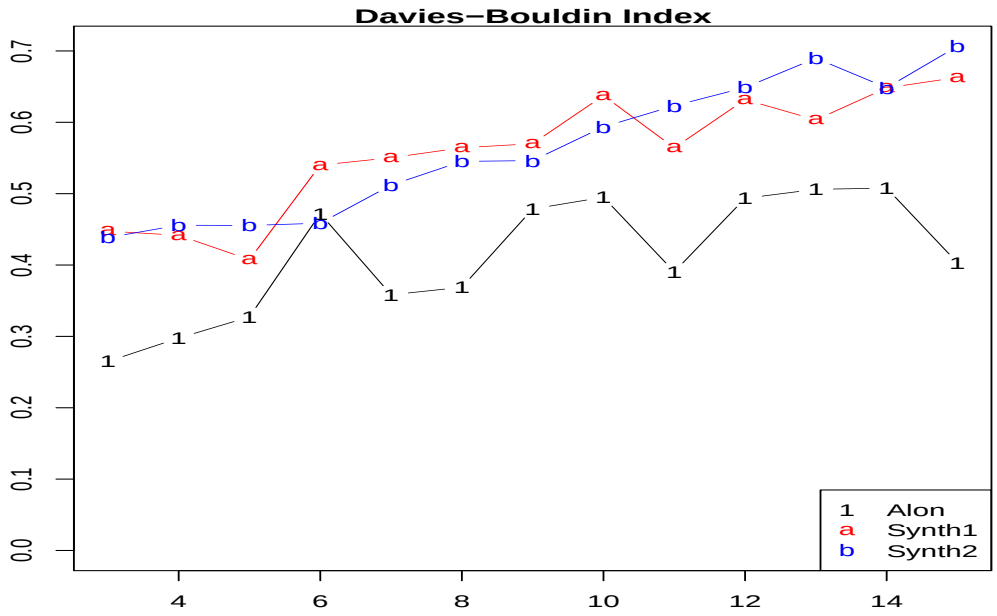


Figure 4.14: K-Means of selected datasets using Davies-Bouldin Index. The x-axis indicates the cluster number while the y-axis indicates the score obtained.

Overall, the range of values and trend for the *SD-Validity* index is similar to that obtained for the Hierarchical application - gradually increasing with K. For Synthetic A dataset, the *SD-Validity* index is minimised at $K = 5$, before it increases (deteriorates) rapidly as K increases. The *SD-Validity* index indicates $K = 3$ for the Alizadeth and the Spellman datasets, before the value increasing rapidly, Fig. 4.15.

Unlike the hierarchical clustering methods, the *Silhouette* index is > 0 for all

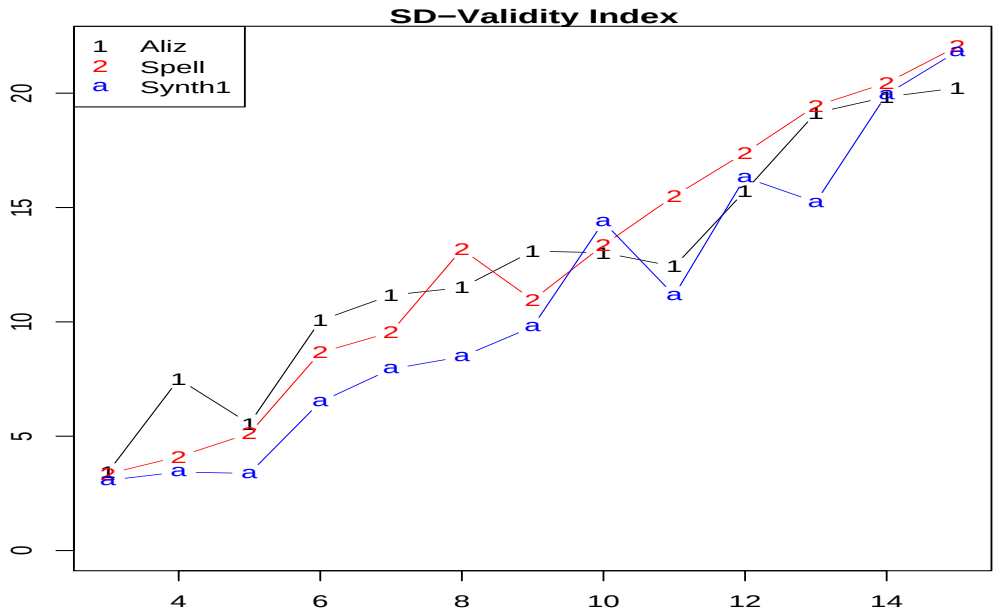


Figure 4.15: K-Means of selected datasets using SD-Validity Index. The x-axis indicates the cluster number while the y-axis indicates the score achieved obtained.

datasets, which suggests a better clustering, Fig. 4.16. This index indicates that the best choice of K is 14 for Synthetic B dataset. The *Silhouette* index indicates correctly that the optimal number of clusters for Synthetic A dataset is at $K = 11$. For the Alon, Hsiao, Steigmaier and West datasets, the *Silhouette* index indicates $K = 3$.

The *C-Index* is minimised at $K = 9$ for Synthetic dataset B, while for Synthetic dataset A the *C-Index* indicates the optimal $K = 5$. For the Alizadeth and Spellman datasets, the C-Index indicates $K = 14$ and 15 respectively. In contrast to hierarchical methods, when this index is used with the K-Means algorithm, no bias towards small K is indicated, Fig. C.13, Appendix C. Results of K-Means internal analysis are summarised in Table 4.6.

Unsurprisingly, the *Connectivity* measure increases with K, (when K is low, each gene vector will more likely be grouped with its nearest neighbour). This

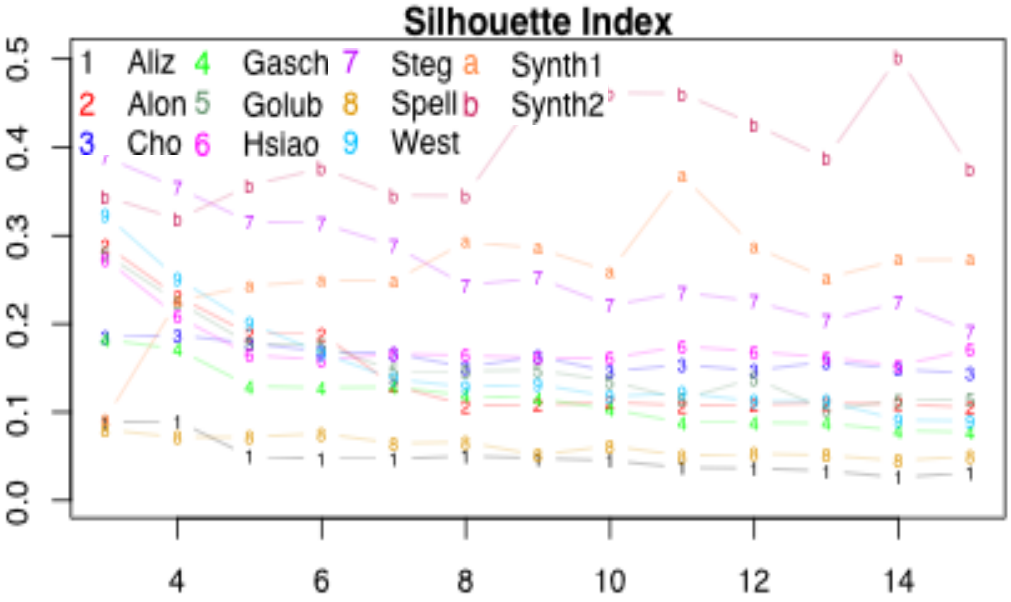


Figure 4.16: K-Means assessment using Silhouette Index. The x-axis indicates the cluster number while the y-axis indicates the score achieved obtained.

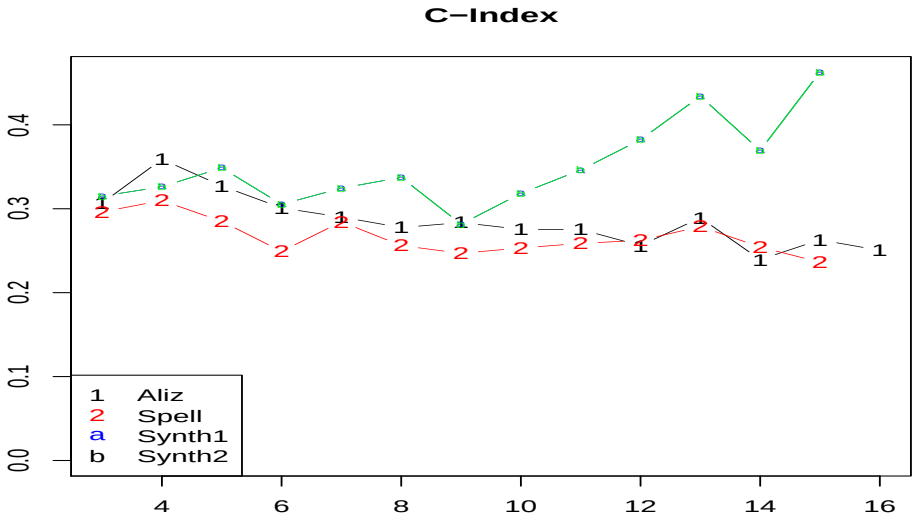


Figure 4.17: K-Means assessment using C-Index for selected datasets. The x-axis indicates the cluster number while the y-axis indicates the score achieved obtained.

indicated 4 clusters for Synthetic dataset B and 5 clusters for Synthetic dataset A. This measure is minimised for all other datasets at $K = 3$.

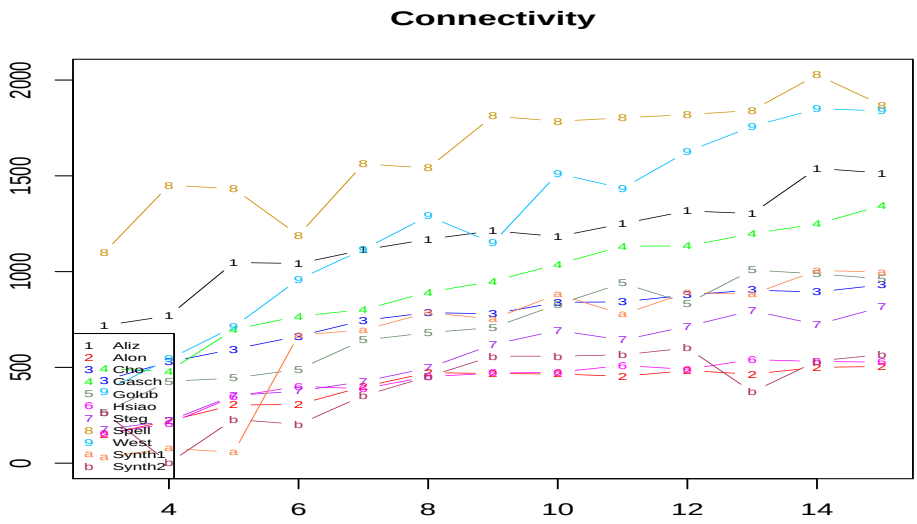


Figure 4.18: K-Means assessment using connectivity index for selected datasets. The x-axis indicates the cluster number while the y-axis indicates the score achieved obtained.

Stability analysis of the K-Means algorithm show that the *AD* measure does not determine K for the K-Means algorithm, although the value is minimised for both Synthetic datasets A and B at $K = 11$, Fig. 4.19. The *ADM* measure selects $K = 8$ for the Spellman dataset and $K = 4$ for the Gasch dataset. Values of this measure are erratic across all the values of K, Fig. 4.20. The *APN* index is minimised at $K = 6$ for Synthetic B dataset, $K = 4$ for the Gash dataset and $K = 5$ for the Alizadeth, Cho and Hsiao datasets, Fig. 4.21. The *FOM* index is not selective of K across any dataset, see Table 4.7 for summary of stability indices.

As for hierarchical clustering, *biological stability* (BSI) and *biological homogeneity* (BHI) were assessed for the clustering produced by K-Means. As for hierarchical clustering, moreover, the biological measures identify no discernible struc-

	Dunn	Davies-Bouldin	SD-Validity	Silhouette	C-Index
Synthetic A	-	-	5	11	5
Synthetic B	4	6	-	14	9
Alizadeth	-	-	-	3	14
Alon	-	-	3	-	-
Hsiao	-	-	-	3	-
Spellman	-	-	3	-	15
Steigmaier	-	-	-	3	-
West	-	-	-	3	-

Table 4.6: Optimal K identified by internal assessment measures for K-Means algorithm.

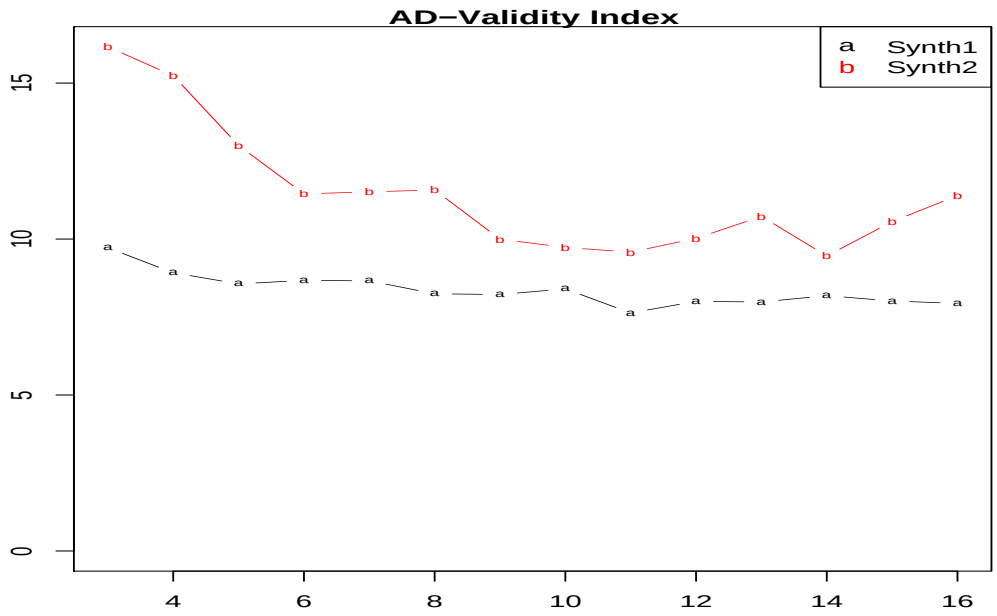


Figure 4.19: K-Means stability analysis using AD index. x-axis indicated cluster number and y-axis, score achieved.

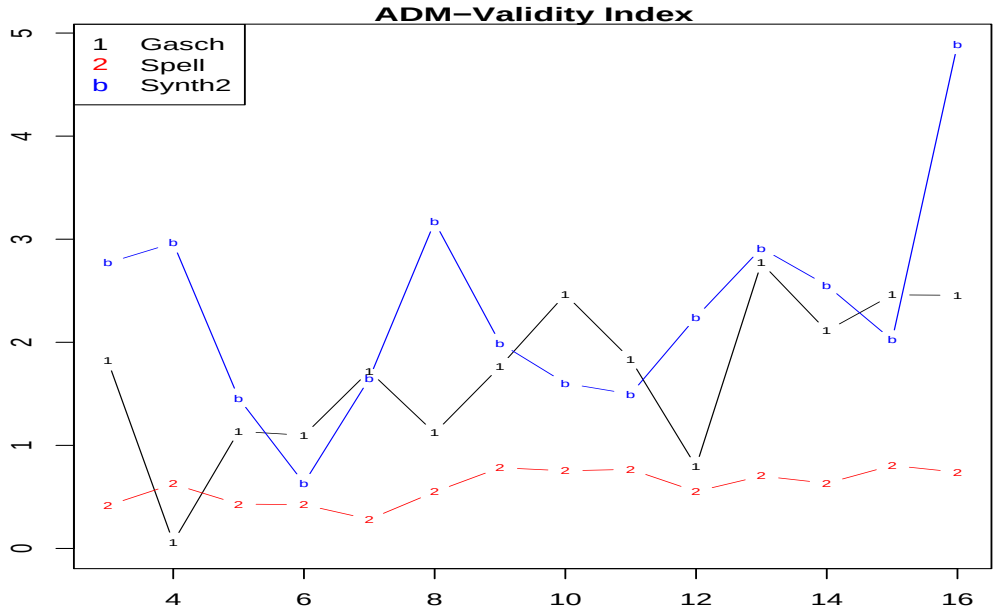


Figure 4.20: K-Means stability analysis using ADM index. x-axis indicated cluster number and y-axis, score achieved.

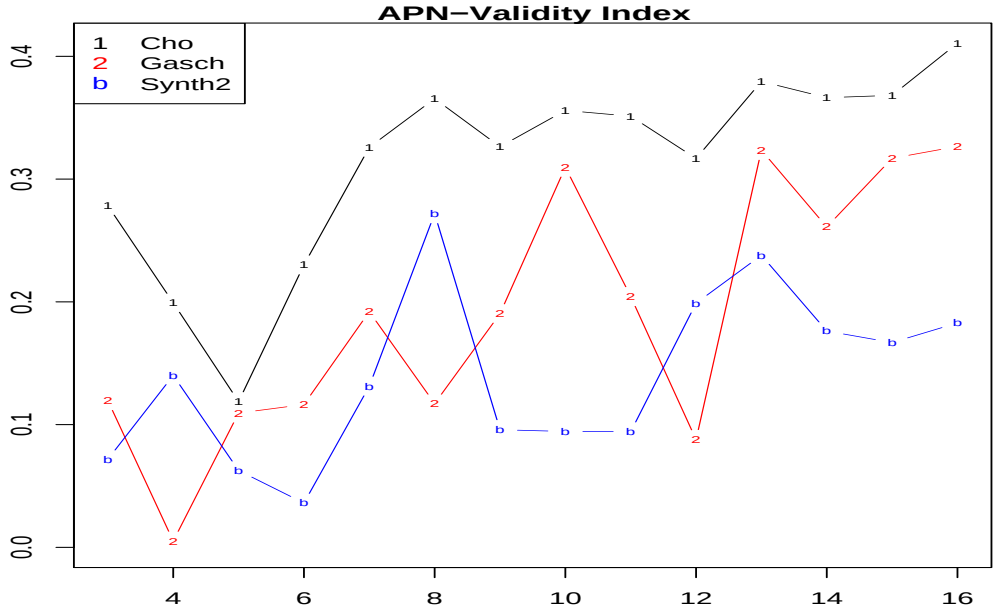


Figure 4.21: K-Means stability analysis using APN index. x-axis indicated cluster number and y-axis, score achieved.

	AD	ADM	APN	FOM
Synthetic A	11	-	-	-
Synthetic B	11	-	6	-
Spellman	-	8	-	-
Gasch	-	4	4	-
Alizadeth	-	-	5	-
Cho	-	-	5	-
Hsiao	-	-	5	-

Table 4.7: Stability summary of K-Means

ture in the datasets. *BSI* values decrease with the number of clusters, K, consistent with the trend for hierarchical clustering, Fig. 4.22. The value of *BHI* increases slightly for the Golub dataset at $K = 15$ (as opposed to $k = 2$ found by hierarchical average linkage for this dataset).

SOTA - Self Organising Tree Algorithm

Details of this technique were given in Section 3.3.2. A partitive technique, it uses self organising maps to discover hierarchical structure in the data. For this analysis, the Euclidean distance measure was used, with an ancestor height of two.

Figure C.15, Appendix C, summarises the internal assessment of the clusters obtained for various values of K. Again, the internal measurements tell us little about these data. The *Silhouette* index indicates the optimum choice of K is 6 for both the Synthetic B and Stegmaier datasets. For all other datasets, the optimum choice is indicated at K=3. Again, this index is > 0, for each datasets and across all K, similar to the results found for the K-Means algorithm, and again the maximum value found is for Synthetic B dataset. There is a large range in *Silhouette* values across K, *for each* dataset, however the trend is to decrease with increasing K (exception is Synthetic B dataset which increases with K and Synthetic A dataset

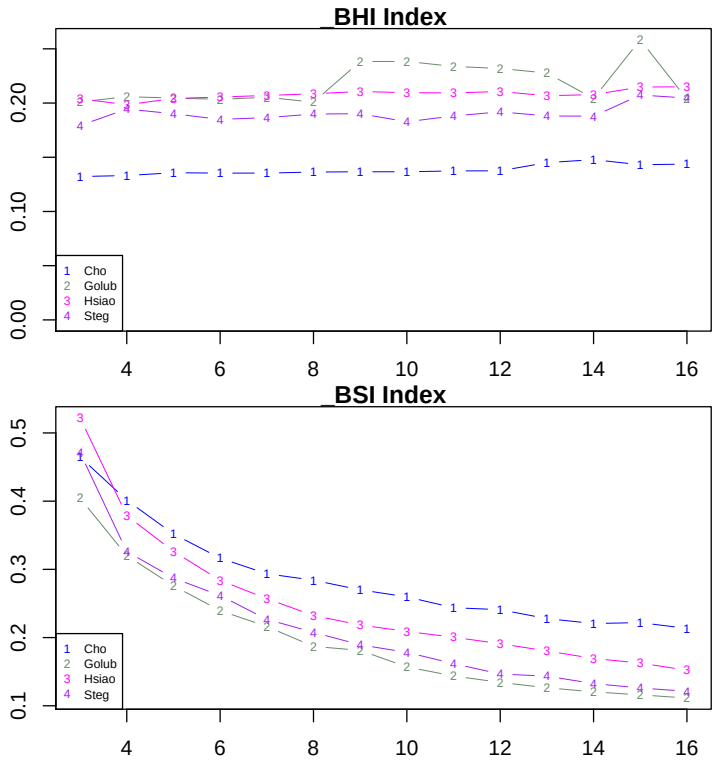


Figure 4.22: KMeans Assessment using Biological Homogeneity and Biological Stability Measures . The x-axis indicates the cluster number while the y-axis indicates the score achieved obtained.

which remains constant). The *Dunn* and *Davies-Bouldin* indices indicate K=8 for the Synthetic B dataset. The C*-index* is minimised for Synthetic dataset B at $K = 7$. This value of this index increases with K for Synthetic dataset A. Optimal K indicated by the internal assessments for various datasets is given in Table 4.8. Unsurprisingly, the *SD-Validity* index tends to increase with K, Fig. 4.24.

	Dunn	Davies-Bouldin	SD-Validity	Silhouette	C-Index
Synthetic B	8	8	-	6	7
Stegmaier	-	-	-	6	

Table 4.8: Optimal K selected with internal assessments and SOTA algorithm

Table 4.9 shows the results for the stability analysis of selected datasets for the

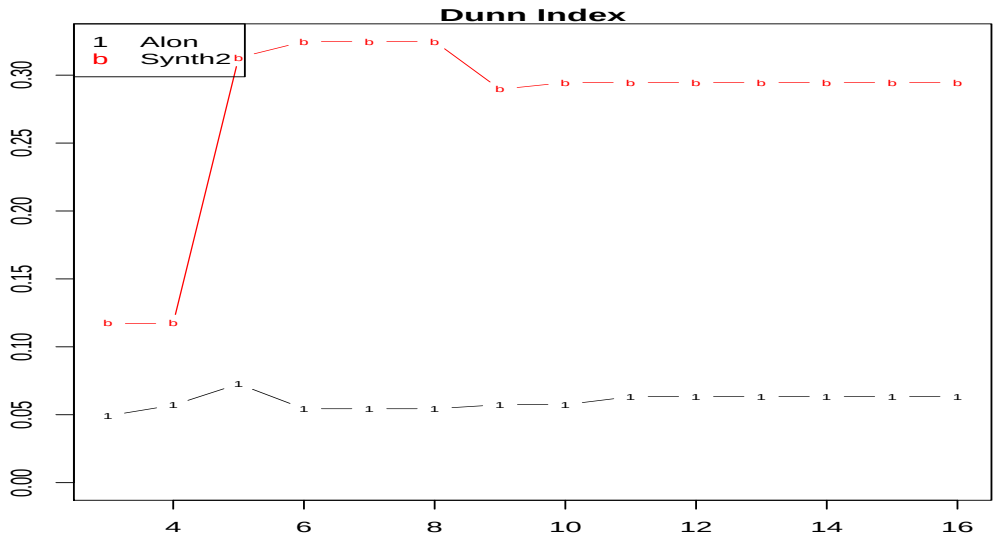


Figure 4.23: Dunn index returned for various K and SOTA algorithm.

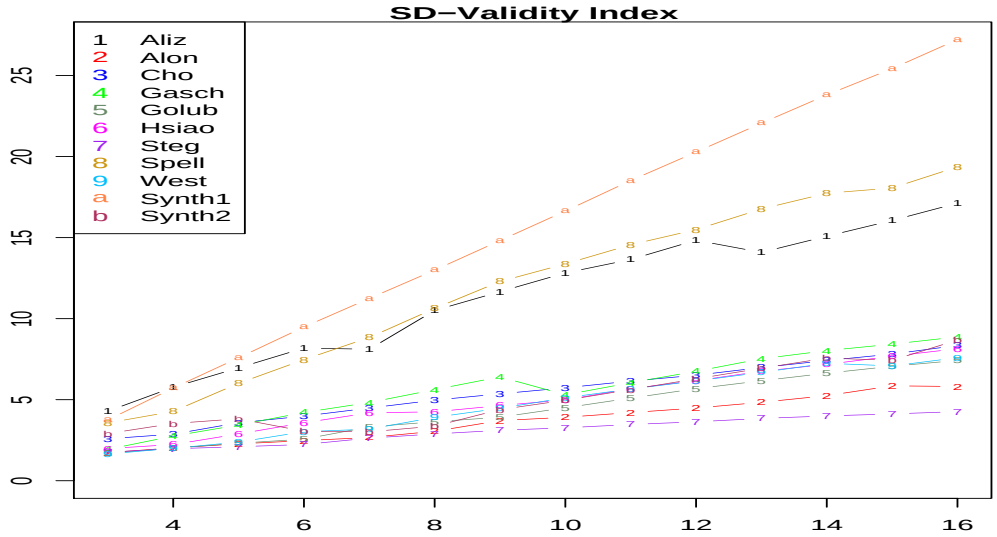


Figure 4.24: SD-Validity returned for various K using SOTA algorithm. This index is biased towards small values of K.

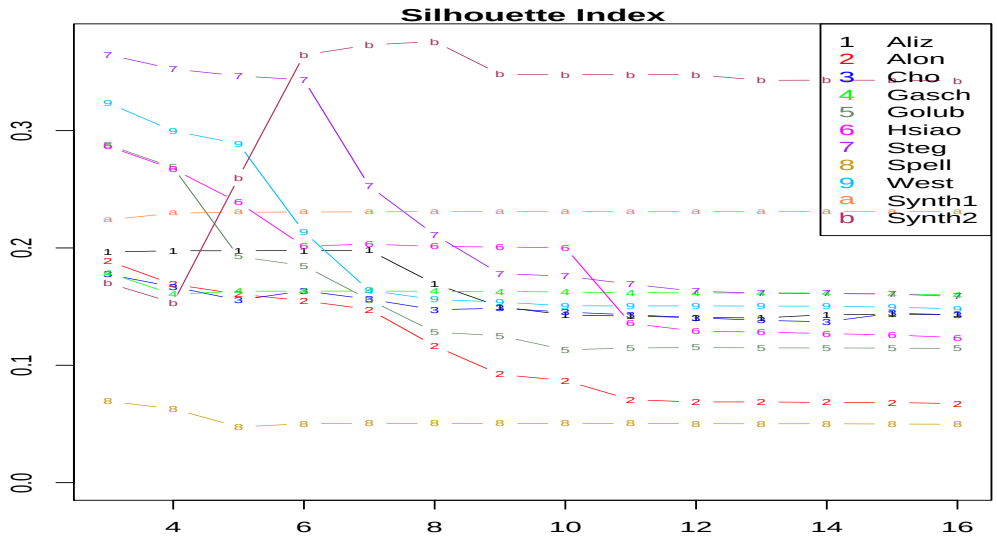


Figure 4.25: Silhouette index for various K using SOTA algorithm. For all datasets this value is > 0

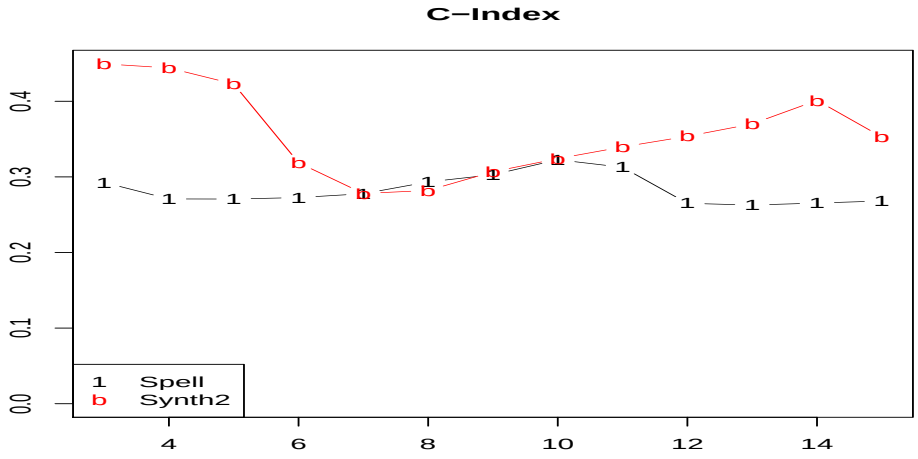


Figure 4.26: C-index for selected datasets for various K, using SOTA algorithm

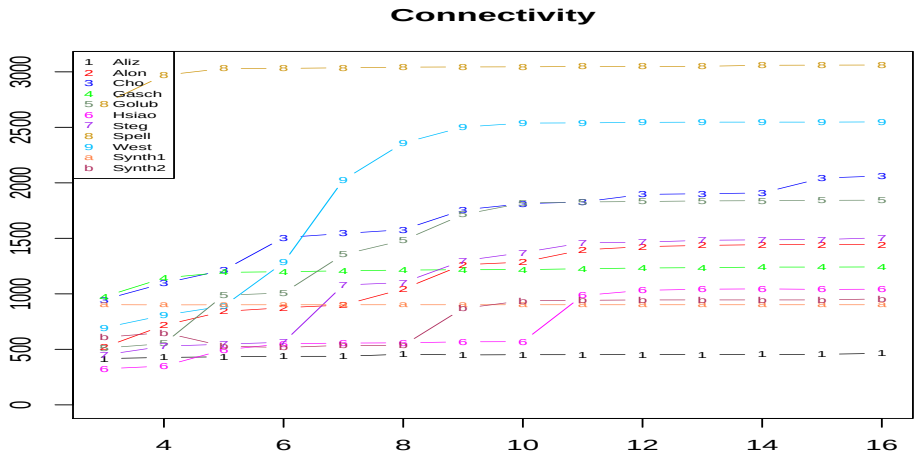


Figure 4.27: Connectivity measurements for selected datasets for various K, using SOTA algorithm

SOTA algorithm. The AD and FOM assessment indices were not informative of K for this technique, hence not shown.

	AD	ADM	APN	FOM
Synthetic B	-	-	5	-
Alizadeth	-	6	-	-
Cho	-	-	6	-
Hsiao	-	8	-	-
Golub	-	-	5	-
Stegmaier	-	-	7	-
West	-	-	6	-

Table 4.9: K selected by stability measurements and SOTA algorithm

Biological Homgeneity (BHI) values, obtained for various values of K with the SOTA algorithm, are of a similar range as those obtained for other clustering algorithms tested, Fig. 4.30. However, with previous techniques, this index did not change with K, whereas with SOTA there is an optimal. For example, this index is optimised at $K = 16$ for the Golub dataset, corresponding to $BHI = 0.25$. This

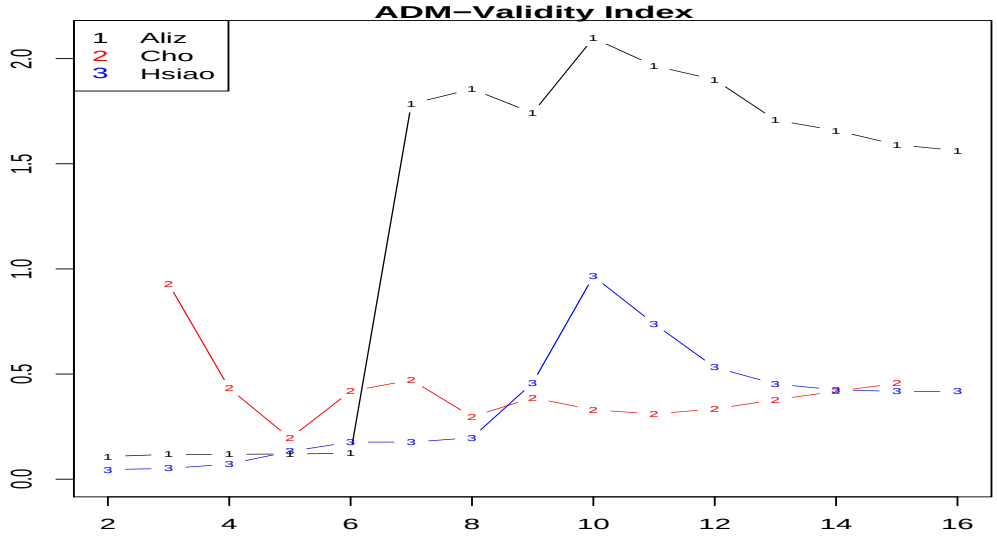


Figure 4.28: SOTA assessment using ADM index for values of K for selected datasets.

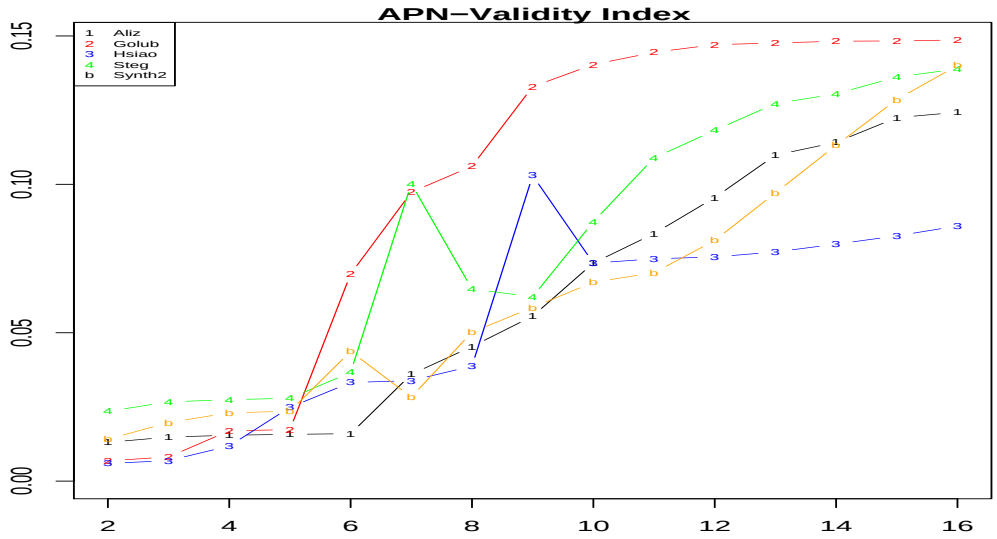


Figure 4.29: SOTA assessment using APN index for values of K for selected datasets.

value does not represent very homogenous clusters, or is it significantly different than the *BHI* value found, for example, with K-Means.

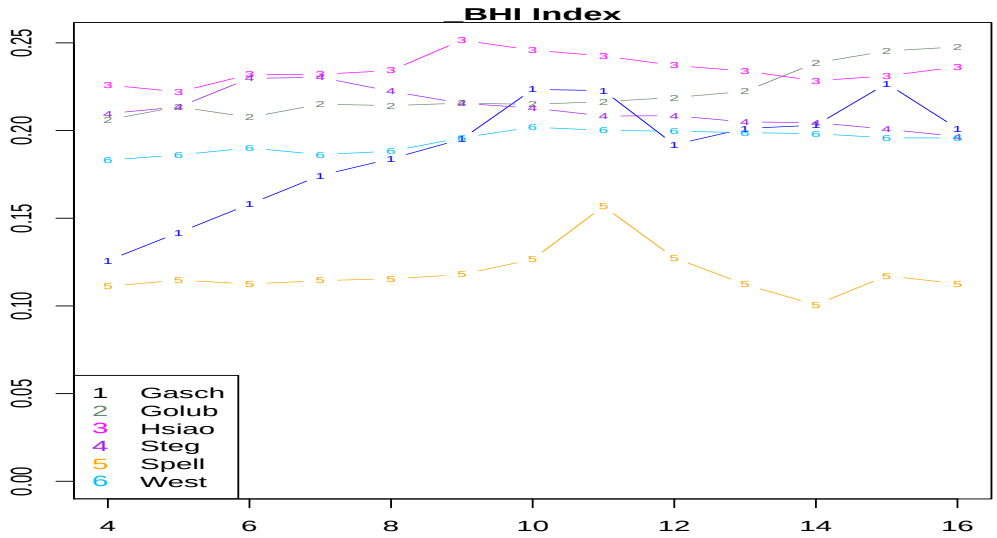


Figure 4.30: SOTA assessment using Biological Homogeneity. The x-axis indicates the cluster number while the y-axis indicates the score obtained.

Partitive Application Summary

A similar bias as observed for hierarchical algorithms occurs for the K-Means and SOTA algorithms whereby *SD-Validity* values increase with K. The Dunn index does not identify any compact and well separated clusters with the K-Means algorithm, again similar to hierarchical techniques. However, in contrast to hierarchical techniques, the silhouette values are > 0 for all datasets across all K. K-Means and SOTA strive to find spherically linked compact clusters, and these results in agreement with Handl et al. (2005) in that the Silhouette index will perform better with these techniques.

4.6.3 Fuzzy Application

Examples of Fuzzy clustering algorithms include Fuzzy CMeans (FCM) and Fuzzy Local Approximation MEthod (FLAME). FCM is the most widely used fuzzy clustering method. We used the FCM implementation available in the *e1071* package of Bioconductor, (Dimitriadou et al., 2008), and similarly to the K-Means analysis, the initial cluster centres were estimated using the technique available in the *clusterSim* package. For these tests we used a 'fuzzification' parameter of 2, (see Section 3.3.2). The developers of Fuzzy Local Approximation MEthod (FLAME) committed code to open-source[3], which we adapted for comparison in R. Recall that this method does not take K as an input parameter, but determines the optimal number from the dataset. However, K is dependent on a number of input parameters, primarily the number of nearest neighbours, knn.

From an analysis of the effect of the knn parameter on FLAME output it was observed that K decreases as knn increases, Table 4.10 (also noted in the original paper (Fu and Medico, 2007)). This relationship arises for two reasons. In this algorithm, a cluster is determined from a 'Cluster Supporting Object' (CSO) and its relationship to knn. Firstly, knn determines the smoothness of the cost function used by this algorithm, which in turn limits the maximum number of CSO's, and secondly, knn determines the range covered by one CSO - a larger value of knn results in a wider CSO range, therefore the fewer the CSO's, (Fu and Medico, 2007). As K is not specified, the assessment indices are presented for various knn. The reader can cross reference with Table 4.10 to check equivalent number of clusters. As this technique creates a dedicated *outlier* cluster, this was removed before assessment.

[3]available from `http://flame-clustering.googlecode.com/svn/trunk/`

knn →	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Alon	35	29	22	22	23	20	19	17	17	15	15	14	14	14	14	14	14	13	13	14	12	11	11	11	10	9
Aliz	48	31	25	16	13	10	10	9	8	6	5	4	4	3	3	3	3	2	2	2	2	2	2	2	2	2
Cho	56	39	35	28	25	23	20	19	21	20	20	19	17	14	11	10	10	10	10	10	10	10	10	10	10	10
Gasch	48	37	31	26	22	21	19	18	17	16	13	13	10	9	9	8	9	6	6	7	7	7	7	7	7	7
Golub	25	22	19	13	15	12	12	11	11	9	9	9	8	7	7	8	7	6	6	6	5	5	5	5	5	5
Hsiao	36	31	28	28	26	25	25	24	21	22	21	20	19	16	16	15	15	15	15	14	13	13	12	12	12	11
Iressa	116	97	83	71	57	54	51	44	41	37	33	32	29	29	29	28	26	23	22	20	18	19	19	18	18	18
Spell	49	37	33	26	23	20	21	19	20	20	20	18	14	13	12	10	9	9	9	9	9	8	8	8	6	5
West	28	29	24	21	16	16	16	14	14	11	11	10	9	7	6	6	7	7	7	7	7	7	7	7	7	7
Synth A	88	64	52	46	44	41	37	37	35	32	25	24	20	18	18	18	18	16	16	15	15	14	14	14	14	14
Synth B	67	59	52	50	36	38	34	30	27	28	28	26	24	25	24	24	21	23	23	21	22	21	21	20	21	20

Table 4.10: FLAME: Effect of knn on K. As the number of nearest neighbours increase, the number of clusters decrease. Due to the property that knn controls K, in further analysis of this method, we we use knn parameter in place of K.

The cluster separation and the *fuzzy* compactness ratio, (*Xie-Beni* index), estimate how spatially separated the clusters are. No well separated or compact clusters found for any $K > 10$ by the FCM technique appearing as very large values of the *Xie-Beni* index, Fig. 4.31. For the Alon, Golub, Stegmaier and West datasets, this index indicates $K = 3$ is optimal, while, for the FLAME method, it points to the partitioning again being more stable for small K (i.e. for large values of knn in the graph). For large values of K, (i.e. small knn) no compact or well separated clusters are found, Fig. 4.32.

The *Partition Coefficient* measures the amount of overlap, or sharing of gene membership, between clusters. As this value approaches unity the clustering approaches a crisp partitioning. For FCM this value monotonically decreases with K (memberships of genes is spread roughly equally amongst the clusters), while for FLAME it remains roughly constant for each dataset for all values of K. This highlights the contrast FCM and FLAME membership requirements for a cluster. For the former, for a given gene, the membership value to a set of clusters is proportional to its similarity to cluster mean, and the latter requires membership of a cluster, i, to be determined by the weighted similarity of the gene to its K-nearest neighbours, and *their* membership of cluster i. The assignment of genes with few neighbours to a dedicated outlier cluster (not considered with this analysis) also contributes to stable results. This implies that for FCM that each gene does not particularly belong to any cluster, while for FLAME, memberships are similarly spread among a subset of clusters, regardless of the cluster number.

Similarly, the *partition entropy* measure monotonically increases with K for FCM, where a small value indicates a good clustering, while for FLAME this value remains roughly constant for the majority of datasets. Again, highlighting the difference in the membership weighting schemes of the techniques. For example, in

the Alizadeth dataset, there is a minimisation of this value at $knn = 23$, which corresponds to $K = 2$ (Table 4.10), while for the Spellman dataset this value is minimised at $knn = 30$, which corresponds to $K = 5$.

4.6.4 Biclustering Application

Here the Plaid biclustering model is investigated[4], significant because it uses a statistical model to capture the biclusters in the data.

Plaid Model

Described in Sections 3.3.1 and 3.3.3, this method requires few input parameters. Code used for this method was made available by the original authors (Lazzeroni and Owen, 2000) and was used in this analysis. An advantage of this approach is the determination of K directly from the dataset, however this algorithm uses a K-Means start to initialize clusters, thus it is sensitive to starting positions and different runs may produce different results. The number and size (% volume of the dataset) of clusters returned for each dataset by the Plaid algorithm is given in Table 4.11

Notable of this technique is the large range in the size of the clusters returned. The average volume of clusters is quite high for each dataset. For example, the average volume of clusters found in the Alon dataset is 16.4% of the volume of the entire dataset and the largest bicluster found in the Stegmaier dataset is 36.6% of the total dataset size. This means that large proportions of the datasets are grouped into one cluster. Table 4.12 gives description of biclusters scores obtained with

[4]In the next sub-section the SAMBA algorithm, a technique which also finds biclusters in the data, (although categorised here as a graphical application) is examined.

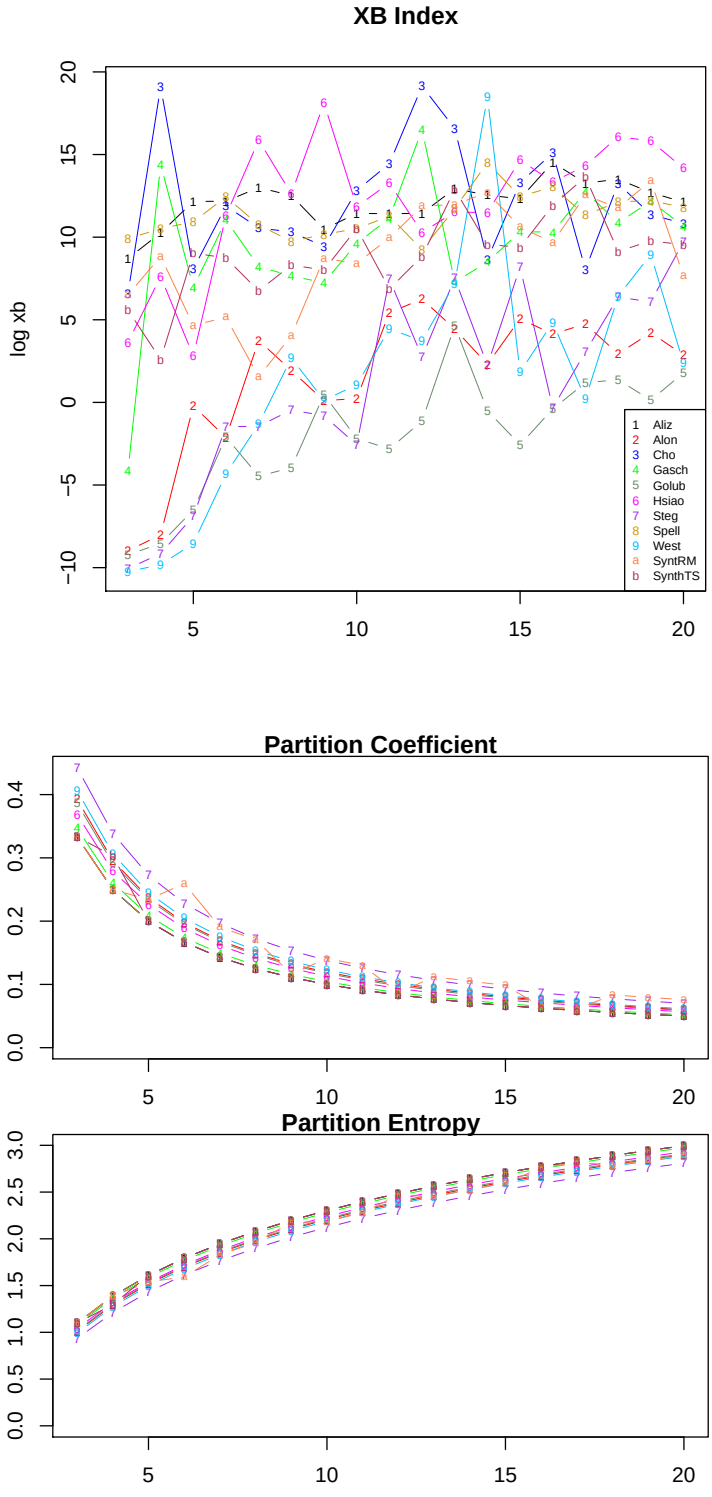


Figure 4.31: FCM Assessment using Internal measures. The x-axis indicates K while the y-axis indicates the score achieved obtained.

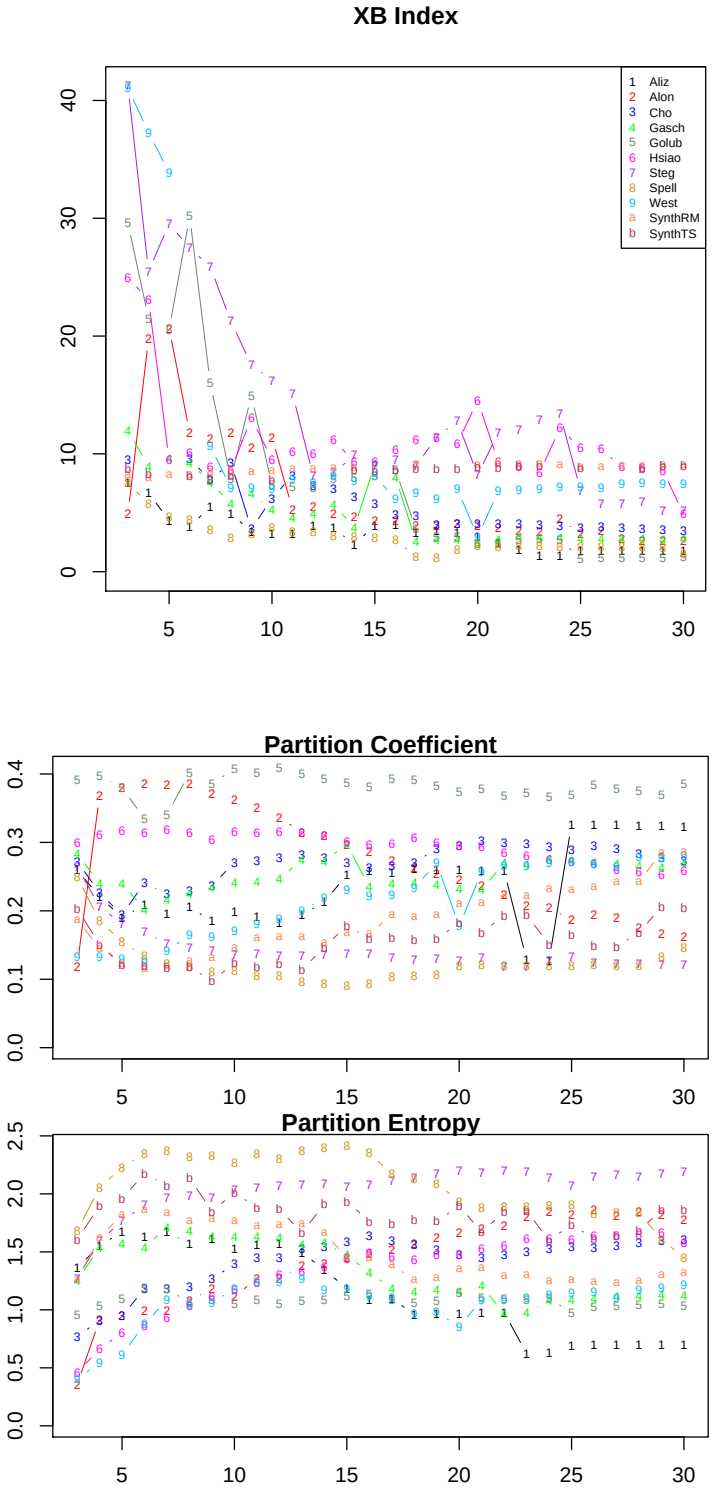


Figure 4.32: FLAME Assessment using Internal measures. The x-axis indicates knn while the y-axis indicates the score achieved obtained.

Dataset	K	Avg. Vol	Min. Vol	Max. Vol
Aliz	3	27866 (9.6%)	12208 (4.2%)	41360 (14.3%)
Alon	37	20280.49 (16.4%)	11400 (9.2%)	32190 (25.9%)
Cho	9	5656.67 (11.1%)	1484 (2.9%)	9904 (19.4%)
Gasch	56	58626.75 (10.8%)	2087 (0.4%)	131442 (24.4%)
Golub	10	39528.9 (15.4%)	27144 (10.5%)	58240 (22.6%)
Hsiao	8	11359 (9.1%)	1788 (1.4%)	27081 (21.7%)
Spell	150	33423.83 (13.3%)	1256 (0.5%)	52836 (21.1%)
Stegmaier	9	10510.44 (15.1%)	3980 (5.7%)	25395 (36.6%)
West	63	23846.37 (14.6%)	2238 (1.4%)	38280 (23.4%)
Synth. Data A	8	21979.25 (12.2%)	5168 (2.8%)	34804 (19.3%)
Synth. Data B	7	20102.71(8.3%)	12250(5.1%)	30818(12.8%)

Table 4.11: Plaid cluster statistics. % indicates the percentage volume of the dataset.

this technique. Of note is the range of scores for biclusters found in the Spellman dataset. This dataset also presented the largest number of biclusters, 69 of which had a score > 150. Biclusters with large scores were found in Synthetic B dataset, however these represented the smallest in terms of % volume. However, $K = 3$, found in Synthetic dataset B, is considerably less than the number designed. Biological homogeneity is again low for the biclusters returned by the Plaid Model, and is in fact lower than that obtained from more traditional methods (analysed above), Table 4.13.

4.6.5 Graphical Application

These evaluations were performed using software developed by the original authors and available with the Expander v4.3 suite of analysis tools[5], (Sharan et al., 2003). Both algorithms examined here, CLICK and SAMBA, score clusters (biclusters in the case of SAMBA), using a probabilistic based scoring scheme (Section 3.3.4).

[5]available for download at `http://acgt.cs.tau.ac.il/expander`

Dataset	K	Avg. Score	Min. Score	Max. Score
Aliz	3	12426.97	10893.62	13960.33
Alon	37	603.17	179.51	1727.81
Cho	9	3068.08	1444.87	4517.52
Gasch	56	6314.35	735.99	66365.16
Golub	10	2621.29	1433.33	4780.91
Hsiao	8	3136.63	1625.79	5768.42
Spell	150	306.44	38.13	2517.25
Stegmaier	9	1475.67	270.29	4155.55
West	63	774.28	238.38	3799.12
Synth. A	8	6653.08	3257.03	8576.35
Synth. B	7	38919.86	17629.02	68774.17

Table 4.12: Plaid cluster scores, a large score indicates a more significant cluster.

	BHI
Gasch	0.110
Golub	0.197
Stegmaier	0.179
Spellman	0.103
West	0.176

Table 4.13: BHI values for selected datasets obtained with Plaid model algorithm

As these methods determine the optimal K from the dataset, clusterings for multiple K are not investigated.

CLICK

Dataset	K	Avg. Score	Min. Score	Max. Score
Alon	8	0.578	0.521	0.614
Aliz	13	0.535	0.423	0.57
Cho	24	0.725	0.615	0.788
Gasch	14	0.671	0.491	0.749
Golub	23	0.492	0.406	0.525
Hsiao	20	0.645	0.538	0.734
Spell	13	0.546	0.344	0.6
Stegmaier	18	0.702	0.583	0.78
West	14	0.501	0.447	0.553
Synth. Data A	-	-	-	-
Synth. Data B	7	0.844	0.685	0.94

Table 4.14: CLICK Cluster Statistics

This is a partial clustering method, in that not all the genes must be put into a cluster, however genes are grouped over all samples, i.e. *global* clustering. Notable from the results, and contrary to results obtained from previous methods, a structure was not identified in Synthetic dataset A (time series synthetic dataset), Table 4.14. Structure was, however, identified in 'real' time-series datasets, (Cho, Spellman, see Appendix B). Structure was found in Synthetic dataset B (7 clusters) with a higher significance compared to all real datasets. There is a small range in scores of clusters found in each of the datasets and the number of clusters found in each dataset is larger than those suggested by assessment indices for other global clustering techniques (hierarchical, partitive and fuzzy).

As this is a partial clustering method, and therefore clusters should in theory

only contain relevant genes, it is surprising that the Biological Homogeneity Index has a similar range to those obtained for traditional clustering techniques. Again, the largest BHI value is associated with the Golub dataset, Table 4.15 and these values are consistent with those obtained for other clustering algorithms tested.

	BHI
Gasch	0.121
Golub	0.200
Stegmaier	0.185
Spellman	0.115
West	0.178

Table 4.15: BHI values obtained for selected datasets using the CLICK algorithm

SAMBA

Dataset	K	Avg. Score	Min. Score	Max Score
Alon	37	468.11	113.14	1217.62
Aliz	101	304.83	62.65	1114.9
Cho	-	-	-	-
Gasch	118	834.04	82.4	5780.29
Golub	49	490.78	32.91	1477.37
Hsiao	20	702.96	307.64	1452.83
Spell	88	348.2	83.43	984.93
Stegmaier	9	570.32	311.2	809.48
West	36	274.78	12.7	881.04
Synth. Data A	49	1823.27	39.55	5777.94
Synth. Data B	28	1027.81	55.78	4216.82

Table 4.16: SAMBA Cluster Scores

On average finds much smaller clusters compared to the Plaid technique, with most clusters $< 1\%$ of the total volume of the dataset. This is due to the negative effect of non-edges on bicluster scores, thus making larger clusters unfavourable, as

these would inevitably include more non-edges. Notable among the results of the SAMBA algorithm is the absence of structure found in the Cho dataset.

Dataset	K	Avg. Vol	Min. Vol	Max. Vol
Alon	37	1432.3 (1.1%)	264(0.2%)	3956(3.1%)
Aliz	101	468.8 (0.1%)	60 ($2e^{-3}$%)	2050 (0.7%)
Cho	-	-	-	-
Gasch	118	1170.34 (0.2%)	96($1e^{-4}$%)	9454(1.7%)
Golub	49	1699.08 (0.6%)	45 ($1e^{-4}$%)	5190(2.0%)
Hsiao	20	1636.3 (1.3%)	495 ($3e^{-3}$%)	5264(4.2%)
Spell	88	482.2 (0.2%)	91($3e^{-4}$%)	1648(0.6%)
Stegmaier	9	1704.44 (2.4%)	931(1.3%)	2592 (3.7%)
West	36	1067 (0.6%)	35 ($2e^{-4}$%)	3302 (2.0%)
Synth. Data A	49	1726.14 (0.6%)	55 ($2e^{-4}$%)	6000(2.2%)
Synth. Data B	28	896.78 (0.4%)	104 ($3e^{-4}$%)	3400(1.2%)

Table 4.17: SAMBA Cluster Statistics

Although, structures found with this biclustering technique are smaller compared to Plaid model, the *BHI* values are marginally larger. As with all techniques examined, analysis of the Golub dataset for biological homogeneity returns the best result, albeit still small. Similar BHI results were found for this technique, when compared to more traditional methods, Table 4.18.

	BHI
Gasch	0.178
Golub	0.209
Stegmaier	0.195
Spellman	0.157
West	0.180

Table 4.18: BHI values for selected datasets for biclusters obtained from SAMBA algorithm

Graphical Methods Summary

The probabilistic scoring scheme employed by both models is a strength of these techniques, as it provides an estimate of significance of results. The final score of a cluster is a summation of significance of edges when using SAMBA, whereas for CLICK it is the average. Graph theoretic methods provide an intuitive method of gene expression analysis, owing to the modularity and inter-connectedness of gene expression in the cell. Moreover, the SAMBA technique is innovative as it finds local structures in a bipartite graph. In this analysis, this algorithm was applied to gene expression data only, however, it has been applied to compendium of biological data, (Tanay et al., 2005). The variation of clusterings found between the CLICK and the SAMBA algorithm, (although the former finds global structures, and the latter, local), highlight the need to investigate the techniques used to score gene interactions in the graphical domain. For example, the SAMBA algorithm identifies no biclusters in the Cho dataset, however, when applied to the same dataset the CLICK algorithm identifies clusters which has the highest overall homogeneity score of any *real* dataset tested. SAMBA edge-weighting is thoroughly investigated in Chapter 5. Using graphical techniques also presents the opportunity to explore the organisational properties of gene interaction using tools and ideas from classical graph theory.

4.7 Summary

It is important to understand that use of analytical validation techniques solely is not sufficient, but that an understanding of the working principles of clustering algorithms, validation measures and their intrinsic biases is critical to enable fair

and objective cluster validation. As shown in this investigation, many validation techniques are intrinsically biased, hence a careful analysis of the results obtained is required, and results should be corroborated using alternative validation techniques. Although research suggests (Chapters 2 and 3 and references therein) that clustering techniques which find a local structure in the data are more suitable for gene expression data, the topic of internal assessment measures for biclustering is not well developed and is of fundamental concern.

This analysis shows that the FCM algorithm does not partition the data well, where the membership of the genes is spread evenly across all clusters for all K, as indicated by the *partition coefficient* and the *partition entropy* measures. The *Xie-Beni* index has been found to be very eratic. However, the FLAME algorithm, which performs a partial clustering of the data by having a committed 'outlier' cluster, returns more stable results. FLAME was found to be more suitable for gene expression data clustering, which often contains a large noise component or irrelevant measurements. Most clustering techniques do not provide estimates of the significance of results returned. This is a strength of the exceptions CLICK, SOTA, Plaid and Samba algorithms, which do estimate significance in cluster scoring, as validation methods are inherently biased and misleading. A key strength of graphical techniques is that it transforms gene expression data into a network. These resulting networks can and should therefore be examined further using classical network analysis methods, to highlight the properties of gene interactions.There is no method investigated here which is optimal across all datasets, as indicated by the assessment indices. There is always a tradeoff between algorithm and assessment measure used, e.g. SD-Validity, BSI, Silhouette.

In this chapter assessment indices were evaluated for a range of cluster numbers K, for each dataset, to detect biases and trends. With many cluster validation

packages available, the returned result is simply the minimum or maximum for a particular range of values for K. However, we have shown that the trend and range of the values need to be accounted for. This analysis was carried out on comparably large sets of genes than typically reported in literature. It is expected that better results would have be obtained by limiting the number of genes being clustered to a small subset ($100 \sim 600$) for computational and visualization purposes, however eliminating perhaps interesting genes in the filtering process.

Assessment indices measure the extent of a clustering algorithms's ability to find structures in a dataset. However, for clustering gene expression data, it is reasonable to consider external measures that use existing biological knowledge. Internal measures by themselves may not be suitable for gene expression data which are often subject to many sources of noise, (also argued by Handl et al. (2005)). The BHI index was used to quantify the association of gene expression profiles in a cluster with functional classes. The BHI index is, of course, greatly influenced by the annotation used. Here, *all* (Biological Process, Cellular Component and Molecular Function) categories from Gene Ontologies annotation database was used to define functional classes. If, for e.g. FunCat or EASE were used to determine functional classes the results may vary slightly.

Past studies have concluded that clustering of the gene expression profiles show that functionally similar genes are grouped together. This is often concluded by manually inspecting genes in a cluster. Validation of a clustering result in this manner is tricky, however, from our investigation, the biases and limitations of assessment indices suggest that validation through expert knowledge is the ideal, although time consuming and subjective. Manual interrogation will be aided as more advanced ontologies and detailed annotation databases become mainstream. External indices would be preferable over internal indices when there is a substan-

tial biological knowledge about the genome under investigated (i.e. proportion of annotated genes).

CHAPTER 5

IN DEPTH: CONSTRUCTING AND EXPLORING GENE EXPRESSION BI-PARTITE GRAPHS

Statistical methods can identify specific genes and groups of genes through co-expression analysis but are less reliable in terms of identifying pattern and dynamics of interaction. A natural representation of such inter-connectivity and an aid to its evaluation is a network. Practically, such a network can be considered as a (weighted) graph. Here, we introduce a new method for extracting graph structure from a gene expression dataset. We explore the organisational properties of graphs obtained, such as node degree distributions, edge distributions, clustering co-efficient information and amongst others and compare these for empirical data to those generated by a *random graph* model. We also describe, in detail, the motivation and implementation of a new edge-weighting scheme for the graph extracted from the data. Finally, we present results of an analysis of this weighting scheme and compare its performance with the well-known Tanay scheme, (Tanay et al., 2002, 2005).

5.1 Introduction

Traditionally, graph theory has been used to study complex networks[1] and is proving a useful tool in the analysis of large complex biological datasets. For instance, protein-protein interactions can be modelled as an undirected graph (Pereira-Leal et al., 2004), where nodes represent proteins and an edge connects two nodes if the proteins *physically combine*. Transcription factor binding sites can also be identified through the use of undirected weighted graphs, where the weights of edges capture the similarity between aligned nucleotides in an input set of promoters, (Reddy et al., 2007). Additionally, metabolic networks can be represented as *bi-partite graphs*: In this case, an edge connects a reaction to a compound node, representing either substrate or product relationships, (Bourqui et al., 2007). As noted (Section 3.3), gene expression can also be modelled as a weighted bi-partite graph, where the two node types represent genes and samples. An edge is taken to exiss between a gene and a sample node, with the weight of the edge representing the effect of the experimental condition on the expression of the gene, (Tanay et al., 2002). Bi-partite graph structures have also been been studied in a wide variety of contexts: for instance, in reference to company boards, (Robins and Alexander, 2004), to film actor social contacts, (Newman et al., 2002), to financial networks, (Caldarelli et al., 2004), in investigating word occurrences (i Cancho and Sole, 2001), in peer-to-peer networks, (Blond et al., 2005) and with respect to scientific co-authoring, (Newman, 2001b,a), amongst others.

The analysis of the gene expression graphs of interest to us is carried out on three levels, Fig. 5.1. Our analysis begins by extracting bi-partite graphs to investigate reactivity of genes at various response levels. Genes may be coherently

[1] A system composed of interconnecting parts that as a whole exhibit complex properties not evident from the properties of the individual parts.

expressed at these different levels across samples: the next step is thus to investigate the bi-partite graph properties for combined levels of response. In Chapter 6, (mentioned here to give the complete roadmap), we also investigate the properties of, and extract coherent gene modules from, one-mode[2] gene expression graphs.

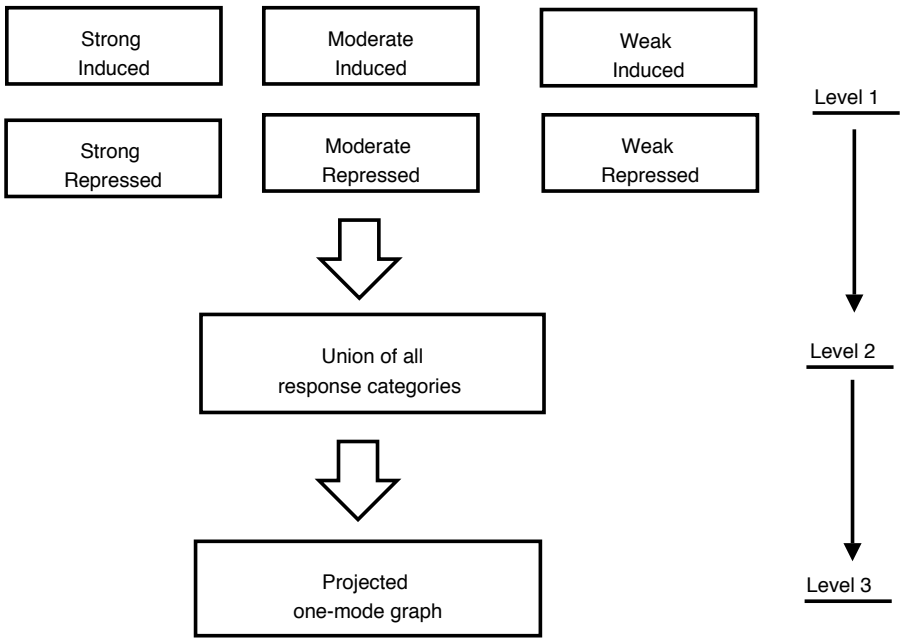


Figure 5.1: Three level analysis of gene expression graphs on three levels.

5.2 Extracting a Graph from a Gene Expression Dataset

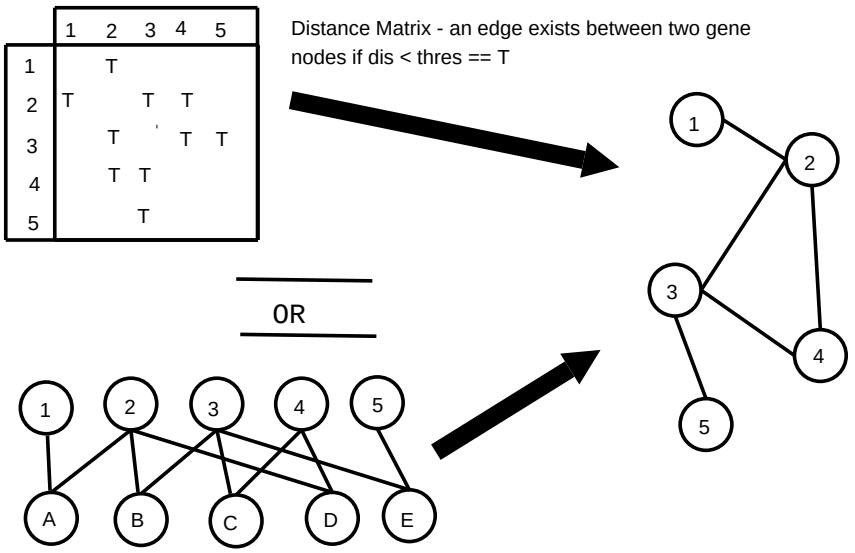
A gene expression graph can be organised as a *bi-partite model*, $G = (\top, \bot, E)$ or a *one-mode model*, $G = (V, E)$, (Section 3.3.4). A large, and powerful, set of tools and ideas exist for one-mode graphs. A one-mode gene expression graph can be extracted from a dataset by e.g. applying a threshold to the distance matrix, and creating an edge between genes whose similarity exceeds this threshold, Section

[2]Defined in Section 3.3.4 to be a graph where an edge can exist between any two nodes

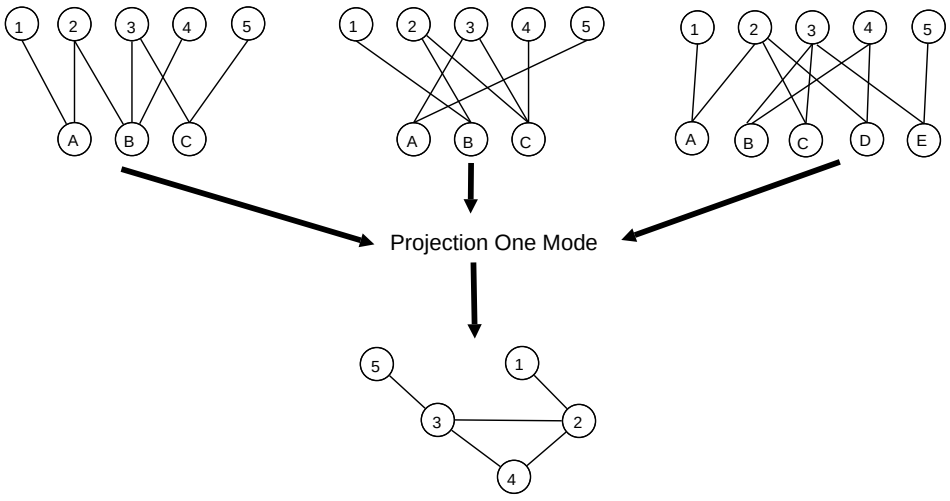
3.3.4 and Zhang and Horvath (2005); Carlson et al. (2006), illustrated in Fig. 5.2a. However, this technique extracts a graph which encodes *global* structures from the dataset, so that it inherits drawbacks of the associated distance/similarity measure. Alternatively, two gene nodes can be linked in a one-mode *projection* of the bi-partite graph, if they have a sample neighbour in common, Fig. 5.2a. However, some information encoded in the bi-partite graph may be lost by this projection, such as details on *which* or for *how many* samples gene expression is similar. In illustration, three gene nodes e.g. can form a *clique* even though not expressed under the same samples, (nodes $2, 3$ and 4 in Fig. 5.2a). This means that a number of bi-partite gene-sample graphs can give rise to the same one-mode projection, (e.g. Fig. 5.2b). Additionally, each sample node of degree d, (the number of edges incident to a node), can inflate to $\frac{d(d-1)}{2}$ edges in a one-mode projection, which can limit the number and type of the computations that are feasible. Bi-partite graphs capture local structures in the dataset, enabling identification and examination of meaningful local groups of gene-interactivity.

5.2.1 Node and Edge Definition

Critical to this analysis is the definition of a node and, in particular, the relationship between nodes defined through an edge. The edge definition and derivation will influence the final graph generated and hence the information retrieved. Both hard and soft threshold strategies can be used for definition of an edge in the network. A *hard threshold* is an all-or-nothing approach, where an edge is said to exist if the score of the gene in a particular sample exceeds a certain threshold. A *soft threshold* approach assigns an edge between *each* gene and sample node according to a function, $f(x) \rightarrow [0, 1]$. This effectively ranks *all* nodes in a network. If a list



(a) Two methods of extracting a one-mode graph for gene expression. Top: extracting a graph from direct threshold analysis of the distance matrix, (dis = distance determined by some distance function, thres = threshold, T = TRUE). Bottom: Projecting bi-partite graph into one-mode by retaining links through second degree neighbours, e.g 2-D-4 produces link 2 – 4 in one-mode projection.



(b) Multiple bi-partite graphs can give result in the same one-mode graph in a projection.

Figure 5.2: One-Mode Gene Expression Graphs

of neighbours of a particular node is required, a threshold must be determined for all edges so that these become the investigation focus.

There are two sets of nodes in our representation of a gene expression graph, $\top$ = the set of all genes and $\perp$ = the set of all samples. Fundamental to the method proposed here is that, for a given sample, genes having either high or low expression, (equivalent to induction or repression), are more likely to contribute to a function, or have a functional response, than for those for which expression values remain unaffected. Affected expression values can thus be extracted for further analysis. An edge (i, j) is thus defined for the bi-partite graph, if gene $i \in \top$ is deemed to show significant change in expression relative to its normal level under sample $j \in \perp$. The edge set here is created using an empirical-based scheme. (The weight of the edge should then reflect how 'interesting' this change in expression is relative to other gene expression changes in that particular sample. Implementation of such a scheme is described, Section 5.3.) This approach differs from previous work, (Zhang and Horvath, 2005; Tanay et al., 2002, 2005), based, respectively, on (i) a soft thresholding approach applied to a distance matrix, based on an assumed power-law distribution, and (ii) defined nodes in a graph based on "properties" rather than samples.

In our method, a high/low expression value for gene i, under sample j, is determined relative to other expression values in gene vector i, (i.e. *across* rather than *within* samples). The motivation for this, (Chapter 2.5), is that, for microarray technology, direct comparison of expression measures within arrays is problematic, because fluorescent intensities are not the same across genes. While measured intensities are roughly proportional to mRNA abundance, the proportionality factor is different for each gene. Specifically, this means that *between-sample, within-gene comparisons are appropriate, but within-sample, between-gene comparisons are*

not straightforward, (Gentleman et al., 2005b).

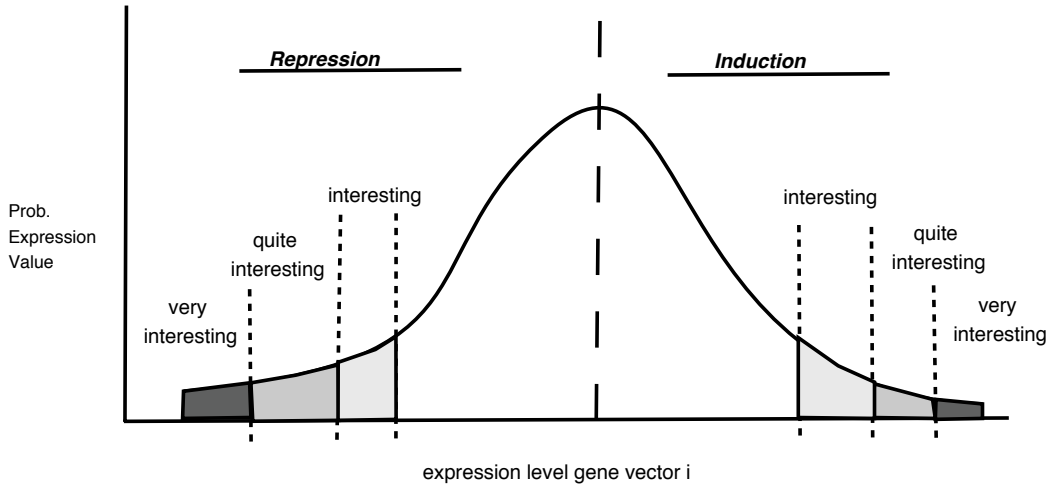


Figure 5.3: Extracting expression values from a gene vector for further analysis

Hence, an edge (i, j) exists when the i^{th} gene shows "significant" induction or repression, relative to its mean level of expression, for sample j, Fig. 5.3. To estimate this significance, we make use of Chebyshev's inequality (Chebyshev, 1867), as no distributional form of expression values for each gene is assumed. Chebyshev's inequality, for any real number $\kappa > 0$, can be written:

$$Pr(|X - \mu| \geq \kappa\sigma) \leq \frac{1}{\kappa^2} \tag{5.1}$$

with random variable X, μ the expected value of X and σ^2 the variance.

Those expression values $X = x_{ij}$, $(i = 1 \ldots n,\ j = 1 \ldots p)$, of interest, for a given sample j, are taken to be $\geq \kappa\sigma$ from the mean expression of gene vector i. From Eq. 5.1, for example, the associated probability of an expression value $\geq 4.47\sigma$ from the mean of gene i is less than 0.05. Expression values $\geq 4.47\sigma$ from the mean would indicate a strong response of gene i to sample j.

Clearly, categories can be established to highlight those expression values which

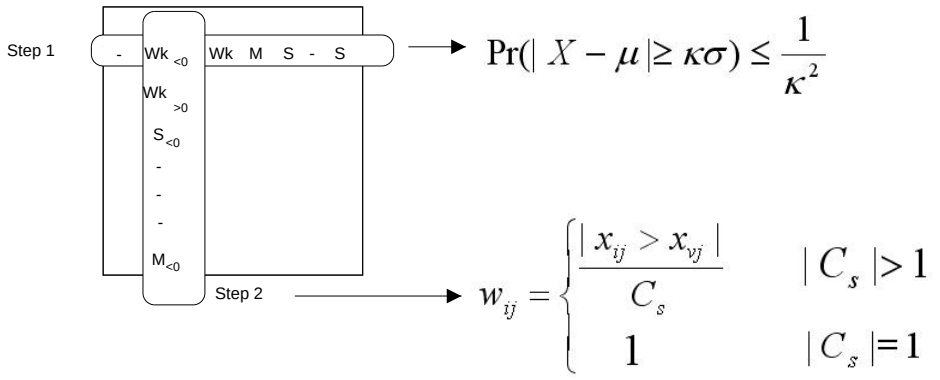


Figure 5.4: Two-step process of empirical scheme. Step 1: A univariate analysis of each gene vector is carried out to determine strength of response. Step 2: A univariate analysis of each sample vector is performed, used to order gene response and assign weights (Section 5.3).

indicate a *weak response*, *moderate response* and *strong response*, where κ indicates the threshold between categories, Fig. 5.4. For example, these categories could be defined by grouping expression values which are $\geq 2.58\sigma$, $\geq 3.16\sigma$ and $\geq 4.47\sigma$ from μ, into *non-overlapping* (mutually exclusive) categories of weak, moderate and strong respectively, corresponding to probabilities = 0.15, 0.10 and 0.05, Eq. 5.1. (The number of categories can clearly be extended for fine-grained response.) For the analysis described, the three categories weak, moderate and strong, as defined here were used as also in Tanay et al. (2002), to facilitate comparison. Threshold determination between categories is discussed further in the next subsection.

5.2.2 Threshold Estimation

To decide on thresholds between categories, i.e. the value of κ, graphs from real datasets, $G = (\top, \bot, E)$ are compared to graphs from random datasets[3], $G_{Rand} =$

[3]Random datasets of the same dimension as the input dataset were created, where for each row (gene) i, random numbers were selected from a Normal distribution of mean, $\mu_{i\ Rand} = \mu_{i\ Real}$, and standard deviation, $\sigma_{i\ Rand} = \sigma_{i\ Real}$ (Note: for a gene which does not respond to any sample, gene

$(\top_{Rand}, \perp_{Rand}, E_{Rand})$, for a range of possible thresholds, (Fig. 5.5). The *null model* assumes each edge in the graph was created with

$$probability = (|E_{Rand}|/\text{Number of possible edges})$$

while the alternative model assumes an edge to be created with

$$probability = (|E|/\text{Number of possible edges})$$

The level at which the logarithm ratio of these two probabilities is maximised is taken to be 'optimal' in terms of any real effect observed.

Thresholds between categories are identified sequentially. Firstly, the strong response threshold, κ_{str}, is investigated and identified. One test criterion for maximum κ_{str} for is that at least one edge must be identified in the real graph. Possible thresholds values are then tested in probability increments of 0.02 to determine percentage inclusion of expression values[4]. The threshold is then set at the level at which the log ratio is maximised. Once κ_{str} is found, moderate and weak response thresholds (κ_{mod} and κ_{wk}, respectively) are established, (in that order). The maximum value to test for κ_{mod} is set to κ_{str} and the maximum value to test for κ_{wk} is set to κ_{mod}. In summary, the technique for identifying thresholds is:

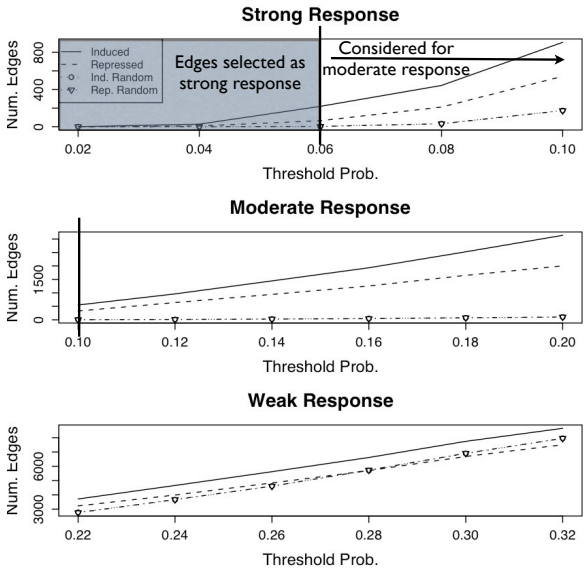
for Induced and repressed categories **do**

- Identify threshold, κ, where at least one gene-sample couple identified. Begin testing from this value.

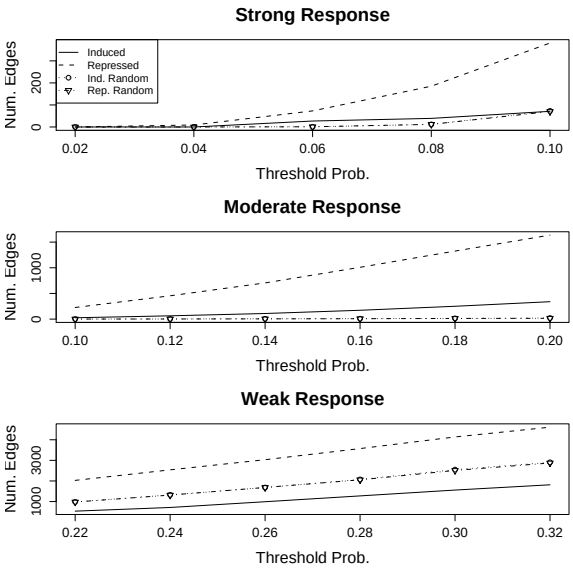
- Test range of possible thresholds for κ_{str}, which correspond to probability

expression is relatively constant.). For each threshold choice in this analysis, 100 random datasets were created to estimate cut-offs, with comparisons based on averaging over these.

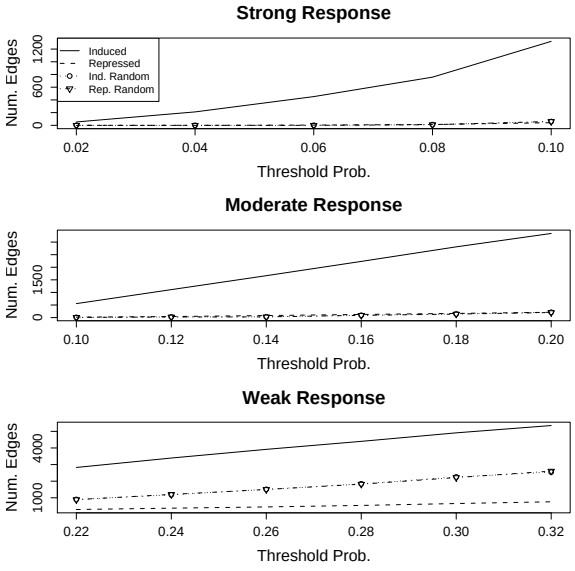
[4]subjectively chosen from analysis of datasets



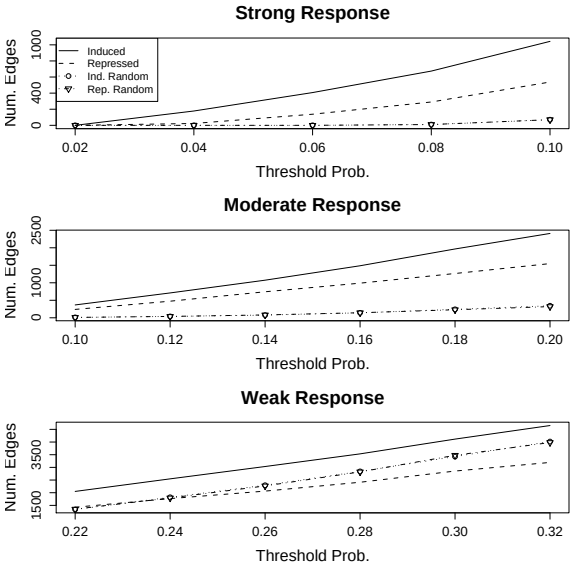
(a) Threshold Analysis of Alizadeth Data



(b) Threshold Analysis of Alon Data



(c) Threshold Analysis of Hsiao Data



(d) Threshold Analysis of West Data

Figure 5.5: Threshold analysis of selected input datasets Alizadeth, Alon, Hsiao and West. The x-axis is the probability threshold i.e. $\frac{1}{\kappa^2}$

increments of 0.02, Eq. 5.1

- Set κ_{str} at the level which maximises the log ratio of probability of edge in real graph to random graph.

(*Gene-sample couples corresponding to strong response have been identified and hence the strong response graph can be created. Remove those gene-sample couples from the dataset.*)

- Starting from κ_{str}, test range of possible thresholds for κ_{mod}, which correspond to probability increments of 0.02, Eq. 5.1.

- Set κ_{mod} at the level which maximises the log ratio of probability of edge in real graph to random graph.

(*The moderate response gene-sample couples have been identified and hence the moderate response graph can be created. Remove those edges, i.e. gene-sample couples, from the dataset*).

- Starting from κ_{mod}, test thresholds for κ_{wk}, which correspond to probability increments of 0.02, Eq. 5.1.

- Set threshold for weak response at level where log ratio is maximised, κ_{wk}.

(*The weak response gene-sample couples have been identified and hence the weak response graph can be created. Remove those edges from the dataset*).

end for

Fig. 5.5 illustrates the results of a threshold analysis of four test datasets, where the probability of an edge for a range of values of κ is plotted for each of the categories of response. Note that for each dataset, thresholds were identified sequentially, therefore the strong response threshold was identified and set before carrying out analysis for the moderate threshold, likewise for weak response. The x-axis in each of the plots indicats the corresponding probabilities for the κ tested. Induced, repressed and random cases are shown in each plot. For the random case, there is

an expected gradual increase in the number of edges identified as κ approaches 0, mimicking the underlying normal distribution. The Alizadeth, West and, to some extent, Hsiao, are somewhat similar to the random profile, however differences exist on the vertical scale, indicating that there are more genes identified as repressed or induced across samples in the real datasets, compared to the random. In terms of comparative performance of real data versus random in the weak response category, the former, in general, led to identification of relatively more co-expressed genes, although the difference is less distinct: in some cases more edge were picked out by random selection.

The point at which the difference in the vertical scale is maximised is taken as the optimal threshold. In, for example, the Alizadeth dataset, Fig. 5.5a, this point occurs in the strong response category at probability = 0.06 (for induced and repressed), corresponding to $\kappa_{str} = 4.08$, Eq. 5.1. All gene expression values which are $\geq 4.08\sigma$ from the mean are taken to be evidence of strong response and are then extracted as strong response from the dataset. All gene expression values which are $< 4.08\sigma$ from the mean are considered for inclusion in the moderate response category. Beginning at $\kappa = 4.08$, possible values for κ_{mod} are tested and theκ_{mod} value at which the log ratio is maximised is identified - this occurs at probability = 0.10 (for induced and repressed), corresponding to $\kappa_{mod} = 3.162$. All gene expression values which, fall in the moderate category, i.e. are $\geq 3.162\sigma$ (and $< 4.08\sigma$) from the mean are then extracted from the dataset. Beginning at $\kappa = 3.162$ possible thresholds for weak response are tested, the threshold for the maximal log ratio is identified, and occurs at probability = 0.20 (for induced and repressed), corresponding to $\kappa_{wk} = 2.23$. All gene expression values in the weak response category: $\geq 2.23\sigma$ (and $< 3.162\sigma$) s.d. from the mean are then extracted from the dataset.

For a number of datasets the thresholds are clear (e.g. Alizadeth and West datasets), however for others it is not so evident. For instance, the Alon dataset, (which represents data from a colon cancer study, see Appendix B), has anomalous behaviour, in that repressed genes are more evident, generating a large number of edges, while the induced genes closely follow random behaviour. Conversely, the Hsiao dataset, (which represents data from a compendium of normal tissue samples), evidently has more genes induced than expected while repressed genes are again close to random.

Table 5.1 shows $\kappa_{str}, \kappa_{mod}$ and κ_{wk} representing the thresholds identified for each of the test datasets, i.e. for which the ratio of edges in real vs. random graphs is a maximum. For the majority of datasets thresholds are common, with $\mu \pm 3(4)\sigma$ a significant deviation the gene vector mean. The Cho and Stegmaier datasets are the smallest in terms of number of samples, (17 and 22 respectively, Appendix B), hence thresholds are less markedly difference from 0. The anomalies of the Alon dataset are again reflected here in the lower threshold for the weak induced category, and, similarly, Hsiao dataset, with lower thresholds for strong response. These thresholds were applied for the respective datasets and six subgraphs were extracted for each test case, where an edge indicates a significant change in expression *in that category.*

5.2.3 Properties Of Gene Expression Graphs

Complex networks are usually analysed for a specific list of properties, whether bi-partite or one-mode, although basic analysis tools for bi-partite graphs have not previously been applied to bi-partite gene expression graphs.

In the following discussion, *weak repressed, moderate repressed, strong re-*

	Repressed			Induced		
	Strong	Moderate	Weak	Strong	Moderate	Weak
Aliz	4.08	3.16	2.23	4.08	3.16	2.23
Alon	4.08	3.16	2.23	4.08	3.16	1.82
Cho	3.16	2.88	2.23	3.16	2.88	2.23
Gasch	4.08	3.16	2.23	4.08	3.16	2.23
Golub	4.08	3.16	2.23	4.08	3.16	2.23
Hsiao	3.58	3.16	2.23	3.58	3.16	2.23
Spell.	4.08	3.16	2.23	4.08	3.16	2.23
Steg.	3.16	2.50	1.82	3.16	2.50	2.04
West	3.58	3.16	2.23	3.58	3.16	2.23

Table 5.1: κ, the number of s.d. units from the mean, that represent thresholds identified for each of the tested datasets.

pressed, weak induced, moderate induced and strong induced subgraphs, are the categorised subgraphs. A *one-mode graph* refers to the graph obtained by projection of the bi-partite graph. The term *'all-in-one'* refers to a graph which is not split into categories, i.e. contains all nodes and edges from all subgraphs $G \in (\top, \bot, E)$.

The main properties of the bi-partite gene expression graphs extracted from the test datasets are outlined in Table 5.2. Despite the size of the table, it is worthwhile considering it as a whole in order to identify common features across a wide range of data, as well as highlighting distinction.

Table 5.2: $n_\top$ = the active set of genes, $n_\bot$ = active set of samples, m = number of edges, $d_\bot$ = average degree of sample nodes, $d_\top$ = average degree of gene nodes, δ = bi-partite density i.e the fraction of existing links with respect to possible ones, cc = clustering coefficient (Eq. 5.3), cc_{min} = minimum clustering coefficient (Eq. 5.5)

Graph Cat.	$n_\top$	$n_\bot$	m	$d_\top$	$d_\bot$	δ	$cc_\top$	$cc_{min\top}$	$cc_\bot$	$cc_{min\bot}$
Alizadeth										
Str. Induced	209	46	219	1.04	4.70	0.02	0.95	0.99	0.35	0.71
Mod. Induced	476	43	550	1.15	12.79	0.03	0.81	0.98	0.05	0.20
	Continued on Next Page...									

Graph Cat.	$n_\top$	$n_\bot$	m	$d_\top$	$d_\bot$	δ	$cc_\top$	$cc_{min\top}$	$cc_\bot$	$cc_{min\bot}$
Wk. Induced	2088	79	3713	1.77	47.00	0.02	0.44	0.81	0.02	0.09
Str. Repressed	63	16	64	1.01	4.00	0.06	0.97	1.00	0.06	0.33
Mod. Repressed	296	13	327	1.10	25.15	0.08	0.88	0.99	0.06	0.15
Wk. Repressed	1889	53	3227	1.70	60.89	0.03	0.48	0.85	0.02	0.12
Alon										
Str. Induced	27	14	27	1	1.92	0.071	1	1	0	0
Mod. Induced	25	10	26	1.04	2.6	0.104	0.92	1	0.11	0.25
Wk. Induced	1167	28	1813	1.55	64.75	0.055	0.54	0.88	0.03	0.09
Str. Repressed	73	31	73	1	2.35	0.032	1	1	0	0
Mod. Repressed	221	29	226	1.02	7.79	0.035	0.96	0.99	0.04	0.11
Wk. Repressed	1338	56	2028	1.51	36.21	0.026	0.55	0.88	0.03	0.09
Cho										
Str. Induced	351	9	351	1	39	0.111	1	1	0	0
Mod. Induced	227	9	227	1	25.22	0.111	1	1	0	0
Wk. Induced	805	12	2051	1.02	68.83	0.212	0.96	0.99	0.02	0.074
Str. Repressed	19	5	19	1	3.8	0.2	1	1	0	0
Mod. Repressed	21	2	21	1	10.5	0.5	1	1	0	0
Wk. Repressed	122	8	826	1.06	16.25	0.84	0.91	1	0.11	0.25
Gasch										
Str. Induced	434	96	564	1.3	5.87	0.014	0.73	0.97	0.18	0.51
Mod. Induced	970	89	1644	1.69	18.47	0.019	0.55	0.89	0.08	0.30
Wk. Induced	2031	138	7017	3.45	50.84	0.025	0.25	0.59	0.04	0.19
Str. Repressed	645	93	848	1.31	9.12	0.014	0.74	0.97	0.22	0.51
Mod. Repressed	1258	85	2217	1.76	26.08	0.013	0.52	0.87	0.10	0.36
Wk. Repressed	2552	136	8448	3.31	62.11	0.024	0.27	0.61	0.03	0.16
Golub										
Str. Induced	279	60	315	1.13	5.25	0.019	0.82	0.98	0.12	0.42
Mod. Induced	408	60	528	1.29	8.8	0.021	0.69	0.95	0.09	0.25
Wk. Induced	1641	70	2898	1.73	41.4	0.025	0.47	0.85	0.03	0.07
	Continued on Next Page…									

Graph Cat.	$n_\top$	$n_\bot$	m	$d_\top$	$d_\bot$	δ	$cc_\top$	$cc_{min\top}$	$cc_\bot$	$cc_{min\bot}$
Str. Repressed	118	46	122	1.03	2.65	0.022	0.95	1	0.42	0.65
Mod. Repressed	312	45	348	1.12	7.73	0.025	0.84	0.99	0.09	0.30
Wk. Repressed	1411	68	2509	1.77	36.89	0.026	0.43	0.81	0.02	0.07
Hsiao										
Str. Induced	576	57	758	1.32	13.3	0.023	0.72	0.96	0.13	0.35
Mod. Induced	399	57	555	1.39	9.7	0.024	0.68	0.95	0.15	0.36
Wk. Induced	1308	59	2826	2.16	47.89	0.036	0.41	0.81	0.06	0.13
Str. Repressed	13	9	13	1	1.44	0.111	1	1	0	0
Mod. Repressed	17	7	17	1	2.4	0.142	1	1	0	0
Wk. Repressed	22	17	296	1.34	17.41	0.079	0.65	0.94	0.05	0.16
Spellman										
Str. Induced	207	36	218	1.05	6.05	0.029	0.93	0.99	0.07	0.45
Mod. Induced	597	35	657	1.1	18.77	0.031	0.86	0.99	0.03	0.19
Wk. Induced	2193	59	3661	1.67	62 .05	0.028	0.49	0.84	0.02	0.11
Str. Repressed	316	49	323	1.02	6.59	0.02	0.96	0.99	0.06	0.17
Mod. Repressed	731	47	814	1.11	17.31	0.024	0.85	0.99	0.04	0.16
Wk. Repressed	2283	62	3934	1.72	63.45	0.028	0.45	0.82	0.02	0.09
Stegmaier										
Str. Induced	9	4	9	1	2.25	0.25	1	1	0	0
Mod. Induced	60	4	60	1	15	0.25	1	1	0	0
Wk. Induced	1037	20	1249	1.2	62.45	0.06	0.88	0.99	0.04	0.12
Str. Repressed	19	9	19	1	2.21	0.111	1	1	0	0
Mod. Repressed	179	8	180	1	22.5	0.126	0.99	1	0.03	0.08
Wk. Repressed	705	19	832	1.18	43.79	0.062	0.69	0.96	0.03	0.08
West										
Str. Induced	583	47	674	1.16	14.34	0.025	0.81	0.98	0.09	0.28
Mod. Induced	350	44	367	1.04	8.34	0.024	0.93	0.99	0.10	0.38
Wk. Induced	1489	48	2051	1.38	42.72	0.029	0.63	0.93	0.02	0.08
Str. Repressed	266	29	290	1.09	10	0.038	0.88	0.99	0.24	0.50
	Continued on Next Page...									

Graph Cat.	$n_\top$	$n_\perp$	m	$d_\top$	$d_\perp$	δ	$cc_\top$	$cc_{min\top}$	$cc_\perp$	$cc_{min\perp}$
Mod. Repressed	195	24	239	1.22	9.95	0.051	0.74	0.97	0.09	0.28
Wk. Repressed	866	37	1426	1.64	38.54	0.045	0.49	0.85	0.03	0.12

Descriptive Statistics:

Unsurprisingly, in all cases, the weak response categories have larger gene, ($n_\top$), sample, ($n_\perp$) and edge, m sets, as thresholds are lower. Those datasets with a small number of samples tend to have a small $n_\top$ in the resulting graphs, e.g. Cho and Stegmaier datasets, which limits the type and amount of information that can be inferred from it. For a number of datasets there is a preference towards either *induced* or *repressed* categories, in terms of the number of edges identified, e.g. Alizadeth, Alon and Hsiao sub-graphs.

The average degree, $d_\top$, of the gene node set in the majority of datasets is ~ 2, increasing slightly for the weak response categories. This indicates that the majority of genes do not participate in more than two samples in each category (the Gasch dataset has the highest number of samples, with $d_{top} > 2$). However, this still exceeds the expected degree (i.e. n_{top}/m). Again, for datasets with small number of samples, $d_\top \sim 1$. Unsurprisingly, the average degree of the sample node set, $d_\perp$ increases from *Strong* - *Weak* categories, meaning that there are more genes in a given sample responding in the weak category compared to the strong categories.

The density measure, δ evaluates overall how sparse the graph is, i.e. the number of edges compared to the number of possible edges, $m/n_\top n_\perp$. In most cases the gene expression graphs with $\delta < 0.06$. This value increases in certain cases for $n_\top$ and $n_\perp$ small, indicating that most of the sample and gene nodes in the graph are connected, (e.g. Cho, Stegmaier and Alon repressed sub-categories). This measure

is more reflective of the number of gene and sample nodes in the graph and doesn't reveal much organisational information.

Clustering:

The inherent tendency of real networks to form cliques (or clusters) is quantified by the *clustering coefficient*, cc. Traditionally, this is defined, for each node in a one-mode graph, $u \in N$, with at least two neighbours (i.e. degree, $d(u) \geq 2$), as the proportion of edges between its neighbours, (Guillaume and Latapy, 2004):

$$c(u) = \frac{|\{x, y\}, x, y \in N(u)|}{\binom{d(u)}{2}} \tag{5.2}$$

where $N(u)$ is the set of neighbours of u. The clustering coefficient of a graph is the average over all nodes $u \in N$. Equivalently, in the case of a bi-partite graph, the clustering co-efficient, for two nodes u and v, for either the gene or sample node set is:

$$c(u, v) = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|} \tag{5.3}$$

This captures the probability that two nodes in the same node set (i.e. gene or sample) have a neighbour in common. The clustering for one node between all its neighbours is then:

$$c(u) = \frac{\sum_{v \cup N(N(u))} c(u, v)}{|N(N(u))|} \tag{5.4}$$

where $|N(N(u))|$ is the number of second degree neighbours (i.e. neighbourhood of neighbour nodes). For the bi-partite case, a clustering coefficient can be obtained for the two sets of nodes separately, $(\top, \bot)$, by taking the average of Eq. 5.4 over all nodes in each set, resulting in $cc_\top$ and $cc_\bot$. The definition pre-

sented in Eq. 5.3 for two nodes, has the disadvantage that the node with the largest neighbourhood will dominate the results, even if the smaller neighbourhood node is completely encapsulated in the larger. An alternative, introduced by (Latapy et al., 2006), is to scale by the *minimum neighbourhood*:

$$c(u,v)_{min} = \frac{|N(u) \cap N(v)|}{min(|N(u)|, |N(v)|)} \tag{5.5}$$

This defines a more precise relationship between two nodes, and is more appropriate for gene expression data, given the fact that genes may participate in multiple samples and need not be co-active under all samples, with some samples having participation from more genes than others. Again, a summary value can be calculated for each node set separately by taking the average of Eq. 5.5 for each node, $cc_{\top min}$ and $cc_{\perp min}$.

In each of the subgraphs, Table 5.2, the clustering coefficient for gene nodes, $cc_{\top}$, is quite large, indicating that if two gene nodes participate in the same sample node, then a large proportion of their neighbourhood will be similar. The clustering coefficient for the sample nodes, $cc_{\perp}$, captures the overlap between gene subsets participating in the samples and is, unsurprisingly, smaller. Each node in this set has a much higher average degree (i.e. each sample node is typically connected to a large number of genes), whereas the average degree for gene nodes is ~ 2. Thus if two genes show similar expression in two or more samples, this reflects the 'entire neighbourhood' of each gene node. In a random Erdös-Rényi Model graph, (Section 3.3.4), edges are distributed randomly between nodes with a probability p, so the clustering coefficient is $cc = p$, (probability that two nodes are connected). In real networks the clustering coefficient is typically *much larger* than it is for a corresponding random network, (i.e. one having the same number of nodes and

edges as the real network). Although the subgraphs are individually quite sparse (bi-partite density $\delta < 0.1$), the relatively high $cc_\top$ and $cc_\bot$ indicate that there are denser local structures (more evident, as noted, for the gene node set), i.e. $cc > p$. The exceptions are the Cho and Stegmaier subgraphs and *strong/moderate repression* Hsiao subgraphs. These subgraphs also have the smallest $n_\top, n_\bot$ and m, and are quite dense, and $cc_\bot$ indicates that there is no overlap between gene subsets participating in the samples.

The minimum clustering coefficients, $cc_{\top min}$ and $cc_{\bot min}$, capture the local interactions of two genes or two samples. This measure, for the sample nodes, $cc_{\bot min}$, is significantly larger than for $cc_\bot$ (at least double in most cases). This captures the fact that the $cc_\bot$ measure is dominated by sample nodes of high degree, while neighbourhoods of lower nodes of smaller degree do in fact overlap with other sample nodes.

Degree Distribution:

To further understand the basis for gene expression networks, we examine the distribution of the node degrees. The spread in the node degrees is characterized by a probability distribution function, $P(d)$, which gives the probability that a randomly selected node, u, has a degree of exactly $d(u)$. For bi-partite graphs, there are two degree distributions, one for each node set, $\{\top\}, \{\bot\}$. Given there are substantially fewer samples than genes in the datasets, tail distribution statistics are unlikely for this case. For a large number of examples of real networks, the node degree distribution follows a power-law $P(d) \sim d^{-\gamma}$ and these networks are said to be *scale-free*. In an Erdös-Rényi model and Watts-Strogatz graph the edges are placed randomly, and degree distribution of the nodes is close to a Poisson distribution $p(d) = \lambda^d e^{-\lambda}/d!$ (special case of Watts-Strogatz model), (Section 3.3.4)

Each subgraph may of course have alternative forms. Figures C.18 - C.21 (Ap-

pendix C) are the distribution plots found when the sub-graphs of strong - weak induction and strong - weak repression are extracted, for both sample and gene sets. It must be noted that clear identification of node degree distribution is difficult when the node set size in the graph is small. For subgraphs with larger node sets, (Aliz, Alon, Gasch, Golub, Hsiao, West), the node degrees are skewed to the right and are more accurately modelled by a Poisson distribution, most evidently in weak response categories, where $n_\top$ is largest. Overall, $n_\perp$ is small for a large number of the subgraphs: however, where large, (e.g. Figures C.20a, C.20b, C.21a,C.21c weak response categories) it again follows a Poisson distribution.

All In One Graph

To examine whether there is a coherent response of genes across all categories, an *all-in-one* graph can be constructed for each dataset. For example, a gene with expression across samples in the moderate induced sub-graph, may be coherent with one expressed in the moderate repressed graph. This is captured in the *all-in-one* graph which is constructed from the union of all sub-graphs, Table 5.3. The average degree of the gene nodes increases in the *all-in-one* graph indicating that genes identified are participating across categories. Likewise the average degree of the sample nodes increases, indicating that samples have overlapping gene sets across categories. The size of the edge set is simply the sum of the cardinality of each edge set from each sub-category.

Clustering Coefficient:

Once again, we use the clustering coefficient information to examine the *degree of sharing* among node neighbourhoods. In the *all-in-one* graph the clustering coefficient of the sample nodes, $cc_\perp$, is ~ 0.02 for all dataset graphs (with the exception

Dataset	$n_\top$	$n_\bot$	m	$d_\top$	$d_\bot$	δ	$cc_\top$	$cc_{min\top}$	$cc_\bot$	$cc_{min\bot}$
Aliz.	2870	84	8100	2.82	96.42	0.03	0.25	0.56	0.02	0.09
Alon	1857	57	4193	2.25	73.56	0.04	0.32	0.68	0.02	0.09
Cho	1537	12	3495	1.02	131.17	0.19	0.96	0.99	0.01	0.05
Gasch	3114	151	20738	6.66	137.33	0.05	0.18	0.35	0.02	0.13
Golub	2933	71	6720	2.29	94.65	0.03	0.31	0.67	0.02	0.05
Hsiao	1749	59	4465	2.55	75.68	0.04	0.39	0.78	0.06	0.13
Spell.	3024	71	9607	3.18	135.31	0.04	0.22	0.48	0.02	0.12
Steg.	1776	22	2349	1.32	106.77	0.06	0.67	0.94	0.02	0.07
West	2748	48	5047	1.84	105.15	0.03	0.43	0.79	0.02	0.07

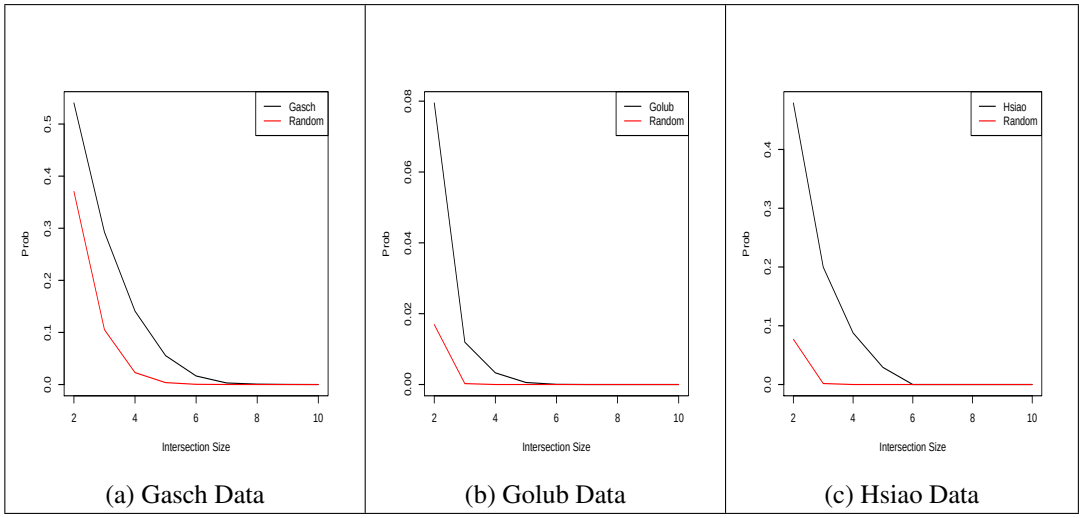
Table 5.3: Statistics of All in One Graph. $n_\top$ = the active set of genes i.e. $d > 0$, $n_\bot$ = active set of samples, m = number of edges, $d_\bot$ = average degree of sample nodes, $d_\top$ = average degree of gene nodes, δ = bi-partite density i.e the fraction of existing links with respect to possible ones, cc = clustering coefficient (Eq. 5.3), cc_{min} = minimum clustering coefficient (Eq. 5.5)

of the Hsiao graph), regardless of sample size - lower than the overall graph densities. This supports the idea that the clustering coefficient is independent of the size of the graph for most real world networks, (Guillaume and Latapy, 2004). However, the minimum clustering coefficients, $cc_{\top min}$ and $cc_{\bot min}$, are relatively high. $cc_{\bot min}$ trebles in most cases and is higher than the overall graph densities. This indicates again that the sample nodes of high degree are dominating the $cc_\bot$ measure, suggesting that there is a level of organization in the graph, such that if two nodes are linked, the neighbourhood of the smaller node will intersect the neighbourhood of the larger.

Intersection Ratios:

Although the clustering coefficient information reveals the degree of sharing among node sets, information on the exact size of the intersection is lost. This level of organisation can be assessed through an examination of the size of the intersection neighbourhoods of the datasets. If two gene nodes have a sample neighbour in common, the size of this intersection neighbourhood will be significantly greater

than for a random graph[5] of the same size and degree distribution, see Fig. 5.6.



(a) Gasch Data (b) Golub Data (c) Hsiao Data

Figure 5.6: Red = random, Black = real. Cumulative distribution of degree of intersection of neighbourhoods of gene nodes, i.e. if two gene nodes have a sample neighbourhood in common, there is a significantly greater chance that this neighbourhood will be larger than expected by chance. The plot shows, for each value i on the x axis, the ratio of all intersections greater or equal to i.

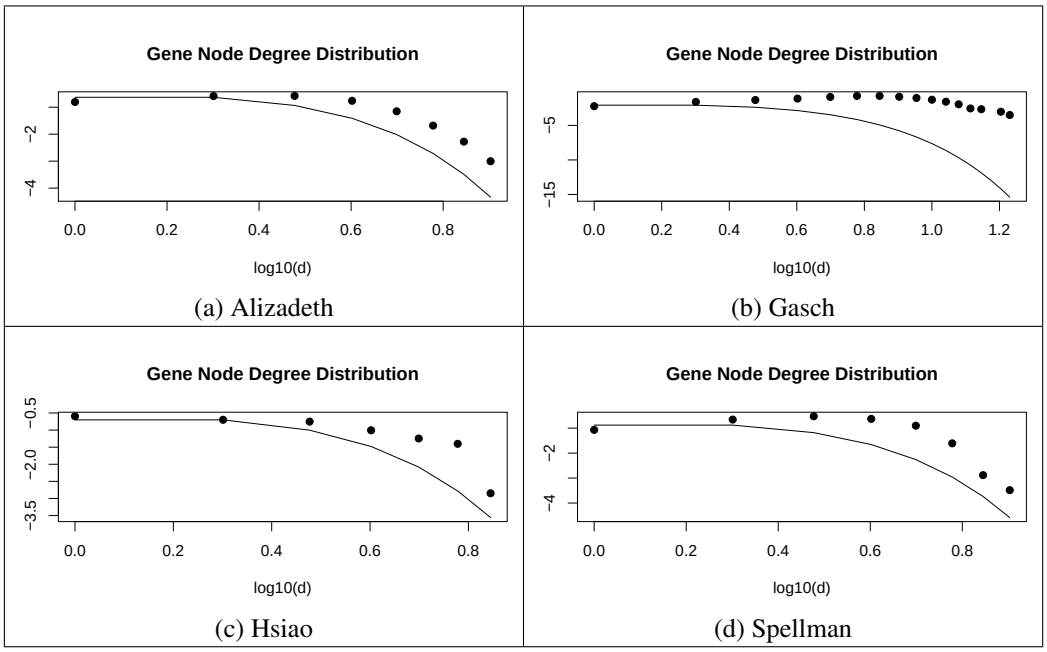
Degree Distribution:

In terms of gene node degree distribution of the *all-in-one* gene expression graphs, (Fig.C.22 - C.23, Appendix C), these are in general right skewed: this is less evident in the Cho and Iressa datasets, which have the smallest number of sample nodes. Right-skewness is largely due to the higher cardinality of the gene sets in each dataset. For the gene node sets, the distributions approach the Normal - in those graphs where the average degree is high - and hence not particularly heterogeneous. The sample node distributions, in most cases, follow the Poisson. In Fig. 5.7 a plot of gene node degree distribution for four test datasets is given. In most cases the

[5]The random graphs were Monte Carlo switching process (Maslov et al., 2004), to create random graphs with the same degree distributions. This proceeds by picking two random edges (x,y) and (u, v) uniformly with x, y, u, v distinct nodes. If (x, u) and (y, v) are not edges, then adding the edges (x, u), (y, v) and delete edges (x, y), (u, v). This process is repeated $m \times 100$ times

gene node distribution is close to the Poisson in shape (solid line in plots), although tail effects are less extreme, indicating that approximation by the Normal is likely to be reasonable. The Gasch dataset, for example, is approximated well by a Normal, Table 5.3, and has the largest set of sample nodes: Note, Alizadeth $n_{\perp} = 84$, Gasch $n_{\perp} = 151$, Hsiao $n_{\perp}$=59, Spellman $n_{\perp} = 71$. While this suggests that as the cardinality of the sample node set increases, the gene node degree distribution tends to Normal, the limited sample and gene set for most datasets means that we cannot conclude this is the case in general. Many other examples of real world one-mode networks approximate a power law. However, in support of our findings here it should be noted that a poor fit to a power-law distribution was also observed in other *bi-partite* models of real world complex networks, (Latapy et al., 2006; Guillaume and Latapy, 2004). Although, projection of the gene nodes into a one-mode graph is shown to follow a power-law in general (see Chapter 6), this conclusion cannot be drawn for the bi-partite case. However, we can conclude that the node degrees in a single graph can have alternative distributions.

Correlations between the bi-partite gene node degree and the degree in a one-mode projection of the gene nodes, (Chapter 6), captures the notion of overlap, Fig. 5.8. The degree of a node in the one-mode projection is the sum of the degrees of the sample nodes to which it is connected in the bi-partite graph, *minus* the number of nodes in common in the neighbourhood of these nodes. That is to say, if a node, u has a high degree in the bi-partite graph and a lower degree in the one-mode projection, there is an overlap between u and its neighbour nodes, i.e. they are over/under expressed in the *same* samples, Fig. 5.8. Correlations between node degree in the bi-partite and one-mode graphs suggest that while some genes show little or no common sample activity, others are co-expressed across their entire sample neighbourhoods.



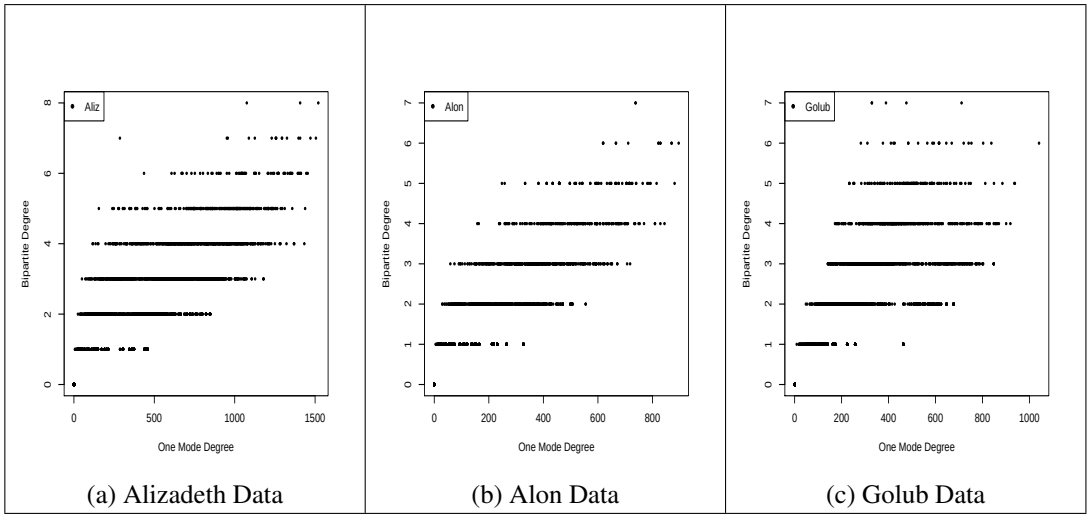
(a) Alizadeth (b) Gasch

(c) Hsiao (d) Spellman

Figure 5.7: Distribution of gene node degrees for four test datasets. The solid line indicates a Poisson distribution. The distributions depart from the Poisson, despite shape similarity in part, with a higher % distribution in the bulk and less extreme tail effect.

5.3 Edge Weights

The weighting scheme for a gene expression graph plays an important role in cluster determination, and thus merits independent investigation. To this end, we introduce the second step of our graph theoretic approach - a univariate analysis of expression values, within each sample vector, used to empirically weight the edges. We also discuss performance measures for edge-weighting schemes, and present a comparative evaluation based on application to several real datasets.



(a) Alizadeth Data (b) Alon Data (c) Golub Data

Figure 5.8: Correlation between node degrees in the bi-partite and one-mode projection for three test datasets. There is a large spread in the data indicating that although for some gene nodes there is overlap, for others there is none.

5.3.1 Definition of Assessment Properties

For evaluation of any clustering technique and for its reusability, it is important that weighting schemes are validated independently of the subsequent network analysis and good quality results reflect a well-designed weighting scheme, together with a reasonably robust and efficient search algorithm. Both aspects are usually susceptible to considerable refinement. Consequently, we propose an edge weighting assessment procedure, based upon four properties, as detailed below.

Discrimination:

Ability of the method to "rate" highly those gene-sample couples which contribute to a cluster. The range and distribution of edge weights establish how well a given scheme distinguishes between relevant and irrelevant gene-sample couples.

Reusability:

Independence of the proposed scheme and the subsequent clustering technique. This deals with how/if the weighting scheme must change to reflect additional layers of analysis.

Robustness:

Ability of a given weighting scheme to deal with noise and missing values. This involves investigating the distortion of edge weights caused by different levels of noise and missing values.

Noise and missing values were added to the dataset to replicate measurement error of differing amounts. Noise was randomly "added/subtracted" to each value in the dataset as a percentage (up to 10%) of the original value. To replace data with missing values, up to 10% of expression values from the original dataset were randomly selected and removed. Commonly, in cluster analysis, missing values in the gene expression matrix are replaced by zeroes or by an average expression level of the gene, ("row average"). More sophisticated options include methods of K-Nearest Neighbour (KNN) and Support Vector Decomposition type, (Troyanskaya et al., 2001). To test our weighting schemes the common practice of replacing missing values by the row mean was adopted. Replacing missing values with the gene vector mean, in this scheme, equate the expression as unresponsive.

For this analysis we define *"Average Absolute Variation"* as the average difference in edge weights compared to 0% noise/missing values, while *"Stable weights"* are defined to be those weights for which the variation is less than the $\%$ level of noise/missing values added.

Parameter Influence:

The weighting scheme ideally should require minimal specification of input parameters. This includes consideration of input parameter influence on discrimination and robustness, as well as on the distribution of weights themselves.

5.3.2 Edge Weights in Bipartite Subgraphs

For each sample vector, j, expression values x_{ij} which indicate strong response of the i^{th} gene under j, (as determined in step one, Section 5.2.1) are selected. Similarly, genes which show moderate and weak response under j can be identified. For each of the three "strength of response" categories, a gene may be repressed or induced relative to the mean expression value; ($X - \mu < 0$, or $X - \mu > 0$ respectively), giving six sub-categories in total. For each sub-category, Cat_s, $s = 1 \ldots 6$ and for each sample variable, $j = 1 \ldots p$, the empirical probability of $x_{ij} \in Cat_s$ is calculated as:

$$|x_{ij} \geq x_{vj}|/|Cat_s|,\ x_{vj} \in Cat_s,\ i \neq v \tag{5.6}$$

(probability $= 1$ if $|Cat_s| = 1$) and hence the edge weight in the bi-partite gene expression graph is obtained.

The weight is thus, directly related to obtaining a given expression level *in a specific response category* for a particular sample. So, if many genes react strongly in the sample, the weight is smaller, while if, more interestingly, only a few react strongly, the weight will be larger. Note that, with this weighting scheme, a given sample (experiment) may also have no reacting genes.

The graph is broken down into an *independent subgraph for each sub-category*.

Gene-sample edges in these different response groups may have similar weight "values" but are distinguished, in terms of absolute levels of expression[6], by the category into which they fall, so that strength of response is important overall. Within a category however, the weight can be interpreted directly in terms of the relative probability of a gene-sample response. Thus the *higher the weight*, the more confidence that a relationship exists between gene and sample *in that category*.

5.3.3 Edge Weights in *All In One* graph

Obviously, transformation of the data into an *all-in-one* graph, requires scaling of edge weights to reflect the additional information in the graph.; Due to the design of the weighting scheme, an edge within the weak response category may have a larger weight (e.g. $e_{ij} = 1$), than an edge within the moderate response category (e.g. $e_{ij} = 0.2$), Fig. 5.9. These weights are rescaled in the *all-in-one* graph, to reflect the significance of an edge *for a particular sample* overall.

Each edge weight is rescaled to reflect the additional information in the graph. This is achieved simply by, (Fig. 5.10):

for all Sample nodes, j **do**

for all Induction or Repression Category, Cat_s **do**

if j is in Cat_s **then**

Multiply each edge weight incident to j by the degree of the sample node in the sub-graph

Add the degree of j in each of the lower sub-graphs, e.g. for strong sub-graphs weights, add the degree of j in the moderate and weak sub-

[6]If the dataset was not categorised, a weak response and a strong response gene-sample couple would have very different weights, with the consequence that the strong response couple would dominate the analysis and obscure more subtle patterns.

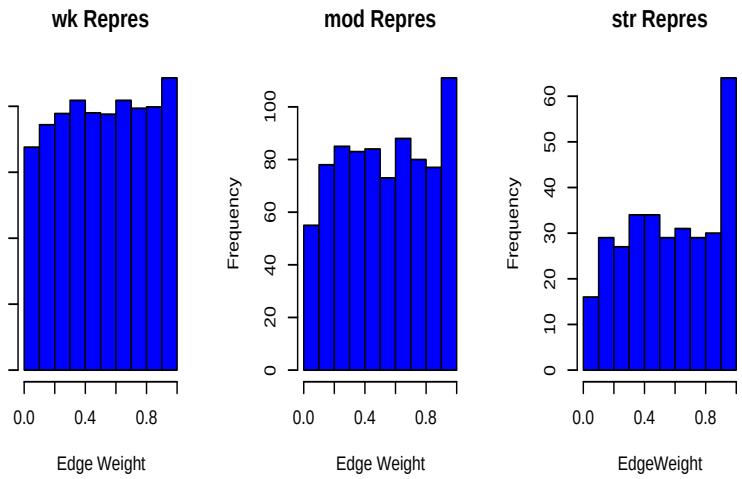


Figure 5.9: Edge Weight Distribution of each Spellman subgraph for Induced genes

graphs.

Likewise, for the moderate sub-graphs, add the degree of j in the weak subgraph

end if

end for

Divide by the degree of j in the *all-in-one* graph

end for

These edge weights still reflect the probability of getting that level of expression relative to the other genes being expressed *in that sample*.

5.4 Scheme evaluation

In this Section, we use the 4-point framework introduced above to analyse our novel weighting scheme, and to compare this with the scheme introduced in Tanay et al. (2002), (described in Section 3.3.4).

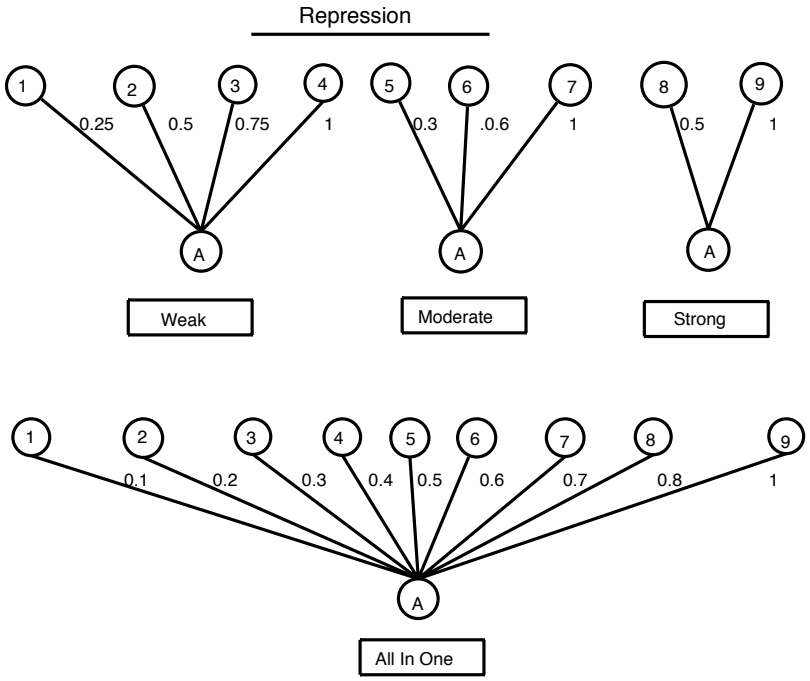


Figure 5.10: Rescaling edge weights from either Induced or Repressed Categories to weights in *all-in-one* graph, to reflect additional information.

5.4.1 Reusability

The empirical weighting scheme proposed here results in a partially-connected graph for each sub-category, since genes which do not show a significant change for a given sample do not generate an edge. Subsequent clustering techniques need to allow for this non-edge set, by optimisation of the objective function excluding nodes not connected by an edge. This scheme requires a dedicated algorithm for subsequent biclustering, which maintains the lists of gene-sample couples in each category, (i.e. strongly induced, moderately induced, etc.).

The Tanay scheme is independent of the subsequent clustering technique, as it results in positive edges for "interesting" gene-sample couples and negative edges for "non-interesting" gene-sample couples. Indeed, the scheme was specifically designed for an additive scoring system, where the sum of the edge weights in a subgraph corresponds to its statistical significance, ((Tanay, 2005) for more details). This scheme has also been applied to a compendium of information, and not just to gene expression data, (Tanay et al., 2005).

5.4.2 Parameter influence

The weighting scheme introduced here is controlled by a single parameter and is, therefore, easily configurable while offering some flexibility. This parameter is κ, (Eq. 5.1), which determines thresholds between categories. Table 5.4 illustrates the results of the threshold analysis for the Alizadeth dataset. The maximum threshold for which any gene-sample couple was identified in the real dataset was $\kappa = 7.07$ (probability ≤ 0.02, (Eq. 5.1)). Thresholds of $\kappa = 5, 4.08, 3.58$ and 3.162, (i.e. probabilities $(\frac{1}{\kappa^2}) \leq 0.04, 0.06, 0.08$ and 0.10 respectively) were then tested. The

Strong					
κ	7.07	5	4.08	3.58	3.16
P	0.02	0.04	0.06	0.08	0.10
$log(\frac{P(X=Edge)}{P_{rand}(X=Edge)})$	undef	undef	1.69	1.02	0.62
Moderate					
κ	3.162	2.88	2.67	2.5	2.35
P	0.10	0.12	0.14	0.16	0.18
$log(\frac{P(X=Edge)}{P_{rand}(X=Edge)})$	1.94	1.76	1.62	1.55	1.45
Weak					
κ	2.23	2.13	2.04	1.96	1.88
P	0.20	0.22	0.24	0.26	0.28
$log(\frac{P(X=Edge)}{P_{rand}(X=Edge)})$	0.11	0.08	0.04	0.04	0

Table 5.4: Threshold Analysis, Alizadeth data. κ= the number of standard deviations from mean (Eq. 5.1). P = $\frac{1}{\kappa^2}$, is the probability that values are $\kappa\sigma$ from mean. The maximum log-ratio is taken as the threshold between categories.

log ratio of the probabilities is maximised at $\kappa = 4.08$, and this was taken to be the strong response threshold. Thresholds for moderate and weak response were then similarly deduced to be $\kappa = 3.162$ and $\kappa = 2.23$, respectively. Table 5.1 provides results of the threshold analysis for three test datasets.

The main parameters for the Tanay weighting scheme are, similarly, thresholds between categories, and P_c, (the constant probability that an edge appears in a bicluster, Eq. 3.10). Thresholds between categories are arbitrarily chosen, based on normalized ranked values within each sample and are, therefore, not directly data dependent. This 'hard thresholding' has consequences for the deterioration of the scheme when noise and missing values are added to the data. As the threshold parameter is lowered, a higher percentage of edges will be identified, even if none exist. (For a more detailed discussion of parameter P_c see Tanay et al. (2002)).

5.4.3 Robustness

The influence of noise and missing values are summarised respectively in Table 5.5 for the Alizadeth dataset. Results for other datasets are not displayed here but are broadly consistent with those presented.

The absolute variation in weights is extremely low for the empirical scheme since the technique examines extreme values, i.e. values which appear in the tail of the distributions of each gene variable. In addition, weights are not based directly on a given expression value, but on that expression value *relative* to other values in the category for a particular sample (Step 2 of scheme, see Fig. 5.4). The category is also defined *relative* to expected value of the gene variable, (Step 1 of scheme, see Fig. 5.4). As "missing" values are replaced by the row mean, this does not greatly affect extreme values. Equally, even noise added at 10% level of the original values does not affect *relative* values, so that, perturbations in the data have small effect on weights assigned.

Similar to results shown in Table 5.5, for the Stegmaier dataset, average absolute variation in edge weights is $\sim 0.26\%$ for an added noise level of 10% (not shown), while denoting 10% of the dataset as 'missing', gives average absolute variation in values $\sim 0.3\%$, while 99.5% of weights are stable. For the Cho dataset, the corresponding values for 10% noise added were: $\sim 0.22\%$ (absolute variation) and $\sim 99.65\%$ (stable weights); and for missing values at 10% was: $\sim 0.15\%$ (absolute variation) and $\sim 99.69\%$ (stable weights).

Using the Tanay weighting scheme, perturbations in the data have very little effect on weights derived: (similar results for all tested datasets). We were surprised by this result and tested missing values up to a level of 80% however the effect was still minimal (0.01%, average variations and 99.99% stable weights). It may be

% Noise level	1.5	2.5	5	10
Empirical Based				
% Average Absolute variation	6×10^{-2}	2×10^{-2}	3×10^{-2}	5×10^{-2}
% "stable" weights	99.66	99.68	99.68	99.65
Tanay Scheme				
% Average Absolute variation	2.7×10^{-3}	2.6×10^{-3}	2.3×10^{-3}	3.2×10^{-3}
% "stable" weights	99.98	99.98	99.99	99.99
%missing values	1.5	2.5	5	10
Empirical Scheme				
% Average absolute variation	0.04	0.04	0.09	0.16
% of "stable" weights	99.70	99.72	99.64	99.62
Tanay Scheme				
% Average absolute variation	0.0023	0.0027	0.0028	0.0069
% of "stable" weights	99.98	99.98	99.99	99.99

Table 5.5: Influence of noise level and missing values on weights assigned

the case that this scheme identifies 'interesting' gene-sample couples, even if none exist, due to the 'hard' threshold nature of the scheme and its reliance on a ranking system. If 10% noise is added to the dataset, thresholds still depend *on ranking and not a calculated mean level*, thus approximately the same gene-sample couples are selected as interesting as the ranked position is not changed. Missing values also have little effect, as these are replaced by the mean, and the same thresholds used so that decisions are more conservative if anything.

5.4.4 Discrimination

For this analysis, a 'random graph' refers to a graph created from a random dataset, as described in Section 4.5.2.

From the threshold analysis for the Empirical scheme, described above, maximum discrimination between empirical and random graphs is obtained. As expected, the largest number of gene-sample couples falls in the weak response cat-

egory. Discriminating between gene responses clearly depends on the category thresholds used and, while threshold derivation described is based on statistical considerations, this can obviously be augmented by biological information. From Table 5.6, for strong and moderate response, the probability of an edge existing between a gene and sample node in the real graph is greater than that for the random graph, indicating that significant structure is present. For a weak response, the ratio of probabilities is smaller and it is less convincing that real differences exist. Nevertheless, an examination of the average degree of sample nodes in the real graphs indicates that the average number of genes responding is higher than expected. For example, for the weak repression sub-category, a sample node is, on average, connected to $\sim 3.2\%$ of gene nodes compared to $\sim 1.2\%$ in the random graph, $(d_{\perp}/n_{\top})$. The average degree of a sample node in the real graph is much higher than expected, $(> m/n_{\perp}$ with degree $\geq 0 = 96)$. This suggests that, although the ratio of edge probabilities in the weak response category compared to the random graph is not high, some pattern structure is present and the method is capable of identifying 'indicative' gene-sample couples, even in this less-reactive category.

From our analysis of the Tanay et al. scheme (Table 5.7), we observed that (a) a smaller number of total positive weights were identified compared to those identified from corresponding graphs generated from a random dataset (Section 5.2.2), (b) the number of positive weights in each category is roughly equivalent to that for the random case(the exceptions are Stegmaier Moderate and Strongly repressed categories). Observations (a) and (b) imply that the amount of edge 'sharing' of gene-sample couples in the real dataset is considerable, (i.e. gene-sample couples having a positive edge in the weakly induced category, also feature in the strongly induced category) - thus categories are not mutually exclusive. Since 'hard' thresh-

	Repres			Induc		
	Wk	Mod.	Str.	Wk.	Mod.	Str.
$n_\top$	1889 (2004)	296 (6)	63 (5)	2088 (2030)	476 (4)	209 (2)
$n_\perp$	53 (82)	13 (2)	16 (2)	79 (85)	43 (2)	46 (2)
m	3227 (2754.04)	327 (6.01)	64 (2.92)	3713 (2779.27)	550 (5.57)	219 (2.89)
$d_\top$	1.7 (1.4)	1.1 (1)	1.01 (1)	1.77 (0.98)	1.15 (0.79)	1.04 ($6.6e^{-4}$)
$d_\perp$	60.89 (35.23)	25.15 (1.02)	4 (1)	47 (34.95)	12.79 (2)	4.7 (1)
δ	0.01 (0.003)	0.001 ($1.7e^{-5}$)	$2.2e^{-4}$ ($1e^{-5}$)	0.012 (0.009)	0.002 ($2e^{-5}$)	$7.5e^{-4}$ ($1e^{-5}$)

Table 5.6: Categories for Alizadeth data, created by cut off thresholds 0.20 (weak induction/repression), 0.10 (moderate induction/repression), and 0.06 (strong induction/repression). $n_\top$ = the active set of genes (gene nodes with degree ≥ 1), $n_\perp$ = active set of samples, m = number of edges, $d_\top$ = average degree of active set of genes, $d_\perp$ = average degree of active set of samples, δ = bi-partite density i.e the fraction of existing links with respect to possible ones (i.e. gene nodes with degree $\geq 0\times$ sample nodes with degree ≥ 0). Numbers in brackets indicate corresponding values for random graphs.

Datasets	Aliz.	Cho.	Steg.
% total edges	2.7	3.34	2.13
	(2.9)	(4.3)	(3.6)
Induced			
% strong	1.2	1.1	0.8
	(1.1)	(1.2)	(0.8)
% moderate	3.2	4.1	1.0
	(3.1)	(3.9)	(1.0)
% weak	3.9	5.0	1.1
	(3.9)	(4.8)	(1.1)
Repressed			
% strong	1.2	1.1	1.3
	(1.1)	(1.2)	(1.4)
% moderate	3.2	3.8	3.8
	(3.1)	(3.9)	(3.4)
% weak	4.0	4.7	4.7
	(3.8)	(4.8)	(4.1)

Table 5.7: Tanay scheme - percentage of total possible edges is taken as Number of Positive Edges$\backslash n \times p \times 6$ categories whereas percentage of edges in each category is expressed in terms of the total possible edges in that category i.e. Number of positive Edges$\backslash n \times p$, (n = number of genes, p = number of samples). Bracketed values represent results from random graphs.

olds between categories were used and arbitrarily chosen, the order of the number of edges in each category is $Strong < Moderate < Weak$. Note also that those gene-sample couples, evaluated as strongly reacting, will have a magnified impact on any clustering procedure for the resulting graph, due to the overlap between categories, (i.e. weak influences are a subset of strong influences).

5.5 Summary

We have presented an empirical-based method for the extraction of a bi-partite graph from a gene expression dataset. This method is important because it builds the gene expression graph in a data dependent manner. Analysing gene expression

datasets is important as it allows for fine grained analysis of the reactivity of genes at various response levels.The scheme for constructing graphs presented here results in independent non-overlapping sub-graphs, each representing a strength/type of response category. These can then be used to construct, what we have called an "all-in-one" graph, where the interactions and reinforcement of response groupings from combined categories can be analysed. This can be used to determine subtle coherence in patterns of co-expression. Using tools and notation for classic network analysis for extraction, identification and analysis, we have uncovered organisational structure in graphs, constructed with this scheme. To our knowledge, this is the first time basic network analysis techniques for bi-partite graphs have been applied to the analysis of gene expression. Clustering coefficients, cc, were obtained for gene and sample node sets individually with that for gene nodes found, unsurprisingly, to be much larger than that for sample nodes, due to the large disparity in average degree between sets. In the latter, the cc were dominated by sample nodes of high degree. Consideration of the minimum clustering coefficient measure removes large neighbourhood bias, and reveals there more subtle, local interactions in the data. We examined the size of the neighbourhood intersections for genes, and found that, genes react more coherently in larger neighbourhoods than expected by chance. An examination of the gene node degree distribution of the extracted graphs suggests that has some Poisson characteristics, but suggests that for large gene sets is well-approximated a Normal distribution. However, the size of the sample node set for most datasets is relatively low, so that degree distribution is less easy to establish and appears to depend on the observed data itself.

The issue of data-dependent threshold estimation is addressed in the empirical-scheme presented, but is non-trivial, as numerous thresholds need to be assessed, which is computationally expensive. The scheme presented here is more specific in

that fewer gene-sample couples are identified than in the Tanay et al. scheme. For example, for the Alizadeth dataset, our scheme extracts $\sim 3\%$ edges (at optimal threshold levels). The Tanay scheme extracts $\sim 16\%$ for the same dataset. In real terms this means that the denser graphs created with the Tanay scheme, could contain more irrelevant gene-sample couples, while our scheme can more precisely target the gene set of interest.

We have presented a novel edge-weighting scheme for gene expression bipartite graphs. From the investigations, presented in this chapter, it is clear that interpretations of edge weights in graphical gene expression schemes can be difficult and a comparative analysis with the well-known Tanay et al. (2002) scheme is also presented. This analysis was carried out w.r.t four major properties: reusability, parameter influence, robustness and discrimination. Both schemes result in positive weights for interesting gene-sample couples. Our new weighting scheme, for a particular sample j, determines affected genes *relative to other gene expression values for that sample* j. This is an important feature, as absolute level of gene expression is not directly accounted for, but rather the fact that change occurs, together with the significance of this change relative to the majority of genes. Relative evaluation is also an intrinsic feature of the Tanay scheme as the initial probability $\phi(i, j)$, Eq 3.10 is based on ranks. However, the selection of a pre-determined thresholds (between ranks) with the Tanay scheme has a large effect on robustness and discrimintation as it is *not data dependent*. Overall, therefore, weights resulting from the Tanay scheme seem little affected by noise and missing values, which indicates that this scheme could assign high weights to gene-sample couples, even if none are present. Our contention, therefore, is that this ignores the subtleties in the data, and selects 'interesting' gene-sample couples based on absolute values, (including noise levels). With our empirical scheme, however, small sample size (number of

microarray experiments), the performance deteriorates with respect to the random graph comparison basis, (difficult to estimate μ and sd for each gene variable), and the thresholds between categories become increasingly difficult to identify.

We have also demonstrated that edge-weighting schemes should be considered to be independent of the subsequent clustering procedure in the sense that they should satisfy intrinsic requirements and be internally consistent. Alternative edge-weight derivation can be seen as providing different probes for data interrogation leading to complimentary interpretations. This type of assessment framework for weighting is not unique to gene expression data, but is also crucial for other applications, generating large, complex datasets.

Using graphical techniques to extract meaningful information from biological data is both intuitive and valuable. In this chapter, we have limited our investigation to bi-partite graphs, a representation which captures essential properties of gene expression datasets and allows for the extraction of suitable bi-clusters. In the next chapter we investigate how information from these bi-partite graphs can be used to find important gene sets.

CHAPTER 6

PARTITIONING GENE EXPRESSION GRAPHS OF LOCAL INTERACTIONS

In Chapter 5, a framework for creating weighted bi-partite gene expression graphs and highlighting interesting properties of these was outlined in Chapter 5. Here, a method is presented for the construction of a one-mode gene expression network which specifically focuses on *local interactions* of genes, (i.e. across a subset of samples). This approach permits use of classical network analysis tools and adds to previously published work in the area, (Stuart et al., 2003; Carter et al., 2004; Bergmann et al., 2004; Zhang and Horvath, 2005). Specifically, no developed framework exists to date for the construction and examination of a one-mode gene expression network which *captures local structures* of a dataset. In what follows we present our analysis, and validation, of such a framework. We present a method to extract cliques from the network of local interactions, and provide compelling evidence that cliques extracted from a suitably constructed graph, represents meaningful biological structures.

6.1 Introduction

Gene co-expression networks provide a straightforward mechanism to explore system-wide functionality of genes, by using intuitive network concepts to analyse complex interactions. However, it is essential that relationships between genes in the network are captured in a meaningful manner. We address this issue by constructing a gene expression network from the underlying bi-partite graph. Nodes are connected in a one-mode graph, if the corresponding genes are significantly co-expressed across a subset of samples in the bi-partite graph, reflecting local gene interactions. However, in the one-mode projection information, such as the number of samples which a gene shows similar expression can be lost (Chapter 5). This is overcome by using a weighting scheme to capture this information, where edges are weighted in the one-mode graph to reflect the size and significance of the intersection in the bi-partite graph.

This represents a further development and improvement on previous gene expression network analysis, (Stuart et al., 2003; Carter et al., 2004; Bergmann et al., 2004). In their work each node in the network represents an expression profile of a given gene, and an edge represents a significant pairwise expression profile association across *all* samples. Zhang and Horvath (2005) also use an adjacency function to weight the edges, which assumes that connections between nodes approximate a scale free topology. The framework presented here does not require such assumptions and node connections are more biologically plausible as these represent significant pairwise gene expression across a *subset* of samples (i.e. identifies gene groups with similar expression for this set of samples and divergent expression otherwise). We describe an approach where the graph encoding is built from the bi-partite version and designed to reveal local interactions in the data, hence the no-

tion of *similarity*, (weighted edges), is significantly different from previous work. This leads us to define edge weights in the one-mode graph, based on information obtained from the bi-partite graph, and an "intersection" score.

The main purpose of the network analysis is to use gene connectivity information to group genes according to function and to relate to external gene information. For each graph constructed, gene subgroups are identified which show similar expression across a subset of samples. Many graphical clustering procedure have been proposed, Kernighan-Lin algorithm (Kernighan and Lin, 1970), Simulated Annealing (Johnson et al., 1989), Path Optimization (Berry and Goldberg, 1995), Genetic algorithms (Bui and Moon, 1996), MinMax clustering (Ding et al., 2001), heuristic search using hash tables (Tanay et al., 2002) and others. We use a number of graph properties in a clustering procedure to identify biologically meaningful groups.

6.2 Graph Properties

A *one-mode* graph, $G = (U, E, W)$, is constructed, where U is the set of gene nodes and $n = |U|$, $e_{ij} \in E$ is the set of edges - an edge exists between two gene nodes u_i and u_j ($i \neq j$), *if they show similar co-expression across at least two samples* (i.e. their intersection neighbourhood in the underlying ***all-in-one*** bi-partite graph is ≥ 2).[1], $w_{ij} \in W$ is the set of edge weights associated with the edge between nodes u_i and u_j, $\forall e_{ij} \in E, i \neq j$.

Once the network has been constructed several biologically important network concepts can be identified which relate connectivity information to external gene information. Descriptive statistics of the one-mode graphs, extracted for the test datasets, are given in Table 6.1.

[1]This could be easily extended to > 2 sample neighbours. We focus on the all-in-one graph as we want to find coherent bi-clusters i.e. genes which may be alternately expressed in a given sample.

	n	m	d	density	cc	cc_{min}	γ
Aliz	2349	90624	77.15	0.03	0.66	0.69	1.1 (2.09 tail)
Alon	1205	19404	32.2	0.03	0.72	0.72	1.2 (2.33 tail)
Cho	27	91	6.7	0.26	1	0.84	-
Gasch	3086	1348783	874.13	0.28	0.69	0.77	0.76 (1 tail)
Golub	1974	20931	21.2	0.01	0.68	0.68	1.26 (2.45 tail)
Hsiao	1163	42243	72.64	0.06	0.84	0.88	1.18 (1.12 tail)
Spell	2736	176248	122.25	0.04	0.62	0.67	1.89 (2.47 tail)
Steg	549	14189	51.69	0.09	0.94	0.89	1.08 (2.06 tail)
West	1556	102462	131.69	0.08	0.86	0.87	1.09 (1.58 tail)
Random Graphs							
Aliz	2392	16747	117.58	0.005	0.16	0.20	1 (1.6 tail)
Alon	1296	32093	49.52	0.03	0.13	0.18	1.1 (2.15 tail)
Cho	36	56	3.11	0.08	0	0	-
Gasch	3094	3002425	970.40	0.62	0.19	0.25	0.75 (1 tail)
Golub	2052	30457	29.69	0.02	0.05	0.10	1.19 (2.65 tail)
Hsiao	1199	37893	63.43	0.05	0.20	0.26	1.12 (1.12 tail)
Spell	2763	275068	199.10	0.07	0.18	0.20	0.92 (1.67 tail)
Steg	585	10915	37.31	0.06	0.18	0.25	1.16 (2.28 tail)
West	1631	149530	183.36	0.11	0.33	0.40	0.98 (1.56 tail)

Table 6.1: one-mode Graph Properties: For each real dataset: n, the number of active genes, i.e. $U' \in U$ with $d \geq 1$, m, the number of edges between gene nodes, d, the average degree of the graph, cc, the unweighted clustering coefficient for the graph, γ the value of the exponent of the power law that best fits its degree distribution based on maximum likelihood estimation. The value in brackets refers to the power law estimate for values $log(d) > 1$

Descriptive Statisitics:

For these gene expression test datasets, the number of gene nodes, n, is small compared to typical real world complex networks, (see Guillaume and Latapy (2004) for examples). It is clear that the size of the one-mode graph (nodes and number of edges) is dependent on the number of sample nodes in the underlying graph, Table 6.1, Fig. 6.1. Not typical of most real world complex networks is the observation that the average node degree, d, of each graph is quite high compared to n, and varies depending on the dataset under observation. (The Cho dataset is indeterminate, which has low n and d, thus has limited information encoded.) The average degree of gene nodes is independent of the number of sample nodes in the underlying graph. For example, for the Alon data, genes are connected on average to 3% of the nodes, while for the West data genes connect on average to 8% of the nodes - the sample node set size in the underlying graphs are 57 and 48 respectively. There are, on average, more nodes in the random graphs[2], which suggests that there is less overlap in sample node neighbourhoods in the underlying random bi-partite graph compared to real, Fig. 6.2.

The density of all graphs is quite low, with the exception of the Gasch and Cho dataset. These datasets represent the maximum and minimum size graphs in terms of number of nodes. For the Cho dataset, n is small and, as noted, the intersection neighbourhoods in the underlying bi-partite graph are snot extensive, resulting in a larger number of neighbours in the one-mode projection. The Gasch dataset contains a large number samples from experiments under extreme conditions, resulting in a large subset of genes responding. If the number of samples for which

[2]Random graphs were created by projecting uniformly sampled random bi-partite graphs (created with monte carlo edge switching algorithm) into one-mode.

two genes have similar expression is increased to qualify for an edge in one-mode graph (eliminating spurious measurements), to, for e.g., 5, the density of the Gasch graph reduces to 0.03.

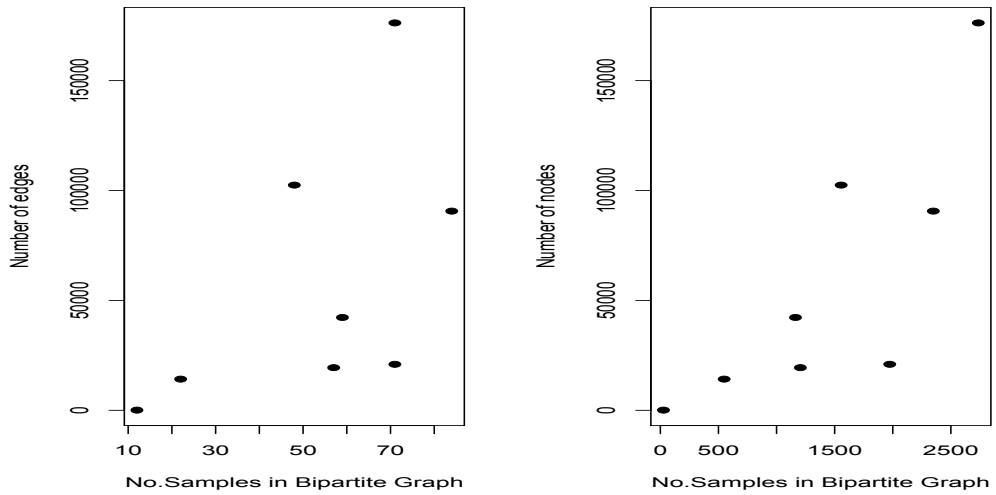


Figure 6.1: The size of the one-mode graph (number of nodes and edges) is dependent on the number of samples in the underlying bi-partite graph.

Clustering:

The clustering co-efficient, cc, in the one-mode gene expression graphs are again high, compared to the corresponding random graphs (Table 6.1), despite most nodes not begin linked (lower density), i.e. this implies that graphs have locally dense structures. Random networks (both the Erdös-Rényi model and the Barabasi-Albert model), result in a small clustering coefficient, which corresponds to cc values generated for the random case. This indicates that there is a *non-trivial* level of organisation in the graph, such that, if two genes are connected, they are more than likely to have similar neighbourhoods.

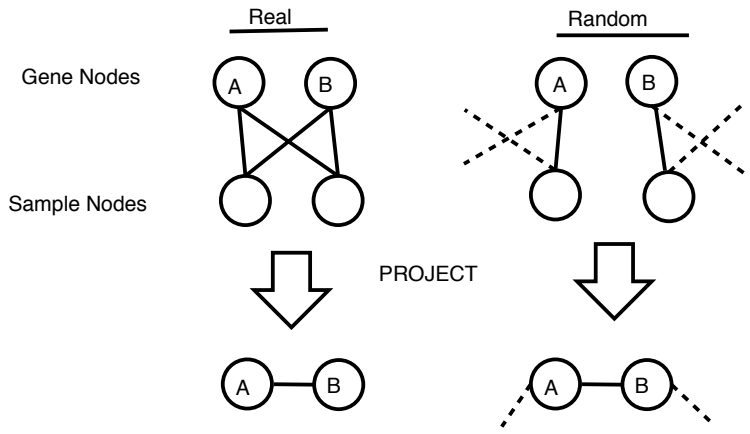
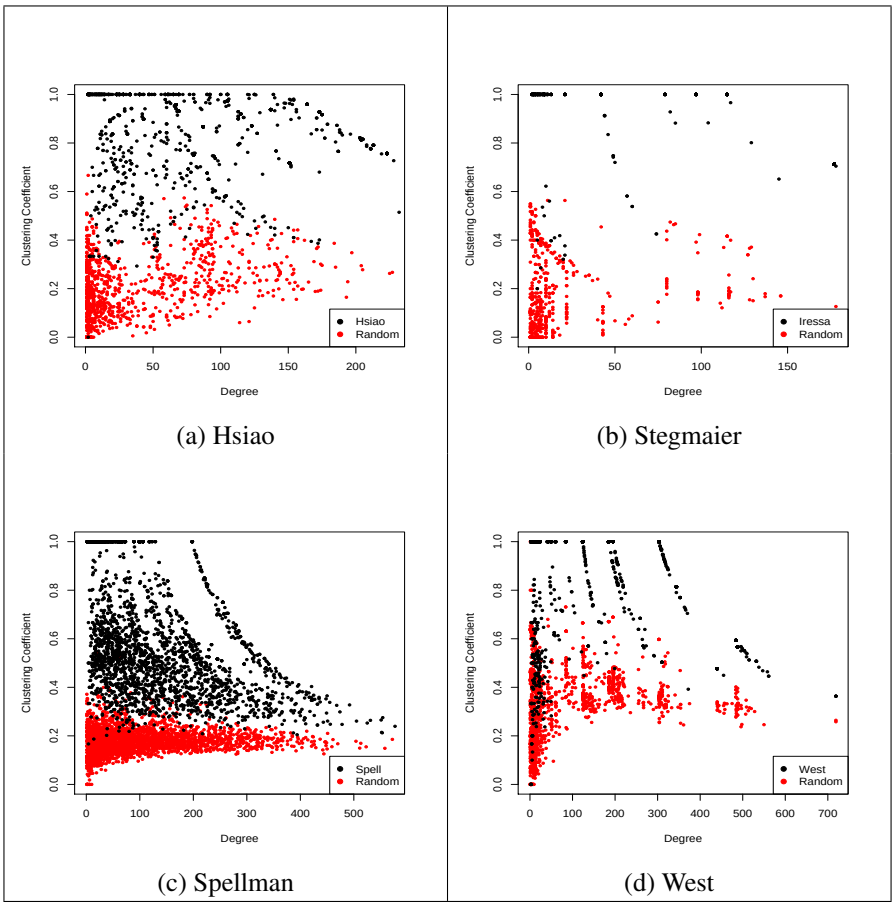


Figure 6.2: There are smaller intersection neighbourhoods in the underlying random graphs compared to real, resulting in a higher average number of nodes.

As suggested in Chapter 5, cc can be dominated by the gene node with the larger neighbourhood. Thus we examine the minimum clustering coefficient, cc_{min}, (Eq. 5.5, (Latapy et al., 2006)). This measure does not deviate much from cc, suggesting that high degree nodes do not dominate. To further investigate this, the relationship between the average clustering coefficient for a node and its degree was analysed. The clustering coefficients, both cc and cc_{min}, have a large spread for gene nodes of small degree, which are the majority, Figures 6.3, 6.4. For the fewer nodes of high degree, cc decreases, while cc_{min} increases (as its denominator is the minimum neighbourhood).

For all graphs, the values for the random graphs are well below the values for the test datasets. This shows that the values of cc and cc_{min} are *larger in real gene expression datasets and the difference is substantial*. There is an inverse relationship between cc and node degree in the real datasets, while it remains roughly constant in the random case. This was also found by Ravasz et al. (2002) in a study of metabolic networks, who suggested that it was due to a hierarchical structure within the network. In the case of cc_{min} there is a positive relationship with node

degree for both the real and the random graphs.

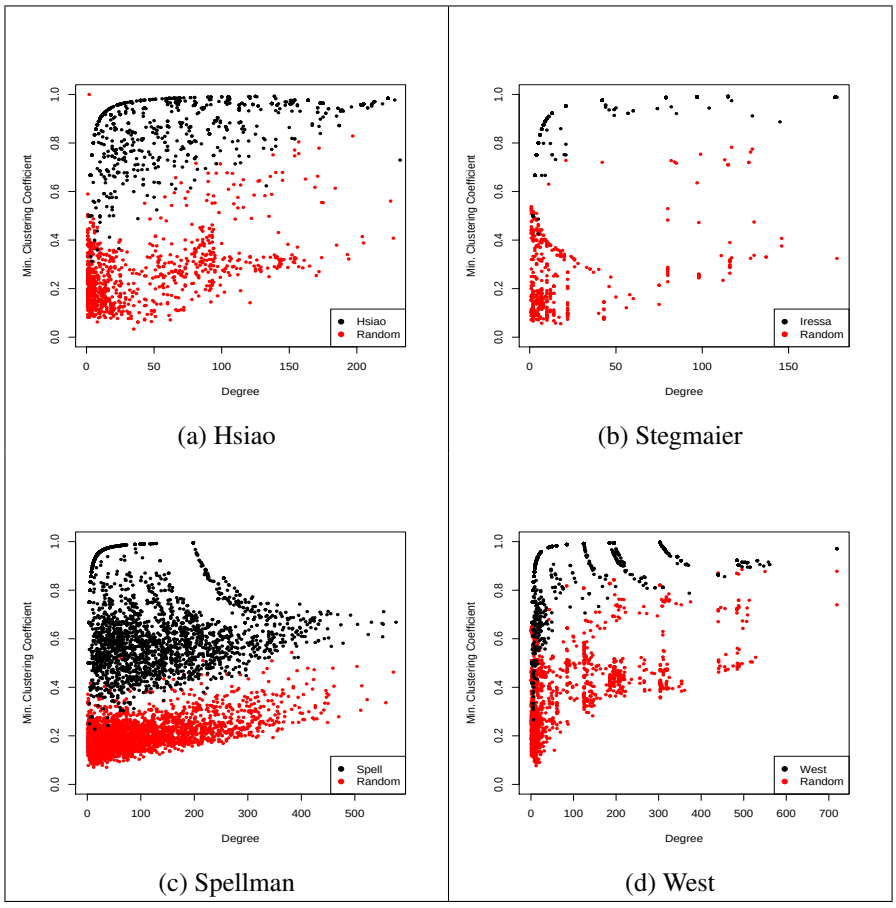


(a) Hsiao (b) Stegmaier

(c) Spellman (d) West

Figure 6.3: Average Clustering Coefficient of Nodes Vs. Degree

In Fig. 6.5 the cumulative distribution of the clustering coefficients for four test datasets are shown. In the random case the value of cc_{min} grows very quickly and are close to unity at low values of cc_{min}. This means that cc_{min} are very small for most nodes. That is, the intersection of the neighbourhoods is quite small compared to the minimum neighbourhood size. The value of cc_{min} grows much less quickly for the real datasets and remain lower than unity for a long time. This means that for an important number of nodes, cc_{min} is large, closer to one in most cases - the neighbourhoods of many nodes significantly or completely overlap with other node



(a) Hsiao (b) Stegmaier

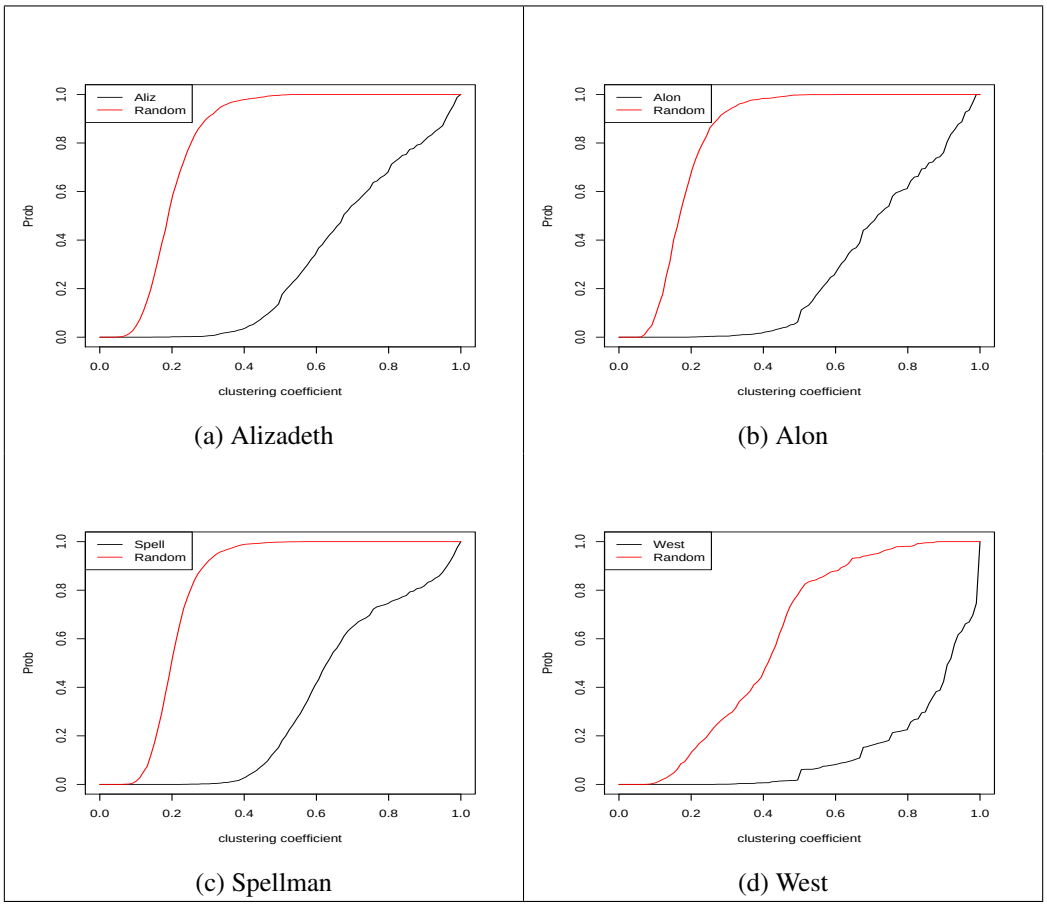
(c) Spellman (d) West

Figure 6.4: Average Clustering Coefficient of Nodes Vs. Degree.

neighbourhoods.

Degree Distribution:

For $\sim log(d) > 1$, (where $d =$ node degree), approximate power-law behaviour is observed for the node degree distributions, implying that its heterogeneous nature is non-trivial, i.e. the networks are scale-free, Figures C.26 - C.27 Appendix C. This implies that the node degree distribution exhibits greater heterogeneity for one-mode compared to bi-partite graphs, with most genes having small degree and only few having high degree. Exceptions include the Gasch and Hsiao datasets (Figures C.26d,C.27b, Appendix C), for all d. Recall that an edge in the one-mode



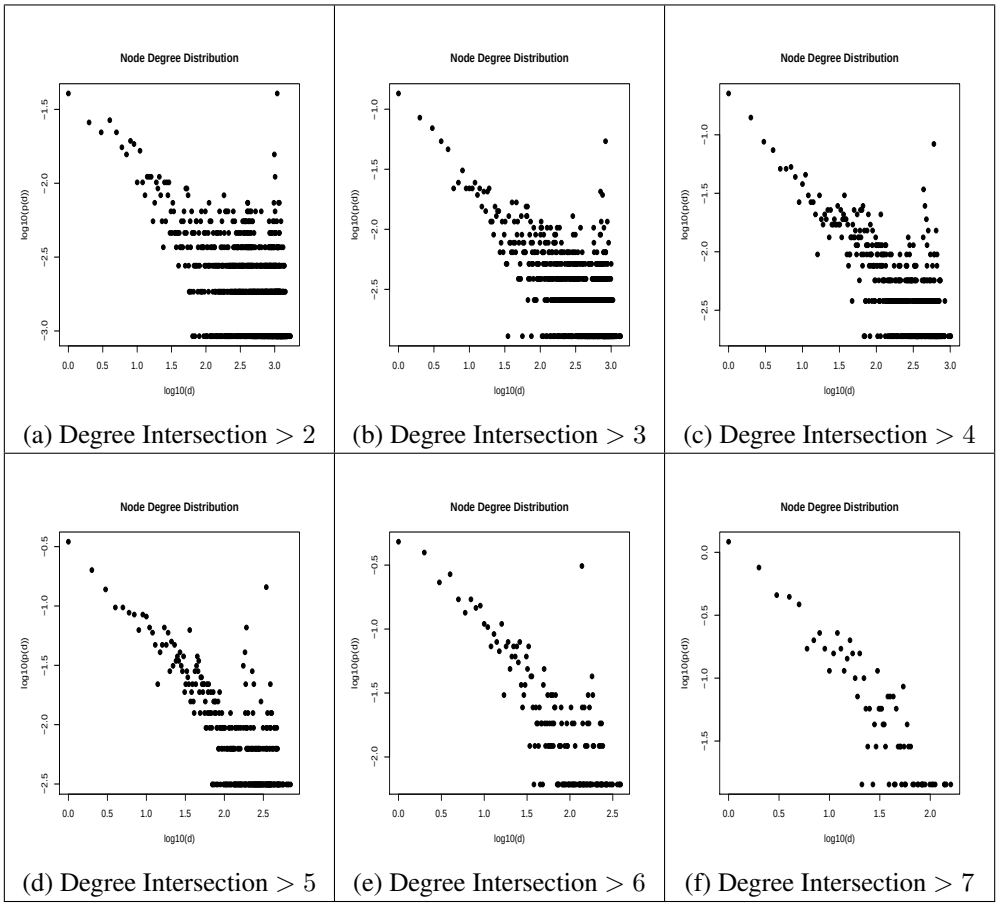
(a) Alizadeth (b) Alon

(c) Spellman (d) West

Figure 6.5: Cumlative distribution of the minimum clustering coefficient. For each value on the x-axis the probability of all the nodes having lower than x for cc_{min}

graph results when two genes show similar expression across samples. As with the density information, if the condition of number of samples for which gene expression is similar is increased to qualify for an edge in the one-mode graph (eliminating spurious measurements) a power-law distribution emerges, Fig. 6.6. Interestingly, Gasch and Hsiao are the datasets with near normal behaviour in the gene node degree distributions in the underlying bi-partite graph, suggesting that the near-normal behaviour of these underlying datasets is the limiting case, and increased sample intersection size requirements need to be applied when creating these graphs.

From this investigation, it is clear that there is structure in the constructed gene



(a) Degree Intersection > 2 (b) Degree Intersection > 3 (c) Degree Intersection > 4

(d) Degree Intersection > 5 (e) Degree Intersection > 6 (f) Degree Intersection > 7

Figure 6.6: As the requirement for intersection neighbourhood increases, the degree distribution conforms to a power law distribution

expression graph, that is not typical of a random network, and projecting into one-mode does reveal new information. There is considerable sharing among neighbourhoods, (high clustering coefficients), indicating that these are co-active in similar samples. Although the underlying gene node distributions in the bi-partite graph were approximated by the Poisson, the node degrees in the corresponding one-mode graph are more closely modelled by a power law. This indicates that the topology of the network is dominated by a few highly connected genes which link the rest of the less connected genes to the network, i.e. genes are preferentially attached to genes of high degree.

6.2.1 Edge Weights in one-mode Graph

We have already discussed how two genes are linked in a one-mode projection of a bi-partite graph if they have a sample neighbour in common. If two genes were truly co-expressed, they would be co-expressed in more than one sample i.e. the intersection neighbourhood of the genes would be large. Here, we define the weighted strength of this interconnection as the "interconnection coefficient", Eq. 6.1.

$$cc_{inter}(u,v) = \frac{\sum_{j \in \{N(u) \cap N(v)\}} w_{uj} w_{vj}}{|N(u) \cap N(v)|} \tag{6.1}$$

where w_{uj} and w_{vj} is the weight of edges $(u, j)$and (v, j) in the *bi-partite graph* respectively, and $N(u)$ and $N(v)$ are the neighbourhoods of gene nodes u and v in the *bi-partite graph* respectively. This quantifies the level of confidence in the co-expression of two genes, defined through edge weights in the bi-partite graph. For example, if two genes have a high significance of expression(i.e. large edge weights) under three samples, the interconnection coefficient will be close to one, on the other hand if one gene has a high significance and the other a low significance under three samples it will be closer to 0.5, and if both genes have a low significance under three sample the interconnection coefficient will be close to 0. Hence, *the weight of a link between two genes u and v in the one-mode projection captures the weight of the intersection of their neighbourhoods*. This weight corresponds to a similarity measure as it is non-negative and symmetric.

6.3 Detecting High Scoring Coherent Modules - *GraphCreate*

A principal objective is to detect meaningful subsets of genes which are tightly connected to each other across a subset of samples, i.e. to detect modules in the one-mode graph and to develop a scoring scheme which takes into account the samples for which these genes show similar expression. Here we introduce a new method, *GraphCreate*, designed to detect high scoring modules. A module of genes is defined by considering nodes with high neighbourhood overlap. We have already shown that if two genes have a neighbour in common, they are more likely to have similar neighbourhoods when compared to a random graph. All potential modules, $1...L$ are identified by considering, for each gene, u, neighbours of u which have an intersection neighbourhood greater than a predefined threshold (I) with u, i.e.

$$L = \{N(u) \cap N(j)\} > t \; {}_{\forall j \in N(u), L=1...n} \tag{6.2}$$

This identifies a maximum of n, (= number of gene nodes) potential modules. Modules are processed to remove those with a high degree of overlap between their nodes.

Which, or the number of, samples gene nodes exhibit co-activity under is considered by a bit string associated with each gene node, u, in the one-mode graph. The bit string has length equal to the number of samples nodes in the bi-partite graph, where a 1 at position s indicates that u shows response in sample s. For each bicluster formed by gene u define:

$$M_\ell = max[sum\{bs(i) \wedge bs(j)\}] \; {}_{\forall i,j \in \ell, \ell=1...L} \tag{6.3}$$

and

$$m_{ij\ell} = \frac{bs(i) \wedge bs(j)}{M_\ell}$$

where $bs(\cdot)$ is the bit string associated with the node $(\cdot)$, $\wedge$ = logical AND of each position in the bit string. (Note: $bs(i) \wedge bs(j)$ is the intersection size of $N(i)$ and $N(j)$ sample neighbourhoods in the bi-partite graph.) M_ℓ is the maximum intersection for the bicluster, ℓ, formed by nodes in the module. It is a factor which scales the weight of an edge between two gene nodes based all gene nodes in the modules, see Fig. 6.7. This scaling factor for an edge weight will change between modules, depending on its membership, i.e. the edge weight will have greater or lesser significance depending on which genes it is grouped with. Thus the weight of a module can be found by:

$$W_{bicluster} = \sum_{i,j \in \ell} w_{ij} m_{ij} \tag{6.4}$$

where w_{ij} is weight of the edge between gene's i and j in the one mode projection, defined in Eq. 6.1. This weighting scheme reflects the intersection of the two genes and the samples for which genes show similar expression.

6.4 Representative Modules Found

In practice, the search space is restricted by searching only gene nodes with a minimum degree $> d_t$. Thus, there are two parameters affecting the number and size of gene groups found by *GraphCreate*: the size of the intersection neighbourhood, I, and d_t. Fig. 6.8 illustrates how these parameters affect the number, size and weight of modules found in the Hsiao dataset. Similar results, found for other datasets, are not given here. In general, the number of modules, K, decreases as d_t and I

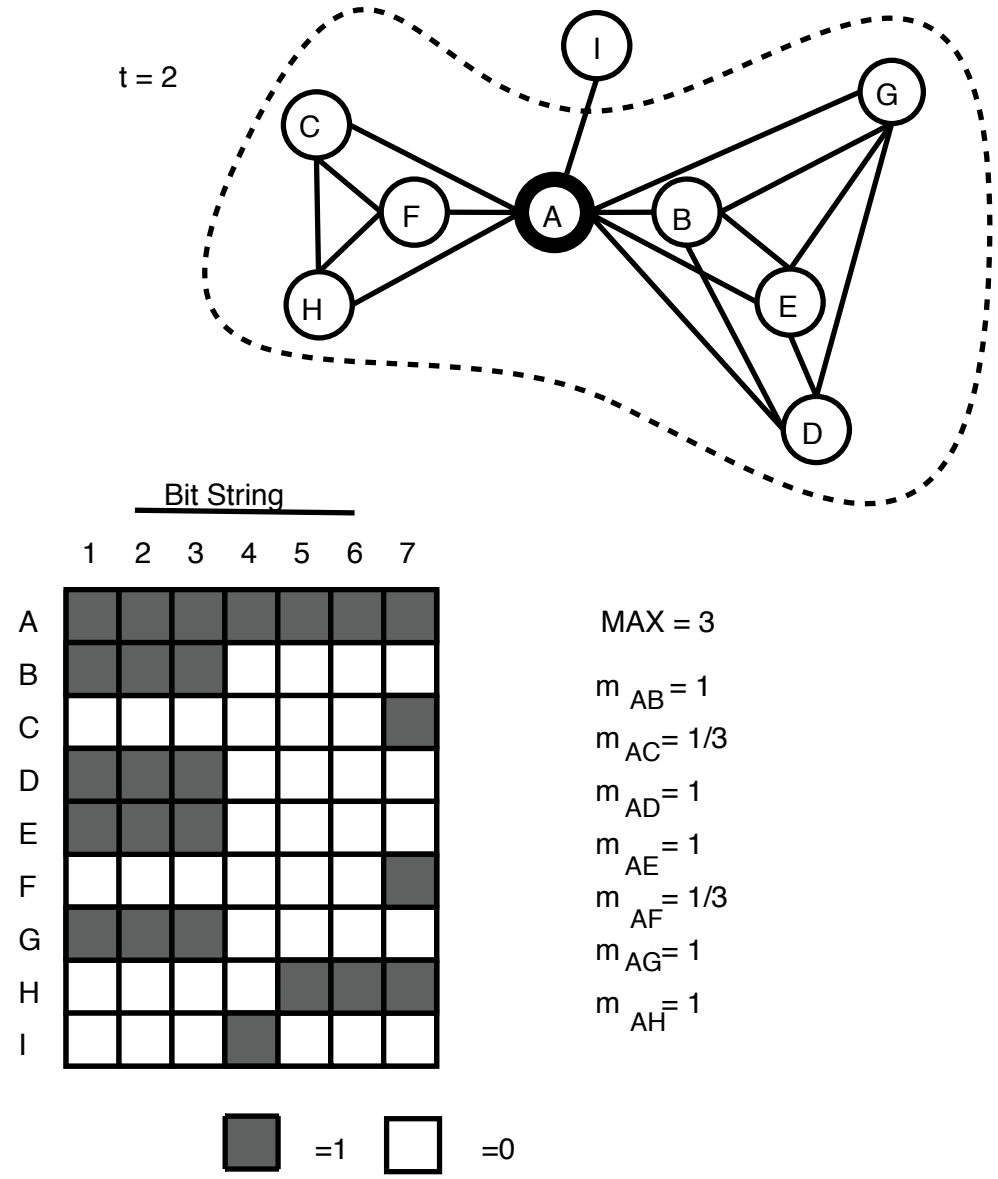


Figure 6.7: Finding groups in the data. Neighbour of A whose neighbourhood intersects with A $\geq t$ are found.The bit string AND operation is then used to weight the edges according to the number of samples the group are expressed similarly in the bi-partite graph.

increase (Fig. 6.8a), while the average weight of the modules increases, (Fig. 6.8b). This is due to the additive nature of the scoring scheme, since as the requirement for d_t and I increases, more edges are included in the module. The average size (i.e. number of gene nodes) in a module tends to increase with d_t, but not strictly so, indicating that as more edges are required the module does not necessarily acquire more nodes.

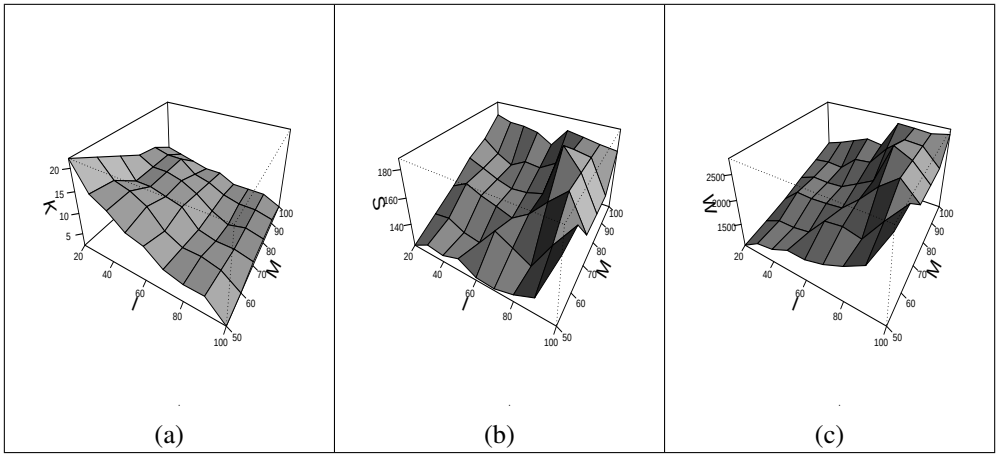


Figure 6.8: Effects on the Number of Clusters (K), Average Size (S) and Average Weight (W) as input parameters (I = Intersection Neighbourhood Size, M = d_t = minimum degree of nodes) change for the Hsiao dataset. Similar results were obtained for all datasets.

Dataset	K	Avg. Size	Min. Size	Max. Size	Avg. Score	Min. Score	Max Score
Aliz	21	275.29	106	566	1457.02	247.61	3348.63
Alon	7	159.57	102	264	425.00	217.32	666.50
Gasch	4	280.25	141	446	4081.05	1455.76	7065.32
Golub	6	106	84	139	303.04	230.97	437.57
Hsiao	4	177	118	232	2665.96	1674.98	3477.03
Steg.	3	130	99	178	1636.66	1161.44	2401.49
Spell	46	285.30	108	728	1801.884	329.63	4990.37
West	11	352.90	135	721	3623.64	637.48	8458.10

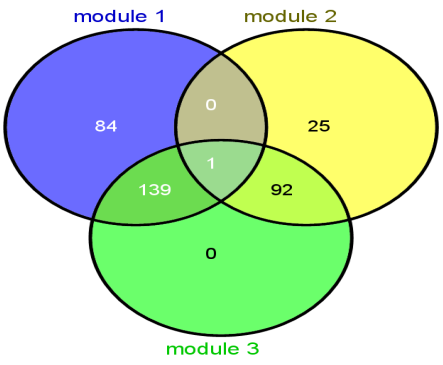
Table 6.2: Descriptive statistics of modules found. The minimum and maximum values are presented to illustrate the range of modules found.

Table 6.2 provides descriptive statistics for modules (i.e. sets of genes exhibiting coherent activity), found in all datasets[3]. The average weight of the modules varies *between* graphs, e.g. Alon and Hsiao modules have very different average weight although the size of the modules are approximately the same. Unsurprisingly, the density of the Hsiao one-mode graph is much greater than the density of the Alon one-mode graph. Indeed, in all cases, higher weight modules tend to be associated with one-mode graphs of higher density. Therefore, although *within* graphs higher weighted modules tend to be associated with larger size, this is not the case *between* graphs.

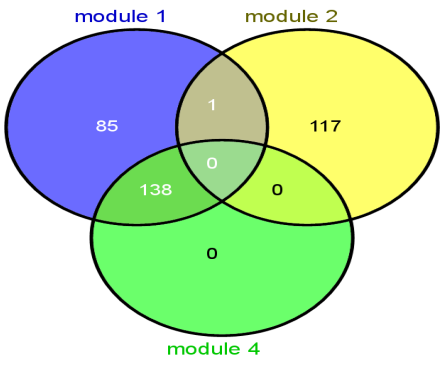
One prediction of *GraphCreate* is that groups of high degree genes will be repeated across modules for a particular dataset, while genes appearing solely in one module are less reactive (have smaller degree). This arises due to the power-law distribution of the gene nodes whereby a few genes are connected to many nodes and many nodes are connected to a few genes. Modules found in all datasets are indeed hierarchical in nature (i.e. there is gene overlap between modules). Gene node memberships of modules found in the Hsiao dataset, are given in Fig. 6.9, which illustrates these overlaps e.g. module 3 is completely formed from subgroups of nodes found in modules 1 and 2, Fig.6.9a. If we examine the degree of gene nodes within a module, x say, it is clear that those gene nodes which overlap with other modules have higher connectivity *within* module x (i.e. considering only these connections within x), than those gene nodes which appear solely in x and nowhere other than x, Fig. 6.10. This indicates that genes found in the overlaps of the modules are highly reactive, either affecting or affected by genes which appear in one

[3]Note that, due to small size and lack of information, the Cho dataset is not considered for further analysis. Modules for the Gasch dataset was extracted from a one-mode graph created whereby two gene nodes must have a common neighbourhood of at least 5 sample nodes in the underlying bipartite graph in order to qualify for an edge in the one-mode graph.

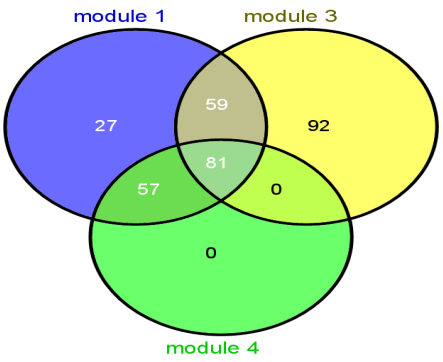
module only.



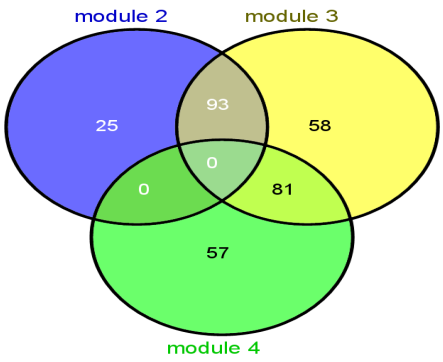
(a) Modules 1, 2 and 3



(b) Modules 1, 2 and 4



(c) Modules 1, 3 and 4



(d) Modules 2, 3 and 4

Figure 6.9: Overlap between modules in the Hsiao dataset.

Groups of genes forming modules in the Hsiao dataset were compared with those gene clusters found by the original authors, Hsiao et al. (2001), who categorised these according to involvement in *house-keeping functions* (HK), *tissue selectivity* (TS) or *tissue variance* (TV). It is clear, (where genes in modules found by *GraphCreate* are annotated by original authors), that overlaps occur for TV and TS categories. Modules 1 and 3 (Fig **??**) have a significant overlap of 127 TS genes

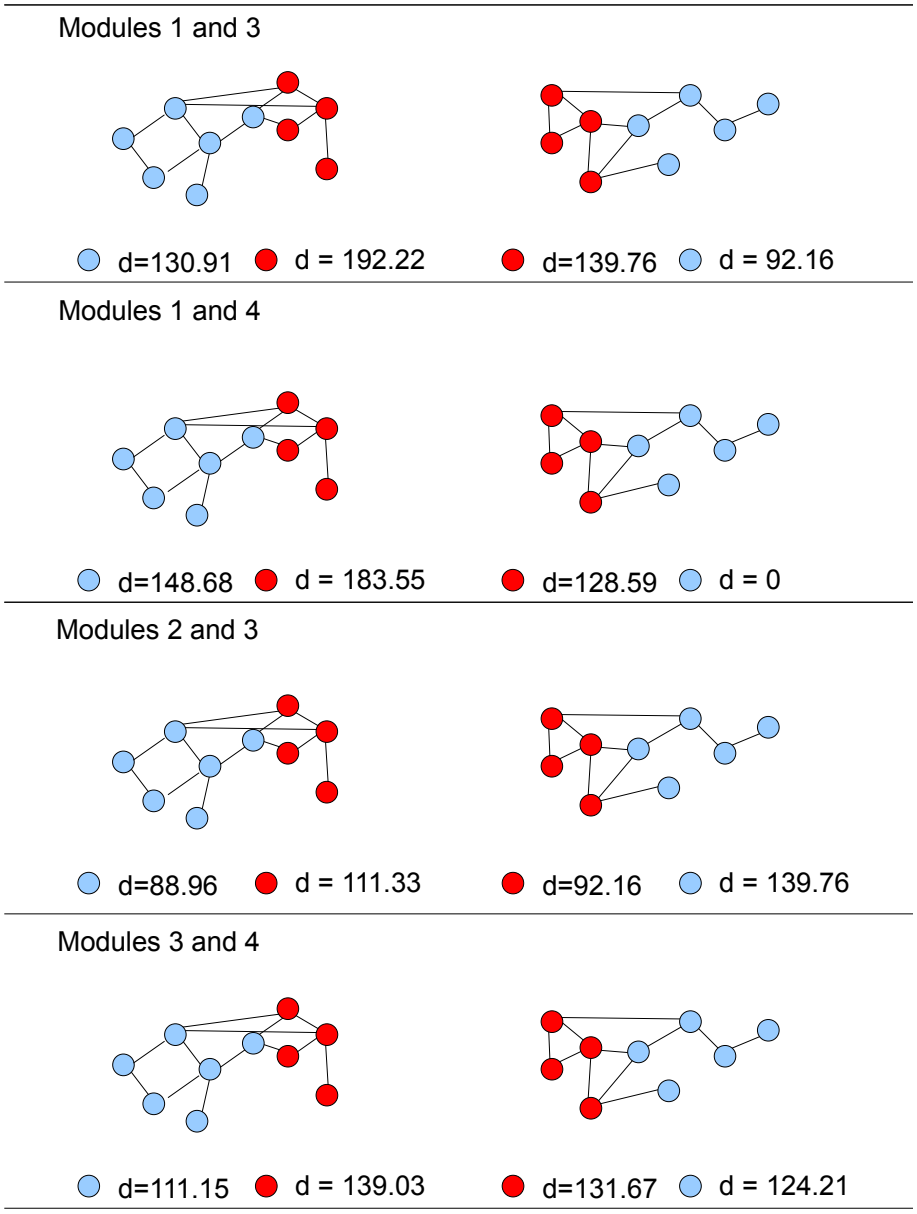


Figure 6.10: Schematic depicting the degree of modules found in Hsiao dataset. In each instance, "modules x and y", blue represents the average degree of the gene nodes only found in x, while red represents the average degree of the nodes which are co-operating in module y. For e.g. "Modules 3 and 4", the average degree of gene nodes found solely in 3 is 111.15, while those co-operating with module 4 also is 139.03. Similarly, the average degree of gene nodes found solely in module 4 is 124.21, while those co-operating with module 3 also is 131.67. In each instance, there is a higher average node degree with the co-operating nodes, (except modules 2 and 3, as module 3 also has large commonality with module 1)

(hypergeometric test[4] $p = 9.3e^{-33}$). Similarly modules 1 and 4 have a TS overlap of 121 ($p = 6.7e^{-33}$), modules 2 and 3 of 85 ($p = 6.5e^{-28}$), and finally modules 3 and 4 of 76 ($p = 1^{e}-22$). These TS genes forming commonalities across modules have the highest variance across samples in the original dataset, and therefore have a higher degree in the underlying bi-partite graph, (and consequently a higher degree in the one-mode graph).

We continue our investigation of modules found by *GraphCreate* by considering the *cancer* and *yeast* datasets separately. From our set of test datasets, four relate to cancer experiments and two relate to experiments on yeast, Table 6.3.

Cancer	Alon (Colon)	Golub (Leukemia)	Stegmaier (Leukemia)	West (Breast)
Yeast	Gasch (Stress Response)	Spellman (Cell Cycle)		

Table 6.3: Datasets used for analysis

6.4.1 Cancer Datasets

A hypergeometric test was used to find groups of over-represented genes in each module[5], (Falcon and Gentleman, 2007). For these tests, the entire set of genes in each dataset was used as the pool to draw from (i.e. *the gene universe*). Genes that mapped to more than one entrezID were removed to avoid Gene Ontology

[4] A hypergeometric test was used to find groups of over-represented genes in each module (Falcon and Gentleman, 2007). This describes the probability of number of successes from n trails, using sampling without replacement.

[5] In the Gene Ontology (GO) annotation hierarchy, each GO term inherits all annotations from its more specific descendants. An analysis for GO term associations can result in the identification of genes associated with directly related GO terms with considerable overlap. To avoid this problem, when analysing the GO ontology graph, the leaves of the GO graph were tested first (nodes with no children), before testing terms whose children have already been tested, and all genes annotated at significant children are removed from the parent's gene list. This continues until all terms were tested.

HK = House Keeping, TS = Tissue Selective, TV = Tissue Variant

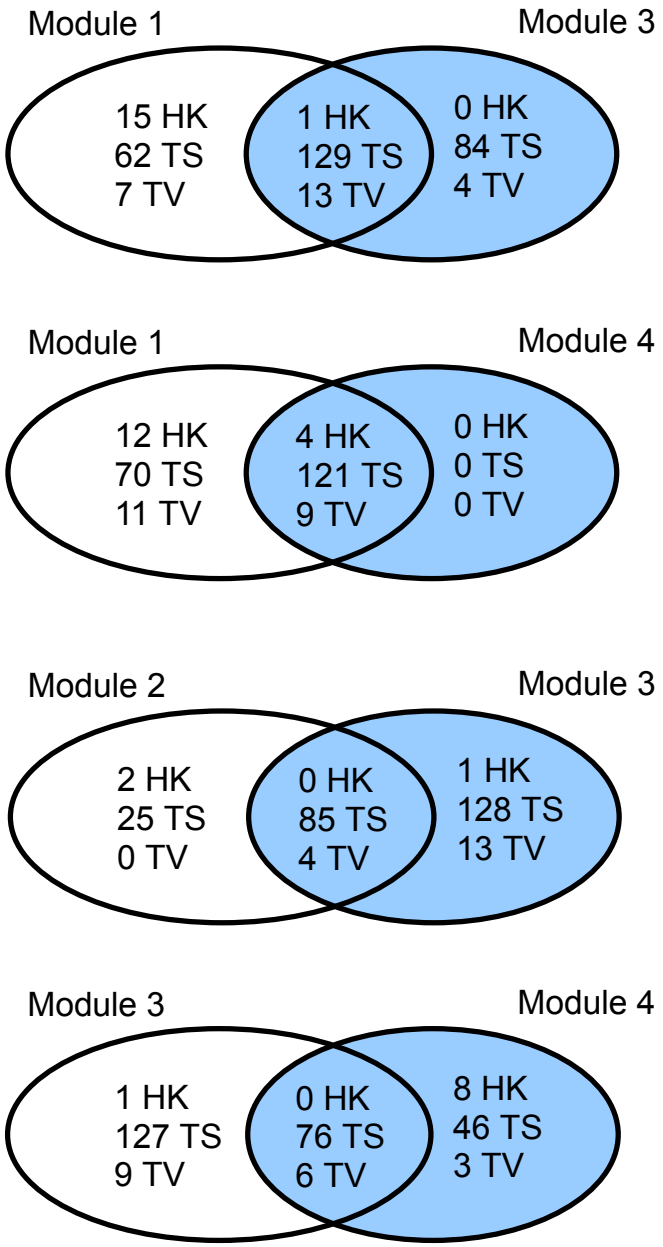


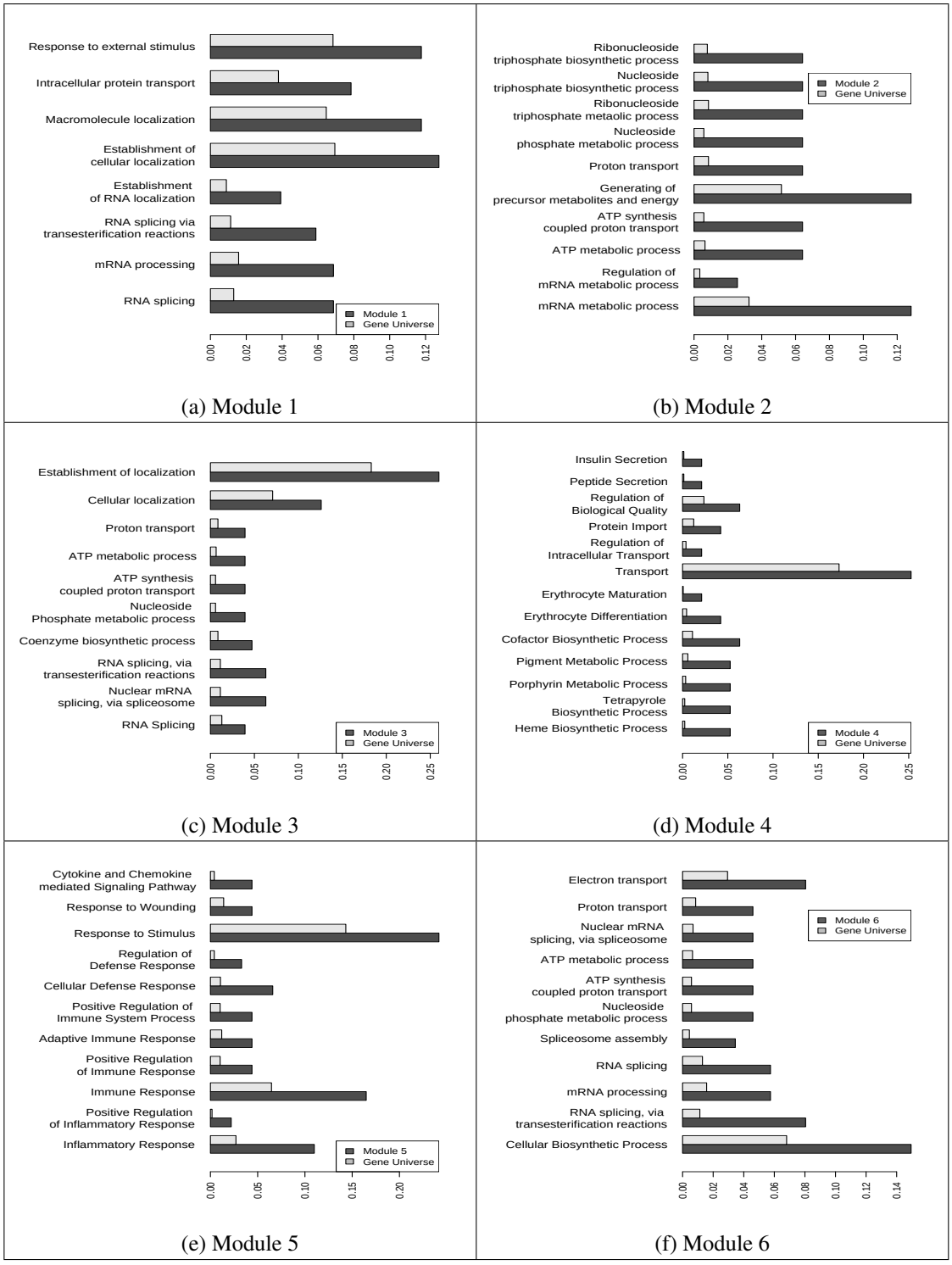
Figure 6.11: Hsiao module comparison with author's annotation.

(GO) categories being counted twice. Only annotated genes in each module were considered for this analysis. Fig. 6.12 shows the most significant GO ontology terms associated with groups of over-represented genes, (using hypergeomtric tests described above), for the Golub dataset. All ontology terms selected had an associated p-value < 0.001. In each of the modules (Fig. 6.12a - 6.12f), it may be seen that the percentage of genes within a module associated with a particular term is greater than the percentage of genes in the 'gene universe' associated with the same term. This illustrates graphically the idea that there is a higher representation of terms in the modules, or gene subsets, than would be expected by chance given the number of terms in the gene universe.

Although a few of the modules extracted from the Golub dataset have a diverse categorisation of genes, other are quite specific. For instance, Module 5 quite clearly contains genes associated with changing the state of a cell as a result of a stimulus. Module 4 on the other hand contains genes involved in transport of substances, biosynthetic processes (reactions resulting from the formation of substances), and erythrocyte (red blood cell) development. There is overlap among modules for genes involved in RNA processing events and pathways involving ATP (a universal coenzyme and enzyme regulator). Not all genes found in each module were associated with the most significant GO terms. Table 6.4 illustrates the percentage of annotated genes in each module which were associated with the most significant GO terms.

Module	1	2	3	4	5	6
%	55%	54%	54%	57%	65%	56%

Table 6.4: Percentage of genes annotated in associated with the most significant GO categories for each module in the Golub dataset



(a) Module 1

(b) Module 2

(c) Module 3

(d) Module 4

(e) Module 5

(f) Module 6

Figure 6.12: GO annotations of modules found for Golub Dataset, using Biological Process Ontology of GO database. GO categories with p-value < 0.001 where chosen.

Associated GO categories of over-represented genes for the remaining three cancer test datasets are shown in Table 6.5. Clearly, there are more genes with associated significant GO annotations in the Stegmaier dataset compared to both the West and Alon datasets. In the Stegmaier dataset, many of the genes are associated with cell-cycle ontology terms (this dataset consists of leukemia samples obtained at 6 hr and 24 hr time points). The edge set from the underlying bi-partite graph (used to generate the one-mode graph) can be used to determine to which samples the genes in a module, M, are coherently responding, i.e. $E' \subset E$, (E = edge set in bi-partite graph), where $(x, y) \in E'$ iff $x \in M$, then $y \in Y (=$ subset of interesting samples). For example, it was found that in Module 2 of the Steigmaier dataset all genes had an edge to sample 8, 9 and 10 *only* (Kasumi cell line, Genetifib treated at 24 hrs) in the underlying bi-partite graph, (Fig. 6.13), indicating a significant alteration in expression of all genes in module 2 at this time point.

Modules in the West dataset have quite a low $\%$ of annotated genes associated with the most significant GO terms (Module 11 has an insignificant amount, hence not shown). However, a noteworthy significant identification is the RAS protein signal transduction, identified in modules 6, 9, and 10. This forms part of the MAPK (mitogen-activated protein kinase) pathway, the activity of which was found to be high in breast cancers (Maemura et al., 1999), and which is correlated to the degree of RAS activation (von Lintig et al., 2004). RAS protein signal transduction is hyper-activated in breast cancer by overexpression of growth factor receptors which signal through it. RAS involvement in breast cancer has been well documented (von Lintig et al., 2004; McGlynn et al., 2009), and, although not mutated itself, is abnormally activated in breast cancers overexpressing the ErbB-2 receptor, (von Lintig et al., 2004). ErbB2 was also identified in modules 6, 9 and 10, as was GPR30, (G protein-coupled estrogen receptor 1). The fact that these

	1	2	3	4	5	6	7	8	9	10
West										
Cell Proliferation	19		49				28			52
Cell localization	13									
Intracellular Transport	11			20						
Biopolymer metabolic process		251	147		102		97			185
Nucleoside/tide/base		189				136	70	25	63	137
Transcription		127				91		25	65	90
RAS protein signal transduction						14			12	12
Regulation of biological processes						184				
Regulation of apoptsis								11		
%	28%	88%	49%	7%	40%	73%	86%	31%	23%	85%
Stegmaier										
Biopolymer metabolic process	81	57	51							
Cell Cycle	42	33	28							
Regulation of apoptsis	14		9							
Response to Stress	22									
Nucleoside/tide/base metabolic process		48	44							
%	93%	91%	90%							
Alon										
Celular Macromolecule metabolic process	30	31	42	39	56	46	34			
Regulation of biological quality			10	9			8			
Organelle Organization and biogenesis				13	22	16	12			
Prgrammed cell death					19					
Regulation of apoptsis					15					
Cell developement					22					
Negative regulation of programmed cell death						7				
%	68%	53%	68%	59%	57%	43%	54%			

Table 6.5: Number of of genes annotated in top GO categories for each module in the datasets. $\%$ = percentage of annotated genes in the module which are associated with the top GO terms.

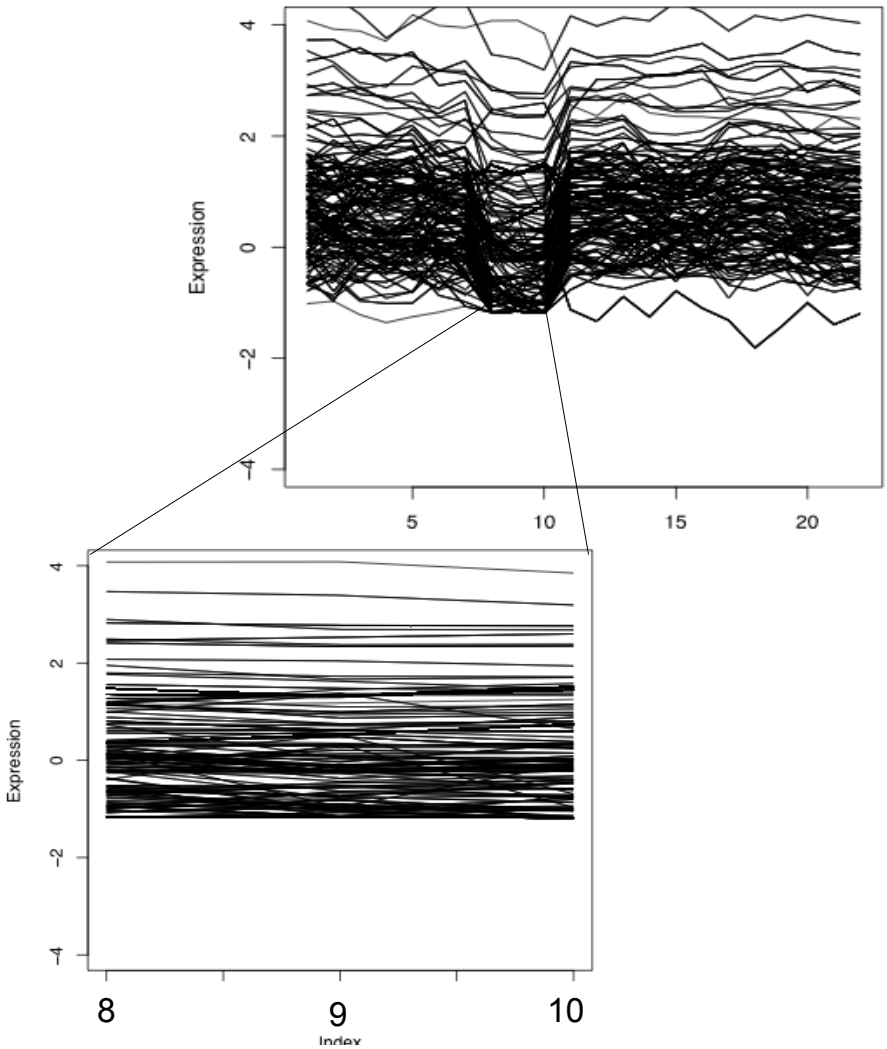


Figure 6.13: Genes in module 2 of Stegmaier dataset where induced or repressed in samples 8, 9 and 10, found through an examination of the edges in the bi-partite graph. These correspond to genetifib treated Kasumi cell line sample at 24hrs.

genes where identified indicates proliferation and migration of the breast cancer cells (Pandey et al., 2009). Other genes involved in MAPK pathway (i.e. MAPK, MAP2K, MAPKAPK2, MAPK14 (p38 isoform)) which play a central role in invasive breast cancer were also found in modules 6, 9 and 10 (Maemura et al., 1999; Han et al., 2002). These genes were not identified in modules that did not have genes associated with the RAS signal transduction.

The gene nodes involved in the RAS signal transduction pathway (RAS genes) also contribute heavily to the weight of the modules. For example, the 14 nodes associated with RAS genes in module 6 make up 2.8% of the total gene nodes of the module, however the weight of the RAS gene nodes (i.e. weight of all edges incident to these nodes) make up 40% of the total weight of the module. The average degree of RAS gene nodes is higher than the rest of the gene nodes in module 6 (1354.64:1333.614, RAS:non-RAS).

West et al. (2001), (the authors of the original analysis of this dataset), made available the top 100 genes found to be most discriminatory (DS genes) between ER+ and ER- [6] (based on a Bayesian regression model). Of these, five DS genes were found in modules 6 and 9, while seven DS genes where found in Module 10. From an analysis of the edges in the bi-partite graph, the RAS genes in Module 6 were responsive in either ER+LN- or ER-LN+ samples[7]. Fig. 6.14 shows the pattern of expression of RAS genes across the subset of samples (ER+LN- or ER-LN+) in Module 6. The genes in these modules did not discriminate between ER+ and ER- tumors, (because the technique was applied to a graph which contained edges for both induction and repression).

[6]ER+ Estrogen Receptor positive samples, which ER- represents Estrogen Receptor negative samples

[7]ER+LN- represents Estrogen Receptor positive and Lymph Node negative samples, which ER-LN+ represents Estrogen Receptor negative and Lymph Node positive samples

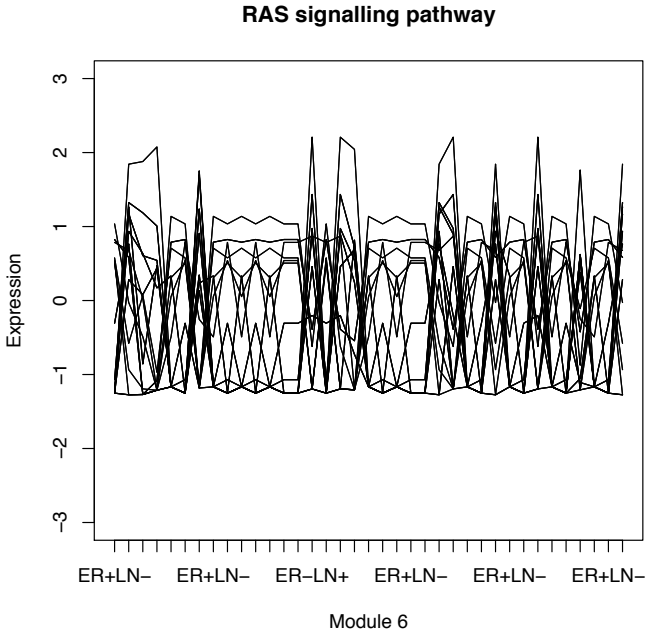


Figure 6.14: Pattern of coherent expression (i.e. induced or repressed coherently) of genes involved in RAS protein signalling pathway in module 6 found in the West dataset, across a subset of 37 samples in which change of expression was identified. Each of the 37 samples were either ER+LN- or ER-LN+.

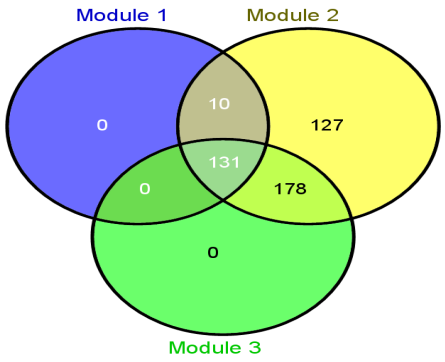
6.4.2 Yeast Datasets

There are a core set of genes in yeast that are transcribed under numerous stressful conditions, representing a general yeast response set (Mager and Kruijff, 1995; Ruis and Schuller, 1995). The Gasch dataset represents a compendium of gene response in yeast under a variety of stressful conditions[8]. The structure of the modules found in this dataset is presented in Fig. 6.15. Again, a hierarchical organisation is indicated, where Module 2 is a super-module, containing nodes from all other modules, i.e. the general stress response. The genes in Module 2 where found to be responsive under a wide variety of samples.

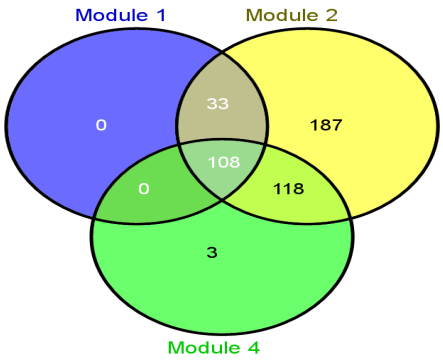
The most significant GO term associations for genes in modules found in the Gasch dataset are given in Table 6.6 (all terms identified with $p < 0.01$). The wide range of GO terms found illustrates the large scale effect of stress conditions on gene expression in yeast cells. (Note that although Module 2 is a superset of all other modules significant GO terms can alter between module 2 and other modules found, due to the nature of the hypergeometric test.) Genes in Module 1, with known function, are mainly involved in fatty acid metabolism, while genes in Module 4 are associated both with these and also with cell wall organisation and modifications. GO terms identified in Module 3 largely intersect those of Module 2, however a large percentage of genes are also significantly over-represented in cellular macromolecule and protein metabolic processes.

Genes which are responsive to stress can also be isolated to a particular stressful condition. For instance, 22 genes were found to be responsive in the heat-shock (25° to 30°) samples of the Gasch dataset (HS genes). All of these genes where members

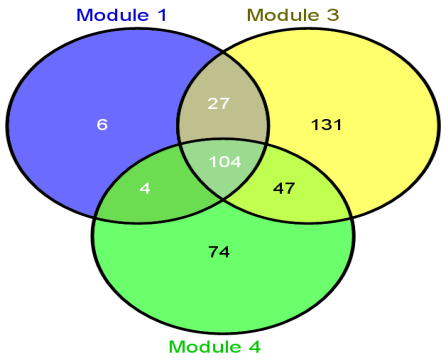
[8]Heat Shock, Hydrogen Peroxide treatment, Menadione exposure, Diamide treatment, DTT Exposure, Hyper-osmotic shock, Hypo-osmotic shock, Amino acid starvation, Nitrogen depletion, Stationary phase, Steady state growth on alternative carbon sources. Each represent time course experiments.



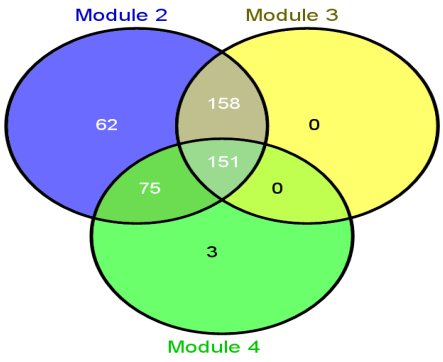
(a) Modules 1, 2 and 3



(b) Modules 1, 2 and 4



(c) Modules 1, 3 and 4



(d) Modules 2, 3 and 4

Figure 6.15: Overlap between modules in the Gasch dataset.

	1	2	3	4
Lipid Metabolic Proces	15			20
Fatty Acid Metabolic Process	6			7
Cell Wall Organisation and Biogenesis				17
Ergosterol Biosynthetic Process			20	6
Carboxylic acid metabolic process	20			
tRNA aminoacylation for protein translation	5	12	11	
Secretion		38	29	
Secretory pathway		13	11	
Intracellular transport		62	38	
Cellular localisation		67	53	
Amino acid activation		12	11	
Ergosterol biosynthetic process		12	10	
Vesicle-mediated transport		35		
Steroid biosynthetic process	5	11	10	6
Sterol metabolic process		12	10	
Post-Golgi vesicle-mediated transport		12		
ER to Golgi vesicle-mediated transport		12		
External encapsulating structure organization and biogenesis		26		
Protein amino acid glycosylation		11		
Glycoprotein metabolic process		11		
Macromolecule localization		37	30	
Protein import		13		
Nitrogen compound metabolic process		38		
Cellullar lipid metabolic process			24	
Vaculor Transport			10	
Macromolecule biosynthetic process			49	
Cellular macromolecule metabolic process			76	
Protein metaboic process			74	
%	25%	48%	64%	21%

Table 6.6: Significant GO associations in modules found in the Gasch dataset. These where identified via a hypergeometric test. All annotations have p-value < 0.01

of Module 3 (subset of Module 2) only. Notably, the application of sudden heat shock elicited large and rapid alterations in the expression of these genes before they returned to a steady state, Fig 6.16. From an examination of the underlying bi-partite graph, the HS genes where identified in samples 2, 3, 4 and 6 (10 mins, 15 mins, 20 mins and 40 mins respectively) of the heat shock samples. Thus, this rapid alteration in gene expression is captured in the bi-partite graph, translated to the one-mode graph and captured in modules by *GraphCreate*. Notably, the HS genes are also a subset of genes found to be activated in the stationary phase, Fig. 6.17. Interestingly, the HS genes were not identified as significantly responding under any other conditions. This relationship between HS and stationary phase genes was also identified by Gasch et al. (2000).

As a second example from the Gasch dataset, 130 genes which are responsive under nitrogen depletion were found in Module 4 (ND genes). From an examination of the underlying bi-partite graph, the ND genes were identified across all time points in the experiment (1hr, 2hr, 4hr, 8hr, 12hr, 1day, 2 day, 3 day, 5 day), Fig. 6.18. Notably, this condition elicited a large response in the later stages of nitrogen depletion, Fig. 6.18.

The most significant GO terms associated with the modules in the Spellman (Yeast Cell Cycle) dataset are given in Table 6.7, (all 41 modules not shown). Clearly, there are distinct modules associated with (i) rRNA processing and ribosome assembly, and (ii) protein catabolic processes, (a similar distinction between rRNA processing and protein catabolic processes was found in a study by Carlson et al. (2006) of yeast cell cycle data). These processes are fundamental processes in the general function of the cell. The pattern of GO terms associated with many of the modules is unique, indicating a unique relationship between genes in individual modules. For instance, Modules 2 and 34 both have strong associations with rRNA

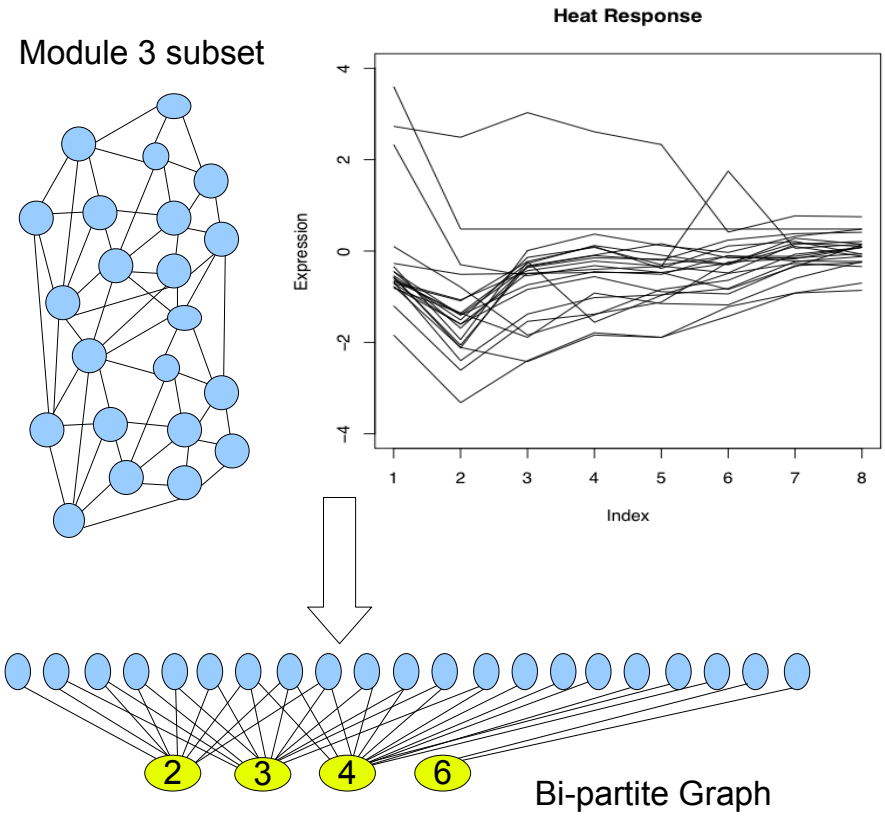


Figure 6.16: Responsive genes to heat shock (25° to 30°) in the Gasch dataset. These genes showed a rapid change in expression in the early stages, which was captured in the bi-partite graph, projected into one-mode graph and grouped into module 3 by *GraphCreate*.

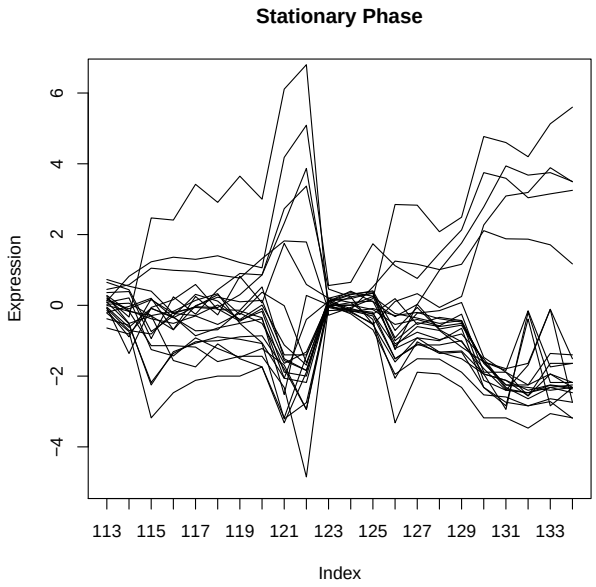


Figure 6.17: Responsive HS genes in stationary phase in the Gasch dataset. Genes which showed a rapid change in expression in heat shock where also identified as a subset of those changing in stationary phase.

processing and ribosome assembly. However, Module 34 also has a strong association with Organelle ATP synthesis coupled electron transport, Metabolic process and Alcohol metabolic process suggesting a possible relationship between genes involved in these and those involved in rRNA processing and ribosome assembly.

Patterns of gene expression are shown in Fig. 6.19, for selected modules. In these examples, Module 22 is the smallest, with genes coherently expressed in 22 samples (out of a possible 82), 12 of which are from the cdc15 strained based time course experiment[9]. (There are 37 cdc15 time-points in the Spellman dataset in total.) Module 12 is the largest with genes showing coherent expression in 47 samples (out of a possible 82). Again, 29 of the samples were cdc15 based, which repre-

[9]The Spellman yeast cell cycle experiment is a compendium of time course experiments based on four synchronisation methods: alpha factor (DBY8724), elutriation-chamber (DBY7286), cdc15-2 (DBY8728), Cln3 and Clb2 (DBY8725, DBY8726).

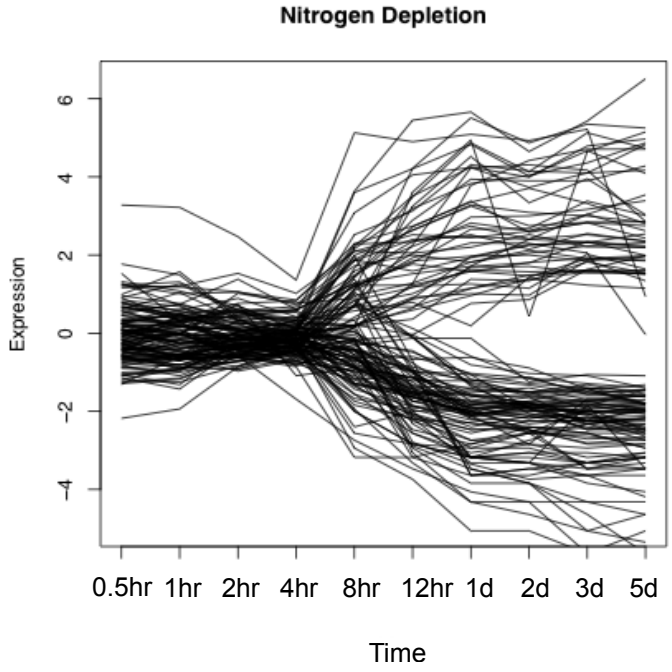


Figure 6.18: Responsive genes to nitrogen depletion found in module 4 in the Gasch dataset. These genes showed a rapid change in expression in the later stages of the time series.

Module →	2	10	11	12	14	16	22	23	31	32	34	36	37	39
rRNA processing	60		21		25		28	60	25		53	15	18	20
RNA metabolic process	104							108			86			
Maturation of SSU-rRNA from tricistronic rRNA transcript	13							13						
Cellular component organization and biogenesis	215				80			196			183		56	
Organelle organization and biogenesis												41		
Ribosome biogenesis and assembly	29				12			28			21	9		11
Ribosomal large subunit biogenesis and assembly	10							10			10			
Ribosomal large subunit assembly and maintenance	16				10		9	16			15			
Ribosome assembly	21						10	22	31		20			
Oxidative phosphorylation	10										10			
Ribonucleoprotein complex biogenesis and assembly			37									18		
Proteolysis involved in cellular protein catabolic process		20		18		22				21				
Protein catabolic process		22		21		24				23				
Ubiquitin-dependent protein catabolic process		19		17		21				20				
Protein folding														14
Modification-dependent macromolecule catabolic process		19				21								
Cellular protein metabolic process		74		59		80				97				
Steroid biosynthetic process			7											
Ergosterol biosynthetic process			6											
Glycolysis			7		7		8					6	6	
Modification-dependent macromolecule catabolic process				17										
Macromolecule catabolic process				28										
Hexose catabolic process							9					6	7	
Alcohol catabolic process							9	15					7	
Alcohol metabolic process											28			
Metabolic process								17			237			
Vacuolar transport										14				
Organelle ATP synthesis coupled electron transport											8			
Tricarboxylic acid cycle											6			
%	87	41	25	40	65	30	75	85	19	33	90	61	70	43

Table 6.7: Significant GO term associations for modules found in the Spellman dataset. All GO categories were chosen with a $p-$value < 0.001. Light grey columns represent rRNA processing and ribosome assembly modules, while dark grey modules represent protein catabolism modules.

sents the majority. However, the entire elutriation-chamber time course was represented in this module, (there are 10 elutriation-chamber experiment time-points in the Spellman dataset in total). Additionally, 24 of the 33 samples in Module 37 are cdc15 based, while Module 14 has representative samples from all time-course experiments.

6.4.3 Overall Summary of Modules Found

It is clear from the evaluation of modules found in the test datasets that significant groups of genes are found by the *GraphCreate* algorithm. To avoid the misleading results arising from overlaps of associations at different levels of the GO ontology tree, a 'conditional' hypergeometric test was used in all cases, whereby the leaves of the GO graph were tested first, before terms with children previously tested with all genes annotated at significant children removed from the parent's gene list, (Falcon and Gentleman, 2007). Regardless, strong associations were found in each dataset, which agreed with published findings, e.g. RAS signal transduction pathway in West dataset. When examining the Gasch dataset, it was insightful to observe how each specific stressful condition affected groups of genes found by *GraphCreate*, and how the patterns of expression for these genes where repeated across stressful conditions.

A drawback of projecting a gene expression bi-partite graph into one-mode is of course the loss of sample information. However, once a module of genes is discovered to be important, it is trivial to search the bi-partite edge-set to highlight those samples for which these gene nodes are coherently expressed. This analysis was carried out on one-mode graphs, projected from bi-partite graphs that contained both induced *and* repressed genes, with the intention of capturing groups of

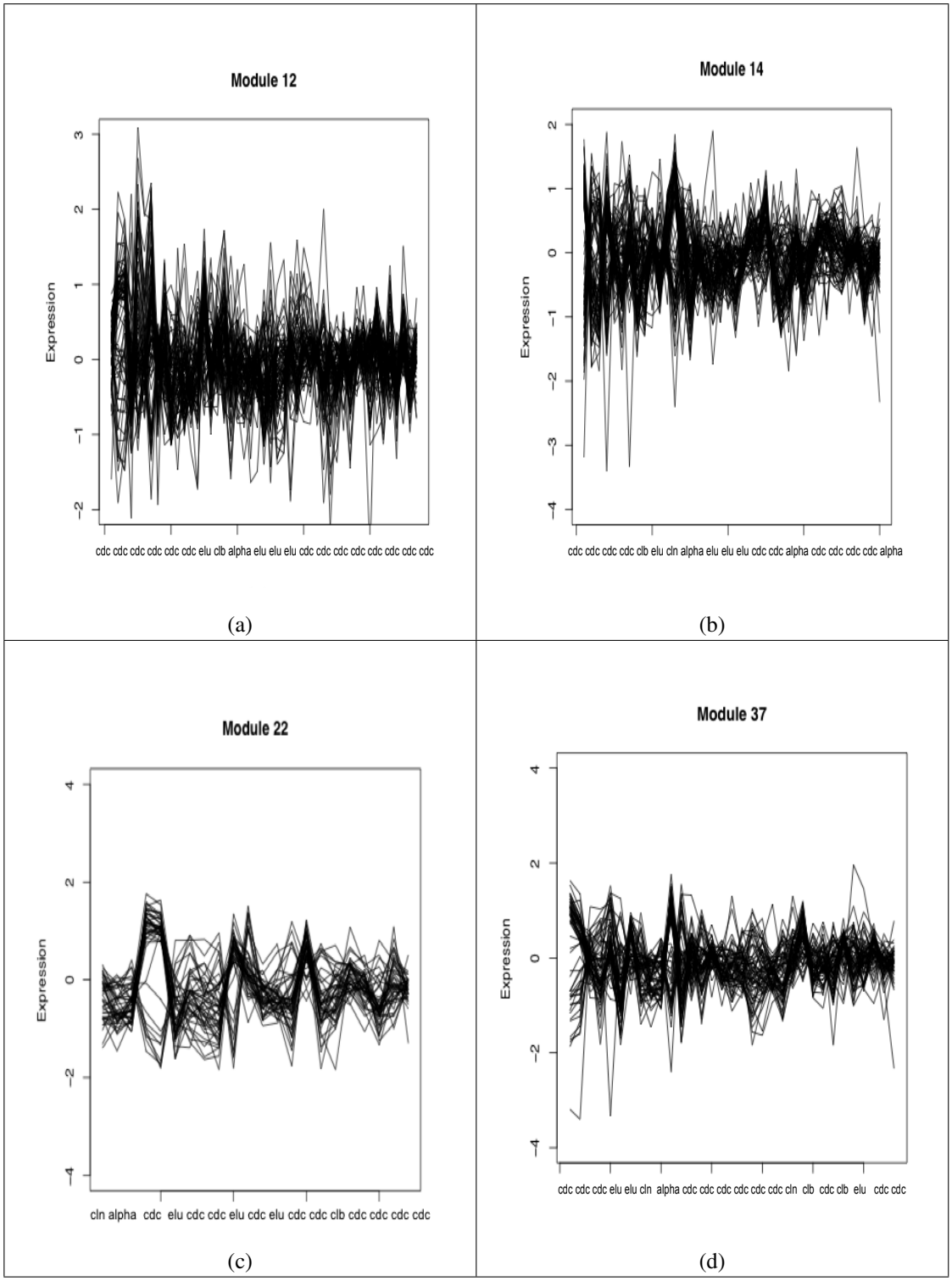


Figure 6.19: Expression pattern of Spellman dataset modules 12, 14, 22 and 37, captured by GraphCreate.

coherently expressed genes. This could easily be adapted to analyse induced *or* repressed genes, simply by projecting the appropriate induced or repressed bi-partite graph (see Chapter 5).

There is a hierarchical structure to the groups found (overlap of modules), which was also suggested in work by Ravasz et al. (2002) on metabolic networks. This structure arises from the power-law distribution of the gene nodes, and it was observed that gene nodes involved in overlaps of modules have a higher degree within the modules than those nodes not not included in overlaps. All this suggests that genes which 'connect' distinct modules either affect or are affected by a number of sub-groups of genes with membership in one module only.

6.5 Summary

In this chapter, we developed a general technique for weighted gene co-expression network construction that can be applied to gene expression datasets based on a bi-partite model, developed in Chapter 5. Microarray-based experiments of gene expression allow for a detailed survey of gene expression across a wide range of experimental samples. An advantage of using the bi-partite model as a basis include the fact that relationships in the gene co-expression network are based on a *subset* of these samples.

A thorough investigation of the gene co-expression network was carried out using important network concepts. This investigation uncovered many important organisational properties in these network not evident in the corresponding random networks of similar size and degree distribution. Chief amongst these was the increased probability that high density cliques, (measured through the clustering coefficients), in the gene co-expression networks together with the finding that the

node degree followed a power-law distribution.

An algorithm, *GraphCreate*, was developed for identification of modules of coherently expressed genes, based on considering nodes with high neighbourhood overlap, (our investigations also indicate that this is more likely in real gene co-expression networks). A *module* of gene nodes identifies those genes which form a common functional group. A scoring scheme for modules was developed, whereby the weight of a module is a direct consequence of the significance of expression of *all* genes *in that module*, (i.e. the weights of the edges are scaled based on other genes in the module). This innovative scoring scheme takes into account the samples under which the genes are co-regulated.

GraphCreate was used to find modules of gene nodes in gene co-expression graphs constructed. Our analysis demonstrates that significant functional groups of genes, which are co-active across a subset of experiments, are identified within the gene co-expression networks by *GraphCreate*. These results were corroborated from various published papers.

We carried out this analysis on a one-mode graph, obtained from a bi-partite all-in-one graph, containing information on both induced and repressed genes, to identify genes which were co-activated coherently. For if induced or repressed gene expression patterns are only required, this investigation can easily be adapted by projecting only the bi-partite subgraph which encodes the required category, (see Chapter 5). For instance, if only strongly induced genes where desired, the strongly-induced bi-partite graph would be projected. This is a powerful method of analysis which allows the researcher to analyse patterns of gene expression at different granularities if desired.

CHAPTER 7

OVERALL DISCUSSION AND FUTURE WORK

7.1 Goals of this Thesis

Understanding the interactions of genes expressed in a cell is critical to elucidating how biological organisms function. In this work efforts were concentrated on identifying these interactions through clustering (unsupervised) techniques to highlight meaningful patterns of gene co-regulation. The hypothesis is that genes which show co-regulated will be grouped together, indicating shared functionality, hence clusters defined at gene level represent biological modules.

The main goals of this thesis have been to: (i) investigate common unsupervised clustering methods and their applicability to gene expression data, (ii) extensively examine the properties of the gene expression data (iii) develop a robust solution to the computational analysis of gene expression data, and (iv) test the solution proposed for diverse datasets.

The view that clustering methods are universally applicable is a common misconception and has provoked controversy among practitioners, (Levsky and Singer, 2003; Shendure, 2008). While traditional global clustering techniques are popular,

biological theory supports the view that bi-clustering methods offer better interpretation in terms of data features and local structure. Limitations of commonly-used algorithms are well documented in the literature, while adoption of new (and hybrid) techniques has been slow among practitioners of microarray experiments and would be catalysed by transparent guidelines and increased availability in specialised software and public dataset repositories.

This work attempts to provide a new framework for analysis of gene expression analysis, which include:

- An numerical assessment of the performance of common unsupervised clustering methods and applicability of associated assessment measures.
- Adoption of graphical and statistical theory concepts to achieve an extensive examination of the properties of gene interactions over a diverse range of datasets, chosen reflect a range of experimental designs, platforms, sizes and objectives.
- Drawing on biological theory to develop a data dependent probabilistic model for weighting gene-sample interactions in bi-partite graph structures.
- Construction of a gene expression network from an underlying bi-partite graph, whereby nodes are connected in the gene expression network, if the corresponding genes are significantly co-expressed across a subset of samples in the bi-partite graph, reflecting local gene interactions.
- Using information of properties of gene interactions, development of *GraphCreate*, which identifies local structures of interaction among genes in a gene co-expression network. Development of a scoring scheme whereby the weight

of each module is a consequence of the significance of expression among all genes *in that* module.

7.2 Summary and Conclusions

We summarise a number of findings related to this work as follows:

1. The set of assessment measures for biological data is incomplete, with omissions for assessment metrics for overlapping local structures. Careful consideration of the compatibility of a particular assessment measure and clustering algorithm is required. Many assessment measures exhibit biases towards a particular algorithm or number of clusters. Internal measures by themselves may not be suitable for gene expression data, and validation through external measures, (although continued development of public annotation databases and metrics is required), is optimal.

2. Using classical network analysis tools, organisational structure was found in bi-partite graphs. To our knowledge, this is the first time basic network analysis techniques for bi-partite graphs have been applied to the analysis of gene expression. Clustering coefficients, cc, were obtained for gene and sample node sets individually with cc of gene nodes found, unsurprisingly, to be much larger than that for sample nodes, due to the large disparity in average degree between sets. Consideration of the minimum clustering coefficient measure removes the bias of nodes with larger neighbourhood, and reveals there more subtle, local interactions in the data. We examined the size of the neighbourhood intersections for genes, and found that, genes react more coherently in larger neighbourhoods than expected by chance. An examination

of the gene node degree distribution of the extracted graphs suggests that, for large gene sets, it is well-approximated a Normal distribution. However, the size of the sample node set for most datasets is relatively low, so that degree distribution is less easy to establish and appears to depend on the observed data itself.

3. The issue of data-dependent threshold estimation for the identification of edges is non-trivial, as numerous thresholds need to be assessed, which is computationally expensive.This work is important as it directly addresses the problem transformation of gene-sample couples into edges of a graph, and associated edge-weights, and has shown that it is critical to successful analysis of gene expression and extraction of meaningful patterns. This transformation process has received little attention in the literature.

4. Investigations into weighting schemes of gene expression networks is often overlooked. Comparison analysis of the empirical-based scheme with the well know Tanay et al. (2002) scheme, has shown that the edge weighting scheme itself deserves careful consideration. Presented an analysis under four major properties, *parameter influence, robustness, reuseability and discrimination* and found that our new empirical based scheme outperforms the Tanay scheme, capturing subtleties in the data, by selecting 'interesting' gene-sample couples in a data dependent manner and based on relative values. The new empirical based scheme presented is more specific in that fewer gene-sample couples are identified than in the Tanay scheme.

5. This investigation one-mode gene co-expression networks uncovered many important organisational properties in these network not evident in corresponding random networks of similar size and degree distribution. Amongst

these was the increased probability that high density cliques, (measured through the clustering coefficients), in the gene co-expression networks. Although gene-node degree distributions in the underlying bi-partite graph are well approximated by the Normal distribution, node degree in the gene co-expression networks more closely follow a power-law distribution, suggesting there a few 'hub' genes which connect to many other genes in the network.

6. Our analysis demonstrates that significant functional groups of genes, which are co-active across a subset of experiments, are identified within gene co-expression networks by *GraphCreate*, and these results were corroborated from various published papers.

Gene expression analysis represents only one parameter by which cells or tissues may be characterised. While clusters found at the transcript level represent potential shared function, the ability to combine RNA and protein expression data to comprehensively profile both transcriptional and post-transcriptional changes is particularly appealing, and will inevitably provide a more complete picture cell function. Although it is more difficult to identify proteins that are differentially expressed, advances in techniques for rapid and reproducible two-dimensional gel protein separation and mass spectrometry-based protein identification make high throughput proteomics feasible as an adjunct to microarray gene expression analysis, (Bowtell, 1999). Consequently, it is important that accurate algorithms and computing techniques are developed to measure and understand gene expression data *before* integration with other levels.

7.3 Future Work

Analysis in Chapter 6 was carried out on a one-mode graph, obtained from a bi-partite all-in-one graph, containing information on both induced and repressed genes. Future work will involve further investigation of the dynamics of gene interaction at various levels of granularity and the framework presented in Chapter 6 can easily be adapted by projecting only the bi-partite subgraph which encodes the required category, (for e.g., if only strongly induced genes where desired, the strongly-induced bi-partite graph would be projected). This method of analysis has enormous potential which allows the researcher to systematically analyse patterns of gene expression at different granularities if desired.

Throughout this thesis, care was taken to implement efficient and robust algorithms. This was achieved through the R and Bioconductor platform, calling external C functions for computationally demanding tasks. However, as with all open-source packages, R is continously improving. There are ongoing projects developing packages for parallel computing and message passing (e.g. Rmpi (Yu, 2009)). As biological data is increasingly complicated, important packages like these are vital to continue to successfully use this platform for Bioinformatics tasks. This work would benefit for these ongoing projects, enabling a larger scale analysis and yet even more thorough investigations.

Providing code developed in this work as an open-source package to the Bioconductor community would ensure that the algorithms and techniques would be thoroughly examined and tested. Although implementations of popular algorithms are sometimes available as standalone implementations, they often have clunky GUI interfaces and particular presentation of results which cannot be directly used as input to other statistical methods. For this reason, we have carried out this analy-

sis using, and developed algorithms which use, command line function calls in R. However, although a GUI has its disadvantages, (clunky and platform specific), it also has benefits for dissemination of the technique to practitioners of microarray experiments, and hence future work would involve a development of user interface for this package. Continued development and addition of algorithms to this package is part of future work.

Gene expression in the genome of a cell an extremely complex and does not act as a predetermined system, but as a system that responds to chemical changes in its environment. Therefore, although it is very important to study the transcriptome in isolation, it can only be understood by taking the complete state of the cell into consideration. The development of Bioinformatics/data-mining tools that span different levels of "omics", and which consider sequence similarity in promoter regions, is a crucial next step in the investigation of gene expression and its role in cell function.

BIBLIOGRAPHY

Al-Shahrour, F., Minguez, P., Vaquerizas, J. M., Conde, L., and Dopazo, J. (2005). Babelomics: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. *Nucleic acids research*, 33(Web Server issue):W460–4.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., et al. (2000). Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Nat. Acad. Sci. USA*, 96(12):6745–6750.

Alwine, J. C., Kemp, D. J., and Stark, G. R. (1977). Method for detection of specific rnas in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with dna probes. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5350–5354.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ring-

wald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat. Genet.*, 25(1):25–29.

Asyali, M. H., Colak, D., Demirkaya, O., and Inan, M. S. (2006). Gene expression profile classification: A review. *Curr. Bioinformatics*, 1(1):55–73.

Bagirov, A. M. and Mardaneh, K. (2006). Modified global k-means algorithm for clustering in gene expression data sets. In *WISB '06*, pages 23–28, Darlinghurst, Australia. Australian Computer Society, Inc.

Bar-Joseph, Z., Demaine, E. D., Gifford, D. K., Srebro, N., Hamel, A. M., and Jaakkola, T. S. (2003). K-ary clustering with optimal leaf ordering for gene expression data. *Bioinformatics*, 19(9):1070–1078.

Barabasi, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science (New York, N.Y.)*, 286(5439):509–512.

Barabasi, A. L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature reviews.Genetics*, 5(2):101–113.

Ben-Dor, A., Chor, B., Karp, R., and Yakhini, Z. (2003). Discovering local structure in gene expression data: the order-preserving submatrix problem. *J. Comput. Biol.*, 10(3-4):373–384.

Ben-Dor, A., Shamir, R., and Yakhini, Z. (1999). Clustering gene expression patterns. *J. Comput. Biol.*, 6(3-4):281–297.

Ben-Hur, A., Elisseeff, A., and Guyon, I. (2002). A stability based method for discovering structure in clustered data. *Pacific Symposium on Biocomputing.Pacific Symposium on Biocomputing*, pages 6–17.

Bergmann, S., Ihmels, J., and Barkai, N. (2004). Similarities and differences in genome-wide expression data of six organisms. *PLoS biology*, 2(1):E9.

Berry, J. and Goldberg, M. (1995). Path optimization and near-greedy analysis for graph partitioning: an empirical study. In *SODA '95: Proceedings of the sixth annual ACM-SIAM symposium on Discrete algorithms*, pages 223–232, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics.

Bezdek, J. C. (1973). Cluster validity with fuzzy sets. *Cybernetics and Systems*, 3(3):58–73.

Bezdek, J. C. (1974). Numerical taxonomy with fuzzy sets. *Journal of Mathematical Biology*, 1(1):57–14.

Birzele, F., Csaba, G., and Zimmer, R. (2007). Alternative splicing and protein structure evolution. *Nucleic acids research*, 36(2):550–558.

Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., Sampas, N., Dougherty, E., Wang, E., Marincola, F., Gooden, C., Lueders, J., Glatfelter, A., Pollock, P., Carpten, J., Gillanders, E., Leja, D., Dietrich, K., Beaudry, C., Berens, M., Alberts, D., and Sondak, V. (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406(6795):536–540.

Blatt, M., Wiseman, S., and Domany, E. (1996). Superparamagnetic clustering of data. *Physical Review Letters*, 76(18):3251–3254.

Blond, S., Guillaume, J., and Latapy, M. (2005). Clustering in p2p exchanges and consequences on performances.

Bolshakova, N., Azuaje, F., and Cunningham, P. (2006). Incorporating biological domain knowledge into cluster validity assessment. In *EvoWorkshops*, Lecture Notes in Computer Science, pages 13–22. Springer.

Bourqui, R., Cottret, L., Lacroix, V., Auber, D., Mary, P., Sagot, M. F., and Jourdan, F. (2007). Metabolic network visualization eliminating node redundance and preserving metabolic pathways. *BMC Systems Biology*, 1:29.

Bowtell, D. D. (1999). Options available–from start to finish–for obtaining expression data by microarray. *Nature genetics*, 21(1 Suppl):25–32.

Branden, C. I. and Tooze, J. (1999). *Introduction to protein structure*, volume 1. Taylor and Francis.

Brock, G., Pihur, V., Datta, S., , and Datta, S. (2008). *clValid: Validation of Clustering Results*. R package version 0.5-7.

Brown, T. A. (2002a). *Accessing the Genome*, chapter 8, pages 221–239. Genomes. John Wiley and Sons Inc., United States of America, 2 edition.

Brown, T. A. (2002b). *Assembly of the Transcription Initiation Complex*, chapter 9, pages 239–272. Genomes. John Wiley and Sons Inc., United States of America, 2 edition.

Brown, T. A. (2002c). *Genome Replication*, chapter 13, pages 384–415. Genomes. John Wiley and Sons Inc., United States of America, 2 edition.

Brown, T. A. (2002d). *The Humane Genome*, chapter 1, pages 3–23. Genomes. John Wiley and Sons Inc., United States of America, 2 edition.

Brown, T. A. (2002e). *Synthesis and Processing of RNA*, chapter 10, pages 273–311. Genomes. John Wiley and Sons Inc., United States of America, 2 edition.

Brown, T. A. (2002f). *Synthesis and Processing of the Proteome*, chapter 11, pages 313–346. Genomes. John Wiley and Sons Inc., United States of America, 2 edition.

Brown, T. A. (2002g). *Transcriptomes and Proteomes*, chapter 3, pages 70–91. Genomes. John Wiley and Sons Inc., United States of America, 2 edition.

Bryan, K., Cunningham, P., and Bolshakova, N. (2006). Application of simulated annealing to the biclustering of gene expression data. *IEEE T-ITB*, 10(3):519–525.

Bui, T. N. and Moon, B. R. (1996). Genetic algorithm and graph partitioning. *IEEE Trans. Comput.*, 45(7):841–855.

Caldarelli, G., Battiston, S., Garlaschelli, D., and Catanzaro, M. (2004). Emergence of complexity in financial networks.

Campello, R. J. G. B. (2007). A fuzzy extension of the rand index and other related indexes for clustering and classification assessment. *Pattern Recognition Letters,*, 28(7):833–841.

Cano, C., Adarve, L., Lopez, J., and Blanco, A. (2007). Possibilistic approach for biclustering microarray data. *Comp. Biol. Med.*, 37(10):1426–1436.

Carlson, M. R., Zhang, B., Fang, Z., Mischel, P. S., Horvath, S., and Nelson, S. F. (2006). Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC genomics*, 7:40.

Carter, S. L., Brechbuhler, C. M., Griffin, M., and Bond, A. T. (2004). Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics (Oxford, England)*, 20(14):2242–2250.

Chebyshev, P. L. (1867). Des valeurs moyennes,. *Journal de Mathmatique Pures et Appliques*, 12(2):177–8.

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

Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J., and Davis, R. W. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular cell*, 2(1):65–73.

Choi, J., Choi, J., Kim, D., Choi, D., Kim, B., Lee, K., Yeom, Y., Yoo, H., Yoo, O., and Kim, S. (2004). Integrative analysis of multiple gene expression profiles applied to liver cancer study. *FEBS Letters*, 565(1-3):93–100.

Churchill, G. A. (2002). Fundamentals of experimental design for cdna microarrays. *Nature genetics*, 32 Suppl:490–495.

Consortium, I. H. G. S. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945.

Datta, S. and Datta, S. (2006). Evaluation of clustering algorithms for gene expression data. *BMC bioinformatics*, 7 Suppl 4:S17.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence.*, 1(2):224–227.

Dembele, D. and Kastner, P. (2003). Fuzzy c-means method for clustering microarray data. *Bioinformatics*, 19(8):973–980.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *J Royal Statistical Soc B (Methodological)*, 39(1):1–38.

Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., and Lempicki, R. A. (2003). David: Database for annotation, visualization, and integrated discovery. *Genome biology*, 4(5):P3.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., , and Weingessel, A. (2008). *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*. R package version 1.5-18.

Ding, C., Xiaofeng, H., Hongyuan, Z., Ming, G., and Simon, H. D. (2001). A min-max cut algorithm for graph partitioning and data clustering. pages 107–114.

Dudek, M. W. A. (2008). *clusterSim: Searching for optimal clustering procedure for a data set*. R package version 0.36-3.

Dudoit, S., Fridlyand, J., and Speed, T. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97:77–87.

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *J Cybernetics and Systems*, 4(1):95–104.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci USA*, 95(25):14863–14868.

Erdös, P. and Rényi, A. (1959). On random graphs. I. *Publ. Math. Debrecen*, 6:290–297.

Erlich, H. A. (1989). *PCR technology principals and applications for DNA amplification*, volume 1. Stockton Press.

Falcon, S. and Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23(2):257–258.

Forti, A. and Foresti, G. L. (2006). Growing hierarchical tree som: an unsupervised neural network with dynamic topology. *Neural networks*, 19(10):1568–1580.

Fridlyand, J. and Dudoit, S. (2001). Applications of resampling methods to estimate the number of clusters and to improve the accuracy of a clustering method. Technical Report 600, Department of Statistics, University of California, Berkeley.

Fu, L. and Medico, E. (2007). Flame, a novel fuzzy clustering method for the analysis of dna microarray data. *BMC Bioinformatics*, 8:3.

Gamberoni, G., Storari, S., and Volinia, S. (2006). Finding biological process modifications in cancer tissues by mining gene expression correlations. *BMC Bioinformatics*, 7(6):8.

Gasch, A. P. and Eisen, M. B. (2002). Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome biology*, 3(11):RESEARCH0059.

Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., and Brown, P. O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Molecular biology of the cell*, 11(12):4241–4257.

Gentleman, R., Carey, V., Huber, W., Irizarry, R., and Dudoit, S. (2005a). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor (Statistics for Biology and Health)*, volume 1, chapter 1, pages 3–12. Springer-Verlag New York, Inc, Secaucus, NJ, USA, 1 edition.

Gentleman, R., Carey, V., Huber, W., Irizarry, R., and Dudoit, S. (2005b). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor (Statistics for Biology and Health)*, volume 1, chapter 11, pages 183–4. Springer-Verlag New York, Inc, Secaucus, NJ, USA.

Gentleman, R. C., Carey, V. J., Bates, D. M., et al. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80.

Getz, G., Levine, E., and Domany, E. (2000). Coupled two-way clustering analysis of gene microarray data. *Proceedings of the National Academy of Sciences of the United States of America*, 97(22):12079–12084.

Giancarlo, R., Scaturro, D., and Utro, F. (2008). Computational cluster validation for microarray data analysis: experimental assessment of clest, consensus clustering, figure of merit, gap statistics and model explorer. *BMC bioinformatics*, 9(1):462.

Gibbons, F. D. and Roth, F. P. (2002). Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res.*, 12(10):1574–1581.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537.

Gowda, M., Jantasuriyarat, C., Dean, R. A., and Wang, G. L. (2004). Robust-longsage (rl-sage): a substantially improved longsage method for gene discovery and transcriptome analysis. *Plant Physiology*, 134(3):890–897.

Guillaume, J. and Latapy, M. (2004). Bipartite graphs as models of complex networks. In *Aspects of Networking*, pages 127–139. Springer.

Halkidi, M., Vazirgiannis, M., and Batistakis, Y. (2000). Quality scheme assessment in the clustering process.

Han, Q., Leng, J., Bian, D., Mahanivong, C., Carpenter, K., Pan, Z., Han, J., and Huang, S. (2002). Rac1-MKK3-p38-MAPKAPK2 Pathway Promotes Urokinase Plasminogen Activator mRNA Stability in Invasive Breast Cancer Cells. *Journal of Biological Chemistry*, 277(50):48379–48385.

Handl, J., Knowles, J., and Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15):3201–3212.

Hartigan, J. A. (1972). Direct clustering of a data matrix. *JASA*, 67(337):123–129.

Hartl, D. L. and Jones, E. W. (2002). *DNA structure, replication and manipulation*, volume 1 of *Essential Genetics: A Genomics Perspective*, chapter 6, pages 202–38. Jones and Bartlett Publishers, 3 edition.

Hartuv, E., Schmitt, A. O., Lange, J., Meier-Ewert, S., Lehrach, H., and Shamir, R. (2000). An algorithm for clustering cdna fingerprints. *Genomics*, 66(3):249–256.

Herrero, J., Valencia, A., and Dopazo, J. (2001). A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, 17(2):126–136.

Hsiao, L., Dangond, F., Yoshida, T., Hong, R., Jensen, R. V., Misra, J., Dillion, W., Lee, K. F., Clark, K. E., Haverty, P., Weng, Z., Mutter, G. L., Frosch, M. P., Donald, M. E. M., Milford, E. L., Cru, C. P., Bueno, R., Pratt, R. E., Mahadevappa, M., Warrington, J. A., Stephanopoulos, G., and Gullans, S. R. (2001). A compendium of gene expression in normal human tissues. *Physiological Genomics*, 7(2):97–104.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1):193–218.

Hubert, L. and Schultz, J. (1976). Quadratic assignment as a general data-analysis strategy. *British Journal of Mathematical and Statistical Psychology*, 29:190–241.

i Cancho, R. F. and Sole, R. V. (2001). The small world of human language. *Proceedings: Biological Sciences*, 268(1482):2261–2265.

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bulletin de la Socit vaudoise des sciences naturelles.*, 44:223–47.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: A review. *ACM CSUR*, 31(3):264–323.

Johnson, D. S., Aragon, C. R., McGeoch, L. A., and Schevon, C. (1989). Optimization by simulated annealing: An experimental evaluation; part i, graph partitioning. *Operations Research*, 37(6):865–892.

Kaiser, S., Santamaria, R., Theron, R., Quintales, L., and Leisch., F. (2007). *biclust: BiCluster Algorithms*. R package version 0.5.

Kasturi, J. and Acharya, R. (2005). Clustering of diverse genomic data using information fusion. *Bioinformatics*, 21(4).

Kernighan, B. W. and Lin, S. (1970). An efficient heuristic procedure for partitioning graphs. *The Bell system technical journal*, 49(1):291–307.

Kerr, M. K. and Churchill, G. A. (2001). Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 98(16):8961–8965.

Kim, D. W., Lee, K. H., and Lee, D. (2005). Detecting clusters of different geometrical shapes in microarray gene expression data. *Bioinformatics*, 21(9):1927–1934.

Klinkenberg, L. G., Mennella, T. A., Luetkenhaus, K., and Zitomer, R. S. (2005). Combinatorial repression of the hypoxic genes of saccharomyces cerevisiae by dna binding proteins rox1 and mot3. *Eukaryotic cell*, 4(4):649–660.

Kluger, Y., Basri, R., Chang, J. T., and Gerstein, M. (2003). Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Res.*, 13(4):703–716.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480.

Krishnapuram, R. and Keller, J. M. (1993). A possibilistic approach to clustering. *IEEE TFS*, 1(2):98–110.

Kuhn, K., Baker, S. C., Chudin, E., Lieu, M. H., Oeser, S., Bennett, H., Rigault, P., Barker, D., McDaniel, T. K., and Chee, M. S. (2004). A novel, high-performance

random array platform for quantitative gene expression profiling. *Genome research*, 14(11):2347–2356.

Kustra, R. and Zagdanski, A. (2006). Incorporating gene ontology in clustering gene expression data. *CBMS*, 0:555–563.

Lander, E. S. et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Latapy, M., Magnien, C., and Vecchio, N. D. (2006). Basic notions for the analysis of large affiliation networks / bipartite graphs.

Lazzeroni, L. and Owen, A. B. (2000). Plaid models for gene expression data. *Statistica Sinica*, 12(1):61–86.

Lee, C. and Wang, Q. (2005). Bioinformatics analysis of alternative splicing. *Briefings in bioinformatics*, 6(1):23–33.

Levine, E. and Domany, E. (2001). Resampling method for unsupervised estimation of cluster validity. *Neural computation*, 13(11):2573–2593.

Levsky, J. M. and Singer, R. H. (2003). Gene expression and the myth of the average cell. *Trends in cell biology*, 13(1):4–6.

Li, C. and Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome biology*, 2(8):RESEARCH0032.

Li, H., Zhang, K., and Jiang, T. (2004). Minimum entropy clustering and applications to gene expression analysis. *Proc. IEEE CSB.*, pages 142–151.

Liu, J., Wang, W., and Yang, J. (2004). Gene ontology friendly biclustering of expression profiles. In *CSB '04*, pages 436–447, Washington, DC, USA. IEEE Computer Society.

Liu, X., Cheng, G., and Wu, J. X. (2002). Analyzing outliers cautiously. *IEEE Trans. Knowl. Data Eng.*, 14(2):432–437.

Lorkowski, S. and Cullen, P. M. (2006). *Basic Concepts of Gene Expression*, volume 1 of *Analysing Gene Expression, A Handbook of Methods: Possibilities and Pitfalls*, pages 1–78. Wiley-VCH.

Lu, Y., Lu, S., Fotouhi, F., Deng, Y., and Brown, S. J. (2004). Incremental genetic k-means algorithm and its application in gene expression data analysis. *BMC bioinformatics*, 5:172.

Lukashin, A. V. and Fuchs, R. (2001). Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters. *Bioinformatics*, 17(5):405–414.

Luo, F., Khan, L., Bastani, F., Yen, I. L., and Zhou, J. (2004). A dynamically growing self-organizing tree (dgsot) for hierarchical clustering gene expression profiles. *Bioinformatics*, 20(16):2605–2617.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In Cam, L. M. L. and Neyman, J., editors, *Proc. Fifth Berkeley Symp. Math. Stat. Prob.*, pages 281–297. University of California Press.

Madeira, S. C. and Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 1(1):24–45.

Maechler, M., Rousseeuw, P., Struyf, A., and Hubert, M. (2005). *cluster: Cluster Analysis Basics and Extensions*. R package version 1.11.11. Rousseeuw et al provided the S original which has been ported to R by Kurt Hornik and has since been enhanced by Martin Maechler.

Maemura, M., Iino, Y., Koibuchi, Y., Yokoe, T., and Morishita, Y. (1999). Mitogen-activated protein kinase cascade in breast cancer. *Oncology*, 57(Suppl. 2):37–44.

Mager, W. and Kruijff, A. J. D. (1995). Stress-induced transcriptional activation. *Microbiol. Rev.*, 59(3):506–531.

Maslov, S., Sneppen, K., and Zaliznyak, A. (2004). Pattern detection in complex networks: Correlation profile of the internet. *PHYSICA A*, 333:529.

Matsumura, H., Reich, S., Ito, A., Saitoh, H., Kamoun, S., Winter, P., Kahl, G., Reuter, M., Kruger, D. H., and Terauchi, R. (2003). Gene expression analysis of plant host-pathogen interactions by supersage. *Proceedings of the National Academy of Sciences of the United States of America*, 100(26):15718–15723.

McGlynn, L. M., Kirkegaard, T., Edwards, J., Tovey, S., Cameron, D., Twelves, C., Bartlett, J. M., and Cooke, T. G. (2009). Ras/raf-1/mapk pathway mediates response to tamoxifen but not chemotherapy in breast cancer patients. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 15(4):1487–1495.

Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a dataset. *Psychometrika*, 50:159–179.

Modrek, B. and Lee, C. (2002). A genomic view of alternative splicing. *Nature genetics*, 30(1):13–19.

Moller-Levet, C., Cho, K. H., Yin, H., and Wolkenhauer, O. (2003). Clustering of gene expression time-series data. Technical.

Newman, M. E. J. (2001a). Scientific collaboration networks. i. network construction and fundamental results. *Physical Review E*, 64(1):016131.

Newman, M. E. J. (2001b). Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical Review E*, 64(1):016132.

Newman, M. E. J., Watts, D. J., and Strogatz, S. H. (2002). Random graph models of social networks. *Proceedings of the National Academy of Sciences*, 99(90001):2566–2572.

Nieweglowski., L. (2008). *clv: Cluster Validation Techniques*. R package version 0.2.

Oliver, J. L. and Marn, A. (1996). A relationship between gc content and coding-sequence. *Journal of Molecular Evolution*, 43:216–223.

Pandey, D., Lappano, R., Albanito, L., Madeo, A., Maggiolini, M., and Picard, D. (2009). Estrogenic gpr30 signalling induces proliferation and migration of breast cancer cells through ctgf. *EMBO J*, 28(5):523–532.

Pennisi, E. (2000). Human genome project. and the gene number is...? *Science (New York, N.Y.)*, 288(5469):1146–1147.

Pereira-Leal, J. B., Enright, A. J., and Ouzounis, C. A. (2004). Detection of functional modules from protein interaction networks. *Proteins*, 54(1):49–57.

Petsko, G. A. and Ringe, D. (2004a). *Control of protein function*, volume 1 of *Protein Structure and Function*, chapter 3, pages 86–126. New Science Press, 1 edition.

Petsko, G. A. and Ringe, D. (2004b). *From sequence to structure*, volume 1 of *Protein Structure and Function*, chapter 1, pages 2–44. New Science Press, 1 edition.

Petti, A. A. and Church, G. M. (2005). A network of transcriptionally coordinated functional modules in saccharomyces cerevisiae. *Genome research*, 15(9):1298–1306.

Pham, T. D., Wells, C., and Crane, D. I. (2006). Analysis of microarray gene expression data. *Curr. Bioinformatics*, 1(1):37–53.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabasi, A. L. (2002). Hierarchical Organization of Modularity in Metabolic Networks. *Science*, 297(5586):1551–1555.

Reddy, T. E., DeLisi, C., and Shakhnovich, B. E. (2007). Binding site graphs: a new graph theoretical framework for prediction of transcription factor binding sites. *PLoS computational biology*, 3(5):e90.

Rijsbergen, C. V. (1975). *Information Retrieval, 1st edition*. Dept. of Computer Science, University of Glasgow.

Robins, G. and Alexander, M. (2004). Small worlds among interlocking directors: Network structure and distance in bipartite graphs. *Computational and Mathematical Organization Theory*, 10(1):69–94.

Romesburg, C. (2004). *Cluster Analysis for Researchers*. Lulu Press, Morrisville.

Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational Applied Mathematics*, 20(1):53–65.

Ruis, H. and Schuller, C. (1995). Stress signaling in yeast. *BioEssays*, 17(11):959–965.

Saha, S., Sparks, A. B., Rago, C., Akmaev, V., Wang, C. J., Vogelstein, B., Kinzler, K. W., and Velculescu, V. E. (2002). Using the transcriptome to annotate the genome. *Nature biotechnology*, 20(5):508–512.

Saramaki, J., Kivela, M., Onnela, J. P., Kaski, K., and Kertesz, J. (2007). Generalizations of the clustering coefficient to weighted complex networks. *Physical review.E, Statistical, nonlinear, and soft matter physics*, 75(2 Pt 2):027–105.

Sharan, R., Maron-Katz, A., and Shamir, R. (2003). Click and expander: a system for clustering and visualizing gene expression data. *Bioinformatics (Oxford, England)*, 19(14):1787–1799.

Sharan, R. and Shamir, R. (2000). Click: a clustering algorithm with applications to gene expression analysis. *ISMB '00*, 8:307–316.

Shendure, J. (2008). The beginning of the end for microarrays? *Nature methods*, 5(7):585–587.

Sokal, R. R. and Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon*, 11(2):33–40.

Southern, E. M. (1975). Detection of specific sequences among dna fragments separated by gel electrophoresis. *Journal of Molecular Biology*, 98(3):503–517.

Speed, T. (2000). *Statistical Analysis of Gene Expression Data*, volume 1.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell*, 9(12):3273–3297.

Stegmaier, K., Corsello, S. M., Ross, K. N., Wong, J. S., Deangelo, D. J., and Golub, T. R. (2005). Gefitinib induces myeloid differentiation of acute myeloid leukemia. *Blood*, 106(8):2841–2848.

Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003). A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science*, 302(5643):249–255.

Sturn, A. (2001). Cluster analysis for large scale gene expression studies.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA*, 96(6):2907–2912.

Tanay, A. (2005). *Computational analysis of transcriptional programs: function and evolution. Ph.D. Thesis.* PhD thesis.

Tanay, A., Sharan, R., and Shamir, R. (2002). Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18 Suppl 1:S136–44.

Tanay, A., Steinfeld, I., Kupiec, M., and Shamir, R. (2005). Integrative analysis of genome-wide experiments in the context of a large high-throughput data compendium. *Mol. Sys. Biol.*, 1:2005.0002.

Thalamuthu, A., Mukhopadhyay, I., Zheng, X., and Tseng, G. C. (2006). Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, 22(19):2405–2412.

Toronen, P. (2004). Selection of informative clusters from hierarchical cluster tree with gene classes. *BMC bioinformatics*, 5:32.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525.

van Bakel, H. and Holstege, F. C. (2004). In control: systematic assessment of microarray performance. *EMBO reports*, 5(10):964–969.

Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995). Serial analysis of gene expression. *Science (New York, N.Y.)*, 270(5235):484–487.

von Lintig, F. C., Dreilinger, A. D., Varki, N. M., Wallace, A. M., Casteel, D. E., and Boss, G. R. (2004). Ras activation in human breast cancer. *Breast Cancer Research and Treatment*, 62(1):51–62.

Walker, M. R. and Rapley, R. (1997). *Formation of the DNA double helix*, volume 1 of *Route Maps in Gene Technology*, chapter 5, pages 18–2. Blackwell Publishing, Oxford.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J., Marks, J. R., and Nevins, J. (2001). Predicting the clinical status

of human breast cancer by using gene expression profiles. *PNAS*, 98(20):11462–11467.

Xie, X. L. and Beni, G. (1991). A validity measure for fuzzy clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(8):841–847.

Yang, J., Wang, H., Wang, W., and Yu, P. (2003). Enhanced biclustering on expression data. *BIBE '03*, page 321.

Yang, Y. H. and Speed, T. (2002). Design issues for cdna microarray experiments. *Nature reviews.Genetics*, 3(8):579–588.

Yeung, K. Y., Haynor, D. R., and Ruzzo, W. L. (2001). Validating clustering for gene expression data. *Bioinformatics*, 17(4):309–318.

Yeung, K. Y. and Ruzzo, W. L. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics (Oxford, England)*, 17(9):763–774.

Yip, A. M. and Horvath, S. (2007). Gene network interconnectedness and the generalized topological overlap measure. *BMC bioinformatics*, 8:22.

Yu, H. (2009). *Rmpi: Interface (Wrapper) to MPI (Message-Passing Interface)*. Rmpi provides an interface (wrapper) to MPI APIs. It also provides interactive R slave environment.

Zakharkin, S. O., Kim, K., Mehta, T., Chen, L., Barnes, S., Scheirer, K. E., Parrish, R. S., Allison, D. B., and Page, G. P. (2005). Sources of variation in affymetrix microarray experiments. *BMC bioinformatics*, 6:214.

Zhang, B. and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4:Article17.

Glossary

Chapter 2

Array

See microarray

Base-pairing

The process by which two nucleotides become connected by hydrogen bonds. The only permissible pairs are the nitrogenous base Adenine(A) pairing with the nitrogenous base Thymine(T), similar Guanine(G) pairing with Cytosine(C).

Coding-RNA (non-coding RNA)

Coding RNA refers to RNA which will be translated into a protein. Non-coding RNA refers to RNA which carries out various essential functions in the cell.

Complementary Strand

In double stranded DNA or RNA structures, each strand is complementary to the other in that base pairing occurs between the nitrogenous bases on each strand. Since since each base can only pair with one other possible type of base, the complementary strand can be reconstructed from any single strand.

DNA (Deoxyribonucleic Acid)

The collective term for polymers nucleotides. Each nucleotide consists of a deoxyribose sugar, a phosphate group and one of four nitrogenous bases; Adenine,

Guanine, Cytosine and Thymine .

End Modifications

This refers to 5' capping, 3' polyadenylation which are modifications made to the mRNA strand (referred to as pre-mRNA before end-modifications) before it is translated.

Exon

DNA sequences with a gene which code for a protein.

Expression

This refers to the process of converting the information encoded in a gene sequence to information encoded in an RNA sequence.

Gene

Sequence of DNA in the genome of a cell, which is transcribed into RNA.

Intron

DNA sequences within a gene which do not code for a protein.

Microarray

A technology which consists of glass or silicon chip with thousands of DNA oligonucleotides or cDNA sequences immobilised at distinct known positions.

Polymerase (RNA Polymerase)

An enzyme involved in DNA replication. RNA Polymerase is an enzyme involved in the transcription of RNA from DNA.

Probes

DNA RNA segments, immobilized at distinct positions, which 'probe' a sample if interest, i.e. bind to a complementary strand in a sample of interest. See also: Target

RNA (Ribonucleic Acid)

The collective term for polymers nucleotides. Each nucleotide consists of a ribose sugar, a phosphate group and one of four nitrogenous bases; Adenine, Guanine, Cytosine and Uracil .

Splicing

Alternative removal of mRNA coded for by introns, resulting in alternative protein products.

Target

DNA RNA extracted from a sample of interest which is to be analysed.

Transcription

The process of realisation of the RNA coded for in the genome.

Translation

The process of realisation of the proteins coded for in the mRNA strands.

Chapter 3

Bi-clusters

A cluster defined over a subset of attributes and subset of samples.

Cluster

A group of object which are more similar in some properties compared to groups in other clusters.

Complete clustering

A cluster structure in which every variable and attribute are in some cluster.

Crisp Membership

An object is assigned membership of a cluster with a certainty of 1. In the case of overlapping clusters, a cluster can be a member of ≥ 1 cluster with a membership of 1.

Exclusive (hard) clusters

Each object is a member of only one cluster.

Feature extraction

Identifying latent features in the dataset which can differentiate groups.

Feature selection

Selecting features of the variables most distinguishable in grouping objects.

Fuzzy Membership

An object is assigned to a cluster with a membership value which indicates the associativity with that cluster.

Global clusters

Structure in the dataset defined over all attributes and all samples.

Local structure

Structure in the dataset defined over a subset of attributes (samples) and variables (genes).

Overlapping clusters

Two or more clusters which contain common objects.

Partial clustering

A clustering structure found in the dataset where every variable and attribute are in a cluster.

Chapter 4

Assessment

Formative in nature which is ongoing and used to improve process. It is also diagnostic, i.e. identify areas for improvement.

Cluster Validation

The process of determining the correctness of grouping genes in a cluster, and of the cluster structure overall.

Cophenetic Distance

This refers to how similar two objects have to be in order to be grouped together in the same cluster. The cophenetic correlation measure, measures the correlation between the cophenetic distance matrix and the distance matrix.

Evaluation

Summative in nature which is final and used to gauge quality. Judgemental i.e. arrive at an overall grade/score.

Validation

The process of evaluating techniques and determining suitability based on statistical evidence and/or meeting user needs and requirements.

Chapter 5

All In One Graph

A bipartite graph, $G = (\top, \bot, E)$ which contains independent subgraphs from all subcategories, i.e. weak repressed $G_{wkR} = (\top_{wkR}, \bot_{wkR}, E_{wkR})$, moderate repressed $G_{modR} = (\top_{modR}, \bot_{modR}, E_{modR})$, strong repressed $G_{strR} = (\top_{strR}, \bot_{strR}, E_{strR})$, weak induced $G_{wkI} = (\top_{wkI}, \bot_{wkI}, E_{wkI})$, moderate induced $G_{modI} = (\top_{modI}, \bot_{modI}, E_{modI})$, strong induced $G_{strI} = (\top_{strI}, \bot_{strI}, E_{strI})$, such that $\{\top\} = \{\top_{wkR} \cup \top_{modR} \cup \top_{strR} \cup \top_{wkI} \cup \top_{modI} \cup \top_{strI}\}$, $\{\bot\} = \{\bot_{wkR} \cup \bot_{modR} \cup \bot_{strR} \cup \bot_{wkI} \cup \bot_{modI} \cup \bot_{strI}\}$, $\{E\} = \{E_{wkR} \cup E_{modR} \cup E_{strR} \cup E_{wkI} \cup E_{modI} \cup E_{strI}\}$.

Bipartite Graph

A graph, $G = (\top, \bot, E)$, is considered to be bipartite if there are two disjoint subsets of vertices, $\top$, $\bot$, and there is no edge between two vertices in the same subset. A gene expression dataset can be modelled in this way, where $\top$ vertices represent genes and $\bot$ vertices represent samples. An edge $w_{ij} \in W$ is the weight matrix, were $w_{ij} \neq 0$ if there is an edge between $i \in \top$ and $j \in \bot$.

Clustering Coefficient

Clustering Coefficient refers to a measure of the degree of sharing of neighbourhoods among nodes.

Degree

Degree of a node refers to how many edges are incident to the node.

Density

Density of a graph refers to the number of edges in the graph proportional to the total possible number of edges.

Edge

An edge connects two nodes/vertices in a graph. Each pair node can be connected by an edge, referred to as a complete graph, or a subset of nodes can be connected referred to as a partial graph.

Edge-Weight

Value assigned to an edge in a graph which can have various interpretations depending on the problem, e.g. distance, cost, length, capacity. In the problems represented in this thesis the weights represent significance of expression.

Node

Graphs comprise of a set of fundamental units, referred to as nodes. These nodes are the connecting points for edges in the graph. Nodes of the graph have various interpretations depending on the problem. In this thesis nodes represent genes and/or samples in the gene expression dataset under consideration.

Node Neighbourhood

The neighbourhood of a node, i, refers to the set of nodes an i has a connection to.

One-mode Graph

A graph, $G = (V, E)$, is considered to be in one mode if an edge can exist between any two vertices, $v \in V$. A gene expression dataset can be modelled in this way where vertices in the graph represent genes and an edge exists between two gene vertices if they show common expression (e.g. measured as a distance function).

One-Mode Projection

A bipartite graph can be projected to a one-mode graph. Two nodes in a bi-partite graph are linked in a one mode projection if they have a sample neighbour in common, A bipartite graph is either projected with the $\top$ set of nodes, or the $\perp$ set of nodes.

Random Graph

A random graph is a graph in which properties such as the number of vertices, edges and/or connections between are determined randomly.

Vertex

Also known as /textitnode, see above

APPENDIX A

DISTANCE AND ASSESSMENT METRICS

Distance Formulas

Minkowski Distance

This is a measure of distance in euclidean space. The Minkowski distance of order p between two points of n dimension, $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$, is:

$$\text{p norm dis} = \left(\sum_{i=1}^{n} |x_i - y_i| \right)^{\frac{1}{p}} \tag{A.1}$$

Of special interest arise when $p = 1$ (*Manhattan distance*), $p = 2$, (*Euclidean distance*), and $p = \infty$, (*Chebychev distance*).

Correlation distance

Correlation measures the linear relationship between two variables. *Pearson's Correlation* is defined as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1)s_x s_y} \tag{A.2}$$

where:

- x and y are two data points of n-dimension.
- s_x and s_y are the standard deviations of data points x and y respectively.
- $\overline{x}$ and $\overline{y}$ are the mean values of x and y respectively.

Spearman's correlation is a non-parametric correlation measure, where x_i and y_i are converted to rankings. It is defined as:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d^2 i}{n(n^2-1)} \tag{A.3}$$

where:

- $d_i = x_i - y_i =$ the difference in ranks.

An alternative non-parametric measure is *Kendal's tau*:

$$\tau = \frac{n_c - n_d}{n(n-1)/2} \tag{A.4}$$

where:

- n_c is the number of concordant pairs, i.e. pairs ordered the same way
- n_d is the number of discordant pairs i.e. pairs ordered differently

Assessment measures

Average Distance (AD)

$$AD = \frac{1}{NP}\sum_{i=1}^{N}\sum_{\ell=1}^{P}\frac{1}{n(K_{i,0})n(K_{i\ell})}\left(\sum_{i\in K_{i,0}, j\in K_{i,\ell}} dist(i,j)\right) \tag{A.5}$$

where:

- $K_{i,0}$ represents the cluster containing observation i using the original clustering (based on all available data),
- $K_{i,\ell}$ represents the cluster containing observation i where the clustering is based on the dataset with column ℓ removed.

Average Distance between Means (ADM)

$$ADM = \frac{1}{NP}\sum_{i=1}^{N}\sum_{\ell=1}^{P}\left(dist(x_{K_{i,\ell}}, x_{K_{i,0}}\right) \tag{A.6}$$

where:

- $x_{K_{i,0}}$ is the mean of the observations in the cluster which contain observation i, when clustering is based on the full data,
- $x_{K_{i,\ell}}$ is the mean of the observations in the cluster containing observation i where the clustering is based on the dataset with column ℓ removed

Average Proportion of non-overlap (APN)

$$APN = \frac{1}{NP}\sum_{i=1}^{N}\sum_{\ell=1}^{P}\left(1 - \frac{n(K_{i,\ell} \in K_{i,0})}{n(K_{i,0})})\right) \tag{A.7}$$

where:

- $K_{i,0}$ represents the cluster containing observation i using the original clustering (based on all available data),
- $K_{i,\ell}$ represents the cluster containing observation i where the clustering is based on the dataset with column removed.

C-Index, (Hubert and Schultz, 1976)

$$CI = \frac{S - S_{min}}{S_{max} - S_{min}}$$

where S = the sum of the distances over all pairs of patterns from the same cluster. Let that number of patterns in a cluster $= l$, then S_{min} is the sum of the l smallest distances if all pairs of distances are considered, similarly S_{max} is the sum of the l largest distances. Hence, a small CI indicates a good clustering.

Connectivity, (Handl et al., 2005)

$$Conn = \sum_{i=1}^{N} \sum_{j=1}^{L} x_{i,nn_{i(j)}} \tag{A.8}$$

where:

- $nn_{i(j)}$ is the j^{th} neighbour of observation i.
- $x_{i,nn_{i(j)}}$ is 0 is both i and $nn_{i(j)}$ are in the same cluster, otherwise $x_{i,nn_{i(j)}}$ is $1/j$.
- L is an input parameter which determines the number of neighbours that contribute to the measure.

Cophentic correlation

$$COPH = \frac{\sum_{i<j}(x_{ij} - x)(z_{ij} - z)}{\sqrt{\sum_{i<j}(x_{ij} - x)^2 \sum_{i<j}(z_{ij} - z)^2}} \tag{A.9}$$

where:

- Z is a symmetric matrix of size $N \times N$, where, for a hierarchical clustering, each entry $(z_{ij})$of the matrix indicates the level at which genes i and j where put into a cluster together.
- X is the original distance matrix

Davies-Bouldin Index, (Davies and Bouldin, 1979)

$$DB = \frac{1}{K}\sum_{i=1}^{K} \frac{max_{i \neq j}(diam(K_i) + diam(K_j))}{dist(K_i, K_j)} \tag{A.10}$$

where:

- $K =$ the number of clusters.
- $diam(K_x)$, is the diameter of cluster x.
- $dist(K_i, K_j)$ is the distance between clusters K_i and K_j.

Dunn Index, (Dunn, 1974)

$$Dunn = min_{1 \leq i \leq K}\left\{min_{1 \leq j \leq K, j \neq i}\left(\frac{dist(X_i, X_j)}{max_{1 \leq c \leq K} diam(X_c)}\right)\right\} \tag{A.11}$$

where

$$diam(X_i) = max_{x,y \in X_i}\{d(x,y)\} \qquad \text{and}$$
$$dist(X_i, X_j) = min_{x \in X_i, y \in Y_i}\{d(x,y)\} \tag{A.12}$$

Diam and dist can be severely affected by noisy values.

Figure of Merit

This is a measure of the predictive power of the algorithm. This measure assesses predictive power of each sample.

$$FOM(e,k) = \sqrt{\frac{1}{n} \times \sum_{i=1}^{K} \sum_{x \in K_i} (R(x,e) - \mu_{K_i}(e))^2} \tag{A.13}$$

where:

- e = the sample that is being assessed for predictiveness
- K = the number of clusters
- n = the number of genes
- K_i = the i^{th} cluster
- R(x,e) = the expression level of gene x in sample i.
- $\mu_{K_i}(e)$ = the average expression level in sample e of genes in cluster K_i.

The *aggregate figure of merit* assesses the total predictive power of the algorithm over all the samples for K clusters.

$$FOM(k) = \sum_{e=1}^{p} FOM(e,k) \tag{A.14}$$

This figure is biased towards larger number of clusters, due to the fact that (a) smaller clusters will tend to be more homogenous and (b) increasing the number of clusters will decrease the FOM(e,k) equation. The FOM equation can be *adjusted* to account for (b), see Yeung et al. (2001) for details.

$$FOM(e,k)_{adjusted} = FOM(e,k)/\sqrt{\frac{n-K}{n}} \tag{A.15}$$

Rand Index

The rand index is the proportion of concordant gene pairs in two partitions of a gene expression matrix, (two genes are concordant if they appear in the same cluster in both partitions or different clusters in both partitions) Rand (1971).

$$Rand_1 = \frac{a+d}{a+b+c+d} \tag{A.16}$$

where:

- a = the number of pairs of genes which are in the same cluster in both partitions
- b = the number of pairs of genes which are in different clusters in both partitions
- c = the number of pairs of genes which are in the same cluster in partition 1 and different clusters in partition 2

- d = the number of pairs of genes which are in different clusters in partition 1 and different clusters in partition 2

This can be standardised to an expected value of zero if the partitions are randomly generated and takes a maximum value of 1 if the partitions are perfectly correlated ($Rand_{adjusted}$), Hubert and Arabie (1985).

SD-Validity Index

This index is composed of the average *Scattering* of the clustering (measures the variance/compactness of the clusters)and the total *Separation* of the clustering (distance between cluster centres).

Average scattering of the clustering is given by:

$$Scatt = \frac{1}{n_c} \sum_{i=1}^{n_c} \left\| \frac{\sigma(v_i)}{\sigma(v_x)} \right\| \tag{A.17}$$

and Separation is given by:

$$Dis = \frac{max_{i,j=1...n_c}(\|v_j - v_i\|)}{min_{i,j=1...n_c}(\|v_j - v_i\|)} \sum_{k=1}^{n_c} \left(\sum_{j=1,k\neq j}^{n_c} \|v_j - v_k\| \right)^{-1} \tag{A.18}$$

and the SD-Validity index is defined to be:

$$SD = \alpha \cdot Scatt + Dis \tag{A.19}$$

where:

- α is a weighting factor which is equal to the total separation of the maximum number of input clusters.

Silhouette Width

For each observation i, the *Silhouette Value* is:

$$S(i) = \frac{b_i - a_i}{max(b_i, a_i)} \tag{A.20}$$

where:

- a_i is the average distance between i and all other observations in the same cluster.
- b_i is the average distance between i and all observations in the nearest neighbouring cluster.

The Silhouette Width of a cluster is the average of the silhouette values of each observation.

$$Sil = \frac{1}{N} \sum_{i=1}^{N} S(i) \tag{A.21}$$

Xie-Beni Index, (Xie and Beni, 1991)

$$XB = \frac{\sum_{i \in N} \sum_{j \in K} u_{ij}^2 \cdot d^2(X_i, K_j)}{n \cdot min_{i,j} d^2(K_i, K_j)} \tag{A.22}$$

where:

- u_{ij}^2 is the membership of gene i to cluster j.
- $d^2(X_i, K_j)$ is the distance between i and cluster j centroid.

APPENDIX B

DATASETS

Alizadeth - Lymphoma data

The dataset downloaded from `http://llmpp.nih.gov/lymphoma/data.shtml`, and contains data pertaining to the seminal paper of Alizadeh et al. (2000).

The dataset consists of 4026 genes and 96 samples. It represents data from a two-colour spotted array. The available values are ratio values that were log-transformed (base 2). The available data were centered by subtracting (in log space) the median observed value for each gene, (Alizadeh et al., 2000)

Alon - Colon Cancer

The dataset was downloaded from `http://microarray.princeton.edu/oncology/affydata/index.html`, which contains data pertaining to a colon cancer study by Alon et al. (1999).

The dataset contains measurements for 2000 genes across 62 samples (40 tumor samples, 22 normal tissue samples). An oligonucleotide array was used and the available data represented raw intensity values (i.e. unprocessed). The data was log transformed and scaled to have mean 0 and standard deviation of one for each

sample.

Cho - Yeast Cell Cycle

Data was downloaded from: `http://genomics.stanford.edu/yeast_cell_cycle/full_data.html`. The purpose of this experiment data was the characterization of mRNA levels during the cell cycle of the yeast *Saccharomyces cerevisiae*.

The dataset contains 6601 genes and across 17 time points. The downloaded dataset was normalized between timepoints with respect to each other, and represents information from 4 chips. The provided values for each of the 17 time points data for each gene are the normalized fluorescence between 0 and 160 minutes after cell cycle initiation from time 0. Data was normalized similar to the technique used in: `http://www.nature.com/ng/journal/v22/n3/full/ng0799_281.html`.

Gasch Dataset - Yeast Stress

Dataset was retrieved from `http://genome-www.stanford.edu/yeast_stress/data.shtml` and contains measurements of mRNA from an experiment monitoring yeast expression under various stressful conditions.

The data contains $6,152$ genes under 173 samples. The available data represents normalized, background-corrected log2 values of the Red/Green ratios measured on spotted DNA microarrays. Details of materials and methods can be found at `http://genome-www.stanford.edu/yeast_stress/materials.pdf`

Golub - Leukemia data

The dataset was downloaded from http://www.broad.mit.edu/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=43. and contains data pertaining to an experiment to monitor human acute leukemia's by Golub et al. (1999).

The dataset consists of 7129 genes across 72 samples, and represents unprocessed data. The data was processed according to technique outlined in Dudoit et al. (2002)

Hsiao data - Human tissue

Data was downloaded from http://www.biotechnologycenter.org/hio/databases/index.html, (HuGE Index) and represents a compendium of gene expression data for normal human tissues, (Hsiao et al., 2001).

The dataset represents 7,070 genes over 59 samples. As the data is from a collection of sources the units of expression level are arbitrary. All data was processed by the curators identically so that data can be compared across samples and tissues, (see http://www.biotechnologycenter.org/hio/faq/index.html).

Spellman - Yeast Cell cycle.

Data downloaded from http://genome-www.stanford.edu/cellcycle/data/rawdata/, complementing the work of Spellman et al. (1998). The dataset represents processed measurements from various experiments, with the aim to to identify all genes whose mRNA levels are regulated by the cell cycle in the yeast

Saccharomyces cerevisiae.

The dataset contains 6,178 genes across 82 sample points (and includes analysis of Cho et al. (1998) data). Spotted two cDNA arrays were used and details of materials and methods can be found at: `http://www.molbiolcell.org/cgi/content/full/9/12/3273#MaterialsMethods`.

Stegamaier - Kasumi data

Raw data was downloaded from `http://www.broad.mit.edu/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=117`. and pertains to myeloid differentiation in acute leukemia. The data represents HL-60 and Kasumi-1 cells treated in replicates of 3 with getinib at 10 μM or DMSO, and RNA was prepared at 6 and 24 hours for hybridization to Affymetrix U133A microarrays. (Stegmaier et al., 2005)

The dataset contains 22,283 probes over 22 samples and represented unprocessed values. The data was log transformed and scaled to have mean 0 and standard deviation of 1 across samples. Detail of materials and methods can be found at above url.

West - Breast Cancer

The data was downloaded from `http://data.cgt.duke.edu/west.php`. It contains data resulting from an experiment to analyse gene expression assays from breast cancer tissue, with the aim of identifying potential prognostic and/or predictive factors.

It contains 7,129 probes for 49 samples. The collection of tumors for RNA

extraction consisted of 13 estrogen receptor (ER) + lymph node (LN) + tumors, 12 ER−LN+ tumors, 12 ER−LN− tumors and 12 ER+LN− tumors. The level of RNA transcripts was measured using oligonucleotide array's (HuGeneFL Genechip array). The data was processed by scaling to have mean 0 and standard deviation of one. The data was log transformed.

Synthetic Dataset - Repeated Measures

The following functions were used to create the 11 groups in the synthetic repeated measures dataset. A 12^{th} group containing random data was also created.

1	Treatment effect on group 1, group 2 and group 3 similarly but to different levels.
2	Affect first individual in each group differently from others individuals
3	Affects each individual in similar way across treatment groups, apart from 3rd treatment group
4	Affects treatment groups 1 and 3 similarly, has no effect on 2nd treatment group (random numbers)
5	Affects treatment groups 1 and 3 in opposite way, has no effect on 2nd treatment group (random numbers)
6	Affects 4 individual in treatment groups 1 and 2 and 3 individuals in group 3. Affects each treatment of an individual equally not the same across treatment groups were it affects the last treatment of each individual differently.
7	Affects 4 individuals in groups 1 and 2; 2 individuals in group 3. Affects each treatment of individuals differently.
8	Affects only individuals in group 1, affects each treatment of an individual equally.
9	Affects first individual in each treatment group only, affects each treatment equally.
10	Affect each treatment, affects all individuals of each treatment equally.
11	Linear combination of groups 2 and 10, such that $j = (g2^3 + g10)/10$
Noise	Random

Synthetic Dataset - Time Series

The following functions were used to create the 11 groups in the synthetic time series dataset. For sine wave functions, $y(t) = A.sin(\omega \times t + \theta)$, $\omega \times t$ = frequency, θ = phase shift. $t1$ = time-series one, $t2$ = time-series two. A 12^{th} group containing random data was also created.

	T1	T2
1	$\omega = 2 \times pi/5, \theta = -10$ $sin(2 \times 3.14159 \times \frac{1:30}{5} - 10)$	$\omega = 2 \times pi/10, \theta = -10$ $sin(2 \times 3.14159 \times \frac{1:30}{10} - 10)$
2	$\omega = 2 \times pi/5, \theta = +10,$ $sin(2 \times 3.14159 \times \frac{1:30}{5} + 10)$	$A = 1, \omega = 2 \times pi/10, \theta = +10,$ $sin(2 \times 3.14159 \times \frac{1:30}{10} + 10)$
3	group 1 + group 2	
4	$A = 0.4, \omega = 2 \times pi/20, \theta = +0$ $0.4 \times sin(2 \times 3.14159 \times \frac{1:30}{20}) + 0.6$	random
5	$A = 1.5, \omega = 2 \times pi/40, \theta = +10$ $1.5 \times sin(2 \times 3.14159 \times \frac{1:30}{20} + 10)$	
6	$A = 1, \omega = 2 \times pi/40, \theta = +10$ $sin(2 \times 3.14159 \times \frac{1:30}{40} + 10)$	random
7	$A = 0.8, \omega = 2 \times pi/10, \theta = +0$ $0.8 \times sin(2 \times 3.14159 \times \frac{1:30}{10})$	$A = 1, \omega = 2 \times pi/3, \theta = +0$ $sin(2 \times 3.14159 \times \frac{1:30}{3})$
8	(i,j) $\rightarrow$ i = $\frac{j}{15}$	(i,j) $\rightarrow$ i = $\frac{-j}{15}$
9	(i,j) $\rightarrow$ i = $\frac{j^2}{1000}$	(i,j)$\rightarrow$ 0.5
10	random	$A = 2, \omega = 2 \times pi/40, \theta = +20$ $2 \times sin(2 \times pi \times (1:30)/40 + 20) + 0.5$
11	(group 7 + group 8)/10	
Noise	random	

APPENDIX C

GRAPHS, CHAPTERS 4, 5, AND 6

Chapter 4 - Method Analysis

Hierarchical Clustering - Internal Analysis

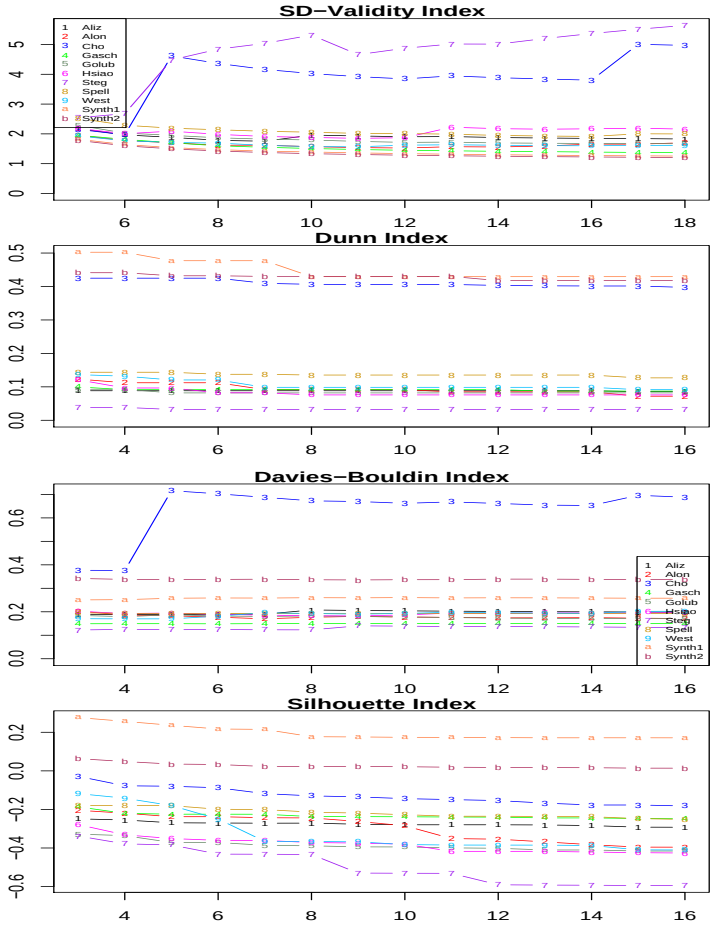


Figure C.1: Hierarchical Assessment (Single) using Internal measures. The x-axis indicates the cluster number while the y-axis indicates the score achieved obtained. The majority of indices for 'real' datasets remain constant across all K, however the Cho and Stegmaier datasets reveal information.

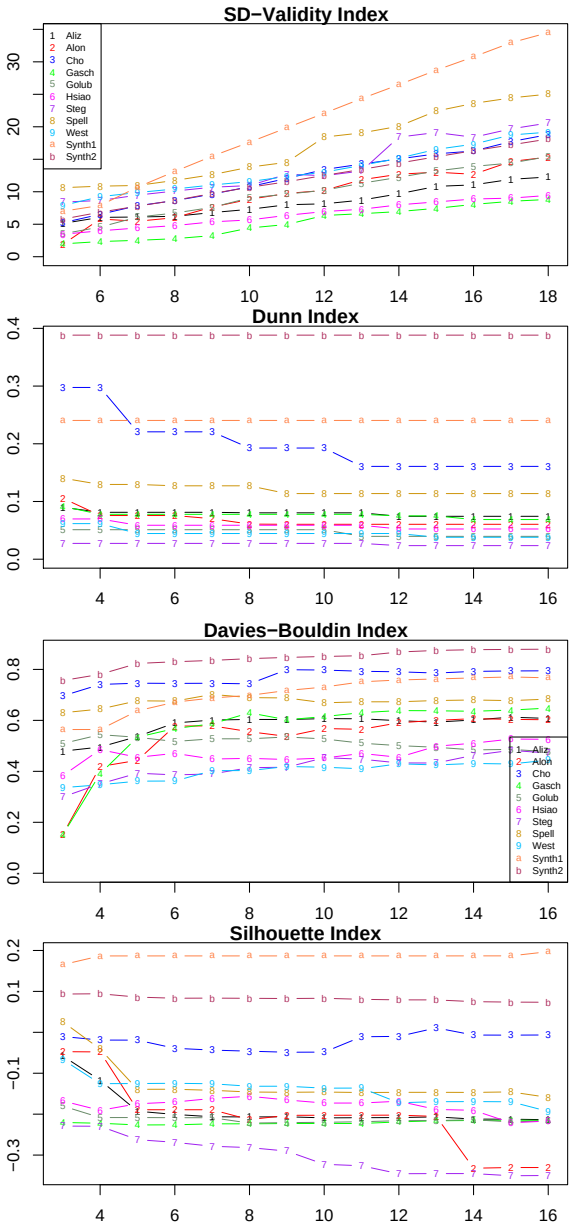


Figure C.2: Hierarchical Assessment (Average) using Internal measures. The x-axis indicates the cluster number while the y-axis indicates the score achieved obtained. Similarly to all linkage methods, the Silhouette index is < 0 for real datasets, indicating a bad clustering.

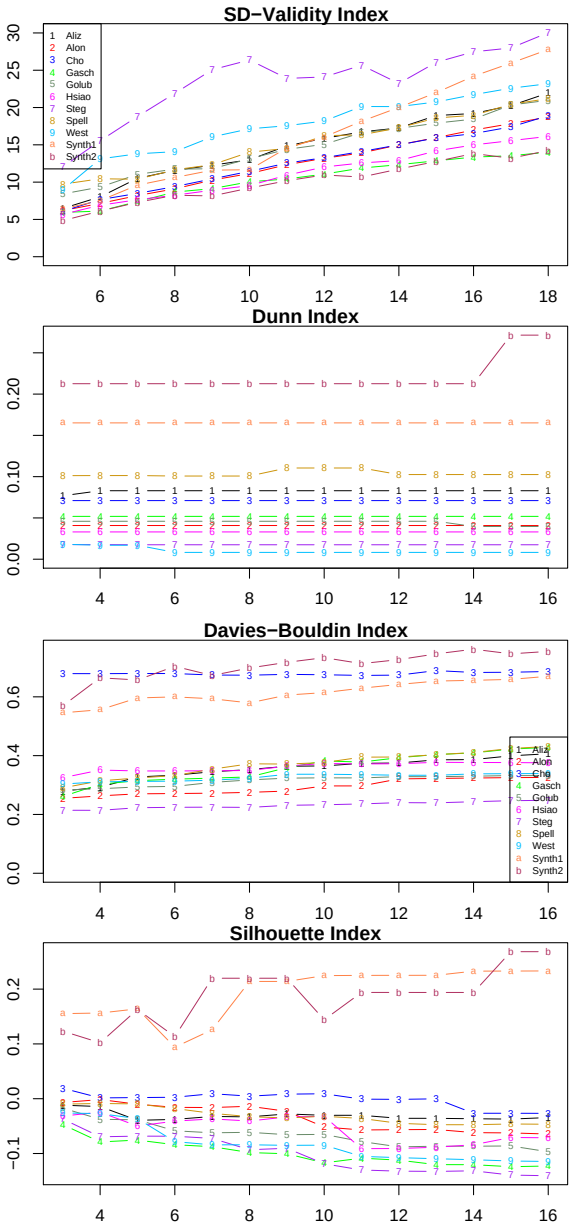


Figure C.3: Hierarchical Assessment (Complete) using Internal measures. The x-axis indicates the cluster number while the y-axis indicates the score achieved obtained. The Dunn index indicates a better clustering in Synthetic datasets compared to 'real' datasets.

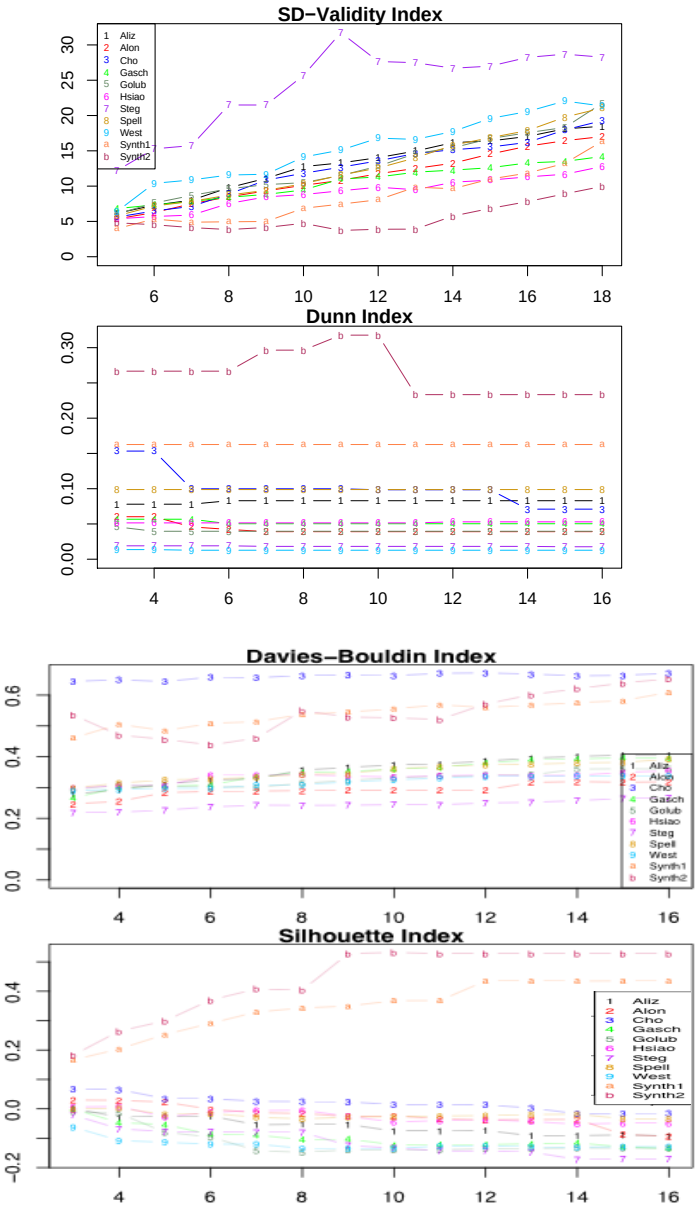


Figure C.4: Hierarchical Assessment (Ward) using Internal measures. The x-axis indicates the cluster number while the y-axis indicates the score achieved obtained. As with all linkage methods, the SD-Validity index increases with K, (small values indicate a better clusterings.)

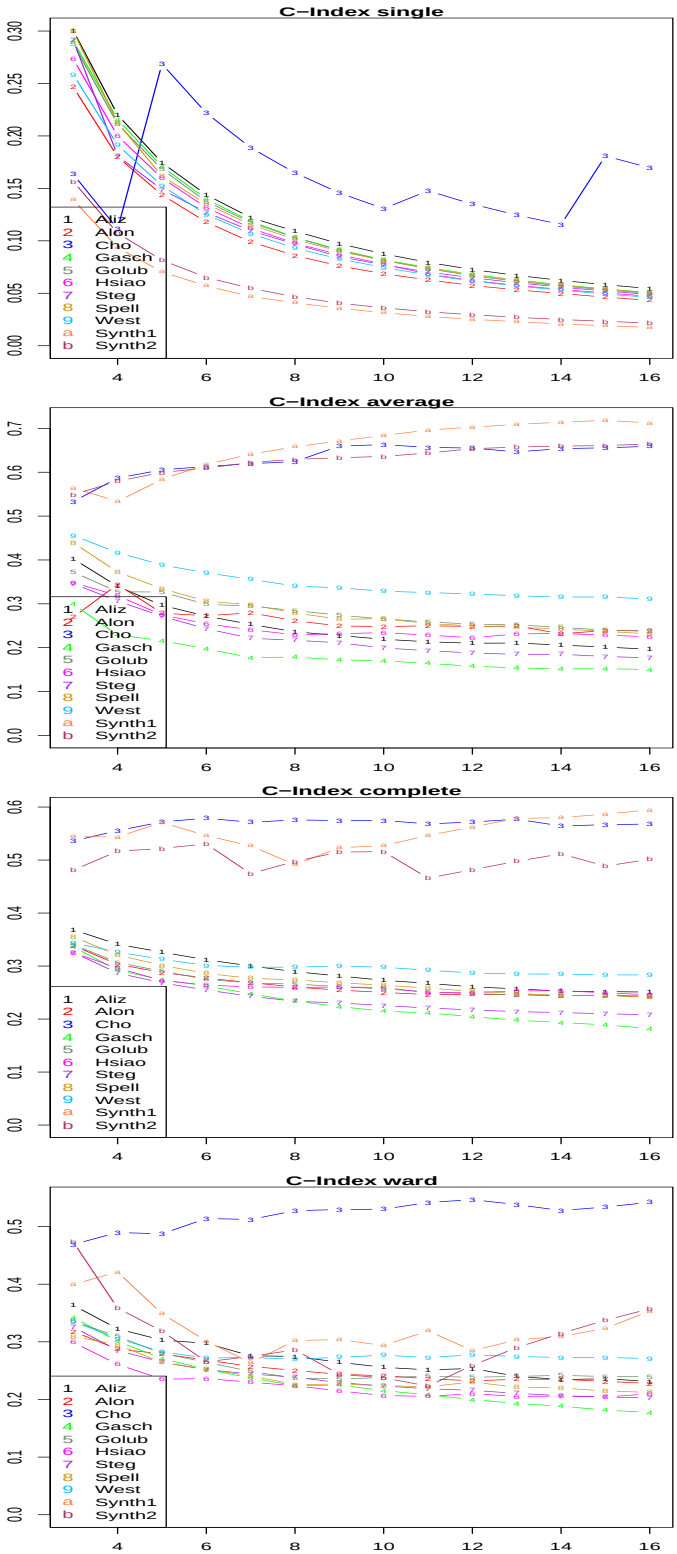


Figure C.5: Hierarchical Assessment using C-Index. The x-axis indicates the cluster number while the y-axis indicates the score achieved obtained. Contrary to other assessment metrics, this index indicates a better structure in the 'real' datasets compared to Synthetic (small values indicate a better clustering).

Hierarchical Clustering - Stability Analysis

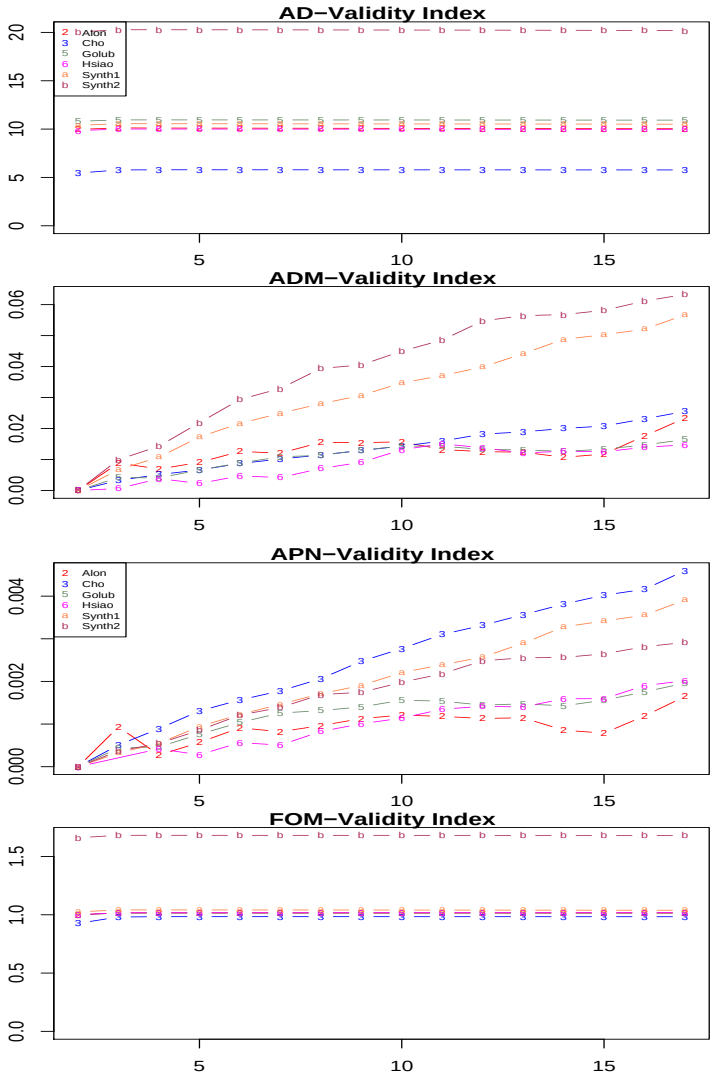


Figure C.6: Hierarchical Assessment (single linkage) using AD,ADM, APN and FOM stability Measures. The x-axis indicates the cluster number while the y-axis indicates the score achieved obtained.

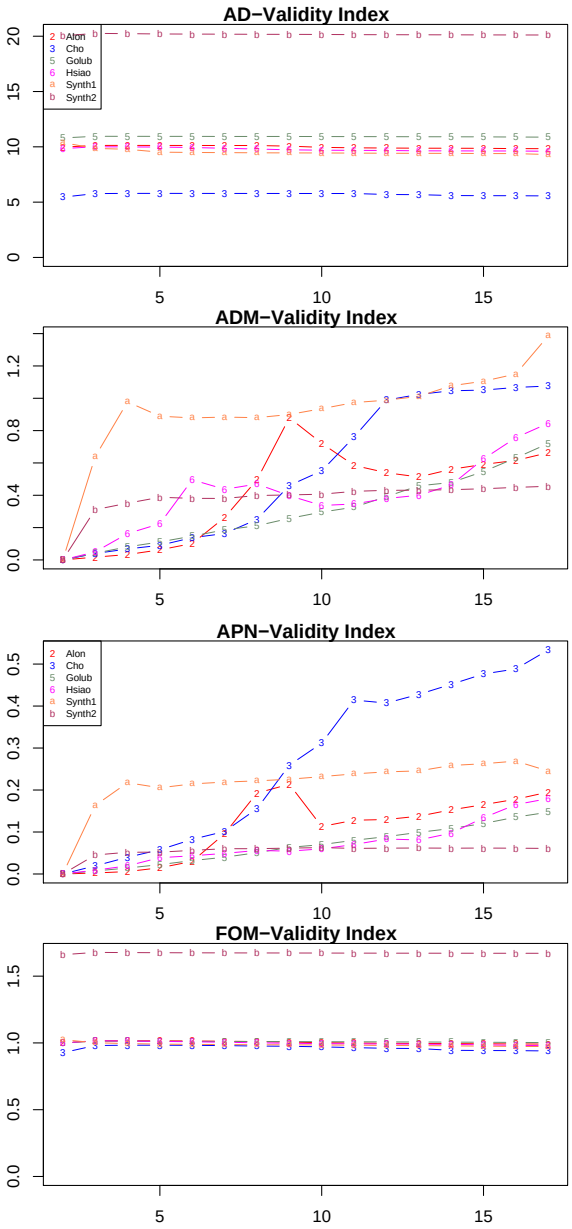


Figure C.7: Hierarchical Assessment (average linkage) using AD, ADM, APN and FOM stability Measures. The x-axis indicates the cluster number while the y-axis indicates the score achieved obtained. The APN index reveals a marked deterioration in stability with the Cho dataset as the value of K increases. AD and FOM indices reveal little information.

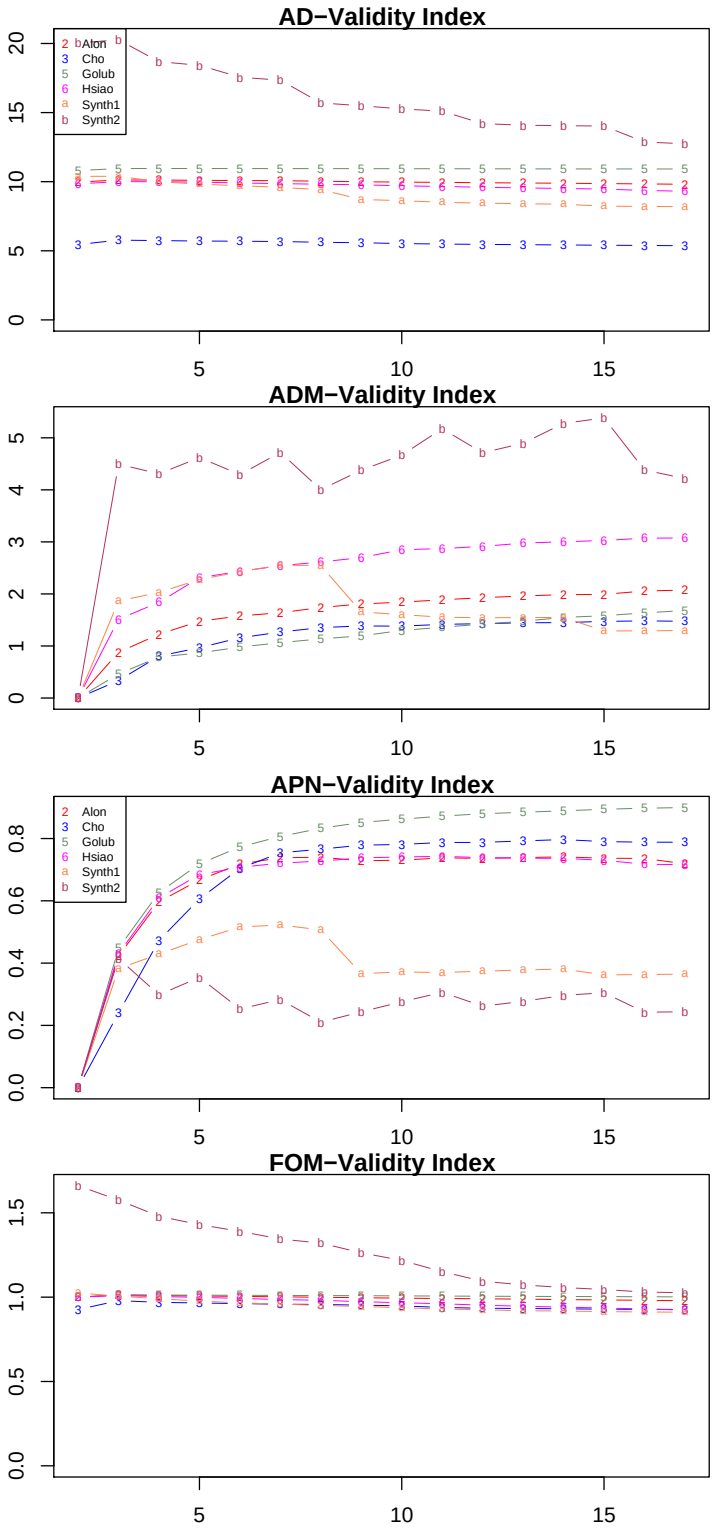


Figure C.8: Hierarchical Assessment (complete linkage) using AD, ADM, APN and FOM stability Measures. The x-axis indicates the cluster number while the y-axis indicates the score achieved obtained.

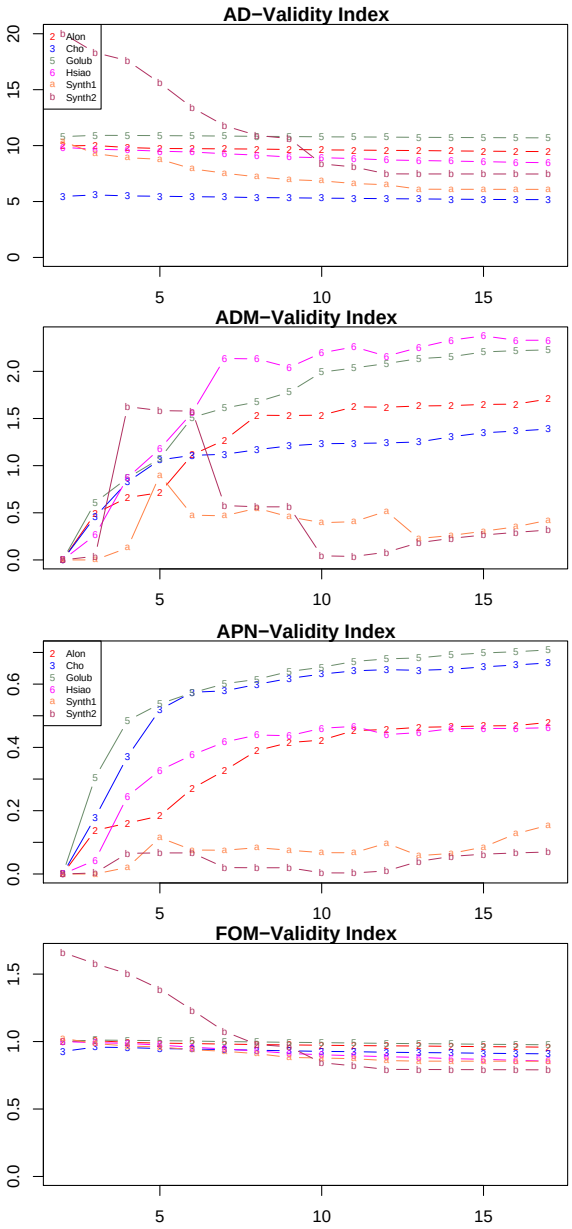


Figure C.9: Hierarchical Assessment (Ward linkage) using AD, ADM, APN and FOM stability Measures. The x-axis indicates the cluster number while the y-axis indicates the score achieved obtained. The APN values in Ward and Complete linkage increase exponentially with K, compared to a linear increase observed with Single and Average linkage.

Hierarchical Clustering - External Analysis

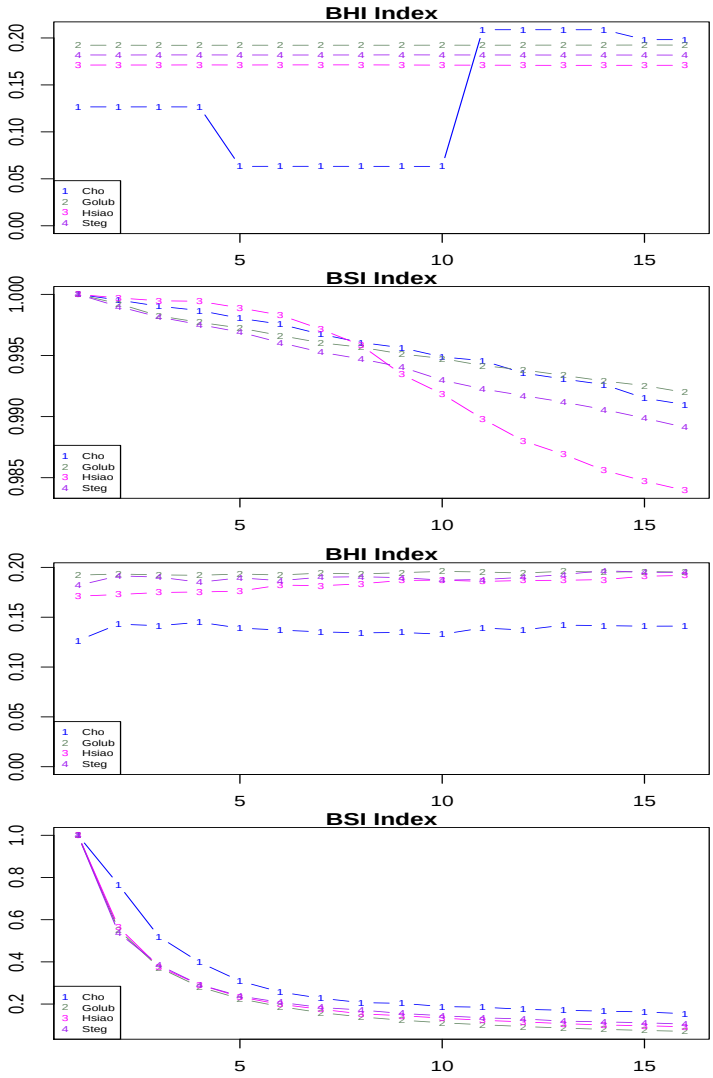


Figure C.10: Hierarchical Assessment (Single and Complete linkage) using Biological Homogeneity and Biological Stability Measures . The x-axis indicates the cluster number while the y-axis indicates the score achieved obtained.

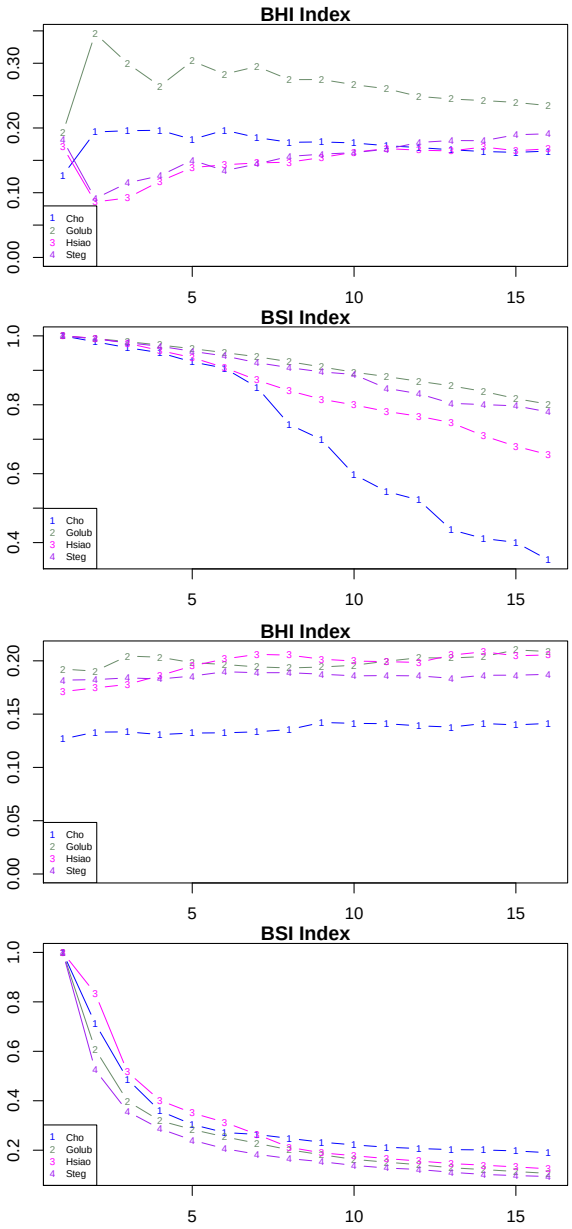


Figure C.11: Hierarchical Assessment (Average and Ward linkage) using Biological Homogeneity and Biological Stability Measures . The x-axis indicates the cluster number while the y-axis indicates the score achieved obtained. The BSI stability indices exponentially decrease in Ward and Complete linkage, while the linearly decrease with Single and Average linkage.

KMeans - Internal Analysis

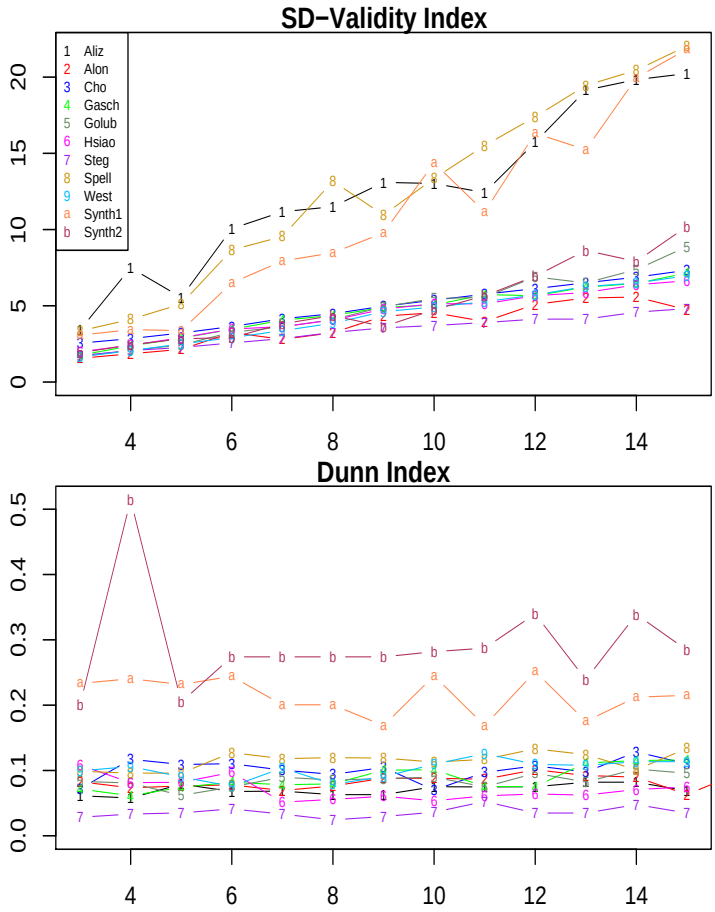


Figure C.12: K-means Assessment using Internal measures. The x-axis indicates the cluster number while the y-axis indicates the score achieved obtained. Better clustering are obtained for the Synthetic datasets when assessed with Dunn metric (larger values preferred). Again, SD-Validity increases with K (smaller values preferred).

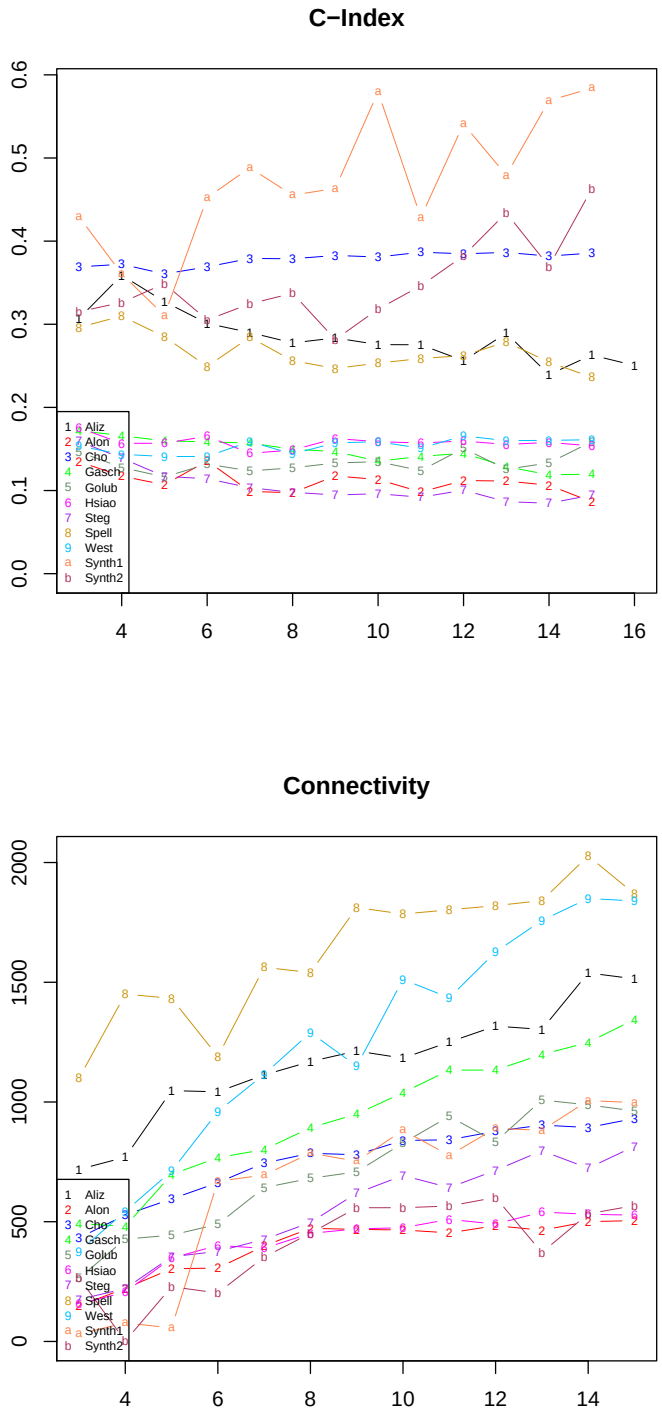


Figure C.13: K-means Assessment using Internal measures c-index and connectivity. The x-axis indicates the cluster number while the y-axis indicates the score achieved obtained.

KMeans - Stability Analysis

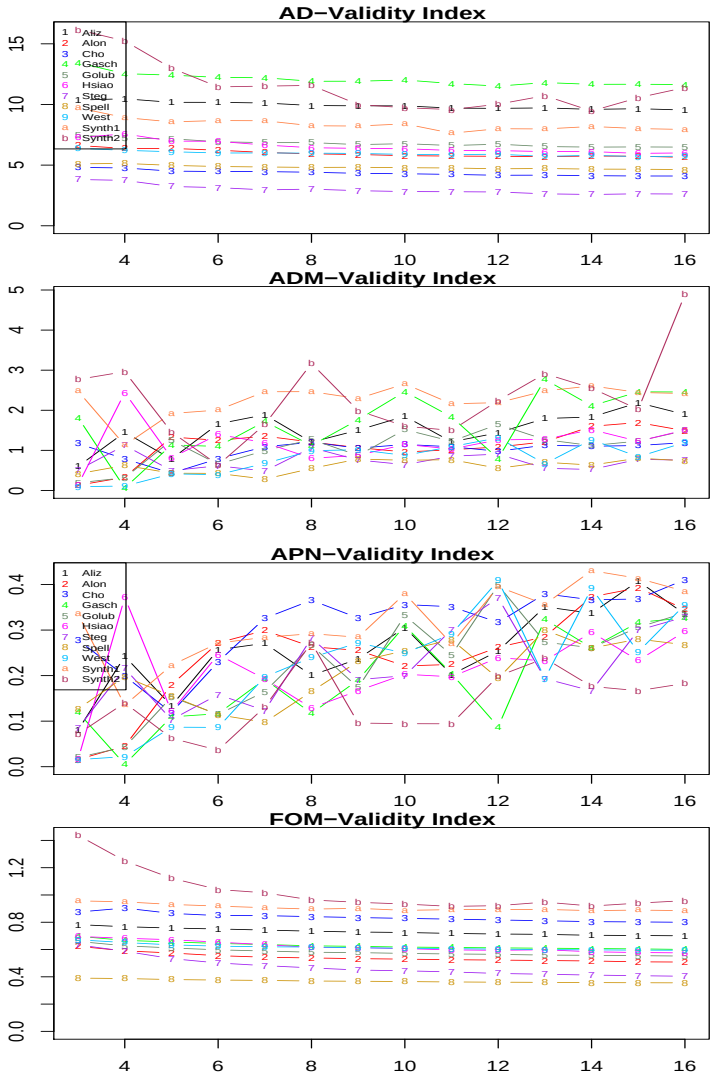


Figure C.14: K-means Assessment using Stability measures. The x-axis indicates the cluster number while the y-axis indicates the score achieved obtained.

SOTA - Internal Analysis

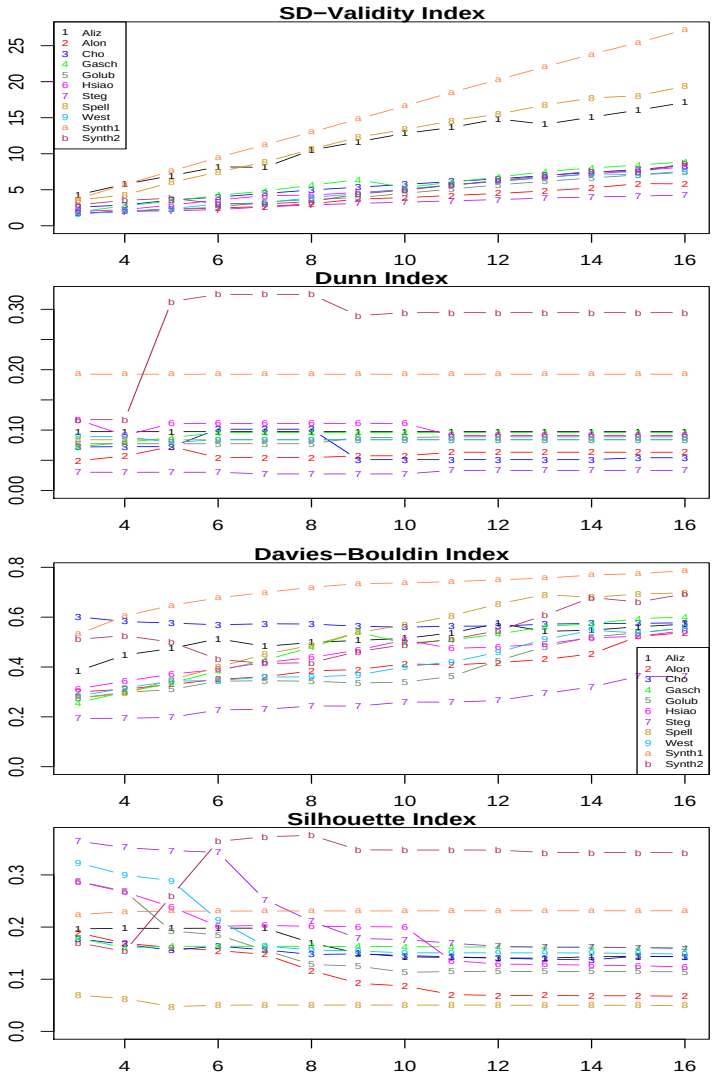


Figure C.15: SOTA Assessment using Internal measures. The x-axis indicates the cluster number while the y-axis indicates the score achieved obtained.

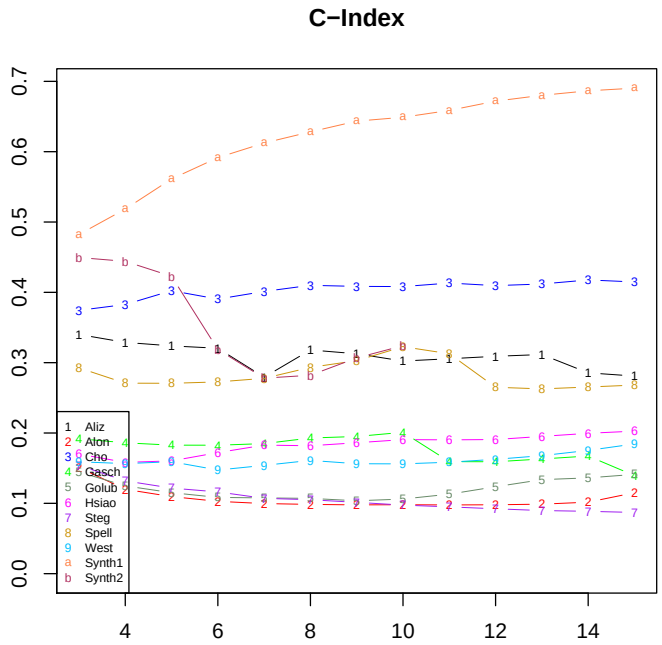


Figure C.16: SOTA Assessment using Internal measure - c-index. The x-axis indicates the cluster number while the y-axis indicates the score achieved obtained.

SOTA - Stability Analysis

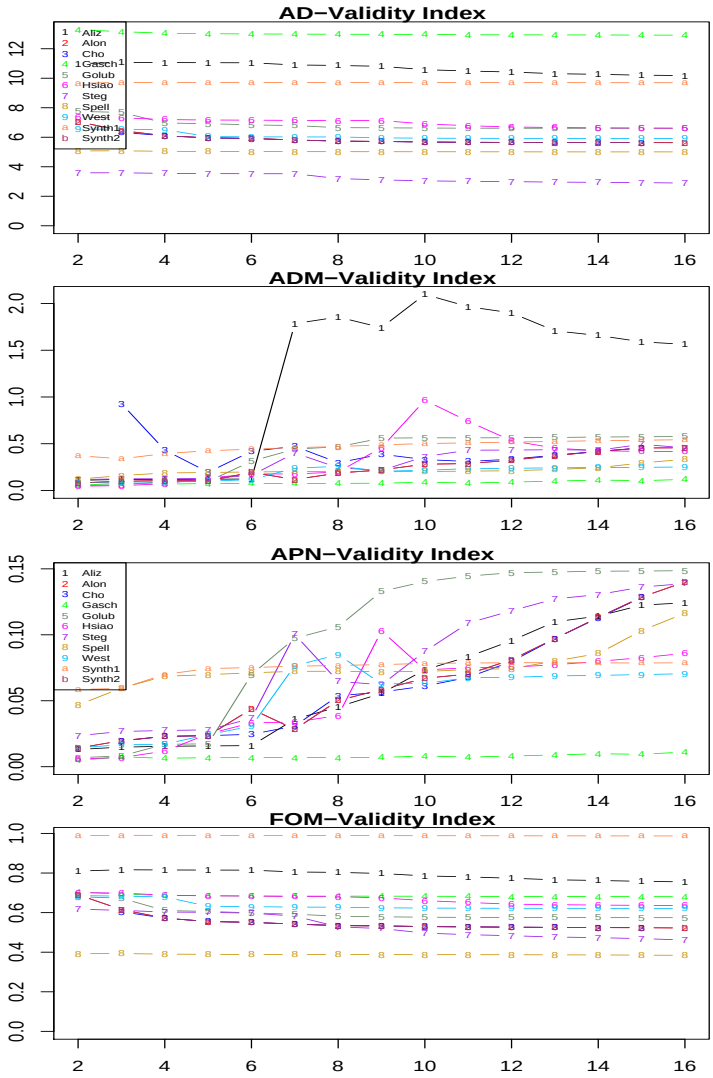
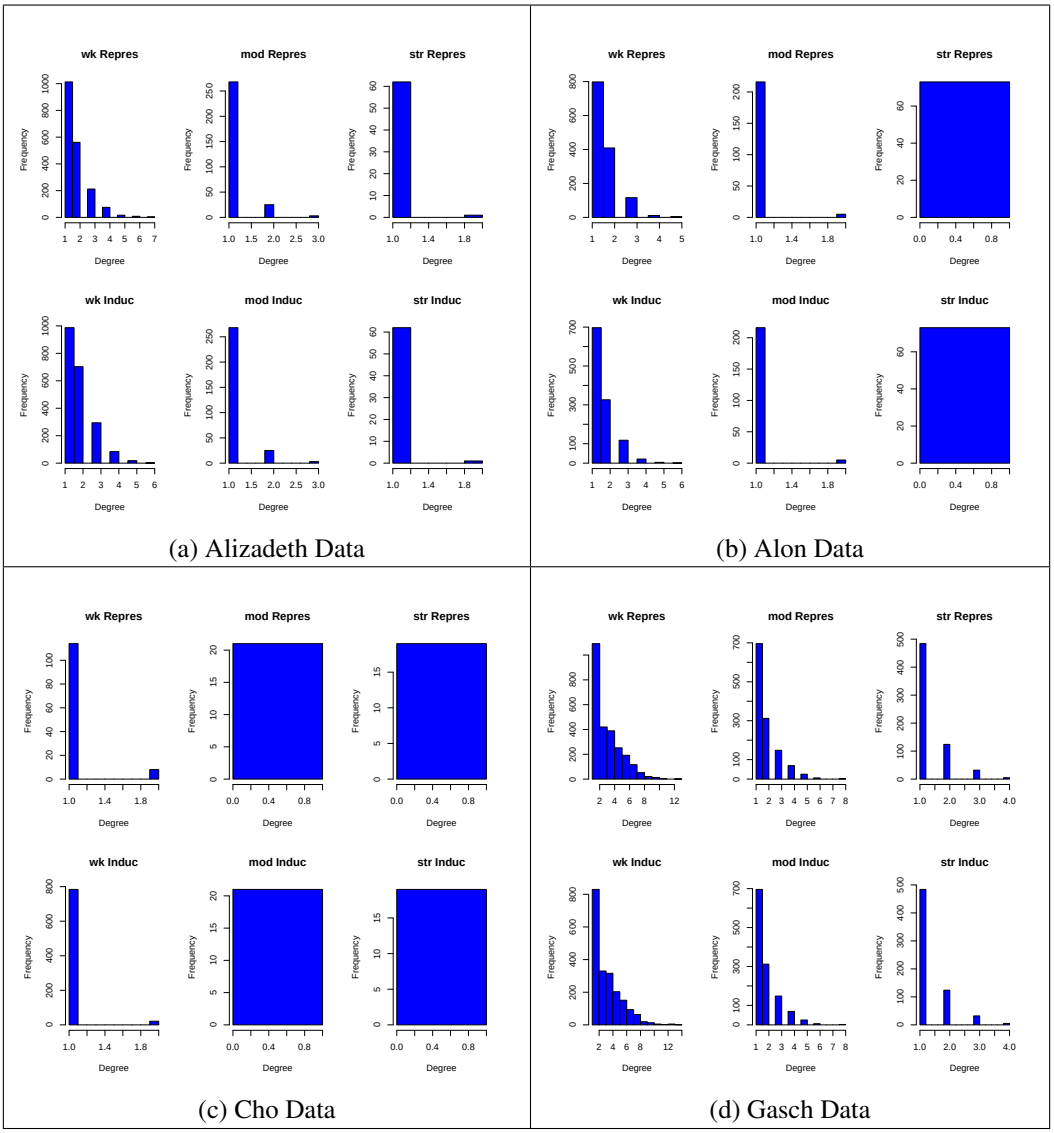


Figure C.17: SOTA Assessment using Stability measures. The x-axis indicates the cluster number while the y-axis indicates the score achieved obtained.

Chapter 5

Node Degree -Bipartite SubGraphs



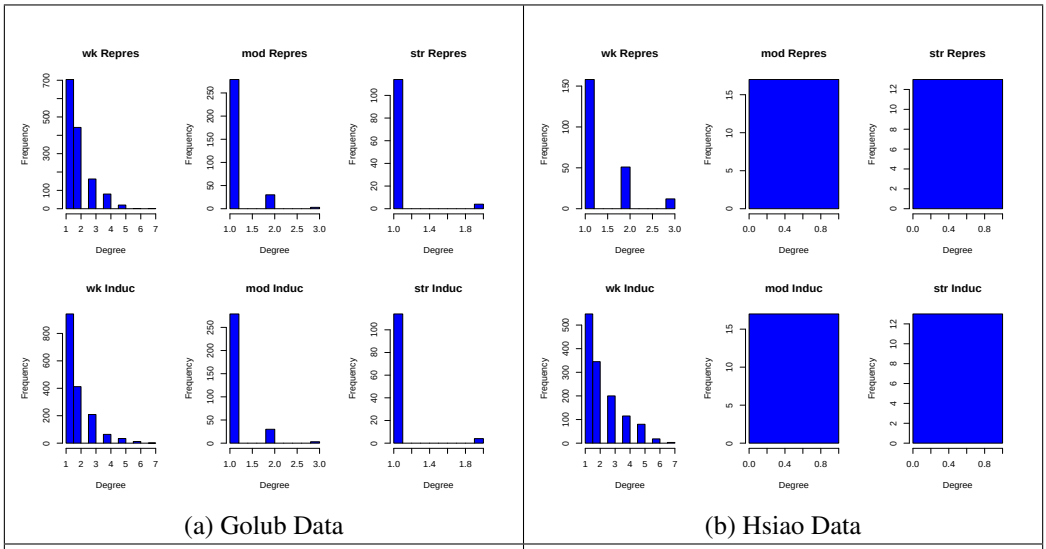
(a) Alizadeth Data

(b) Alon Data

(c) Cho Data

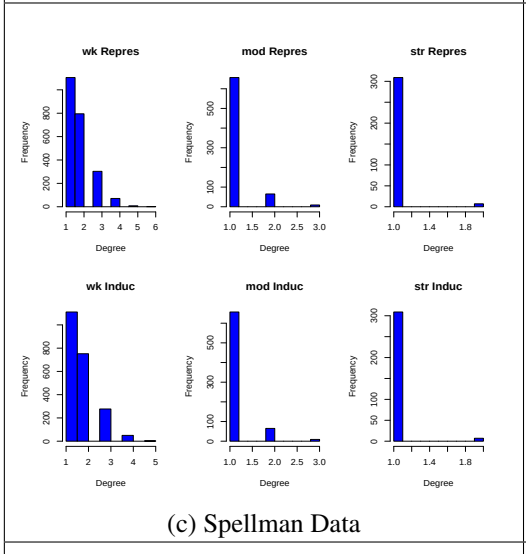
(d) Gasch Data

Figure C.18: Gene Degree Distribution of Alizadeth, Alon, Cho and Gasch subgraphs

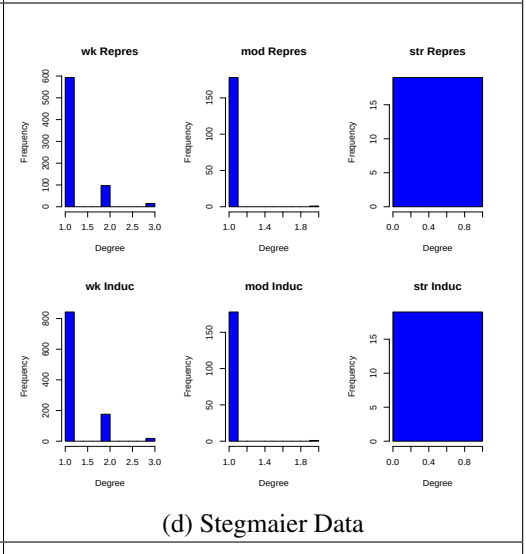


(a) Golub Data

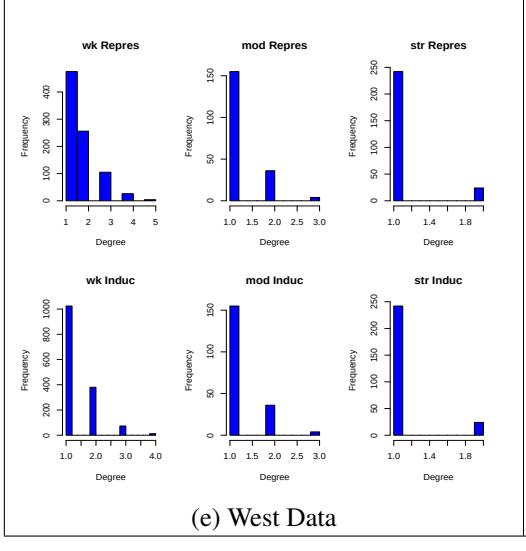
(b) Hsiao Data



(c) Spellman Data

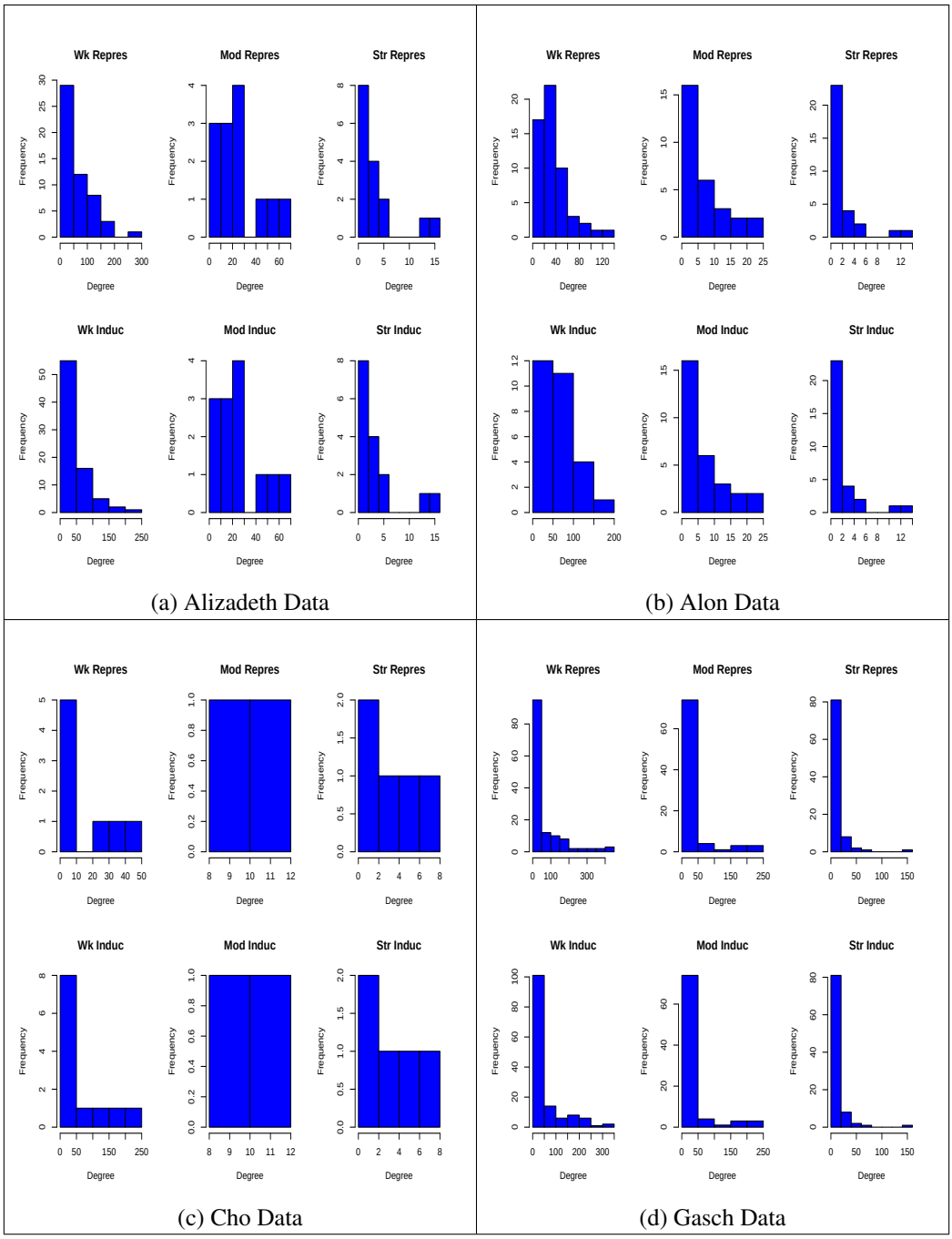


(d) Stegmaier Data



(e) West Data

Figure C.19: Gene Degree Distribution of Golub, Hsiao, Spellman, Stegmaier and West subgraphs



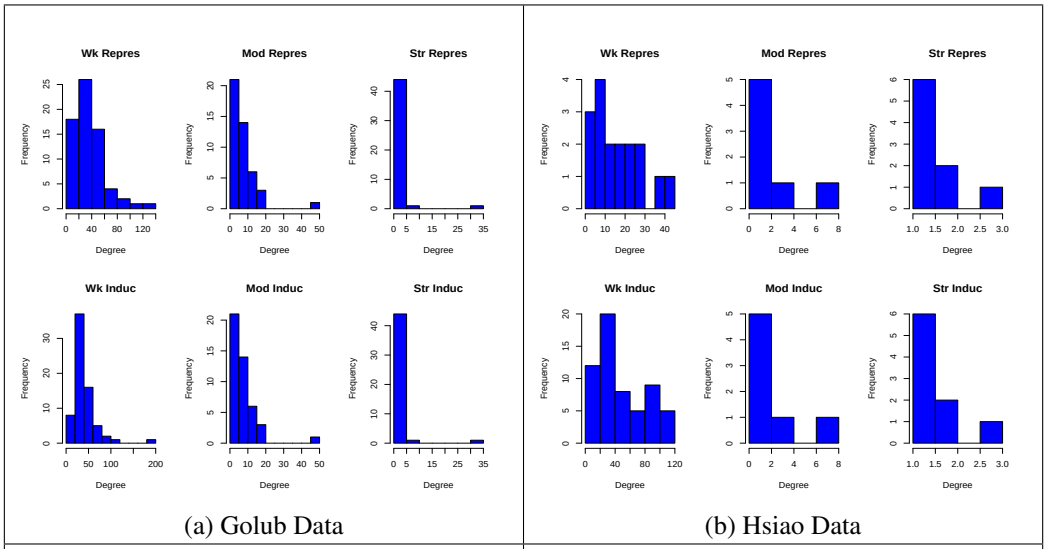
(a) Alizadeth Data

(b) Alon Data

(c) Cho Data

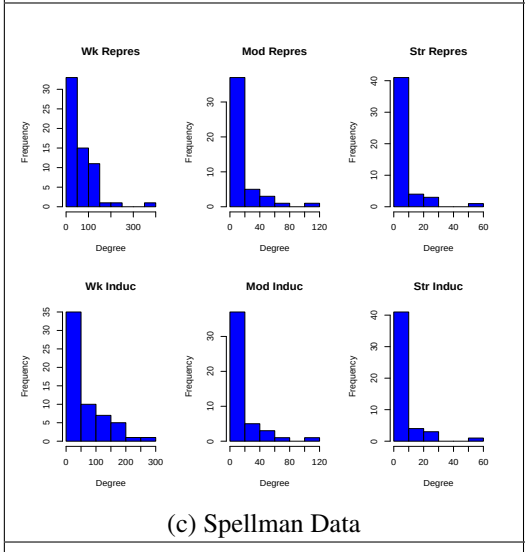
(d) Gasch Data

Figure C.20: Sample Degree Distribution of Alizadeth, Alon, Cho and Gasch subgraphs

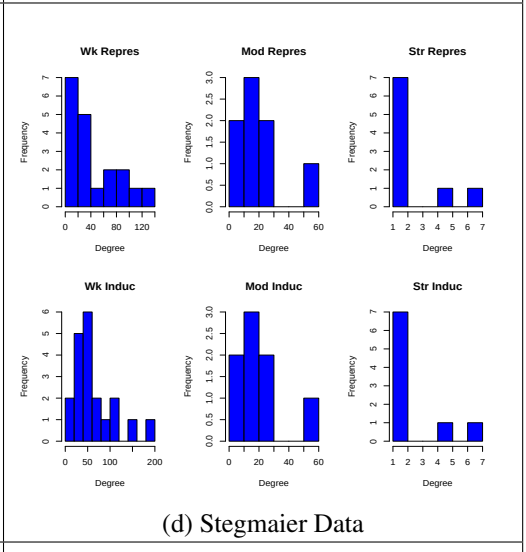


(a) Golub Data

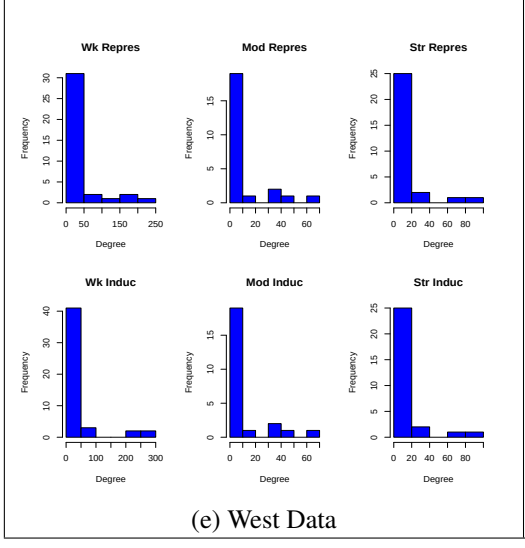
(b) Hsiao Data



(c) Spellman Data



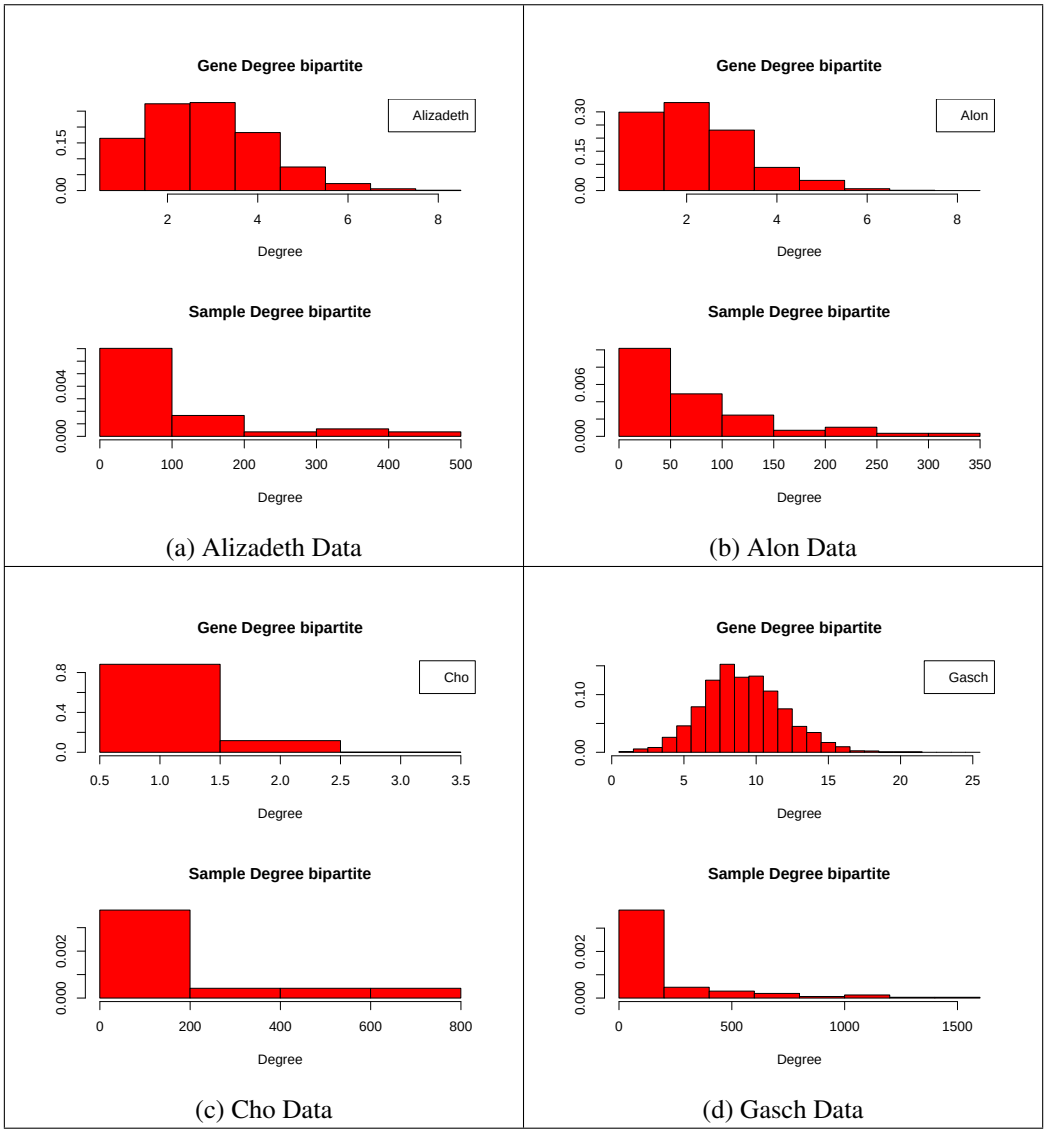
(d) Stegmaier Data



(e) West Data

Figure C.21: Sample Degree Distribution of Golub, Hsiao, Spellman, Stegmaier and West subgraphs

Node Degree Distributions -Bipartite *All In One* Graphs



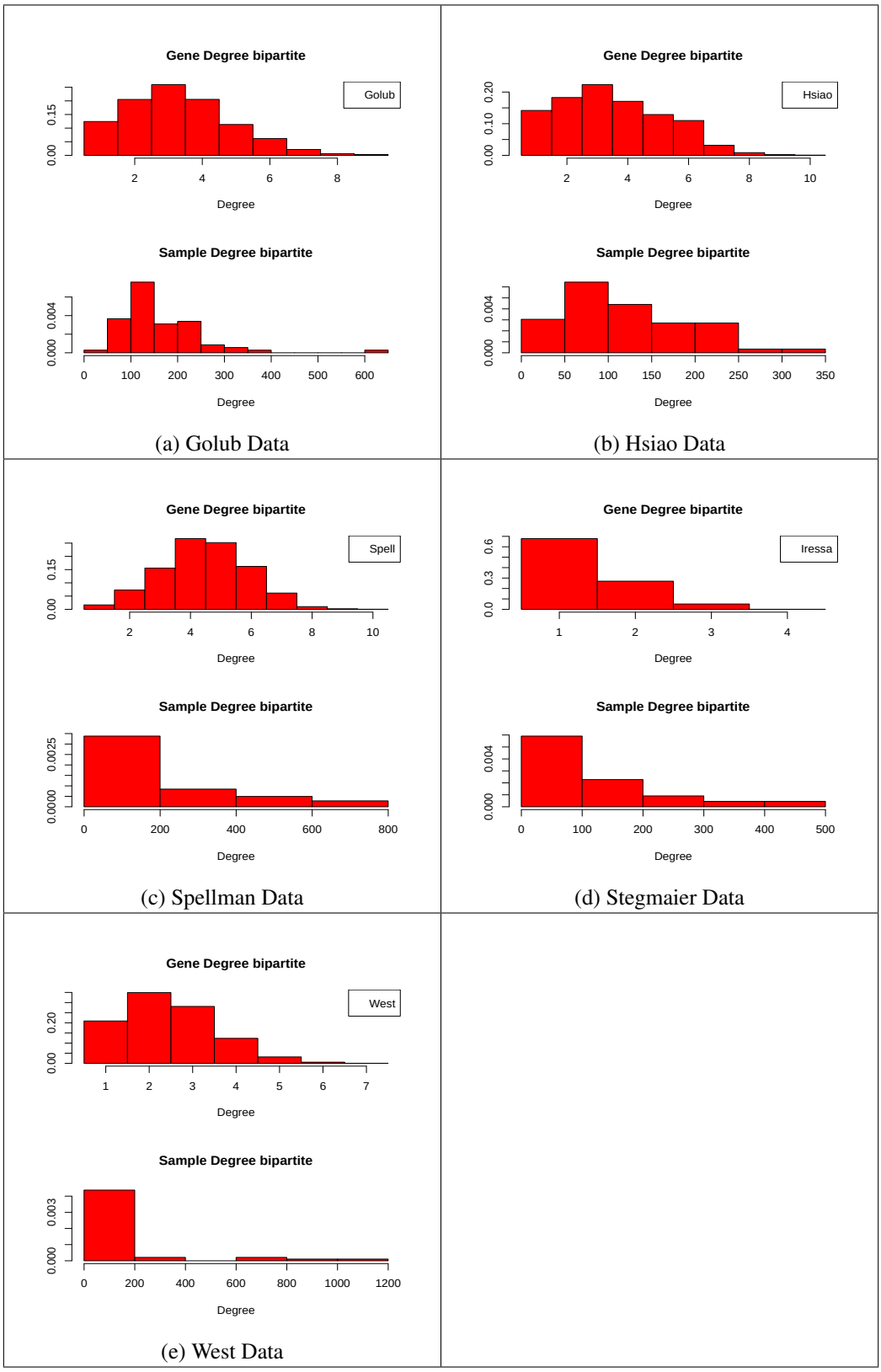
(a) Alizadeth Data

(b) Alon Data

(c) Cho Data

(d) Gasch Data

Figure C.22: Degree Distribution of Alizadeth, Alon, Cho and Gasch datasets



(a) Golub Data

(b) Hsiao Data

(c) Spellman Data

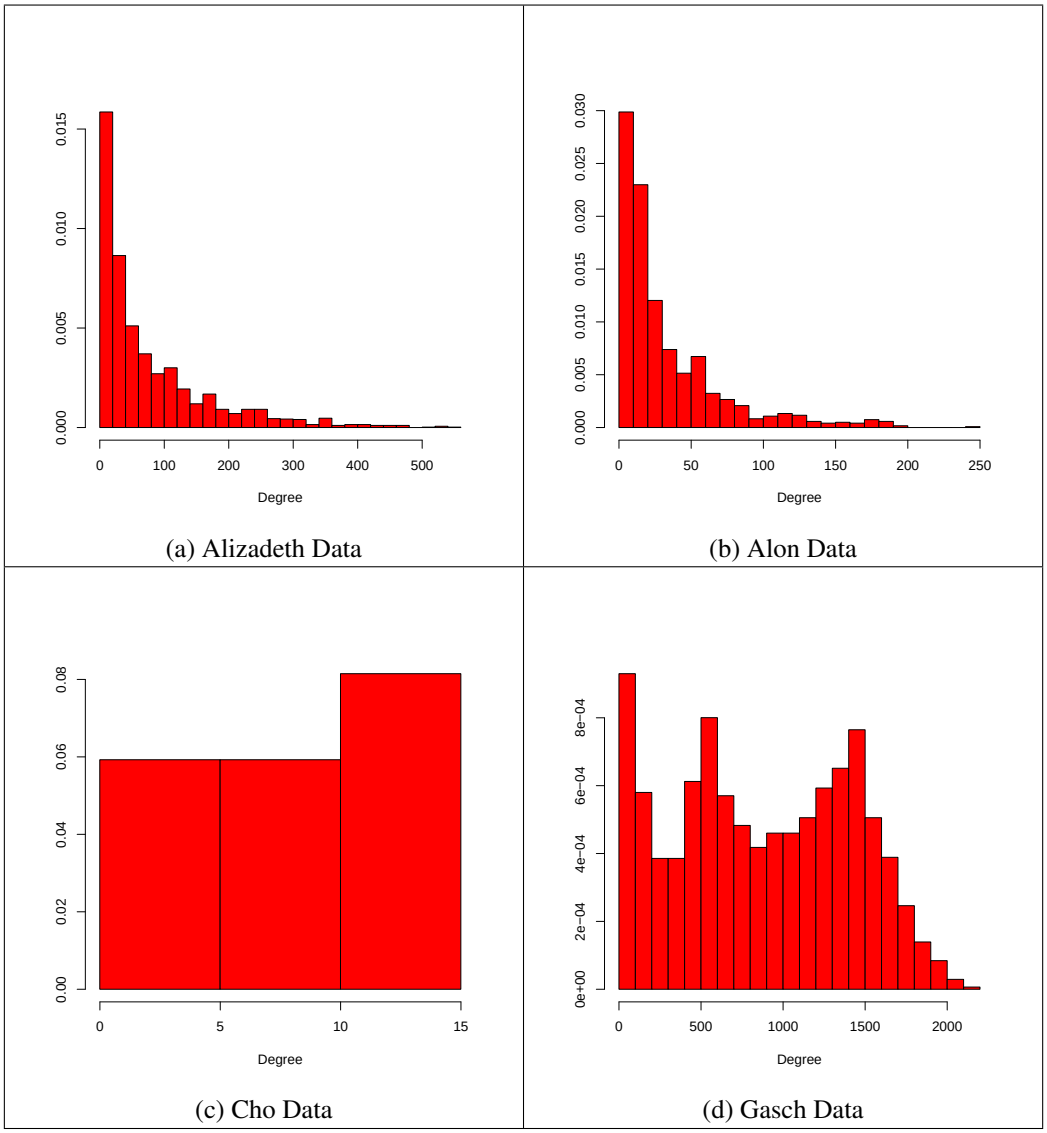
(d) Stegmaier Data

(e) West Data

Figure C.23: Degree Distribution of Golub, Hsiao, Spellman, Stegmaier and West datasets

Chapter 6

Node Degree Distributions



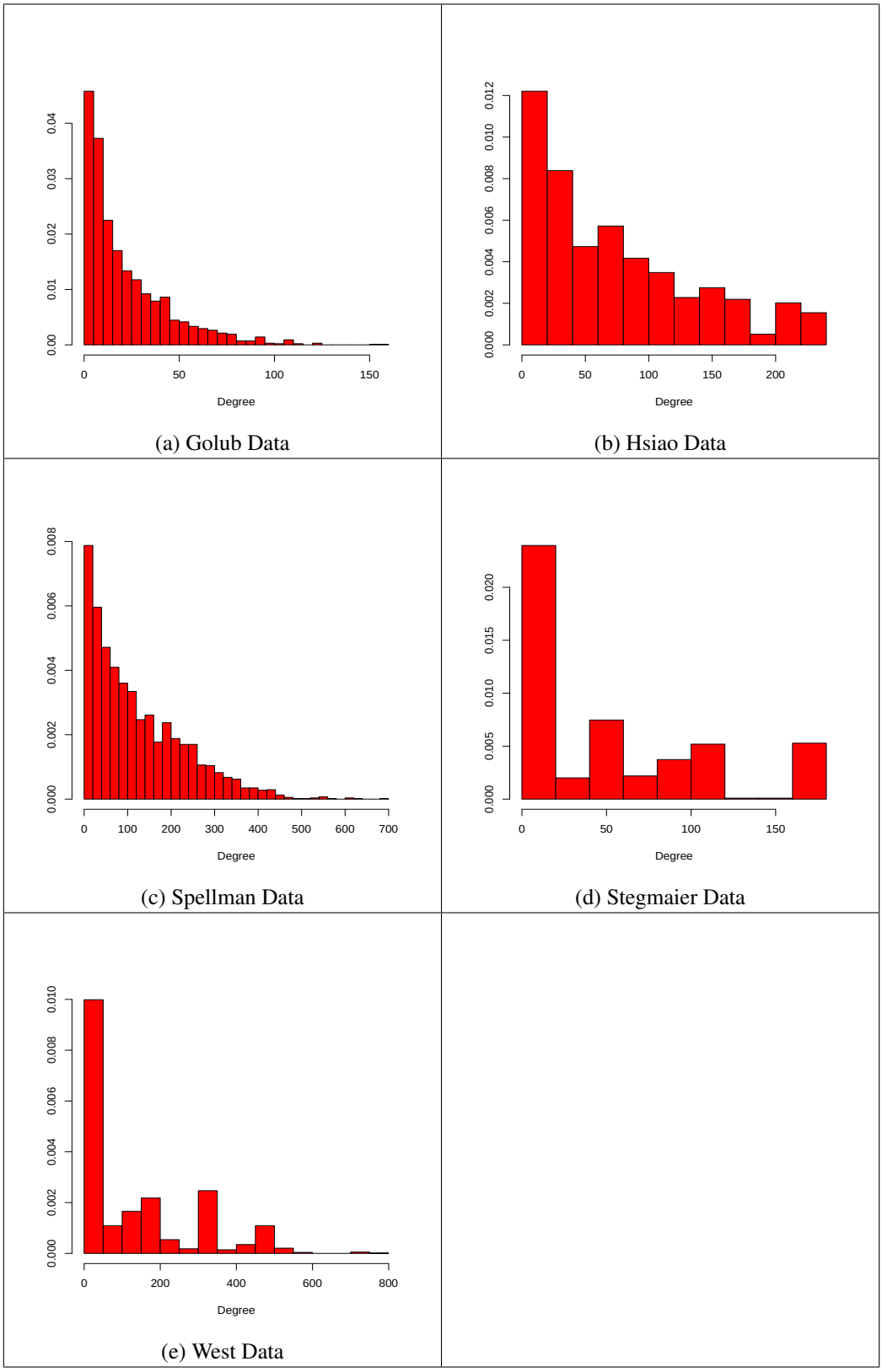
(a) Alizadeth Data

(b) Alon Data

(c) Cho Data

(d) Gasch Data

Figure C.24: Degree Distribution of Alizadeth, Alon, Cho and Gasch datasets



(a) Golub Data

(b) Hsiao Data

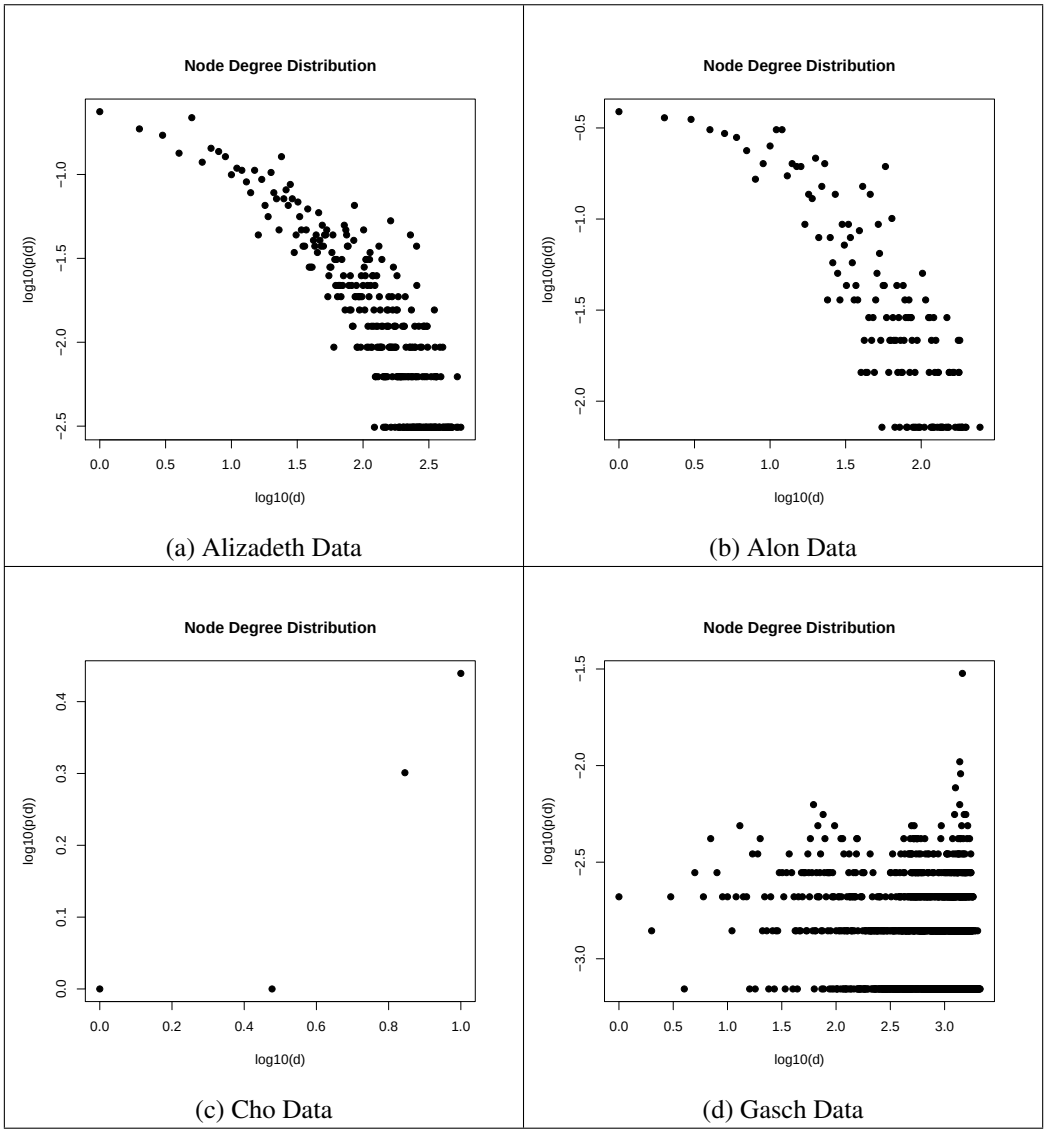
(c) Spellman Data

(d) Stegmaier Data

(e) West Data

Figure C.25: Degree Distribution of Golub, Hsiao, Spellman, Stegmaier and West datasets

Degree Distribution One Mode



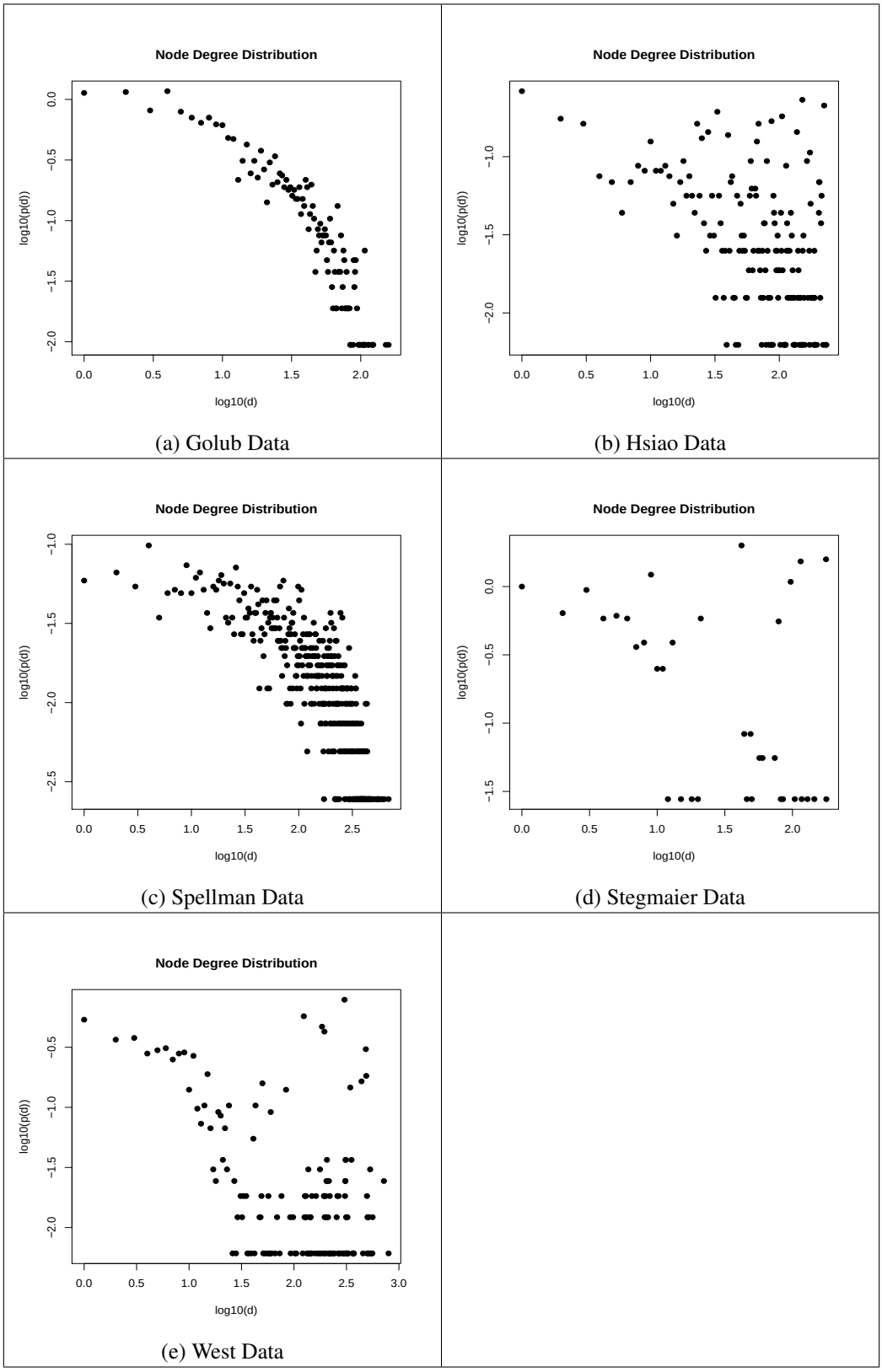
(a) Alizadeth Data

(b) Alon Data

(c) Cho Data

(d) Gasch Data

Figure C.26: Degree Distribution of Alizadeth, Alon, Cho and Gasch datasets



(a) Golub Data

(b) Hsiao Data

(c) Spellman Data

(d) Stegmaier Data

(e) West Data

Figure C.27: Degree Distribution of Golub, Hsiao, Spellman, Stegmaier and West datasets

List of Publications:

Kerr G., Ruskin H. J., Crane M., Doolan P. (2007), Techniques for Clustering Gene Expression Data, *Computers in Biology and Medicine*, Nov 30, 2007.

Kerr G., Ruskin H.J. and Crane M., Pattern Discovery in Gene Expression Data, in Wang, H.F. (Ed), *Intelligent Data Analysis: Developing New Methodologies Through Pattern Discovery and Recovery,* Idea Group Publishing Ltd, 2008, ISBN 978-1599049823.

Kerr G., Perrin P., Ruskin H. J., Crane M. (Accepted Manuscript) Edge Weighting of Gene Expression Graphs, *Advanced Complex Systems.*

This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.

This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.

This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.

This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.

This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.

This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.

This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.

This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.

This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.

This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.

This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.

This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.

This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.

This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.

This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.

This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.

This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.

This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.

This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.

This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.

This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.

This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.

This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.

This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.

This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.

This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.
This content has been removed due to third party copyright.