University of Strathclyde

Department of Bioengineering

Novel Network Approaches for the Interrogation of Large Data Sets with Relevance to Schizophrenia

Martin McDonald

A thesis presented for the degree of
Doctor of Engineering

December 2013

This thesis is the result of the authors original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Signed:

Date:

Acknowledgements

I am grateful to everyone involved in funding and supporting my time spent on this project, whether they were seen or unseen. It is a privileged position to be able to study for the length of time required to complete a project such as this, I realise that and I realise that I am fortunate to have been afforded many opportunities I would not have met otherwise as a result.

I would like to thank my trio of supervisors, Des Higham, Ben Pickard and Judy Pratt. I am thankful for the humour, confidence and passion in their approach, which has contributed immensely to the process. Neil Dawson for our collaboration which formed much of the early parts of my work. Keith Vass for many enjoyable discussions, and for helping guide my way of thinking for the future. The neuroscience group at the university for listening to my talks, and giving me a little more insight into the complicated world of biology from listening to theirs. PsyRING and CeNsUS have supported in this vein and more.

My office mates in the Department of Mathematics & Statistics also deserve a mention. In particular, I would like to thank Alan for beginning and continuing many interesting office distractions. My sense of skepticism is now fortified with a little statistics knowledge. Eusebio and his wife Marita for introducing me to some of the culinary delights from their home, Mexico. Wei Liu with his inspired work ethic, and for his words about a corner of the world I have not yet experienced. Steven for, despite what little time he spent in the office, we have had a number of interesting, provocative discussions. This multicultural office was a true perk, and is one I have enjoyed.

The Bioengineering Department Doctoral Training Centre in Medical Devices has supported this project, and introduced me to a number of people I am pleased to

have shared a classroom with. I have had many interesting chats with Jamie over lunch, I am sure he will see an idea meet success.

My family have all contributed in one way or another to the position I am in today and I would like to thank each of them. I appreciate the many productive TD chats with my friend Graeme, and Craig as a source of confirmation bias. Finally, Julie, for love and support and for finding us a home and making us slave to a duo of cats. I am proud of everything you have accomplished, and happy you have shared it with me - we are lucky indeed.

Abstract

Complex networks are an important tool for the study of biological data. There are two main aims in this data-driven work, which are explored in tandem. We study (1) the nature of schizophrenia and (2) utility in novel additions to traditional network based spectral clustering methods. More specifically, we explore three facets of schizophrenia. First, we study functional brain data in animal models of relevance to the condition. Second, we examine the impact of antipsychotic medication on gene expression in humans, and third we assess whole blood for potential as a suitable alternative to brain tissue. With regard to spectral clustering, we employ the Singular Value Decomposition and the Generalized Singular Value Decomposition in a way that allows us to incorporate additional information into the clustering problem. This work is of interest in the life sciences due to the complex heterogeneous nature of schizophrenia, which has created desire for analysis of large amounts of data. In addition, development of network based approaches is a timely area of study in general given recent explosions in the amount of data produced across many subject areas.

Our interdisciplinary work leads to four main conclusions: (a) network approaches for functional brain animal model studies can produce results that are biologically meaningful in humans, (b) a novel node-weighted version of the Laplacian is a flexible tool that allows multiple sources of network information to be combined, (c) antipsychotic medication, used routinely to treat schizophrenia, has a dominant effect on gene expression as compared to the control state, masking the underlying nature of the disease and (d) human whole blood is useful for the study of gene expression in schizophrenia.

Contents

List of Figures

List of Tables

Chapter 1

Mathematics Introduction

Networks offer a method of observing and studying holistic, emergent, global and local properties of systems in full context. This is a timely area of study, with a number of important applications in a wide range of areas from computer science and physics to sociology and biology. Recent growth in the field owes much to increases in computational power, digitisation of data and development of novel high-throughput methods for gathering and processing data. This is particularly relevant in a time when it is recognised that unexpected qualities emerge as a consequence of interactions within a network - this ‘systems level’ thinking embraces the idea that interactions cause the whole to be greater than the sum of the constituents.

In this chapter we provide a brief overview for some important concepts at the intersection of network science and applied linear algebra, focussing on methods used for clustering that are most relevant to the problems we will confront in this thesis. Chapter 6 provides a more detailed discussion of some of these key ideas when it is necessary.

1.1 Complex networks

A network is a mathematical tool that can be described as a collection of objects and their relationships. As an interdisciplinary field, various approaches, tools and terminology have been developed to suit complex networks - for example, in network chemistry the objects of study may be *sites* and the connection/interaction between them *bonds*. In sociology the terms commonly used are *actors* and *ties*. For this thesis, we will use *node* to describe all objects in the various networks explored, and *edge* to refer to the connections/interactions. As implied by these splits in terminology, networks are useful in a variety of circumstances, for example in biology to understand cellular function [13], in sociology to understand dynamics of human relationships [193], in finance to mine market data [18] and in technology to explore structure of the World Wide Web [22]. The field itself has origins in the mathematics of graph theory (the area of mathematics concerning the study of networks) - first demonstrated in Euler's famous 1736 paper solving the problem of traversing the Seven Bridges of Königsberg [59]. In this paper, Euler realised that information can be condensed and abstracted to core components - in a problem of finding a route to traverse bridges the key information is simply a collection of *nodes* (the points of land) and *edges* between them (the bridges).

There are also means to include additional information into network interpretations of data. For example, networks can be *directed* where links (edges) form connections between websites (nodes), as in the case of the internet where links go from one website to another and not in the other direction. Networks can also be *undirected*, where the structure is symmetric - in a network describing handshakes where both people (nodes) participate. Networks may be weighted, where edges are given a number that describes the strength of the relationship e.g., in an academic citation network where it may not only be desirable to show which node (academic) has cited (edge) which other nodes but also to include the number of citations (edge weight) between academics or unweighted where the edges are binary and the number of citations would not be counted. These options allow for utility and flexibility that makes networks adaptable to a multitude of situations.

1.1.1 Biological Networks

Networks are useful in a broad range of biological settings. We will focus on two specific types; functional brain networks and microarray data. Steps have already been made towards this end within network science, providing understanding of how structural and functional subsystems sit with the anatomy of the human brain [26, 90, 196, 185]. Networks have also been used to develop models for the brain - for example, network models can be tailored such that they replicate specific features of the real case in parameters such as density or clustering [102]. In understanding how networks must be connected in order to emulate the real case, inferences can be made about the possible real world mecahnisms, giving a unique perspective to a biological problem.

The nature of the biological networks we will study will become clearer with Chapter 2 which will provide an introduction to relevant biological background.

1.2 Notation

In this section we will introduce some key definitions and concepts used throughout this thesis. This will provide a brief outline to ensure consistency of these ideas when they are explored in later chapters.

1.2.1 The Adjacency Matrix

The creation of a matrix representation of networks is important in allowing for application of techniques from linear algebra to explore structure within the data. One convenient way to do this is with a so-called adjacency matrix. The creation of an adjacency matrix, A, of a network G involves setting the components $a_{ij} = 1$ where nodes i and j share an edge, and 0 otherwise. Figure 1.1 shows an example of a directed network (the edges have arrows indicating direction) and the corresponding adjacency matrix. The adjacency matrix is then a $(0, 1)$ matrix.

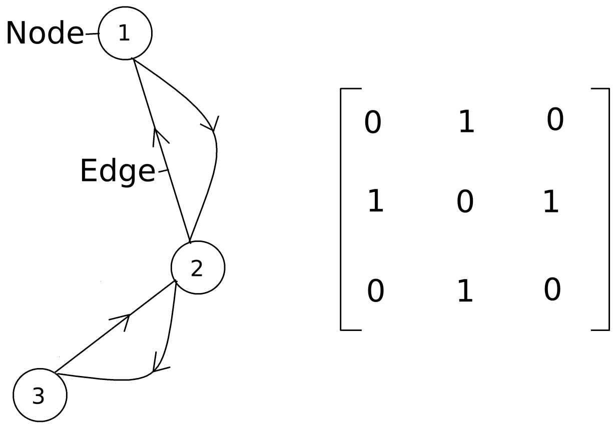

Figure 1.1: A network represented graphically (left) and the corresponding adjacency matrix (right).

1.2.2 The Graph Laplacian

There are a number of other matrices used to describe networks, depending on the structures we wish to explore. The Laplacian is one such matrix, related to the adjacency matrix, and is the difference between the degree, deg_i, (the degree of node i is the number of connections it has to other nodes) matrix and the adjacency matrix of the network. For an undirected network $A \in \mathbb{R}^{n \times n}$, the Laplacian, L, is defined as:

$$L = D - A \tag{1.1}$$

where $D = diag(deg_i)$. Note that the Laplacian has smallest eigenvalue 0 with $0 = \lambda_1 \leq \lambda_2 \leq .. \leq \lambda_n$ with corresponding eigenvectors labelled $v^{[1]}, v^{[2]}, v^{[3]}, .., v^{[n]}$ where we set $v^{[1]} \propto 1$. If the network is connected, so that every node can be reached from every other node by traversing edges, then $\lambda_2 > 0$ [57].

1.2.3 Fiedler Vector

For a connected network, the eigenvector $v^{[2]}$ is called the Fiedler vector of L. The Fiedler vector was first explored in [48] and developed by the eponymous Fiedler [61, 62] and is useful in spectral partitioning (the division of data into smaller components with specific properties) problems, as we will see in Chapter 6.

1.2.4 Normalised Laplacian

An extension of the graph Laplacian, another matrix of interest is:

$$\hat{L} = D^{-\frac{1}{2}}(D - A)D^{-\frac{1}{2}} \tag{1.2}$$

the *normalized graph Laplacian* of A. As with the Laplacian in Equation 1.1, this new matrix has a 0 eigenvalue. This time, the eigenvalues have range [0,2] and are labelled as $0 = \mu_1 < \mu_2 < .. \leq \mu_n \leq 2$. The corresponding eigenvectors are given labels $w^{[1]}, w^{[2]}, w^{[3]}, ...w^{[n]}$ see [85] for details.

This normalisation rescales the eigenvectors such that $w^{[1]} \propto D^{\frac{1}{2}}1$. As the term Fiedler vector refers to this particular unit vector $w^{[2]}$ of the Laplacian, the normalized Fielder vector is similarly used to describe the unit eigenvector corresponding to the second smallest eigenvalue of the normalized graph Laplacian [211]. That is, $D^{-\frac{1}{2}}w^{[2]}$ is the equivalent of the Fiedler vector - the *normalized Fiedler vector*. The normalized Laplacian then has the effect of reducing possible skew introduced by nodes with very large weights by rescaling according to $D^{-\frac{1}{2}}$.

1.3 Spectral Methods

Spectral graph theory is the study of matrix representations of networks or graphs, specifically through their eigenvectors and related quantities. In the way that chemistry can deduce the constituent components of a material through observing an energy spectrum (spectroscopy), spectral methods explore the principal properties of a network through the spectrum of eigenvalues. As we have seen, there are approaches to generating matrix representations of networks - we will now introduce methods for finding their spectra. We can say that, broadly, methods for spectral decomposition of matrices have the effect of giving a lower rank, same dimension, least squares estimate of the original.

1.3.1 Principal Component Analysis: PCA

Principal component analysis (PCA) is a method for dimensionality reduction that performs a linear mapping of data to a lower dimension in a way that the variance in the lower dimension is maximised - that is, the dimension is reduced while as much of the structure as possible within the data is retained. This property means that the eigenvectors corresponding to the largest principal components (eigenvalues) are good approximations of much of the variance in the original data set. As a result, PCA can be used to identify important/prominant features in data e.g., where the eigenvectors can be used for extraction of important features in image processing or image compression [5, 181, 47].

1.3.2 Singular Value Decomposition: SVD

The singular value decomposition (SVD) is a closely related generalisation of PCA, which is equivalent to applying the SVD to a covariance matrix of the data.

As an example, the SVD of a matrix $A \in \mathbb{R}^{m \times n}$ is as follows:

$$A = U \Sigma V^T$$

where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal and $\Sigma = diag(\sigma_1, \sigma_2, ..., \sigma_m)$ a diagonal matrix whose nonzero entries are the corresponding eigenvalues of A. The columns of U and V are referred to as the *left* and *right singular vectors* of A, respectively.

The SVD has a rich history as a fundamental of spectral clustering, and has seen use in many different types of data to identify patterns of connectivity between subsections of the data [159]. In forming a lower dimensional approximation, the SVD allows large amounts of information to be essentially compressed into a single summary that is easier to interpret and so is particularly useful in large data sets [36].

1.3.3 Generalized Singular Value Decomposition: GSVD

The generalized singular value decomposition (GSVD) is an extension of the SVD - decomposing a pair of rectangular matrices but allowing for the account of constraints on rows and columns. The result here is again an optimal lower rank, (weighted) least square estimate of the original [187]. The GSVD has been justified as an approach to computational network reordering - and is especially useful because it can be used to reorder two networks in such a way so as to identify mutually exclusive clusters. See [220] for details of the derivation and justification of the algorithm.

We provide a basic outline of the mathematics of the algorithm: if we have a matrix $A \in \mathbb{R}^{m \times n}$, $m > n$, and a matrix $B \in \mathbb{R}^{p \times n}$ then the GSVD takes the two as expressing the required constraints on rows and columns, and factorizes them to give:

$$A = UCX^{-1} \qquad\qquad B = VSX^{-1}$$

with $U \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{p \times p}$. $C \in \mathbb{R}^{m \times n}$ and $S \in \mathbb{R}^{p \times n}$ are nonnegative diagonal matrices, with $C \in diag(c_1, c_2, ...c_n)$ and $S \in diag(s_1, s_2, ...s_n)$. As with the SVD, U and V are orthogonal. $X^{-T} \in \mathbb{R}^{m \times m}$ is the invertible ordering matrix [73, 220]. The reason that X^{-T} is able to perform this function is that, by the nature of the GSVD, the columns of X^{-T} are stationary points of the function:

$$f : \mathbb{R}^n \mapsto \mathbb{R} \text{ where: } f(x) = \frac{||Ax||_2}{||Bx||_2}.$$

It is then argued in [219] (see Algorithm 2) that columns in X^{-T} highlight orderings for finding clusters in the original matrices A and B. The first columns of X^{-T} highlight orderings that are good for B and poor for A and vice-versa. This is particularly useful in areas where experimental comparisons are the ultimate goal, such as is often the case in the life sciences where test versus control experiments are common.

1.4 Reordering and Clustering

In large data sets there is often a desire to divide nodes within a network into groups, based on the patterns the corresponding set of edges forms. The problem of dividing a network into groups is often known as *graph partitioning*. Most commonly, nodes are grouped such that they are dense with edges, leaving a smaller number of edges to connect groups [166]. The problem of partitioning data is complex and there are many approaches, each suited to a different situation. In this thesis we are interested exploring matrix representations of networks with their SVD and GSVD. These methods fall under the term of spectral clustering, since the clustering is performed using the matrix spectra (singular values in these cases, though eigenvalue methods are also appropriate). Spectral clustering methods derive from graph cutting problems in spectral graph theory - in that context a cut is a partition of a set of nodes into disjoint sets. This problem involves the creation of a graph with weighted edges that measure the similarity between nodes.

Spectral clustering techniques are also heuristic in nature, providing a potentially non-optimal solution to the problem of discretely clustering data. The heuristic nature of the problem means that the method must validated for new types of data, hence much of the data we examine in this thesis is microarray gene expression where spectral methods are well established, for example, spectral approaches on microarray data has helped to guide cancer treatment by allowing for stratification of patients [113].

1.4.1 Cluster Verification

In this thesis we will also have to deal with the question of what it means for a data set to have a *good* cluster. This can be visually very difficult, particularly when the size of data sets increase. In Chapters 3 and 4 we will adopt an automated processing technique and quantify the statistical significance of a proposed cluster. This involves using some measure to determine the quality of a cluster - then testing the likelihood that our proposed cluster could have a particular score by chance.

We will use a variation of the ‘Cluster Validation’ approach as outlined in [220], including determination of a p-value in an attempt to place a level of significance on the result ($p < 0.05$). In brief, this method involves:

1. Calculate the density (mean degree) of the cluster of nodes in the reordered network, relative to the density of edges outside the cluster.
2. Randomize the order of the network.
3. Calculate the same relative density for a group of nodes the same size as the cluster in Step 1.
4. Repeat steps 2 and 3 999 times.
5. The frequency with which randomized relative density is calculated to be more than in our deliberately ordered network is the p-value: a probability that our result occurred by chance.

For an individual network A the cluster quality measure $c(A)$ is the ratio:

density of edges within the cluster in A / density of edges outside the cluster in A.

To calculate p-values we take the same measure for a randomly permuted version of the same matrix, $c(A_{rand})$ then the p-value is simply the proportion of $c(A_{rand}) > c(A)$. This is extended appropriately for use with the GSVD, with networks A and B, $c(A, B)$ is:

$$\frac{(\text{density of edges within the cluster in } A)/\ (\text{density of edges outside the cluster in } A)}{(\text{density of edges within the cluster in } B)/\ (\text{density of edges outside the cluster in } B)}.$$

Chapter 2

Biological Background

Reductionism has long been an important approach in the life sciences. The reductionist philosophy involves dissecting complicated systems in order to understand them through their constituent parts - and has had great success in understanding living processes through chemistry. Though we continue to learn from examination of individual components such as molecular structures and DNA sequences, there has been a recent shift in the paradigm towards understanding the systems level through complexity. In this context complexity involves interactions between entities - it is now understood that there is a limit to that which can be gained from examination of single biological components. Understanding the wider interplay between these elements is essential to understanding overall biological function and outcome - almost always the emergent properties are a result of interactions across the system rather than any one individual component [111, 116].

In recent years there has been rapid development and deployment of high-throughput technology across the life sciences. The vast outputs that these approaches yield has created an environment where the data are now available to tackle questions on the systems level. This advent of ‘big data’ has resulted in a surge in computer applications to address complexity within the biological sciences, leading to an explosion in the field of bioinformatics.

2.1 Microarray

One of the technologies central to this data revolution is the DNA microarray. A DNA microarray (or biochip) is a piece of experimental equipment that can be used to carry out multiple genetic tests in parallel. The key to microarray technology is the use of nucleic acid hybridization - this is done by so-called hybridization probes - fragments of DNA complimentary to that which the user wishes to detect. The target substance (sample) is complementary with some detectable molecular marker, such as a radioactive material or simply a molecule which fluoresces. The target is then exposed to the microarray probes and, after hybridization, non-hybridized sequences are washed off. Next, the DNA spot where the probes for a feature are located is scanned and the strength level of the molecular marker is noted at each probe. These response levels are then normalized for dynamic factors that affect hybridization conditions such as temperature.

There are different surface platforms to use when attaching the probes - for instance, the Illumina branded chip attaches the probes to a microscopic bead structure whereas the Affymetrix branded chip uses a glass or silicon base [1, 96].

In both surface platforms, the result is that microarray technology allows all features to be included on a small, convenient, plate. The decrease in size compared with traditional techniques also means that smaller amounts of biological material are necessary to yield results. This is good from research and diagnostics perspectives where smaller and less invasive approaches are more desirable - though the recent increased interest in microarray technology is mainly related to the ability to simultaneously quantify a vast number of sequences/probes. There are multiple uses for microarray kits - expression profiling being one of them, where others include: pathogen detection and characterization, comparative genome hybridization (CGH), genotyping and whole genome resequencing.

Gene Expression

Almost all cells in the human body contain the complete set of chromosomes and genes. The difference in anatomy and function across cell or tissue types and their response to stimuli or insult (e.g.,disease) is explained in part by differences in the transcription of these genes. The Central Dogma of molecular biology states that DNA can self-replicate (replication), the information within DNA can be transferred to create mRNA (transription) and mRNA can be used to create proteins in a process called translation. Gene expression describes the level at which, if at all, each gene is transcribed into the corresponding mRNA and from these onwards to the functional protein product in an individual.

Gene expression is then an intermediate step - residing between the gene sequence and phenotype and is controlled through a particular set of proteins known as Transcription Factors (TFs, which act the stage of transcription mentioned above), which can either up-regulate (increase) or down-regulate (decrease) expression levels according to function or need. These TFs activate or repress gene expression in response to both internal and external states or stimuli, with examples ranging from sleeping to disease. This process of transcription is then highly variable across individuals, and given the number of states and stimli that affect the outcome, is also highly dynamic for a particular person.

2.1.1 Approaches to Microarray Expression Analysis

As a result of the relationship between gene expression and disease states, microarray analysis is topical and timely - microarrays present a data driven approach to understanding changes in gene expression and disease states. The typical data output of a microarray gene expression study is a data set of $O(10^3)$ genes with $O(10)$ samples. The set of genes that can be measured in an experiment covers virtually all genes in the genome, which gives a close-to-complete context to the nature of the phenotype being studied.

The output from traditional microarray analysis comes in the form of lists of genes and their associated expression levels - the difficulty with which, due to the large number of genes involved, lies in the biological interpretation. In order to address these problems, there have been some attempts to integrate information a priori in order to focus the analysis and provide a more specific direction for interpretation. Examples include a co-clustering approach [80] and network component analysis [129, 67]. Clustering approaches are relevant to this thesis, where we use spectral clustering approaches on biological data. There are two commonly used types of clustering techniques used in the life sciences - hierarchical and non-hierarchical methods.

Hierarchical methods are either agglomerative (where a cluster is constructed in a bottom-up fashion - initially each node is considered as a separate cluster and progressively larger groups are formed from there) or divisive (which does the opposite - starting with all nodes in a single cluster and creating progressively smaller groups). Hierarchical methods do not result in any definitive clusters however, and no objective function is minimized [101].

Non-hierarchical methods include the k-means method where, initially, nodes are arbitrarily assigned to any one of a predetermined k clusters [81]. The number of clusters can be guessed, randomized, or implemented based on some previously performed hierarchical method. The procedure is then iterative and involves calculation of a centroid of each cluster, and re-assignment of the nodes based on those results. Although k-means methods scale well for large data, they will not necessarily converge [19]. As well as the difficulty in choosing a predefined number of clusters, k-means methods suffer because they tend to generate clusters of similar size - since nodes are grouped towards the nearest centroid.

One issue with these classical clustering techniques is that they require genes within a single cluster to have the same response over all samples in an experimental group. This is not the case with spectral methods, which can accommodate a multitude of responses across groups of samples. Applying clustering methods to large gene lists generates a selection of smaller clustered lists that, depending on the clustering approach, may be related to interesting features of the data.

The automatic translation of a gene list into a specific biological interpretation is very challenging. As such, although there are a variety of tools, there are many limiting factors - for instance there is a circular effect in utilization of knowledge based approaches: the knowledge must already be present in some fashion for it to appear in a tool. Manual interpretation of microarray experiments is a very time consuming task, however with a list of thousands of genes, there may be hundreds that are significantly clustered. Often the relation between these clustered genes and the experimental parameters under examination is not clear, thus there is an innate subjectivity in the analysis.

2.1.2 Gene Ontology Project

Gene Ontology (GO) is a bioinformatics project (`http://www.geneontology.org/`) that attempts to provide summary biological interpretation for large genetic analysis project. In particular, GO brings together numerous identification and description methods used for genes and gene products [8], aiming to provide tools that can annotate the genes and their products in a consistent way. There are three different elements covered in the GO project:

1. Cellular component
2. Molecular function
3. Biological process

Cellular component refers to parts of a cell or parts of the extracellular environment (e.g.,membranes, proteins, nucleic acid). Molecular function covers function of the gene product at a molecular level, e.g.,protein binding or enzyme catalysis. Biological process describes molecular processes that can be mapped with definitive start and end points. The biological process level covers chemical reactions within an integrated system, e.g.,cells, tissues.

In this vein of thinking, it is possible that a gene product may have a seemingly insignificant level of differential expression but be highly relevant in the disease state when considered as part of an interaction network. Overrepresentation analysis is one approach where genes are selected based on some criteria and assigned $p-$values. There is also functional class scoring which uses either $p-$values from a t-test, analysis of variance (e.g.,ANOVA) or fold change (a ratio of means).

GOrilla

We will use two of the many GO tools that have been created to date. The first we will see is Gene Ontology enRIchment anaLysis and visuaLisAtion (GORILLA:-`http://cbl-gorilla.cs.technion.ac.il/`) [52, 51]. GORILLA is an enrichment tool that accepts ranked lists of genes and generates a directed acyclic graph such as in Figure 2.1, illustrating steps in any or all of the three GO elements mentioned above.

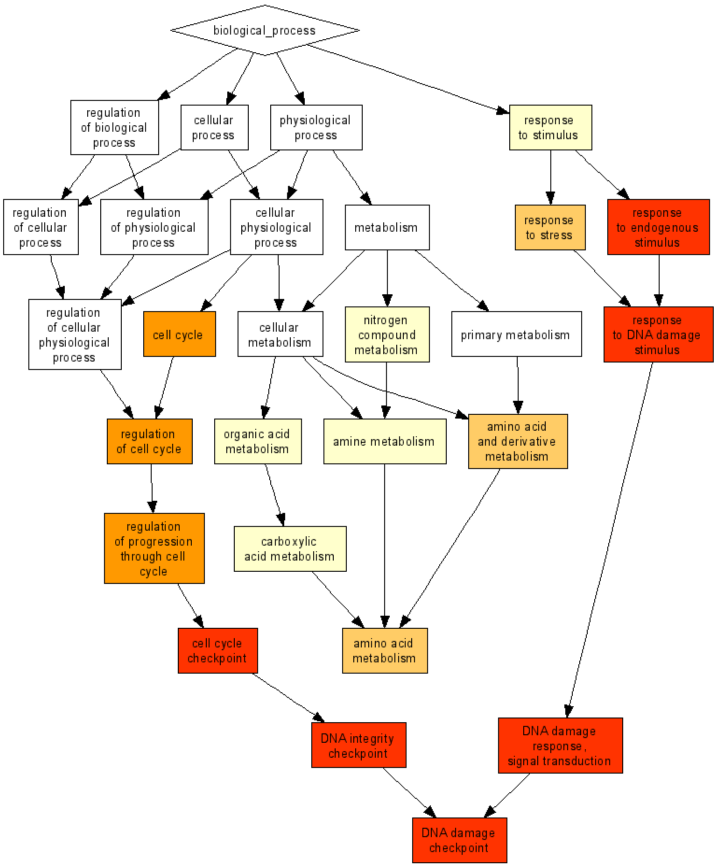
biological_process
regulation of biological process
cellular process
physiological process
response to stimulus
regulation of cellular process
regulation of physiological process
cellular physiological process
metabolism
response to stress
response to endogenous stimulus
regulation of cellular physiological process
cell cycle
cellular metabolism
nitrogen compound metabolism
primary metabolism
response to DNA damage stimulus
regulation of cell cycle
organic acid metabolism
amine metabolism
amino acid and derivative metabolism
regulation of progression through cell cycle
carboxylic acid metabolism
cell cycle checkpoint
amino acid metabolism
DNA integrity checkpoint
DNA damage response, signal transduction
DNA damage checkpoint

Figure 2.1: Example of Gene Ontology tool GOrilla output. This illustrates a biological process tree where each lower level tier is an increase in specificity. Colour indicates p-value for enrichment in a sub-list of genes for a specific ontology term, relative to the complete list of genes in the experiment.

2.1.3 REVIGO

The GO project is still in early stages - efforts are scattered and some terms are loosely defined, this is particularly difficult with the arbitrary cutoffs for tiers in the ontology tree. REduce and VIsualise Gene Ontology (REVIGO:-`www.revigo.irb.hr`) [200] is an effort to find overlap between GO terms, allowing for the identification and removal of possible redundancies. For instance, the example Figure 2.1 from GOrilla shows repeat of terms such as "cell cycle", "regulation of cell cycle", "regulation of progression through cell cycle" and "cell cycle checkpoint" - in generating a large list of ontology terms this pattern of repetition can lead to an unmanageable amount of information to process. In cases of exploratory analysis we may be satisfied with a broad representation of ontology characteristics, such as representing the previous example as "regulation of cell cycle".

In addition, REVIGO aims to simplify, and increase the specificity of, GO terms in cases where potentially uninformative umbrella terms show enrichment due to enrichment of their lower tier constituents. This is seen in Figure 2.1 where "cell cycle" is enriched, but is a general, non-specific term. The enrichment is as a result of more significant enrichment of component terms such as "regulation of progression through the cell cycle". The balance between breadth and depth is important in the usage of GO terms.

2.2 Schizophrenia

This thesis will focus on investigation of schizophrenia using the previously mentioned clustering and ontology methods, among others, in an attempt to validate animal models and explore stratification of human patients. Schizophrenia is a debilitating psychiatric disorder that affects around 1% of the global population, ranking as the third most disabling condition (behind quadriplegia and dementia) [209]. A high impact condition, schizophrenia often appears in late adolescence and can cause impairment of social interaction, executive function, creative thought and emotive expression, making it difficult to complete education and often disrupting early career progression. As a result of these extraneous factors, on top of the costs of treatment to manage symptoms, in 2007 alone there was an estimated societal cost of £6.7bn to England [140] - with a total cost of brain disorders in the UK during 2010 at £53bn. On a wider scale, brain disorders are now the largest contributor to 'all cause morbidity' in the EU and the Global Burden of Disease studies has found that the global economic and health burden attributed to brain disorders is likely to continue to increase in the future [137].

Schizophrenia is also highly complex, spanning multiple psychological domains with a high level of heterogeneity in presentation. Key symptoms involve hallucinations and delusions. Owing to the complex nature of the condition, there is also a huge diversity in investigation - e.g.,psychiatric, genetic, and epidemiological study. Genetic and environmental factors impact on brain development which can lead to abnormalities in the formation of brain networks. These abnormalities from both genetic and environmental sources may then ultimately combine to produce the overall abnormal behaviour [31]. Despite this understanding we are still presented with a limited knowledge of specifically how the disease affects brain connectivity [91, 204, 136, 15]. The network science approach to data analysis has already made significant steps towards remedying this, with demonstrations that patients exhibit disruptions in small-world network properties of the brain [136]. These small-world properties are important in terms of information transfer and functional integration of different areas of the brain, and disruption of those factors in functional brain imaging is consistent with other studies that have found dysfunctional integration of the brain in schizophrenia [128, 152].

2.2.1 Diagnosis

In psychiatry, diagnosis is always based on symptom rather than results from repeatable, objective biological tests. Symptoms come in three classes and are described as positive (in addition to normal experience), negative (missing compared to normal experience) or as deficits in cognitive function. Positive symptoms are those that exist in addition to individuals from the general population. Common examples include hallucination (auditory, visual, olfactory and tactile), delusion, interruption in logical processing and disorganised speech. In contrast, negative symptoms describe reductions in experience compared with the general population. Negative symptoms include anhedonia (reduced capability or inability to experience pleasure), avolition (lack of motivation) and alogia (reduced fluency of thought or speech). The final category of symptoms, a long list of cognitive dysfunctions, have been noted in schizophrenia, with many causing difficulties in every day living, such as problems with social integration, working memory, attention and executive function. Due to the variety in the presentation of the symptomatology of schizophrenia clinical diagnosis is based on matching symptoms with an agreed clinical standard. This varies across medical bodies, with European countries commonly adopting the ICD-10 criteria and the United States the DSM-V critera, though there is broad agreement between standards [98]. There have also been attempts to differentiate between subtypes of the condition (e.g., disorganised, catatonic, paranoid) by grouping patients based on the similarity of their symptoms. One issue with this approach is that symptoms are dynamic, varying over time, categorisation is also reliant on accurate self-reporting which is difficult with an illness that often accompanies lack of motivation and other cognitive deficits. To alleviate these issues a consistent, repeatable biological test would be beneficial to provide objectivity in diagnosis as well as to explore potential benefits to division of patient groups.

Currently, the main antipsychotics only treat the positive symptoms associated with the disease, having little to no effect on negative symptoms or cognitive deficits. Cognitive enhancers are being explored as a possible treatment modality to address this but none are currently in standard use.

2.2.2 Treatment

There are two generations of antipsychotics used in the treatment of schizophrenia [126]. Antipsychotics such as chlorpromazine and haloperidol act as dopamine receptor antagonists. The site of action for these drugs is commonly at the dopamine D_2 receptor subtype, acting in the mesolimbic and nigralstriatal dopamine pathways [35]. Unfortunately, the same mechanism of dopamine blockade that significantly alleviates positive symptoms also results in a new class of side-effects - movement disorders. Specifically, nigralstriatal dopamine blockade induces movement issues resembling Parkinson's disease (extrapyramidal side effects - in this case the inability to remain motionless) [179].

The second generation of antipsychotic drugs were developed as a response to this difficult profile of side effects, including sedation and hypotension. This next generation of so-called atypical antipsychotics, or AAPs, are used in the treatment of schizophrenia as well as other conditions such as mania, bipolar disorder and psychotic agitation, targetting a wider range of receptors than just dopamine. The result is that these drugs often have similar results to the first generation neuroleptics with fewer extrapyramidal side effects but this comes at the cost of introducing new issues such as weight gain and potential links to diabetes. There are other successes here however, with drugs such as clozapine successful in many treatment resistant cases.

Recent evidence has shown that distinguishing the drugs as two generations could be misleading to patients - treatment success may not have improved with the second generation, and claims to a reduction in side effects are exaggerated [208, 126, 131]. In any case there are multiple avenues of investigation required in the treatment of schizophrenia. For one, there is a clear need for general improvement and individualisation in antipsychotic pharmacology - currently, side effects are severe to the point that a large part of the psychiatric process is to minimise side effects.

In particular, both generations of antipsychotics have a limited ability to treat cognitive deficits or negative symptoms. There are very few successes in the development of drugs to address the negative and cognitive aspects of the disease. Recent initiatives have included the Measurement and Treatment Research to Improve Cognition in Schizophrenics (MATRICS) [74], CATIE Project (Clinical Antipsychotic Trials of Intervention Effectiveness) [109] and Cognitive Neuroscience Treatment Research to Improve Cognition in Schizophrenia (CNTRICS) [29]. Secondly, the action and relationship between the therapeutic effect of antipsychotic medication and the underlying biology of the condition require further investigation. Understanding specific effects of antipsychotics is important as it is unclear if medication masks or eases symptoms in a way that is related to the underlying cause of the condition.

2.2.3 Genetic Factors

Schizophrenia is a highly heritable condition with estimates of approximately 80% [130]. This suggests that genetic factors play a key role in the risk of developing the condition. Twin studies have also shown incidence rates of 41-65% in monozygotic cases and 0-28% in dizygotic cases [28], and siblings of affected individuals have a 10-fold increase in risk [127]. The complex nature of the schizophrenic phenotype is mirrored in the complexity of the aetiology - there are both genetic and environmental factors with many genes implicated.

2.2.4 Biomarkers

The discovery of an indicator biomarker is therefore sought after for many reasons. In pharmacological terms, biomarkers may allow for personalisation of treatment paradigms and, in broader terms, there is potential for predictive biomarkers for early diagnosis or pre-onset intervention. It is possible that biological indicators are dynamic and could map course of illness, treatment efficacy, and relationship with side effects - leading to an individualisation of care. In addition, biomarkers present the opportunity to understand disease mechanisms and so develop new drug targets or diagnostic assays. This diagnostic element is important for schizophrenia, particularly considering the aforementioned issues associated with subjective diagnostics in psychiatry. Unfortunately the specific biological understanding of many psychiatric conditions is limited - which decreases the likelihood of development of a viable biological marker. In psychiatric terms until recently in the DSM there were five sub-classifications of schizophrenia:- paranoid type, disorganised, residual, undifferentiated and catatonic. These sub-types were based solely on presentation of symptoms, in the hope that common symptom profiles would be informative guiding the treatment regime of newly diagnosed patients. However, with the release of the DSM-V (2013) the recommendation of specifying a schizophrenia sub-type has been removed - the conclusion being that psychiatric sub-types have low diagnostic stability, low reliability and poor validity [9]. This highlights the need for an objective biomarker based approach that is not reliant on subjective assessment of symptoms.

Whilst there is still a lack of understanding of the underlying aetiology of schizophrenia, there is still potential to divide schizophrenia patients into smaller, more specific groups - so-called ‘patient stratification’. Once identified, potentially distinct subgroups can be informative in guiding treatment and making a prognosis - as was the case with stratification of breast cancer patients based on the *HER2* gene [89].

2.2.5 Animal Models

Genes can be implicated when investigating risk factors for schizophrenia. However, to prove a definitive link and understand the underlying disease mechanisms, biological evidence and understanding is required. Rodent models are one of the most effective ways of exploring this higher level element, and have a valuable role in treatment development through verifying and suggesting areas for drug targeting [180]. There are a number of different types of animal models of relevance to schizophrenia - created through different means, with different goals in mind [54, 132]. Because of the complexity of schizophrenia, and the fact that there is no animal equivalent to the human diagnostic process which is based on self-reporting of symptoms, development of a coherent and comprehensive animal model is a major challenge. Hence, animal models tend to focus on a subset of symptoms of the disorder e.g., behavioural, genetic or neurological. These models are then used with an understanding that some elements of the phenotypic presentation will be model specific.

Pharmacological models are a focus of this thesis, where we explore two different drug models. Pharmacological models are those where a pharmacological agent is administered, often in an attempt to mimic altered neurotransmitter function as seen in an individual with schizophrenia. There is strong evidence implicating N-methyl-D-aspartic (NMDA) acid receptor hypofunction as having a role in schizophrenia [164]. This hypothesis was suggested, and supported, by multiple studies showing that repeated dosing of NMDA receptor antagonists causes a symptom profile, in healthy humans, that induces a variety of positive, negative and cognitive deficit symptoms, and intensifies symptoms in patients with schizophrenia [33, 122, 188, 121]. Additionally, acute exposure to ketamine (an NMDA receptor antagonist) increases the severity of symptoms in schizophrenia patients [122]. As a result, animal models created through dosing of an NMDA receptor antagonist such as phencyclidine (PCP) or ketamine, are an approach that is often utilised to model translationally relevant symptom profiles in schizophrenia [32, 158]. As a result of the success of PCP in mimicing the symptoms of schizophrenia, the glutamate hypothesis of schizophrenia was proposed as a mechanism for psychotic episodes [149].

There are animal models for many other aspects of the disorder, including pharmacological models to explore the dopamine hypothesis [147], serotonin association [215] and GABAergic system involvement [21, 179]. Models are also created through the introduction of brain lesions, toxins or other insults to explore neurodevelopmental factors [134, 135], though we will not consider any such models in this thesis.

There are also genetic animal models that can be used to investigate results from gene association studies: specific genes may be associated or otherwise implicated in a disease state but uncovering the role a particular gene plays in the overall biology is often complicated [2].

2.2.6 Summary

This chapter has provided an overview of gene expression and microarray technology, and the conditon that will be the focus of this thesis, schizophrenia. The scene has been set in terms of highlighting some of the difficulties that patients and their clinicians face in dealing with and treating the condition. The main points are that microarray data can be clustered (using methods seen in Chapter 1) in a way that provides biologically relevant results, related to differences between disease and control samples and that animal models can be informative in exploring underlying aetiology of the disease state in humans.

2.3 Outline of Thesis

Schizophrenia is a complex neurological disorder. In this thesis we explore different approaches to understanding the nature of the condition through application of a variety of spectral clustering based techniques. The opening chapters, Chapters 3 and 4, introduce data from preclinical animal model studies of relevance to schizophrenia. We examine differences between model and control animals using spectral clustering methods (the GSVD in particular), and introduce a new approach to quantify differences in structure between sample groups. This novel variation in spectral clustering allows us to explore the relevance of animal models to the human state of disease.

Next, in Chapter 5 we see the construction of two new networks representing interactions in human metabolism. These networks are formed based on shared involvement of metabolites in defined metabolic pathways and as such are a variation on more common metabolic networks which represent interaction between metabolites. These two new structures are each formed from a different database of human metabolism, and were developed as part of the aims of the following Chapter 6. In this chapter we develop a variant, named the node-weighted Laplacian, of existing spectral methods that can be used to merge information between two networks with a common set of nodes. This novel method is tested with synthetic data and the principle is illustrated on real data, framed for use to explore metabolic involvement of schizophrenia in human data in later chapters.

The study of schizophrenia then moves from preclinical animal data to microarray data taken from samples of human whole blood in Chapter 7. In this chapter we cluster gene expression measurements in order to investigate the potential of blood as a tissue for the study of schizophrenia, in place of the far more commonly used, but much more difficult to obtain, brain tissue. Since it is suspected that antipsychotic treatment may have a significant role in altering gene expression levels in patients, this chapter is also used as an opportunity to study differences in expression between medicated and antipsychotic-free patients. We also apply the previously developed node-weighted Laplacian in combination with metabolic networks from Chapter 5 to test the approach on more real data and explore metabolic involvement in schizophrenia.

The promising approach from the previous chapter is tested in Chapter 8, where a second data set with an antipsychotic-free cohort is subject to the same approach and examined for consistency across experiments. In addition to the previous analysis at the gene-level, in this chapter results are generated at the function and process levels with Gene Ontology terms, which illustrates broad differences in results between the SVD and GSVD.

Schizophrenia is a disorder in the brain, and so for the final chapter of work the same methods as in the preceeding chapters are applied to microarray human brain data. Chapter 9 shows these results, giving context to the earlier studies of data from whole blood.

Finally, conclusions are made in Chapter 10 where the context of this work within the greater field is discussed, giving opportunity for the examination of potential improvements and future work.

2.4 Publications and Presentations

Much of the material presented in Chapter 3 was presented in a poster at Scottish Neuroscience Group conference (2010), given as a talk at the 2011 Biennial Conference on Numerical Analysis and appears in the article

- Dawson N, McDonald M, Higham DJ, Morris BJ, Pratt JA. Subanesthetic Ketamine Treatment Promotes Abnormal Interactions between Neural Subsystems and Alters the Properties of Functional Brain Networks, *Neuropsychopharmacology*, 2014 [39].

A selection of the work related to modafinil in Chapter 4 appears as part of the article

- Dawson N, Xiao X, McDonald M, Higham DJ, Morris BJ, Pratt, JA. Sustained NMDA receptor hypofunction induces compromised neural systems integration and schizophrenia-like alterations in functional brain networks, Cerebral Cortex, 2012 [42].

Much of the work from Chapter 5 and all of the material in Chapter 6 combine to form the special issue article

- McDonald M, Higham DJ, Vass JK. Spectral algorithms for heterogeneous biological networks, Briefings in functional genomics, Vol. 11 No. 6, 2012 [146]

Chapter 3

Neural Subsystems to Interrogate Ketamine as a Translational Model for Schizophrenia

The work presented in this chapter forms part of a collaborative publication:-

Dawson N, McDonald M, Higham DJ, Morris BJ, Pratt JA. Subanesthetic Ketamine Treatment Promotes Abnormal Interactions between Neural Subsystems and Alters the Properties of Functional Brain Networks, *Neuropsychopharmacology*, 2014 [39].

This chapter presents the material from that manuscript concerned with the use of spectral methods to analyse the data.

3.1 Background and the Ketamine Model

An important development in the growth of the field of systems biology is the application of network theory to biology [57, 195, 26]. This chapter aims to add to this by providing a quantitative assessment of acute ketamine exposure on brain imaging data, using network based methods. This study is carried out in mice with the goal of understanding the effects of ketamine on functional connectivity of neural subsystems. As mentioned in Chapter 2, NMDA receptor blockade produces behavioural alterations and a symptomological profile that resembles schizophrenia. Though there are phenotypic similarities, the mechanism and the link between NMDA receptor blockade and this end result have not yet been adequately explored. The NMDA receptor is a subtype of glutamate receptors - these receptors recognise the excitatory neurotransmitter glutamate.

Complex network-based approaches provide an opportunity to uncover differences in network clustering between the ketamine model and control animals, making inferences about the action of ketamine. We will take lead from strong interest recently in the literature in characterising interactions between distinct neural subsystems in the context of complex networks, including some works aimed at elucidating the aetiology of the ketamine model [168, 40]. A particular point of interest with regards to the model for schizophrenia is functional integration of the prefrontal regions of the brain - to test the so-called ‘hypofrontality hypothesis’ - a decrease of activity in the prefrontal regions of the brain. The hypofrontality hypothesis has been supported by many brain imaging studies on patients with schizophrenia [4]. This is consistent with negative and cognitive symptoms in the disease state, since the prefrontal cortex is known to have involvement in complex cognitive behaviours, decision making and moderating behaviour [221].

3.1.1 Data: 2-Deoxyglucose (2-DG) Autoradiography and Suitability for Network Science

Chapters 3 and 4 take a network approach to interpreting 2-deoxyglucose (2-DG) autoradiography measurements in mice. 2-DG imaging provides a measure of localised regional metabolism (glucose utilisation) which is directly correlated to neuronal activity [115, 194]. The basis for this protocol is that brain cells metabolise glucose as an energy source - as glucose uptake occurs so does 2-DG, but 2-DG remains detectable in the cells afterwards as 2-DG phosphate since it is not completely metabolised. In brief, this process involves injecting the mice with an isotope ^{14}C 2-DG. They are left for 45 minutes to allow the isotope to distribute before being sacrificed. The brains are dissected out and divided according to some pre-determined mapping of areas of desired measurement. After freezing and coronal sectioning, autoradiograms are taken of each individual brain segment. Optical intensity of silver graph deposition on X-ray film corresponds to recent cerebral glucose utilization in that region, which can be expressed with respect to the average level throughout the whole brain of the animal [40]. Thus the process of 2-DG autoradiography provides an index of regional cerebral metabolism over a time period, similar to techniques such as fludeoxyglucose positron emission tomography (FDG PET) which has yielded significant results in human schizophrenia studies [23, 192] as well as other neurological conditions such as Huntington's disease [169, 60].

If a functional interaction exists between measured brain regions across time, it is supposed that this functional interaction can be detected by observing correlated changes in metabolism between these regions across subjects or states - one of our objectives is to test whether this assumption leads to reasonable conclusions.

Interpreting these results in a network sense is appropriate because the brain does not operate as a collection of individual elements, but as a complex, dynamic network where functional interaction between regions is highly important in determining neural operation [26]. In humans, these functional interactions between regions have previously been elucidated through functional brain imaging methods such as fMRI, where the correlated activity of brain regions across time supports a functional interaction between regions [15, 26, 10, 97].

Given that strong functional interactions are more likely to exist between anatomically connected neural systems, a major focus of this chapter is the determination of differences in clustering of functional neural subsystems between a control and an experimental group treated with ketamine.

3.1.2 Introducing the Data

The data for this chapter are previously unpublished and were provided through a collaboration with Dr Neil Dawson, who carried out all of the 2-DG data gathering process (Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde) [39]. This 2-DG study involved two groups of mice - one acutely treated with NMDA receptor antagonist ketamine (n=9, $30mgkg^{-1}$, intraperitoneally [i.p.]), the other being a control set given physiological saline (n=9, $2mgkg^{-1}$, [i.p.]). This experiment followed a published protocol [41]. ^{14}C 2-DG uptake ratios were measured for 66 regions of interest across the mouse brain. After gathering the data, the inter-regional Pearson's correlation coefficients (partial correlation) were taken as a metric of functional association between regions of interest, as has been done with previous studies [40, 42]. The data was also Fisher z-transformed to increase normality.

We were not involved in this laboratory collection phase, instead we received the data in Pearson's r/Fisher z processed form as two square, symmetric arrays. We have $A \in \mathbb{R}^{66\times 66}$ from the ketamine treated animals ($n = 9$) and, $B \in \mathbb{R}^{66\times 66}$ from the control animals ($n = 9$). The arrays have have a common set $G = \{g_1, g_2, ...g_{66}\}$ of nodes representing brain regions. The array entries a_{ij} and b_{ij} represent the edge weight between regions i and j. More precisely, these edge weights correspond to the normalised partial correlation co-efficients, across mice.

The cross-correlations are two-signed and real valued; a_{ij} or $b_{ij} > 0$ means the profile 2-DG levels for regions i and j agree across mice, and a_{ij} or $b_{ij} < 0$ means there is an anti-correlation. Additionally, $a_{ii} = 0$ and $b_{ii} = 0$. Note that both real-valued weights and directed graphs have previously been studied in brain networks [36, 102]. From a networks perspective, this can be re-stated by saying that there are both positive and negative edge weights between nodes and the zero diagonals indicate that there are no self-loops. The existence of both positive and negative edge weights has the consequence in spectral analysis that the first singular vector becomes relevant - this is in contrast to common spectral analysis of networks where edges are assumed to be non-negative and the Fiedler vector (2nd) and above are relevant [87]. Often in analysis of brain networks a threshold is applied to the matrix of weighted pairwise associations between regions. Though this thresholding approach allows for the formation of a binary adjacency matrix which simplifies the process of analysis, the sparsity, among other parameters, depends on the threshold value. This means that important structure and information can be lost with a poor choice of threshold [26]. The clustering approach taken in this chapter allows for the use of real values, and so the full weighted network can be used and the thresholding problem avoided.

Figure 3.1 shows initial plots of the data, along with a colour bar. This illustrates the heat map type format that will be used at various stages of this analysis, where values are represented as colours. This format is widely used in the life sciences, where visual assessment of clustering methods is beneficial [50, 217]. There are no obvious visual features or patterns (e.g., collections of similarly weighted/coloured subsections) in the data at this point, which makes sense since the ordering in which the brain regions are presented is arbitrary.

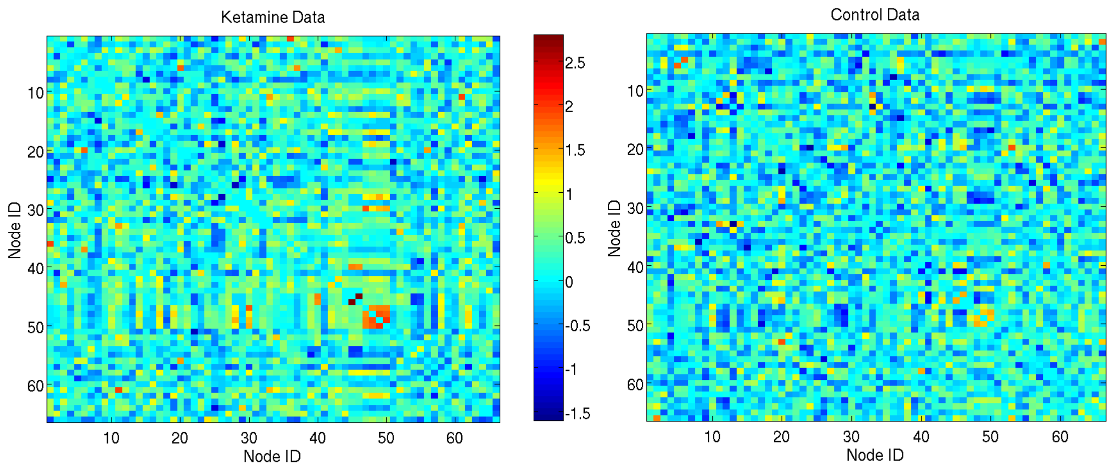

Figure 3.1: The ketamine and control data with their initial arbitrary orderings. The X/Y dimensions are node IDs, which correspond to brain regions. These node IDs are initially assigned to nodes based on the order in which the experimenter collected the data.

Allocating Brain Regions to Functional Groups

The connections in a brain network can take various forms - e.g., functional networks where edges represent some measure of activity (e.g., fMRI data) or anatomical networks, where edges are formed from known structural connections between areas of the brain [210, 65]. In this chapter, the nodes are anatomical and edges represent function. Since the goal of this experiment was to uncover functional differences between the two sets, it is useful to assign each member of the node set G to a subset depending on known biology.

Each of the brain regions were allocated to a discrete neural system on the basis of established neuroanatomy, connectivity or known functional similarities. While the neuroanatomy of some of these functional subsystems is established in the literature [88, 40], others are less well defined or understood. For example, the connectivity of the thalamic regions of the brain is so wide-ranging and complicated that assigning these regions to subsystems based on connectivity is not currently possible. Thus, thalamic nuclei have been designated a thalamic subsystem, due to their functional similarities and high anatomical interconnectivity [139]. Similarly, we group regions of the prefrontal cortex and hippocampus on the basis of functional similarity and known anatomical connecitivty. The neuromodulatory group is comprised of neurotransmitter systems that regulate neuronal activity in many neural subsystems (e.g., serotonin/noradrenaline) and the multimodal group is a series of brain regions that interact with many different neural systems. The complete set of functional groups, provided by Dr Neil Dawson, is shown in Table 3.1.

Functional Group	Number of Regions
Thalamus	11
Hippocampus	6
Prefrontal Cortex	6
medial Prefrontal Cortex	4
Cortex	6
Mesolimbic	4
Amygdala	3
Septum/DB	4
Basal Ganglia	6
Neuromodulatory	6
Multimodal	10

Table 3.1: Functional groups of 66 regions of interest

To recap, in this subsection an experimental paradigm has been described, and detail provided of an experiment using this protocol. Since the purpose of this experiment is to investigate an animal model of relevance to schizophrenia, as outlined in Section 2.2.5, our aim is to find differences in function between the control state mouse brain and the acutely treated ketamine model mouse brain. The suggestion is then that the observable differences between these two sets of mice is related to the aetiology of schizophrenia. The first objective of this analysis is to evaluate the use of acute administration of ketamine as an animal model for schizophrenia, that is, to explore whether ketamine disrupts functional integration of neural operation in a similar manner to that which is known in humans. The second, complimentary, objective is to show any other differences between the two groups, exploring the possible role of NMDA receptor hypofunction in schizophrenia.

3.2 Method of Analysis

The methods taking focus in this thesis are spectral clustering approaches - use of the SVD and the GSVD. These spectral approaches are heuristic in nature, so it is important to validate results on novel data types. To this end, we employ a cluster verification technique (Section 3.2.1) and compare results with other calculations on this data, such as small-world and other traditional network parameters (e.g., degree, path length - as performed by Dr Neil Dawson in his own assessment of the data).

As described in Section 1.3.3, the GSVD can identify mutually exclusive structures between a pair of networks that share a common node set. In the notation of Section 1.3.3, taking the GSVD of (A, B) the first few columns of X^{-T} are expressive of the structures in the control (B) data, and the last few columns of X^{-T} show the structures of the ketamine (A) data - as according to [219].

These columns are then sorted and used for reordering and clustering of the original data. So, node i in the original data is assigned a new ID as node k and the overall reordering can be represented by a permutation of the integers 1 to 66. To visualise the clustering process, Figure 3.2 shows the networks (from Figure 3.1) reordered using vectors from the GSVD, with potential clusters highlighted in red: the clusters can be identified as areas of heat (orange/red in this case), and appear in the expected areas given the mutually exclusive clustering ability of the GSVD.

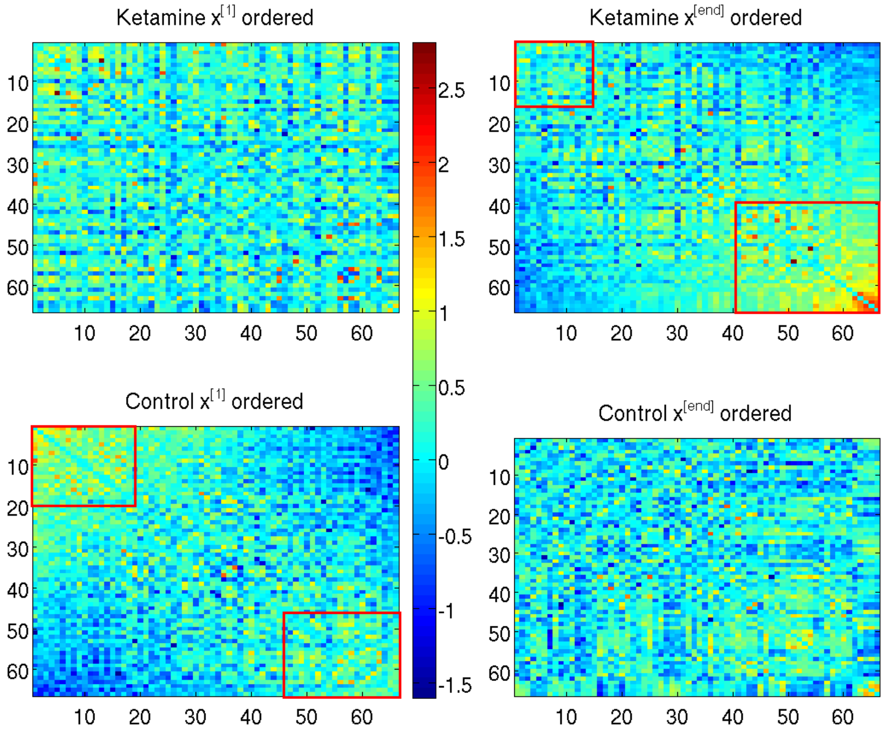

Figure 3.2: Heat maps showing GSVD reordered data. Clusters, as highlighted in red, can be seen as concentrated areas of heat in the end vector for ketamine data and start for control.

In this case clusters are seen in the control data in the first column $x^{[1]}$ and with the last vector $x^{[end]}$ in the ketamine data. Then, according to this interpretation of the GSVD, the nodes pushed into these clusters are those driving structure within the data. The next step is to quantify the significance of the clusters.

3.2.1 Cluster Quality Measure

Though visual assessment is useful, and often convincing, we will utilise a previously developed approach (see 'c_2' in Chapter 3 [219] and [220], and Section 1.4.1 in this thesis) to give a quantitative estimate on the quality of a cluster. There are numerous possible definitions of a cluster, in this method the cluster is defined in terms of relative density - an area with a density (mean value) higher than that in the rest of the network is a cluster.

Figure 3.3 illustrates how cluster quality is calculated according to this measure. The ratio of the density in area 1 ($x \times x$) of network A and the density of the rest of the edges in network A, is compared with the same ratio in network B - with both networks having been reordered by a vector from the GSVD. This value is assigned a bootstrap style p-value, where the p-value represents the frequency with which a ratio of this level, or higher, occurs in 10,000 random permutations of the same networks.

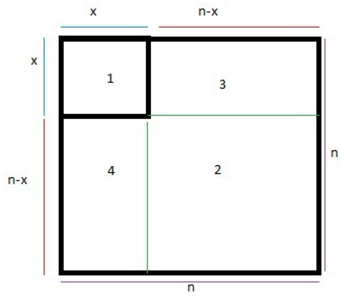

Figure 3.3: A sample network with a candidate cluster highlighted as area (1)

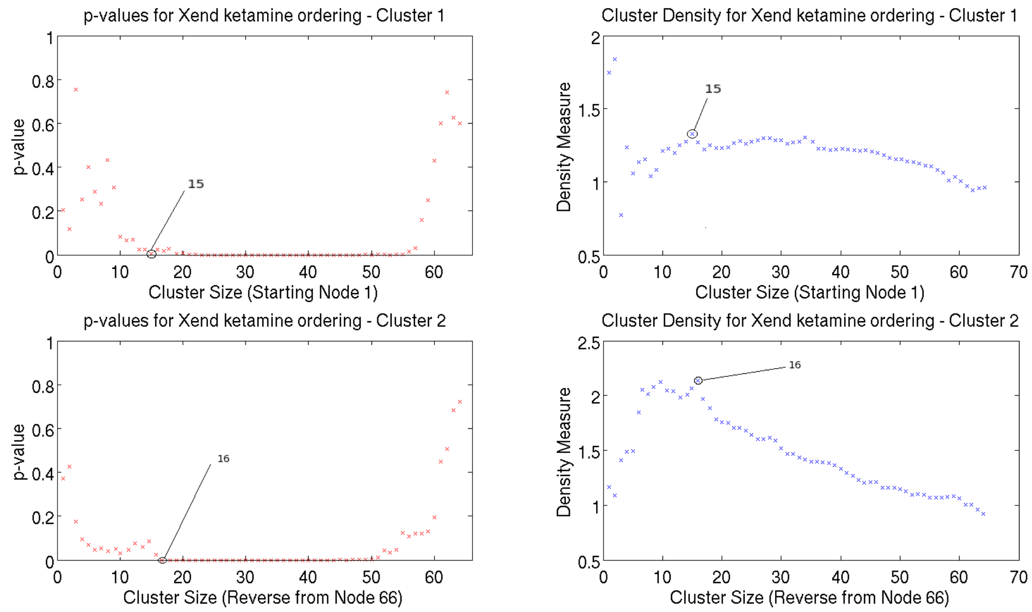

Figure 3.4: p-value and relative density measure for varying cluster size. Top images show values for a cluster starting at the first node, while the bottom images show those starting at the end node.

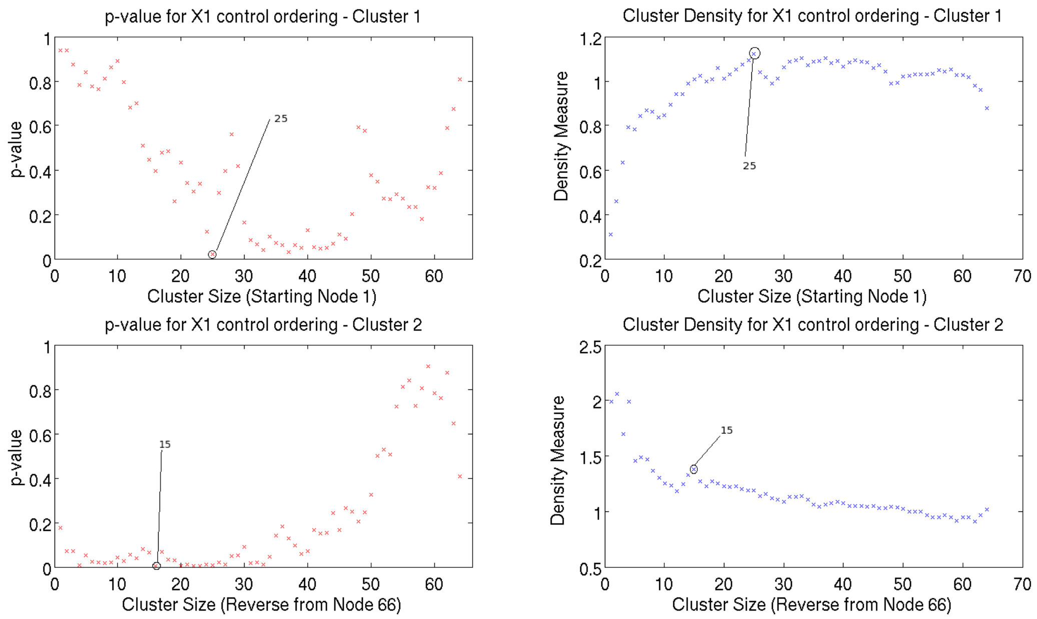

Figure 3.5: p-value and relative density measure for varying cluster size. Top images show values for a cluster starting at the first node, while the bottom images show those starting at the end node.

From Figure 3.4 the obvious cluster size for the $x^{[end]}$ ordering in the ketamine data is to have cluster one as nodes $1-15$ ($p = 0.0048$) and cluster two as nodes $66-50$ ($p < 0.0001$). This is a point where the p-value reaches significance, cluster size is reasonable and the cluster density is locally optimal. To expand on this, the top left figure shows that after approximately a cluster size of 20, as this cluster size increases to encompass progressively more of the data (until close to the end) the p-values remain significant. This is explained by the fact that, as seen on the top right density measure plot, most of the density is located on that side of the ordering, thus it is highly likely any group containing those high density nodes will show significance.

Similarly, from Figure 3.5 the choice clusters are nodes $1-25$ ($p = 0.0027$) and $66-52$ ($p < 0.0001$). Having decided on cluster size, the next step is to check for representation within these clusters for each of the neural subsystems. A reordered list of brain regions with colour coding for functional group is shown in Figure 3.6. Visual assessments are difficult and uninformative with this number of regions, as can be seen in the complicated colour diagram of reordered lists in Figure 3.6. To achieve a quantitative result, the level of representation can be calculated with a hypothesis test in order to determine the likelihood of a chance result, in this case a hypergeometric probability test (see [157] for more details) [220, 42]. In words, we are testing whether a functional group is present in a cluster at a level that is very unlikely to arise if the members of the group were distributed at random throughout the list of nodes.

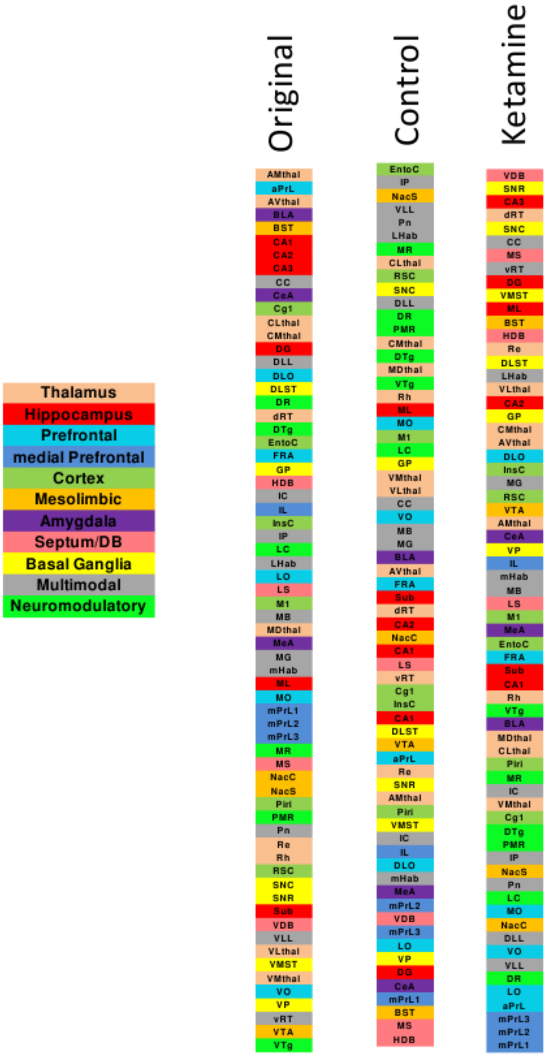

Figure 3.6: Left: Colour key for brain subsystems. Right: Lists of brain regions (nodes) for which measurements were given. Control and ketamine lists show $x^{[1]}$ and $x^{[end]}$ reorderings of the list respectively, according to the GSVD.

Tables 3.2 and 3.3 show the results for hypergeometric probability testing of the number of regions appearing in a particular cluster for each functional group, for the control and ketamine treated groups respectively. Functional groups that are significantly represented in a cluster are posited to be functionally relevant in the data for that particular animal group.

Control			
Cluster One	Nodes 1-25		
Brain subsystem	Regions in subsystem (N)	Regions in cluster (n)	Hypergeometric probability (P($x \geq n$)
Thalamus	11	6	0.152
Hippocampus	6	1	0.942
Prefrontal Cortex	6	1	0.942
medial Prefrontal Cortex	4	0	1.000
Cortex	6	3	0.378
Mesolimbic	4	1	0.845
Amygdala	3	0	1.000
Septum/DB	4	0	1.000
Basal Ganglia	6	2	0.718
Neuromodulatory	**6**	**6**	**1.48e-3**
Multimodal	10	5	0.266
Cluster Two	Nodes 52-66		
Thalamus	11	0	1.000
Hippocampus	6	0	1.000
Prefrontal Cortex	6	2	0.414
medial Prefrontal Cortex	**4**	**4**	**1.89e-3**
Cortex	6	0	1.000
Mesolimbic	4	0	1.000
Amygdala	3	2	0.130
Septum/DB	**4**	**3**	**0.034**
Basal Ganglia	6	1	0.800
Neuromodulatory	6	0	1.000
Multimodal	10	2	0.722

Table 3.2: p-values from a hypergeometric probability distribution hypothesis test of finding at least the number of observed elements of a given subsystem within the clusters generated by the GSVD.

Ketamine			
Cluster One	Nodes 1-15		
Brain subsystem	Regions in subsystem (N)	Regions in cluster (n)	Hypergeometric probability (P($x \geq n$)
Thalamus	11	3	0.480
Hippocampus	6	3	0.125
Prefrontal Cortex	6	0	1.000
medial Prefrontal Cortex	4	0	1.000
Cortex	6	0	1.000
Mesolimbic	4	1	0.653
Amygdala	3	0	1.000
Septum/DB	**4**	**3**	**0.034**
Basal Ganglia	**6**	**4**	**0.021**
Neuromodulatory	6	0	1.000
Multimodal	10	1	0.939
Cluster Two	Nodes 50-66		
Thalamus	11	0	1.000
Hippocampus	6	0	1.000
Prefrontal Cortex	**6**	**4**	**0.034**
medial Prefrontal Cortex	**4**	**3**	**0.049**
Cortex	6	0	1.000
Mesolimbic	4	2	0.271
Amygdala	3	0	1.000
Septum/DB	4	0	1.000
Basal Ganglia	6	0	1.000
Neuromodulatory	**6**	**4**	**0.034**
Multimodal	10	4	0.228

Table 3.3: p-values from a hypergeometric probability distribution hypothesis test of finding at least the number of observed elements of a given subsystem within the clusters generated by the GSVD.

In the control data, the neuromodulatory group shows significance in cluster one ($p = 0.00148$) and septum/diagonal band and medial prefrontal cortex appear in cluster two together ($p = 0.00189$, $p = 0.034$, respectively). This changes in the ketamine treated group - septum/diagonal band and the basal ganglia now appear in a cluster ($p = 0.034$, $p = 0.021$ respectively) and the prefrontal cortex, medial prefrontal cortex and neuromodulatory regions show significance ($p = 0.034$, $p = 0.049$, $p = 0.034$ respectively).

The transition of the septum/diagonal band from the same cluster as the medial prefrontal cortex in the control ordering to separate clusters in the ketamine ordering suggests that the functional interactions of these two subsystems is compromised following ketamine treatment. In the opposite direction, the transition of the neuromodulatory subsystem from a separate cluster from the medial prefrontal cortex to the same cluster across reorderings suggests these subsystems have become more functionally integrated in the ketamine treated animals.

3.2.2 Calculating the Variance

Working with this data set in particular was interesting in that it presented an opportunity to test an alternative method for incorporating known information (i.e., functional groups) into the study. In this subsection we therefore take a different approach to analysing the reordered set of nodes, with the aim of investigating whether the orderings uncover clusters nodes that form particular functional groups (as outlined in Section 3.1.2). The measure used here involves reordering the brain regions with the vectors from the GSVD as before, and calculating the variance of the components in the reordered vectors that pertain to each functional group. For example, there are 11 regions assigned to the thalamus - each of the 11 regions has a corresponding component within the reordering vector from X^{-T}. The variance of these 11 components throughout the ordering is a measure of how close together

Functional Group	Control Exclusive		Ketamine Exclusive	
	variance	p-value	variance	p-value
Thalamus	0.0027	0.0653	**0.0033**	**0.0467**
Hippocampus	0.0023	0.1291	0.0011	0.3813
Prefrontal Cortex	0.0041	0.4503	**0.0029**	**0.0334**
medial Prefrontal Cortex	**0.0005**	**0.0303**	0.0033	0.4420
Cortex	0.0047	0.5553	0.0021	0.3040
Mesolimbic	0.0073	0.8409	0.0047	0.6275
Amygdala	0.0035	0.4602	0.0006	0.0952
Septum/DB	0.0034	0.3893	0.0032	0.4245
Basal Ganglia	0.0045	0.5278	0.0018	0.1121
Neuromodulatory	**0.0001**	**0.0001**	**0.0005**	**0.0041**
Multimodal	0.0044	0.4770	0.0038	0.4089

Table 3.4: Single neural subsystem in control and ketamine-treated mice. Columns show the variance, and corresponding p-value, of singular value components from GSVD orderings for nodes within each functional group. Significant values are in bold.

they have been mapped. Then when the variance of nodes within a functional group is relatively small the component nodes are positioned closely together in the ordering, suggesting they are functionally interacting.

Measuring Statistical Significance: p-values

As part of validating this method we need to quantify the level of significance of our results. As an example, to test the significance of the variance of the aforementioned 11 thalamic regions, we will randomize the distribution of singular vector components $1,000$ times and take the variance of the blocks of 11 each time. The p-value is then the frequency with which the experimental variance is smaller than those from the randomizations.

Table 3.4 shows the variance of the singular vector components for each of the given neural subsystems, and an assigned p-value where significant values are in bold.

These results show that ketamine interferes with the connectivity of the medial prefrontal cortex, and promotes activity in the prefrontal cortex and thalamus. The neuromodulatory region shows significance in both ketamine and control groups. This mirrors the results from hypergeometric probability testing of the ketamine data ordering in an earlier section in Table 3.3 where the prefrontal cortex shows significance. The septum/diagonal band was also significantly overrepresented in hypergeometric probability testing in Table 3.3 - this significance does not show in the variance measure in this section. This shows the septum/diagonal band regions are part of a larger cluster involving other brain regions (since they are significantly over-represented in the cluster), but the individual component regions are not clustered together.

Similarly, Table 3.2 shows the neuromodulatory and medial prefrontal cortex functional groups as having significant over-representation where their component regions are also significantly grouped in Table 3.4.

3.2.3 Two-way Neural Subsystem Interaction

To further quantify the alterations in the neural subsystem interaction induced by acute ketamine treatment, the significance of bipartite interactions between all investigated subsystems was also determined, through examination of the variance of the components of all regions across a pair of subsystems. For example, to quantify interaction between the thalamus and the hippocampus, the variance of these 17 (11 thalamic, 6 hippocampal) components measures how close together they have been mapped. As with the previous subsection, $p-$values are calculated through randomization of the distribution of singular vector components. The variance of the block of 17 components is stored each time, and the $p-$value is the frequency with which the experimental variance is smaller than those from the randomizations.

This measurement of clustering across subsystems is applied and shown in Tables 3.5 and 3.6 In this way, the observation in Section 3.2 that the Septum/DB-mPFC show significant interaction in the control network but not in the ketamine-treated animals gains weight through verification from a different angle. In addition, in the control network the neuromodulatory subsystem is significantly functionally coupled to the thalamus ($p = 0.004$) and hippocampus ($p = 0.003$). In the ketamine treated animals these interactions are lost and the neuromodulatory subsystem becomes more functionally coupled to the PFC ($p = 0.007$), mPFC ($p = 0.022$), cortex ($p = 0.003$) and the amygdala ($p = 0.004$). Finally, the functional coupling of thalamic regions to the amygdala ($p = 0.012$), basal ganglia ($p = 0.019$) and cortical regions ($p = 0.005$) is enhanced by ketamine treatment. In control animals, a significant interaction existed between the hippocampus and thalamus neural systems ($p = 0.013$) that was not altered by ketamine treatment ($p = 0.033$).

Control	Thalamus	Hippocampus	Prefrontal Cortex	medial Prefrontal Cortex	Cortex	Mesolimbic	Amygdala	Septum/DB	Basal Ganglia	Neuromodulatory	Multimodal
Thalamus	-	0.003	0.003	0.004	0.003	0.004	0.003	0.005	0.003	0.002	0.003
Hippocampus	**0.013**	-	0.003	0.003	0.003	0.004	0.003	0.004	0.003	0.002	0.003
Prefrontal Cortex	0.076	0.089	-	0.004	0.004	0.005	0.004	0.005	0.004	0.003	0.005
medial Prefrontal Cortex	0.388	0.117	0.317	-	0.006	0.005	0.002	0.002	0.004	0.006	0.006
Cortex	0.063	0.196	0.464	0.833	-	0.005	0.005	0.007	0.005	0.003	0.004
Mesolimbic	0.199	0.293	0.582	0.590	0.741	-	0.005	0.006	0.005	0.004	0.005
Amygdala	0.158	0.108	0.327	**0.048**	0.683	0.661	-	0.003	0.004	0.004	0.005
Septum/DB	0.674	0.367	0.618	**0.026**	0.953	0.845	0.248	-	0.005	0.007	0.007
Basal Ganglia	0.115	0.124	0.330	0.313	0.575	0.637	0.375	0.613	-	0.004	0.005
Neuromodulatory	**0.004**	**0.033**	0.163	0.793	0.042	0.321	0.345	0.943	0.272	-	0.004
Multimodal	0.070	0.255	0.528	0.932	0.397	0.726	0.735	0.985	0.625	**0.046**	-

Variance

p-values

Table 3.5: Two-way neural subsystem positioning in control mice. These results show the variance (top right triangle) and p-values (bottom left triangle, significant in bold), for the $x^{[1]}$ control ordering.

Ketamine	Thalamus	Hippocampus	Prefrontal Cortex	medial Prefrontal Cortex	Cortex	Mesolimbic	Amygdala	Septum/DB	Basal Ganglia	Neuromodulatory	Multimodal
Thalamus	-	0.003	0.004	0.005	0.002	0.003	0.002	0.003	0.002	0.003	0.003
Hippocampus	0.330	-	0.006	0.008	0.003	0.005	0.003	0.003	0.002	0.005	0.005
Prefrontal Cortex	0.401	0.909	-	0.003	0.002	0.004	0.003	0.007	0.006	0.002	0.004
medial Prefrontal Cortex	0.744	0.989	0.199	-	0.003	0.005	0.004	0.009	0.008	0.002	0.004
Cortex	**0.005**	0.163	0.083	0.291	-	0.002	0.001	0.003	0.003	0.001	0.003
Mesolimbic	0.104	0.601	0.395	0.689	0.070	-	0.003	0.005	0.004	0.002	0.004
Amygdala	**0.012**	0.234	0.156	0.424	**0.002**	0.202	-	0.003	0.003	0.001	0.003
Septum/DB	**0.049**	0.200	0.946	0.997	0.231	0.708	0.332	-	0.002	0.005	0.005
Basal Ganglia	**0.019**	0.073	0.936	0.994	0.129	0.537	0.144	0.054	-	0.005	0.005
Neuromodulatory	0.116	0.648	**0.007**	**0.022**	**0.003**	0.083	**0.004**	0.737	0.706	-	0.004
Multimodal	0.130	0.614	0.317	0.596	0.062	0.339	0.155	0.707	0.627	0.079	-

Variance

p-values

Table 3.6: Two-way neural subsystem clustering in ketamine-treated mice. These results show the variance (top right triangle) and p-values (bottom left triangle, significant in bold) for clustering across two functional subsystems, for the $x^{[end]}$ ketamine ordering.

3.3 Discussion

In this chapter we have verified the ability of the GSVD to find mutually exclusive clusters in 2-DG type data. We have illustrated biologically relevant results, and can now briefly compare with results observed in humans.

In humans, MRI studies have shown abnormal densities in many of these neural subsystems - including an increase in basal ganglia and decrease in amygdala, frontal, cortical and thalamic regions [55, 70]. These results from meta-analyses included studies with patients studied at point of onset (and so medication free), which is important as treatment may mask the underlying disease mechanism. Though these results are not directly transferrable, since they are from a different data type, as well as being from human study, it is clear that schizophrenia has a wide-ranging effect across the whole brain - this is consistent with the brain wide disruption observed in this chapter.

The two-way interaction analysis shows that ketamine substantially alters interaction between subsystems - increased activity in the PFC and thalamus between other subsystems is seen, results which parallel those (published in [42]) found with another NMDA receptor antagonist, PCP, in the following chapter.

Our clustering results mirror those found through the alternative approach of examination of traditional topological network parameters - clustering coefficients, closeness and betweenness centrality, mean degree and average path length, as found by our co-authors on the same data in the submitted publication. This suggests that the novel approach to network clustering and the method of calculating the variance of spectral vector components developed here can yield biologically relevant results. Further biological interpretation and detailed discussion of the implications for the ketamine model have been provided by other members of the collaboration in the manuscript [39].

There is however an obvious potential disconnect in that this work has been carried out on a rodent mouse model (that is, there is an issue of transferrability of results to humans). However, animal models are important for various aforementioned reasons and, most importantly, the purpose of an animal model in this context is to mimic some aspect of the condition. While the presentation may be legitimate, at a system level the biology differs from that seen in humans - questioning the utility of the model. In particular, hypofrontality is a key neurological correlate for individuals with schizophrenia - the results seen in this chapter suggest that ketamine induces increased activity in the frontal regions, a state of hyperfrontality. It is therefore possible that the alterations in the network in the acute phase may not be representative of the long term changes in network topology associated with the chronic condition.

We also note that although we have computed $p-$values to quantify significance, the sample size of $n = 9$ mice per group is not ideal. More data would be desirable in order to allow further patterns to be discovered. Ideally these results will be taken as a point of interest for further work. Finally, this chapter has established the general utility of the GSVD/variance clustering approach - and this approach will be used in the following chapter.

Chapter 4

Integration of Neural Subsystems in Rats Administered Modafinil and PCP

The results in this chapter form part of a publication:-

Dawson N, Xiao X, McDonald M, Higham DJ, Morris BJ, Pratt, JA. Sustained NMDA receptor hypofunction induces compromised neural systems integration and schizophrenia-like alterations in functional brain networks, Cerebral Cortex, 2012 [42].

This chapter presents the material from that manuscript concerned with the use of spectral methods to analyse the effects of modafinil on integration of neural subsystems in the presence of PCP or vehicle; specifically clustering and hypergeometric probability test results.

4.1 Background and the PCP Animal Model

Following successful collaboration with Dr Neil Dawson in Chapter 3, we continued efforts, now exploring an alternative animal model. The previous chapter revealed changes in the functional integration of neural subsystems as a response to acute dosing of the NMDA receptor antagonist ketamine. This chapter provides analysis of data from a similar $2-DG$ study, this time with subchronic phencyclidine (PCP) dosing of Lister Hooded rats. PCP is also an NMDA receptor antagonist and like ketamine can produce schizophrenia-like symptoms in healthy volunteers and exacerbate psychosis in patients with schizophrenia [100, 218]. The animal model of subchronic PCP treatment effectively reproduces in rats much of the symptomatology observed in human schizophrenia [53, 198], notably cognitive deficits [33, 41]. We begin with the assumption that the schizophrenia-like symptoms induced by the PCP are recreating the differential neural operation that is present in the disease state. If this assumption is valid then, as with the previous chapter, we can observe alterations induced by administration of PCP as compared with a control group.

We will examine data in the context of functional integration of neural subsystems. Integration between these subsystems is essential for overall cognitive function - complex tasks are comprised of components that localize to different areas of the brain. This functional integration is compromised in schizophrenia, a fact which is proposed to account for cognitive deficits seen in the condition [175, 91]. Specific proposals include compromised integration between prefrontal regions, both hippocampal [151] and cortical [110] regions.

The psychoactive drug modafinil has been shown to reverse some of the cognitive deficits observed in animals treated with PCP [41]. Similarly, modafinil has been shown to restore deficits in cognitive processing in schizophrenia [41], suggesting modafinil is a candidate for a translationally relevant treatment for cognitive deficits.

4.2 Introducing the Data

The approach of analysis in this chapter is similar that used in the previous chapter with the ketamine model. The data is from a 2-DG study, and was provided through a further collaboration with Dr Neil Dawson (who carried out the experiments and processed the data: processing involved forming the correlations across animals and taking the Fisher-z transform to increase normality). This time there are four data sets:-

- $A_{ctrl} \in \mathbb{R}^{64\times64}$ control data ($n = 7$, 0.9% physiological saline [i.p.])
- $B_{pcp} \in \mathbb{R}^{64\times64}$ PCP ($n = 9, 2.58mgkg^{-1}$ PCP.HCl [i.p.])
- $C_{mod} \in \mathbb{R}^{64\times64}$ modafinil ($n = 6$, $64mgkg^{-1}$ modafinil in 0.5% methylcellulose [o.p.])
- $D_{pcpmod} \in \mathbb{R}^{64\times64}$ PCP + modafinil ($n = 7$)

Arbitrarily ordered heat maps of the data are shown in Figure 4.1.

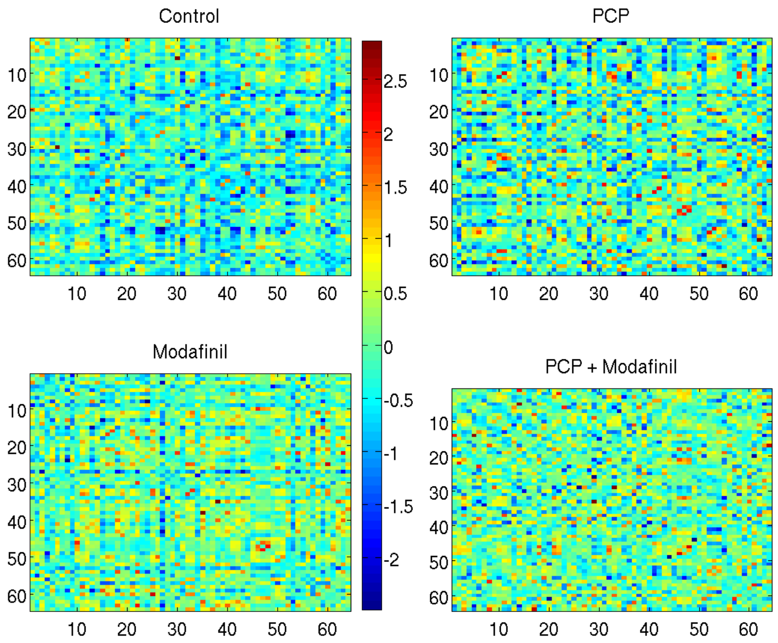

Figure 4.1: Heat maps of control, PCP, modafinil and PCP + modafinil data

The four data sets can be compared in six different combinations, with twelve orderings - GSVD of (A,B) results in an ordering for both A and B. This chapter will provide the following comparisons:-

Control	$<->$	Modafinil
Control	$<->$	PCP
Control	$<->$	PCP + Modafinil

The others:-

PCP	$<->$	Modafinil
Modafinil	$<->$	PCP + Modafinil
PCP	$<->$	PCP + Modafinil

are less informative in terms of the goals of this experiment, hence are not included in this study.

The same functional subsystem allocations (Section 3.1.2) are used as with the previous data set, with the slight difference that there are 64 regions instead of 66. The number of regions per functional subsystem in this chapter is shown in Table 4.3.

Functional Group	Number of Regions
Thalamus	11
Hippocampus	5
Prefrontal Cortex	6
medial Prefrontal Cortex	4
Cortex	6
Mesolimbic	4
Amygdala	3
Septum/DB	4
Basal Ganglia	6
Neuromodulatory	5
Multimodal	10

Table 4.3: Functional groups of 64 regions of interest

4.3 Hypergeometric Probability Testing of Subsystem Representation

We will test GSVD reordered data for clusters and overrepresentation of functional subsystems in each of the combinations of data. In the results, this time we omit any subsystems that have zero regions within the cluster (and so hypergeometric probability of $p = 1.000$). Clusters were verified using the same cluster quality measure as in Section 3.2.1. The resultant cluster size and $p-$value are given in each section as appropriate.

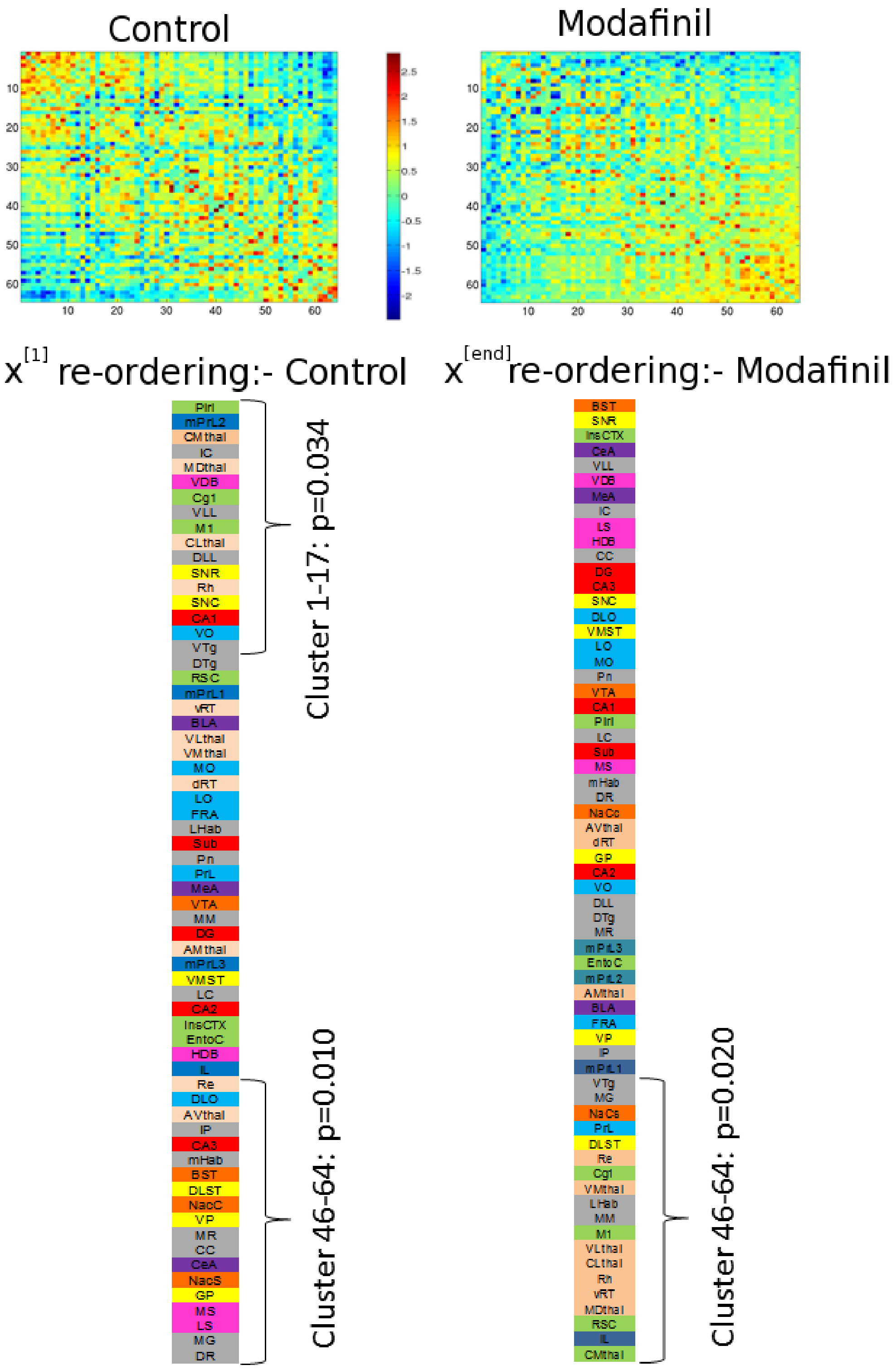

Figure 4.2: Top: Heat maps of control and modafinil data reordered according to GSVD. Bottom: reordered brain region list with $p-$values highlighting significant clusters.

Modafinil and Control

Figure 4.2 shows a reordering of the modafinil and control data according to the GSVD. In the modafinil ordered heat map (top right) we can already see visually that the hot nodes are concentrated towards one end of the ordering - this is confirmed when significant clusters are elucidated through the cluster quality measure where, in the control data, nodes $1-17$ ($p = 0.034$) and nodes $46-64$ ($p = 0.010$) are clustered. Similarly, in the modafinil data there is a cluster of nodes $46-64$ ($p = 0.020$).

Control			
Cluster One	Nodes 1-17		
Brain subsystem	Regions in subsystem (N)	Regions in cluster (n)	Hypergeometric probability (P($x \geq n$))
Thalamus	11	4	0.321
Hippocampus	5	1	0.799
Prefrontal Cortex	6	1	0.857
medial Prefrontal Cortex	4	1	0.719
Cortex	6	3	0.185
Septum/DB	4	1	0.719
Basal Ganglia	6	2	0.509
Neuromodulatory	5	1	0.799
Multimodal	10	1	0.966
Cluster Two	Nodes 46-64		
Thalamus	11	2	0.905
Hippocampus	5	1	0.840
Prefrontal Cortex	6	1	0.891
Mesolimbic	4	3	0.075
Amygdala	3	1	0.659
Septum/DB	4	2	0.341
Basal Ganglia	6	3	0.242
Neuromodulatory	5	2	0.469
Multimodal	10	4	0.334

Table 4.4: p-values from a hypergeometric probability distribution hypothesis test of subsystem representation in a significant cluster in control not modafinil rat data.

Modafinil			
Cluster One	Nodes 1-15		
Brain subsystem	Regions in subsystem (N)	Regions in cluster (n)	Hypergeometric probability (P($x \geq n$)
Thalamus	**11**	**7**	**0.012**
Prefrontal Cortex	6	1	0.891
medial Prefrontal Cortex	4	1	0.766
Cortex	**6**	**4**	**0.049**
Mesolimbic	4	1	0.766
Basal Ganglia	6	1	0.891
Neuromodulatory	5	1	0.840
Multimodal	10	3	0.624

Table 4.5: p-values from a hypergeometric probability distribution hypothesis test of subsystem representation in a significant cluster in modafinil not control rat data.

Table 4.5 shows the modafinil exclusive ordering. The modafinil data presents one significant cluster including clustering between thalamic ($p = 0.012$) and cortical ($p = 0.049$) regions. Additionally, in the control exclusive ordering (Table 4.4) it is notable that the thalamic regions are divided across clusters, where those regions appear in the same cluster in the data from modafinil treated rats.

PCP and Control

Table 4.6 with results for control exclusive (vs PCP) data has two clusters, nodes 1-10 ($p = 0.046$) and nodes 54-64 ($p = 0.026$), as determined by the cluster quality measure. The PCP exclusive ordering also shows two significant clusters:- nodes 1-15 ($p = 0.031$) and nodes 53-64 ($p = 0.016$). In the control ordering, we see that the basal ganglia and thalamus are significantly overrepresented.

Control			
Cluster One	Nodes 1-10		
Brain subsystem	Regions in subsystem (N)	Regions in cluster (n)	Hypergeometric probability (P($x \geq n$)
Prefrontal Cortex	**6**	**3**	**0.044**
Septum/DB	4	1	0.502
Basal Ganglia	**6**	**4**	**0.004**
Multimodal	10	2	0.494
Cluster Two	Nodes 54-64		
Thalamus	**11**	**6**	**0.002**
Hippocampus	5	1	0.624
Cortex	6	1	0.694
Amygdala	3	1	0.438
Basal Ganglia	6	1	0.694
Multimodal	10	1	0.871

Table 4.6: p-values from a hypergeometric probability distribution hypothesis test of subsystem representation in a significant cluster in control, not PCP rat data.

PCP			
Cluster One	Nodes 1-15		
Brain subsystem	Regions in subsystem (N)	Regions in cluster (n)	Hypergeometric probability (P($x \geq n$)
Thalamus	11	5	0.071
Prefrontal Cortex	6	1	0.814
medial Prefrontal Cortex	4	2	0.211
Cortex	6	3	0.135
Multimodal	10	3	0.430
Cluster Two	Nodes 53-64		
Hippocampus	5	1	0.659
Mesolimbic	4	2	0.157
Amygdala	3	2	0.088
Septum/DB	4	1	0.574
Basal Ganglia	6	2	0.313
Neuromodulatory	5	1	0.659
Multimodal	10	3	0.276

Table 4.7: p-values from a hypergeometric probability distribution hypothesis test of subsystem representation in a significant cluster in PCP not control rat data.

PCP + Modafinil and Control

Tables 4.8 and 4.9 show results for control exclusive ordering compared to PCP + modafinil data and vice versa, respectively. There is one significant cluster in the control exclusive ordering- nodes $1 - 14$ ($p = 0.012$) and two clusters in the PCP + modafinil exclusive ordering $1 - 11$ ($p = 0.048$) and nodes $46 - 64$ ($p = 0.022$).

Control			
Cluster Two	Nodes 1-14		
Brain subsystem	Regions in subsystem (N)	Regions in cluster (n)	Hypergeometric probability (P($x \geq n$)
Thalamus	11	1	0.950
Hippocampus	5	3	0.065
Mesolimbic	4	1	0.638
Septum/DB	4	1	0.638
Basal Ganglia	6	2	0.392
Neuromodulatory	5	2	0.299
Multimodal	10	2	0.701

Table 4.8: p-values from a hypergeometric probability distribution hypothesis test of subsystem representation in a significant cluster in control (not PCP + modafinil) rat data.

PCP + Modafinil			
Cluster One	Nodes 1-11		
Brain subsystem	Regions in subsystem (N)	Regions in cluster (n)	Hypergeometric probability (P($x \geq n$)
Thalamus	11	4	0.085
Hippocampus	5	1	0.624
medial Prefrontal Cortex	4	1	0.539
Basal Ganglia	6	2	0.273
Neuromodulatory	5	2	0.201
Multimodal	10	1	0.871
Cluster Two	Nodes 46-64		
Thalamus	11	3	0.701
Prefrontal Cortex	6	2	0.582
Cortex	6	2	0.582
Mesolimbic	4	3	0.075
Amygdala	3	1	0.659
Septum/DB	4	1	0.766
Basal Ganglia	6	1	0.891
Neuromodulatory	5	2	0.469
Multimodal	10	4	0.334

Table 4.9: p-values from a hypergeometric probability distribution hypothesis test of subsystem representation in a significant cluster in PCP + modafinil exclusive (not control) rat data.

According to the hypergeometric probability test there are no significantly overrepresented subsystems in these orderings. This is interesting in that it suggests the disintegration induced through administration of PCP is counteracted by modafinil, with no particular subsystems differentiating between the two data sets.

4.4 Clustering of Individual Neural Subsystems

In this section we use the same data reordering from the GSVD, measuring the variance of the singular vector components for nodes in individual neural subsystems. Smaller values for variance correspond to functional groups with their constituents placed relatively closer together. The approach for calculating the variance of components in the appropriate reordering vector from X^{-T} is outlined in Section 3.2.2. Similarly, $p-$values are assigned where the components of the desired ordering vector are randomly permuted and a variance calculated for a group the same size as those in the test-subsystem (or pair). The $p-$value is the frequency with which the experimental variance is smaller than those from the random permutations.

In examination of the modafinil and control data, we see in Table 4.10 that thalamic and cortical regions are clustered in modafinil, and the prefrontal cortex is clustered only in control animals. This is in agreement with the hypergeometric probability testing in Table 4.5 which identified overrepresentation of thalamic and cortical regions.

Functional Group	Control Exclusive		Modafinil Exclusive	
	variance	p-value	variance	p-value
Thalamus	0.0017	0.1846	**0.0005**	**0.0030**
Hippocampus	0.0011	0.1411	0.0005	0.0548
Prefrontal Cortex	**0.0007**	**0.0322**	0.0012	0.1963
medial Prefrontal Cortex	0.0030	0.6913	0.0005	0.1027
Cortex	0.0026	0.5895	**0.0012**	**0.0056**
Mesolimbic	0.0006	0.0890	0.0068	0.9310
Amygdala	0.0019	0.4863	0.0043	0.7998
Septum/DB	0.0055	0.9678	0.0009	0.2179
Basal Ganglia	0.0025	0.5281	0.0039	0.7351
Neuromodulatory	0.0033	0.7409	0.0005	0.0063
Multimodal	0.0032	0.8631	0.0026	0.5127

Table 4.10: Single neural subsystem in control and modafinil-treated rats. Columns show the variance, and corresponding p-value, of singular value components from GSVD orderings for nodes within each functional group. Significant values are in bold.

Next, in Table 4.11 we see that the thalamus ($p = 0.0363$) and prefrontal regions (prefrontal cortex, $p = 0.0208$ and medial prefrontal cortex, $p = 0.0435$) are clustered in control animals and the effect is lost after PCP treatment. These results are consistent with the hypergeometric test results in Table 4.6, only we add medial prefrontal cortex involvement and lose the basal ganglia. The final single subsystem results are shown in Table 4.12 for animals that have received both modafinil and PCP. These results show that the mesolimbic region is the only subsystem that shows significant grouping in the control animals and not PCP + modafinil treated. The lack of difference, particularly in the frontal regions, is interesting since we find no subsystem oriented differences between the two groups. This result could suggest that on the level of neural subsystems, the combination of modafinil and PCP produces a similar neurological output as would be expected in healthy animals.

Functional Group	Control Exclusive		PCP Exclusive	
	variance	p-value	variance	p-value
Thalamus	**0.0011**	**0.0363**	0.0049	0.6329
Hippocampus	0.0012	0.2827	0.0019	0.1663
Prefrontal Cortex	**0.0009**	**0.0208**	0.0019	0.3589
medial Prefrontal Cortex	**0.0003**	**0.0435**	0.0003	0.114
Cortex	0.0006	0.0334	0.0022	0.1797
Mesolimbic	**0.0001**	**0.0030**	0.0113	0.9615
Amygdala	0.0004	0.1526	0.0009	0.1713
Septum/DB	0.0030	0.5920	0.0012	0.1158
Basal Ganglia	0.0036	0.5941	0.0016	0.0855
Neuromodulatory	0.0052	0.3455	0.0035	0.4355
Multimodal	0.0027	0.5220	0.0060	0.7798

Table 4.11: Single neural subsystem in control and PCP-treated rats. Columns show the variance, and corresponding p-value, of singular value components from GSVD orderings for nodes within each functional group. Significant values are in bold.

Functional Group	Control Exclusive		PCP + Modafinil Exclusive	
	variance	p-value	variance	p-value
Thalamus	0.0038	0.9410	0.0031	0.9135
Hippocampus	0.0020	0.8852	0.0017	0.7478
Prefrontal Cortex	0.0008	0.1526	0.0020	0.8143
medial Prefrontal Cortex	0.0023	0.5923	0.0027	0.6113
Cortex	0.0031	0.0435	0.0028	0.3243
Mesolimbic	**0.0005**	**0.0029**	0.0023	0.7291
Amygdala	0.0032	0.3020	0.0012	0.1802
Septum/DB	0.0050	0.8497	0.0026	0.4531
Basal Ganglia	0.0080	0.4510	0.0011	0.0705
Neuromodulatory	0.0044	0.3455	0.0029	0.3851
Multimodal	0.0040	0.9228	0.0046	0.6359

Table 4.12: Single neural subsystem in control and PCP + modafinil-treated rats. Columns show the variance, and corresponding p-value, of singular value components from GSVD orderings for nodes within each functional group. Significant values are in bold.

4.5 Clustering and the Two-way Interaction of Neural Subsystems

This next section extends the approach of taking the variance across subsystems to the two dimensional case - each subsystem is grouped with each other subsystem in turn and variance measured.

Modafinil and Control

In the comparison of modafinil and control data sets, Tables 4.13 and 4.14 show that the prefrontal cortex is significantly activated in both data sets - interacting with the thalamus, hippocampus and amygdala in the control animals, and the thalamus and hippocampal groups in the modafinil treated animals. Though the GSVD identifies mutually exclusive clusters, we have found that the prefrontal cortex is activated strongly in both. This could be an indication that the groups are clustered in a different fashion, migrating positions and playing a different role within the network. The modafinil animals exhibit clustering between medial prefrontal cortex and the thalamus. In addition, the multimodal region is more central, working with the thalamus, hippocampus and medial prefrontal cortex. With induction of clustering across many multimodal combinations, modafinil has a wide-ranging effect across the whole system.

Ctrl vs Mod	Thalamus	Hippocampus	Prefrontal Cortex	med Prefrontal Cortex	Cortex	Mesolimbic	Amygdala	Septum/DB	Basal Ganglia	Neuromodulatory	Multimodal	
Thalamus	-	0.002	0.001	0.002	0.002	0.002	0.002	0.003	0.002	0.003	0.003	
Hippocampus	0.074	-	0.001	0.002	0.002	0.001	0.001	0.003	0.002	0.002	0.002	
Prefrontal Cortex	**0.016**	**0.007**	-	0.001	0.002	0.001	0.001	0.003	0.002	0.002	0.002	
med Prefrontal Cortex	0.162	0.258	0.080	-	0.003	0.003	0.002	0.004	0.003	0.003	0.003	
Cortex	0.167	0.381	0.124	0.538	-	0.003	0.003	0.004	0.003	0.003	0.003	
Mesolimbic	0.420	**0.028**	0.062	0.569	0.805	-	0.001	0.003	0.002	0.003	0.002	Variance
Amygdala	0.187	0.063	**0.030**	0.499	0.594	0.105	-	0.004	0.002	0.003	0.002	
Septum/DB	0.857	0.656	0.593	0.961	0.987	0.613	0.853	-	0.003	0.004	0.004	
Basal Ganglia	0.392	0.168	0.116	0.670	0.795	0.176	0.356	0.838	-	0.003	0.003	
Neuromodulatory	0.553	0.471	0.363	0.813	0.885	0.649	0.680	0.972	0.760	-	0.003	
Multimodal	0.550	0.312	0.280	0.782	0.890	0.329	0.549	0.925	0.594	0.867	-	
						p-values						

Table 4.13: Variance (top right) and significance (bottom left) of GSVD ordering vector components for two-way neural subsystem comibinations. This table shows significant combinations of subsystems in the control (not modafinil) $x^{[end]}$ ordering.

Mod vs Ctrl	Thalamus	Hippocampus	Prefrontal Cortex	med Prefrontal Cortex	Cortex	Mesolimbic	Amygdala	Septum/DB	Basal Ganglia	Neuromodulatory	Multimodal	
Thalamus	-	0.002	0.001	0.001	0.002	0.003	0.003	0.003	0.003	0.002	0.001	
Hippocampus	0.078	-	0.001	0.001	0.003	0.003	0.002	0.001	0.002	0.002	0.001	
Prefrontal Cortex	**0.020**	**0.028**	-	0.001	0.003	0.003	0.002	0.002	0.002	0.002	0.001	
med Prefrontal Cortex	$<$ **0.001**	0.193	0.105	-	0.003	0.005	0.004	0.003	0.003	0.002	0.001	
Cortex	0.296	0.568	0.530	0.647	-	0.005	0.005	0.004	0.004	0.003	0.003	
Mesolimbic	0.659	0.552	0.638	0.875	0.966	-	0.005	0.003	0.005	0.004	0.003	Variance
Amygdala	0.473	0.268	0.433	0.744	0.924	0.916	-	0.002	0.004	0.003	0.003	
Septum/DB	0.438	**0.047**	0.222	0.602	0.831	0.712	0.369	-	0.003	0.003	0.002	
Basal Ganglia	0.450	0.368	0.419	0.677	0.898	0.892	0.750	0.532	-	0.003	0.002	
Neuromodulatory	0.195	0.227	0.216	0.360	0.720	0.797	0.639	0.419	0.603	-	0.002	
Multimodal	**0.002**	**0.023**	0.227	**0.013**	0.515	0.686	0.485	0.277	0.463	0.226	-	
						p-values						

Table 4.14: Variance (top right) and significance (bottom left) of GSVD ordering vector components for two-way neural subsystem comibinations. This table shows significant combinations of subsystems in the modafinil (not control) $x^{[1]}$ ordering.

PCP and Control

In the two-way comparison of animals treated with PCP and the control group, we find in Table 4.15 that the hippocampus and prefrontal cortex play a significant role in the control animals, exhibiting numerous two-way interactions across the brain. There is a marked difference in the PCP treated animals in Table 4.16. Instead, this group show activity across the thalamus - with diminished hippocampal and prefrontal activity.

Ctrl vs PCP	Thalamus	Hippocampus	Prefrontal Cortex	med Prefrontal Cortex	Cortex	Mesolimbic	Amygdala	Septum/DB	Basal Ganglia	Neuromodulatory	Multimodal	
Thalamus	-	0.005	0.003	0.003	0.003	0.006	0.005	0.005	0.004	0.005	0.005	Variance
Hippocampus	0.530	-	0.002	0.004	0.004	0.005	0.002	0.001	0.002	0.004	0.003	
Prefrontal Cortex	0.175	**0.022**	-	0.001	0.002	0.005	0.002	0.002	0.002	0.004	0.003	
med Prefrontal Cortex	0.367	0.419	**0.009**	-	0.002	0.008	0.006	0.004	0.004	0.006	0.007	
Cortex	0.319	0.295	**0.010**	0.013	-	0.007	0.005	0.004	0.003	0.005	0.006	
Mesolimbic	0.945	0.686	0.636	0.947	0.890	-	0.006	0.005	0.005	0.007	0.006	
Amygdala	0.747	**0.037**	0.110	0.721	0.574	0.815	-	0.001	0.001	0.005	0.002	
Septum/DB	0.599	**0.018**	**0.033**	0.509	0.367	0.713	**0.014**	-	0.001	0.005	0.002	
Basal Ganglia	0.461	**0.008**	**0.010**	0.349	0.235	0.591	**0.022**	**0.011**	-	0.004	0.003	
Neuromodulatory	0.849	0.521	0.385	0.828	0.741	0.955	0.705	0.557	0.451	-	0.005	
Multimodal	0.876	0.123	0.309	0.865	0.768	0.858	0.088	0.086	0.083	0.793	-	
	p-values											

Table 4.15: Variance (top right) and significance (bottom left) of GSVD ordering vector components for two-way neural subsystem comibinations. This table shows significant combinations of subsystems in the control (not PCP) $x^{[1]}$ ordering.

PCP vs ctrl	Thalamus	Hippocampus	Prefrontal Cortex	med Prefrontal Cortex	Cortex	Mesolimbic	Amygdala	Septum/DB	Basal Ganglia	Neuromodulatory	Multimodal	
Thalamus	-	0.001	0.002	0.001	0.001	0.001	0.001	0.003	0.005	0.002	0.003	
Hippocampus	**0.035**	-	0.002	0.001	0.001	0.001	0.001	0.004	0.005	0.002	0.003	
Prefrontal Cortex	0.052	0.140	-	0.002	0.002	0.001	0.002	0.003	0.004	0.002	0.003	
med Prefrontal Cortex	**0.013**	0.074	0.192	-	0.000	0.001	0.000	0.005	0.006	0.003	0.004	
Cortex	**0.010**	0.041	0.146	**0.003**	-	0.001	0.000	0.005	0.006	0.003	0.003	Variance
Mesolimbic	**0.004**	0.064	0.089	0.175	0.124	-	0.001	0.002	0.003	0.002	0.003	
Amygdala	**0.022**	0.099	0.249	**0.005**	**0.004**	0.170	-	0.005	0.006	0.003	0.004	
Septum/DB	0.612	0.683	0.586	0.910	0.856	0.332	0.904	-	0.003	0.003	0.005	
Basal Ganglia	0.904	0.878	0.787	0.976	0.974	0.456	0.979	0.484	-	0.004	0.005	
Neuromodulatory	0.170	0.228	0.216	0.386	0.320	0.155	0.362	0.549	0.729	-	0.003	
Multimodal	0.270	0.477	0.509	0.634	0.508	0.367	0.689	0.855	0.938	0.545	-	
					p-values							

Table 4.16: Variance (top right) and significance (bottom left) of GSVD ordering vector components for two-way neural subsystem comibinations. This table shows significant combinations of subsystems in the PCP (not control) $x^{[end]}$ ordering.

PCP + Modafinil and Control

The final comparison is between PCP + modafinil treated animals with the control. Table 4.17 shows some mesolimbic involvement with the hippocampus and cortex, which is not present in the PCP + modafinil ordering. In fact, the PCP + modafinil ordering shows almost no significant clustering - the exception being in a pairing of the multimodal regions and the amygdala. This further suggests that any differences between the PCP + modafinil and control data sets are not visible on a neural subsystems level, thereby vindicating the role of modafinil as an agent to restore subsystem capacity.

Ctrl vs PCPMod	Thalamus	Hippocampus	Prefrontal Cortex	med Prefrontal Cortex	Cortex	Mesolimbic	Amygdala	Septum/DB	Basal Ganglia	Neuromodulatory	Multimodal	
Thalamus	-	0.002	0.003	0.002	0.002	0.003	0.001	0.002	0.002	0.002	0.002	Variance
Hippocampus	0.886	-	0.006	0.002	0.002	0.006	0.001	0.003	0.001	0.001	0.003	
Prefrontal Cortex	0.670	0.005	-	0.004	0.005	0.004	0.003	0.004	0.004	0.001	0.004	
med Prefrontal Cortex	0.951	0.860	0.342	-	0.002	0.005	0.001	0.003	0.001	0.002	0.003	
Cortex	0.914	0.797	0.083	0.739	-	0.005	0.002	0.003	0.002	0.003	0.003	
Mesolimbic	0.493	**0.014**	0.377	0.097	**0.044**	-	0.005	0.006	0.004	0.008	0.004	
Amygdala	0.987	0.945	0.560	0.912	0.818	0.150	-	0.002	0.001	0.001	0.003	
Septum/DB	0.881	0.636	0.169	0.627	0.523	**0.022**	0.731	-	0.002	0.004	0.003	
Basal Ganglia	0.986	0.972	0.239	0.959	0.927	0.193	0.988	0.852	-	0.001	0.002	
Neuromodulatory	0.935	0.881	0.949	0.618	0.513	0.035	0.859	**0.305**	0.944	-	0.003	
Multimodal	0.947	0.682	0.139	0.789	0.736	0.169	0.836	0.628	0.903	0.636	-	

p-values

Table 4.17: Variance (top right) and significance (bottom left) of GSVD ordering vector components for two-way neural subsystem comibinations. This table shows significant combinations of subsystems in the control (not PCP + modafinil) $x^{[1]}$ ordering.

PCPMod vs Ctrl	Thalamus	Hippocampus	Prefrontal Cortex	med Prefrontal Cortex	Cortex	Mesolimbic	Amygdala	Septum/DB	Basal Ganglia	Neuromodulatory	Multimodal	
Thalamus	-	0.003	0.003	0.003	0.003	0.004	0.004	0.003	0.003	0.004	0.004	Variance
Hippocampus	0.466	-	0.002	0.001	0.002	0.003	0.003	0.002	0.002	0.002	0.004	
Prefrontal Cortex	0.545	0.934	-	0.001	0.002	0.002	0.002	0.001	0.002	0.002	0.004	
med Prefrontal Cortex	0.532	0.958	0.965	-	0.002	0.003	0.002	0.001	0.002	0.001	0.004	
Cortex	0.312	0.771	0.908	0.839	-	0.002	0.002	0.002	0.003	0.002	0.004	
Mesolimbic	0.116	0.443	0.713	0.551	0.890	-	0.002	0.002	0.003	0.001	0.004	
Amygdala	0.195	0.634	0.811	0.702	0.811	0.824	-	0.002	0.003	0.003	0.004	
Septum/DB	0.431	0.885	0.945	0.934	0.942	0.909	0.859	-	0.002	0.001	0.004	
Basal Ganglia	0.291	0.793	0.837	0.809	0.648	0.329	0.478	0.757	-	0.003	0.004	
Neuromodulatory	0.201	0.776	0.801	0.856	0.720	0.786	0.465	0.815	0.445	-	0.005	
Multimodal	0.050	0.238	0.183	0.053	0.595	0.350	**0.037**	0.798	0.826	0.417		-

p-values

Table 4.18: Variance (top right) and significance (bottom left) of GSVD ordering vector components for two-way neural subsystem comibinations. This table shows significant combinations of subsystems in the PCP + modafinil (not control) $x^{[end]}$ ordering.

4.6 Discussion

In this chapter we have seen that the NMDA receptor antagonist PCP compromises functional integration between neural subsystems - and that the cognitive promoter modafinil counteracts some of this effect by promoting particular subsystems. We have seen that consistent results can be obtained through different approaches to the analysis of the GSVD reordering of data. Hypergeometric probability testing has been shown again to yield biologically relevant results when used in combination with spectral clustering. The results from measuring the variance of components in spectral vectors are consistent with those obtained from the hypergeometric probability testing, which further validates our approach of exploratory cluster analysis when individual nodes can be categorized according to known information. In a comparison with the literature, we find that the PCP-induced compromised functional integration between the hippocampus and prefrontal cortex, perhaps driven through a reduction in connectivity of the thalamus. These findings concur with results from brain imaging studies of schizophrenia [151, 17]. The data shows that PCP induces segregation of functional subsystems, with the hippocampus, prefrontal cortex and medial prefrontal cortex acting as discrete clusters in PCP treated animals but not in control animals. These findings in this translational model also support prior hypotheses that schizophrenia is a disconnection syndrome [25]. We also show that modafinil has a dramatically different effect on subsystem clustering than PCP, failing to show many of the clusters present in the PCP-treated animals. This provides further evidence that modafinil has little interaction with NMDA receptors [153], which is important in considering reasons for the reason for efficacy of modafinil as a treatment for cognitive deficits. In addition, we have shown that the combined treatment of PCP + modafinil results in few differences from control animals, suggesting that modafinil has the ability to restore functional deficits introduced by PCP.

This work, combined with the previous chapter, establishes the approach of using the variance to measure clustering of a priori known sub-groups within a data set. The process of calculating the p-values can be adjusted depending on the data and desired stringency; it is felt that the basic approach taken here, in combination with visual assessment, is sufficient to make a strong statement about the viability of the method.

Chapter 5

Constructing Metabolic Networks from the KEGG and MetaCyc databases

One of the aims of this thesis is to develop tools to incorporate a priori known information into spectral clustering problems in order to increase flexibility in the approach to biological data analysis.

To that end, in this chapter we create a novel pair of metabolic networks from publically available metabolic databases in preparation for the next chapter where we propose and test a novel method of combining networks. The work in forming the KEGG metabolic network and testing is contained in the publication:-

Martin McDonald, Desmond J Higham, J Keith Vass. Spectral algorithms for heterogeneous biological networks, Briefings in functional genomics, Vol. 11 No. 6, 2012 [146]

5.1 Metabolic Networks

Metabolism describes the process through which cells break down and reassemble food and other nutrients. Due to the complexity of human metabolic pathways, there is still much work outstanding in uncovering the complete set of metabolic pathways and their constituent metabolites. For reference, the complete genome of E.coli (Escherichia coli MG1655), a widely studied bacterium, has been sequenced and contains $4,405$ genes - the Escherichia coli Metabolome Database (ECMDB, `http://www.ecmdb.ca`), currently contains > 2600 metabolites with links to 1500 genes and is still being updated [78].

Traditionally, a metabolic network consists of a collection of individual chemicals (the nodes) and their interactions (the edges) [120]. We have taken a different interpretation of the metabolic network, described in the following section, to meet our interest in the integration of microarray and metabolic networks. There are several efforts aimed at completing the human metabolome, we will use information from two such efforts that differ significantly.

5.2 KEGG

The Kyoto Encyclopedia of Genes and Genomes (KEGG:- `http://www.genome.jp/kegg/`) is a knowledge base for analysis of gene function, representing the efforts of Kanehisa Laboratories [106, 105]. KEGG contains a database of metabolic pathways (PATHWAY) which contains representations of different types of cellular processes, in this case we will consult the portion of the database dedicated to human metabolism. These pathway mappings are manually curated, with some interactions calculated computationally [106, 105].

The KEGG database contains pathway maps representing interactions and reactions between molecules - these maps are available for individual metabolic processes (e.g., glycolysis). Each individual pathway is comprised of a number of interactions and reactions, and has a unique ID. Through the knowledgebase of the Database for Annotation, Visualization and Integrated Discovery (DAVID [94, 93]) we have access to a list of genes with a corresponding KEGG metabolic pathway identifier - we specifically select all genes that are linked with a metabolite in a KEGG metabolic pathway.

The list we are using from DAVID associates each gene with a number of these KEGG pathway IDs, depending on which metabolic pathways the gene has product in. Each gene is counted once per pathway regardless of the number of gene products present in that pathway i.e., if a gene product has influence at multiple points in an individual pathway it is marked only as being present. With this information we can identify genes that have products with pathway IDs in common to declare which genes have known effets on metabolic processes in common. Using this, we have constructed a gene-gene metabolic adjacency matrix where gene i is connected to gene j if they appear at least once in the same metabolic pathway. We then construct B, an $\mathbb{R}^{N \times N}$ size weight matrix where b_{ij} has an integer weight corresponding to the number of times genes i and j appear in the same pathway as one another.

We used this viewpoint to construct a gene-gene co-incidence matrix whose (i, j) entry is a non-negative integer recording the number of times genes i and j appear in the same pathway. That is, if gene i is present in pathways x, y and z and gene j is present in pathways x and z then (i, j) will have a score of 2 for the number of matches. This also means the resultant metabolic network will be symmetric.

Rather than containing data on magnitudes of reactions, our metabolic network acts as an illustration of how well-connected genes are with each other in the

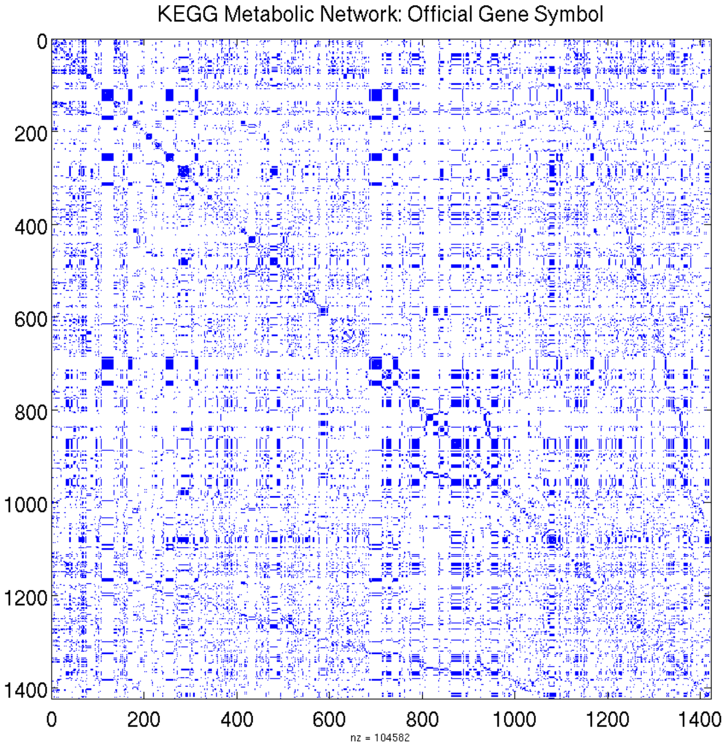

Figure 5.1: Binary representation of the KEGG metabolic network: blue signifies the presence of an edge.

sense of sharing metabolic pathways. We will use this knowledge to gain a new perspective on biological data to make statements about the metabolomics of a disease state.

5.2.1 KEGG Network Features

Figure 5.1 shows a MATLAB `spy` plot of the matrix form of the KEGG network. Each dot indicates a nonzero value in the matrix, for simplicity weights are not visible in this figure.

There are some block structures and repeating patterns visible in Figure 5.1, particularly around the diagonal. These structures are the result of closely related genes and cases where multiple gene IDs are present for the same gene. We will now outline some properties of the gene metabolic network we have constructed. Firstly, there is an overdispersion in the degree distribution (shown in Figure 5.2) - this is not expected to cause problems since we are aiming to focus on the high degree nodes, of which there are few. It is noted at this point that we are also limited by the completeness of the database from which we obtain the metabolic pathway information. The definition of pathways within KEGG also plays a key role in the structure - there may be bias towards specific metabolic pathways, of particular concern given the fact that metabolic pathways are not uniform in size. Some are much larger than others which means they automatically contain more gene products - thus genes in those high volume pathways have higher weights, effectively becoming more important.

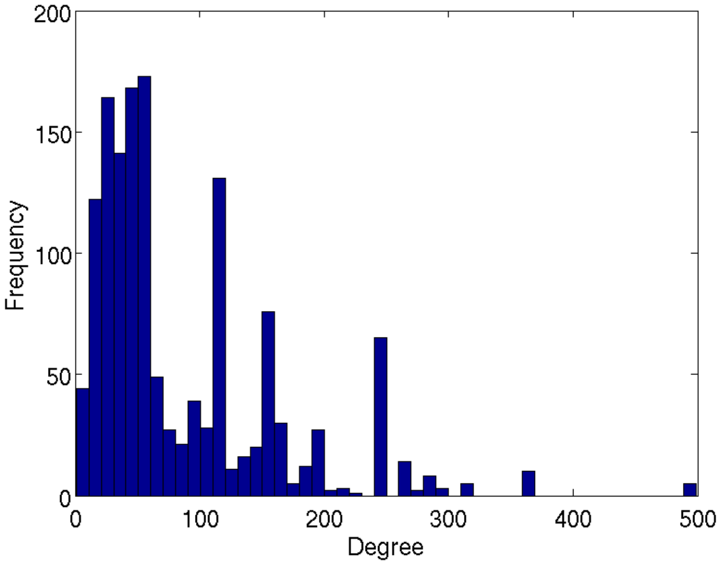

Figure 5.2: Histogram with 50 bins illustrating the degree distribution in the KEGG metabolic network.

The final network contains $N = 1422$ genes, according to the official gene symbol ID, with 5.10% of entries nonzero. The average (mean) nonzero entry is 1.162 and maximum (i, j) entry of 14.

5.3 MetaCyc: A Second Metabolic Network

We are also using an additional source of information to create a second metabolic network that will be used in parallel, allowing us to compare two efforts at cataloguing the human metabolome. This time the MetaCyc (http://www.metacyc.org/ [30]) database will provide the list of genes. The MetaCyc project contains a list of experimentally derived metabolic pathways and a corresponding list of genes. The MetaCyc database is also purposely limited to enzymatic, metabolic genes where the KEGG database contains genes outside this strict criteria of metabolism, thus using the two databases covers different perspectives on metabolism.

The MetaCyc database provides publically a gene ID (official gene symbol), associated pathways and reactions. There are 411 unique genes (using the official gene symbol, the gene ID also used for the KEGG network) spanning 187 pathways. The network is again undirected (symmetric) and contains 5106 edges - the maximum degree is 98. The same binary representation, as with the KEGG network, of the final network is shown in Figure 5.3. This structure contains $N = 411$ genes with 3.02% of entries nonzero. The average (mean) nonzero entry is 1.69 and maximum (i, j) entry of 11. A histogram of node degree is shown in Figure 5.4 - we see here that as with the KEGG case the distribution is overdispersed with a small number of high-degree nodes.

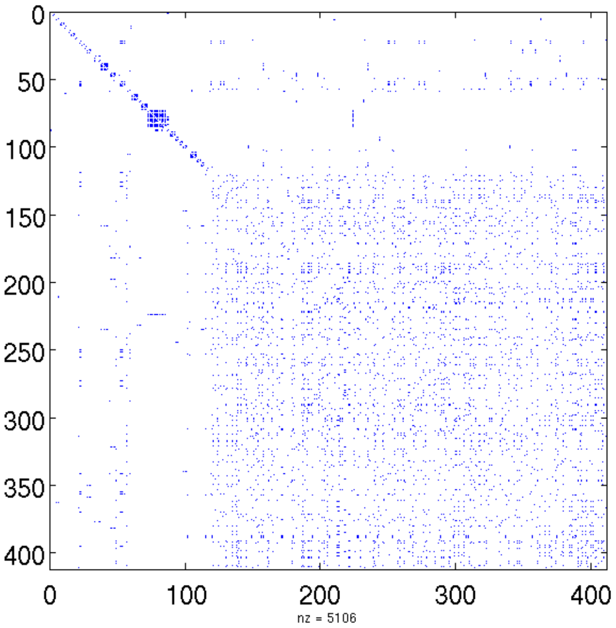

Figure 5.3: Binary representation of MetaCyc network: blue signifies the presence of an edge.

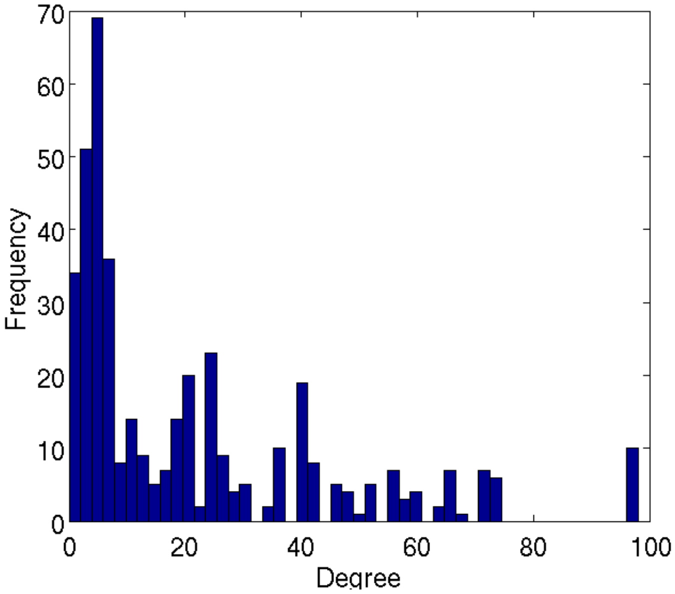

Figure 5.4: Histogram with 50 bins illustrating the degree distribution in the MetaCyc metabolic network.

5.4 Network Comparison: KEGG + MetaCyc

In this section we will see a brief comparison of vital statistics between the two networks. There are 311 genes in common, meaning the majority of genes within MetaCyc are also present in the KEGG database, other summary details are similarly displayed in Table 5.1.

	KEGG	MetaCyc
# Genes	1422	411
Pathways	158	187
Percentage nonzero	5.10%	3.02%
Mean nonzero	1.16	1.69
Max nonzero	14	11
Max Degree	499	98

Table 5.1: Summary information on KEGG and MetaCyc networks for comparison.

The two databases have significant overlap yet have a significantly different number of entries. One possible reason for this is with ambiguity in the official gene symbol - a difficulty that is not uncommon in genetics where gene sequences often have a large number of proprietary and public identifiers. For example, the KEGG database may include repeats of closely related gene symbols involving the same product where MetaCyc does not. In order to test this idea, we plot ordered (in degree) lists for both networks, in Figure 5.5 and Figure 5.6. Repeated or highly related gene symbols will have the same degree, and appear as flat lines on a plot of degree versus gene, which should make large groups visually identifiable. Figure 5.5 shows a significant amount of this structure, as well as small clusters (highlighted in green) with high degree. Figure 5.6 shows a relatively smooth increase in degree, in very similar pattern to the KEGG network. This similarity, along with the degree distribution histograms (Figures 5.2 and 5.4) and summary Table 5.1, shows that the networks have multiple features in common.

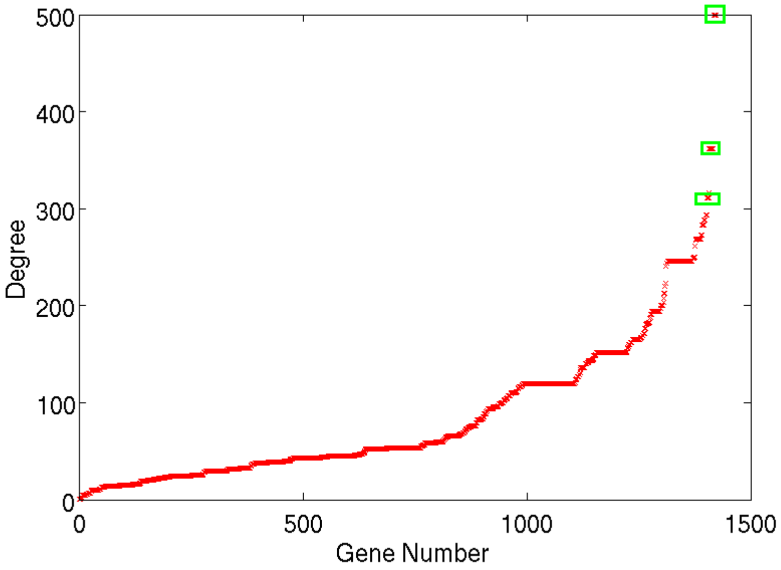

Figure 5.5: Degrees of KEGG network increasing order, highlighting examples of unlikely duplicate genes in green boxes.

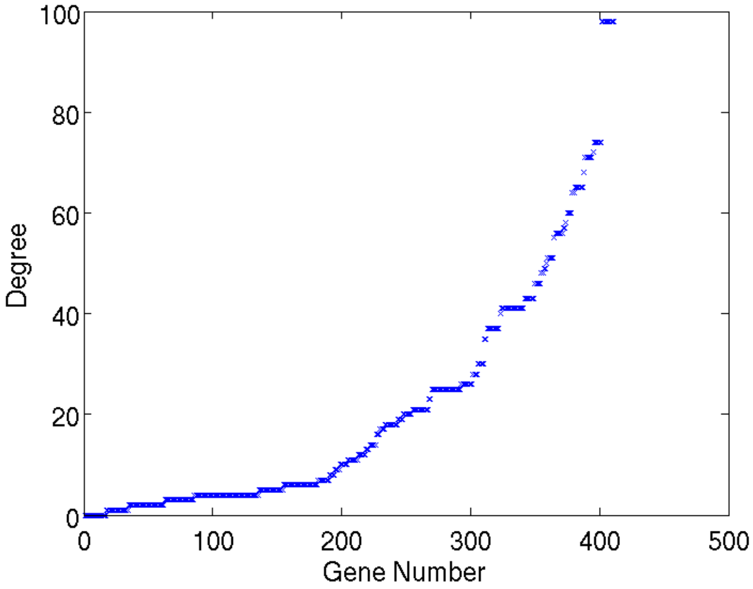

Figure 5.6: Degrees of MetaCyc network increasing order.

It would be time consuming with, considering our goals, little product to investigate each gene within these two databases. A simple observation to aid in better understanding the structures is instead to check the highest degree genes - since any overlap is less likely to be coincidental, given their separation from the rest of the group. To this end, Table 5.2 shows the top 15 genes for each network. In the KEGG list it is clear there is a lot of similarity, e.g., aldehyde dehydrogenases (*ALDH*) and UDP glucuronosyltransferases (*UGT*). In the MetaCyc list this is still present to a smaller extent with e.g., dihydrolipoamide S-acetyltransferase (*DLST*) and dihydrolipoamide S-succinyltransferase (*DLAT*).

Rank	Top KEGG OGS IDs	Degree	Top MetaCyc OGS IDs	Degree
1	*ALDH3A2*	499	*DBT*	98
2	*ALDH9A1*	499	*BCKDHB*	98
3	*ALDH7A1*	499	*BCKDHA*	98
4	*ALDH2*	499	*PDHN*	98
5	*ALDH1B1*	499	*DLAT*	98
6	*UGT1A3*	362	*PDHA1*	98
7	*UGT2B7*	362	*DLST*	98
8	*UGT2B4*	362	*PDHX*	98
9	*UGT2B28*	362	*OGDH*	98
10	*UGT2B17*	362	*DLD*	74
11	*UGT2B15*	362	*INPP5J*	74
12	*UGT2B11*	362	*SYNJ1*	74
13	*UGT12B10*	362	*INPP5B*	74
14	*UGT2A3*	362	*INPP5K*	74
15	*UGT2A1*	362	*OCRL*	74

Table 5.2: Top 15 genes in degree for both KEGG and MetaCyc.

This investigation into the construction of the networks is not necessarily important in isolation, but may be useful in terms of setting expectations for future use. For example, the KEGG network appears to have more significant overlap in closely related genes and products. This provides a slightly different perspective on future analysis - the KEGG network may prove more effective at identifying highly profilic gene products. Otherwise, the main difference between the two constructions is that the KEGG network has results with significantly more gene IDs. The other statistics provided in the summary in Table 5.1 are of similar order, which is expected given that they contain the same class of information (i.e.,metabolic). It is interesting that despite the lower number of genes, the MetaCyc database spans a greater number of pathways. Pathway IDs are not standardized, thus the number of identifiers depends on the protocol outlined by the creator. The greater number of pathways has visible effect on the degree distribution. As seen in comparing Figures 5.2 and 5.4 the KEGG network has many nodes at higher degrees than seen in the MetaCyc case. This is as expected; with a greater number of pathways the MetaCyc network is in effect more specialised: the genes are spread thinly across more specific metabolic pathways, decreasing the likelihood that they overlap.

5.5 Discussion

In this chapter we have seen the creation of two separate, but equivalent, novel metabolic networks, from different sources of data. There are numerous simplifications and assumptions made in the construction of these networks. For instance, a gene is indicated as having a product in a particular pathway or not but there is no indication of the ‘importance’ (since that is not easily quantifiable) of that particular gene product. In a similar way, all pathways are given the same degree of importance, whereas from observing visualisations of specific pathways on the KEGG database it is clear that the number of gene products per pathway varies greatly. Thus, no weight is given to critical (or not), larger or smaller pathways. However, the straightforward network-level representation of this information offers many conceptual and computational advantages.

A complete and unambiguous knowledgebase of the human metabolome would go some way to alleviating some of the shortcomings, but others are inherent in the process. Labelling of metabolic pathways is, in a sense, a subjective process, and their importance within an extremely complex system cannot be reliably rated. These facts mean it should be clear that these networks represent a very specific case of interest and are likely not suited to making general statements about metabolism. More specifically, the connections in this analysis are built on counting the total number of edges for each gene - this favours hub genes. The suitability of this as a metric depends on the desired output. Rather than containing data on magnitudes of reactions, our metabolic network presents a view of how well-connected the genes are.

The variation in size between these two networks may cause issue in later applications on real data. Our approach is therefore to add information from each of these networks, in turn, to specific cases of real data (as per the method outlined in the following chapter) - so the results from utilising the KEGG network can be compared with those from the MetaCyc network. The disparity in the number of genes presents a challenge - to combine the metabolic networks with real data the genes must match on both sides. So, real data will be truncated differently for each network, meaning that included experimental data will be different in addition to differences in the metabolic networks.

In summary, our aim here is to incorporate metabolic data in a systematic and quantifiable manner, while acknowledging the simplifications and approximations involved. We will show that, despite the limitations, the networks that we construct hold valuable information that can lead to new insights.

Chapter 6

The Node-Weighted Laplacian

This chapter proposes, investigates and validates a novel approach to combining multiple sources of data. Specifically, we are interested in establishing an alternative graph Laplacian that allows for integration of information from a second, related network. This will allow us to explore biological data in later chapters in a novel way. The work presented in this chapter (and some of Chapter 5) is the result of the following publication:-

Martin McDonald, Desmond J Higham, J Keith Vass. Spectral algorithms for heterogeneous biological networks, Briefings in functional genomics, Vol. 11 No. 6, 2012 [146]

6.1 Spectral Methods and Graph Laplacians

In this chapter, in order to appropriately frame the problem, we revisit, expand and elaborate on some previously introduced elements of spectral methods. The work developed here will be tested and used in conjunction with the metabolic networks of Chapter 5.

In this section we motivate and explain how the Laplacian and normalized Laplacian can be used to find structure in a network. The next subsection introduces a key result from linear algebra: we refer to [85] for more information.

6.1.1 Rayleigh-Ritz Theorem

The following lemma, which is a special case of the Rayleigh-Ritz Theorem [92, Theorem 4.2.2], will be used to justify the spectral algorithms in the following sections.

Lemma 1 *Let* $M \in \mathbb{R}^{N\times N}$ *be a symmetric positive semi-definite matrix with eigenvalues* $0 = \gamma_1 < \gamma_2 < \gamma_3 \leq \gamma_4 \leq \cdots \leq \gamma_N$*, and corresponding eigenvectors* $\mathbf{r}^{[1]}, \mathbf{r}^{[2]}, \ldots, \mathbf{r}^{[N]}$*. Then the problem*

$$\begin{array}{cc} \min & \mathbf{y}^T M \mathbf{y} \\ \mathbf{y} \in \mathbb{R}^N & \\ \mathbf{y}^T \mathbf{r}^{[1]} = 0 & \\ \mathbf{y}^T \mathbf{y} = 1 & \end{array} \tag{6.1}$$

is uniquely solved by $\mathbf{y} = \mathbf{r}^{[2]}$*.*

Proof. The matrix M has the spectral decomposition $M = R\Gamma R^T$, where $\Gamma \in \mathbb{R}^{N\times N}$ is diagonal with (i,i)th entry γ_i and $X \in \mathbb{R}^{N\times N}$ has jth column $\mathbf{r}^{[j]}$. The eigenvectors are mutally orthogonal, so we may take $R^T R = I$. Letting $\mathbf{z} = R^T \mathbf{y}$, the problem (6.1) becomes

$$\begin{array}{cc} \min & \mathbf{z}^T\Gamma\mathbf{z}. \\ \mathbf{z}\in\mathbb{R}^N & \\ \mathbf{z}^T R\mathbf{r}^{[1]}=0 & \\ \mathbf{z}^T\mathbf{z}=1 & \end{array}$$

The constraint $\mathbf{z}^T R\mathbf{r}^{[1]}=0$ simplifies to $z_1=0$, so the problem becomes

$$\begin{array}{cc} \min & \sum_{i=2}^{N}\gamma_i z_i^2. \\ \mathbf{z}\in\mathbb{R}^N & \\ \mathbf{z}^T\mathbf{z}=1 & \end{array}$$

Since $\gamma_2<\gamma_3\leq\gamma_4\leq\cdots\leq\gamma_N$, it is clear that $z_2=1$ and $z_i=0$ for $i=3,4,\ldots,N$ uniquely solves the problem. Hence, we have $\mathbf{y}=\mathbf{r}^{[2]}$ as required.

6.1.2 Clustering and Reordering

We have applied the SVD and GSVD as methods of spectral clustering in Chapters 3 and 4. Here we outline and expand on the mechanics of this approach to give context to the new approach formed later in the chapter. Let $A\in\mathbb{R}^{N\times N}$ be a symmetric matrix with non-negative entries. From a network perspective, we think of $a_{ij}=a_{ji}\geq 0$ as representing the pairwise similarity between nodes i and j, where a larger value indicates a greater similarity.

Suppose we wish to divide the vertices in two disjoint *clusters*, where a pair of nodes within a cluster are typically well-connected and a pair of nodes in different clusters are not. One way to judge the quality of a partition is to count the total weights in the edges that span the two clusters. Introducing the indicator vector $\mathbf{y}$, so that $y_i = -\frac{1}{2}$ if node i is in one set and $y_i = \frac{1}{2}$ if node i is in the other, the total weight across the clusters may be written

$$\tfrac{1}{2} \sum_{i,j} (y_i - y_j)^2 a_{ij}. \tag{6.2}$$

In matrix-vector form, this expression becomes

$$\mathbf{y}^T (D - A) \mathbf{y}, \tag{6.3}$$

where $D \in \mathbb{R}^{N \times N}$ is the diagonal degree matrix with $D_{ii} = \text{deg}_i$, and $\text{deg}_i := \sum_j a_{ij}$ is the degree of node i. Asking for $\mathbf{y}$ to minimize this quantity is not reasonable, since it leads us to the trivial solutions $y_i \equiv \frac{1}{2}$ and $y_i \equiv -\frac{1}{2}$; that is, put all nodes into a single cluster. It therefore makes sense to add a balancing constraint that limits the mismatch between cluster sizes. In general, however, it is not feasible to tackle the discrete problem (6.3) directly, and hence it is standard practice to allow the y_i to take any real values; thereby *relaxing* the problem. Using $\mathbf{y} \in \mathbb{R}^N$, a suitable balancing constraint is $\mathbf{y}^T \mathbf{1} = 0$, where $\mathbf{1} \in \mathbb{R}^N$ is the vector with all entries equal to one, and in order to avoid the trivial solution $y_i \equiv 0$, we add the extra constraint $\mathbf{y}^T \mathbf{y} = 1$. This leads us to the optimization problem

$$\min_{\substack{\mathbf{y} \in \mathbb{R}^N \\ \mathbf{y}^T \mathbf{1} = 0 \\ \mathbf{y}^T \mathbf{y} = 1}} \quad \mathbf{y}^T (D - A) \mathbf{y}. \tag{6.4}$$

As we discuss further in subsection 6.1.3, Lemma 1 shows that this problem can be solved via a spectral decomposition; that is, by computing appropriate eigenvectors and eigenvalues.

At this stage, it is worth pointing out that after the relaxation step, where we move from $y_i \in \{-\frac{1}{2}, \frac{1}{2}\}$ to $\mathbf{y} \in \mathbb{R}^N$, we are in the realm where each node is assigned a position on the real line. We can recover clusters by picking a threshold, such as 0, and assigning nodes to the same cluster if they lie on the same side of the threshold. However, rather than interpreting (6.4) as a problem that approximates a discrete analogue, we could use it as starting point, and take the viewpoint that nodes are being mapped to points on the real line in such a way that nearby nodes are well-connected. Because the solution of (6.4) may be expressed in terms of a spectral decomposition, this idea may be taken further. Using the fact that the power method iteration converges to a dominant eigenvector, we may argue that solving (6.4) is equivalent to placing the nodes on the real line in random locations and then iteratively "shuffling" them, based on their pairwise affinities, until an equilibrium state is reached; see [77] for details.

Rather than taking a hard clustering approach through thresholding, it is also possible to use the real-valued solution $\mathbf{y}$ to relabel the nodes. In this way a permutation vector $\mathbf{p} \in \mathbb{R}^N$ is constructed, whose components consist of the integers from 1 to N, so that node i gets mapped to position p_i, with

$$p_i \leq p_j \quad \Leftrightarrow \quad y_i \leq y_j. \tag{6.5}$$

In words, $\mathbf{y}$ places the nodes on the real line, and we relabel them according to their position, the left-most becomes node 1 and the right-most becomes node N. Returning to the matrix interpetation of the data set A, we have equivalently performed a symmetric permutation that reorders the rows and columns of the matrix. Viewing the reordered matrix is often a very useful way to visualize interesting patterns in the data [75, 85, 103, 220], as we have shown previously in e.g., Chapter 3.

6.1.3 Graph Laplacian

The matrix $L = D - A \in \mathbb{R}^{N \times N}$ appearing in (6.4) is known as the *Graph Laplacian* matrix for the network. This symmetric positive semi-definite matrix has smallest eigenvalue 0 and corresponding eigenvector $\mathbf{1}$. We suppose that the network is connected (every pair of nodes may be joined by at least one set of edges with non-zero weights), in which case all other eigenvalues of the Laplacian are positive; see, for example, [45, 211]. We also suppose that there is a unique smallest non-zero eigenvalue, and order the eigenvalues so that $0 = \lambda_1 < \lambda_2 < \lambda_3 \leq \cdots \leq \lambda_N$. We denote the corresponding eigenvectors $\mathbf{v}^{[1]}, \mathbf{v}^{[2]}, \ldots, \mathbf{v}^{[N]}$. These are orthogonal, and we assume that they have Euclidean norms of unity. The eigenvector $\mathbf{v}^{[2]}$ corresponding to the first non-zero eigenvalue of the Laplacian plays an important role in many areas of graph theory and network science, and is referred to as the *Fiedler vector* [61]. It now follows from Lemma 1 that the solution of the relaxed problem (6.4) is given by the Fiedler vector, $\mathbf{v}^{[2]}$. Note that in cases where eigenvalues are similar e.g., $\lambda_2 \approx \lambda_3$ we may need to look in more than one direction to get a good description of the data.

6.1.4 An Alternative Form of Clustering and Reordering

Next we note that the constraint $\mathbf{y}^T\mathbf{1} = 0$ in (6.3) aims to balance the number of nodes in each group. As an alternative, we may wish to quantify the size of each node i in terms of its degree and aim to balance the overall size of the clusters. An appropriate balancing constraint is then $\mathbf{y}^T D\mathbf{1} = 0$. Further, rather than normalizing with $\mathbf{y}^T\mathbf{y} = 1$, so that all nodes are treated equally in terms of distributing the locations on the real axis, we may prefer $\mathbf{y}^T D\mathbf{y} = 1$, which enourages high degree nodes to be placed nearer the origin. From the reordering viewpoint, this may be interpreted as an attempt to reduce the influence of "promiscuous" nodes, encouraging them away from the extremes of the ordering range. These issues of calibration can be important when there is a high degree of variance among the interaction weights, a circumstance that is common for gene expression data. These two changes convert the relaxed problem (6.4) to

$$\begin{array}{cl} \min & \mathbf{y}^T (D - A) \mathbf{y}. \\ \mathbf{y} \in \mathbb{R}^N & \\ \mathbf{y}^T D \mathbf{1} = 0 & \\ \mathbf{y}^T D \mathbf{y} = 1 & \end{array} \tag{6.6}$$

Changing variable to $\mathbf{x} = D^{\frac{1}{2}} \mathbf{y}$, this problem becomes

$$\begin{array}{cl} \min & \mathbf{x}^T D^{-\frac{1}{2}} (D - A) D^{-\frac{1}{2}} \mathbf{x}, \\ \mathbf{x} \in \mathbb{R}^N & \\ \mathbf{y}^T D^{-\frac{1}{2}} \mathbf{1} = 0 & \\ \mathbf{x}^T \mathbf{x} = 1 & \end{array} \tag{6.7}$$

where we make the reasonable assumption that all node degrees are non-zero.

6.1.5 Normalized Graph Laplacian

The matrix $D^{-\frac{1}{2}} (D - A) D^{-\frac{1}{2}} \in \mathbb{R}^{N \times N}$ appearing in (6.7) is known as the *Normalized Graph Laplacian.* Like the (unnormalized) Laplacian in subsection 6.1.3, this symmetric positive semi-definite matrix has an eigenvalue 0 and, in the case of a connected graph, a unique smallest nonzero eigenvalue. The eigenvalues lie in the interval $[0, 2]$, see, for example, [211], and we label them $0 = \mu_1 < \mu_2 < \mu_2 \leq \cdots \leq \mu_N$, with corresponding eigenvectors $\mathbf{w}^{[1]}, \mathbf{w}^{[2]}, \ldots, \mathbf{w}^{[N]}$. By construction, we have $\mathbf{w}^{[1]} = D^{\frac{1}{2}} \mathbf{1} / || \, D^{\frac{1}{2}} \mathbf{1} \, ||$.

We refer to $D^{-\frac{1}{2}}\mathbf{w}^{[2]}$ as the *normalized Fiedler vector.* Lemma 1 now shows that the relaxed problem (6.7) is solved by $\mathbf{x} = \mathbf{w}^{[2]}$ and hence the required solution of (6.6) is the normalized Fiedler vector $\mathbf{y} = D^{-\frac{1}{2}}\mathbf{w}^{[2]}$; see [85, Corollary 1] for further details.

At this stage it is worth making a few points about the spectral approach.

1. Eigenvalues and eigenvectors are invariant under permutation, in the sense that
$$A\mathbf{x} = \lambda\mathbf{x} \quad \Leftrightarrow \quad PAP^T\mathbf{x} = \lambda P\mathbf{x}$$
for $\mathbf{x} \in \mathbb{R}^N$, $\lambda \in \mathbb{R}$ and any permutation matrix of both rows and columns [1] $P \in \mathbb{R}^{N\times N}$. It follows that spectral algorithms are oblivious to the way that nodes are labelled—for example, relabelling the nodes simply reorders the elements of the Fielder vector accordingly. As a consequence, when we test spectral algorithms on synthetic data where known structures have been deliberately created, it is reasonable to label the nodes of A in any convenient manner.

2. Whether we use the vector $\mathbf{y} \in \mathbb{R}^N$ for hard clustering or for reordering, it is clear that we should be unconcerned about two types of transformation

   translation: where $y_i \mapsto y_i + c$, for a constant c that is independent of i,

   rescaling: where $y_i \mapsto \alpha y_i$, for a constant $\alpha \neq 0$ that is independent of i.

   In particular, the map $\mathbf{y} \mapsto -\mathbf{y}$ coincides with relabelling the two clusters or to reversing the node ordering, and we note that eigenvectors are uniquely defined only up to a ± factor.

3. Using the translation and scaling operations above, we can show that the same Fiedler vector solutions arise for a very wide range of balancing constraints—

[1] A permutation matrix is found by permuting the rows of an identity matrix according to some permutation of the numbers 1 to N.

we do not need to ask for exactly equal cluster sizes in the original discrete formulation; see [85].

4. Because a symmetric matrix has orthogonal eigenvectors, moving beyond the Fiedler cases and using $\mathbf{v}^{[3]}$, $\mathbf{v}^{[4]}$, ... and $D^{-\frac{1}{2}}\mathbf{w}^{[3]}$, $D^{-\frac{1}{2}}\mathbf{w}^{[3]}$, ... to cluster or reorder the data can reveal further information about the data; see [85].

6.1.6 Singular Value Decomposition (SVD)

In the case of a bipartite network, we have two separate groups of nodes and the weight a_{ij} represents the pairwise affinity between node i in the first group and node j in the second group. If the groups contain M and N nodes respectively, then $A \in \mathbb{R}^{M\times N}$. Spectral information is now contained in the Singular Value Decomposition (SVD)

$$A = U\Sigma V^T,$$

where $U \in \mathbb{R}^{M\times M}$ and $V \in \mathbb{R}^{N\times N}$ are orthogonal and $\Sigma \in \mathbb{R}^{M\times N}$ is diagonal with diagonal elements $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$. The columns of U and V are referred to as the *left* and *right singular vectors* of A, respectively.

Analogously to the development in section 6.1.2, we may introduce two indicator vectors, $\mathbf{p} \in \mathbb{R}^M$ and $\mathbf{q} \in \mathbb{R}^N$, and consider the quantity

$$\tfrac{1}{2}\sum_{i,j}(p_i - q_j)^2 a_{ij}. \tag{6.8}$$

After adding appropriate constraints and relaxing to real-valued vectors $\mathbf{p}$ and $\mathbf{q}$, it may be shown that the left and right singular vectors of A can be used to reorder the two groups of nodes. Similarly, the SVD of the normalized data $D_{\text{out}}^{-\frac{1}{2}} A D_{\text{in}}^{-\frac{1}{2}}$ arises if we generalize the $\mathbf{y}^T D\mathbf{1} = 1$ alternative in (6.6). We refer to [103] for full details.

We note that the left and right singular vectors of $A \in \mathbb{R}^{M \times N}$ are equivalent to the eigenvalues of $A^T A$ and AA^T, respectively, and this forms a natural bridge to the methods described in subsections 6.1.2–6.1.4. For example, we may regard the operation of forming $A^T A$ as correlating across the second group of nodes in order to form a pairwise affinity matrix for the first group. A spectral method could then be applied directly to $A^T A$ in order to cluster or reorder the first group.

6.2 Formulation

In Section 1.3.2 we motivated spectral methods by setting up appropriately constrained optimization problems. This approach offers a lot of flexibility, a fact that we now exploit to derive an alternative Laplacian style matrix. A network is said to be *assortative* if connections are more likely between nodes of similar degree (where the degree of a node is the number of edges that are connected to it) [57, 166]. Many authors have considered the issue of quantifying the overall level of assortativity in a network, relative to some null model [165, 167]. However, here we consider an inverse problem that also has practical relevance—given a network, can we identify specific patterns of assortativity? More precisely, can we find a set of nodes that

(a) form a strong cluster, and

(b) possess similar degrees?

Where a strong cluster is a collection of nodes that shows significance in terms of weight density, as tested for by the cluster quality measure (as described in Section 1.4.1). We note that this is a partially local concept—it is possible for a *substructure* of this type to be present in a network that is not categorized as being assortative by a global measure. We also note that this type of substructure has a very natural generalization; the condition **(b)** could be extended to the case where nodes possess an independent measure of "size" and we seek clusters that involve nodes of comparable size.

We therefore suppose that a positive weight w_i is associated with each node i. In order to look for nodes that are well-connected and size-compatible, we may replace the starting point (6.2) with

$$\tfrac{1}{2}\sum_{i,j}(\sqrt{w_i}y_i - \sqrt{w_j}y_j)^2 a_{ij}. \tag{6.9}$$

Letting $D_w \in \mathbb{R}^{N\times N}$ denote the diagonal matrix with iith entry w_i, this expression may be written

$$\mathbf{y}^T D_w^{\frac{1}{2}} \left(D - A\right) D_w^{\frac{1}{2}}\mathbf{y}. \tag{6.10}$$

Here we emphasize that D is the original diagonal degree matrix arising from the data matrix, but the diagonal matrix D_w may contain any appropriate set of nodal weights.

To focus on the case where we prioritize nodes with large weights, we take $\mathbf{y}^T D_w^{-1}\mathbf{y} = 1$ as our normalizing constraint. This encourages the highly weighted nodes to take values at the extreme ends of the range of y_i values. Changing variable to $\mathbf{z} = D_w^{-\frac{1}{2}}\mathbf{y}$, the expression (6.10) then becomes

$$\mathbf{z}^T D_w \left(D - A\right) D_w\mathbf{z}, \tag{6.11}$$

with $\mathbf{z}^T\mathbf{z} = 1$. We will refer to the matrix

$$L_w := D_w \left(D - A\right) D_w \tag{6.12}$$

appearing in (6.11) as the *node-weighted Laplacian*. By construction, L_w has a zero eigenvalue with corresponding eigenvector $D_w^{-1}\mathbf{1}$. This vector depends only on w-

information; it ignores the network connectivity. Hence, by analogy with the Fiedler vector and normalized Fiedler vector approaches, we propose to reorder/cluster for this generalized notion of assortativity in terms of $\mathbf{x}^{[2]}$, the eigenvector of L_w corresponding to the smallest positive eigenvalue. From Lemma 1, this becomes the required minimizer of (6.11) if we add the balancing constraint[2] $\mathbf{z}^T D_w^{-1}\mathbf{1} = 0$. Converting back to $\mathbf{y} = D_w^{\frac{1}{2}}\mathbf{z}$, we therefore propose to take $D_w^{\frac{1}{2}}\mathbf{x}^{[2]}$ as our network reordering vector. The first step to verifying this new algorithm is to verify performance on a synthetic data set: an artificially manufactured network that is constructed to specifically contain the type of structure we are hoping to identify. After dealing with this basic test, the new algorithm is tested on real data where we have an expectation as to how the algorithm will perform.

6.3 Synthetic Testing

We now examine the behaviour of the new normalisation of the graph Laplacian on a synthetic network that is designed to contain nodes with the desirable property. The network, A, shown in Figure 6.3 contains $1,000$ nodes with random weights between $0-100$ on all edges a_{ij} (though A is made to be symmetric so $a_{ij} = a_{ji}$). The aim of this synthetic test is to examine whether or not the new normalisation for the graph Laplacian can uncover clusters whose nodes also have high degrees, in general. To create this structure, we force nodes $1-100$ to become a strong cluster - the edges between these nodes are given weight 100. Then, in order that the degrees of the nodes in this cluster are assured to be higher than average, the edges between this cluster and the rest of the network are boosted by an increase in weight of 50. To make the test more realistic, another cluster is introduced to the data, which instead is comprised of nodes that have generally low overall degrees,

[2] In terms of the original variable, $\mathbf{y}$, this balancing constraint is $\mathbf{y}^T D_w^{-\frac{3}{2}}\mathbf{1} = 0$, which further encourages highly weighted nodes away from the origin.

as is obvious in the heat map in Figure 6.3. For this, nodes $101 - 200$ are clustered in the same way as $1 - 100$ - each edge in this group is given weight 100. Then the weights between this second cluster and the rest of the network are decreased by 50.

It is also worth noting that the network in Figure 6.3 is purposefully ordered such that the structures are readily visible - this is done only as a matter of convenience. As mentioned in Section 1.3.2, the algorithm is invariant to the initial ordering of the data - the result will be the same regardless of how the data might be permuted - so we simply choose one that allows for a simple visual assessment of results.

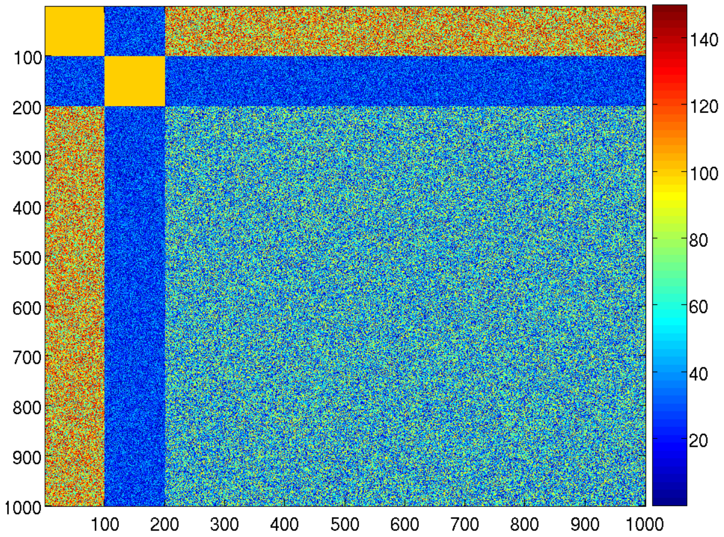

Figure 6.1: Weight matrix for the synthetic network 'A' showing both well and poorly connected clusters.

Now, as per Section 6.2, we calculate the node-weighted Laplacian (equation 6.12). In this case, $D_w = \sum_{i=1}^{N} a_{ij}$ - the degrees of the nodes in the network A. The next step is to examine the components of the new suggested ordering vector $D^{\frac{1}{2}}x^{[2]}$ and compare these results to the components from the Fiedler vector of the unnormalised Laplacian, $u^{[2]}$, and the components of normalised Fiedler vector from the traditional normalised Laplacian, $D^{-\frac{1}{2}}w^{[2]}$. Success for the algorithm would result in components of the singular vectors following the distribution of the node degrees.

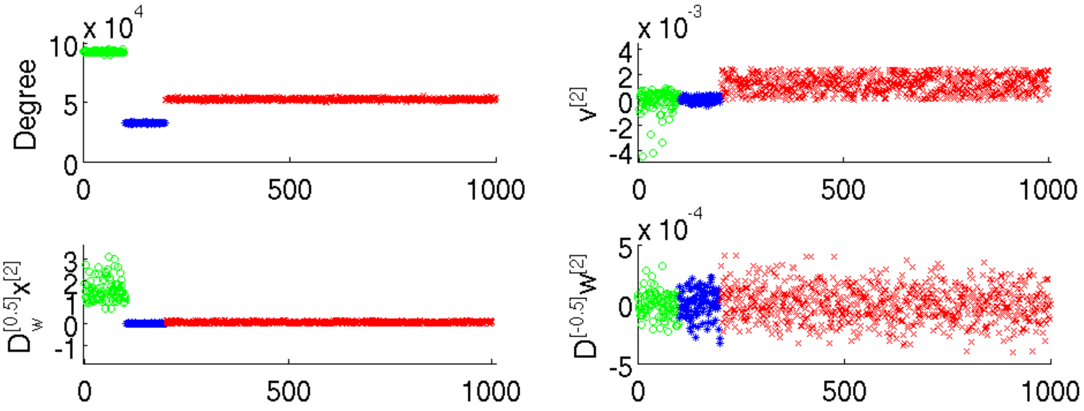

Figure 6.2: The top left picture shows the weights of the nodes in the synthetic network. The remaining three pictures show the results for the Fiedler vectors from the various approaches.

In Figure 6.2, we highlight the high-degree cluster by using the 'o' marker and the second cluster that has low degree nodes with the '$*$' marker. The desired structure was identified clearly with the new normalised Laplacian $D^{\frac{1}{2}}x^{[2]}$, whereas both clusters were identified in a single set by $u^{[2]}$. The normalised Laplacian $D^{-\frac{1}{2}}w^{[2]}$ had the effect of supressing nodes from the high degree structure and separating those in the low degree cluster. Degree is plotted in this figure to give a reference point since we expect the singular vectors will be impacted by this information.

In order to visualise what the placement of nodes in these real valued vectors means for the recovery of information from our original network A, we assign all nodes new IDs based on their position in the sorted vector (Section 1.4 has details on data reordering). If we then use a heat map to output the reordered data, we find an intuitive image where clusters will be represented as significant areas of hot or cold (i.e. high or low) values. Figure 6.3 shows the initial network after reordering by each of the three approaches. The image on the left is from the unnormalised case and, as we expected from previous examination of the vectors, the unnormalised

Fiedler vector $v^{[2]}$ fails to separate the two clusters - instead forming one mixed grouping. In the middle we have the new node-weighted Laplacian approach - which completely recovers the desired structure from Figure 6.3. The last plot in Figure 6.2 from the traditional normalised Laplacian is perhaps more difficult to interpret from vector alone - from this image (Figure 6.3) we can see that the result of the reordering is scattered such that we have lost most of the information present in the initial network.

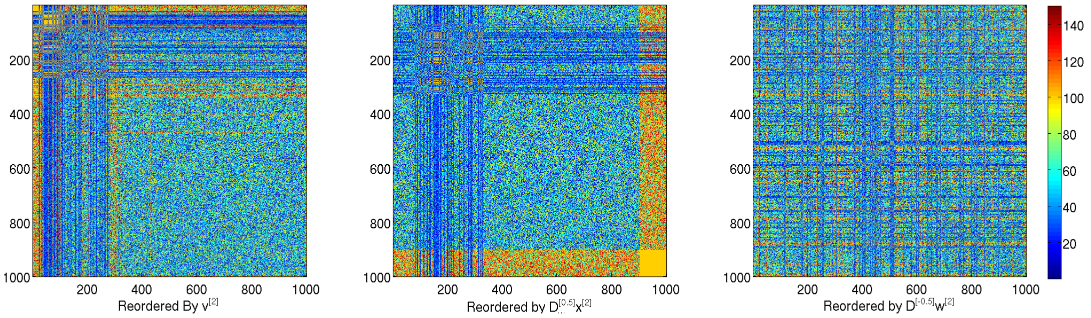

Figure 6.3: The left picture shows synthetic network reordered by $v^{[2]}$, middle by node-weighted Laplacian $D^{[0.5]}x^{[2]}$ and right by traditional Laplacian $D^{[-0.5]}w^{[2]}$.

6.3.1 Merging Two Data Sets: Synthetic Testing

This next section involves an extension of the idea to the case where the components of D_w are from a matrix independent of A - that is, we introduce a second data set that spans the same set of nodes. The proposal is that this time D_w is given a range of values in order to test how much of the ordering of node-weighted Laplacian is influenced by structure in A versus weights in D_w. A is constructed as in the previous example, and the first 50 values in D_w are given a high weight 20, the next 50 given a low weight 1. This pattern of 50 high and 50 low is repeated for the next 200 nodes, giving D_w and results shown in Figure 6.4.

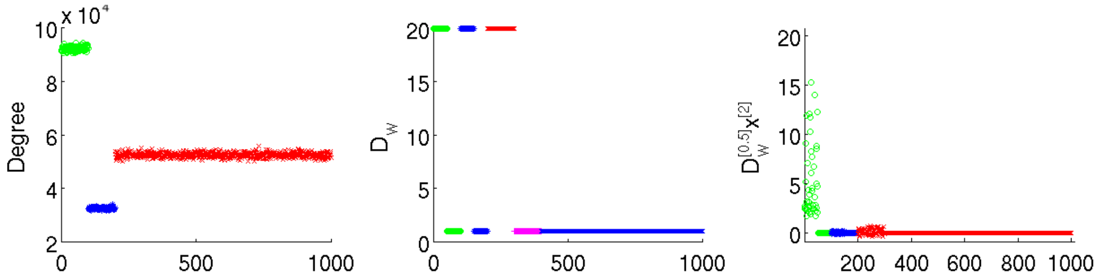

Figure 6.4: Synthetic network, original node ordering. Left: nodal degrees. Middle: components of D_w for the node-weighted Laplacian. Right: components of the vector $D_w^{\frac{1}{2}}\mathbf{x}^{[2]}$ arising from the node-weighted graph Laplacian.

The left picture in Figure 6.4 shows the original order of the nodes with degree. The middle picture shows the new values we are using in D_w, from a second synthetic network. The right picture shows the result from the node-weighted Laplacian. In this case, we can see that the node-weighted Laplacian clearly separates the first 50 nodes - those that were well-connected in matrix A and have high values in D_w. The 50 nodes that are well-connected but have low values in D_w are not separated - illustrating the fact that the result of the node-weighted Laplacian is a combination of both the information in the original network, and the values in the rescaling vector D_w. In addition to Figure 6.4, we also show the network reordered according to the normalised Fiedler vector, in Figure 6.5. We see here the strongly-weighted/high D_w nodes are pushed to the end of the ordering. The strongly-weighted/low D_w nodes have been pushed away from the end.

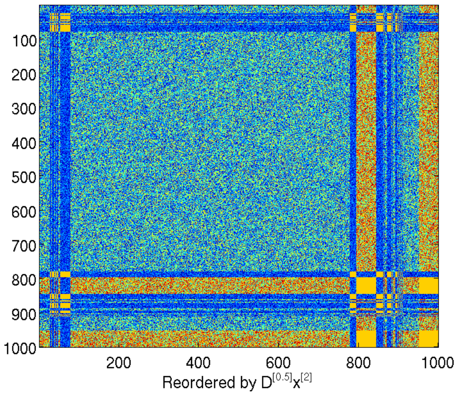

Figure 6.5: Heatmap for network reordering applied to synthetic network in Figure 6.3 with the node-weighted Laplacian.

In summary, we have shown that the output from the node-weighted Laplacian is influenced by both the original network and the values used for the components of D_w. This new approach has then been shown to hold value in identifying assortive structures within a single network, as well as providing a new method of combining information when multiple edge sets exist for a single pair of nodes. There are many possible uses for this method of incorporating extra-information into an analysis, in the following section we will explore an application pertinant to this thesis - a combination of metabolic and microarray networks. This will allow a specific type of analysis that brings together different elements of this project so far - we have previously created a pair of metabolic networks (see Chapter 5) from publically available databases, and in later chapters we will examine microarray data as an exploratory tool for features in schizophrenia.

6.4 Node-Weighted Laplacian: Microarray Application

For this section we will use one publically available microarray data set, adipose tissue measurements from decode study GSE7965, for 296 samples [56]. This study contained a cohort of 701 individuals - though this data set contained both male and female test subjects, for this study we selected only the male contingent, which is 296 samples. The reason for this is that there are significant levels of variance between the male and female cohorts - and we wish to examine content of the gene expression from the adipose tissue, rather than discovering a list of genes that may differentiate between males and females. Additionally, specific interpretation is not key to this chapter - we are initially testing the node-weighted Laplacian on real data.

These data may be regarded as a rectangular array whose (i, j) entry records the expression level of gene i in sample j, for $23,765$ genes across 296 samples. We use the absolute value of the expression data, so that all data entries are non-negative; with this approach we treat under and over-expression as equivalent, on the grounds that both indicate a deviation from basal behaviour, [112]. Hence, a larger weight is taken to denote a higher level of activity. It is, of course, possible to retain the distinction between under/overexpression using a signed network [87], as we will do later when specific analysis is important.

We make one other modification to the data based on the metabolic network we will use later. In order to make the two networks compatible, and form a node-weighted Laplacian, they must have the same number of genes. We then take the KEGG metabolic network formed in Chapter 5 and check which genes have been measured in the probe set of this microarray experiment. Note at this point that probes in microarray kits commonly measure only a portion of the sequence for an individual gene - there is often overlap such that multiple probes measure the same gene. We calculate the degrees in the metabolic network, and repeat assign the degree to each relevant probe.

Overall, then, in this section we have a non-negative real valued microarray data set size 4567 by 296 and a non-negative integer valued 4567 by 4567 array of KEGG metabolic pathway co-incidence data.

6.4.1 SVD: The Rectangular Case

We perform an initial analysis of the microarray data set using the aforementioned reordering for clustering from the SVD, without the node-weighted Laplacian. Preserving the rectangular form of the data means that we maintain the ability to simultaneously reorder the samples as well as the genes (as with [86, 112]). In the language of Section 1.3.2, the matrix A is the rectangular array of microarray data with genes and patients as the rows and columns. Then, the matrices U and V provide the left and right singular vectors - preserving the rectangular form of the data gives us the ability to reorder the samples as well as the genes (where in Chapters 3, 4 we acted on symmetric matrices); this type of bi-clustering is commonly performed on microarray data [103, 112]. Although this SVD is acting purely on the microarray data, in order to maintain consistency between analyses we also trim this data set as described previousy, leaving a rectangular $N \times M$ matrix that contains $N = 4567$ genes that that are also present in the metabolic analysis later, and $M = 296$ samples.

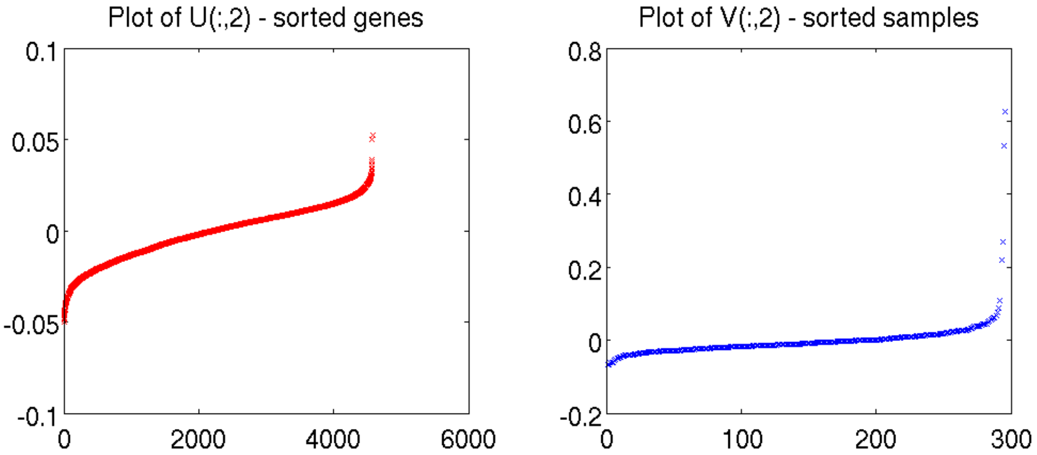

Figure 6.6: The Fiedler vectors for each of the two dimensions of the microarray data - also known as eigengenes and eigensamples.

The gene that codes for the hormone leptin is of particular interest as a measure in this adipose tissue data set. The reason for this is that leptin resistance is a good indicator for obesity - so it should be expected that there is a pattern for leptin levels in a data set of this type [24]. In our reordered data, the gene that codes for leptin appears at the end of the ordering vector ($u^{[2]}$) from the SVD - identifying the leptin gene as key to explaining variance in the data set. We can check the levels of leptin across samples (ordered by $v^{[2]}$), in Figure 6.4.1. As expected, there is a trend when checking the expression levels of the leptin-gene in the ordered sample list. This shows that the clustering identified thorugh reordering the sample list roughly follows a trend in the values of the leptin gene. This is a good check for the analysis since we have prior information about the impact leptin would be expected to have on data of this type. We also note that in this, and following sections within this chapter, we do not show a reordered plot of the full data set. This is because in a data set, such as this, with thousands of rows it can be difficult to identify any structure visually.

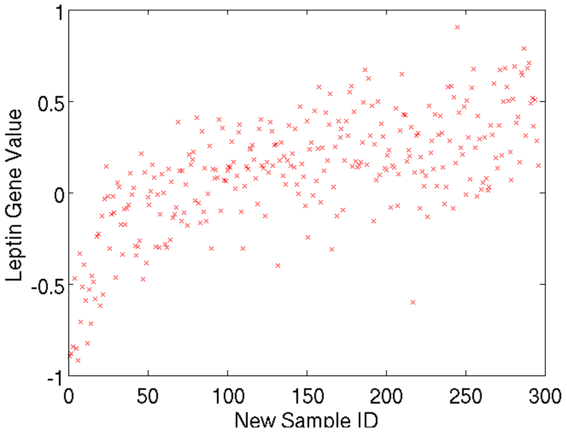

Figure 6.7: Leptin expression level from simultaneous ordering of genes and samples using $u^{[2]}$ and $v^{[2]}$ showing leptin variation per patient.

6.4.2 KEGG Metabolic Network Application

Having shown that the SVD can produce biologically meaningful results on this data we next show the case where the KEGG metabolic network is used to provide the node weights in the nodeweighted Laplacian. For the nodeweighted Laplacian we use $L_b = D_b(D - AA^T)D_b$ where B is the KEGG metabolic network (see Chapter 5 for details), components b_{ij}. So, $D_b = \sum_{i=1}^{N} b_{ij}$ - and the node-weighted Laplacian will re-scale the weights of each node in a network A based on information gleamed from a separate network B. We treat this metabolic construction as additional infromation for the microarray problem. Overall, we have (a) a non-negative real-valued 4567 by 296 array of gene expression data, and (b) a non-negative integer-valued 4567 by 4567 array of metabolic pathway co-incidence data (with 4.4% percent of entries nonzero, mean nonzero entry is 1.35 and maximum entry is 45).

In this test, we are now seeking to uncover structures that are well-connected within the microarray network A but also have a high degree in metabolic network B. The viewpoint we take here is that we are enhancing the presence of the genes that have high degrees in the metabolic network, and are thus selecting for clusters in the microarray network that are composed of genes that are also strongly active in a metabolic sense. A sensible first step after producing the node-weighted Laplacian and calculating the reordering vectors using the SVD is to check if the resultant ordering vector produces any visible structure when used to reorder the vector of metabolic degrees. The expectation is that if the metabolic degrees are having an influence on structure within the microarray data, the reordering vector will have a tendency to push high degree nodes towards either end of the vector.

In the left of Figure 6.8, we show the degree in the metabolic network, D_b for the genes, when ordered by the vector $\mathrm{D}_w^{\frac{1}{2}} x^{[2]}$ from the node-weighted Laplacian, does indeed give preference to genes with high metabolic weights - placing them at the extremes of the list. As a minor verification we also produce the right hand picture which, by contrast, using the Fiedler vector arising from the Laplacian matrix, does not incorporate any metabolic information. Naturally, in this case, we do not see any metabolic pattern. This is a basic check to ensure it was not simply the case that high degree genes in the KEGG network happen to be more important in the structure of the microarray data.

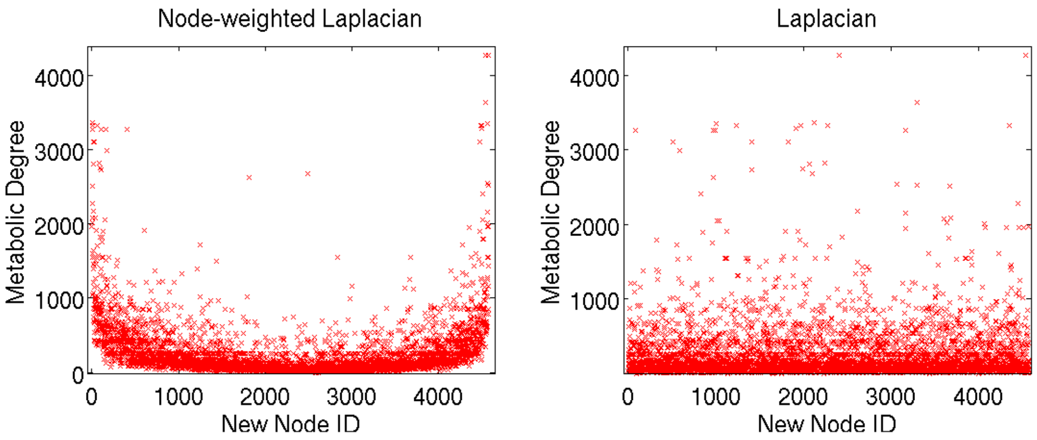

Figure 6.8: Metabolic degree of reordered genes. Left: ordered by vector from the node-weighted Laplacian. Right: ordered by Fiedler vector.

Having confirmed that the metabolic information has affected the ordering, we now check whether $D_w^{\frac{1}{2}}x^{[2]}$ has identified structure (i.e. a cluster) in the microarray data. We may do this by first inspecting the reordered microarray correlation matrix and choosing an appropriate range of contiguous nodes from the end of the ordering. In our case, 200 genes appeared to form a strong group. This is our putative cluster, whose quality can then be measured. There are, of course, many competing measures of cluster quality. Here, we follow the approach of [220], as used in Chapter 3, Section 3.2.1.

Using this approach, the 200 genes in clusters at both ends of the data were tested, producing p-values below 0.01. Overall this confirms that (a) the data contains a set of nodes with high expression correlation and high metabolic activity, and (b) the customized spectral approach was able to identify this structure.

6.4.3 Interpreting the Results

Factorizing metabolic pathway data together with gene-expression data is a way of adding known large-scale biological information to the analysis. This approach does not attempt to prejudice the outcome, but asks if prior knowledge can add any useful information.

We are able to add a basic biological narrative to some of the observed genes that appear at both ends of the matrix. Along with leptin, a signalling molecule produced in adipose tissue, we find acyl-CoA oxidase 1, palmitoyl, the first enzyme in fatty-acid beta oxidation; malonyl-CoA decarboxylase, involved in both fatty-acid bio-synthesis or, more plausibly here, scavenging odd-length dicarboyxlic acid fatty-acids.

At the other end of the matrix, we find the gene for argininosuccinate lyase, traditionally linked to low food availability [176]. This is implausible in this cohort, both from the social background and internally. Our analysis also finds ketohexokinase; the presence of this enzyme has been linked to a high fructose diet and its role is to use this sugar as both an energy source and, in adipose tissue, as source for precursors of fatty-acids [114]. Ketohexokinase initiates the pathway through which most dietary fructose is metabolised [46, 14]. Traditionally this was described as an energy store, but now is usually viewed as leading to undesirable fat and obesity. Fructose, in developed countries, is a common ingredient in most diets from the addition of corn syrup [12].

Our analysis has also led us to discover patterns with high probability of relevance to metabolic syndrome, obesity and type-2 diabetes, which has been linked to fructose intake [12]. The availability of relevant biometric information would allow us to place these observations into more specific biological context.

6.5 Discussion

Our aim was to motivate and illustrate spectral methods for network analysis. We used a first principles, linear algebra setting in order to show that by varying specific choices in the algorithm design we can generate a range of spectral methods. In particular we derived a simple, novel extension that can uncover assortative substructure. We finish by mentioning two key areas of current interest. First, for a large complex network, that is perhaps noisily defined, it may be of interest to identify substructures that go beyond simple clusters. For example, algorithms can be devised that discover subpatterns of bi-partivity [58], periodicity [76] or hierarchy [37], using spectral means. Second, a more systematic spectral approach for dealing with two or more related data sets can be developed through the use of the Generalized Singular Value Decomposition [124, 177, 189, 220], as we have also demonstrated in Chapters 3 and 4.

Since the principle has been established, it is not important to verify that the approach works in individual cases - that is, since the method was successful with the KEGG metabolic network we know the MetaCyc metabolic network will allow for equivalent results.

For the interest in this thesis, however, specifics in this adipose microarray data set are not important. Instead, this chapter has validated the node-weighted Laplacian, showing that it is possible to identify such structures in synthetic data and microarray gene expression data. The next step is to implement the approach as part of a live analysis on schizophrenia data pertinant to this thesis.

Chapter 7

Exploring the Effect of Antipsychotic Medication on Gene Expression in Human Whole Blood Schizophrenia Data

7.1 Introduction and Motivation

The application of network theory to biological systems is particularly well suited to the case of gene expression. Gene expression experiments provide a huge amount of data which is impractical to process using reductionist approaches but is convenient to represent in network form. In this chapter we create a setting for the exploration of gene expression data in human whole blood by providing a brief picture of issues with the more common profiling methodologies of brain tissue. Gene expression profiling in schizophrenia is typically performed in post-mortem brain tissue; and there is a lack of a consistent outcome across studies. There are many reasons for this inconsistency, including factors that will vary from sample to sample such as the post-mortem interval, sample pH and possible degradation of RNA [203, 133]. For instance, false levels of differential expression can arise in cases where sample parameters alter expression in a subset of genes in opposing directions. Mitochondrial defects have been implicated as playing a role in both schizophrenia and bipolar disorder, but it has been shown that a decrease in brain tissue sample pH has the effect of both decreasing mitochondrial gene expression and increasing apoptotic pathway expression, resulting in a differential not related to the condition [213]. There are also potential issues with abundance - low abundance transcripts may not be detectable or have easily measurable dynamic range [155].

In addition, post-mortem tissue is difficult to obtain, limiting the potential for large scale study [190]. There is then room to explore potential differential expression in different tissues. A desirable choice in terms of diagnostics is whole blood. Blood collection is non-invasive and needs little specialist training - this means it is relatively straightforward to obtain large numbers of samples, potentially for the same individuals through different stages of the disease, though this was not the case with the data set considered here.

In this chapter the analysis is exploratory with the objectives being to investigate the viability of blood as a sample source, and to uncover differences between gene expression of different sample sub-groups. To this end different reorderings from the SVD and GSVD clustering techniques will be used and results can be examined for potential biological merit. We take this approach since it is unclear if schizophrenia will influence gene expression measurements in blood - that is, schizophrenia is a brain disorder and the transferrability and applicability of blood gene expression on brain disorders is currently under examination [20, 71, 205]. Since there is potential for the disease to have limited effect on gene expression within whole blood we note that there may be factors within the data that are more important for clustering than the disease state. Consequently, it is sensible to adopt a wide ranging approach.

Genome Wide Association Study: GWAS

Variations in the genome of an individual are one factor that contributes to variations in their phenotype. GWAS studies measure genetic features physically associated with an individual [27]. This includes insertions and deletions of genetic code, variations in the number of times segments of code are repeated (copy-number variation) and variations in individual nucleotides, single nucleotide polymorphisms (SNPs). The GWAS approach examines a particular genetic variation (allele) across large samples and calculates an odds ratio of association. This odds ratio gives the probability that the genetic variant is linked to the phenotype. Thus, in combination with high throughput technology, GWAS studies can be used in case-control studies to identify physical genetic variants that are related to a disease state. The result is that GWAS study provides a reliable (through objective determination of the allele) measure of genetic involvement in a condition.

Although the variations in genetics as investigated in GWAS studies are not inexorably linked to expression levels, in this chapter we will in part treat the gene expression study as phenotypic data and examine overlap between this and GWAS results. As described in Chapter 2, gene expression is a highly variable process that is affected by a large number of external and internal factors. This means that definitively linking differential gene expression to a disease state can be difficult - and so there is more weight in identifying cases where gene expression levels are altered in genes that have previously been shown to have an association (in separate studies) with the condition.

7.1.1 Blood Background

Blood presents an interesting investigative opportunity for schizophrenia. As well as the aforementioned reasons of availability and cost, there have been suggestions of lymphatic involvement in an intermediate role between the nervous system and immune system [69]. Though the publication numbers are limited compared to brain tissue studies, there have been a number of recent efforts in characterising the viability of whole or partial blood samples as representatives for neurological conditions [118]. Results are currently uncertain and more study is required, but there are some encouraging studies that have shown the reproducability of brain gene expression profiles in perhipheral blood [99, 199, 183], and our work provides further positive support. In addition, blood shows stable gene expression profiles, giving potential for repeat study [148].

7.2 Introducing the Data

Analysis in this section is carried out on a publically available (`http://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-38484`) microarray dataset [43]. This set was chosen for its large sample size, and for the fact that there are a number of antipsychotic-free patients included in the cohort. The inclusion of antipsychotic-free patients is of interest since it is unclear how much of a normalizing (or, more likely, disruptive) effect antipsychotics will have on an individual's expression profile.

The data, summarised in Table 7.1, consists of 118 controls, 92 medicated diagnosed schizophrenia patients and 29 antipsychotic-free (or 'unmedicated') schizophrenia patients. All individuals are from Denmark or the Netherlands. Age and gender data is also available for each.

	Samples	Mean Age	Num. males	Num. females
Schizophrenia	92	40.86	66	26
Antipsychotic-free	14	31.21	10	4
Control	96	39.31	42	54

Table 7.1: Summary information on the data set adapted from [43]

The data was gathered on an Illumina microarray kit, model:

- Illumina HumanHT-12 v3.0 Expression BeadChip

There are measurements for 48, 743 probes in this data set, comprising a smaller number of genes - since the relationship between gene and probe is one to many. Figure 7.1 shows an unordered list of patient (kept in the order they were given in the data set) total expression levels (across all probes), colour coded for experimental state.

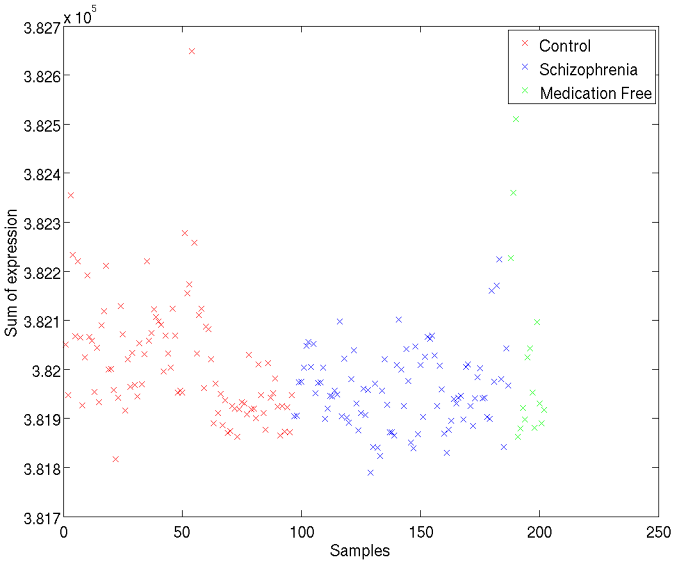

Figure 7.1: Sum of expression across all probes for each patient.

Figure 7.1 shows a similar range across each experimental group with a small number of outliers. Samples $1-57$ of the control group appear higher on total expression than the rest of the controls, and other experimental groups. This could be an experimental artefact or a result of some unmentioned change in the experimental protocol. For example, this pattern may be explained by one of the systematic variations that affect measurement of gene expression - such as variation in detection of the flourescent dyes or differing amounts of mRNA material in the samples before measurements are taken. Since we take an unsupervised approach in the analysis, we will need to take care to check that our results are releated to the disease state, rather than differences between these control measurements.

7.3 Analysis: SVD and GSVD

In the following sections we will carry out a process of data exploration that utilises a spectral clustering approach with the SVD and GSVD, as outlined in Chapter 3, and expands to use other statistical approaches in the search for novel genes implicated in the disease state.

The first analysis will be with the SVD, using a clustering (in this case we can perform a biclustering of samples and probes) and reordering to generate lists of probes and genes that are responsible for variation within the data set. There is an intrinsic assumption here that the disease state is the main factor for the variation within the data. The genes partitioned towards the ends of the orderings will then be examined. In Section 7.4.4, a hypothesis test and measurement of fold change of gene expression are calculated to generate a subset of genes that meet a significance level ($p < 0.05$) and surpass a pre-defined level of fold change (1.5+).

The following sections will show results from analysis of the complete data set, and selected subsets (e.g., schizophrenia + control samples with antipsychotic free removed). This will allow investigation of potential differences between medicated and antipsychotic-free samples and provide insight into their differences with the control group. We will also apply the GSVD which, as outlined previously, can be used to identify mutually exclusive structures across networks. This will be applied to give three comparisons:

- Antipsychotic-free vs control
- Antipsychotic-free vs treated
- Medicated vs control

In addition, the data will be combined with the KEGG and MetaCyc metabolic networks using the novel node-weighted Laplacian approach developed in Chapter 6 with the goal of uncovering structures that have high degree in either the KEGG or MetaCyc networks but also have significant structure in the microarray data.

7.4 Clustering with the SVD

This section contains analysis where the data is reordered according to the SVD. Each of the following subsections contain plots of singular vectors where, based on classic spectral analysis arguments, as discussed in Chapter 6, we take the view that (a) nearby genes in the ordering exhibit similar behaviour and (b) the genes at the end of the orderings are responsible for driving structure of the ordering.

Whole Data

The SVD was carried out on the data as in previous chapters and for illustrative purposes we show plots of the Fiedler vector and the next left singular vector (corresponding to probes) in Figure 7.2. The probe IDs in Figure 7.2 have been reordered according to these singular vectors, showing groups at both ends with components deviating from zero. This indicates that these probes are responsible for driving the structure the SVD uncovers.

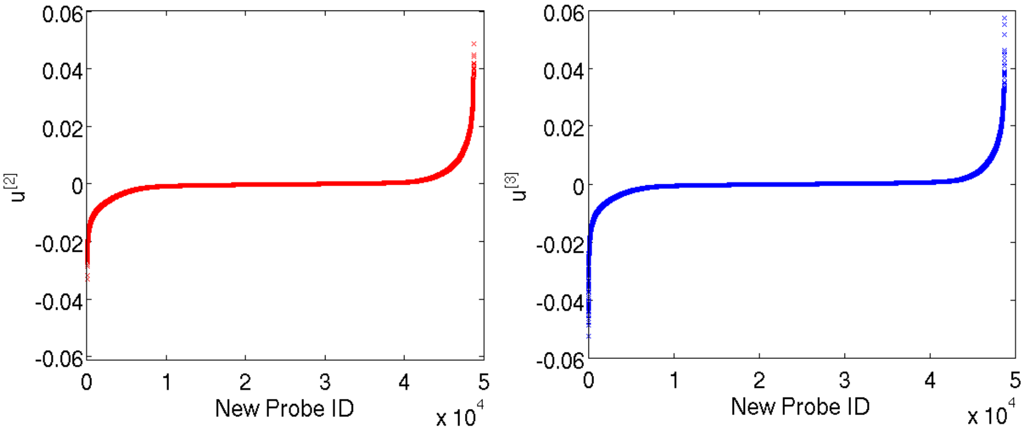

Figure 7.2: Components of the left singular vector $u^{[2]}$ and $u^{[3]}$ in increasing order from an SVD of the data.

Following on from the plots of the left singular vectors, we show in Figure 7.3 the reordering of samples according to the sorted components of the right singular vectors (corresponding to samples) from the SVD of the data with the complete sample cohort. For interest, in this figure we have provided different labels for schizophrenia patients (labelled SCZ), antipsychotic-free schizophrenia patients (medication/antipsychotic free patients - MFP), controls (CTRL) and males/females. Gender was split in this way such that we could check whether gender had a significant effect on gene expression, necessitating separation of males and females in future analyses.

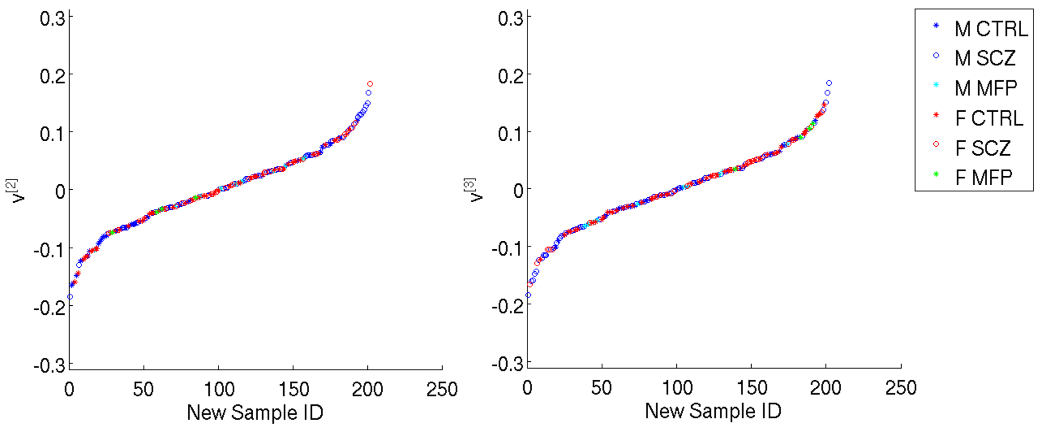

Figure 7.3: Components of right singular vectors $v^{[2]}$ and $v^{[3]}$ in increasing order from an SVD of the data. Each component represents an individual in the data: each point is labelled according to gender and disease state.

The results in Figure 7.3 show little separation in either singular vector between males and females, suggesting that gender is of limited importance in driving the ordering of the samples - and so gender will not be used to separate this data in future sections. For additional clarity on the division of disease status, we next plot the singular vectors without the male/female distinction in Figure 7.4.

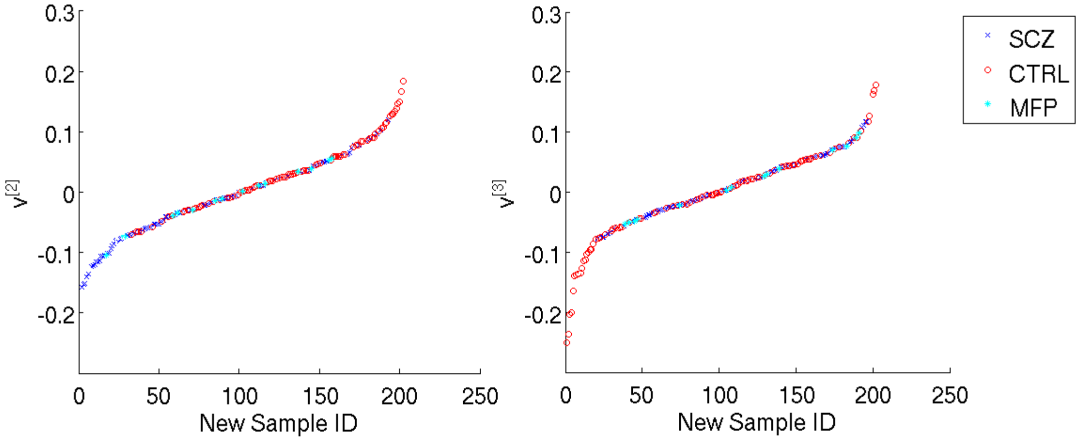

Figure 7.4: Components of right singular vector $v^{[2]}$ and $v^{[3]}$ in increasing order with disease state highlighted.

In this case, the plot on the left with the right singular vector $v^{[2]}$ efficiently distinguishes the schizophrenia patients from the control. The antipsychotic-free patients are spread across the ordering - a possible explanation here is that the relative number of antipsychotic-free patients in this set is small, so the variance may be dwarfed by the other two sets.

Antipsychotic-free and Controls

Next we present figures where the antipsychotic-free and control patients have been selected. Figure 7.5 shows the left singular vectors $u^{[2]}$ and $u^{[3]}$ from this comparison, showing again that a large number of probes drive a division in the ordering of the data.

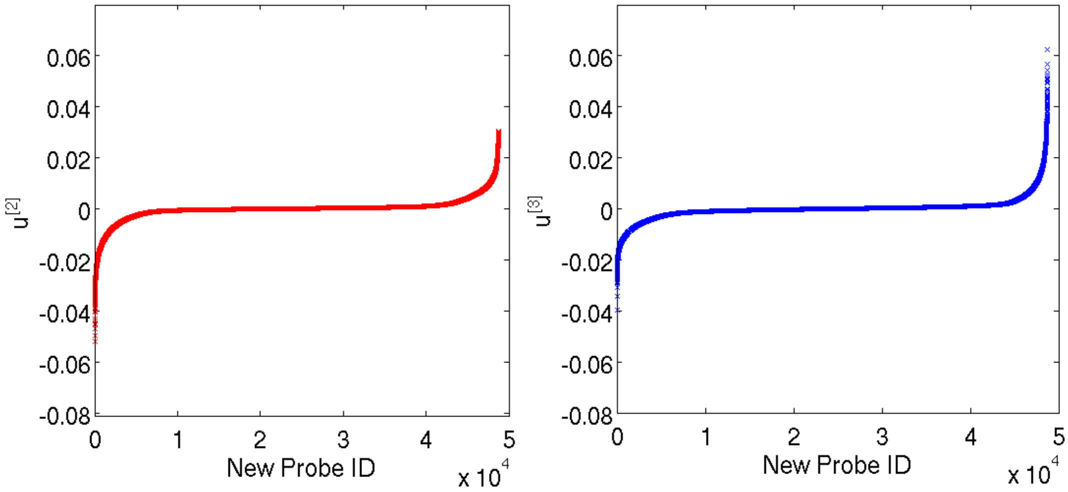

Figure 7.5: Components of $u^{[2]}$ and $u^{[3]}$ from SVD of dataset containing control (CTRL) and antipsychotic-free (MFP) samples.

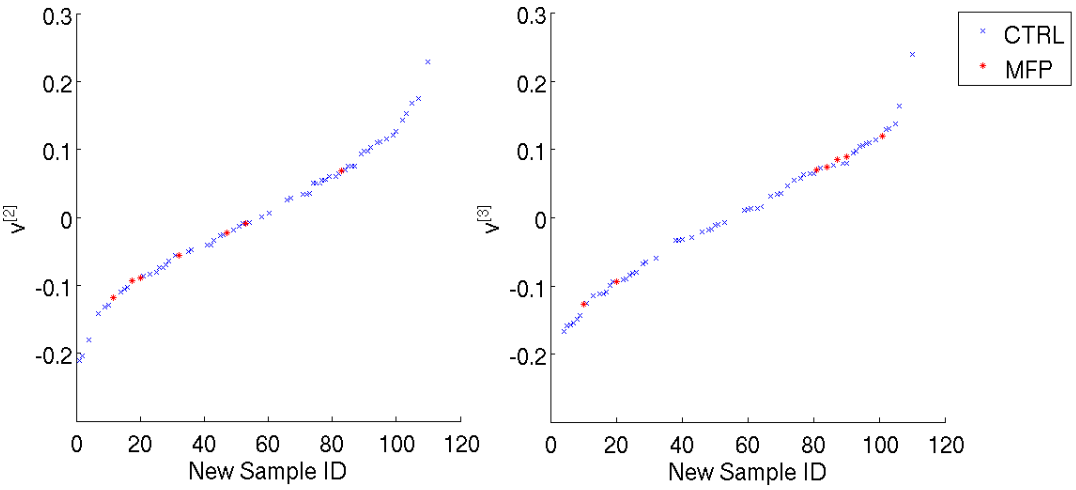

Figure 7.6: Components of $v^{[2]}$ and $v^{[3]}$ from SVD of dataset containing control and antipsychotic-free samples.

In Figure 7.6 the disparity in sample numbers is obvious - $v^{[2]}$ orders antipsychotic free samples towards the left, where $v^{[3]}$ orders those towards the right. The antipsychotic free samples are not perfectly clustered, which suggests that there are factors other than the differences due to their sample type driving the structure. It could be the case that the antipsychotic-free samples are not vastly different in gene expression, and so driving features within the data are from any of the factors that affect gene expression in all individuals (see Gene Expression in Section 2.1 for related discussion). This idea will be explored in later sections, with for instance a comparison with schizophrenia literature in Section 7.4.2.

Schizophrenia and Control samples

This final comparison, illustrated in Figures 7.7 and 7.8 results from the SVD on the data from schizophrenia patients and control samples. There is a weak trend in $v^{[2]}$ where the schizophrenia samples appear more often on the right hand side of the plot, similarly the left hand side in $v^{[3]}$, but once again there are discrepancies where samples are not grouped according to their disease status.

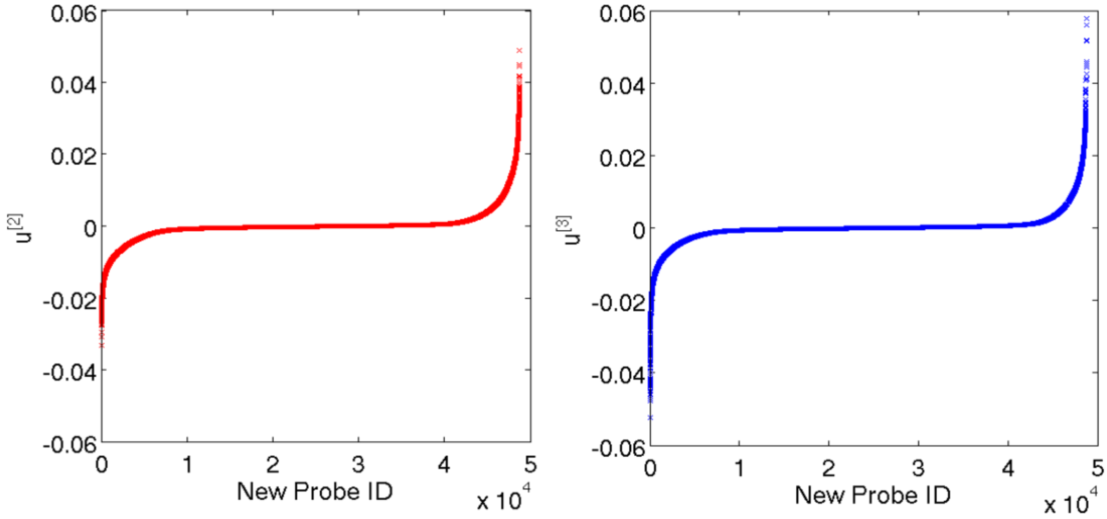

Figure 7.7: Components of $u^{[2]}$ and $u^{[3]}$ from SVD of dataset containing schizophrenia and control samples.

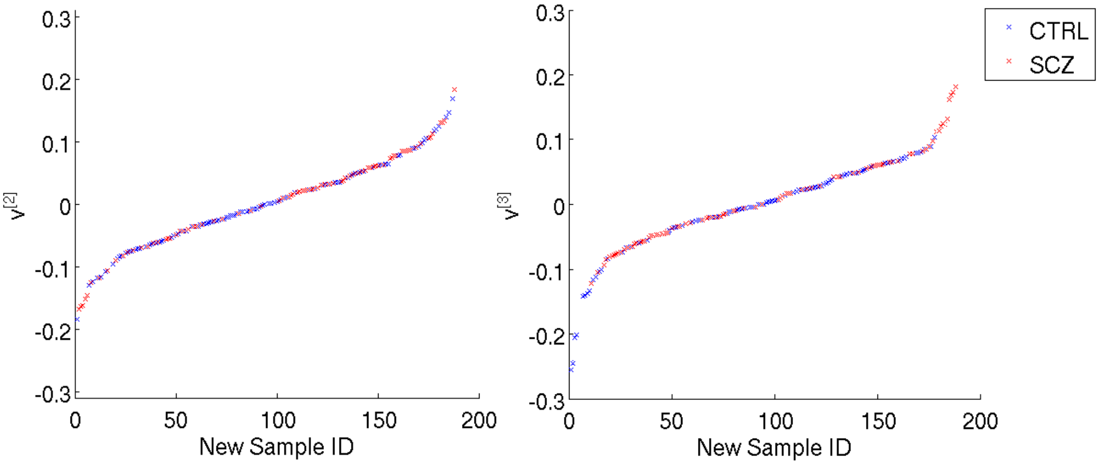

Figure 7.8: Components of $v^{[2]}$ and $v^{[3]}$ from SVD of dataset containing schizophrenia and control samples. Separation of sample groups is not perfect yet there are some visible collections of common sample groups.

These results suggest that over the whole probe set there are probes that vary naturally across control individuals, providing some discriminating structure between samples. The fact that many schizophrenia samples are grouped on the right hand side of this picture may be indicative of the fact that those samples share some common structure in expression. The fact that samples are not grouped perfectly according to disease state may also be suggestive of the fact that schizophrenia is not one specific condition - there is a wide variety of samples, variety of onset points and treatment modalities. Each of these factors can contribute to the measured expression levels.

7.4.1 Distribution of Variance

Often in microarray analysis low variance (that is the spread of expression level across all samples, control and disease) genes will be excluded [43], this potentially limits the scope for discovery of novel genes across sample groups - equivalent to the issue of discarding genes with low fold changes in expression (as mentioned in Section 7.4.4). Motivated by this aspect of traditional microarray analysis, the distribution of variance across probes within the data was checked. The variance used to create these figures is the the variance across all sample groups. Whilst there are small differences in the distributions of variance between the sample subsets, the results are transferrable enough that individual sample groups are not included. Figures 7.9- 7.10 show a variance histogram and the relationship between variance and mean expression. Figure 7.11 shows how the SVD responds to variance by showing some data reorderings.

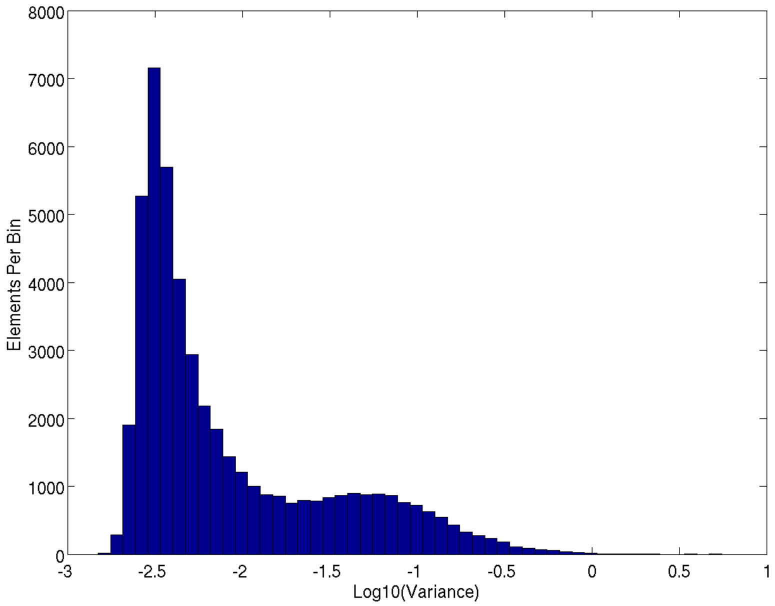

Figure 7.9: Frequency plot of sample variance across probes in the complete data set with 50 bins. There is a visible peak and hump in the elements across variance.

In Figure 7.9 it is clear that the majority of probes are contained within a small range of variance (we use $log10$ as a more appropriate scale). This highlights one of the issues with arbitrary variance level selection procedures; if we were to introduce a cutoff somewhere in the range where the majority of probes are located, we would discard many potentially interesting points located close to this cutoff.

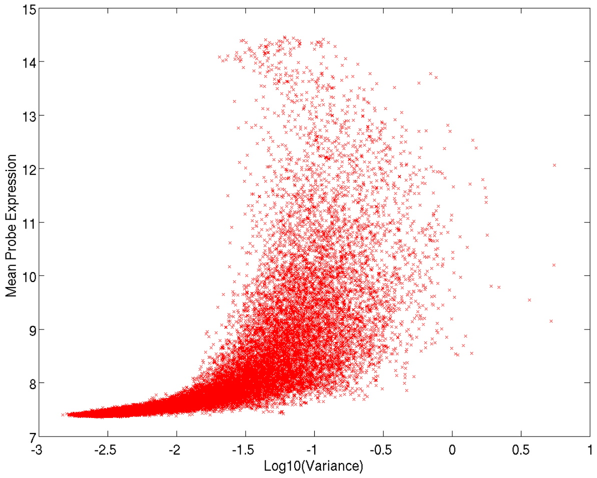

Figure 7.10: Scatter of mean expression for each probe against log10(Variance) showing that high variance probes tend to have higher mean expression

Figure 7.10 shows that high variance probes in this data tend to have high mean expression levels. This suggests that significant genes in this data set will likely be more abundant overall. As the mean expression level increases, the number of probes begins to decrease, until there is a collection at a ceiling point of expression between 14 – 15. This is explainable as saturation of the microarray kit. The second feature of note is the dense beam at the low end of mean expression levels - this may again signify limitations of the equipment.

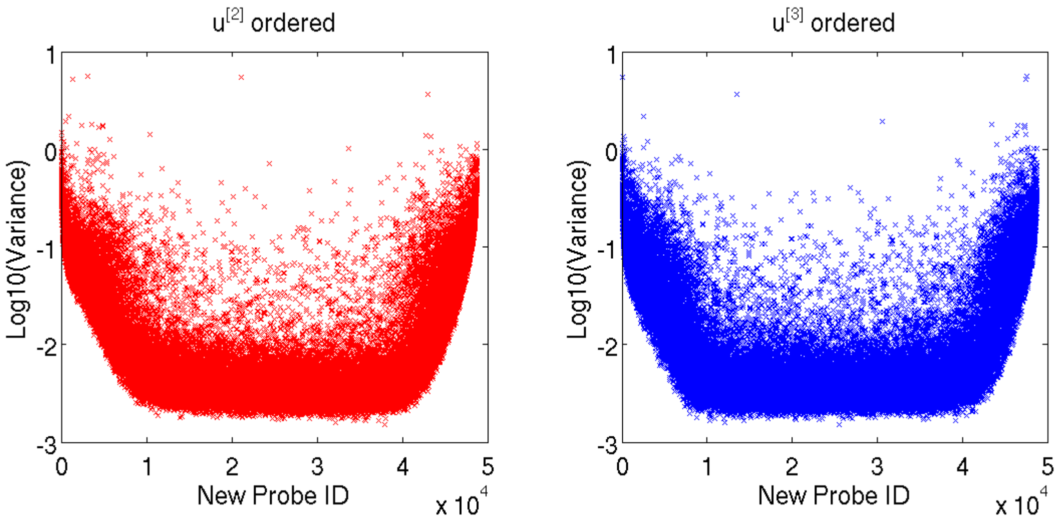

Figure 7.11: Variance of probe ID across samples reordered according to increasing value of the components of the left singular vectors $u^{[2]}$ and $u^{[3]}$. There are few low variance probes at the end of the ordering.

In Figure 7.11 there is the interesting result that whilst the high variance probes have a tendency to move towards the end of the ordering, there is a scattering in the distribution (that is, there are high variance probes located throughout the orderings). This shows that variance across a probe is not the only driver in forming the reordering from the SVD: another point which suggests it is undesirable to simply discard low variance probes, particularly in a clustering approach.

7.4.2 Comparison with Gene List Database

As a starting point towards assessing the viability of blood for schizophrenia diagnostics, we can compare our gene lists with information available from the literature. There have been many genes implicated in schizophrenia, as a result categorising the literature and building an objective list of genes would be an enormous challenge, beyond the scope of this thesis.

However, the Schizophrenia Gene Resource SZGR:-

`http://bioinfo.mc.vanderbilt.edu/SZGR/index.jsp`

is an effort to collect information on all genes that have been studied in relation to schizophrenia - there are three main approaches and categories built in organising the database: association, expression and literature studies, these are outlined individually in the following sub-sections.

In this section we compare genes from the reordering of the SVD with the SZGR database to find out which of the genes that have previously been implicated in schizophrenia are present in our clusters. In addition, it is interesting to consider that at least the majority of the genes present in the SZGR database will be results from studies in brain tissue - in comparing our experiment to this database we are, in a sense, testing for overlap between blood and brain tissues, an important step towards demonstrating the practical relevance of a blood based diagnostic assay.

Genome Wide Association Studies

In order to form the association set, SZGR extracted gene information and study details from the SchizophreniaGene (SZGene:- `http://www.szgene.org/`) database. There is then an odds ratio calculated and assigned for each of the genes, with a score of 3 indicating a p-value < 0.001, 2 is p $[0.001 - 0.05)$ and 0 otherwise. Genes implicated from association experiments are likely to be the most reliable in terms of definite connection with disease.

Gene Expression

The expression section of the database is constructed from a set of meta-analyses. This includes a compendium of 12 gene expression data sets spanning 988 arrays in a study compiled by the Stanley Medical Research Institute (SMRI) [84] and a comparison of expression profiles (according to Gene Atlas [107]) of genes collected from meta analyses available in the literature.

The genes gathered from these studies for entry to SZGR are required to meet a significance threshold ($p < 0.05$).

Literature

The literature portion of the database is formed based on a principle of co-occurence of search terms. The keywords “schizophrenia”, “schizophrenias”, “schizophrenic”, “schizophrenics”, “schizotypy” and “schizotypal” are searched in the NCBI Entrez search utility for NCBI PubMed. If a gene and one of the above keywords co-occur in a publication then a hit counter is incremented. Then the maximum score a gene can have is 6 - where the gene co-occurs with each of the above keywords. This is likely to be the least reliable in terms of definite disease link. This approach is used as a basic mining strategy for large data and has been previously used in other areas of systems biology [182].

Comparison Between Lists

An obvious step to take with these lists in place before integrating them into the experimental analysis is to compare them with each other for overlap. Tables 7.2 and 7.3 show how many genes are present (according to the official gene symbol) in the aforementioned sources: off diagonals show the number of matches between the lists. For information, Table 7.3 shows the corresponding result for probe ID rather than official gene symbol.

	Association	Expression	Literature
Association	277	15	250
Expression	15	656	70
Literature	250	70	1599

Table 7.2: Number of matches between SZGR lists in official gene symbol IDs. There is overlap between all combinations of list, with limited cross-over between association and expression.

	Association	Expression	Literature
Association	3856	262	3551
Expression	262	5054	1048
Literature	3551	1048	15124

Table 7.3: Number of matches for Illumina probe IDs between each SZGR list. There are significantly more probes than gene symbols, with overlap between all lists.

An interesting result is that there is a small nontrivial overlap between association and expression studies - this is unexpected due to the difference between the nature of genetic variants (as measured by association studies) and gene expression. Any such overlap is of interest as there is an implication of a relationship between specific genetic variants and expression levels for these genes. We can also see from these tables that the literature subset is the largest, as expected since it has the smallest degree of specificity and, again as expected, the literature set has significant overlap with the other types.

7.4.3 Significance Testing of Ordering Results

The number of matches between SZGR and the SVD ordering are assigned p-values using a bootstrap type approach (where the original list is resampled repeatedly [157], though in this case the list corresponds to IDs), estimating the probability of the result appearing by random chance.

We will do two things:

- Check the ends of the singular vectors, to see how many genes match up with those in the database.

- Calculate a p-value for the middle of the ordering (which should be not significant) as a sanity check to verify a base level of agreement with the literature.

The p-values are calculated according to the process in Section 3.2.2 and provide a basic assessment of whether the SVD reordering of the data is identifying genes that are known to be implicated in schizophrenia. This is of interest since the two main goals of this chapter are to test the value of blood for expression study in schizophrenia and report any correlation with current knowledge. A significant p-value (< 0.05) suggests that the reordering procedure is able to identify a significantly greater number of genes implicated in the disease state than random chance.

The p-value results are found in Tables 7.4, 7.5 and 7.6, with $p < 0.05$ values highlighted in bold. Also since, as mentioned in Section 7.2, there are often multiple probes to measure one gene, we provide the number of genes present in the ordering *Sum* (number of genes), as well as the number of probes *Sum with repeats* (which is the number of probes). For the initial results, we simply chose a selection of $1,000$ probes from either end of the ordered singular vectors. Thus, '$u^{[2]}$ Start' is the first $1,000$ nodes from the left singular vector $u^{[2]}$, '$u^{[2]}$ Mid' is $1,000$ nodes selected from the middle of the singular vector and '$u^{[2]}$ End' is $1,000$ nodes selected from the end of $u^{[2]}$, and similarly $u^{[3]}$ results are those from the left singular vector $u^{[3]}$. The singular vectors used to generate these tables are as shown in Section 7.4.

We can see that the results are very similar between Tables 7.4 and 7.6, which makes sense since schizophrenia and control populations make up the majority of the data set. It is particularly notable within these tables that there are a number of genes from association studies present in either end of the orderings. This replication across study types could suggest these genes are of a high level of

Complete Data						
Association	$u^{[2]}$ Start	$u^{[2]}$ Mid	$u^{[2]}$ End	$u^{[3]}$ Start	$u^{[3]}$ Mid	$u^{[3]}$ End
Sum	13	6	9	3	9	15
Sum with repeats	15	6	9	3	9	17
p-value	**0.045**	0.932	0.687	0.997	0.673	**0.027**
Expression	$u^{[2]}$ Start	$u^{[2]}$ Mid	$u^{[2]}$ End	$u^{[3]}$ Start	$u^{[3]}$ Mid	$u^{[3]}$ End
Sum	30	10	29	23	14	24
Sum with repeats	41	11	36	38	16	36
p-value	**0.003**	0.999	**0.016**	**0.015**	0.978	**0.026**
Literature	$u^{[2]}$ Start	$u^{[2]}$ Mid	$u^{[2]}$ End	$u^{[3]}$ Start	$u^{[3]}$ Mid	$u^{[3]}$ End
Sum	34	23	42	31	24	47
Sum with repeats	73	38	103	77	47	116
p-value	**0.021**	0.997	$<$ **0.001**	**0.003**	0.912	$<$ **0.001**

Table 7.4: Comparison of selected genes from left singular vectors $u^{[2]}$ and $u^{[3]}$ (Figure 7.2) of SVD of the complete data set. Significant numbers of genes appear in almost all orderings and gene lists. No significance in Mid ranges as expected.

Antipsychotic-free and Control						
Association	$u^{[2]}$ Start	$u^{[2]}$ Mid	$u^{[2]}$ End	$u^{[3]}$ Start	$u^{[3]}$ Mid	$u^{[3]}$ End
Sum	14	7	11	8	9	13
Sum with repeats	14	7	12	9	9	15
p-value	0.137	0.804	0.305	0.709	0.584	0.091
Expression	$u^{[2]}$ Start	$u^{[2]}$ Mid	$u^{[2]}$ End	$u^{[3]}$ Start	$u^{[3]}$ Mid	$u^{[3]}$ End
Sum	30	19	33	29	22	31
Sum with repeats	39	21	46	43	25	38
p-value	**0.006**	0.761	$<$ **0.001**	$<$ **0.001**	0.464	**0.015**
Literature	$u^{[2]}$ Start	$u^{[2]}$ Mid	$u^{[2]}$ End	$u^{[3]}$ Start	$u^{[3]}$ Mid	$u^{[3]}$ End
Sum	33	17	49	32	26	39
Sum with repeats	69	38	110	76	52	97
p-value	**0.039**	0.997	$<$ **0.001**	**0.005**	0.730	$<$ **0.001**

Table 7.5: Comparison of genes from left singular vectors $u^{[2]}$ and $u^{[3]}$ (Figure 7.5) of SVD from data of antipsychotic-free and control samples only. Many significant results, with high significance in expression studies. No significance for association orderings in this case.

Schizophrenia and Control						
Association	$u^{[2]}$ Start	$u^{[2]}$ Mid	$u^{[2]}$ End	$u^{[3]}$ Start	$u^{[3]}$ Mid	$u^{[3]}$ End
Sum	13	10	9	4	8	15
Sum with repeats	14	10	9	6	8	17
p-value	0.1434	0.5656	0.6854	0.935	0.799	**0.027**
Expression	$u^{[2]}$ Start	$u^{[2]}$ Mid	$u^{[2]}$ End	$u^{[3]}$ Start	$u^{[3]}$ Mid	$u^{[3]}$ End
Sum	29	19	32	27	10	26
Sum with repeats	36	21	43	38	13	36
p-value	**0.019**	0.828	< **0.001**	**0.008**	0.997	**0.022**
Literature	$u^{[2]}$ Start	$u^{[2]}$ Mid	$u^{[2]}$ End	$u^{[3]}$ Start	$u^{[3]}$ Mid	$u^{[3]}$ End
Sum	34	23	42	31	24	49
Sum with repeats	73	38	105	77	47	116
p-value	**0.021**	0.997	< **0.001**	**0.003**	0.912	< **0.001**

Table 7.6: Comparison of genes from left singular vectors $u^{[2]}$ and $u^{[3]}$ (Figure 7.7) of SVD from data of schizophrenia and control samples only (no antipsychotic-free). Results are similar to previous table, with significance in all expression and literature SZGR lists.

interest since they are altered in gene expression and show some genomic variation in association study.

Finally, these results provide only an indication of the usefulness of this approach. The generation of p-values here is somewhat arbitrary, and should be taken only as a guide. For clarity, a non-significant p-value only indicates insignificant overlap between the ordering and SZGR database. It does not make a statement about the importance of the genes present in the ordering.

7.4.4 Hypothesis Testing: T-test and Fold Change

We next apply a hypothesis test to genes that have appeared in the clusters of the SVD orderings. This approach will identify genes that potentially do not appear in the previous database lists.

The hypothesis tests in this section compare the schizophrenia samples with the control samples, and the antipsychotic-free samples with the controls (that is, the disease state is tested for similarity with the control). The result is a p-value for each Illumina probe that will give an indication of the likelihood that, for that probe and the gene it measures, the schizophrenia sample differs in some way (dependent on the test used) from the control samples. We then also compare the treated schizophrenia samples with the antipsychotic-free patients. In comparing the probes that are significant in the experimental orderings in this way we are also focussing on probes that differ between experimental groups - and checking that the approach is capable of identifying results that are as such. This will make sure the separation of control samples seen in Figure 7.1 is not an overriding feature of the data.

This process is performed for multiple SVD orderings, and the most significant genes are highlighted, with their positions in the cluster noted. The results of the hypothesis testing are then used in the generation of volcano plots, as explained later in this section.

T-Test

For background information on the t-test see [157]. The t-test is commonly used in applications related to gene expression data [16, 201, 205], though there have also been noted concerns about the ability of the t-test to manage false positives - including the fact that the t-test is reliant on the data holding to the normal distribution, which may not always be the case. The t-test is used to compare the means of sample groups, where the p-value is the probability of obtaining a t value at least as extreme as observed. The sampling distribution is known for Student's t-test so the p-value is given by the sum of the probabilities of events more extreme than observed.

Mann-Whitney Wilcoxon test

Following on from the concern that the t-test may not be suitable due to issues of normality within the data, we calculated a measure of skewness (a measure of how skew a distribution is, or tendency for data to lie on a particular side of the mean) for each of the probes in this data set and also use the more robust non-parametric Mann-Whitney test. The results from the Mann-Whitney test are broadly in agreement with those of the t-test - with none or very few instances where the p-value significance threshold is affected. As a result, in the tables that follow we show only the t-test results.

Volcano Plot

Volcano plots are a graphical method of examining results in fold change and statistical significance simultaneously [38], allowing for quick identification of points that meet a dual threshold criteria, particularly in relation to microarray data where there is a large amount of data [3]. Volcano plots are typically formed as a plot of some log scaled p-value (usually $-log_{10}$) versus a scaled fold change (commonly log_2), as in the example Figure 7.12 - the name 'Volcano' comes from the typical shape of the result. The usage in this data is appropriate since we can plot p-values for a particular gene (from the hypothesis test comparing the disease sample to control) and a ratio of the probe mean for each sample group - this will allow us to identify any genes that are significantly different between experimental groups, and have a large fold change - the upper corners of a volcano plot contain genes that have both significant $p-$values and a high fold change. In each of the following tables the fold change is always illustrated as disease state relative to control, so a negative value indicates downregulation.

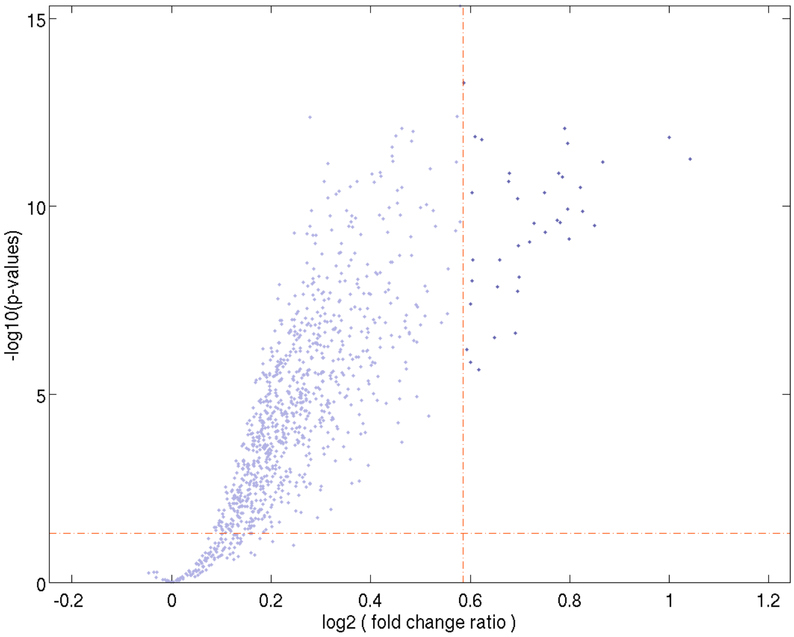

Figure 7.12: Illustrative example of a volcano plot formed form one end of reordered components of $u^{[2]}$ from the SVD of the schizophrenia and control patients data set. Red dashed lines correspond to fold change > 1.5 and $p-$value < 0.05 thresholds.

Each point on a volcano plot of this type represents an individual probe or gene -the top left and top right of the diagram highlight the areas of most interest- where the p-values are low and the fold change is high. After generating a volcano plot, we then place a threshold of a fold-change for each gene (typically 1.5 or 2 [207, 141]). Rather than displaying volcano charts for each ordering and each subset of data, the significant results have been extracted and displayed in Tables 7.7 - 7.25.

An average gene expression level is calculated for each gene, for each sample group - the ratio of this is known as the gene expression ratio. The advantage of using the log2 transform on a set of gene expression ratios is that differential up-regulation and down-regulation are treated the same, e.g. $log2(2) = 1$, log2(0.5) = -1. The log2 mapping is then desirable as it creates a continuous space that allows for direct comparison of up and down regulation.

Gene Expression Abundance

Since fold change is a ratio of over or underexpression, it is interesting to include a measure of the relative abundance, in terms of the overall experimental measurements, of the expression of a particular probe. This will allow an assessment of how the algorithms respond to the data, as a follow-on from Figure 7.10 in Section 7.4.1 which shows a trend whereby high mean expression probes tend to have higher variance. The purpose of this addition is twofold:-

- There is more biological information in knowing fold change levels alongside a relative expression level (the abundance measure).
- The approach is expected to favour high variance probes, in calculating a measure of abundance we can identify low abundance and, in addition as implied by Figure 7.10, low variance probes.

To calculate the percentile abundance the MATLAB function `tiedrank` is used - ranking each of the probes according to average expression level (of the control samples). In this section we show the results from fold change and t-test measurements on genes from the end of the corresponding left singular vectors.

Complete Data

Tables 7.7 and 7.8 show results that meet the dual threshold criteria of the t-test (comparing schizophrenia, including antipsychotic-free samples, with controls) and gene fold change. Cluster one in orderings from $u^{[2]}$ and $u^{[3]}$ are very similar, notably genes such as *NDUFA4*, *HINT1*, *COX7C* and *EVI2A* appear in both orderings.

Official Gene Symbol	Illumina ID	p-Value	Fold Change	Percentile Abundance
Cluster One				
PRF1	ILMN1740633	2.13e-08	-1.58	96
NRGN	ILMN1705686	4.30e-08	-1.55	96
SMOX	ILMN2367258	1.51e-07	-1.58	89
GPR56	ILMN2352097	7.86e-07	-1.59	95
GPR56	ILMN2384122	2.11e-06	-1.57	95
ZNF683	ILMN1678238	3.31e-06	-1.56	88
EPB49	ILMN1671686	4.71e-05	-1.59	95
GZMH	ILMN1731233	1.38e-04	-1.57	97
RNF213	ILMN1749722	3.78e-04	-1.56	98
Cluster Two				
RPL9	ILMN1750507	1.23e-10	1.82	98
RPS17P16	ILMN1664610	2.04e-10	1.52	98
NDUFA4	ILMN1751258	2.62e-10	1.52	95
HINT1	ILMN1807710	3.02e-10	1.60	96
EVI2A	ILMN1733579	4.22e-10	1.59	88
COX7C	ILMN1798189	8.32e-10	1.60	96
RPS17	ILMN2207539	1.49e-09	1.56	97
RSL24D1	ILMN2175465	2.39e-09	1.60	90
COMMD6	ILMN1777378	3.05e-09	1.65	91
RPL31	ILMN1754195	1.83e-08	1.51	95
RPL9P25	ILMN2408415	8.78e-07	1.50	90

Table 7.7: $u^{[2]}$ complete data (ordering of all schizophrenia and control samples) set results: t-test on first and last 1000 nodes (named cluster one and two, respectively) of $u^{[2]}$ ordering from SVD. Genes that meet fold change > 1.5 and $p-$value < 0.05 are shown.

Official Gene Symbol	Illumina ID	p-Value	Fold Change	Percentile Abundance
Cluster One				
RPL9	ILMN1750507	1.23e-10	1.82	98
RPS17P16	ILMN1664610	2.04e-10	1.52	98
NDUFA4	ILMN1751258	2.62e-10	1.52	95
HINT1	ILMN1807710	3.02e-10	1.65	96
EVI2A	ILMN1733579	4.22e-10	1.59	88
COX7C	ILMN1798189	8.32e-10	1.65	96
RPS17	ILMN2207539	1.49e-09	1.56	97
RSL24D1	ILMN2175465	2.39e-09	1.60	90
COMMD6	ILMN1777378	3.05e-09	1.65	91
RPL31	ILMN1754195	1.83e-08	1.51	95
EIF1AY	ILMN1755537	3.01e-07	1.65	87
RPL9P25	ILMN2408415	8.78e-07	1.50	90
RPS4Y1	ILMN1783142	5.09e-05	2.50	96
Cluster Two				
ORM1	ILMN1696584	3.73e-04	1.51	91
RNF213	ILMN1749722	3.78e-04	-1.56	98

Table 7.8: $u^{[3]}$ complete (ordering of all schizophrenia and control samples) expression data set results: t-test on first and last 1000 nodes (named cluster one and two, respectively) of $u^{[3]}$ ordering from SVD showing genes with significant p-value and fold change.

Antipsychotic free and Control

The next comparison deals with the ($n = 14$) antipsychotic-free and the ($n = 96$) control samples. Table 7.9 shows the second ordering and Table 7.10 the third. The first thing to note is the scarsity of genes in these results (3 genes in $u^{[2]}$ and 4 from $u^{[3]}$). There are two likely explanations - the antipsychotic free contingent is much smaller than the controls, and so the majority of the matrix information comes from the control samples. The control samples are from a single sample group, and so there are little-to-no consistent differentiating genes that would survive a t-test between disease and control groups.

The second explanation is that the samples likely to provide variance in the data, the antipsychotic-free group, simply have a limited consistent alteration in expression as compared to controls - of the 2000 genes tested in each ordering only 3 and 4, respectively, meet the threshold criteria.

Official Gene Symbol	Illumina ID	p-Value	Fold Change	Percentile Abundance
Cluster One				
PABPC1	ILMN2136133	1.92e-02	-1.60	95
RPS4Y1	ILMN1783142	4.63e-02	-2.63	96
Cluster Two				
CDC14B	ILMN1733559	6.14e-03	-1.62	99

Table 7.9: $u^{[2]}$ ordering from SVD of all antipsychotic-free and control sample expression data results: t-test on first and last 1000 nodes (named cluster one and two, respectively) of $u^{[2]}$ ordering showing genes with significant p-value and fold change.

Official Gene Symbol	Illumina ID	p-Value	Fold Change	Percentile Abundance
Cluster One				
EOMES	ILMN1760509	1.15e-04	1.51	94
Cluster Two				
DDX17	ILMN2371590	2.23e-03	-1.57	95
CDC14B	ILMN1733559	6.14e-03	-1.62	99
PABPC1	ILMN2136133	1.92e-02	-1.60	95

Table 7.10: $u^{[3]}$ ordering from SVD of all antipsychotic-free and control sample expression data results: t-test on first and last 1000 nodes (named cluster one and two, respectively) of $u^{[3]}$ ordering showing genes with significant p-value and fold change.

Schizophrenia and Control

Tables 7.11-7.12 show comparison of schizophrenia samples with control (not including antipsychotic-free). This list is much more comprehensive than that seen in the previous sub-section with antipsychotic-free and control samples. As expected, there is significant overlap with Tables 7.7-7.8, since the majority of the data is the same. Interestingly, the total number of genes passing both thresholds has increased significantly - from 20 in $u^{[2]}$ and 13 in $u^{[3]}$ of the complete data, to 36 in $u^{[2]}$ and 28 in $u^{[3]}$ in this section. This is likely because the antipsychotic-free group was having the effect of normalising either fold change or t-test results towards the levels seen in control patients. This would mean that the schizophrenia with treatment and antipsychotic-free groups have very different presentations in terms of gene expression, a point which we test by analysing those two groups in the following section.

We also plot, limited to this section, the total expression level for all significant genes in cluster 2 of Table 7.11. This plot, in Figure 7.13, illustrates a stratification of patients within a sample group, highlighted in orange. This is a somewhat circular observation as, since the t-test is a comparison of the mean, the clustered genes are significant *because* these patients have higher levels of expression. However, plotting the individual sample means is informative in terms of further investigation. In future projects where the profile of an individual is more complete, identifying which samples are responsible for the variation holds value. For example, perhaps this sub-group suffers from a particular presentation of the condition - this information could potentially be linked to particular medication strategies, ultimately leading to a diagnostic paradigm using a specific differential expression profile.

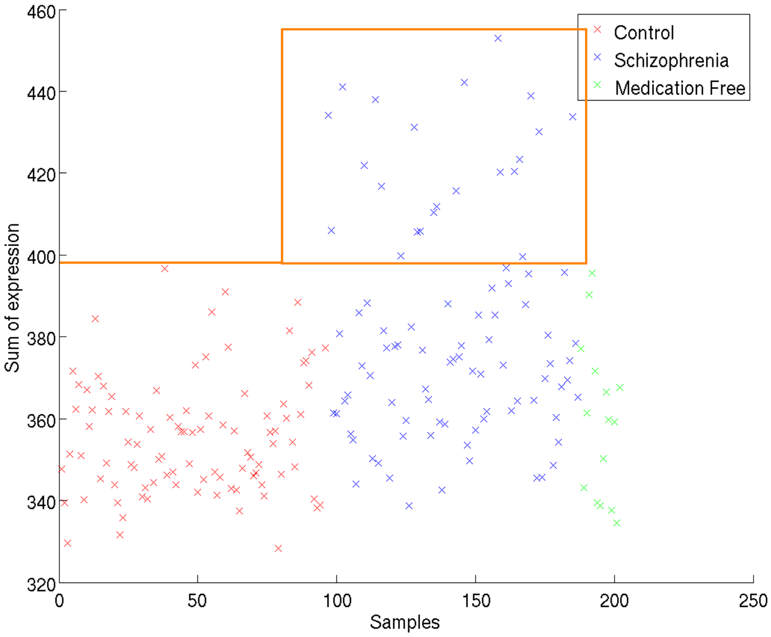

Figure 7.13: Total expression across probes in Cluster 2 of Table 7.11 for each sample. This shows a group of schizophrenia samples (orange box) where the total expression is increased relative to all other samples - this is the reason the t-test is significant and shows that significance is due to a limited number of patients.

Official Gene Symbol	Illumina ID	p-Value	Fold Change	Percentile Abundance
Cluster One				
CD3E	ILMN1739794	2.53e-10	-1.51	97
SMOX	ILMN2367258	1.01e-09	-1.56	97
NRGN	ILMN1705686	1.59e-09	-1.52	98
PRF1	ILMN1740633	6.30e-09	-1.51	92
UCP2	ILMN1685625	6.56e-08	-1.51	93
KIR3DL2	ILMN2190842	9.33e-08	-1.55	97
GPR56	ILMN2352097	1.43e-07	-1.54	99
ZNF683	ILMN1678238	3.01e-07	-1.51	96
GPR56	ILMN2384122	4.05e-07	-1.52	93
EPB49	ILMN1671686	1.60e-06	-1.58	95
NPRL3	ILMN1733581	2.84e-06	-1.51	99
RNF213	ILMN1749722	2.16e-05	-1.57	98
Cluster Two				
RPLP0	ILMN1709880	5.23e-14	1.50	86
UQCRQ	ILMN1666471	1.38e-12	1.53	92
RPL9	ILMN1750507	1.49e-12	2.00	91
PSMA6	ILMN2151818	1.64e-12	1.54	93
RPL9	ILMN1769277	5.61e-12	2.06	92
RPS17	ILMN1664610	1.30e-11	1.60	84
HINT1	ILMN1807710	1.34e-11	1.71	90
COX7C	ILMN1798189	1.66e-11	1.72	89
NDUFA4	ILMN1751258	2.20e-11	1.60	92
RPS17	ILMN2207539	4.28e-11	1.68	81
RPS17	ILMN2207533	4.42e-11	1.52	79
RSL24D1	ILMN2175465	2.69e-10	1.72	95
EVI2A	ILMN1733579	2.84e-10	1.66	99
COMMD6	ILMN1777378	3.17e-10	1.80	79
RPL7	ILMN1687738	4.86e-10	1.68	90
RPL31	ILMN1754195	1.11e-09	1.62	97
CAPZA2	ILMN1768870	2.65e-09	1.52	92
LY96	ILMN1724533	2.66e-09	1.58	87
ARGLU1	ILMN1788468	9.77e-09	1.52	89
RPL23	ILMN1755115	1.36e-08	1.57	96
RPS27	ILMN1652955	2.31e-07	1.61	91
RPL31	ILMN1659405	6.32e-07	1.51	89
TMEM123	ILMN1724139	1.42e-06	1.52	84
DEFA1	ILMN1693262	9.41e-05	1.53	94

Table 7.11: $u^{[2]}$ ordering from SVD of schizophrenia and control sample expression data results: t-test on first and last 1000 nodes (named cluster one and two, respectively) of $u^{[2]}$ ordering showing genes with significant p-value and fold change.

Official Gene Symbol	Illumina ID	p-Value	Fold Change	Percentile Abundance
Cluster One				
RPLP0	ILMN1709880	5.23e-14	1.50	99
UQCRQ	ILMN1666471	1.38e-12	1.53	96
RPL9	ILMN1750507	1.49e-12	2.00	98
PSMA6	ILMN2151818	1.64e-12	1.54	96
RPS17P16	ILMN1664610	1.30e-11	1.60	98
HINT1	ILMN1807710	1.34e-11	1.71	96
COX7C	ILMN1798189	1.66e-11	1.72	96
NDUFA4	ILMN1751258	2.20e-11	1.60	95
RPS17	ILMN2207539	4.28e-11	1.68	97
RPS17	ILMN2207533	4.42e-11	1.52	99
RSL24D1	ILMN2175465	2.69e-10	1.72	90
EVI2A	ILMN1733579	2.84e-10	1.66	88
COMMD6	ILMN1777378	3.17e-10	1.80	91
RPL31	ILMN1754195	1.11e-09	1.62	95
LY96	ILMN1724533	2.66e-09	1.58	91
ARGLU1	ILMN1788468	9.77e-09	1.52	96
RPL23	ILMN1755115	1.36e-08	1.57	92
RPL31	ILMN1659405	6.32e-07	1.51	89
EIF1AY	ILMN1755537	1.15e-06	1.64	87
RPL9	ILMN2408415	2.21e-06	1.53	90
RPS4Y1	ILMN1783142	1.03e-04	2.48	96
Cluster Two				
DNAJC25	ILMN1757074	1.16e-05	1.51	92
GNG10	ILMN1757074	1.16e-05	1.51	92
RNF213	ILMN1749722	2.16e-05	-1.57	98
DEFA1B	ILMN2102721	2.87e-04	1.51	98
DEFA1	ILMN1725661	3.19e-04	1.50	98
DEFA3	ILMN2165289	3.35e-04	1.58	98
DEFA1B	ILMN2193213	3.68e-04	1.53	99

Table 7.12: $u^{[3]}$ schizophrenia and control sample expression data results: t-test on first and last 1000 nodes (named cluster one and two, respectively) of $u^{[3]}$ ordering showing genes with significant p-value and fold change.

In terms of specific results, a group of defensin genes (*DEFA1*, *DEFA1B*×2, *DEFA3*) appear with increased expression in cluster 2 of the $u^{[3]}$ ordering. This result mirrors a result found in a study of peripheral blood (the cellular component of whole blood) and plasma where multiple α-defensins showed increased expression in individuals with schizophrenia [34]. The authors of this study also remark that the increased gene expression of α-defensins was observed in in the asymptomatic (monozygotic) twin, suggesting that the result represents a potential diagnostic opportunity for susceptibility of schizophrenia.

Additionally, neurogranin (*NRGN*) appears at the very end location of this ordering - this is important as neurogranin has strong links with schizophrenia, and has appeared with significance in GWAS studies [197]. This cross-over of results from GWAS and gene expression is unusual and potentially important.

Schizophrenia and Antipsychotic free

Tables 7.13 and 7.14 show results for t-test comparing schizophrenia with antipsychotic-free and values for mean fold change. This time there is significant overlap between cluster two in the $u^{[2]}$ ordering and cluster $u^{[3]}$ ordering. We can also see in these results that the number of genes differentiating between treated schizophrenia and antipsychotic-free patients (29 in $u^{[2]}$ and 19 in $u^{[3]}$ vs 3 and 4 respectively) is much larger than those between the antipsychotic-free and control groups. These numbers could signify that medication has a significant impact on gene expression, hence there is much variation between these two cohorts which the t-test and fold change can identify. This is an interesting proposition, as it becomes unclear if the genes identified are due to the condition, due to the medication or due to some extraneous factor that alters gene expression but could be associated with the medication. As mentioned in the introduction to this thesis, gene expression is a highly dynamic process that changes in response to a multitude of stimuli and states. The medication given to treat schizophrenia has a wide array of adverse effects - many of which (e.g., lethargy, weight gain) could have result in second order changes on gene expression.

It is also worth re-stating the fact that antipsychotic-free patients are likely in very early stages of the condition (since they have not yet been assigned treatment), and so have a different biological presentation because the condition progresses afterwards.

Official Gene Symbol	Illumina ID	p-Value	Fold Change	Percentile Abundance
Cluster One				
FOLR3	ILMN1730454	3.53e-05	1.67	96
KIR3DL2	ILMN2190842	9.07e-05	-1.71	89
CDC14B	ILMN1733559	2.37e-04	-2.00	99
ATP6V0C	ILMN1773849	4.81e-04	-1.52	96
SEMA3E	ILMN2154322	6.16e-04	-1.53	97
BCL2L1	ILMN1654118	1.06e-03	-1.52	98
RNF213	ILMN1749722	1.98e-03	-1.76	98
GATSL3	ILMN2098418	2.19e-03	-1.77	94
EPB49	ILMN1671686	3.02e-03	-1.65	95
SMOX	ILMN2367258	8.30e-03	-1.50	89
Cluster Two				
CCDC72	ILMN1707783	5.74e-07	1.53	92
UQCRQ	ILMN1666471	7.78e-07	1.78	96
C17ORF61	ILMN2201533	1.01e-06	1.51	94
NDUFB2	ILMN2117330	1.65e-06	1.52	96
COMMD6	ILMN1777378	3.51e-06	1.90	91
TMCO1	ILMN1793829	4.12e-06	1.55	93
RPL35P5	ILMN1788742	1.39e-05	1.51	99
RSL24D1	ILMN2175465	2.26e-05	1.72	90
RPL31P17	ILMN1754195	3.57e-05	1.72	95
PSMA6	ILMN2151818	3.74e-05	1.54	96
LY96	ILMN1724533	2.52e-04	1.58	91
RPS17	ILMN2207539	3.37e-04	1.74	97
COX7C	ILMN1798189	3.79e-04	1.78	96
RPL9	ILMN1750507	4.93e-04	2.01	98
TPT1	ILMN1789614	6.68e-04	1.61	99
HINT1	ILMN1807710	7.04e-04	1.69	96
RPS29P18	ILMN1739263	7.99e-04	1.60	98
CD52	ILMN2208903	2.15e-03	1.58	98
CLC	ILMN1654875	3.86e-02	1.52	97

Table 7.13: $u^{[2]}$ ordering from SVD of schizophrenia and antipsychotic-free sample expression data results: t-test on first and last 1000 nodes (named cluster one and two, respectively) of $u^{[3]}$ ordering.

Official Gene Symbol	Illumina ID	p-Value	Fold Change	Percentile Abundance
Cluster One				
CCDC72	ILMN1707783	5.74e-07	1.53	92
UQCRQ	ILMN1666471	7.78e-07	1.78	96
C17ORf61	ILMN2201533	1.01e-06	1.51	94
NDUFB2	ILMN2117330	1.65e-06	1.52	96
COMMD6	ILMN1777378	3.51e-06	1.90	91
RPS9	ILMN2038772	7.54e-06	1.55	99
RPL35P5	ILMN1788742	1.39e-05	1.51	99
RSL24D1	ILMN2175465	2.26e-05	1.72	90
RPL31	ILMN1754195	3.57e-05	1.72	95
PSMA6	ILMN2151818	3.74e-05	1.54	96
RPS17	ILMN2207539	3.37e-04	1.74	97
COX7C	ILMN1798189	3.79e-04	1.78	96
RPL9	ILMN1750507	4.93e-04	2.01	98
TPT1	ILMN1789614	6.68e-04	1.61	99
HINT1	ILMN1807710	7.04e-04	1.69	96
RPS29P18	ILMN1739263	7.99e-04	1.60	98
CD52	ILMN2208903	2.15e-03	1.58	98
LY6E	ILMN1695404	4.02e-02	-1.51	96
Cluster Two				
FCGR3B	ILMN2134453	1.44e-04	-1.64	95

Table 7.14: $u^{[3]}$ ordering from SVD of schizophrenia and antipsychotic-free sample expression data results: t-test on first and last 1000 nodes (named cluster one and two, respectively) of $u^{[3]}$ ordering showing genes with significant p-value and fold change.

7.5 GSVD: Identification of Mutually Exclusive Features in Sample Groups

The SVD is useful for identifying clusters that drive structure within a data set, but in carrying on with the theme of this thesis of examining combinations of networks the GSVD, will be applied as in previous sections. This allows the comparison of network A with network B - identifying structures that are mutually exclusive. To use the GSVD on this data, the complete network was split into sub-samples based on the sample status. That is, schizophrenia and antispychotic-free samples can initially be paired in network A, with the controls comprising network B. In addition, a similar comparison was made between further sub-groups, e.g., schizophrenia vs antipsychotic-free and both, in turn, vs control. This provides an interesting opportunity to compare the output from exploratory analyses of the SVD and a more direct result (which makes account of the sample state by dividing the patient groups).

7.5.1 Singular Vectors from the GSVD

This section shows singular vectors from the GSVD of multiple combinations of data sub sets. From these figures we can visually check which ends of the vectors have significant deviation from the midpoint - and compare this with results from comparison with the gene list database in the following section.

Figure 7.14 shows results for GSVD of schizophrenia (including antipsychotic-free) and control subsets.

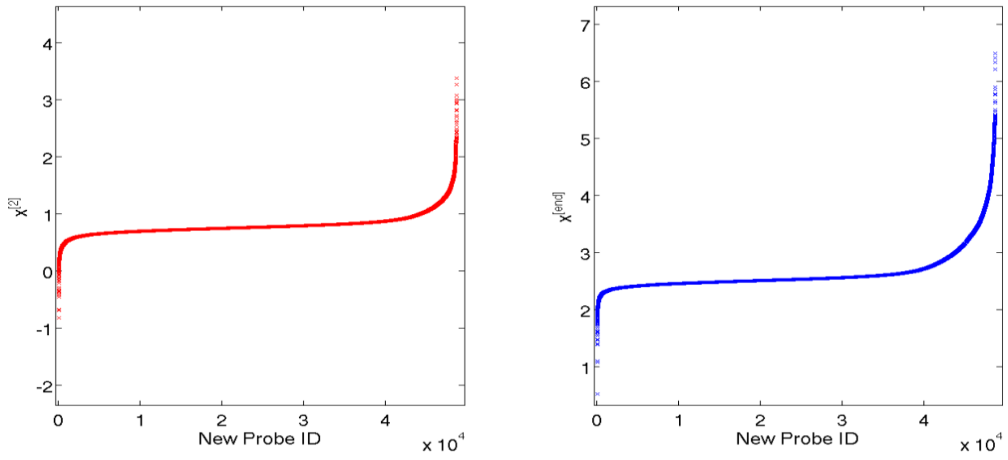

Figure 7.14: Components of the singular vector $x^{[2]}$ for control and $x^{[end]}$ for schizophrenia in increasing order from GSVD of the data.

Figure 7.15 has plots for schizophrenia (not including antipsychotic-free) compared with control. These results show a large bias towards nodes on the left hand side - particularly for $x^{[2]}$. This suggests that significant genes are more likely to be found on this side, we can confirm this in the next section.

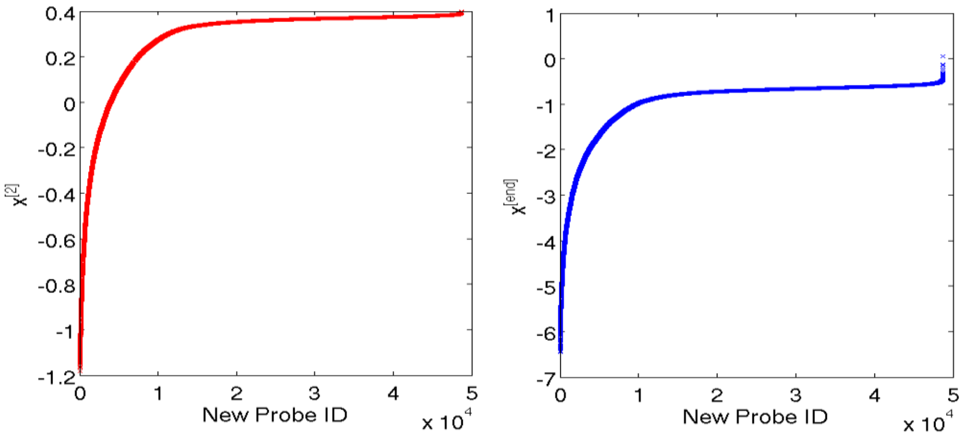

Figure 7.15: Components of the singular vector $x^{[2]}$ for control and $x^{[end]}$ for schizophrenia in increasing order from GSVD of the data.

Figure 7.16 shows antipsychotic-free vs control vectors. There is a large number of nodes separated on both ends of each vector.

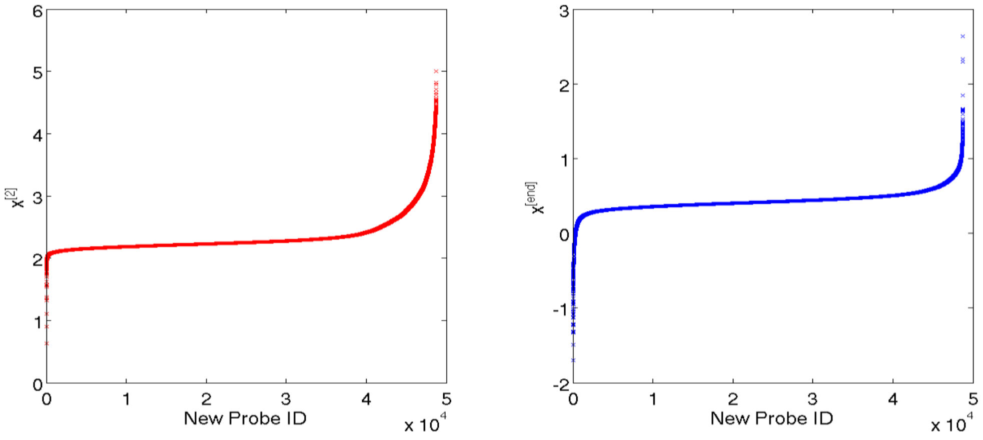

Figure 7.16: Components of the singular vector $x^{[2]}$ for control and $x^{[end]}$ for antipsychotic-free in increasing order from GSVD of the data.

The final comparison in Figure 7.17 is with schizophrenia and antipsychotic-free subsets. As with the comparison of schizophrenia and control sets in Figure 7.15, there is a large bias in separation towards one side of the vector. The next section will show if this matches probability testing.

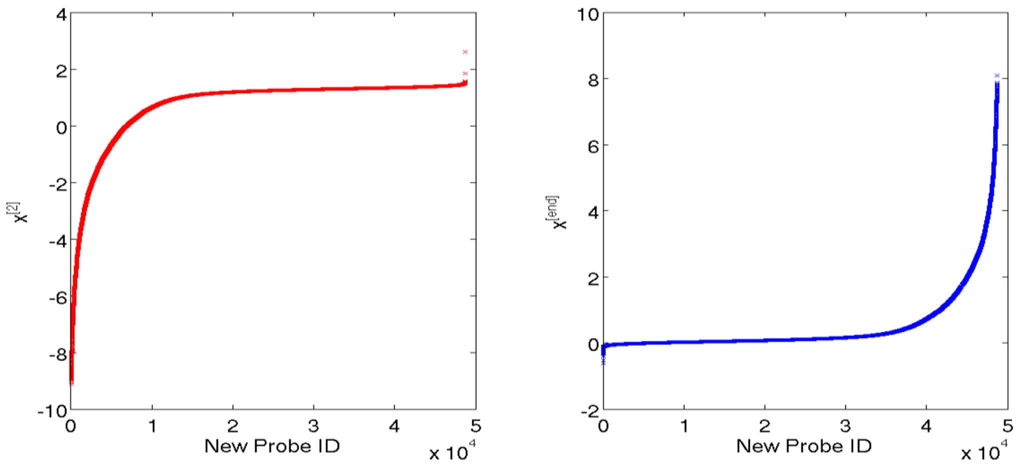

Figure 7.17: Components of the singular vector $x^{[2]}$ for schizophrenia and $x^{[end]}$ for antipsychotic-free in increasing order from GSVD of the data.

7.5.2 Distribution of Variance

In a comparison with Section 7.4.1, we can check the orderings generated from the GSVD to examine the behaviour of the variance. Figure 7.18 shows variance reordered using the vectors for reordering the combination of schizophrenia + antipsychotic-free (A) and control (B).

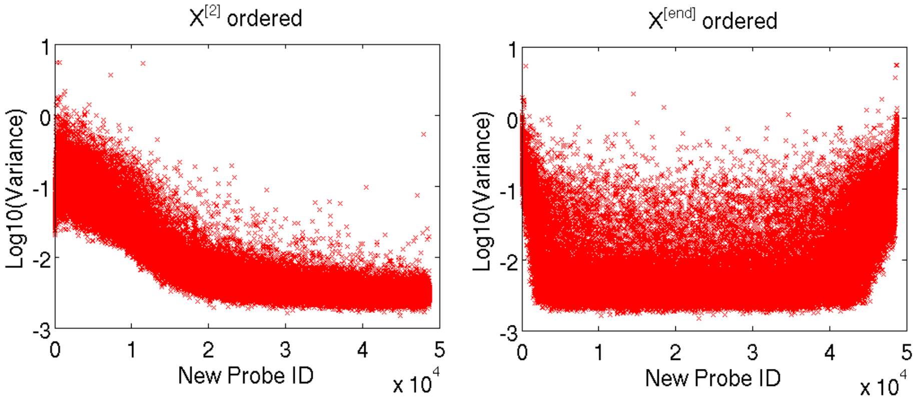

Figure 7.18: $x^{[2]}$ and $x^{[end]}$ reordering of variance across samples for each probe, showing a mixture of low and high variance genes are responsible for structure in the data.

This time, as compared to Figure 7.11, we can see that in finding differences between the two data sets, the GSVD allows for a more significant number of low variance probes to appear at the ends of the vectors. In fact, if we zoom in on the ends of Figures 7.11 and 7.18 we actually see that, at the extreme end, the SVD has no genes from the lowest percentiles of variance. The GSVD, however, does- we suggest that this is an illustration of the GSVD identifying different structures that are perhaps even less driven purely by variance.

7.5.3 Comparison with Gene List Database

As with with the SVD in Section 7.4.2, this section contains comparisons of the probe orderings from the GSVD with the Association, Expression and Literature lists from SZGR database.

In Table 7.16 it is interesting that there are no significant results for any of the lists in the $x^{[2]}$ End. Figure 7.15 shows that the important probes in this $x^{[2]}$ ordering have been driven to one side - the start of the vector. This was as expected given the separation of genes to this side of the vector seen in Figure 7.15 in the previous section. Table 7.17 shows multiple significant results, including for association studies. As compared to the SVD results for antipsychotic-free and control samples, the GSVD has identified more association derived genes. In Table 7.18 we see no significance in the end of $x^{[2]}$ ordering, and no significance in the start of the $x^{[end]}$ ordering. We can see from the components of those vectors in Figure 7.17 that these portions show little deviation, meaning they are of limited importance in describing the data - and so the lack of significance is in line with our expectations. This is an extra verification for this approach.

Association	$x^{[2]}$ Start	$x^{[2]}$ Mid	$x^{[2]}$ End	$x^{[end]}$ Start	$x^{[end]}$ Mid	$x^{[end]}$ End
Sum	15	2	6	9	4	11
Sum with repeats	17	2	6	9	4	13
p-value	**0.0274**	0.993	0.889	0.709	0.939	0.122
Expression	$x^{[2]}$ Start	$x^{[2]}$ Mid	$x^{[2]}$ End	$x^{[end]}$ Start	$x^{[end]}$ Mid	$x^{[end]}$ End
Sum	37	12	15	11	3	32
Sum with repeats	48	13	12	13	3	42
p-value	$<$ **0.001**	0.968	0.932	0.799	1.000	$<$ **0.001**
Literature	$x^{[2]}$ Start	$x^{[2]}$ Mid	$x^{[2]}$ End	$x^{[end]}$ Start	$x^{[end]}$ Mid	$x^{[end]}$ End
Sum	46	15	23	24	12	44
Sum with repeats	128	34	41	52	19	127
p-value	$<$ **0.001**	0.983	0.869	0.738	1.000	$<$ **0.001**

Table 7.15: Comparison of selected genes from left singular vectors $x^{[2]}$ and $x^{[end]}$ of GSVD of schizophrenia + antipsychotic-free vs control samples. $x^{[2]}$ orders for controls not schizophrenia patients, $x^{[end]}$ for schizophrenia patients not controls.

Association	$x^{[2]}$ Start	$x^{[2]}$ Mid	$x^{[2]}$ End	$x^{[end]}$ Start	$x^{[end]}$ Mid	$x^{[end]}$ End
Sum	15	6	11	11	7	3
Sum with repeats	17	7	14	16	7	3
p-value	**0.0274**	0.932	0.146	**0.041**	0.786	0.998
Expression	$x^{[2]}$ Start	$x^{[2]}$ Mid	$x^{[2]}$ End	$x^{[end]}$ Start	$x^{[end]}$ Mid	$x^{[end]}$ End
Sum	32	16	15	25	12	32
Sum with repeats	44	16	20	39	19	42
p-value	$<$ **0.001**	0.969	0.900	$<$ **0.001**	0.874	$<$ **0.001**
Literature	$x^{[2]}$ Start	$x^{[2]}$ Mid	$x^{[2]}$ End	$x^{[end]}$ Start	$x^{[end]}$ Mid	$x^{[end]}$ End
Sum	44	22	23	38	31	43
Sum with repeats	124	47	52	87	55	92
p-value	$<$ **0.001**	0.894	0.730	$<$ **0.001**	0.555	$<$ **0.001**

Table 7.16: Comparison of selected genes from left singular vectors $x^{[2]}$ and $x^{[end]}$ of GSVD of schizophrenia vs control samples. $x^{[2]}$ orders for controls not schizophrenia, $x^{[end]}$ for schizophrenia not controls.

Association	$x^{[2]}$ Start	$x^{[2]}$ Mid	$x^{[2]}$ End	$x^{[end]}$ Start	$x^{[end]}$ Mid	$x^{[end]}$ End
Sum	13	13	12	15	9	3
Sum with repeats	17	13	15	15	10	3
p-value	**0.036**	0.155	0.091	0.094	0.446	**0.050**
Expression	$x^{[2]}$ Start	$x^{[2]}$ Mid	$x^{[2]}$ End	$x^{[end]}$ Start	$x^{[end]}$ Mid	$x^{[end]}$ End
Sum	12	20	31	24	16	32
Sum with repeats	15	22	45	35	19	52
p-value	**0.985**	0.883	< **0.001**	**0.036**	0.883	< **0.001**
Literature	$x^{[2]}$ Start	$x^{[2]}$ Mid	$x^{[2]}$ End	$x^{[end]}$ Start	$x^{[end]}$ Mid	$x^{[end]}$ End
Sum	27	27	39	47	21	21
Sum with repeats	70	46	115	124	53	43
p-value	**0.040**	0.919	< **0.001**	< **0.001**	0.664	0.984

Table 7.17: Comparison of selected genes from left singular vectors $x^{[2]}$ and $x^{[end]}$ of GSVD of antipsychotic-free vs control samples. $x^{[2]}$ orders for controls not antipsychotic-free, $x^{[end]}$ for antipsychotic-free not controls.

Association	$x^{[2]}$ Start	$x^{[2]}$ Mid	$x^{[2]}$ End	$x^{[end]}$ Start	$x^{[end]}$ Mid	$x^{[end]}$ End
Sum	12	13	14	11	9	13
Sum with repeats	14	13	15	12	9	15
p-value	0.140	0.138	0.0860	0.341	0.578	0.097
Expression	$x^{[2]}$ Start	$x^{[2]}$ Mid	$x^{[2]}$ End	$x^{[end]}$ Start	$x^{[end]}$ Mid	$x^{[end]}$ End
Sum	36	14	23	15	16	30
Sum with repeats	56	19	26	15	24	48
p-value	< **0.001**	0.868	0.463	0.987	0.554	< **0.001**
Literature	$x^{[2]}$ Start	$x^{[2]}$ Mid	$x^{[2]}$ End	$x^{[end]}$ Start	$x^{[end]}$ Mid	$x^{[end]}$ End
Sum	54	23	34	26	28	54
Sum with repeats	122	47	62	56	50	130
p-value	< **0.001**	0.903	0.226	0.561	0.790	< **0.001**

Table 7.18: Comparison of selected genes from left singular vectors $x^{[2]}$ and $x^{[end]}$ of GSVD of antipsychotic-free vs schizophrenia samples. $x^{[2]}$ orders for schizophrenia not antipsychotic-free, $x^{[end]}$ for antipsychotic-free not schizophrenia.

7.5.4 Hypothesis Testing: T-test and Fold Change

We now show results (Tables 7.20-7.25) from the GSVD. This analysis involves comparing subsets of the data set, in the following order:-

Schizophrenia (including antipsychotic-free)	$< - >$	Control
Schizophrenia (excluding antipsychotic-free)	$< - >$	Control
Antipsychotic-free	$< - >$	Schizophrenia
Antipsychotic-free	$< - >$	Control

Schizophrenia (including antipsychotic-free) and Controls

Results in Tables 7.20 and 7.21 show a number of interesting genes. Neurogranin *NRGN* has been linked to schizophrenia in males in gene association study [184]. In a postmortem study of the dorsolateral prefrontal cortex and thalamus it was found that the histidine triad nucleotide-binding protein (*HINT1*) is downregulated in the dorsolateral prefrontal cortex and upregulated in the thalamus [212]. NADH Dehydrogenase (Ubiquinone) 1-Alpha Subcomplex (*NDUFA4*) is overexpressed in the prefrontal cortex, and *HINT1* shows decreased expression [214]. Decreased levels of expression of spermine oxidase (*SMOX*) have been found in suicide completers and there is also an association between SNPs of *SMOX* and mood disorder [63, 64]. Finally, probability testing of the SVD of the complete data set yielded 33 hits across (including genes that appear more than once across both orderings) two orderings.

This time, the GSVD has 21 genes meet the test thresholds. This reduction may be as a result of the mutually exclusive nature of GSVD orderings - the SVD identifies structurally interesting genes where the GSVD acts to distinguish between data subsets.

Official Gene Symbol	Illumina ID	p-value	Fold Change	Percentile Abundance
Cluster One				
NRGN	ILMN1705686	4.30e-08	-1.55	96
EIF1AY	ILMN1755537	3.01e-07	1.65	87
RPS4Y1	ILMN1783142	5.09e-05	2.50	96
Cluster Two				
RPL9	ILMN1750507	1.23e-10	1.82	98
RPS17P16	ILMN1664610	2.04e-10	1.52	98
NDUFA4	ILMN1751258	2.62e-10	1.52	95
HINT1	ILMN1807710	3.02e-10	1.60	96
EVI2A	ILMN1733579	4.22e-10	1.59	88
COX7C	ILMN1798189	8.32e-10	1.60	96
RPS17	ILMN2207539	1.49e-09	1.56	97
RSL24D1	ILMN2175465	2.39e-09	1.60	90
COMMD6	ILMN1777378	3.05e-09	1.65	91
RPL31	ILMN1754195	1.83e-08	1.51	95

Table 7.20: $x^{[2]}$ ordering from GSVD for control not schizophrenia (including antipsychotic-free) sample expression data results: t-test on first and last 1000 nodes (named clusters one and two, respectively). This table shows genes with a significant p-value and fold change.

Schizophrenia (not including antipsychotic-free) and Controls

Table 7.22 shows an upregulation of differential expression of *DEFA1B* in peripheral blood mononuclear cells (PBMC) has been found in schizophrenia patients compared to controls [68] (fold change 1.73 - here we have found a fold change of 1.51).

Official Gene Symbol	Illumina ID	p-value	Fold Change	Percentile Abundance
Cluster One				
RPLP0	ILMN1709880	5.23e-14	1.50	99
RPS17P16	ILMN1664610	1.30e-11	1.60	98
RPS17	ILMN2207533	4.42e-11	1.52	99
NRGN	ILMN1705686	1.59e-09	-1.52	96
RPS4Y1	ILMN1783142	1.03e-04	2.48	96
DEFA1B	ILMN2102721	1.87e-03	1.51	98

Table 7.22: $x^{[2]}$ ordering from GSVD for control not schizophrenia (not antipsychotic-free) sample expression data results: t-test on first and last 1000 nodes (named clusters one and two, respectively). This table shows genes with a significant p-value and fold change.

Official Gene Symbol	Illumina ID	p-value	Fold Change	Percentile Abundance
Cluster One				
SMOX	ILMN2367258	1.51e-07	-1.51	98
Cluster Two				
RPL9	ILMN1750507	1.23e-10	1.82	98
RPS17P16	ILMN1664610	2.04e-10	1.52	95
NDUFA4	ILMN1751258	2.62e-10	1.52	96
HINT1	ILMN1807710	3.02e-10	1.60	96
COX7C	ILMN1798189	8.32e-10	1.60	90
RPL9	ILMN2408415	8.78e-07	1.50	96
RPS4Y1	ILMN1783142	5.09e-05	2.50	89

Table 7.21: $x^{[end]}$ ordering from GSVD for schizophrenia (including antipsychotic-free) not control sample expression data results: t-test on first and last 1000 nodes (named clusters one and two, respectively). This table shows genes with a significant p-value and fold change.

Official Gene Symbol	Illumina ID	p-value	Fold Change	Percentile Abundance
Cluster One				
SMOX	ILMN2367258	1.01e-09	-1.51	89
RPB49	ILMN1671686	1.60e-06	-1.55	95
RNF213	ILMN1749722	2.16e-05	-1.47	98
DEFA1B	ILMN2102721	1.87e-03	1.51	98
DEFA1	ILMN1725661	2.19e-03	1.50	98
DEFA3	ILMN2165289	2.35e-03	1.51	98
DEFA1	ILMN1679357	3.41e-03	1.52	98
Cluster Two				
RPLP0	ILMN1709880	5.23e-14	1.50	99
UQCRQ	ILMN1666471	1.38e-12	1.53	96
RPL9	ILMN1750507	1.49e-12	2.00	98
PSMA6	ILMN2151818	1.64e-12	1.54	96
RPS17P16	ILMN1664610	1.30e-11	1.60	98
HINT1	ILMN1807710	1.34e-11	1.71	96
COX7C	ILMN1798189	1.66e-11	1.72	96
NDUFA4	ILMN1751258	2.20e-11	1.60	95
RPS17	ILMN2207539	4.28e-11	1.68	97
RPS17	ILMN2207533	4.42e-11	1.52	99
RSL24D1	ILMN2175465	2.69e-10	1.72	90
EVI2A	ILMN1733579	2.84e-10	1.66	88
COMMD6	ILMN1777378	3.17e-10	1.80	91
RPL31	ILMN1754195	1.11e-09	1.62	95
CAPZA2	ILMN1768870	2.65e-09	1.52	94
LY96	ILMN1724533	2.66e-09	1.58	91
ARGLU1	ILMN1788468	9.77e-09	1.52	96
RPL23	ILMN1755115	1.36e-08	1.57	92
RPL31	ILMN1659405	6.32e-07	1.51	89
EIF1AY	ILMN1755537	1.15e-06	1.64	87
TMEM123	ILMN1724139	1.42e-06	1.52	96
RPL9	ILMN2408415	2.21e-06	1.53	90
RPS4Y1	ILMN1783142	1.03e-04	2.48	96

Table 7.23: $x^{[end]}$ ordering from GSVD for schizophrenia (not antipsychotic-free) not control sample expression data results: t-test on first and last 1000 nodes (named clusters one and two, respectively). This table shows genes with a significant p-value and fold change.

In an ordering for schizophrenia not control, cluster one is comprised mainly of defensin related genes - with multiple probes highlighting *DEFA1/B* and *DEFA3* - this mirrors the result found in the $u^{[3]}$ ordering from the SVD of schizophrenia + control samples in Table 7.12 where multiple *DEFA1* repeats were found. As with the previous note, increased expression of *DEFA1B* has previously been shown in schizophrenia [68]. *SMOX* appears in cluster one, as with the analysis in Section 7.5.4. Similarly with Section 7.5.4 *NDUFA4* appears again.

Antipsychotic-free and Schizophrenia (not including controls)

SEMA3E appears in the ordering for schizophrenia, not antipsychotic-free in Table 7.24 as well as the complimentary ordering for antipsychotic-free not schizophrenia in Table 7.25. *SEMA3E* has been shown, in a closely controlled and matched study, to have significantly decreased expression in measurements from the prefrontal cortex [6] and this table shows similarly decreased expression in blood. *GCNT1* is linked with blood presssure, and *RNF213* with Moyamoya disease [104], a disease whereby arteries in the brain are constricted causing reduced blood flow. It has been reported that patients with Moyamoya are commonly misdiagnosed as having schizophrenia [119]. *RPS9P4* has appeared in an association study of panic disorder [172].

Official Gene Symbol	Illumina ID	p-value	Fold Change	Percentile Abundance
Cluster One				
RPS9	ILMN2038772	7.54e-06	1.55	99
BCL6	ILMN1654118	1.06e-03	1.52	98
RPL35P5	ILMN1788742	1.39e-05	1.51	99
CDC14B	ILMN1733559	2.37e-04	-2.00	99
SEMA3E	ILMN2154322	6.16e-04	-1.53	97
XRCC2	ILMN2204909	6.41e-04	-1.53	98
TPT1	ILMN1789614	6.68e-04	1.61	99
RNF213	ILMN1749722	1.98e-03	-1.76	98
CLC	ILMN1654875	3.86e-02	1.52	97

Table 7.24: $x^{[2]}$ ordering from GSVD for antipsychotic-free not schizophrenia (no controls) sample expression data results: t-test on first and last 1000 nodes (named clusters one and two, respectively). This table shows genes with a significant p-value and fold change.

Official Gene Symbol	Illumina ID	p-value	Fold Change	Percentile Abundance
Cluster One				
RPS9	ILMN2038772	7.54e-06	1.55	99
RPS17	ILMN2207539	3.37e-04	1.74	97
CLC	ILMN1654875	3.86e-02	1.52	97
Cluster Two				
RPL31	ILMN1754195	3.57e-05	1.72	95
PSMA6	ILMN2151818	3.74e-05	1.54	96
KIR3DL2	ILMN2190842	9.07e-05	-1.71	89
FCGR3B	ILMN2134453	1.44e-04	-1.64	95
CDC14B	ILMN1733559	2.37e-04	-2.00	99
HIST1H4C	ILMN2075334	1.67e-03	1.56	98
FCAR	ILMN2279367	3.42e-04	-1.52	96
SEMA3E	ILMN2154322	6.16e-04	-1.53	97
XRCC2	ILMN2204909	6.41e-04	-1.53	98
RPS29P18	ILMN1739263	7.99e-04	1.60	98
BCL2L1	ILMN1654118	1.06e-03	-1.52	98
DDX17	ILMN2371590	1.22e-03	-1.63	95
PABPC1	ILMN2136133	1.36e-03	-2.01	95
GATSL3	ILMN2098418	2.19e-03	-1.77	94
EPB49	ILMN1671686	3.02e-03	-1.65	95
SMOX	ILMN2367258	8.30e-03	-1.50	89
LY6E	ILMN1695404	4.02e-02	-1.51	96

Table 7.25: $x^{[2]}$ ordering from GSVD for schizophrenia not antipsychotic-free (no controls) sample expression data results: t-test on first and last 1000 nodes (named clusters one and two, respectively). This table shows genes with a significant p-value and fold change.

Antipsychotic-free and Controls

The SVD analysis of the antipsychotic-free and control group yielded a small number of genes, and we proposed that this could be due to the low statistical power of such a small sample cohort (as is seen with the relatively much larger $p-$values) or a lack of variation between sample groups. Tables 7.24-7.25 show that the GSVD has similar results. A small number of genes meet the dual threshold. *CDC14B*, *PABPC1*, *RPS4Y1* and *DDX17* appear in both the SVD and GSVD orderings showing that there is an agreement between the approaches, though the genes are clustered differently - in the SVD, *PABPC1* and *CDC14B* appear at opposing ends of the vector, where they have migrated to the same cluster in the GSVD. Two of the genes that appear in the orderings in this section that were not found in the SVD (*LYPD2* and *SCGB3A1*) are both less abundant than any genes we have found so far.

Official Gene Symbol	Illumina ID	p-value	Fold Change	Percentile Abundance
Cluster One				
LYPD2	ILMN1724266	1.20e-04	1.53	79
Cluster Two				
DDX17	ILMN2371590	2.23e-04	-1.57	95
CDC14B	ILMN1733559	6.14e-04	-1.62	99
PABPC1	ILMN2136133	1.92e-03	-1.60	95
RPS4Y1	ILMN1783142	4.63e-03	-2.63	96

Table 7.26: $x^{[2]}$ ordering from GSVD for control not antipsychotic-free (no medicated schizophrenia samples) sample expression data results: t-test on first and last 1000 nodes (named clusters one and two, respectively). This table shows genes with a significant p-value and fold change.

Official Gene Symbol	Illumina ID	p-value	Fold Change	Percentile Abundance
Cluster One				
CDC14B	ILMN1733559	6.14e-04	1.62	99
Cluster Two				
SCGB3A1	ILMN1679666	1.22e-03	-1.57	76

Table 7.27: $x^{[end]}$ ordering from GSVD for antipsychotic-free not control (no medicated schizophrenia samples) sample expression data results: t-test on first and last 1000 nodes (named clusters one and two, respectively). This table shows genes with a significant p-value and fold change.

7.6 Node-Weighted Laplacian

This final section shows application of the node-weighted Laplacian, as developed in Chapter 6 to the pair of networks created from the microarray data and the KEGG/MetaCyc metabolic pathway networks. The merging of these networks necessitates that we work with a cross-correlation (across samples) of the microarray data:- instead of $M \in \mathbb{R}^{48743\times202}$ we use MM^T.

This network is then truncated to include only the probes measuring genes that are present in a second, metabolic network, B, components b_{ij}. So the degree of node j in the metabolic network B is $(D_b)_j = \sum_{i=1}^{N} b_{ij}$.

7.6.1 KEGG Metabolic Network

We now show results from the SVD of the node-weighted graph Laplacian constructed from the microarray data and the KEGG metabolic network. $B_K \in \mathbb{R}^{2312\times 202}$ The node-weighted Laplacian is formed:-

$$L_b = D_b(D - MM^T)D_b \tag{7.1}$$

Figure 7.19 show the ordered components of vectors from the node-weighted Laplacian - in the left picture a pair of nodes are separated at the lower end and a small group of nodes are separated at the higher end. In the right picture the results are similar, but reversed.

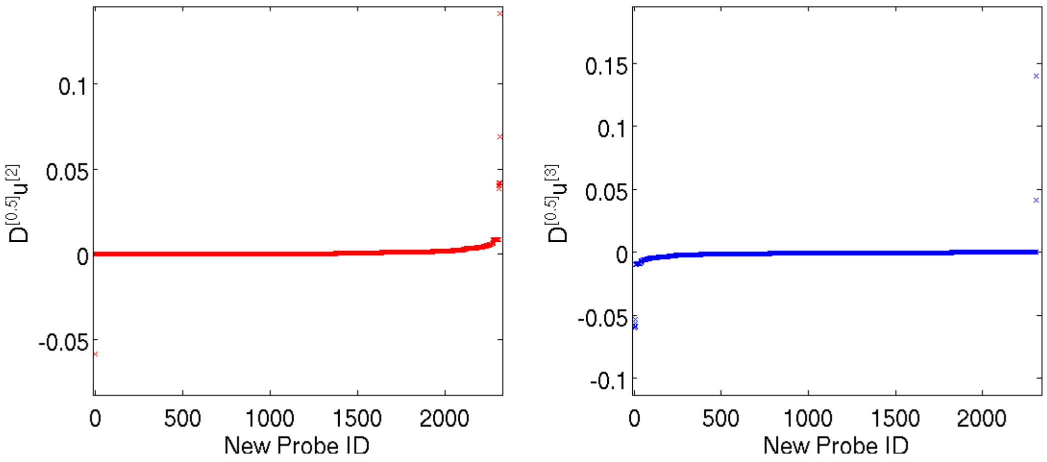

Figure 7.19: Components of $D^{[0.5]}u^{[2]}$ and $D^{[0.5]}u^{[3]}$ from SVD of complete data set, order by increasing size.

In Chapter 6 a reordering of the degree of the metabolic network was used to indicate whether the ordering is influenced by the data, rather than metabolic information. Figure 7.20 shows the reordering of D_b. Where we see a strong scattering of nodes in terms of degree, though there is a general trend line following increased degree. The shelf-like structures are explained in some way by the fact that many of the degrees used to weight the Laplacian are repeated - a feature which is increased since the metabolic networks are constructed using the official gene symbol and the experiments measure Illumina probe IDs. This means there are often multiple probe IDs for each official gene symbol.

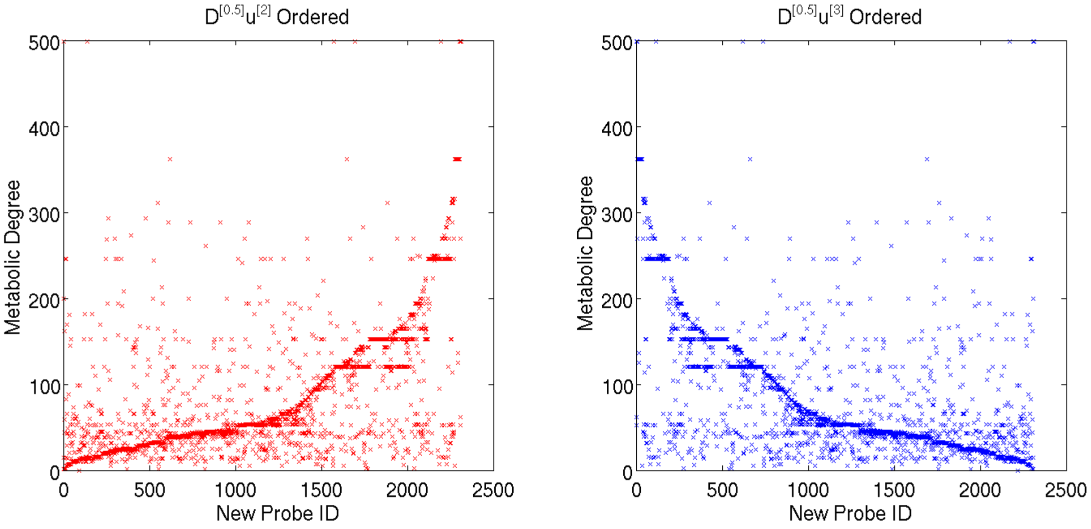

Figure 7.20: Metabolic degree reordered by vectors from the node-weighted Laplacian. There is clear structure in both pictures where high degree nodes are pushed towards one of the ends.

Next, results are shown for the hypothesis testing and fold change approach as used in previous sections. No probes met the fold change threshold (1.5), which is not unexpected given that the metabolic set of genes is a small subset of the original data so a reduced threshold of fold change 1.3 is chosen to allow for some observable results.

In Table 7.28 there are some lower percentile abundance genes than have been seen in previous results. Results in previous sections found genes with expression levels in often the 90+ percentile - this was explained by showing that variance across samples in this data tends to increase with increasing abundance. So the appearance of low abundance (*NFS1* at 42%, *PKM2* at 62% and *GSTM1* at 57%) is as a result of the influence of the metabolic network.

Official Gene Symbol	Illumina ID	p-value	Fold Change	Percentile Abundance
Cluster One				
NFS1	ILMN1703079	1.05e-07	-1.36	42
ST6GALNAC4	ILMN2297453	5.46e-05	-1.37	85
ST6GALNAC4	ILMN2413064	1.68e-04	-1.34	85
Cluster Two				
POLR2A	ILMN1737704	7.38e-13	-1.33	92
ATIC	ILMN1673991	4.45e-08	-1.30	90
PKM2	ILMN1672650	1.12e-06	-1.39	62
GSTM1	ILMN1762255	1.87e-02	-1.32	57
GSTM2	ILMN1713162	3.07e-02	-1.33	92

Table 7.28: Results from t-test and fold change threshold of $D^{[0.5]}u^{[2]}$. Test on first and last 200 nodes.

7.6.2 MetaCyc Metabolic Network

Equivalent results are now shown with the metabolic network B_M constructed from the MetaCyc database. The original MetaCyc network had the form $B \in \mathbb{R}^{411\times411}$. This time the node-weighted Laplacian is $L_b \in \mathbb{R}^{672\times672}$ due to the one-to-many relationship between probes and genes. Figure 7.21 shows components of the vectors from the node-weighted Laplacian, where we can see a small number of nodes deviating from the mid point at either end.

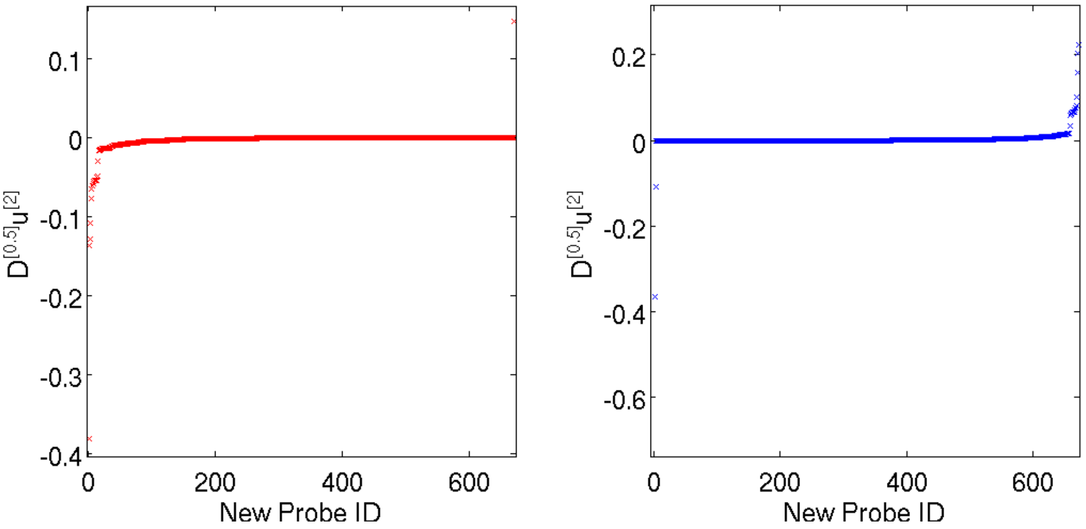

Figure 7.21: Components of $D^{[0.5]}u^{[2]}$ and $D^{[0.5]}u^{[3]}$ from SVD of complete data set

Figure 7.22 shows that the nodes very closely follow a pattern with ascending degree, with a pair of nodes separated in the top right corner of the left picture and top left of the picture on the right.

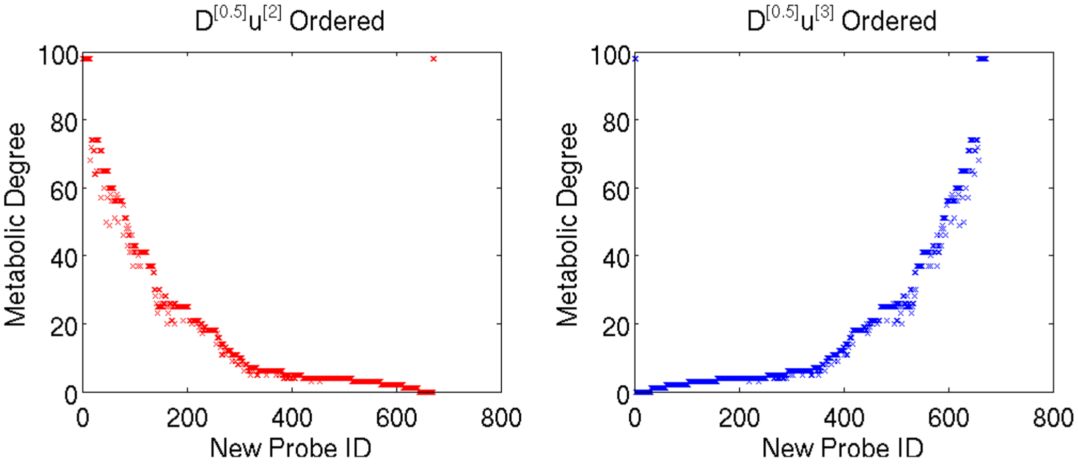

Figure 7.22: Metabolic degree reordered by vectors from the node-weighted Laplacian.

In the MetaCyc analysis the genes that appear at the ends of the ordering do not meet any reasonable thresholds for fold change. As seen in Figures 7.21-7.22, the nodes separated at the very end of the ordering match the highest metabolic degree nodes, suggesting that the experimental data does not contribute much to the structure.

7.7 Discussion

The inclusion of antipsychotic-free samples in this cohort is important since medication may introduce therapeutic effects (or adverse effects) through the alteration of expression profiles. There is a point of comparison for the medicated patients to develop understanding of drug interactions, as well as the areas of potential therapeutic interest which could pave the way for introduction of more closely targetted drug treatments. The inclusion of antipsychotic-free patients is particularly relevant when considering blood vs brain studies.

There are difficulties with antipsychotic-free samples in general, however. First, the fact that it is unethical to withold treatment from individuals in a disease state for the purpose of study means that, naturally, there are far fewer identified antipsychotic-free samples available. Related to this is the fact that it is likely that the majority of antipsychotic-free samples are taken at the point of diagnosis. As a result, there will be a lack of representation of progressed disease states and so the picture is skewed towards early stage symptoms. This could be useful to develop early-stage biomarkers that may be relevant for diagnosis, since there is some discussion of schizophrenia as a progressive illness [144].

The issue of sample size is apparent in the cohort used for this chapter, with a far smaller number of antipsychotic-free samples than either of the other experimental groups. One approach to alleviate this concern would be to perform a matching process whereby control samples are matched according some parameters (e.g., age and sex). Rather than taking additional steps to match, we have chosen to analyse a second data set that also presents a number of antipsychotic-free samples, with comparative numbers of control samples. Chapter 8 will deal with a second data set that contains only matched antipsychotic-free and control samples.

The approach we used to threshold on fold change and p-value produced some interesting results. Many genes are altered significantly in expression level across this large sample group. It is also noted that testing fold change has disadvantages in that we will only identify as significant genes that have a very high fold change across groups. This means that we will miss genes with a small, but potentially significant fold change, as well as potentially identifying false positives with natural variation, though the reasonably large sample size in this study should help to minimize that effect. Another previously mentioned issue is with using the t-test to examine large groups of genes. Ideally, the distribution of each probe in the microarray kit would be examined for normality, but when dealing with $O(10^3)$ probes this is impractical. As a compromise, we performed the non-parametric Mann-Whitney-Wilcoxon hypothesis test and compared results with the t-test. In the end there were few differences between the statistical testing techniques. This is potentially because the sample sizes are reasonably large - which encourages the p-values from the statistical testing to be very small with any difference. This means that the limiting factor with significance from the volcano plot approach is almost exclusively because of the threshold in fold change.

The first goal of this chapter was to establish whether blood is a viable as a medium for studying gene expression alterations relevant to schizophrenia. To do this, we compared clustering results with genes that have been previously implicated in schizophrenia. Results were positive - many of the orderings used attained significance (in p-value).

The SVD and GSVD have shown some consistency - there is clear overlap in orderings, yet the GSVD tends to cluster more statistically significant genes, as was expected since the GSVD takes information about the sample groups. The last goal of this chapter was potentially stratify patients or develop a diagnostically relevant list of genes. We have shown across the different orderings and comparisons that the antipsychotic-free samples are vastly different from medicated schizophrenia patients. This suggests that using gene expression as a diagnostic tool will be difficult - the vast majority of data in the public domain is on medicated schizophrenia patients, and if their gene expression profiles are as far removed from unmediated (or pre-diagnosis) schizophrenia patients as our results suggest then there needs to be further focus and investigation on the antipsychotic-free contingent. Whilst this is true from a diagnostic perspective in terms of investigating the condition as a whole, the disparity between medicated and unmedicated patients generates several points of interest for potential future study. Unfortunately with this study we lacked information about current patient states (for instance, how effectively their current medication paradigm was at tackling their symptoms). This extra information could allow for identification of the nature, in terms of expression, of therapeutic pathways. The information on which medication were currently and previously prescribed for the patients receiving treatment was also unavailable. This could be used in a similar way to potentially identify therapeutic propensity, and better understand therapeutic action of specific drugs.

The node-weighted Laplacian generated some results, but was of limited success. This suggests that the types of structure the node-weighted Laplacian identifies are not present within this data set - that is, subsets of genes that have high degrees in a metabolic network and drive structure in a gene expression network. There are of course metrics other than the degree we could use to assess the metabolic 'importance' of a gene, but this question is left open.

As a final note, it may be possible that the SVD identifies genes in the list of those implicated in schizophrenia simply because those genes vary more than others within the data. This would mean the p-value is less informative than it otherwise could be, as genes could be showing up in the clusters due to their natural variance rather than their significance in this data set. It is also possible that gene expression changes may be adaptive responses to a disease state, rather than causative - though even in this case the changes may be able to provide diagnostic insight.

7.8 Summary

In this chapter we have seen spectral clustering with the SVD and GSVD on a data set containing both medicated and unmedicated schizophrenia patients together with a control group. We have also seen application of the node-weighted Laplacian with two different metabolic networks. The results have found commonalities between differential expression in whole human blood and genes that have been previously implicated in the literature, which is typically made up of brain tissue studies.

The final approaches to analysis show differences in hypothesis testing results between medicated and antipsychotic-free samples, suggesting medication has a significant effect on differential expression.

The following chapter will use the same orderings and divisions (that is, the separation of the first and last 1000 nodes for investigation) in a different type of analysis. The focus will shift from individual probes and genes to the summary approach of gene ontology. In this way we can identify overrepresented biological functions, or processes.

Chapter 8

Clustering and Gene Ontology of Antipsychotic-free Schizophrenia: Replicate Results

This chapter is, in part, a replication of the process of analysis as Chapter 7 on a second 'whole blood' data set. The goal of this chapter is to further explore the antipsychotic-free schizophrenia state with a view to exploring possible effect of medication on gene expression measurements. We also compare results with the previous chapter to check for consistency across experiments.

8.1 Introducing the Data

The data used in this chapter was once again obtained from a public repository for microarray data. The data was gathered and uploaded by the same research group as the study in Chapter 7 (see [43]) and so it is reasonable to assume consistency in experimental protocol, which will minimize some sources of error. There are two patient groups in this data set - antipsychotic-free patients with schizophrenia

	Samples	Mean Age	Num. males	Num. females
Antipsychotic-free	15	30.86	11	4
Control	22	29.45	16	6

Table 8.1: Summary information on the data set adapted from [43]: age and gender information

and control groups. This makes the analysis more straightforward than Chapter 7, where more comparisons were there were more factors to consider.

The data was gathered using a different Illumina chip than the other expression data we investigated:-

- Illumina HumanRef-8 v3.0 Expression BeadChip

This chip has a different probe set - there are $24,526$ probes as compared to the $48,743$ measured in Chapter 7. Thus, where [43] chose to merge the two data sets discarding probes that were not present in both, we have chosen to examine the two data sets separately, in their entirety. Table 8.1 provides a summary of sample information for this data. It can be seen that the overall sample size (37 total) is reduced compared to Chapter 7 where there are 202 samples, but that this time the control sample size is similar to the antipsychotic-free group.

Figure 8.1 shows the total expression across all probes, for each sample. We can see that the controls and antipsychotic-free groups have similar levels of variation, which is useful to check as a first point in assuring that there is no obvious inconsistency in the experiment across samples.

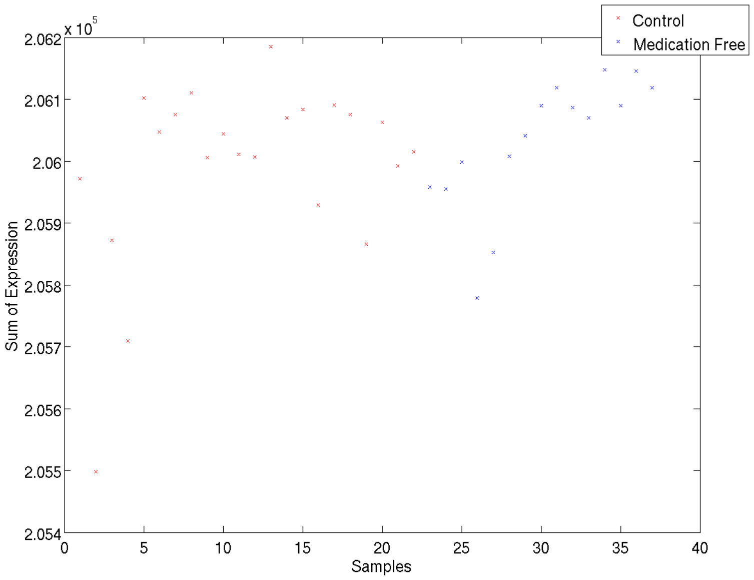

Figure 8.1: Sum of expression across all probes for each patient samples. There is no obvious separation in expression across sample groups.

8.2 Clustering with the SVD

We first take the SVD of the complete data set (both sample groups and all genes), and show the Fiedler vector and next singular vector in Figure 8.2. We can see sharp deviations at the ends of both vectors - which we will investigate as clusters.

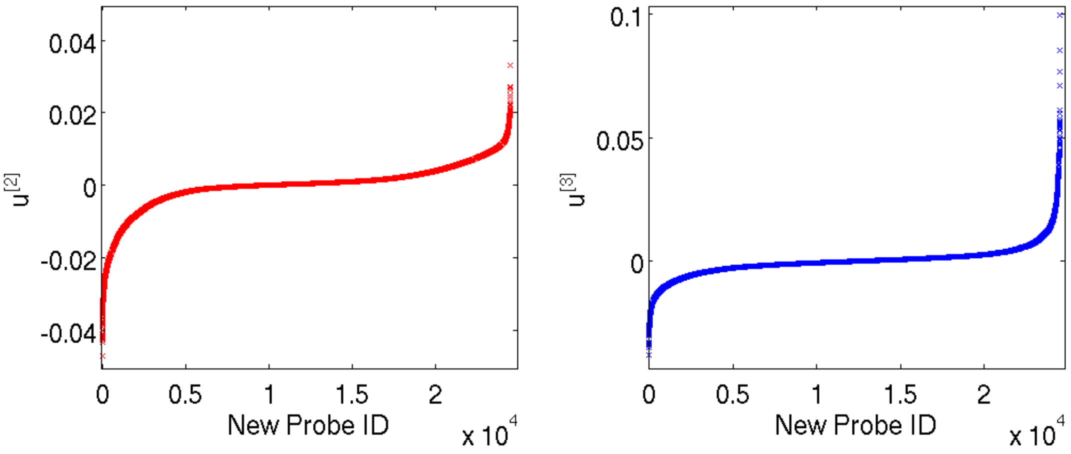

Figure 8.2: Components of the left singular vector $u^{[2]}$ and $u^{[3]}$ in increasing order from an SVD of the data.

For the following sections we will define a cluster as being the nodes that are pushed to the end of the ordering. In the analysis in Chapter 7, we used the first and last $1,000$ nodes (for ease of comparison) for each ordering. In this chapter, since the data set is significantly smaller, to maintain ease of comparison we will select proportionally a smaller amount, the first and last 600 nodes of the orderings.

8.2.1 Comparison with Gene List Database

As with Chapter 7, the results from the ends of the singular vectors with a gene list database are compared. This will give further assessment of the viability of whole blood as a tissue to investigate schizophrenia. The SZGR database is used as outlined in Section 7.4.2 - we can then directly compare results across the two antipsychotic-free cohorts. The reduced number of probes in this data set means that we will compare p-values rather than the number of hits. The number of probes that appear in the SZGR database will be affected (the ratio could be more or less), but the p-value approach controls for this by sampling randomly only from the list of genes present in the data.

Table 8.2 shows the first results from the ends of the cluster. This time, the ‘Start’, ‘Mid’ and ‘End’ portions of the singular vector are each 600 nodes in size.

We see significant results for expression and literature sets in one end of $u^{[2]}$. In both orderings the mid-points have, relatively, a very low number of hits for all three lists. This is indicative of the fact that genes in the middle of the ordering are likely to be of very low variance - if the expression levels are stable across all samples then it is unlikely that many approaches to microarray analysis would identify those as significant - that is, low variance genes are less likely to appear in the database lists.

Association	$u^{[2]}$ Start	$u^{[2]}$ Mid	$u^{[2]}$ End	$u^{[3]}$ Start	$u^{[3]}$ Mid	$u^{[3]}$ End
Sum	11	9	10	7	4	8
Sum with repeats	11	9	13	8	4	10
p-value	0.711	0.889	0.487	0.932	0.999	0.791
Expression	$u^{[2]}$ Start	$u^{[2]}$ Mid	$u^{[2]}$ End	$u^{[3]}$ Start	$u^{[3]}$ Mid	$u^{[3]}$ End
Sum	27	12	21	16	14	17
Sum with repeats	37	14	26	26	18	19
p-value	**0.020**	0.995	0.608	0.631	0.982	0.961
Literature	$u^{[2]}$ Start	$u^{[2]}$ Mid	$u^{[2]}$ End	$u^{[3]}$ Start	$u^{[3]}$ Mid	$u^{[3]}$ End
Sum	22	15	26	28	21	24
Sum with repeats	89	49	53	69	50	73
p-value	$<$ **0.001**	0.995	0.968	0.423	0.992	0.201

Table 8.2: Comparison of selected genes from left singular vectors $u^{[2]}$ and $u^{[3]}$ (Figure7.2) of SVD of the complete data set.

8.2.2 Hypothesis Testing: T-test and Fold Change

Next we present results from a t-test and thresholding of fold change of the same ‘First’ and ‘Last’ sections of the ordering vectors as used in the previous section. In Tables 8.3 and 8.4 we see a small number of genes meet the dual significance condition of $p = 0.05$ and fold change > 1.5. Hypothesis tests are again carried out with the t-test. The small number of significant genes is in line with expectations from the antipsychotic-free and control SVD results in Tables 7.9 and 7.10.

Official Gene Symbol	Illumina ID	p-Value	Fold Change	Percentile Abundance
Cluster One				
ACSL1	ILMN1684585	1.77e-03	1.58	90
LAMP2	ILMN2279961	2.40e-03	1.51	90
SUZ12	ILMN1797813	3.61e-03	1.57	91
Cluster Two				
HBE1	ILMN1651358	1.10e-03	-1.76	93
GATSL3	ILMN2098418	1.57e-03	-1.55	92

Table 8.3: $u^{[2]}$ ordering from SVD results: t-test on first and last 600 nodes of $u^{[2]}$ ordering (clusters one and two, respectively). This table shows genes with a significant p-value and fold change.

Official Gene Symbol	Illumina ID	p-Value	Fold Change	Percentile Abundance
Cluster One				
MX1	ILMN1662358	9.08e-03	-1.50	98
LYZ	ILMN2162972	5.90e-03	-1.56	81
Cluster Two				
HBE1	ILMN1651358	1.10e-02	-1.76	93
HBG2	ILMN2084825	1.11e-03	-1.87	99
HBG1	ILMN1796678	1.18e-03	-1.89	99
CDC14B	ILMN1733559	1.19e-03	-2.47	93
GATSL3	ILMN2098418	1.57e-03	-1.55	92
RNF213	ILMN1749722	3.79e-03	-1.51	98

Table 8.4: $u^{[3]}$ ordering from SVD results: t-test on first and last 600 nodes of $u^{[3]}$ ordering (clusters one and two, respectively). This table shows genes with a significant p-value and fold change.

8.3 Clustering with the GSVD

The data is now split into a pair of networks, the control and antipsychotic-free groups, $A \in \mathbb{R}^{24526\times22}$ and $B \in \mathbb{R}^{24526\times15}$ respectively. Taking the GSVD, the components of two ordering vectors $x^{[2]}$ and $x^{[end]}$ are shown in ascending order in Figure 8.3.

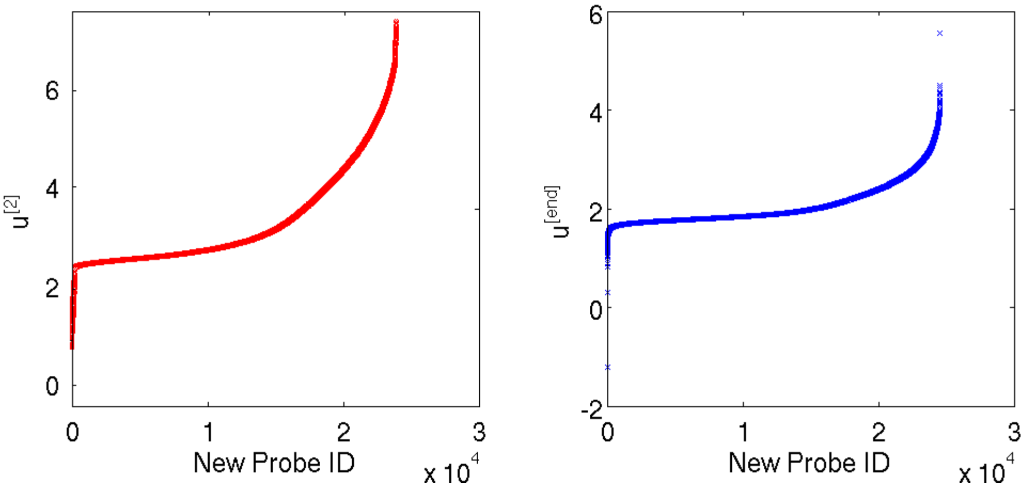

Figure 8.3: Components of $x^{[2]}$ and $x^{[end]}$ in ascending order from GSVD of antipsychotic-free and control samples

8.3.1 Comparison with Gene List Database

Table 8.5 shows occurrences and p-values for nodes in the first, middle and last 600 positions in $x^{[2]}$ and $x^{[end]}$ positions from the GSVD. The ordering from the GSVD shows significant p-values at both ends for both expression and literature groups. In particular, for the list of Expression genes the GSVD ordering has a total of 41 (cluster one) + 36 (cluster two) = 77 matches in the ends of the $x^{[2]}$ ordering, versus $37 + 26 = 63$ in the $u^{[2]}$ ordering from the SVD in Table 8.2. Similarly, the number of nodes in the ends of the $x^{[end]}$ ordering increase from $26 + 19 = 45$ in the $u^{[2]}$ SVD ordering to $24 + 40 = 64$. Equivalent numbers are present in the literature based comparisons. This GSVD is then once again highlighting more genes that have been implicated in schizophrenia - we can compare the results from the hypothesis test of GSVD orderings in the next section to the SVD results and observe any changes in the predictive abilities of the approaches.

Association	$x^{[2]}$ Start	$x^{[2]}$ Mid	$x^{[2]}$ End	$x^{[end]}$ Start	$x^{[end]}$ Mid	$x^{[end]}$ End
Sum	11	6	11	10	9	7
Sum with repeats	13	10	14	13	9	12
p-value	0.484	0.833	0.404	0.508	0.673	0.595
Expression	$x^{[2]}$ Start	$x^{[2]}$ Mid	$x^{[2]}$ End	$x^{[end]}$ Start	$x^{[end]}$ Mid	$x^{[end]}$ End
Sum	30	26	27	19	14	27
Sum with repeats	41	34	36	24	16	40
p-value	**0.003**	0.118	**0.046**	0.742	0.978	**0.014**
Literature	$x^{[2]}$ Start	$x^{[2]}$ Mid	$x^{[2]}$ End	$x^{[end]}$ Start	$x^{[end]}$ Mid	$x^{[end]}$ End
Sum	32	23	42	25	21	29
Sum with repeats	93	38	103	67	65	93
p-value	< **0.001**	0.997	< **0.001**	0.522	0.621	< **0.001**

Table 8.5: Comparison of selected genes from left singular vectors $u^{[2]}$ and $u^{[end]}$ of GSVD. Significant results for expression and literature.

8.3.2 Hypothesis Testing: T-test and Fold Change

Table 8.6 shows genes that meet the dual threshold criteria for significance we have previously described. *CDC14B*, *MX1* and *RNF213* appear in both the SVD and these results from the GSVD, but there are a number of differences.

Official Gene Symbol	Illumina ID	p-Value	Fold Change	Percentile Abundance
Cluster One				
MMP25	ILMN1717207	3.68e-04	1.59	98
MYO1F	ILMN1681239	2.86e-03	1.59	96
IFI6	ILMN1687384	4.21e-03	-1.63	97
MX1	ILMN1662358	9.08e-03	-1.72	98
CDC14B	ILMN1733559	1.15e-02	-2.47	93
BCL6	ILMN1737314	1.42e-02	1.52	98
LY6E	ILMN1695404	2.03e-02	-1.54	95
RNF213	ILMN1749722	3.79e-02	-1.64	98
MXD1	ILMN2214678	4.05e-02	1.53	97

Table 8.6: $x^{[2]}$ ordering from GSVD of antipsychotic-free not control expression data results: t-test on first and last 600 nodes of $x^{[2]}$ ordering (clusters one and two, respectively). This table shows genes with a significant p-value and fold change.

Official Gene Symbol	Illumina ID	p-Value	Fold Change	Percentile Abundance
Cluster One				
HBG2	ILMN2084825	1.11e-03	-1.87	99
HBG1	ILMN1796678	1.18e-03	-1.89	99
Cluster Two				
MX1	ILMN1662358	9.08e-03	-1.72	98
RNF213	ILMN1749722	3.79e-02	-1.64	98

Table 8.7: $x^{[end]}$ ordering from GSVD of control not antipsychotic-free expression data results: t-test on first and last 600 nodes of $x^{[end]}$ ordering (clusters one and two, respectively). This table shows genes with a significant p-value and fold change.

8.4 Comparison with Chaper 7 Results

A key reason for the analysis of a second data set was to generate comparison for the previous results, since such results are lacking in the literature. In the SVD of antipsychotic-free and control subjects in Chapter 7 (Table 7.9), *CDC14B* appears in both $u^{[2]}$ and $u^{[3]}$ orderings - each with a high negative fold change, as is the case in this data. This gene also appears in schizophrenia + antipsychotic-free SVD and GSVD comparisons. Similarly, *RNF213* appears in multiple orderings from both medicated and unmediated individuals with schizophrenia in both experiments with the same probe each time.

LY6E appears in clusters with similar levels of fold change in antipsychotic-free + schizophrenia (Table 7.14 at −1.51) and in the GSVD of schizophrenia and antipsychotic free data (Table 7.25 at −1.51). In this experiment we see *LY6E* at a fold change of −1.54. *RNF213* appears in Table 7.7, Table 7.8, Table 7.13 and in schizophrenia vs antipsychotic-free from the GSVD. In both experiments *RNF213* is downregulated by a similar amount (≈ -1.6). A summary of the genes that appear in both experiments is shown in Table 8.8

Gene ID	**Chapter 7**		**Chapter 8**	
	Fold Change	*p*-value	Fold Change	*p*-value
CDC14B	**-1.61**	**6.14e-03**	**-2.47**	**1.15e-02**
RNF213	-1.56	3.78e-04	-1.51	3.79e-03
LY6E	**-1.51**	**4.02e-02**	**-1.54**	**2.03e-02**
GATSL3	-1.77	2.13e-03	-1.55	1.57e-03
BCL6	1.52	1.06e-03	1.52	1.42e-02

Table 8.8: Fold change and *p*-value of genes that appear in results from both Chapter 7 and 8. Genes in bold appear only in results from antipsychotic-free patients.

8.5 Gene Ontology (GO): Annotation of Gene Lists

We now take a different approach to the analysis. The previous approaches of comparing clusters to the SZGR database, and hypothesis testing of individual genes has produced some interesting results but has reduced interpretation of large numbers of data to discussion of individual genes. A more expansive approach to analysis of gene lists, Gene Ontology (as introduced in Section 2.1.2 terms allow for identification of over-represented biological processes or functions. In this section we use the web tool GOrilla [51, 52] to identify the GO terms assigned to each of the genes in the clusters identified in the SVD and GSVD. A separate service REVIGO [200] is then used to reduce these typically large lists of ontology terms by identifying overlap and removing redundancies.

The results from this stage are much more general than the previous hypothesis tests. This approach has both advantages and disadvantages - including all nodes within a cluster ensures that each gene is represented in some way, regardless whether the gene meets significant thresholding of t-test and *p*-value. This is important in recognising the interactivity (and so complexity) between the cluster nodes that is missed when carrying out a single element test such as the t-test.

It is possible that gene ontology analysis may not be informative in this data set - ontology informs of large scale enrichment of particular biological processes or functions. The genes implicated in schizophrenia may not be prevelant or plentiful enough to result in enrichment of GO terms.

Gene Ontology enRIchment anaLysis and visuaLizAtion: GOrilla

GOrilla is an enrichment tool that accepts ranked lists of genes and generates a directed acyclic graph illustrating steps, if any, of three gene ontology elements - cellular component, molecular function and biological process.

Enrichment is the identification of GO terms that are in some way significantly over-represented in a list of genes. In our case, we input two gene lists - the genes from the ends of the re-ordering and the background list of genes of all those used in the experiment. The result comes with a *p*-value that tells us the probability of obtaining the observed hypergeometric probability score.

reduce + visualise Gene Ontology: REVIGO

In order to simplify this list to assist with the presentation and interpretation of results, REVIGO was used. This tool applies a clustering approach in combination with similarity measures to discard uninformative, redundant ontology terms. In using this we substantially reduce the number of ontology terms generated from each list - and avoid much of the traditionally problematic overlap.

8.5.1 GO: Results from SVD

Figure 8.4 shows gene ontology enrichment analysis. We have chosen plots of log($p-$value) vs log(size) - where the p-values signify over-representation of specific ontology terms in a gene subset, relative to the complete set of genes in the data set. The size of the marker on the plot corresponds to the number of genes listed under that ontology term, in total. The x-axis shows the size of the ontology term - ontology terms with more entries will tend to be more general. Interestingly, the first 600 probes in the reordered list show far more significant results in both process and functional term enrichment. This signifies the fact that of the 600 probes entered from the end of the ordering there is little to no consistent functional purpose. This parallels earlier results in Table 8.2 where significant results only appear in the first 600 probes.

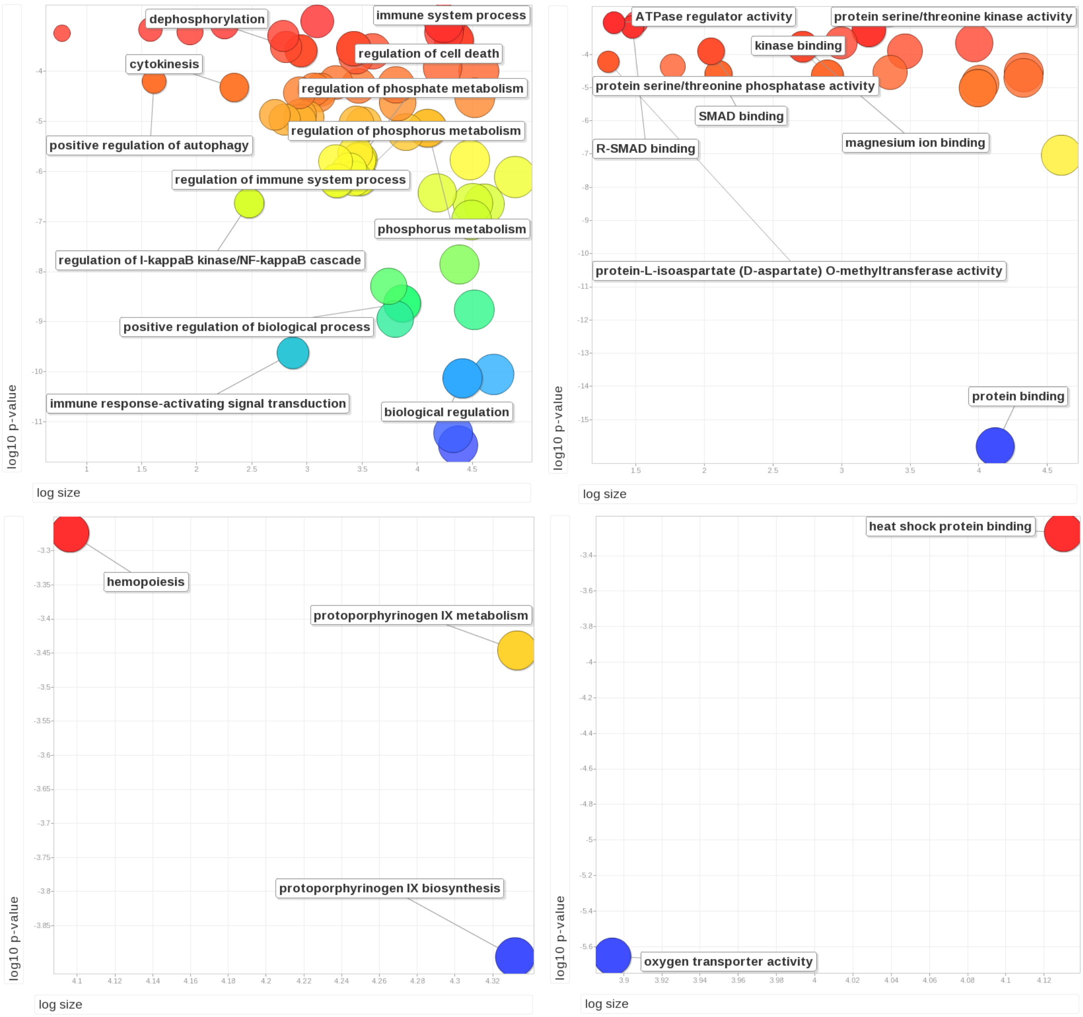

Figure 8.4: Gene ontology results from $u^{[2]}$ ordering. Top row: First 600 probes, left picture is GO:process, right is GO:function. Bottom row: End 600 probes, left is GO:process and right is GO:function. Each plot is log(p-value) vs log(size).

In Figure 8.5 we see both process and function ontology terms differ significantly from the $u^{[2]}$ ordering This is expected from, and mirrored by, the lack of overlap in the hypothesis testing. Both orderings in the SVD do, however, show enrichment of multiple immune system related processes:- regulation of immune system process and immune system process terms are common to both, with other immune related terms appearing exclusively in each. This highlights an advantage of using REVIGO - likely some of the related ontology terms appear because of common genes, but REVIGO ensures that the output is not dominated by features like this. Essentially, after processing with REVIGO we can be confident that related terms differ significantly.

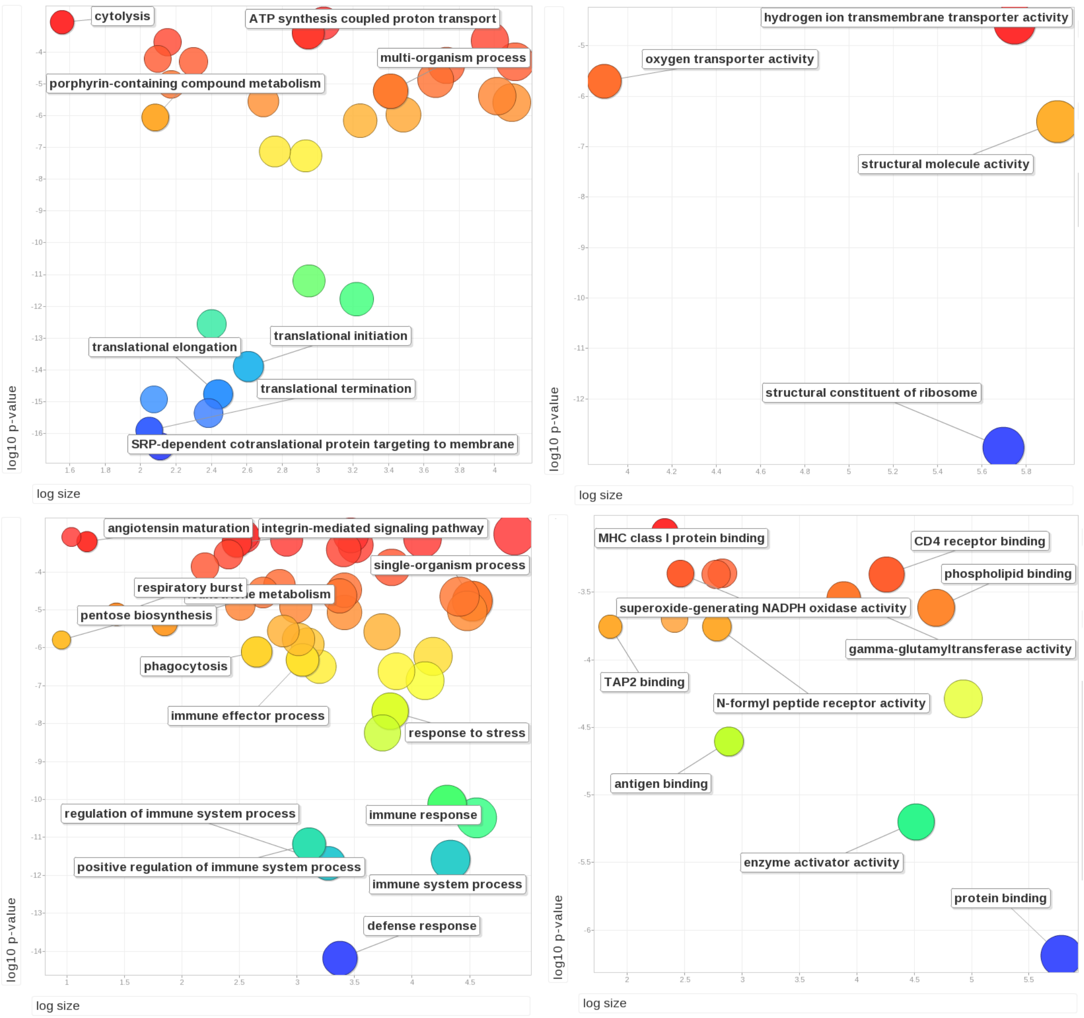

Figure 8.5: Gene ontology results from $u^{[3]}$ ordering. Top row: First 600 probes, left picture is GO:process, right is GO:function. Bottom row: End 600 probes, left is GO:process and right is GO:function. Each plot is log(p-value) vs log(size).

8.5.2 GO: Results from GSVD

The fact that Gene Ontology provides an approach to make summary assessments of results is particularly interesting here where we are also comparing different algorithms in the SVD and GSVD. Whilst we have repeatedly shown overlap in specific hypothesis testing results (as well as differences in results), genes that meet somewhat arbitrary hypothesis testing thresholds are a small subset of the genes driven to the ends of the orderings. Using GO to describe the complete set of genes from the ends of these orderings will then give an indication as to how expansive the difference between the two sets of results are. Figures 8.6 and 8.7 show process and function results from ordering vectors $x^{[2]}$ and $x^{[end]}$ in turn. The bottom row in Figure 8.6 has features common with the $u^{[3]}$ ordering in the previous section - namely multiple immune system processes, and MHC protein binding. We can also see that the last 600 probes in this ordering have a high degree of diversity - there are relatively a much higher number of ontology terms. In Figure 8.7 we see that, again, immune system processes are prevelant, and MHC protein binding in terms of function. The fact that there are many genes involved in immune system processes is interesting, as immune molecules like the MHC and immune system regulators/pro-inflammatory cytokines are suggested to have involvement in neurological processes [163, 44]. Additionally, recent GWAS work has shown association of MHC variants with schizophrenia [173]. It is also worth mentioning that we are dealing with data from whole blood, which contains mostly white blood cells - immune cells. Exploring all of the specific results conferred from GO analysis is outwith the scope of this project, but this simple test is an effective illustration of how ontology terms might be used to guide anaylsis, and provides an overview highlighting differences between the SVD and GSVD.

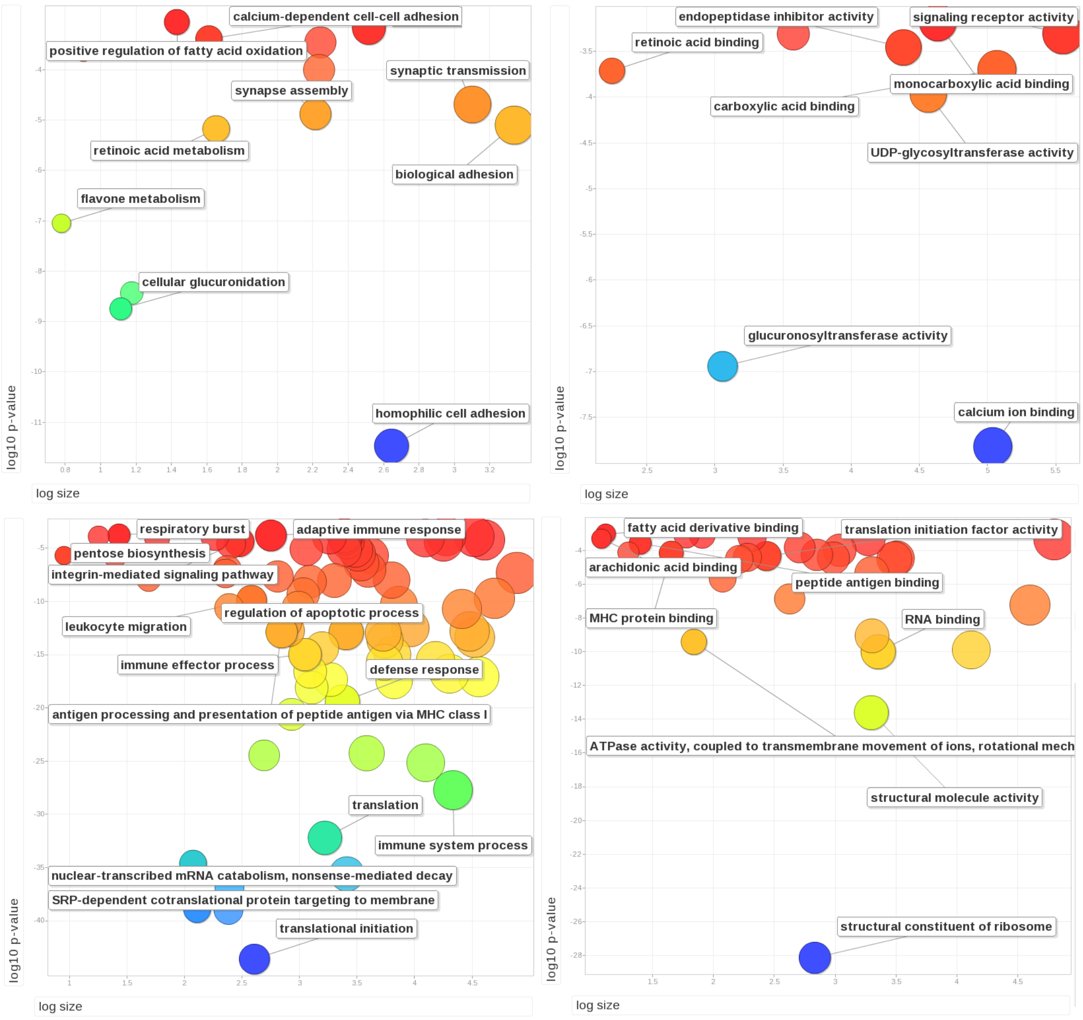

Figure 8.6: Gene ontology results from $x^{[2]}$ ordering. Top row: First 600 probes, left picture is GO:process, right is GO:function. Bottom row: End 600 probes, left is GO:process and right is GO:function.

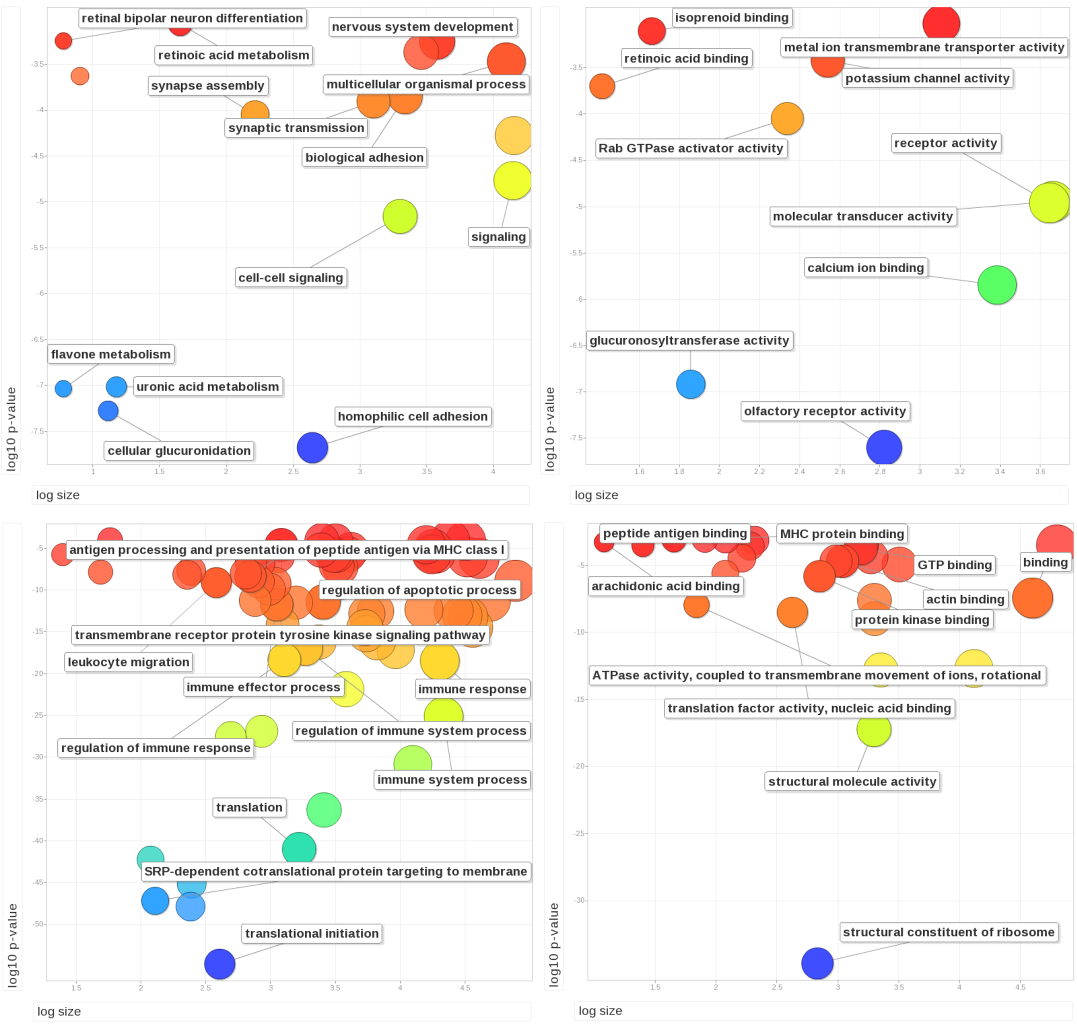

Figure 8.7: Gene ontology results from $x^{[end]}$ ordering. Top row: First 600 probes, left picture is GO:process, right is GO:function. Bottom row: End 600 probes, left is GO:process and right is GO:function.

8.6 Node-Weighted Laplacian

In Chapter 7 we saw a test of the node-weighted Laplacian on whole blood gene expression data. Results were limited, and we concluded that it was likely the type of structures the node-weighed Laplacian is able to identify were not present in the data. In this section we employ the same method from the initial data $M \in \mathbb{R}^{24526\times37}$ producing a pair of metabolic networks $B_K \in \mathbb{R}^{1945\times37}$ and $B_M \in \mathbb{R}^{572\times37}$. We then take MM^T for these matrices in turn which, in conjunction with the degrees of the metabolic networks, produces two node-weighted Laplacian matrices. Since the inital data is smaller than the set used in the previous chapter, the resultant metabolic networks B_K and B_M are also smaller. This time, we aim to both validate the previous finding for a different data set of the same type and investigate further the influence of the degree weighting in the node-weighted Laplacian. In essence, we wish to find out if the same genes appear as were present in the previous analysis. Finally, since (due to previous results) we are now less interested in the exploratory possibilities in node-weighted data of this type the analysis is kept to the equivalent Fiedler vector only.

8.6.1 KEGG Metabolic Network

An ordering vector from the node-weighted Laplacian in Figure 8.8 shows that again there are a small number of separated components. Figure 8.9 uses this vector to reorder the degrees of the KEGG network - showing that the ordering has a trend to follow the degree.

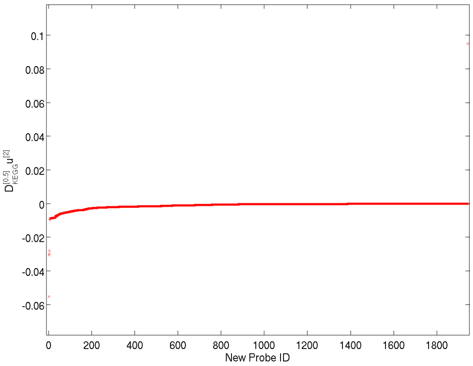

Figure 8.8: Components of $D^{[0.5]}u^{[2]}$ from SVD of complete data set.

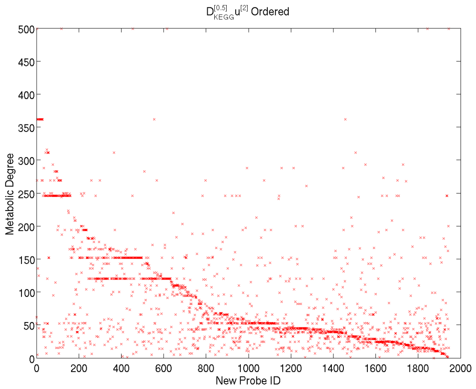

Figure 8.9: Metabolic degree reordered by $u^{[2]}$ from the KEGG node-weighted Laplacian.

Hypothesis testing of 100 nodes from each end of the vector (based on the distribution of the ordered components of the singular vector) yields two significant results that also satisfy the minimum fold change requirement. Table 8.9 shows these two genes, which do not match those found in the previous chapter.

Official Gene Symbol	Illumina ID	p-value	Fold Change	Percentile Abundance
Cluster One				
NT5E	ILMN1775480	3.68e-03	-1.30	62
Cluster Two				
CSAD	ILMN1791576	4.79e-02	1.31	42

Table 8.9: Results from t-test and fold change threshold of $D^{[0.5]}u^{[2]}$ from SVD of node-weighted Laplacian. Clusters one and two correspond to significant (in t-test and fold change) genes from the first 100 and last 100 nodes in the SVD ordering, respectively.

8.6.2 MetaCyc Metabolic Network

We also calculate the node-weighted Laplacian using the MetaCyc metabolic network. After reducing the data to contain only nodes present in the MetaCyc network, we have an array: $M_{MetaCyc} \in \mathbb{R}^{572\times37}$. The ordered components of the singular vector in Figure 8.10 show similarly limited deviation, as compared to the trial with the KEGG network in the preceeding section.

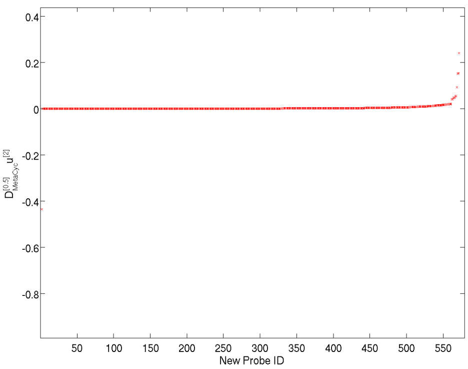

Figure 8.10: Components of $D^{[0.5]}u^{[2]}$ from SVD of complete data set

The reordered metabolic degrees for the MetaCyc network in Figure 7.22 almost exactly ordered for ascending/descending degree. This suggested that the microarray data was of little importance in any structure present in the data. This time the metabolic degree is less of an indicator as to the order of the data.

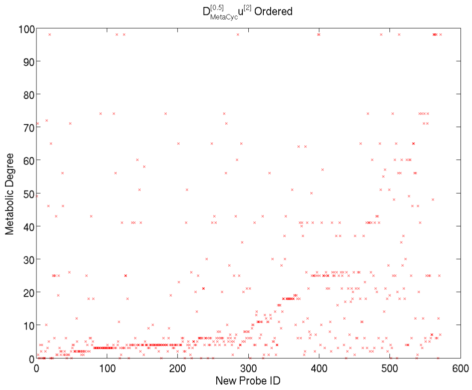

Figure 8.11: Metabolic degree reordered by $u^{[2]}$ from the MetaCyc node-weighted Laplacian.

Results from the t-test in Table 8.10 show three significant genes. This is in contrast to the previous chapter where no significant results were found. This difference can also be observed in the previous reordering of metabolic degrees - in contrast to the previous chapter we see that structure from the microarray data likely contributes to the ordering, the source of that structure may be in part explained by the genes in the following table. *NUDT1* has previously been found to show increase at the transcript level in the schizophrenia brain [178]. *SMOX* is a gene that has been previously implicated in this thesis, e.g., in Table 7.7 and discussed in the literature [63].

Official Gene Symbol	Illumina ID	p-value	Fold Change	Percentile Abundance
Cluster One				
NUDT1	ILMN1775480	2.62e-02	1.33	82
SMOX	ILMN2367258	4.48e-02	-1.37	90
Cluster Two				
CHSY1	ILMN1791576	2.34e-02	1.34	92

Table 8.10: Results from t-test and fold change threshold of $D^{[0.5]}u^{[2]}$ from SVD of the node-weighted Laplacian.

8.7 Discussion

In this chapter differences between antipsychotic-free and control microarray gene expression measurements are reported. The results from the SVD and GSVD have much in common, but there are some key differences. This is visible in part with the results of hypothesis (t-test) testing, but is particularly noticable with vast differences in ontology terms. There are also a number of consistencies between this chapter and results from Chapter 7, from both the SVD and the GSVD. This gives further weight to the conclusion that gene expression in whole blood is a viable source in which to study schizophrenia. It is also interesting that all matching results were present in Chapter 7 in SVD or GSVD only when antipsychotic-free data was included. In partciular, *CDC14B*, *LY6E* are not highlighted in any comparisons of samples from patients receiving antipsychotic treatment, suggesting that they highlight some feature of the data that is unique to antipsychotic-free samples. These genes then may better reflect the true disease state rather than effects as a result of treatment.

Gene Ontology

An analysis using an approach based on Gene Ontology was also conducted. Through this we have seen that, as with our other approaches, the SVD and GSVD have results in common and that there are also key differences. The ontology results seem to hold value in summarizing the results of these clustering approaches. Spectral clustering approaches yield large amounts of data for analysis afterwards. One solution to dealing with this is to use statistical methods, such as with the t-test we have used here. However, information on individual genes may be sparse - and it is difficult to observe higher level features that clustering methods aim to unveil. In a sense, there is a parallel to the difference between reductionist and systems level approaches - ontology approaches may identify major points of interest in the system, such as with the appearance of many immune related terms with our analysis.

In addition to the successes, there are issues with gene ontologies that are worth mentioning. First, gene ontology results are limited to the information present in the database. This means investigation with GO tools is limited to that which is available from current understanding, thus eliminating an angle of potentially novel result. Secondly, many genes have a multitude of functions in GO classifications - often the highest occurring function will be presented first in a GO output, this does not necessarily mean it is the most appropriate one.

Third, related to the other two points is the fact that no GO database is complete. It is possible to miss important functional detail about a group of genes if either the function is unknown or if it simply is not present in the database one consults. A similar point of note is that GO databases are, by their nature, biased towards larger areas of research. Fields where genetic function is better understood will produce more results in GO databases, and this can be a problem when analysing comprehensive genome wide gene expression data.

In summary, this chapter provides further evidence to the conclusion from Chapter 7 that antipsychotic-free patients exhibit significant differences in gene expression as compared to medicated schizophrenia patients. This highlights the need for further research into the role the variety of prescribed antipsychotic medications play in forming conclusions in studies designed to elucidate on the nature of schizophrenia. In addition, the GO approach shows that on a high level there is significant value added in using both the SVD and the GSVD to probe microarray data. A final comment is that genes identified across Chapters 7 and 8, *CDC14B*, *LY6E*, *RNF213*, *GATSL3* and *BCL6* have not previously been implicated in schizophrenia. This suggests that gene expression changes in blood may be specific to blood, rather than brain - we can investigate this further in the following chapter.

Chapter 9

Spectral Clustering of Brain Gene Expression Data

The previous chapters have shown analysis of microarray data from whole blood of patients with schizophrenia. This was aimed at developing the study of schizohrenia in a source of data that is easily accessible, and convenient. However, schizophrenia is a disorder of the brain, and so the majority of research is carried out on brain tissue. In this chapter we will apply the same methods to a microarray data set from human brain tissue. We expect this to be reflected in higher significance levels (than Chapters 7 and 8 in comparisons with results from the literature.

9.1 Introducing the Data

The data used in this chapter is taken from Gene Expression Omnibus study GEO21138 [161]. There are $N = 30$ subjects with schizophrenia, and $N = 29$ controls, which were age and sex matched in this study. Summary information of the data is presented in Table 9.1, showing the overall effect of age and sex matching. The schizophrenia cohort in this study were chosen to provide a balance

	Samples	Mean Age	Num. males	Num. females
Schizophrenia	30	43.40	24	6
Control	29	44.72	24	5

Table 9.1: Summary information on the data set from [161]

of stages of illness - short (< 5 years), intermediate ($7 - 18$ years) and long (> 18 years), since there is some discussion as to the potentially progressive nature of schizophrenia [144, 216]. Though the stage of illness was not explored directly in this chapter, it is noted that these possible sub-groups of the disease state are each represented in the data. This is potentially important since measurements from antipsychotic-free patients (as appearing in previous chapters) tend to point-of-diagnosis and so will more likely represent short duration patients. We know these patients are represented within this data set, albeit with medication included, but we can use this information to guide comparisons from this chapter with medicated and antipsychotic-free schizophrenia patients.

As a first step, we plot the total expression across all probes for each of the samples, in Figure 9.1. There is no apparent structure in the sum of gene expression values, with a similar midpoint and spread for both sample groups - this is helpful as a basic check on the quality of the data in case experimental setup was altered at all between measurements for each sample group.

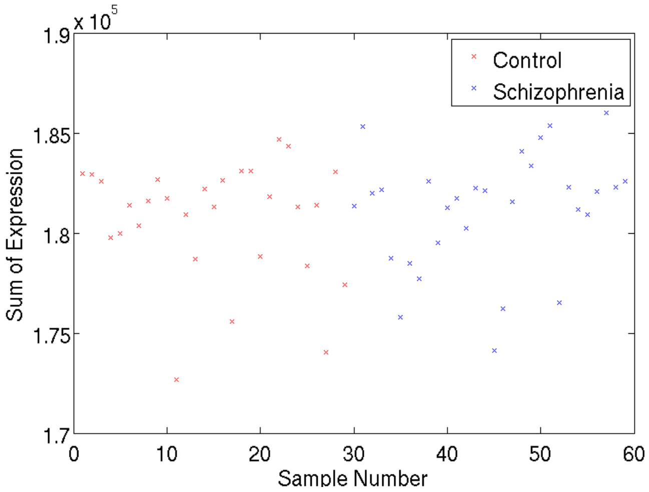

Figure 9.1: Sum of expression across all probes for each patient, showing a level of experimental consistency across patients.

9.2 SVD

Figure 9.2 shows two singular vectors (sorted) from the SVD of the complete data set. We can see separated groups of nodes at either end of the vectors, indicating structure within the data.

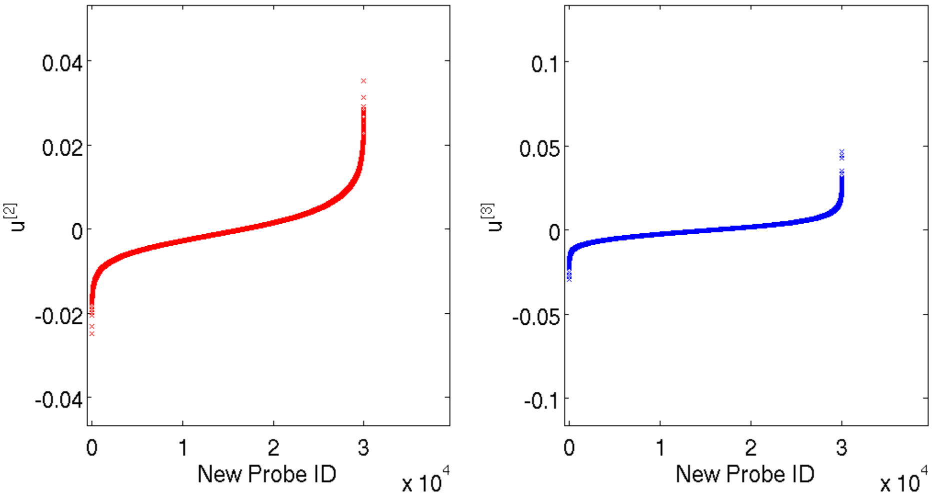

Figure 9.2: Components of the left singular vector $u^{[2]}$ and $u^{[3]}$ in increasing order from an SVD of the data.

9.2.1 Comparison with Gene List Database

As before we compare genes from the ends of the ordering generated by the SVD for matches with the SZGR database (outlined in Section 7.4.2). Since this data set is taken from brain tissue and since the vast majority of studies used to create the SZGR database are from brain tissue, we expect a high (or higher than in Chapters 7 and 8 with data from whole blood) degree of overlap between the two. The overlap is, again, expected to be strongest in expression studies, since this is a study of gene expression. The results from the comparison with SZGR are shown in Table 9.2 and show this to be a sensible postulate when the majority of p-values at either end of the singular vectors are highly significant. In fact, due to high levels of significance, the number of random comparisons used to generate the p-value is increased to $10,000$.

Association	$u^{[2]}$ Start	$u^{[2]}$ Mid	$u^{[2]}$ End	$u^{[3]}$ Start	$u^{[3]}$ Mid	$u^{[3]}$ End
Sum	18	5	29	19	5	18
Sum with repeats	21	6	40	22	3	23
p-value	0.3061	1.0000	$<$ **0.0001**	0.2219	1.0000	0.1605
Expression	$u^{[2]}$ Start	$u^{[2]}$ Mid	$u^{[2]}$ End	$u^{[3]}$ Start	$u^{[3]}$ Mid	$u^{[3]}$ End
Sum	40	15	32	46	19	45
Sum with repeats	98	26	48	85	41	111
p-value	$<$ **0.0001**	0.4981	$<$ **0.0001**	0.6640	0.8596	**0.0066**
Literature	$u^{[2]}$ Start	$u^{[2]}$ Mid	$u^{[2]}$ End	$u^{[3]}$ Start	$u^{[3]}$ Mid	$u^{[3]}$ End
Sum	49	21	52	59	19	59
Sum with repeats	110	35	136	208	32	190
p-value	**0.0092**	0.9878	$<$ **0.0001**	$<$ **0.0001**	1.0000	$<$ **0.0001**

Table 9.2: Comparison of selected genes from left singular vectors $u^{[2]}$ and $u^{[3]}$ from SVD of complete data set.

9.2.2 Hypothesis Testing

We will carry out the same t-test (and threshold based on fold change) on the first and last $1,000$ in the SVD orderings as performed in previous chapters. Table 9.3 shows results from the $u^{[2]}$ ordering, and Table 9.4 from the $u^{[3]}$ ordering.

Appearing in the first line of Table 9.3, there are suggestions in the literature that the allele size of potassium intermediate/small conductance calcium-activated channel gene *KCNN3* is linked to schizophrenia [49, 186] and so it is interesting that expression level is altered in this study. It has been shown that heat shock protein beta-1 (*HSPB1*) is overexpressed in schizophrenia patients [7], and overexpressed here with fold change 1.57.

Hypoxia-inducible factor 3-alpha (*HIF3A*) has shown increased expression levels in brain tissue samples from Brodmann area 46 (BA46) in patients treated with antipsychotics, but not in antipsychotic-free subjects, and in the same study it was found that Neuromedin (*NMU*) levels are decreased in schizophrenia [191] (1.75 and -1.54 respectively here). *NTSR2* is a neurotensin receptor subtype - mice deficient in neurotensin receptor 2 gene *NTSR2* have shown decreased basal glutamate levels [156], which is of interest given the glutamate hypothesis of schizophrenia [72]. One of the most interesting results in this table is the finding of decreased expression of parvalbumin (*PVALB*), validating findings from previous studies implicating *PVALB* in studies of the GARAergic origin of schizophrenia [82, 160]. *PVALB* is a marker of a subtype of GABAergic interneurones which have a profound inhibitory role on glutamatergic pyramidal cells [179].

Decreased expression of T-box, brain, 1 (*TBR1*) has been found in postmortem brain of schizophrenia patients, and is the gene which provides a regulatory protein for *NR2B-NMDA* receptors (as explored in Chapters 3 and 4) [117]. Also in the post mortem temportal conrtex of patients with schizophrenia, phosphodiesterase

Official Gene Symbol	Affymetrix ID	p-value	Fold Change
Cluster One			
KCNN3	244040 AT	2.41e-4	1.57
RGPD6	242712 X AT	4.95e-4	1.53
HSPB1	201841 S AT	5.44e-4	1.57
HIF3A	1555318 AT	7.92e-4	1.75
PITPNC1	239808 AT	2.13e-3	1.59
BCL6	228758 AT	3.12e-3	1.51
A2BP1	1566867 AT	4.06e-3	1.70
NTSR2	206899 AT	4.78e-3	1.52
HGF	209960 AT	5.18e-3	1.60
BAG3	217911 S AT	5.62e-3	1.71
AGAP1	240758 AT	5.93e-3	1.60
EFEMP1	201842 S AT	9.40e-3	1.63
CDKN1A	202284 S AT	9.55e-3	1.59
SLC7A2	225516 AT	1.25e-2	1.53
F3	204363 AT	1.66e-2	1.52
S100A8	202917 S AT	2.35e-2	1.70
AQP1	209047 AT	3.04e-2	1.58
SLC39A	1553126 A AT	3.12e-2	1.60
GJB6	231771 AT	3.81e-2	1.52
Cluster Two			
BTBD11	238692 AT	8.25e-4	-1.58
RXFP1	231804 AT	1.40e-3	-1.53
C5ORF13	201309 X AT	1.48e-3	-1.51
RWDD2B	222614 AT	1.50e-3	-1.63
SLC1A1	206396 AT	1.53e-3	-1.66
KCNC2	240614 AT	2.56e-3	-1.52
C1QTNF3	209426 S AT	2.59e-3	-1.66
NMU	206023 AT	2.67e-3	-1.54
PPM1E	205938 AT	2.69e-3	-1.53
PVALB	205336 AT	2.73e-3	-1.61
TBR1	220025 AT	3.32e-3	-1.52
PDE1A	242789 AT	4.61e-3	-1.67
SEC23A	204344 S AT	5.20e-3	-1.50
TRUB1	235447 AT	5.50e-3	-1.55
SYN2	210315 AT	5.58e-3	-1.54
MAP2K4	203265 S AT	5.78e-3	-1.54
PREPL	212216 AT	6.96e-3	-1.55
ZNF385B	1555800 AT	7.25e-3	-1.60
KCNIP4	224530 S AT	7.40e-3	-1.60
PAK1	209615 S AT	8.11e-3	-1.58
GLRB	205279 S AT	8.60e-3	-1.54
RGS4	204338 S AT	8.66e-3	-1.70
GABRG2	206849 AT	9.22e-3	-1.54
ETNK1	231576 AT	9.59e-3	-1.69
GABRA5	215531 S AT	1.21e-2	-1.57
PLCB1	215687 X AT	1.43e-2	-1.50
SCAMP1	206667 S AT	1.62e-2	-1.52

Table 9.3: $u^{[2]}$ ordering of sample expression data results: t-test on first and last $1,000$ nodes (named cluster one and two, respectively) of $u^{[2]}$ ordering showing genes with significant p-value and fold change.

1A, calmodulin-dependent gene *PDE1A* has shown decreased expression (-1.5, where we have -1.67 in this study) [11]. Synapsin II (*SYN2*) has shown in an association study to potentially confer susceptibility to schizophrenia [125] and so with the altered expression found here becomes an interesting candidate for future study. Regulator of G-protein signaling 4 gene *RGS4* is strongly decreased in expression in schizophrenia patients [154], and is one of the highest fold changes observed in this study at -1.70. Finally from this table, gamma-aminobutyric acid (GABA)-α receptor (*GABRA5*) is another gene that has been identified in association studies for schizophrenia [174] and shows significant down-regulation here.

Next, Table 9.4 shows results from the $u^{[3]}$ reordering. Cluster two in this table is broadly similar to cluster one from the $u^{[2]}$ ordering in Table 9.3. Cluster one in here, however, is small with only one probe surviving the t-test/fold change thresholding, suggesting that the probes separated to the left hand side of $u^{[3]}$ ordering in Figure 9.2 represent some feature of the data unrelated to the disease state of the patients e.g., age, gender. The gene that appears in this cluster *YBX2* (Y box binding protein 2), has been implicated as playing a role in male fertility [79], and so is likely just a chance occurence in passing threshold tests. Interestingly, *BCL6* appears in both $u^{[2]}$ and $u^{[3]}$ orderings - in a comparison of results from both blood chapters and results in this section so far, this is one of the genes highlighted in Table 8.8 showing matches for both blood data sets in Chapters 7 and 8.

That a number of genes identified in this study also appear (with expression levels altered in the same direction) in other studies within the literature is encouraging for the method, and encouraging for the quality and validity of the data set.

Official Gene Symbol	Affymetrix ID	p-value	Fold Change
Cluster One			
YBX2	219704 AT	4.37e-4	1.61
Cluster Two			
KCNN3	244040 AT	2.41e-4	1.57
HIF3A	1555318 AT	7.92e-4	1.75
NEAT1	227062 AT	1.66e-3	1.71
PITPNC1	239808 AT	2.13e-3	1.59
BCL6	228758 AT	3.12e-3	1.51
A2BP1	1566867 AT	4.06e-3	1.70
HGF	209960 AT	5.18e-3	1.60
BAG3	217911 S AT	5.62e-3	1.71
AGAP1	240758 AT	5.93e-3	1.60
PDK4	225207 AT	7.52e-3	1.53
EFEMP1	201842 S AT	9.40e-3	1.63
CDKN1A	202284 S AT	9.55e-3	1.59
MBP	236324 AT	1.10e-2	-1.53
SLC7A2	225516 AT	1.25e-2	1.53
ADM	202912 AT	1.65e-2	1.51
F3	204363 AT	1.66e-2	1.52
S100A8	202917 S AT	2.35e-2	1.70
AQP1	209047 AT	3.04e-2	1.58
SLC39A12	1553126A AT	3.12e-2	1.60
GJB6	231771 AT	3.81e-2	1.52
ATP8A1	231484 AT	4.74e-2	1.54

Table 9.4: $u^{[3]}$ ordering of sample expression data results: t-test on first and last $1,000$ nodes (named cluster one and two, respectively) of $u^{[3]}$ ordering showing genes with significant p-value and fold change.

9.3 GSVD: Schizophrenia and Control Data

The data is now split into a pair of networks, the schizophrenia cohort, $A \in \mathbb{R}^{30024\times 30}$, and the control group $B \in \mathbb{R}^{30024\times 29}$. Figure 9.3 shows the $x^{[2]}$ and $x^{[end]}$ reordered singular vectors.

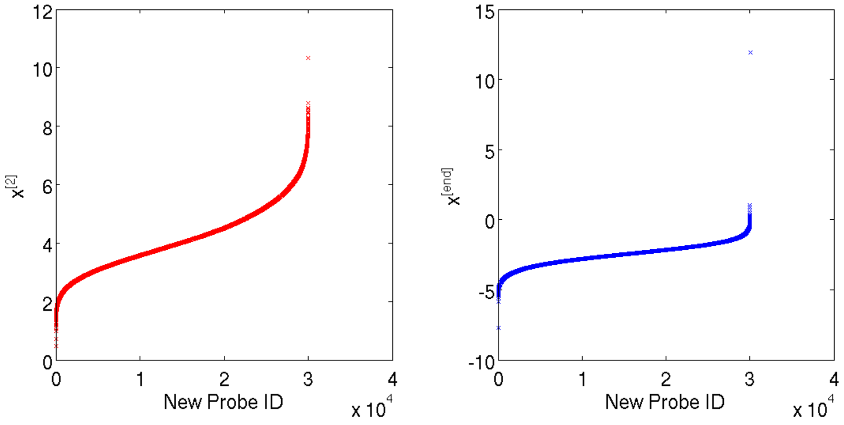

Figure 9.3: Components of the left singular vector $x^{[2]}$ and $x^{[end]}$ in increasing order from the GSVD of the data.

9.3.1 Comparison with Gene List Database

Next we compare the results from the GSVD reordering of the data with the SZGR gene list database. In both Chapters 7 and 8 we have seen that the GSVD clustered a greater total number of SZGR genes than the SVD - in this brain data we have found the opposite to be the case. Table 9.5 shows that in this case the GSVD in fact clustered significantly fewer SZGR genes. For example, in Expression studies Table 9.2 shows 98 (cluster one) + 48 (cluster two) = 146 genes in $u^{[2]}$ and 85 (cluster one) + 111 (cluster two) = 196 in $u^{[3]}$ where the GSVD has separated 55 (cluster one) + 87 (cluster two) = 142 in $x^{[2]}$ and 50 (cluster one) + 65 (cluster two) = 115 in $x^{[end]}$. This suggests that there is value in applying both SVD and GSVD, but this assessment cannot be made on the number of genes alone - further work into the biological interpretation would assist in diagnosing whether the SVD and GSVD are individually particularly good at highlighting specific biological features of the data.

Association	$x^{[2]}$ Start	$x^{[2]}$ Mid	$x^{[2]}$ End	$x^{[end]}$ Start	$x^{[end]}$ Mid	$x^{[end]}$ End
Sum	9	7	30	19	8	19
Sum with repeats	13	7	44	26	11	25
p-value	0.9157	0.9974	< **0.0001**	**0.0441**	0.9490	0.0590
Expression	$x^{[2]}$ Start	$x^{[2]}$ Mid	$x^{[2]}$ End	$x^{[end]}$ Start	$x^{[end]}$ Mid	$x^{[end]}$ End
Sum	32	15	37	31	15	33
Sum with repeats	55	27	87	50	24	65
p-value	0.1487	0.9993	< **0.0001**	0.3823	0.9996	**0.0070**
Literature	$x^{[2]}$ Start	$x^{[2]}$ Mid	$x^{[2]}$ End	$x^{[end]}$ Start	$x^{[end]}$ Mid	$x^{[end]}$ End
Sum	42	22	60	50	22	47
Sum with repeats	91	36	191	140	44	107
p-value	0.4240	1.0000	< **0.0001**	< **0.0001**	1.0000	**0.0225**

Table 9.5: Comparison of selected genes from left singular vectors $x^{[2]}$ and $x^{[end]}$ of GSVD of antipsychotic-free vs control samples. $x^{[2]}$ orders for controls not schizophrenia, $x^{[end]}$ for schizophrenia not controls.

9.3.2 Hypothesis Testing

The final section in this chapter is to apply a t-test and fold change threshold to the GSVD reordered data.

Table 9.6 shows results for the $x^{[2]}$ reordered data, and Table 9.7 for the $x^{[end]}$ reordered data. Copy number variation of *A2BP1* has been linked to significant increase in risk for development of schizophrenia [150]. Myelin Basic Protein (*MBP*) has shown sigifnicant alterations in regulation in both schizophrenia and bipolar disorder patients in a microarray study validated by Q-PCR testing [202]. Interestingly, we see gamma-aminobutyric acid α receptor, gamma 2 gene *GABRG2* here, another gene involved in the GABA system, an attractive area for the study of schizophrenia [95]. A number of key results from the SVD analysis of this data also appear within the GSVD results, including *PVALB*, *RGS4* and *HIF3A*.

Official Gene Symbol	Affymetrix ID	p-value	Fold Change
Cluster One			
HIF3A	1555318 AT	7.92e-4	1.75
A2BP1	1566867 AT	4.06e-3	1.70
MBP	236324 AT	1.10e-2	-1.53
ADM	202912 AT	1.65e-2	1.51
Cluster Two			
C5ORF13	201309 X AT	1.48e-3	-1.51
PREPL	212216 AT	6.96e-3	-1.55
MCTP1	220122 AT	7.01e-3	-1.51
FGF12	214589 AT	8.50e-3	-1.53
GLRB	205279 S AT	8.60e-3	-1.54
RGS4	204338 S AT	8.66e-3	-1.70
GABRG2	206849 AT	9.22e-3	-1.54
ATP6VLA	201971 S AT	1.05e-2	-1.53
F3	204363 AT	1.66e-2	1.52
SLC9A12	1553126 A AT	3.12e-2	1.60
GJB6	231771 AT	3.81e-2	1.52

Table 9.6: $x^{[2]}$ ordering from GSVD for control not schizophrenia sample expression data results: t-test on first and last $1,000$ nodes (named clusters one and two, respectively). This table shows genes with a significant p-value and fold change.

Official Gene Symbol	Affymetrix ID	p-value	Fold Change
Cluster One			
HSPB1	201841 S AT	5.44e-4	1.57
INVS	232759 AT	1.24e-3	-1.58
PVALB	205336 AT	2.73e-3	-1.61
MCTP1	220122 AT	7.01e-3	-1.51
ZNF385B	1555800 AT	7.25e-3	-1.60
GLRB	205279 S AT	8.60e-3	-1.70
RGS4	204338 S AT	8.66e-3	-1.54
GABRG2	206849 AT	9.22e-3	1.56
CITED2	207980 S AT	1.58e-2	-1.52
ATP8A1	231484 AT	4.74e-2	1.54
Cluster Two			
HGF	209960 AT	5.18e-3	1.60
SLC7A	225516 AT	1.25e-2	1.53
SLC39A12	1553126 A AT	3.12e-2	1.60

Table 9.7: $x^{[end]}$ ordering from GSVD for control not schizophrenia sample expression data results: t-test on first and last $1,000$ nodes (named clusters one and two, respectively). This table shows genes with a significant p-value and fold change.

9.4 Comparison of Chapters 7, 8 and 9

Section 8.4 contains a brief comparison between the two chapters with data from whole blood. In this comparison, Table 8.8 highlights a number of genes that meet significance thresholds in both experiments: *CDC14B*, *RNF213*, *LY6E*, *GATSL3* and *BCL6*. Of these genes only one, *BCL6*, appears in all three analyses (and is upregulated in all cases). In this section we provide a brief analysis of all probes present in this brain data for these genes. Figure 9.4 shows expression levels for each patient sample for all probes representing these gene IDs (*CDC14B, RNF213, LY6E, GATSL3* and *BCL6*), red 'x' are from schizophrenia patients and blue 'o' are from controls, as described in Table 9.8

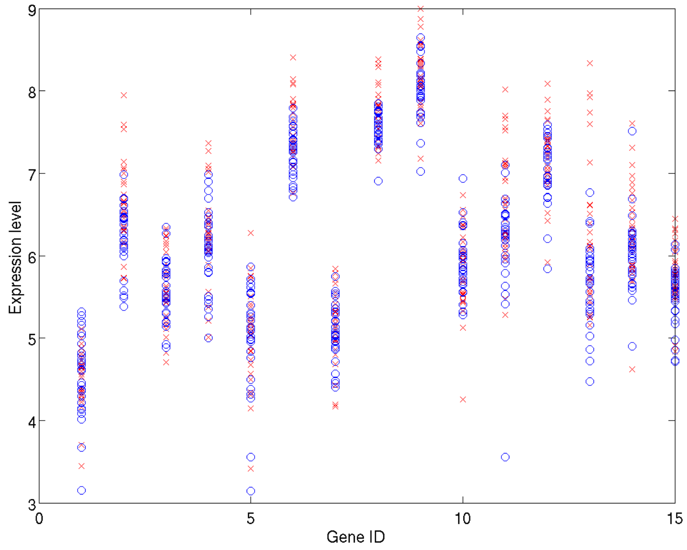

Figure 9.4: Sum expression levels for genes (as measured in this brain data) present in in all of Chapter 7, 8 and 9 analysis. Red ‘x’ is schizophrenia patient data, blue ‘o’ is control. See Table 9.8 for details.

Gene ID	Official Gene Symbol	Affymetrix ID	p-value	Fold Change
1	*RNF213*	1569078 AT	6.74e-1	-1.03
2	*RNF213*	225929 S AT	3.00e-3	1.29
3	*RFN213*	225931 S AT	7.12e-1	1.03
4	*RNF213*	230000 AT	1.55e-1	1.14
5	*RNF213*	241480 AT	3.34e-3	1.22
6	*CDC14B*	208022 S AT	1.91e-3	1.22
7	*CDC14B*	216284 AT	6.30e-1	1.03
8	*CDC14B*	221555 X AT	2.46e-2	1.13
9	*CDC14B*	221556 AT	4.03e-2	1.15
10	*CDC14B*	230887 AT	5.70e-1	-1.05
11	*CDC14B*	234605 AT	2.74e-3	1.45
12	*BCL6*	215990 S AT	1.38e-1	1.12
13	***BCL6***	**228758 AT**	**3.10e-3**	**1.51**
14	*BCL6*	236439 AT	9.08e-2	1.18
15	*GATSL3*	233528 S AT	3.40e-3	1.22

Table 9.8: Fold change and t-test results (as measured in brain data) exploration of genes from Chapter 7, 8 and 9 analysis.

In Table 9.8, we see that there are a variety of fold change values - we would expect genes that are not significant to either experimental group to show an average fold change of 1.00, with error. Whilst *BCL6* is the one gene that meets threshold testing, it is worth noting that an ID 11 gene *CDC14B* has a significant t-test and narrowly misses threshold on fold change. Similarly with ID 2, gene *RNF213* has a significant t-test and low fold change. Finally, the sole probe measuring *GATSL3* expression shows significance in t-test and a minor up-regulation of 1.22. These genes represent an important cross-over point between experiments - being altered in expression in both our results from blood studies and results in this brain data. There are also a number of different probes representing each gene in Table 9.8 - not all of the probes for each gene are over/under expressed in a similar way. This could be indicitive of the fact that each probe measures a specific part of the gene transcript, and alternative splicing points may be expressed differently.

9.5 Discussion

There are points of overlap in our results and the original publication that arose from the data e.g., implication of *PPM1E* as having a key role in the structure of this data [161], which is interesting for verification and validation methods across approaches to analysis.

This chapter again highlights the desirability for an increase in the amount of clinical data taken per patient, in this chapter we have data with length of illness but no details on medication - in Chapters 7 and 8 there were data with medication-free patients but no information about length of illness. Information about side-effects, drug type and drug efficacy would also allow for a formal investigation into possible patient specific symptom vs. management of side-effects, the concern of much of modern psychiatric treatment. A disparity in data depth (i.e., amount of patient information or phenotypic) means that, despite the fact that data gathering rates are at an all-time high, it has proven difficult to convincingly illustrate links between genes and disease states. A solution may be to adopt an approach based on depth rather than breadth, this is difficult in situations where postmortem tissue is used, due to degradation of tissue and other issues of sample quality - but in a tissue such as blood a depth-first approach may be possible, exploring schizophrenia in the same sample group from both genetics and expression viewpoints.

There are various possibilities in terms of future work for a data set of this type. In previous chapters we have discussed the importance of, and analysed, data from antipsychotic-free schizophrenia patients. Data from patients at different stages of the condition offer similar possibilities, and will notably allow for the mapping of potential progression of the disease. There are suggestions that schizophrenia is progressive in nature - with studies measuring brain grey matter demonstrating that grey matter volume decreases from early to chronic stages of the condition [144]. Abnormalities in white matter volume over time have also been observed [216]. These facts parallel our earlier findings in Chapter 4 where we illustrate that PCP, which is used to develop a model relevant to the disease, disrupts brain connectivity which makes this a strong point of interest for future work. Such a study would focus on precision measurements in equivalent brain structures between the model animal and human cases, rather than studying each in isolation.

This also highlights a difficulty that parallels the problem of potential interference of results by antipsychotic medication - progressive elements of schizophrenia may lead to less clarity in results since there is no clear pattern across all schizophrenia patients. It is reasonable to suggest that patients are kept to sample groups appropriate to the stage of their disease, as well as their usage and type of antipsychotic medication (and a number of other factors e.g., smoking). This is in the ideal case, which is difficult to achieve with the small numbers of available samples, particularly in the case of post-mortem brain tissue. The ability to study schizophrenia in a readily available tissue such as blood would help alleviate this issue and allow researchers to obtain data at multiple time points, mapping both the progression of the disease and any effects from long term usage of antipsychotic medication, factors which are both currently not well understood. There has also been evidence that schizophrenia is an umbrella-term encompassing a set of conditions - that is, schizophrenia is heterogeneous in presentation and aetiology [206, 145]. A comprehensive study into one aspect, in this case gene expression, may begin to help separate aspects of the disease.

It would also be interesting study the matches with SZGR (from Tables 9.2.1 and 9.3.1) in further detail. In particular, with more time ideally we would have checked for overlap between the SZGR matches for Chapters 7 and 8 with this chapter. Conclusions in previous chapters that comparisons with SZGR are sensible as an investigative approach for whole blood are validated by strongly the positive results found in this chapter.

It is noted that, though there are a significant number of matches between SZGR and clustered genes in the whole blood studies and an even greater significance in this chapter, there is only one overlapping gene in the last of those that pass hypothesis and fold change thresholding, *BCL6*. This may highlight limitations of assuming a fold-change/statistical testing approach in addition to simply using clustering and SZGR comparisons. There are advantages to using thresholds in addition to clustering, however. With fold change and significance testing, we are simply selecting a specific sub-group of genes that, since thresholds must be met, are more likely to be connected to the sample state.

Finally, we have shown for the first time that *BCL6* is a gene of interest that may overlap between blood and brain expression experiments. This is important and worthy of further investigation with, for example, verification through PCR experiments and functional investigations in modelling. In addition, Section 9.4 shows that there are suggestions of differential expression in other probes. Whilst results here are not conclusive, the suggestion is important enough to warrant further investigation. A concluding remark is that even given a combination of unsupervised and supervised processing approaches, the selection via a clustering approach and further selecting with fold change and t-test, there are still a large number of genes separately in whole blood and brain tissue that appear relevant. After multiple layers of processing (clustering, thresholding), the lack of overlap between these sets of results highlights a combination of the complexity of gene expression data, the complexity of the schizophrenia disease state and inherent differences between blood and brain.

Chapter 10

Conclusions

The overall aims of this thesis were to systematically analyse data pertinant to schizophrenia using spectral clustering methods in a variety of settings. We first investigated animal models relevant to the condition using the generalized singular value decomposition. This approach aims to find structures mutually exclusive to subject test groups, and was augmented with use of a novel technique to assess clustering of subgroups within the test animals. Next we created a pair of new metabolic networks for use with a new matrix, the node-weighted Laplacian, which allows the combination of two sources of data. This was used to explore potential metabolic involvement in schizophrenia across two studies of human whole blood. The final chapter involves application of spectral clustering methods of microarray gene expression data from human brain tissue. In this chapter we summarize the novel developments in the thesis and point to possible future directions.

10.1 Animal Models Relevant to Schizophrenia

The development of animal models useful for studying complex conditions such as schizophrenia is extremely challenging. There are many reasons for this: in humans, the symptomatology is diverse and difficult to profile hence challenging to cover in a single model. The underlying genetic, biological and environmental factors are not well categorised or understood - a combination of these two factors results in a current lack of biomarkers which could be replicated in an animal model [162]. Chapters 3 and 4 presented work on two pharmacological animal models relevant to schizophrenia. With the fact that there is an inability to obtain direct feedback from animals, it is difficult to verify the presence of many of the common symptoms of schizophrenia e.g., hallucinations. Thus, in this work we approach analysis of the ketamine and PCP animal models from the perspective of functional operation of neural subsystems. These measurements do not rely on profiling symptoms which can be unreliable, instead they provide an objective examination of the subsystem level activity in the brain. The analysis of data in the ketamine model found that acute ketamine treatment caused disintegration in the functional interaction between various regions of the brain, relative to control animals. Interestingly, our data suggest that the alterations induced through ketamine treatment are disparate to those seen in chronic schizophrenia. The observed fundamental differences between brain function in the model state versus the desired translational parameters in the human disease state have implications for the relevance of acute ketamine treatment in translational models. Potentially the acute ketamine model is more representative of the early stages of schizophrenia, rather than the chronic stage. Indeed there is evidence that ketamine produces a hyperglutamatergic state which resembles the early phase of schizophrenia but not the chronic state [142].

It is noted that the data set used for this investigation is relatively small, 9 animals in each sample group, which is not ideal when making confident statistical conclusions. That said, there is statistical significance in the results after verifying the quality of clusters (Section 3.2.1), and taking multiple approaches that produce biologically meaningful conclusions. In addition to contributions from this thesis in suggesting caution with the use of ketamine in preclinical models, these results add insight into the functional response of the brain under NMDA receptor blockade. Both of these results make contributions to important areas in neuroscience and warrant further investigation.

The subchronic PCP and modafinil preclinical data show how sustained NMDA receptor hypoactivity affects brain network structure. These results closely parallel those reported in schizophrenia, with direct relevance for understanding the nature of brain dysfunction in the chronic disease state. The quantitative nature of the study, at the systems level, using the GSVD for the first time to compare disease model (PCP), treatment (modafinil) and control animals, reinforce the relevance of NMDA receptor dysfunction models to schizophrenia. Quantifying compromised functional integration at the level of specific subsystems is a first step towards exploring targets for medication, in a similar way that the results of this thesis suggest that modafinil has potential in alleviating cognitive deficits.

10.2 Stratification of Patients and Biomarkers

There are a variety of poorly or un-met clinical needs in schizophrenia [108]. Animal models are one way we may hope to begin to address this issue, stratification of patients is potentially another. Stratification of patients into sub-groups has proven a useful strategy in conditions ranging from cancers to stroke, increasing efficiency in identification of effective treatments for cancers [83, 89], and understanding or mitigating future risks, with stroke [171]. Before a similar stage is reached for a condition such as schizophrenia, there are a multitude of areas where further research is required. Disease mechanisms could be better understood and an efficient, reliable and safe way to achieve results is required. In this thesis, Chapters 7 and 8 make steps towards the later in attempting to validate the use of whole blood as a tissue relevant to schizophrenia.

With the large number of genes studied in expression studies, it is perhaps unsurprising that there is a disparity across studies in the literature as to the genes relevant to the disease. Our study took a new approach by comparing results from spectral clustering with a database of genes implicated in schizophrenia. Finding significant overlap between clustering results from whole blood and a database of genes implicated previously in the condition has two important results (1) that whole blood is of legitimate interest in studying schizophrenia and (2) that results from whole blood replicate, at some level, alterations in the brain. Following the comparison between experiment and database gene lists, it would also be desirable to study the genes that match for any interesting biological results. In particular, genes present in more than one study sub-type (e.g., association + expression) warrant further investigation. For instance, examinations of genetics and gene expression are usually isolated with different patients in each. In line with our proposal to limit the data gathering in terms of the number of patients and instead increase the amount of data for each, a suggestion for future work would be to comprehensively gather genetics (GWAS), gene expression (microarray) and phenotypic data all for a smaller group of patients to allow for correlations across each source of information.

The focus in this thesis was on gene expression data. Other approaches to the study of schizophrenia include structural and functional MRI studies. For example, it has been long established that structural abnormalities are present in patients with schizophrenia, and that structural abnormalities vary over time, i.e., that the disease has a progressive element [170]. There are some questions surrounding this issue, including recent suggestions that progressive changes may be correlated with use of antipsychotics [66] or even substance abuse which is prominant among individuals with schizophrenia [143].

These findings are mirrored in conclusions made in this thesis that antipsychotic-free schizophrenia patients differ significantly in gene expression from their counterparts receiving treatment. This raises serious questions for consideration of the majority of researchers in the field where patients are almost always receiving antipsychotic medication. That this is not already routinely considered is perhaps an issue of pragmatism with the difficulty in gathering medication-free data, or a lack of appreciation for the impact medication may have on results. In either case, in order to improve data quality, future works would benefit from being explicit with details where confounding factors such as medication are present. Given that the vast majority of antipsychotic-free patients involved in schizophrenia research are simply individuals at the point of diagnosis, developing an understanding of the effect of specific medication may allow for the isolation and study of other factors such as disease progression. This is an area where the work in this thesis makes significant contributions - in showing that much of the alterations seen in gene expression are the result of medication we do a number of things: (1) highlight the importance of unambiguously stating medication type and length of treatment to qualify results (2) set the scene for the correlation of gene expression with specific drug treatments to understand therapeutic effect (3) open an avenue for a method of investigation for side effect presentation and severity - measurement of pre-treatment gene expression levels may help to indicate an individuals propensity for particular difficulties or complications. If attributes such as drug type and side effects are to be recorded and reliably studied, the demand for data will increase. Currently, the fact that a significant proportion of research is carried out using post-mortem human brain tissue is a clear barrier to increasing patient numbers, but an increase in phenotypic data may help ensure maximal value.

All of these points come together in relevance to the identification of a disease biomarker. To be useful for diagnostics, biomarkers must be unambiguous, accurate and straightforward to measure.

10.3 Spectral Clustering Methods

In this work we proposed a novel addition to a traditional spectral clustering approach, calculating the variance of spectral vector components to extend the clustering method to the real domain. This allows for increased accuracy of automated clustering assessments, of timely relevance to the recent explosion in the interest of 'big data'.

The proposed algorithm is tested on two real data sets in Chapters 3 and 4 and validated through visual assessment and calculation of $p-$values as a basic check to allow a level of automation. Having shown that this approach can be effective, it can be concluded that, in terms of contributions to spectral clustering, this approach offers additional flexibility as a means to include additional information to give specific perspective to a problem. There are many cases outside of neuroscience where such an algorithm could be useful. For example, clustering methods are utilised in cases where there are interesting potential divisions within the data - in cases where there are many a priori sub-groups of interest (e.g., tissue or cancer types in biology [123], income or age brackets in social networks, institution of author in citation networks) it can be impractical, particularly in large data sets, to evaluate results visually.

A new algorithm is proposed based on the node-weighted Laplacian, giving an extension of existing methods. The node-weighted Laplacian is validated on synthetic data, proving the principle and illustrating that the approach yeilds results satisfying a mesh of two data sources. The algorithm is then tested on real data at numerous points throughout the thesis - results on real microarray gene expression data in Chapters 7 and 8 were generally limited. There are likely a combination of issues here: (1) there may be no metabolic component to discover in schizophrenia data, and it is less likely there is a high-degree metabolic component such as the node-weighted Laplacian would identify, (2) there are clear limitations with the current sources of metabolic information. The KEGG and MetaCyc databases of human metabolism used to create metabolic networks in Chapter 5 differ significantly, highlighting that the complete human metabolome still requires significant effort.

With these difficulties in mind, it would have been of particular interest and perhaps more appropriate to explore wider applications for the node-weighted Laplacian, more specifically in the first instance a case where there is a clear expectation of result. In the field of complex networks there are several commonly used approaches to incorporating extra information, e.g., nodes can be labelled with some extra variable such as personal income in a social network, networks can have multiple edge sets; each of these areas can benefit from a construction like the node-weighted Laplacian which can be adjusted based on relevant questions about the data. Further effort at investigating integration of other network parameters such as clustering coefficient, betweenness, etc., would help to explore sensitivity and scope of the method.

The node-weighted Laplacian is also relevant in a timely sense with the rise of Big Data (see [138]). Huge amounts of data are consistently generated across many fields and exploratory analysis on large data sets with large numbers of variables can be difficult to interpret where many factors can explain variation or observed effects. In providing an approach to include extra information the scope of the analysis can be focussed with a structure such as the node-weighted Laplacian to answer specific questions.

Bibliography

[1] Affymetrix. Affymetrix microarray solutions, July 2013.

[2] B. Alberts, A. Johnson, J. Lewis, M. Raff, and K. Roberts. *Molecular biology of the cell 4th edition*. 2002.

[3] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour. Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics*, 7(1):55–65, 2006.

[4] N. C. Andreasen, K. Rezai, R. Alliger, V. W. Swayze, M. Flaum, P. Kirchner, G. Cohen, and D. S. O'Leary. Hypofrontality in neuroleptic-naive patients and in patients with chronic schizophrenia: assessment with xenon 133 single-photon emission computed tomography and the tower of london. *Archives of general psychiatry*, 49(12):943, 1992.

[5] H. Anton. *Elementary Linear Algebra*. Wiley & Sons, 8th edition, 2000.

[6] D. Arion, S. Horváth, D. A. Lewis, and K. Mirnics. Infragranular gene expression disturbances in the prefrontal cortex in schizophrenia: Signature of altered neural development? *Neurobiology of disease*, 37(3):738–746, 2010.

[7] D. Arion, T. Unger, D. A. Lewis, P. Levitt, and K. Mirnics. Molecular evidence for increased expression of genes related to immune and chaperone

function in the prefrontal cortex in schizophrenia. *Biological psychiatry*, 62(7):711–721, 2007.

[8] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[9] A. P. Association. Highlights of changes from DSM-IV-TR to DSM-5. *American Psychiatric Publishing*, 5, 2013.

[10] L. Astolfi, F. De Vico Fallani, F. Cincotti, D. Mattia, M. G. Marciani, S. Bufalari, S. Salinari, A. Colosimo, L. Ding, J. C. Edgar, W. Heller, G. A. Miller, B. He, and F. Babiloni. Imaging functional brain connectivity patterns from high-resolution EEG and fMRI via graph theory. *Psychophysiology*, 44(6):880–893, Nov. 2007.

[11] C. Aston, L. Jiang, and B. P. Sokolov. Microarray analysis of postmortem temporal cortex from patients with schizophrenia. *Journal of neuroscience research*, 77(6):858–866, 2004.

[12] J. P. Bantle. Dietary fructose and metabolic syndrome and diabetes. *The Journal of nutrition*, 139(6):1263S–1268S, 2009.

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

[14] H. Basciano, L. Federico, and K. Adeli. Fructose, insulin resistance, and metabolic dyslipidemia. *Nutrition & Metabolism*, 2(1):5, 2005.

[15] D. S. Bassett and E. T. Bullmore. Human brain networks in health and disease. *Current opinion in neurology*, 22(4):340, 2009.

[16] D. G. Beer, S. L. Kardia, C.-C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, L. Lin, G. Chen, T. G. Gharib, D. G. Thomas, et al. Gene-expression

profiles predict survival of patients with lung adenocarcinoma. *Nature medicine*, 8(8):816–824, 2002.

[17] S. Benetti, A. Mechelli, M. Picchioni, M. Broome, S. Williams, and P. McGuire. Functional integration between the posterior hippocampus and prefrontal cortex is impaired in both first episode schizophrenia and the at risk mental state. *Brain*, 132(9):2426–2436, 2009.

[18] V. Boginski, S. Butenko, and P. M. Pardalos. Mining market data: a network approach. *Computers & Operations Research*, 33(11):3171–3184, 2006.

[19] L. Bottou and Y. Bengio. Convergence properties of the k-means algorithms. In *Advances in Neural Information Processing Systems 7*. Citeseer, 1995.

[20] N. A. Bowden, J. Weidenhofer, R. J. Scott, U. Schall, J. Todd, P. T. Michie, and P. A. Tooney. Preliminary investigation of gene expression profiles in peripheral blood lymphocytes in schizophrenia. *Schizophrenia research*, 82(2):175–183, 2006.

[21] I. Braun, J. Genius, H. Grunze, A. Bender, H.-J. Möller, and D. Rujescu. Alterations of hippocampal and prefrontal gabaergic interneurons in an animal model of psychosis induced by nmda receptor antagonism. *Schizophrenia research*, 97(1):254–263, 2007.

[22] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Computer networks*, 33(1):309–320, 2000.

[23] M. S. Buchsbaum, R. J. Haier, S. G. Potkin, K. Nuechterlein, H. S. Bracha, M. Katz, J. Lohr, J. Wu, S. Lottenberg, P. A. Jerabek, et al. Frontostriatal disorder of cerebral metabolism in never-medicated schizophrenics. *Archives of General Psychiatry*, 49(12):935, 1992.

[24] C. Buettner, E. D. Muse, A. Cheng, L. Chen, T. Scherer, A. Pocai, K. Su, B. Cheng, X. Li, J. Harvey-White, G. J. Schwartz, G. Kunos, L. Rossetti, and C. Buettner. Leptin controls adipose tissue lipogenesis via central, STAT3-independent mechanisms. *Nat Med*, 14(6):667–675, June 2008.

[25] E. Bullmore, S. Frangou, and R. Murray. The dysplastic net hypothesis: an integration of developmental and dysconnectivity theories of schizophrenia. *Schizophrenia research*, 28(2):143–156, 1997.

[26] E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009.

[27] W. S. Bush and J. H. Moore. Chapter 11: Genome-wide association studies. *PLoS Comput Biol*, 8(12):e1002822, 12 2012.

[28] A. G. Cardno and I. I. Gottesman. Twin studies of schizophrenia: From bow-and-arrow concordances to star wars mx and functional genomics. *American journal of medical genetics*, 97(1):12–17, 2001.

[29] C. S. Carter and D. M. Barch. Cognitive neuroscience-based approaches to measuring and improving treatment effects on cognition in schizophrenia: the cntrics initiative. *Schizophrenia bulletin*, 33(5):1131–1137, 2007.

[30] R. Caspi, T. Altman, K. Dreher, C. A. Fulcher, P. Subhraveti, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, et al. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Research*, 40(D1):D742–D753, 2012.

[31] M. Catani and M. Mesulam. What is a disconnection syndrome? *Cortex*, 44(8):911–913, 2008.

[32] S. M. Cochran, M. Kennedy, C. E. McKerchar, L. J. Steward, J. A. Pratt, and B. J. Morris. Induction of metabolic hypofunction and neurochemical deficits after chronic intermittent exposure to phencyclidine: differential modulation by antipsychotic drugs. *Neuropsychopharmacology*, 28(2):265–275, 2003.

[33] J. Cosgrove and T. G. Newell. Recovery of neuropsychological functions during reduction in use of phencyclidine. *Journal of clinical psychology*, 1991.

[34] R. M. Craddock, J. T. Huang, E. Jackson, N. Harris, E. F. Torrey, M. Herberth, and S. Bahn. Increased α-defensins as a blood marker for schizophrenia susceptibility. *Molecular & Cellular Proteomics*, 7(7):1204–1213, 2008.

[35] I. Creese, D. R. Burt, and S. H. Snyder. Dopamine receptor binding predicts clinical and pharmacological potencies of antischizophrenic drugs. *Science*, 192(4238):481–483, 1976.

[36] J. J. Crofts and D. J. Higham. A weighted communicability measure applied to complex brain networks. *J R Soc Interface.*, 6(33):411–4, 2009.

[37] J. J. Crofts and D. J. Higham. Googling the brain: Discovering hierarchical and asymmetric network structures, with applications in neuroscience. *Internet Mathematics*, 7(4):233–254, 2011.

[38] X. Cui and G. A. Churchill. Statistical tests for differential expression in cdna microarray experiments. *Genome Biol*, 4(4):210, 2003.

[39] N. Dawson, M. McDonald, D. J. Higham, B. J. Morris, and J. A. Pratt. Subanesthetic ketamine treatment promotes abnormal interactions between neural subsystems and alters the properties of functional brain networks. *Neuropsychopharmacology*, 2014.

[40] N. Dawson, B. J. Morris, and J. A. Pratt. Subanaesthetic ketamine treatment alters prefrontal cortex connectivity with thalamus and ascending subcortical systems. *Schizophrenia bulletin*, 39(2):366–377, 2011.

[41] N. Dawson, R. J. Thompson, A. McVie, D. M. Thomson, B. J. Morris, and J. A. Pratt. Modafinil reverses phencyclidine-induced deficits in cognitive flexibility, cerebral metabolism, and functional brain connectivity. *Schizophrenia bulletin*, 38(3):457–474, 2012.

[42] N. Dawson, X. Xiao, M. McDonald, D. J. Higham, B. J. Morris, and J. A. Pratt. Sustained NMDA receptor hypofunction induces compromised neural systems integration and schizophrenia-like alterations in functional brain networks. *Cerebral Cortex*, 2012.

[43] S. de Jong, M. Boks, T. Fuller, E. Strengman, E. Janson, C. de Kovel, A. Ori, N. Vi, F. Mulder, J. Blom, et al. A gene co-expression network in whole blood of schizophrenia patients is independent of antipsychotic-use and enriched for brain-expressed genes. *PloS one*, 7(6):e39498, 2012.

[44] M. Debnath, D. M. Cannon, and G. Venkatsubramanian. Variation in the major histocompatibility complex [MHC] gene family in schizophrenia: Associations and functional implications. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 2012.

[45] I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. *Proceedings of the Seventh ACM SIGKDD Conference*, 2001.

[46] C. Diggle, M. Shires, D. Leitch, D. Brooke, I. Carr, A. Markham, B. Hayward, A. Asipu, and D. Bonthron. Ketohexokinase: expression and localization of the principal fructose-metabolizing enzyme. *Journal of Histochemistry & Cytochemistry*, 57(8):763, 2009.

[47] C. Ding and X. He. K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, New York, NY, USA, 2004. ACM.

[48] W. E. Donath and A. J. Hoffman. Lower bounds for the partitioning of graphs. *IBM Journal of Research and Development*, 17(5):420–425, 1973.

[49] V. Dror, E. Shamir, S. Ghanshani, R. Kimhi, M. Swartz, Y. Barak, R. Weizman, L. Avivi, T. Litmanovitch, E. Fantino, et al. hKCa3/KCNN3 potassium channel gene: association of longer CAG repeats with schizophrenia in israeli ashkenazi jews, expression in human tissues and localization to chromosome 1q21. *Molecular psychiatry*, 4(3):254–260, 1999.

[50] S. R. Eddy. Total information awareness for worm genetics. *Science*, 311(5766):1381–1382, 2006.

[51] E. Eden, D. Lipson, S. Yogev, and Z. Yakhini. Discovering motifs in ranked lists of DNA sequences. *PLoS computational biology*, 3(3):e39, 2007.

[52] E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini. Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC bioinformatics*, 10(1):48, 2009.

[53] A. Egerton, L. Reid, C. E. McKerchar, B. J. Morris, and J. A. Pratt. Impairment in perceptual attentional set-shifting following pcp administration: a rodent model of set-shifting deficits in schizophrenia. *Psychopharmacology*, 179(1):77–84, 2005.

[54] B. Ellenbroek and A. Cools. Animal models with construct validity for schizophrenia. *Behavioural pharmacology*, 1(6):469–490, 1990.

[55] I. Ellison-Wright, D. C. Glahn, A. R. Laird, S. M. Thelen, et al. The anatomy of first-episode and chronic schizophrenia: an anatomical likelihood estimation meta-analysis. *The American journal of psychiatry*, 165(8):1015, 2008.

[56] V. Emilsson, G. Thorleifsson, B. Zhang, A. Leonardson, F. Zink, J. Zhu, S. Carlson, A. Helgason, G. Walters, S. Gunnarsdottir, et al. Genetics of gene expression and its effect on disease. *Nature*, 452(7186):423–428, 2008.

[57] E. Estrada. *The Structure of Complex Networks.* Oxford University Press, Oxford, 2011.

[58] E. Estrada, D. Higham, and N. Hatano. Communicability and multipartite structures in complex networks at negative absolute temperatures. *Physical Review E*, 78(2), 2008.

[59] L. Euler. Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, 8:128–140, 1741.

[60] A. Feigin, K. L. Leenders, J. R. Moeller, J. Missimer, G. Kuenig, P. Spetsieris, A. Antonini, and D. Eidelberg. Metabolic Network Abnormalities in Early Huntington's Disease: An [18F]FDG PET Study. *Journal of Nuclear Medicine*, 42(11):1591–1595, Nov. 2001.

[61] M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(2):298–305, 1973.

[62] M. Fiedler. A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. *Czechoslovak Mathematical Journal*, 25(4):619–633, 1975.

[63] L. M. Fiori, A. Bureau, A. Labbe, J. Croteau, S. Noël, C. Mérette, and G. Turecki. Global gene expression profiling of the polyamine system in

suicide completers. *The International Journal of Neuropsychopharmacology*, 14(05):595–605, 2011.

[64] L. M. Fiori, B. Wanner, V. Jomphe, J. Croteau, F. Vitaro, R. E. Tremblay, A. Bureau, and G. Turecki. Association of polyaminergic loci with anxiety, mood disorders, and attempted suicide. *PloS one*, 5(11):e15146, 2010.

[65] A. Fornito, A. Zalesky, and E. T. Bullmore. Network scaling effects in graph analytic studies of human resting-state fMRI data. *Frontiers in systems neuroscience*, 4, 2010.

[66] P. Fusar-Poli, R. Smieskova, M. Kempton, B. Ho, N. Andreasen, and S. Borgwardt. Progressive brain changes in schizophrenia related to antipsychotic treatment? a meta-analysis of longitudinal MRI studies. *Neuroscience & Biobehavioral Reviews*, 2013.

[67] S. J. Galbraith, L. M. Tran, and J. C. Liao. Transcriptome network component analysis with limited microarray data. *Bioinformatics*, 22(15):1886–1894, Aug. 2006.

[68] E. J. Gardiner, M. J. Cairns, B. Liu, N. J. Beveridge, V. Carr, B. Kelly, R. J. Scott, and P. A. Tooney. Gene expression analysis reveals schizophrenia-associated dysregulation of immune pathways in peripheral blood mononuclear cells. *Journal of Psychiatric Research*, 2012.

[69] A. Gladkevich, H. F. Kauffman, and J. Korf. Lymphocytes as a neural probe: potential for studying psychiatric disorders. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 28(3):559–576, 2004.

[70] D. C. Glahn, A. R. Laird, I. Ellison-Wright, S. M. Thelen, J. L. Robinson, J. L. Lancaster, E. Bullmore, and P. T. Fox. Meta-analysis of gray matter

anomalies in schizophrenia: application of anatomic likelihood estimation and network analysis. *Biological psychiatry*, 64(9):774–781, 2008.

[71] S. J. Glatt, I. P. Everall, W. S. Kremen, J. Corbeil, R. Šášik, N. Khanlou, M. Han, C.-C. Liew, and M. T. Tsuang. Comparative gene expression analysis of blood and brain provides concurrent validation of SELENBP1 up-regulation in schizophrenia. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15533–15538, 2005.

[72] D. C. Goff and J. T. Coyle. The emerging role of glutamate in the pathophysiology and treatment of schizophrenia. *American Journal of Psychiatry*, 158(9):1367–1377, 2001.

[73] G. H. Golub and C. F. Van Loan. *Matrix computations (3rd ed.)*. Johns Hopkins University Press, Baltimore, MD, USA, 1996.

[74] M. F. Green and K. H. Nuechterlein. The matrics initiative: developing a consensus cognitive battery for clinical trials. *Schizophrenia research*, 72(1):1–3, 2004.

[75] P. Grindrod. Range-dependent random graphs and their application to modeling large small-world proteome datasets. *Physical Review E*, 66:066702–1 to 7, 2002.

[76] P. Grindrod, D. J. Higham, and G. Kalna. Perodic reordering. *The Institute of Mathematics and Its Applications (IMA) Journal of Numerical Analysis*, 30:195–207, 2010.

[77] P. Grindrod, D. J. Higham, G. Kalna, A. Spence, Z. Stoyanov, and J. K. Vass. DNA meets the SVD. *Mathematics Today*, 44:80–85, 2008.

[78] A. C. Guo, T. Jewison, M. Wilson, Y. Liu, C. Knox, Y. Djoumbou, P. Lo, R. Mandal, R. Krishnamurthy, and D. S. Wishart. Ecmdb: the e. coli metabolome database. *Nucleic acids research*, 41(D1):D625–D630, 2013.

[79] S. Hammoud, B. R. Emery, D. Dunn, R. B. Weiss, and D. T. Carrell. Sequence alterations in the YBX2 gene are associated with male factor infertility. *Fertility and sterility*, 91(4):1090–1095, 2009.

[80] D. Hanisch, A. Zien, R. Zimmer, and T. Lengauer. Co-clustering of biological networks and gene expression data. *Bioinformatics (Oxford, England)*, 18 Suppl 1:S145–54, Jan. 2002.

[81] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[82] T. Hashimoto, D. W. Volk, S. M. Eggan, K. Mirnics, J. N. Pierri, Z. Sun, A. R. Sampson, and D. A. Lewis. Gene expression deficits in a subclass of gaba neurons in the prefrontal cortex of subjects with schizophrenia. *The Journal of neuroscience*, 23(15):6315–6326, 2003.

[83] D. Y. Heng, W. Xie, M. M. Regan, M. A. Warren, A. R. Golshayan, C. Sahi, B. J. Eigl, J. D. Ruether, T. Cheng, S. North, et al. Prognostic factors for overall survival in patients with metastatic renal cell carcinoma treated with vascular endothelial growth factor–targeted agents: results from a large, multicenter study. *Journal of Clinical Oncology*, 27(34):5794–5799, 2009.

[84] B. Higgs, M. Elashoff, S. Richman, and B. Barci. An online database for brain disease research. *BMC genomics*, 7(1):70, 2006.

[85] D. J. Higham, G. Kalna, and M. Kibble. Spectral clustering and its use in bioinformatics. *J. Computational and Applied Math.*, 204:25–37, 2007.

[86] D. J. Higham, G. Kalna, and J. K. Vass. Analysis of the singular value decomposition as a tool for processing microarray expression data. *Proceedings of Algoritmy*, 2005.

[87] D. J. Higham, G. Kalna, and J. K. Vass. Spectral analysis of two-signed microarray expression data. *IMA Mathematical Medicine and Biology*, 24:131–148, 2007.

[88] O. Hikosaka. Basal ganglia - possible role in motor coordination and learning. *Current Opinion in Neurobiology*, 1(4):638–643, Dec. 1991.

[89] E. R. Hoff, R. R. Tubbs, J. L. Myles, and G. W. Procop. HER2/NEU Amplification in Breast Cancer Stratification by Tumor Type and Grade. *American journal of clinical pathology*, 117(6):916–921, 2002.

[90] C. J. Honey, O. Sporns, L. Cammoun, X. Gigandet, J. P. Thiran, R. Meuli, and P. Hagmann. Predicting human resting-state functional connectivity from structural connectivity. *Proceedings of the National Academy of Sciences*, 106(6):2035–2040, Feb. 2009.

[91] G. D. Honey, E. Pomarol-Clotet, P. R. Corlett, R. A. Honey, P. J. Mckenna, E. T. Bullmore, and P. C. Fletcher. Functional dysconnectivity in schizophrenia associated with attentional modulation of motor function. *Brain*, 128(11):2597–2611, 2005.

[92] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1985.

[93] D. W. Huang, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13, Jan. 2009.

[94] D. W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57, 2009.

[95] M. Ikeda, N. Iwata, T. Suzuki, T. Kitajima, Y. Yamanouchi, Y. Kinoshita, T. Inada, H. Ujike, and N. Ozaki. Association analysis of chromosome 5 $GABA_\alpha$ receptor cluster in japanese schizophrenia patients. *Biological psychiatry*, 58(6):440–445, 2005.

[96] I. Inc. Illumina transcriptome analysis kits, July 2013.

[97] Y. Iturria-Medina, R. C. Sotero, E. J. Canales-Rodríguez, Y. Alemán-Gómez, and L. Melie-García. Studying the human brain anatomical network via diffusion-weighted MRI and Graph Theory. *NeuroImage*, 40(3):1064–1076, Apr. 2008.

[98] K. D. Jakobsen, J. N. Frederiksen, T. Hansen, L. B. Jansson, J. Parnas, and T. Werge. Reliability of clinical ICD-10 schizophrenia diagnoses. *Nordic journal of psychiatry*, 59(3):209–212, 2005.

[99] A. Jasinska, O. Choi, J. DeYoung, O. Grujic, S. Kong, M. Jorgensen, J. Bailey, S. Breidenthal, L. Fairbanks, R. Woods, et al. Identification of brain transcriptional variation reproduced in peripheral blood: an approach for mapping brain expression traits. *Human molecular genetics*, 18(22):4415–4427, 2009.

[100] D. C. Javitt, S. R. Zukin, et al. Recent advances in the phencyclidine model of schizophrenia. *Am J Psychiatry*, 148(10):1301–1308, 1991.

[101] S. C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

[102] M. Kaiser and C. C. Hilgetag. Modelling the development of cortical systems networks. *Neurocomputing*, 58:297–302, 2004.

[103] G. Kalna, J. K. Vass, and D. J. Higham. Multidimensional partitioning and bi-partitioning: analysis and application to gene expression data sets. *Int. J. Comput. Math.*, 85(3-4):475–485, 2008.

[104] F. Kamada, Y. Aoki, A. Narisawa, Y. Abe, S. Komatsuzaki, A. Kikuchi, J. Kanno, T. Niihori, M. Ono, N. Ishii, et al. A genome-wide association study identifies RNF213 as the first moyamoya disease gene. *Journal of human genetics*, 56(1):34–40, 2010.

[105] M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[106] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, 40(D1):D109–D114, 2012.

[107] M. Kapushesky, I. Emam, E. Holloway, P. Kurnosov, A. Zorin, J. Malone, G. Rustici, E. Williams, H. Parkinson, and A. Brazma. Gene expression atlas at the european bioinformatics institute. *Nucleic acids research*, 38(suppl 1):D690–D698, 2010.

[108] R. S. Keefe, R. M. Bilder, S. M. Davis, P. D. Harvey, B. W. Palmer, J. M. Gold, H. Y. Meltzer, M. F. Green, G. Capuano, T. S. Stroup, et al. Neurocognitive effects of antipsychotic medications in patients with chronic schizophrenia in the catie trial. *Archives of general psychiatry*, 64(6):633, 2007.

[109] R. S. Keefe, R. C. Mohs, R. M. Bilder, P. D. Harvey, M. F. Green, H. Y. Meltzer, J. M. Gold, and M. Sano. Neurocognitive assessment in the clinical antipsychotic trials of intervention effectiveness (catie) project schizophrenia trial: development, methodology, and rationale. *Schizophrenia Bulletin*, 29(1):45, 2003.

[110] J. J. Kim, J. H. Seok, H. J. Park, D. S. Lee, M. C. Lee, and J. S. Kwon. Functional disconnection of the semantic networks in schizophrenia. *Neuroreport*, 16(4):355–359, 2005.

[111] H. Kitano. Systems biology: a brief overview. *Science*, 295(5560):1662–1664, 2002.

[112] Y. Kluger, R. Basri, J. T. Chang, and M. Gerstein. Spectral biclustering of microarray cancer data: co-clustering genes and conditions. *Genome Research*, 13:703–716, 2003.

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

[114] A. Korieh and G. Crouzoulon. Dietary regulation of fructose metabolism in the intestine and in the liver of the rat. duration of the effects of a high fructose diet after the return to the standard diet. *Archives internationales de physiologie, de biochimie et de biophysique*, 99(6):455, 1991.

[115] J. L. Kovar, W. Volcheck, E. Sevick-Muraca, M. A. Simpson, and D. M. Olive. Characterization and performance of a near-infrared 2-deoxyglucose optical imaging agent for mouse cancer models. *Analytical biochemistry*, 384(2):254–262, Jan. 2009.

[116] M. Koyuturk, S. Subramaniam, and A. Grama. Introduction to network biology. In *Functional Coherence of Molecular Networks in Bioinformatics*, pages 1–13. Springer, 2012.

[117] L. V. Kristiansen, S. A. Patel, V. Haroutunian, and J. H. Meador-Woodruff. Expression of the *NR2B-NMDA* receptor subunit and its TBR-1/CINAP

regulatory proteins in postmortem brain suggest altered receptor processing in schizophrenia. *Synapse*, 64(7):495–502, 2010.

[118] S. M. Kurian, R. Heilman, T. S. Mondala, A. Nakorchevsky, J. A. Hewel, D. Campbell, E. H. Robison, L. Wang, W. Lin, L. Gaber, et al. Biomarkers for early and late stage chronic allograft nephropathy by proteogenomic profiling of peripheral blood. *PLoS One*, 4(7):e6212, 2009.

[119] S. Kuroda and K. Houkin. Moyamoya disease: current concepts and future perspectives. *The Lancet Neurology*, 7(11):1056–1066, 2008.

[120] V. Lacroix, L. Cottret, P. Thebault, and M.-F. Sagot. An introduction to metabolic networks and their structural analysis. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 5(4):594 –617, oct.-dec. 2008.

[121] A. C. Lahti, H. H. Holcomb, D. R. Medoff, and C. A. Tamminga. Ketamine activates psychosis and alters limbic blood flow in schizophrenia. *Neuroreport*, 6(6):869–872, Apr. 1995.

[122] A. C. Lahti, M. A. Weiler, M. Tamara, A. Parwani, and C. A. Tamminga. Effects of ketamine in normal and schizophrenic volunteers. *Neuropsychopharmacology*, 25(4):455–467, 2001.

[123] C. Lee, D. Higham, D. Crowther, and J. K. Vass. Non-negative matrix factorisation for network reordering. *Monografias de la Real Academia de Ciencias de Zaragoza*, 33:39–53, 2010.

[124] C. H. Lee, B. O. Alpert, P. Sankaranarayanan, and O. Alter. GSVD comparison of patient-matched normal and tumor aCGH profiles reveals global copy-number alterations predicting glioblastoma multiforme survival. *PLoS ONE*, 7:e30098, 2012.

[125] H. J. Lee, J. Y. Song, J. W. Kim, S.-Y. Jin, M. S. Hong, J. K. Park, J.-H. Chung, H. Shibata, Y. Fukumaki, et al. Association study of polymorphisms in synaptic vesicle-associated genes, SYN2 and CPLX2, with schizophrenia. *Behav Brain Funct*, 1:15, 2005.

[126] S. Leucht, C. Corves, D. Arbter, R. R. Engel, C. Li, and J. M. Davis. Second-generation versus first-generation antipsychotic drugs for schizophrenia: a meta-analysis. *The Lancet*, 373(9657):31–41, 2009.

[127] D. F. Levinson, B. J. Mowry, M. A. Escamilla, and S. V. Faraone. The lifetime dimensions of psychosis scale (LDPS): description and interrater reliability. *Schizophrenia bulletin*, 28(4):683–695, 2002.

[128] M. Liang, Y. Zhou, T. Jiang, Z. Liu, L. Tian, H. Liu, and Y. Hao. Widespread functional disconnectivity in schizophrenia with resting-state functional magnetic resonance imaging. *Neuroreport*, 17(2):209–213, 2006.

[129] J. C. Liao, R. Boscolo, Y.-L. Yang, L. M. Tran, C. Sabatti, and V. P. Roychowdhury. Network component analysis: Reconstruction of regulatory signals in biological systems. *Proceedings of the National Academy of Sciences*, 100(26):15522–15527, Dec. 2003.

[130] P. Lichtenstein, B. H. Yip, C. Björk, Y. Pawitan, T. D. Cannon, P. F. Sullivan, and C. M. Hultman. Common genetic determinants of schizophrenia and bipolar disorder in swedish families: a population-based study. *The Lancet*, 373(9659):234–239, 2009.

[131] J. A. Lieberman and T. S. Stroup. The nimh-catie schizophrenia study: what did we learn? *American Journal of Psychiatry*, 168(8):770–775, 2011.

[132] B. K. Lipska. Using animal models to test a neurodevelopmental hypothesis of schizophrenia. *Journal of Psychiatry and Neuroscience*, 29(4):282, 2004.

[133] B. K. Lipska, A. Deep-Soboslay, C. S. Weickert, T. M. Hyde, C. E. Martin, M. M. Herman, and J. Kleinman. Critical factors in gene expression in postmortem human brain: Focus on studies in schizophrenia. *Biological psychiatry*, 60(6):650–658, 2006.

[134] B. K. Lipska, G. E. Jaskiw, and D. R. Weinberger. Postpubertal emergence of hyperresponsiveness to stress and to amphetamine after neonatal excitotoxic hippocampal damage: a potential animal model of schizophrenia. *Neuropsychopharmacology*, 9(1):67–75, 1993.

[135] B. K. Lipska and D. R. Weinberger. To model a psychiatric disorder in animals: schizophrenia as a reality test. *Neuropsychopharmacology*, 23(3):223–239, 2000.

[136] Y. Liu, M. Liang, Y. Zhou, Y. He, Y. Hao, M. Song, C. Yu, H. Liu, Z. Liu, and T. Jiang. Disrupted small-world networks in schizophrenia. *Brain*, 131(4):945–961, 2008.

[137] R. Luengo-Fernandez, J. Leal, and A. Gray. UK research expenditure on dementia, heart disease, stroke and cancer: are levels of spending related to disease burden? *European journal of Neurology*, 19(1):149–154, 2012.

[138] C. Lynch. Big data: How do your data grow? *Nature*, 455(7209):28–29, 2008.

[139] E. A. Maguire, C. J. Mummery, and C. Büchel. Patterns of hippocampal-cortical interaction dissociate temporal lobe memory subsystems. *Hippocampus*, 10(4):475–482, 2000.

[140] R. Mangalore, M. Knapp, et al. Cost of schizophrenia in England. *Journal of Mental Health Policy and Economics*, 10(1):23, 2007.

[141] T. J. Mariani, V. Budhraja, B. H. Mecham, C. C. Gu, M. A. Watson, and Y. Sadovsky. A variable fold change threshold determines significance for expression microarrays. *The FASEB Journal*, 17(2):321–323, 2003.

[142] A. Marsman, M. P. van den Heuvel, D. W. Klomp, R. S. Kahn, P. R. Luijten, and H. E. H. Pol. Glutamate in schizophrenia: a focused review and meta-analysis of 1H-MRS studies. *Schizophrenia bulletin*, 39(1):120–129, 2013.

[143] R. Martın-Santos, A. Fagundo, J. Crippa, Z. Atakan, S. Bhattacharyya, P. Allen, P. Fusar-Poli, S. Borgwardt, M. Seal, G. Busatto, et al. Neuroimaging in cannabis use: a systematic review of the literature. *Psychological medicine*, 40(3):383–398, 2010.

[144] D. H. Mathalon, E. V. Sullivan, K. O. Lim, and A. Pfefferbaum. Progressive brain volume changes and the clinical course of schizophrenia in men: a longitudinal magnetic resonance imaging study. *Archives of General Psychiatry*, 58(2):148, 2001.

[145] J. M. McClellan, E. Susser, and M.-C. King. Schizophrenia: a common disease caused by multiple rare alleles. *The British Journal of Psychiatry*, 190(3):194–199, 2007.

[146] M. McDonald, D. J. Higham, and J. K. Vass. Spectral algorithms for heterogeneous biological networks. *Briefings in functional genomics*, 11(6):457–468, 2012.

[147] W. T. McKinney and E. C. Moran. Animal models of schizophrenia. *Am J Psychiatry*, 138(4), 1981.

[148] E. L. Meaburn, C. Fernandes, I. W. Craig, R. Plomin, and L. C. Schalkwyk. Assessing individual differences in genome-wide gene expression in human

whole blood: reliability over four hours and stability over 10 months. *Twin Research and Human Genetics*, 12(4):372–380, 2009.

[149] J. H. Meador-Woodruff and D. J. Healy. Glutamate receptor expression in schizophrenic brain. *Brain Research Reviews*, 31(2):288–294, 2000.

[150] N. Melhem, F. Middleton, K. McFadden, L. Klei, S. V. Faraone, S. Vinogradov, J. Tiobech, V. Yano, S. Kuartei, K. Roeder, et al. Copy number variants for schizophrenia and related psychotic disorders in oceanic palau: risk and transmission in extended pedigrees. *Biological psychiatry*, 70(12):1115–1121, 2011.

[151] A. S. Meyer-Lindenberg, R. K. Olsen, P. D. Kohn, T. Brown, M. F. Egan, D. R. Weinberger, and K. F. Berman. Regionally specific disturbance of dorsolateral prefrontal-hippocampal functional connectivity in schizophrenia. *Archives of General Psychiatry*, 62(4):379, 2005.

[152] S. Micheloyannis, E. Pachou, C. J. J. Stam, M. Breakspear, P. Bitsios, M. Vourkas, S. Erimaki, and M. Zervakis. Small-world networks and disturbed functional connectivity in schizophrenia. *Schizophrenia research*, 87(1-3):60–66, 2006.

[153] E. Mignot, S. Nishino, C. Guilleminault, and W. Dement. Modafinil binds to the dopamine uptake carrier site with low affinity. *Sleep*, 17(5):436, 1994.

[154] K. Mirnics, F. Middleton, G. Stanwood, D. Lewis, and P. Levitt. Disease-specific changes in regulator of g-protein signaling 4 (RGS4) expression in schizophrenia. *Molecular psychiatry*, 6(3):293–301, 2001.

[155] K. Mirnics, F. A. Middleton, D. A. Lewis, and P. Levitt. Analysis of complex brain disorders with gene expression microarrays: schizophrenia as a disease of the synapse. *Trends in neurosciences*, 24(8):479–486, 2001.

[156] T. Mizuno. Neurotensin receptor (NTSR). *Encyclopedia of Signaling Molecules*, pages 1203–1208, 2012.

[157] D. S. Moore and G. McCabe. *Introduction to the Practice of Statistics.* WH Freeman, 2008.

[158] C. J. Morgan, A. Mofeez, B. Brandner, L. Bromley, and H. V. Curran. Acute effects of ketamine on memory systems and psychotic symptoms in healthy volunteers. *Neuropsychopharmacology*, 2004.

[159] J. L. Morrison, R. Breitling, D. J. Higham, and D. R. Gilbert. A lock-and-key model for protein-protein interactions. *Bioinformatics*, 22(16):2012–2019, Aug. 2006.

[160] K. Nakazawa, V. Zsiros, Z. Jiang, K. Nakao, S. Kolata, S. Zhang, and J. E. Belforte. GABAergic interneuron origin of schizophrenia pathophysiology. *Neuropharmacology*, 62(3):1574–1583, 2012.

[161] S. Narayan, B. Tang, S. R. Head, T. J. Gilmartin, J. G. Sutcliffe, B. Dean, and E. A. Thomas. Molecular profiles of schizophrenia in the cns at different stages of illness. *Brain research*, 1239:235–248, 2008.

[162] E. J. Nestler and S. E. Hyman. Animal models of neuropsychiatric disorders. *Nature neuroscience*, 13(10):1161–1169, 2010.

[163] H. Neumann, H. Schmidt, A. Cavalie, D. Jenne, and H. Wekerle. Major histocompatibility complex (MHC) class I gene expression in single neurons of the central nervous system: differential regulation by interferon (IFN)-γ and tumor necrosis factor (TNF)-α. *The Journal of experimental medicine*, 185(2):305–316, 1997.

[164] J. Newcomer. Ketamine-Induced NMDA Receptor Hypofunction as a Model of Memory Impairment and Psychosis. *Neuropsychopharmacology*, 20(2):106–118, Feb. 1999.

[165] M. E. J. Newman. Mixing patterns in networks. *Phys. Rev. E*, 67:026126, Feb 2003.

[166] M. E. J. Newman. *Networks: An Introduction.* Oxford University Press, New York, 2010.

[167] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, Feb 2004.

[168] M. Niesters, N. Khalili-Mahani, C. Martini, L. Aarts, J. van Gerven, M. A. van Buchem, A. Dahan, and S. Rombouts. Effect of subanesthetic ketamine on intrinsic functional brain connectivity: a placebo-controlled functional magnetic resonance imaging study in healthy male volunteers. *Anesthesiology*, 117(4):868–877, 2012.

[169] T. Ogawa, A. Inugami, J. Hatazawa, I. Kanno, M. Murakami, N. Yasui, K. Mineura, and K. Uemura. Clinical positron emission tomography for brain tumors: comparison of fludeoxyglucose F18 and L-methyl-11C-methionine. *American Journal of Neuroradiology*, 17(2):345–353, Feb. 1996.

[170] B. Olabi, I. Ellison-Wright, A. M. McIntosh, S. J. Wood, E. Bullmore, and S. M. Lawrie. Are there progressive brain changes in schizophrenia? A meta-analysis of structural magnetic resonance imaging studies. *Biological psychiatry*, 70(1):88–96, 2011.

[171] J. B. Olesen, G. Y. Lip, M. L. Hansen, P. R. Hansen, J. S. Tolstrup, J. Lindhardsen, C. Selmer, O. Ahlehoff, A.-M. S. Olsen, G. H. Gislason, et al.

Validation of risk stratification schemes for predicting stroke and thromboembolism in patients with atrial fibrillation: nationwide cohort study. *BMJ: British Medical Journal*, 342, 2011.

[172] T. Otowa, E. Yoshida, N. Sugaya, S. Yasuda, Y. Nishimura, K. Inoue, M. Tochigi, T. Umekage, T. Miyagawa, N. Nishida, et al. Genome-wide association study of panic disorder in the japanese population. *Journal of human genetics*, 54(2):122–126, 2009.

[173] M. J. Owen, N. Craddock, and M. C. O'Donovan. Suggestion of roles for both common and rare risk variants in genome-wide studies of schizophrenia. *Archives of General Psychiatry*, 67(7):667, 2010.

[174] G. Papadimitriou, D. Dikeos, E. Daskalopoulou, G. Karadima, D. Avramopoulos, C. Contis, and C. Stefanis. Association between GABA-α receptor alpha 5 subunit gene locus and schizophrenia of a later age of onset. *Neuropsychobiology*, 43(3):141–144, 2001.

[175] A. Peled, A. B. Geva, W. S. Kremen, H. M. Blankfeld, R. Esfandiarfard, and T. E. Nordahl. Functional connectivity and working memory in schizophrenia: an eeg study. *International Journal of Neuroscience*, 106(1-2):47–61, 2001.

[176] M. Poeze, M. J. Bruins, Y. C. Luiking, and N. E. Deutz. Reduced caloric intake during endotoxemia reduces arginine availability and metabolism. *The American journal of clinical nutrition*, 91(4):992–1001, 2010.

[177] S. P. Ponnapalli, M. A. Saunders, C. F. Van Loan, and O. Alter. A higher-order Generalized Singular Value Decomposition for comparison of global mRNA expression from multiple organisms. *PLoS ONE*, 6:e28072, 12 2011.

[178] S. Prabakaran, J. Swatton, M. Ryan, S. Huffaker, J.-J. Huang, J. Griffin, M. Wayland, T. Freeman, F. Dudbridge, K. Lilley, et al. Mitochondrial

dysfunction in schizophrenia: evidence for compromised brain metabolism and oxidative stress. *Molecular psychiatry*, 9(7):684–697, 2004.

[179] J. Pratt, C. Winchester, N. Dawson, and B. Morris. Advancing schizophrenia drug discovery: optimizing rodent models to bridge the translational gap. *Nature Reviews Drug Discovery*, 11(7):560–579, 2012.

[180] J. Pratt, C. Winchester, A. Egerton, S. Cochran, and B. Morris. Modelling prefrontal cortex deficits in schizophrenia: implications for treatment. *British journal of pharmacology*, 153(S1):S465–S470, 2008.

[181] S. Raychaudhuri, J. M. Stuart, and R. B. Altman. Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 455–466, 2000.

[182] P. M. Roberts. Mining literature for systems biology. *Briefings in bioinformatics*, 7(4):399–406, 2006.

[183] B. Rollins, M. Martin, L. Morgan, and M. Vawter. Analysis of whole genome biomarker expression in blood and brain. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 153(4):919–936, 2010.

[184] D. Ruano, Y. S. Aulchenko, A. Macedo, M. J. Soares, J. Valente, M. H. Azevedo, M. H. Hutz, C. S. Gama, M. I. Lobato, P. Belmonte-de Abreu, et al. Association of the gene encoding neurogranin with schizophrenia in males. *Journal of psychiatric research*, 42(2):125–133, 2008.

[185] M. Rubinov and O. Sporns. Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, 52(3):1059–1069, Sept. 2010.

[186] Q. Saleem, V. Sreevidya, J. Sudhir, J. V. Savithri, Y. Gowda, C. B-Rao, V. Benegal, P. P. Majumder, A. Anand, S. K. Brahmachari, et al. Association

analysis of cag repeats at the kcnn3 locus in indian patients with bipolar disorder and schizophrenia. *American journal of medical genetics*, 96(6):744–748, 2000.

[187] N. Salkind and H. Abdi. *Encyclopedia of Measurement and Statistics - Chapter: Singular Value Decomposition (SVD) and Generalized Singular Value Decomposition (GSVD).* Sage, 1st edition, 2007.

[188] F. Sams-Dodd. Phencyclidine-induced stereotyped behaviour and social isolation in rats: a possible animal model of schizophrenia. *Behavioural pharmacology*, 7(1):3–23, 1996.

[189] A. Schreiber, N. J. Shirley, R. A. Burton, and G. B. Fincher. Combining transcriptional datasets using the generalized singular value decomposition. *BMC Bioinformatics*, 9:335, 2008.

[190] P. A. Sequeira, M. V. Martin, and M. P. Vawter. The first decade and beyond of transcriptional profiling in schizophrenia. *Neurobiology of disease*, 45(1):23–36, 2012.

[191] L. Shao and M. P. Vawter. Shared gene expression alterations in schizophrenia and bipolar disorder. *Biological psychiatry*, 64(2):89–97, 2008.

[192] L. Shihabuddin, M. S. Buchsbaum, E. A. Hazlett, M. M. Haznedar, P. D. Harvey, A. Newman, D. B. Schnur, J. Spiegel-Cohen, T. Wei, J. Machac, et al. Dorsal striatal size, shape, and metabolic rate in never-medicated and previously medicated schizophrenics performing a verbal learning task. *Archives of General Psychiatry*, 55(3):235, 1998.

[193] T. A. Snijders. The statistical evaluation of social network dynamics. *Sociological methodology*, 31(1):361–395, 2001.

[194] L. Sokoloff, M. Reivich, C. Kennedy, M. H. D. Rosiers, C. S. Patlak, K. D. Pettigrew, O. Sakurada, and M. Shinohara. The 14C deoxyglucose method for the measurement of local cerebral glucose utilization: Theory, procedure, and normal values in the conscious and anesthetized albino rat. *Journal of Neurochemistry*, 28(5):897–916, 1977.

[195] O. Sporns. *Networks of the Brain.* The MIT Press, 2011.

[196] O. Sporns, D. Chialvo, M. Kaiser, and C. Hilgetag. Organization, development and function of complex brain networks. *Trends in Cognitive Sciences*, 8(9):418–425, Sept. 2004.

[197] H. Stefansson, R. A. Ophoff, S. Steinberg, O. A. Andreassen, S. Cichon, D. Rujescu, T. Werge, O. P. Pietiläinen, O. Mors, P. B. Mortensen, et al. Common variants conferring risk of schizophrenia. *Nature*, 460(7256):744–747, 2009.

[198] L. J. Steward, M. D. Kennedy, B. J. Morris, and J. A. Pratt. The atypical antipsychotic drug clozapine enhances chronic PCP-induced regulation of prefrontal cortex 5-HT2A receptors. *Neuropharmacology*, 47(4):527–537, 2004.

[199] P. F. Sullivan, C. Fan, and C. M. Perou. Evaluating the comparability of gene expression in blood and brain. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 141(3):261–268, 2006.

[200] F. Supek, M. Bošnjak, N. Škunca, and T. Šmuc. Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS One*, 6(7):e21800, 2011.

[201] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572, 2002.

[202] D. Tkachev, M. L. Mimmack, M. M. Ryan, M. Wayland, T. Freeman, P. B. Jones, M. Starkey, M. J. Webster, R. H. Yolken, and S. Bahn. Oligodendrocyte dysfunction in schizophrenia and bipolar disorder. *The Lancet*, 362(9386):798–805, 2003.

[203] H. Tomita, M. P. Vawter, D. M. Walsh, S. J. Evans, P. V. Choudary, J. Li, K. M. Overman, M. E. Atz, R. M. Myers, E. G. Jones, et al. Effect of agonal and postmortem factors on gene expression profile: quality control in microarray analyses of postmortem human brain. *Biological psychiatry*, 55(4):346–352, 2004.

[204] G. Tononi. Schizophrenia and the mechanisms of conscious integration. *Brain Research Reviews*, 31(2-3):391–400, Mar. 2000.

[205] M. T. Tsuang, N. Nossova, T. Yager, M.-M. Tsuang, S.-C. Guo, K. G. Shyu, S. J. Glatt, and C. Liew. Assessing the validity of blood-based gene expression profiles for the classification of schizophrenia and bipolar disorder: A preliminary report. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 133(1):1–5, 2005.

[206] M. T. Tsuang, W. S. STONE, and S. V. Faraone. Genes, environment and schizophrenia. *The British Journal of Psychiatry*, 178(40):s18–s24, 2001.

[207] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001.

[208] P. Tyrer and T. Kendall. The spurious advance of antipsychotic drug therapy. *The Lancet*, 373(9657):4–5, 2009.

[209] T. B. Üstün, J. Rehm, S. Chatterji, S. Saxena, R. Trotter, R. Room, and J. Bickenbach. Multiple-informant ranking of the disabling effects of different health conditions in 14 countries. *The Lancet*, 354(9173):111–115, 1999.

[210] M. P. van den Heuvel and H. E. Hulshoff Pol. Exploring the brain network: A review on resting-state fMRI functional connectivity. *European Neuropsychopharmacology*, 20(8):519–534, Aug. 2010.

[211] R. Van Driessche and D. Roose. An improved spectral bisection algorithm and its application to dynamic load balancing. *Parallel Computing*, 21(1):29–48, 1995.

[212] J. Varadarajulu, A. Schmitt, P. Falkai, M. Alsaif, C. W. Turck, and D. Martins-de Souza. Differential expression of HINT1 in schizophrenia brain tissue. *European archives of psychiatry and clinical neuroscience*, 262(2):167–172, 2012.

[213] M. Vawter, H. Tomita, F. Meng, B. Bolstad, J. Li, S. Evans, P. Choudary, M. Atz, L. Shao, C. Neal, et al. Mitochondrial-related gene expression changes are sensitive to agonal-pH state: implications for brain disorders. *Molecular psychiatry*, 11(7):663–679, 2006.

[214] M. P. Vawter, J. M. Crook, T. M. Hyde, J. E. Kleinman, D. R. Weinberger, K. G. Becker, and W. J. Freed. Microarray analysis of gene expression in the prefrontal cortex in schizophrenia: a preliminary study. *Schizophrenia research*, 58(1):11–20, 2002.

[215] P. Whitaker-Azmitia, A. Borella, and N. Raio. Serotonin depletion in the adult rat causes loss of the dendritic marker MAP-2: A new animal model of schizophrenia? *Neuropsychopharmacology*, 12(3):269–272, 1995.

[216] T. Whitford, S. Grieve, T. Farrow, L. Gomes, J. Brennan, A. Harris, E. Gordon, and L. Williams. Volumetric white matter abnormalities in first-episode schizophrenia: a longitudinal, tensor-based morphometry study. *American Journal of Psychiatry*, 164(7):1082–1089, 2007.

[217] L. Wilkinson and M. Friendly. The history of the cluster heat map. *The American Statistician*, 63(2), 2009.

[218] J. Wu, M. Buchsbaum, and W. Bunney. Positron emission tomography study of phencyclidine users as a possible drug model of schizophrenia. *Yakubutsu, seishin, Japanese journal of psychopharmacology*, 11(1):47, 1991.

[219] X. Xiao. *Complex Networks and the Generalized Singular Value Decomposition*. PhD thesis, University of Strathclyde, 2011.

[220] X. Xiao, N. Dawson, L. MacIntyre, B. J. Morris, J. A. Pratt, D. G. Watson, and D. J. Higham. Exploring metabolic pathway disruption in the subchronic phencyclidine model of schizophrenia with the generalized singular value decomposition. *BMC Systems Biology*, 5(1):72, 2011.

[221] Y. Yang and A. Raine. Prefrontal structural and functional brain imaging findings in antisocial, violent, and psychopathic individuals: a meta-analysis. *Psychiatry Research: Neuroimaging*, 174(2):81–88, 2009.