Explicating a biological basis for Chronic Fatigue Syndrome

by

Samar A. Abou-Gouda

A thesis submitted to the
School of Computing
in conformity with the requirements for
the degree of Master of Science

Queen's University
Kingston, Ontario, Canada
December 2007

Abstract

In the absence of clinical markers for Chronic Fatigue Syndrome (CFS), research to find a biological basis for it is still open. Many data-mining techniques have been widely employed to analyze biomedical data describing different aspects of CFS. However, the inconsistency of the results of these studies reflect the uncertainty in regards to the real basis of this disease. In this thesis, we show that CFS has a biological basis that is detectable in gene expression data better than blood profile and Single Nucleotide Polymorphism (SNP) data. Using random forests, the analysis of gene expression data achieves a prediction accuracy of approximately 89%. We also identify sets of differentially expressed candidate genes that might contribute to CFS. We show that the integration of data spanning multiple levels of the biological scale might reveal further insights into the understanding of CFS. Using integrated data, we achieve a prediction accuracy of approximately 91%. We find that Singular Value Decomposition (SVD) is a useful technique to visualize the performance of random forests.

Acknowledgments

I would like to thank my supervisor Dr. David Skillicorn for his great support. Thanks to the School of Computing for the financial support. Thanks to my family, colleagues and friends for their love and encouragement.

Contents

List of Figures

List of Tables

List of Acronyms

ANOVA	Analysis of Variance
CAMDA	Clinical Assessment of Microarray Data Analysis
CDC	Center for Disease Control
CFS	Chronic Fatigue Syndrome
GO	Gene Ontology
ICA	Independent Component Analysis
ISF	Insufficient Symptoms for Fatigue
LCA	Latent Class Analysis
MFI	Multidimensional Fatigue Inventory
NF	Non-Fatigued
OOB	Out-Of-Bag
PCA	Principle Component Analysis
SF-36	Survey Short Form-36
SI	Symptom Inventory
SNP	Single Nucleotide Polymorphisms
SVD	Singular Value Decomposition
SVM	Support Vector Machines

Chapter 1

Introduction

1.1 Motivation

A great amount of data is being generated every day by different applications in the medical industry. In order for this data to be beneficial for clinicians, it must be analyzed and processed. The process of analyzing data to discover hidden useful knowledge is called data mining [32]. Different data-mining techniques have been explored to extract new information from the biomedical data of different diseases.

During the last decade, biomedical research has been growing dramatically. This field mostly deals with disorders in biological systems, seeks to develop new drugs and finds treatments for diseases. Moreover, this field studies the human genome to discover genetic causes for diseases which, until now, researchers have not been able to find, such as Chronic Fatigue Syndrome (CFS) [32]. Researchers in this field study the genetic level in organisms trying to find biological markers to diagnose these diseases.

1.2 Problem

CFS is characterized by persistent fatigue accompanied by symptoms describing physical and mental impairments. It is an illness with unknown cause and treatment. CFS has no clear biological markers and its diagnosis depends on self-reported symptoms. Some people who suffer from CFS are not recognized while others are wrongly diagnosed as CFS patients. Due to the ambiguity of CFS, many hypotheses are proposed about its mechanism.

Recent studies have reported that CFS affects between 400,000 and 2.2 million people in the United States [9, 43]. People suffering from this disorder are less productive and face many limitations in their daily life. The fatigue might last for up to 20 years. The national productivity loss resulted from CFS is comparable to other serious diseases [45]. Ongoing research to determine the causes of CFS helps in finding treatments and avoiding factors that might increase the susceptibility to CFS.

1.3 State of the art

While many studies have focused on the correctness of CFS definition [23, 43], others have aimed to explore CFS by analyzing different types of data spanning different levels of the biological scale using different data-mining techniques [8, 16, 27, 52]. The main goal for these studies is to find biological markers whose functional behavior correlates with CFS symptoms. Some studies have been conducted to explore this unexplained disorder by constructing integrative models that combine multiple types of data pertaining to CFS [22, 34, 35, 41, 55]. Clustering and classification analysis of various types of data have been used to discriminate between CFS patients and

controls. To some extent, the inconsistency of the results of these studies reflect the uncertainty about the real basis of this disease.

1.4 Contributions

This research is interested in finding biological markers for CFS. We show that analyzing information from blood profiles, gene expression levels and Single Nucleotide Polymorphisms (SNPs) lead to indications for the existence of CFS. Furthermore, we show that extracting useful information from integrated data spanning these multiple levels of biological scale might reveal further insights into the understanding of CFS. We use Singular Value Decomposition (SVD) to look for clusters of patients who share similar patterns. We also use random forests to classify CFS patients versus non-fatigued controls. Both methods are used to select the most relevant features for CFS.

The contributions of this work can be summarized as follows. We show that CFS has a biological basis that is detectable in gene expression data better than blood and SNP data. The analysis of gene expression data achieves a prediction accuracy of approximately 89%. Our analysis of integrated data achieves a prediction accuracy of approximately 91%. In addition to that, we illustrate the effectiveness of the importance measurement of random forests in selecting the most distinguishable features between CFS patients and controls. We also show that SVD is a useful technique to visualize the structure of the data.

1.5 Organization of Thesis

The remainder of this thesis is organized as follows. In Chapter 2, we provide the background information regarding Chronic fatigue Syndrome(CFS), the Critical Assessment of Microarray Data Analysis (CAMDA) data, and related work. We also give a brief introduction about the two analysis techniques we used in this study, Random Forests and Singular Value Decomposition (SVD). Chapter 3 explains the experimental model we use to analyze the CAMDA dataset. In Chapter 4, we present and discuss the results of our analysis. Finally, Chapter 5 summarizes our analysis and outlines the limitation of this area of study.

Chapter 2

Background

This chapter discusses the background information and work related to this thesis. First, the Chronic Fatigue Syndrome (CFS), the dataset used in our analysis and its source (CAMDA) are discussed. Second, related work conducted to investigate the biological markers for CFS is reviewed. Last, random forests and matrix decompositions techniques, in particular Singular Value Decomposition (SVD) and Independent Component Analysis (ICA), and their applications in the biomedical field are studied.

2.1 Chronic Fatigue Syndrome (CFS)

CFS is a complex disease characterized by deep and persistent fatigue, in addition to other symptoms such as temporary memory loss, severe headache and mental disorders. According to the 1994 case definition, this unexplained fatigue and some of its symptoms must last for at least six months before the patient receive a CFS diagnosis [23]. CFS has no diagnostic signs, tests or risk factors that have been confirmed in scientific studies. As a result, no definitive treatments for it exists [5, 23].

In fact, CFS affects a significant percentage of the population in United States [43], which makes solving the obscurity surrounding it important. Moreover, finding new models to understand the nature of CFS could answer many questions about other unexplained diseases such as cancer.

Scientists believe that complex diseases such as CFS arise from combined genetic factors and environmental factors. Hence, the study of the variation in genetic makeup, together with the clinical assessment of patients, might give useful indications to differentiate between patients and controls [33]. These indications might be detected as local patterns by finding a subset of genes that can be used as biological markers for the disease more efficiently than other genes. On the other hand, global indications can be found by merging more than one type of data ranging from phenotypes to genotypes of the patients. Such integrative models enable us to analyze complex datasets that come from complex diseases.

2.2 CAMDA dataset

The assessment of CFS is complicated, so establishing well-defined measures of this illness is important. The Center for Disease Control (CDC) in Wichita, KS [17] provides data from different sources to help researchers study the disease and test their proposed methods, hence forming a better understanding of the disease. Part of this data is available through the Critical Assessment of Microarray Data Analysis (CAMDA) competition. It consists of clinical assessment data, microarray gene expression data and Single Nucleotide Polymorphism (SNP) data. In addition, the CAMDA website provides a reasonable description of the data [1].

CAMDA dataset was collected from people chosen from the population of Wichita,

KS, USA, from December 2002 to July 2003 [43, 51]. Those people participated in a previous surveillance study of CFS between 1997 and 2000. People were classified using two different schemes based on CFS case definition criteria [23]. The intake classification is based on the case definition algorithm as it had been used during the surveillance study. The empirical classification utilizes a standardized clinically empirical algorithm based on quantitative assessment of the major domains of CFS (impairment, fatigue, and accompanying symptoms) [43, 51]. The categories of the empirical classification are:

1. meeting the CFS research case definition (CFS)
2. meeting the CFS research case definition except that medical or psychological exclusions were identified
3. insufficient number of symptoms or less fatigue severity (ISF)
4. ISF with medical or psychological exclusions
5. non-fatigued controls matched to CFS patients in age, race/ethnicity, gender and body mass index (NF)
6. non-fatigued controls with medical or psychological exclusions.

Medical exclusions include active and inadequately treated thyroid disease, neurological disease, inflammatory disease, C-reactive protein levels two to four times normal, severe anemia, uncontrolled diabetes, cardiac disease, renal disease, breast cancer post-treatment and liver disease. Patients with exclusionary psychiatric conditions have major depressive disorder with melancholic, bipolar disorders and alcohol abuse [43].

2.2.1 Clinical data

To begin with, the diagnosis of diseases that still have unclear causes or biological markers depends on the clinical assessment of the patients. Clinical data usually includes demographic characteristics such as gender and age, and symptoms that doctors observe in most of the patients. These observations are collected through medical examinations done by the doctors or through questionnaires filled in by the patients. Another part of clinical data is the complete blood evaluation. It is defined as a basic evaluation of the blood cells including the numbers, concentrations, and conditions of different types of blood cells [38]. Abnormal increases or decreases of blood counts might give indications about underlying disorders that need further examination [11].

The CAMDA clinical data have two parts; each describes clinical information about 227 patients. The first part of the data is a summery of the information which was collected from the patients by filling the Medical Outcomes Survey Short Form-36 (SF-36), the Multidimensional Fatigue Inventory (MFI) and the CDC Symptom Inventory (SI). The SF-36 measures functional impairment in eight different areas, the MFI assesses different kind of fatigue, while the SI measures CFS symptoms [43, 51]. In addition, this part contains demographic information and the two illness classifications (intake and empirical) that we discussed earlier. The second part of the data contains a complete blood evaluation for the patients [26, 51].

2.2.2 Microarray data

DNA is packaged in units that contain a large number of DNA double strands called chromosomes. Each strand of DNA contains thousands of genes which carry instructions of how every cell operates in our body, and each gene contains information on

how to build a specific protein. This process is done through mRNA which travels from the nucleus of a cell to the machines responsible to produce a specific protein. Thus, gene expression is the process of transcribing the genes into mRNA [2].

Recently developed, microarray technology enables scientists to measure the expression level of thousands of genes simultaneously. In other words, it measures the concentration of mRNA, which is the intermediate step in making proteins. Typically, a microarray experiment is performed on a slide that carries a large number of spots, where each spot contains many identical oligonucleotides, complementary-DNA (cDNA) or other strands. This slide is then washed by a sample where some of its content binds to the slide. The slide is then read by different methods to measure the expression levels of thousands of genes reflected in the mRNA they express. Typically, the results of the experiment are large datasets. These datasets are used to find patterns that differentiate between diseased samples and controls. However, the data generated from such experiments introduces many challenging problems such as the high dimensionality and the noise introduced to the collected data during the experiment.

The microarray data provided by CAMDA, consists of 177 text files and 175 image files (with two missing image files). Each pair represents microarray information (20160 spots) about one sample. The microarray technology that was used to collect such data, the MWG 40K microarray, consists of two glass slides (A and B), each with 20,000 features representing human genes. Only gene expression data for the microarray slide A is provided. In the experiment, each sample was hybridized to a microarray slide. Signals were detected using resonance light scattering. This slide was then imaged using a charge-coupled device (CCD). This was followed by

measuring and analyzing the intensities of the spots using an array analysis tool ArrayVision RLS image analysis software [26]. For each sample, values for each spot represent the average of all the pixels in the spot. These intensities are called artifact-removed density values.

2.2.3 SNP data

Recently, significant efforts have been initiated to study genome polymorphisms, which has led to considerable development in human genetics research. SNPs (Single Nucleotide Polymorphisms) are naturally occurring sequence variations in humans, accounting for 90% of human DNA polymorphisms. These variations are single base changes in the DNA at which different sequence alternatives exist [15]. Researchers hope that the knowledge of such variations will assist in identifying genetic markers for some complex diseases. Such genetic markers help in assessing susceptibility to diseases that might have genetic causes and improve the choice of treatment [37]. However, distinguishing between variations that occur in ill people from those occur in healthy people is challenging [53]. SNP data contains 42 SNPs in 10 different genes for 223 people. Six of the genes (COMT, MAOA, MAOB, SLC6A4, TH and TPH2) are involved in the neurotransmission system, and mutations at those genes can lead to mood disorders. The remaining four genes (CRHR1, CRHR2, NR3C1, and POMC) are involved in the neuroendocrine system [25, 35, 47]. Disturbance of such system leads to various psychiatric illnesses such as anxiety, depression, intolerance to stress and sleep disturbance. The measurement of the genetic variations was performed using TaqMan genotyping assay kits [7]. Details of DNA extraction and genotyping method were reported by Smith and colleagues [47].

2.3 Related research

CFS and other unexplained diseases are argued to be multifactorial [39], that is the interaction between different factors stimulates the disease in the individuals. Therefore, collected data to study such complex diseases is usually of different types. Different environmental and genetic factors have different psychological and physiological effects on the patients. Thus, researchers collect data that covers the abnormalities in the human body, starting from DNA polymorphisms and going through protein levels, mRNA levels, blood evaluation and up to the symptoms. This data promises to provide deep understanding of life on the molecular level and may prove useful in medical diagnosis, treatment and drug design. The research in the area of mining integrated biomedical data is limited since the data required for this kind of research is not often available in practise. Hence, the work discussed here is mostly based on the original CDC dataset and part of the data provided by CAMDA. In the following, we discuss some related work to understand CFS by studying each type of data separately and integrating different types of data at different levels.

2.3.1 Analysis using clinical data

Clinical data, including symptoms and laboratory data, is considered to test the hypothesis of heterogeneity of CFS. Vollmer-Conna and colleagues in [52] study 159 female subjects described by filtered clinical data, using Principle Component Analysis (PCA) [48]. The study results in 38 variables that best describe the variability in the dataset. Those variables are then employed in Latent Class Analysis (LCA) [49] to define five classes of CFS patients. Classes vary along different combination of the effect of obesity, relative sleep hypnoea, sleep disturbance, physiological stress,

depression, interoception and menopausal status. The results in this paper support the hypothesis that CFS is heterogenous.

Bassetti and colleagues [8] use binary Support Vector Machine (SVM) [10] on 57 subjects, each described by a row of clinical data (without blood evaluation). The classification results in completely correct classified subjects. Moreover the authors also looked for biological markers for CFS in the blood data. The classification accuracy of 191 patients ranges from 33 to 50% using k-means clustering.

2.3.2 Analysis using microarray data

Using microarray data, Carmel and colleagues [16] analyze 111 female subjects, each described by 15,315 variables. Each variable is the expression level for a particular gene. To reduce the dimensionality, the authors use the first 110 principal components from PCA. For more dimensionality reduction, the Fisher quotient is maximized to find variables for which the classes are compact and well separated. As a result, the gene scores and threshold algorithms identify 32 and 26 genes capable of discriminating five and six LCA solutions, respectively. Moreover, pairwise comparisons using the discriminant genes for the two LCA solutions identified 17 differentially expressed genes. These genes show connections with immune function, transcription, ubiquitination, signal transduction and transporter.

Bassetti's and colleagues analysis [8] of the microarray data using the conventional t-tests and the Significance Analysis of Microarrays (SAM) [18], identifies a set of 30 differentially expressed genes. K-means clustering on the identified genes results in average classification accuracy of 77% on 130 patients. Some of the identified genes show a consistency in terms of proximity in the Gene Ontology (GO) database [24].

Emmert-Streib and colleagues [21] first obtain a classification of the genes according to their biological processes using the GO database. They aim to detect modifications in biological pathways that genes participate in. The authors detect a significant change in group of genes that participate in the biological process of protein amino-acid ADP-ribosylation.

2.3.3 Analysis using SNP data

Goertzel and colleagues [27] apply SVM and a simple enumerative search approach to the SNP data. They use 28 SNPs profiles, belong to 43 CFS patients and 58 controls. The authors reported that the analysis using SVM achieved lower accuracy than enumerative search. In the enumerative search, they identify all sets of one to five SNPs and evaluate each one as a classification rule by comparing it to a threshold. This method achieves a maximum accuracy of 76% where the accuracy distribution spreads over the range of 58% to 76%. The authors conclude that these results indicate that there are genetic markers observed in the SNP data that might help in diagnosis of CFS. Bassetti and colleagues [8] use different clustering and classification techniques to detect any relationship between the SNP data and CFS patients. The best classification accuracy they achieve is approximately 62% using 5-nearest neighbor classification using 133 patients and 17 selected SNPs. The authors imposed different correctness threshold and include their results from analyzing the SNP data in the "not so bad, but not sufficient to derive conclusions" category.

2.3.4 Integration of microarray and clinical data

Whistler and colleagues [55] find approximately four percent of the studied genes correlate with one or more dimensions of fatigue measured in the multidimensional fatigue inventory (MFI), that is 873 gene features. The list of fatigue-associated genes is then translated into functional profiles using the GO or pathway annotations. While 50% of the genes are mapped to GO, 25% only are mapped to pathway annotations. The authors discusses some of the limitations of this study. First, the measurement of the fatigue (MFI) is self-reported and measurements error might occur. Moreover, probe set information provided by the manufacturers of microarrays are no longer consistent with genes in major public databases.

Fang and colleges [22] consider clinical data that corresponds to two questionnaires (MFI and Zung depression scale) to select samples that have the most or least fatigue or depression. From the microarray data associated with these samples, 24 genes are identified from fatigue and depression-based gene expression analysis. PCA could clearly differentiate between the CFS and NF groups using the identified genes. Further analysis of the selected genes revealed a relation between their functions and the biological processes presumed by the clinicians.

2.3.5 Integration of SNP and microarray data

Lee and colleagues [34] assume that SNP and microarray data will compensate for each other by making up for incomplete information. This work is an attempt to detect the effect of SNPs in different genes on the mRNA level of DNA. The authors could detect significant biological interactions between some SNPs and groups of genes using two-way analysis of variance (ANOVA) model approach [42].

2.3.6 Integration of SNP, microarray and clinical data

Lim and colleagues [35] generate prediction scores separately from the gene expression microarray data and the SNP data for patients with CFS or CFS-like symptoms. Scores for selected genes from microarray data are generated by a kernel-based k nearest neighbor (KNN) classifier, whereas those for the SNP data scores were generated using logistic regression models. These scores of the microarray and SNP data are combined with selected symptoms from clinical data in a final logistic model to improve the predictive power. The results show that the integration of relevant clinical and SNPs data may improve the diagnostic classification accuracy for CFS while the integration with microarray data does not have a noticeable improvement. The overall accuracy of the integrated model is 73%.

Presson and colleagues [41] first construct a weighted gene co-expression network and identify its modules, which are sets of tightly correlated genes. In gene co-expression networks, Pearson correlations in mRNA levels are used to define connectivity and to group genes with similar expression profiles into modules. Therefore, characterizing the modules in terms of clinical traits, their gene ontology information, and SNP marker correlations, simultaneously allows the identification of clinically relevant gene modules and the identification of candidate genes that may contribute to CFS.

2.4 Random forests

Random forests is a classifier that uses an ensemble of decision trees. A decision tree consists of nodes where the data attributes are tested. The outcome of this test is

used to split the training set of objects into subsets, each is passed to a corresponding child node. This process is repeated for all the subsets until each node contains a set of objects belonging to one class only. There are number of decision-tree constructing algorithms that adopt a top-down approach. ID3, C4.5 and CART are the most common ones in the machine learning field [32]. These algorithms are supervised learning methods that build a decision tree from a set of labeled objects. The main concern in these algorithms is choosing the splitting attribute at each node in the tree. The challenge here is to find the most beneficial attribute for the classification of the objects at each node. A commonly used criteria for finding the best split are the information gain (utilized in ID3 and C4.5 algorithms) and the gini index (utilized in CART algorithm) [14].

In the testing phase, an object is evaluated and passes down the tree. At each node, one if its attributes is tested, and the result determines which node the object passes to. In binary decision trees, each node has only two branches. A simple example of a binary decision tree to classify objects with three attributes X, Y and Z is given in Figure 2.1. For example, objects with attribute X greater than 1 and attribute Y less than 0 belong to class 1.

The random forests algorithm [12, 13] is a classification algorithm that grows binary decision trees. The training set for growing each tree is a sample of n objects selected randomly, with replacement, from the original N objects. About one-third of the original objects are left out of the selected sample. The Out-Of-Bag (OOB) objects are used to get an unbiased estimate of the test-set error. The error rate is estimated at the end of the run. Take j to be the class that gets most of the votes every time object n is OOB. The OOB error estimate is the number of times class j

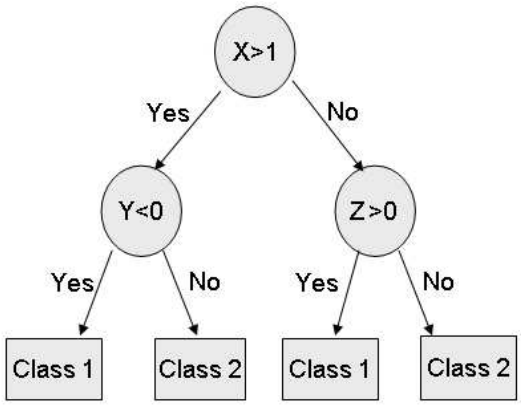


Figure 2.1: A simple binary decision tree

is not equal to the true class of n, averaged over the total number of objects. The best split on m randomly selected attributes out of M input attributes is used to split each node. The splitting criterion employed in the random forests algorithm to select the best split is the gini index, which is defined by

$$gini(D) = 1 - \sum_{i=1}^{k} p_i^2$$

where D is the dataset, k is the number of classes,and p_i is the relative frequency of class i in the dataset D. After splitting on an attribute, say A, the gini index for the two branches is defined as

$$gini(D)_A = \frac{n_1}{n} gini(D_1) + \frac{n_2}{n} gini(D_2)$$

where each branch of the node will have one subset of the dataset, D_1 and D_2, and each subset contains number of objects, n_1 and n_2 respectively. The reduction of impurity from a node to its child nodes is given by

$$\Delta gini(D) = gini(D) - gini(D)_A$$

The best attribute to split on is the one that provides the maximum reduction in impurity $\Delta gini(D)$ [3]. The random forests algorithm is implemented in extended Fortran 77.

2.4.1 Random forests in biomedical research

The random forest classification technique has several features that makes it favorable in the biomedical field and related sciences [30, 50, 56]. In particular, it is a desirable technique in classification problems that involve large datasets such as high throughput microarrays [20, 28, 36]. Among many others features, handling datasets with a larger number of attributes than subjects and returning useful information about the interactions and the importance of the attributes are very attractive features in the biomedical field. However, examples utilizing this technique in the literature are limited comparing to other techniques.

Daz-Uriarte and Alvarez de Andrs [20] investigate the classification of microarray data using random forests. To select the best genes, they iteratively construct random forests. They build a new forest at each iteration after discarding less important attributes. Using simulated and nine real microarray datasets, they show that random forests performs as well as other classification methods such as Diagonal Linear Discriminant Analysis (DLDA), k-nearest neighbor (KNN) and Support Vector Machines (SVM). Moreover, they show that their proposed gene selection method produces small sets of genes that still have good predictive accuracy.

Another example of analyzing microarray data using random forests is the model developed in [36]. In this model, the biological knowledge and random forests are used to build a robust predictive model to treatment response. Using simulated and real

datasets (asparaginase resistant data), the approach proved that irrelevant features to the disease decrease the performance of the classifier, which indicates the importance of prior knowledge of relevant features. Moreover, Pang and colleagues [40] build a similar model based on pathway information and random forests features. As a result of their analysis, they placed random forests within the top two techniques in terms of accuracy.

2.5 Matrix decompositions

Analysis of biomedical data requires mathematical tools that can handle large quantities of data and reduce the complexity of the data to make it more practical. Matrix decompositions are powerful data-mining techniques that are employed to analyze complex data. In the context of biomedical research, matrix decompositions help to analyze data generated from high-throughput medical devises. Each matrix decomposition technique has properties that can be advantageous for analyzing specific type of data. Here, we are interested in the properties of Singular Value Decomposition (SVD) and Independent Component Analysis (ICA). Considering two classes, infected people and controls, SVD helps us to form clusters of people or features in which members of each cluster correlate to each other. ICA helps in capturing and removing noise from data. In the following we define SVD and ICA and briefly discuss their properties. Moreover, we discuss related work to analyze biomedical data using these matrix decomposition techniques.

2.5.1 SVD definition

The singular value decomposition [46] of a matrix A with n rows (objects) and m columns (attributes/features) is given by

$$A = U S V^T$$

If matrix A has r linearly-independent columns, i.e., r is the rank of the matrix A, U is an $n \times r$ matrix, S is an $r \times r$ matrix and V^T is an $r \times m$ matrix. U and V are orthogonal matrices and S is a positive diagonal matrix with non-decreasing singular values. Each row of U gives the coordinates of the corresponding row of A in the new space spanned by new axes represented by the columns of V^T. We call this space, U-space; and the space in which the columns of V^T give the coordinates of the corresponding columns of A, V-space. S entries correspond to the amount of stretching of the new axes, in other words, they represent the importance or the amount of variation captured in each of the new axes. The most variation captured in the data is expressed in the first axis, the second most in the second axis and so on. In the new geometric space, the new axes are orthogonal. This is a very powerful property that distinguishes SVD from other matrix decompositions techniques since the objects can be plotted such that they reflect the relationships between them [46].

SVD captures the most important variations in the data in early dimensions of the transformed space. Thus, one way to remove noise from the data is to consider only few components, say k, out of r components, and reconstruct the matrix A. Moreover, this feature is useful in considering a specific amount of structure in the data. In some settings, important structure can be analyzed more easily when less important structure is removed. On the contrary, sometimes it is useful to choose dimensions that reveal subtle structure. By choosing the first k dimensions, we actually represent

the new r-dimensional space by k-dimensional space that captures as much variation as possible. Different methods are suggested to choose the best k dimensions for best representation [46].

Clustering in the new k-dimensional space makes it easier for us to visualize structure within the dataset. Two objects or attributes are similar if they are close enough from each other, where closeness here is measured by calculating the Euclidean distance between the two objects or attributes. Another way to think about similarity is to consider a vector from the origin to each point in the new space. Vectors in the same direction correspond to correlated objects, vectors in opposite direction correspond to negatively correlated objects and vectors orthogonal to each other are uncorrelated. Moreover, vectors close to the origin correspond to objects correlated or uncorrelated with almost all other objects while vectors far from the origin are those interesting objects that unusually correlate with other objects [46].

Another way to understand the decomposition of the matrix A is to consider the multiplication of each column of U by the corresponding singular value of S and the corresponding row of V^T as a component. Each component represents one of the underlying processes. Capturing the underlying processing is desirable when more than one process are expected to participate in producing the data.

Interpreting SVD

The natural interpretation of SVD is the geometric interpretation of the U-plot and the V-plot. The U-plot shows clusters of objects in a new space spanned by the rows of V, where the rows of U are the coordinates of objects in the new space. From the matrix S, we decide how many axes to keep in order to maintain as much structure as

possible from the data. On the other hand, the V-plot shows clusters of attributes in a new space spanned by the rows of U. The most important variation between objects or different kind of attributes is captured always in the first axis, and the second most important variation in the second axis and so on.

Suppose that we have n objects and m attributes. One way to understand the geometric interpretation of SVD is to consider each data point representing an object as a vector from the origin in m-dimensional space. The direction of these vectors indicates the correlation between objects. SVD locates correlated objects closer to each other and those that correlate in a negative way in the opposite direction. For objects that are largely uncorrelated to each other, SVD locates them orthogonal to each other. However, when the number of dimensions is smaller than the number of uncorrelated objects, SVD pushes them towards the origin. Thus, SVD locates more interesting objects far from the origin, while less interesting ones lie closer to the origin. Objects that are close to the origin are those which have poor correlations with most of other objects or correlate with most other objects. In addition to that, the new space has some properties that make finding clusters more effective. Noise dimensions and less important structure appear at the end of the decomposition. Hence, removing such dimensions and maintaining earlier ones represent the original data in a new, lower dimensional space. One way of locating clusters and interesting objects and attributes is by calculating the Euclidean distance, which is more computationally effective in low-dimensional space.

Visualization

SVD has a unique advantage in visualizing the structure of the data. As we mentioned, SVD captures as much variation as possible in early dimensions. Thus, the truncated U and V matrices at k= 2, 3 enable us to visualize the structure of objects and attributes respectively. In most cases, the discarded part is considered as noise or irrelevant. However, it might sometimes contain useful information about the data. When the analysis requires more dimensions to be visualized, three dimensions at a time can be plotted. The interpretation of the plots in such cases becomes more complicated.

2.5.2 SVD in biomedical research

Alter and colleagues [6] assign the mathematical variables and operations of SVD biological meanings. The work describes the use of SVD in analyzing genome-wide expression data. Researchers use SVD to transform the expression data from the genes×arrays space to a reduced eigengenes×eigenarrays space. In this space, each eigengene is expressed with the corresponding eigenexpression level in the corresponding eigenarray. Using SVD property of capturing the most important variation in the early dimensions, they show that significant eigengenes and the corresponding eigenarrays capture most of the expression information, thus allowing for dimension reduction and estimation of missing data. Interpreting the decomposition in terms of its components suggests the possibility that some of the significant eigengenes (and corresponding eigenarrays) represent independent processes, biological or experimental, which contribute to the overall expression. Moreover, substituting zero instead of a singular value in S and reconstructing the original matrix allows filtering out

anyone of these processes from the data without eliminating genes or arrays. Such filtration may improve any further analysis of the expression data. The geometric interpretation of SVD allows grouping genes of similar regulation and function in the expression of chosen subset of eigengenes (or eigenarrays). The work concludes that SVD provides a useful mathematical framework for processing and modeling genome-wide expression data.

SVD is also employed to winnowing the microarray data [46]. The variation of the expression level might involve part of thousands of genes. Thus, it is preferable to remove genes associated with data that do not help in further analysis. SVD places the objects in the new space such that genes associated with interesting expression patterns tend to be far from the origin. On the other hand, genes with expression patterns such that many or no other genes are similar to them will tend to be close to the origin.

2.5.3 ICA definition

The Independent Component Analysis [46] of a matrix A with n rows (objects) and m columns (attributes/features) is given by

$$A = C\,F$$

where C is $n \times m$ mixing matrix and F is $m \times m$ with m rows of independent components. ICA is based on the assumption that the factors or processes that makeup the dataset are statistically independent. ICA factorization can contain up to one Gaussian component. These properties make this kind of factorization very useful in different applications such as signal processing and other related fields such as biomedical sensing [31].

2.5.4 ICA in biomedical research

One of the most important applications of ICA in biomedical research is removing spatial artifacts from high-throughput biomedical data. As we mentioned earlier, each microarray slide carries a large number of spots that contains identical cDNA or other strands. In some technologies, these spots are printed on the slide in way that creates spatial artifacts. Therefore, the measured intensity depends on the location of the spot it was obtained from. Using ICA, these artifacts can be captured by examining highly non-Gaussian components. The reconstructed data matrix after removing artifacts components from the decomposition, is a new version of the data that contains less spatial structure [46].

2.6 Summary

In this chapter, we gave background information about the Chronic Fatigue Syndrome (CFS), the dataset used in our analysis and its source (CAMDA). We also summarized the related work conducted to investigate the biological markers for CFS. Moreover, random forests and two matrix decomposition techniques, Singular Value Decomposition (SVD) and Independent Component Analysis (ICA), and their applications in the biomedical field were reviewed.

Chapter 3

Experiments

In this chapter we explain the experimental model we use to analyze the CAMDA dataset. First, we discuss the preprocessing procedures that we carried out on each type of the data. Second we discuss different normalization techniques and quality control issues. Last, we present several experiments with different setups. The preprocessing of the dataset is performed using Excel and Java. The plots presented in this work are developed using Matlab.

3.1 Preprocessing

Prior to the analysis, a number of steps were performed on each type of data included in this study. The goal of this data preprocessing is to produce numerical data that is relevant to the analysis and of high quality. Preprocessing data includes translating images into numerical data, excluding unnecessary features, excluding subjects, mapping categorical data to the most appropriate numerical data, dealing with missing data and many other manipulations. Part of the preprocessing on CAMDA dataset

was already performed by the source of the dataset, the CDC Chronic Fatigue Syndrome Research Group. These steps and other steps we carried out are discussed below.

Different data types (clinical, microarray and SNP data) contain data for different numbers of patients. As mentioned earlier, the clinical data contains 227 patients, the SNP data contains 223 patients and the microarray data 177 patients. Obviously, creating a matched set of patients from the three datasets should start from the microarray data which contains the smallest number of patients. Eight out of 177 patients were repeated samples while 10 did not match the patients in other datasets. 159 patients were found to match in the three datasets. We chose to exclude those patients with an insufficient number of CFS symptoms, classified as ISF in the empirical classification. The total number of patients considered in our experiments is 106 patients, 58 CFS cases and 48 controls. We identified 31 patients classified as CFS or NF with medical and psychiatric exclusions. We also consider another set of 75 patients, 38 CFS cases and 37 controls, in which we exclude those patients with medical or psychological exclusions. We conduct the same experiments and analyze the results for both groups of patients.

3.1.1 Clinical data

The first part of the clinical data which includes the classification attributes, the summary scores from SF-36 and the MFI, the CDC SI and other demographic information was preprocessed as follows. First, the classification attributes which include the intake and empirical classification and others, were removed. All attributes that measure the severity and the frequency of the symptoms of CDC SI were excluded

since they contain many missing values. Other missing values were replaced with their column mean.

Other redundant attributes that were also removed from the dataset include years of illness, gender, age, date of birth, race, ethnic groups, BMI, exclusions and MDDm. The Demographic features are considered unnecessary since the initial design of the CDC study matched cases and controls based on them [51]. The attributes that describe the medical and psychiatric features were removed since we examine the effect of these features by having two sets of patients, one set includes only those patients with no exclusions. All attributes that describe the symptoms in CDC SI are categorical with two values, yes and no. Those attributes were mapped to numerical representations using binary mapping. The final number of attributes in the first part of the clinical data is 32. The second part of the clinical data, which is the complete blood evaluation, was preprocessed as follows. The blood collection date and the collection time attributes were removed. All of the attributes were associated with alert flags to indicate low or high blood counts. Alerts flags were considered redundant and eliminated from the data. Two attributes that describe the percentage and the number of one category of the while blood cells, "% basophils" and "# basophils", were also excluded because they contain many "NP" values which was not explained in any of the related documents. A few other "NP" values were found throughout the data and replaced with their column mean. A few "<" symbols were also found and removed. The final number of blood features is 32. To distinguish between the two parts of the clinical data throughout the rest of this study, we refer to the first part that includes the impairment scores, fatigues scores and accompanying symptoms as clinical data, while we refer to the second part which includes the blood counts as blood data.

3.1.2 Microarray data

Artifact-removed density values had been already preprocessed by removing pixels with density values that exceed four median absolute deviations (MADs) above the median. By doing so, the effect of the image artifacts (e.g. dust particles) on density estimation is removed [26]. The background densities had been also subtracted. Out of 20160 spots, 460 spots were labeled as "blank", "mwgaracontrol", "mwghuman" and "mwghumcontrol". We removed these control spots and extracted the densities of the remaining 19700 spots for each patient.

3.1.3 SNP data

The three categories of 42 SNPs were mapped as "allele 1" to 1, "allele 2" to -1, "both" to 0 and the "Undetermined" values were replaced with their column mean.

3.2 Normalization techniques

Since SVD is based on numerical computations, an appropriate normalization technique has to be applied. Normalization includes centering and scaling the values of attributes.

In this study, we deal with both issues, centering and scaling the data, to make the magnitudes of the entries more appropriate for analysis. First, in the case when all attributes in the data have positive values, the first component of the decomposition will capture the most trivial variation by having the first axis connecting the origin to the center of the data. This will create a problem since SVD has the property of orthogonal axes in the new space. Thus, the second axis, which is now orthogonal

to the first, will point in a distorted direction. This problem can be solved by zero centering, where the mean of each column is subtracted from each column entry. Hence, the data is centered at the origin and the axes point to the directions which reflect the correct variation in the data.

The second issue is scaling attributes that have different units, that is, each attribute describes different measurements in the real world. If all the attributes are of the same importance, then the best way to fix this problem is to scale the data into approximately the same range. This can be done by dividing the entries in each attribute by the standard deviation. Values that have been zero-centered and divided by standard deviation are called z-scores. Regardless of what the original values are, when we use the z-transformation, we obtain a standard measurement of the distance from the mean.

Another scaling problem is when non-scaled data have a non-linear significance associated with magnitudes, i.e., large values do not indicate large importance. This results in giving more weights to attributes that have larger values on average even though smaller values may contain an equivalent amount of information. To overcome this issue, we first take the logarithm for each entry, followed by the transformation to z-scores to deal with the non-linear significance of values and to make the values as comparable as possible.

One of the advantages of the random forest technique is that data normalization is not required. This is due to the fact that the choosing the best split on each node depends on the splitting criterion which depends only on the distribution of the values of a single attribute.

3.3 Quality control

There are many factors that could affect the quality of the CAMDA data. In particular, microarray data can be affected by the experimental design, experimental setup and image analysis that might cause misleading variability in the actual intensities. To ensure a solid basis for the validity of the results obtained from our experiments, we attempt to remove outlying objects from each type of the data and clean spatial artifacts from the microarray data using ICA.

3.3.1 Outlier detection

Both techniques, SVD and random forests, have the property of detecting outliers. Objects that correlate with the rest of objects in a very unusual way, are expected to lie somewhere far from other groups of objects in the decomposition. These outlying objects can be detected by looking at the U- and V-plots. This process has to be done carefully by looking at the plots from different angles. Although the random forests algorithm is relatively robust with respect to outliers and noise [12], it has a built-in feature to estimate outliers. Breiman and Cutler [13] define an outlier in class j as an object whose proximity to all other class j objects is small. The following algorithm is used: first, compute the average proximity from object n in class j to the rest of the training objects in class j as the sum of the squares of proximities for all other objects in the same class as n. Second, the inverse is taken so that the value is large when the average proximities is small. Finally, this value is standardized using the median and standard deviation. Objects with large outlier measurements (often higher than 10) are the potential outliers. Throughout our analysis, we use this feature on every type of data to remove outliers.

3.3.2 ICA

SVD can isolate the noise in the last components of the decomposition. However, taking a step to clean the data might improve the detection of important structure by SVD. We use the FastICA algorithm which is available from `www.cis.hut.fi\projects\ica\fastica`. We choose to examine 10 independent components. By looking at the plots of the values of each component, we noticed that some components have a repetitive structure along the values while others have different structure at the beginning or the end of the component. This repetitive structure indicates measurement noise introduced to the collected data during the experiment. Some environmental or experimental factors may vary among different regions of the microarray. For example, manufacturing defects and scratches in the microarray slide might be one source of these spatial artifacts. Another example is washing the microarray slide edges with samples less carefully which may contribute to the different structure detected near the edges [44]. We decide to remove two sharp artifacts components in which a clear repetitive structure can be observed along the values of the components.

3.4 Experimental model

In order to study the effectiveness of our proposed model, we develop several experiments with different setups using SVD and random forests. We start with basic analysis using SVD and random forests algorithms and then we develop the analysis by selecting important attributes in order to improve the classification rate.

3.4.1 Data analysis using SVD

Our first attempt to analyze the CAMDA dataset is to use SVD to look for clusters of patients in the three types of data. Our goal is to be able to compare and analyze the behavior of SVD on each type of data separately and on the integrated dataset. We hypothesize that, by adding the three pieces of information related to the disease, we can gain stronger indications about the biological causes of CFS.

Our analysis depends on the geometric interpretation of the U-plot and the V-plot of SVD. Mainly, we are looking for clusters of CFS patients and controls. In addition, we are looking for attributes that correlate with other attributes in an unusual way. From the SVD point of view, we are looking for attributes that are far from the origin but close to other relevant attributes. Table 3.1 shows an example of a small 6-attribute dataset that we generate to demonstrate the geometric interpretation of the U- and V-plots of SVD. Figure 3.1 is the U-plot for the given data. As shown, SVD can detect the variation between the three groups of objects without any knowledge of their classes. Also, the V-plot in Figure 3.2 demonstrates the negative correlation between attributes 1 and 2.

Objects	Attrib 1	Attrib 2	Attrib 3	Attrib 4	Attrib 5	Attrib 6	Clss
1	2	5	7	9	10	22	1
3	3	6	8	10	12	23	1
2	5	3	2	1	-3	-5	2
4	6	4	3	2	0.5	-3	2
5	5	6	5	6	5	6	3
6	2	3	3	2	3	4	3
7	5	7	5	6	5	7	3

Table 3.1: A small dataset

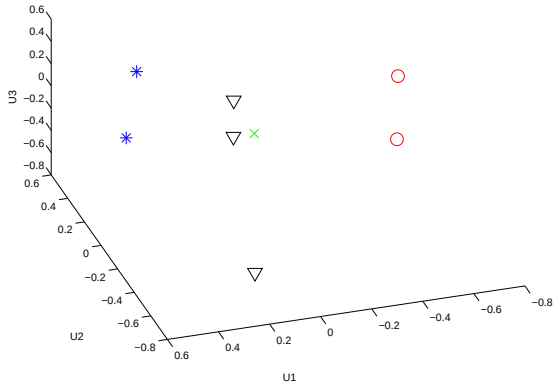


Figure 3.1: 3-dimensional plot of rows of the U matrix from the small dataset. (blue= class 1, red= class 2, black= class 3, green= origin)

SVD analysis of the blood data

Prior to analysis, the blood data is normalized by transforming data into z-scores since the measurement units are different. Then, we apply SVD on the normalized 106×32 and 75×32 data matrices.

SVD analysis of the microarray data

To overcome the emphasis of large measurements, we log-transform the entries. However, Wentzell and colleagues [54] argue that the logarithmic transformation might not be effective in many datasets. Besides that, there is no clear discussion in the literature about the effect of logarithmic transformation on different data distributions. Therefore, we apply SVD on the raw and logarithmic transformed data matrices. Although all attributes in microarray data measure the expression level of genes, values for some genes might have wider ranges than other genes. These genes dominate the

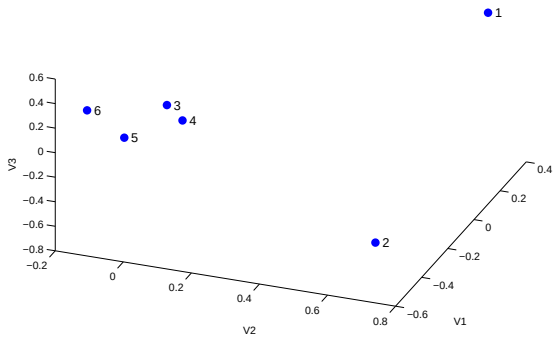


Figure 3.2: 3-dimensional plot of columns of the V^T matrix from the small dataset.

ones with narrower ranges. In order to make the values more comparable, we normalize the microarray data matrix by transforming the entries to z-scores. Now, the expression level is represented as the number of standard deviations above or below the mean of the gene expressions across all objects.

We also apply SVD on the microarray data after cleaning it by removing what we believe are the worst two components of spatial artifacts. Each microarray matrix has two versions with sizes 106×19700 and 75×19700.

SVD analysis of the SNP data

The three categories of SNP data are mapped so that no normalization is required prior to analysis. Thus, we apply SVD directly to the 106×42 and 75×42 data matrices.

SVD analysis of the clinical data

The purpose of this analysis is to confirm the fact that the empirical classification of CFS is based on the assessment of the major domains of CFS, impairment, fatigue, and accompanying symptoms. We first normalize the data by z-scoring since the original data consists of numerical and categorial attributes. Then, we apply SVD to the normalized 159×32 (including ISF patients), 106×32 and 75×32 data matrices.

SVD analysis of the integrated data

Data integration does not seem to be effective at this stage of the analysis because the number of attributes in the three types of data is significantly different. While the microarray data has 19700 attributes, the SNP data has 42 attributes and the blood data has only 32 attributes. Thus, the correlations of microarray data attributes dominate other attribute correlations and the plot of CFS patients and controls is expected to look much the same as the resulting plot from the microarray data.

3.4.2 Data analysis using the random forests algorithm

The analysis in this section takes a form of a supervised classification problem. In order for the analysis to be reliable, the experiments are conducted over a wide range of parameters.

We utilize the random forests algorithm [12, 13] to classify patients in the CAMDA dataset into two classes, CFS patients and controls. Since the number of patients is already small, we use OOB (out-of-bag) data to get an unbiased estimate of the classification error to test the performance of our model without the need to split the dataset into training and testing data. In each tree, about 35 of the original number

of patients and controls are chosen randomly and left out of the selected sample for testing.

To achieve the best performance for the random forests algorithm, we try different settings for two parameters; *mtry*, the number of input variables to try at each split and *nt*, the number of trees to grow for each forest. In this part of the study, we perform our experiment on a wide range of these parameters in order to evaluate the effect of changing them on the classification performance. For the results to be as comparable as possible, we fix the random seed so that the m randomly selected attributes out of M input attributes is the same for all the runs. The default value for *mtry* is $\sqrt{\text{number of attributes}}$. We multiply the default value by the factors 0.25, 0.5, 1, 2, 3, and 4 to get new values of *mtry*. Throughout our analysis, we refer to these multiplicative factors of the default value of *mtry* as *mtry factors*. We also try different values of *nt* depending on the number of attributes in different datasets. Fewer attributes need fewer trees.

Random forests analysis of the blood data

Since the number of blood attributes is only 32, we choose to examine the performance of random forests using *nt* of 2000, 1000, 500, 250, 200, 100 and 50. For each setting of *nt*, we perform these experiments using $mtryfactor$ of 0.25, 0.5, 1, 2, 3, and 4.

Random forests analysis of the microarray data

In this experiment, we set *nt* to 10000, 7000, 5000, 2000, 1000 and 500. For each experiment, we use $mtryfactor$ of 0.25, 0.5, 1, 2, 3, and 4. We also perform the same experiment with the same settings on the data matrix after cleaning with ICA.

Random forests analysis of the SNP data

Since the number of SNP data attributes is close to those of blood data, we perform our experiments using similar settings. We choose to examine the performance of random forests using nt of 2000, 1000, 500, 250, 200, 100 and 50. For each value of nt, we perform the same experiment using all the values of $mtryfactor$.

Random forests analysis of clinical data

Studying the classification performance of random forests on the non-biological clinical data is not one of our main goals. However, examining and comparing the performance of the classifier on each of the three parts of clinical data describing impairment, fatigue and symptoms might be interesting in evaluating the contribution of each part to the empirical classification. We perform our experiment using the default value of $mtry$ and nt of 500.

Random forests analysis of integrated data

As we mentioned earlier, integrating different data types when they differ in their number of attributes significantly, is not efficient. Thus, the performance of the random forests algorithm on the integrated dataset is expected to achieve a maximum accuracy as good as the accuracy obtained from the dominating type of data, which is the microarray data.

3.4.3 Attribute selection

Attribute selection is the process of discovering and discarding the irrelevant features [29]. Selecting the most relevant features to the analysis is one way of reducing

the dimensionality of high-throughput biomedical data describing a disease. In this section, we examine the performance of SVD using different number of selected attributes from each type of the data. We also continue to study the effect of different settings of *mtry factor* and *nt* on the classification rate of random forsts when we select different numbers of attributes. We identify a small set of features from each type of data that represents the most relevant features to CFS. We integrate these features using two techniques, voting and combination of different types of the data in one matrix.

Attribute selection using SVD

As we noted before, two kinds of objects lie close to the origin, those which correlate with many of the other objects and those which correlate with almost none of the other objects. More interestingly, the same applies to attributes. This property of SVD can be utilized to get rid of uninteresting blood features, genes and SNPs. Moreover, ranking the attributes according to their distance from the origin might provide us with an intuition about the importance of each attribute in terms of its relevance to the prediction of the disease.

SVD analysis of the selected blood features and SNPs There is no clear way to decide how many features we should keep to achieve the best possible clustering. One way is to keep all those features whose distance from the origin is greater than the mean distance. Because the number of blood and SNP attributes are relatively small, another way is to first rank attributes based on their distance from the origin and then remove one by one from the bottom of the list.

SVD analysis of the selected genes For the microarray data, the large number of attributes in addition to the three versions of the data matrices, the original, the log-transformed and the ICA-cleaned z-scores, make the process of finding the optimal number of selected attributes harder. We choose to look at the U-plots of the farthest 500 genes from the origin. Since SVD has the property of capturing the most important structure in early dimensions, we expect that smaller groups of genes, removing those which lie at the bottom of the ranked list, do not change the U-plots significantly.

Attribute selection using random forests

It has been reported earlier that irrelevant features decrease the performance of random forests [36]. The random forests algorithm has two attributes-importance measurements that are built in. The first measurement is based on gini index that we have discussed earlier. The gini importance detects the improvement of the splitting criterion associated with each attribute. By adding up the gini decrements for each attribute over all the trees in the forest, we get a sensible estimate of attribute importance. However, Breiman and Cutler [12, 13] recommend a more complex measurement, the permutation importance measure. The underlying principle of this method is permutation. In every tree grown in the forest, the number of votes for the correct class using the OOB objects is counted. To measure the importance of each attribute, the values of the attributes in the OOB objects are permuted randomly and used again in the tree. The difference between the number of votes for the correct class in the permuted OOB data and that for the correct class in the original OOB data averaged over all the trees in the forest is the raw importance score for each

attribute. The raw scores are then divided by the standard deviation to get z-scores. Estimating the attribute importance using this method can be used to perform a second run using the most important selected attributes only. To overcome the effect of the irrelevant attributes in our analysis, we use the more sophisticated measurement, the permutation importance measure.

Random forests analysis of the selected blood features To select the most important blood features that differentiate between our two classes, we perform our experiments using different combination of $mtryfactor$ (0.25, 0.5, 1, 2, 3, 4) and nt (250, 100, 50) and selecting different numbers of the most important attributes (25, 20, 18, 16, 14, 12, 10, 7, 5, 3). We also use SVD to evaluate the selected features.

Random forests analysis of the selected genes In this experiment we have two versions of the microarray data, the original raw data and the cleaned data using ICA. Each have two sets of patients, 106 and 75. We use different combination of $mtryfactor$ (1, 2) and nt (5000, 2000) and selecting different numbers of the most important attributes (1000, 500, 200, 100, 50, 45, 40, 35, 30, 25, 20, 15). Then, we use SVD to evaluate the significantly differentially expressed genes.

Random forests analysis of the selected SNPs Similar to the blood data experimental settings, we perform our experiments using different combination of $mtryfactor$ (0.25, 0.5, 1, 2, 3, 4) and nt (250, 100, 50) and selecting different numbers of the most important attributes (30, 25, 20, 18, 16, 14, 12, 10, 5, 3). Again, we use SVD to evaluate the selected SNPs.

Analysis of the integration:

Our goal from this experiment is to improve the diagnosis of CFS by combining different types of data that describe different aspects of CFS. We use two techniques to integrate the classification power of different types of data, voting and combination.

Voting is the process of combining the classification results of multiple classifiers to generate one overall result [2]. In this experiment, each random forest classifier is trained using the most important attributes of the same patients from different type of data. At testing, each patient will have three votes, each vote comes from one type of data. The result of the voting is the class that obtains at least two votes.

The second technique is to combine the most important attributes from each type of the data in one matrix. In other words, the new matrix consists of the attributes that contribute the most to the discrimination between the two classes.

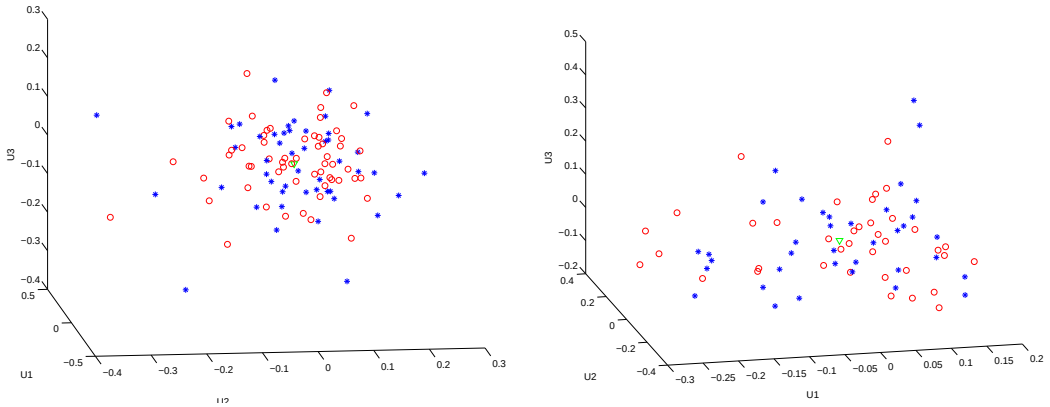
3.5 Summary

In this chapter we illustrated the experimental model we use to test the performance of SVD and random forests using the CAMDA dataset. We also discussed the preprocessing procedures conducted to prepare the data for analysis. We presented several experiments with different setups to examine the performance of random forests over a wide range of parameters. Finally, we explained the experiments conducted on the integrated dataset. In the next chapter, we will show and discuss the results of the analysis.

Chapter 4

Results and analysis

In this chapter, we analyze the performance of our experimental model using SVD and random forest techniques on the CAMDA dataset. In the first section, we show the results of clustering analysis using SVD on each type of the data of two different sets of patients; the full set of 106 patients, and a subset of 75 patients with no medical or psychiatric exclusions. The results of the classification analysis using random forests on each type of data are shown in the second section. In the third section, we present the results of utilizing SVD and the permutation importance measurement of the random forests algorithm to select the most relevant features to CFS. We also show the results of random forests analysis for the integrated data. We discuss the results obtained from different experiments in Section 4. Overall, the results support the previous outcomes in the literature which conclude that there is a biological basis to the etiology of CFS.



(a) 106 patients (b) 75 patients

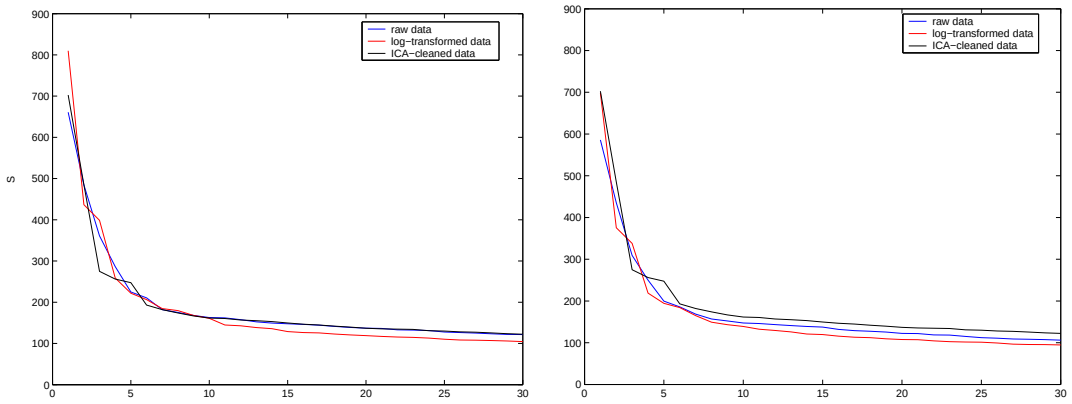
Figure 4.1: 3-dimensional plots of rows of the U matrix from the blood data (red = CFS, blue = NF, green = origin)

4.1 SVD analysis results

To analyze the results of applying SVD to each type of data, we show the plots of the entries of the matrix U truncated to three dimensions to provide the best possible visualization of the structure of the data. In the 3-dimensional plots of rows of the U matrix, we look for clusters of patients that share the target classes, CFS patients and controls.

4.1.1 SVD analysis results for the blood data

The blood data does not seem to have any strong indications for clustering of the two classes as shown in Figure 4.1. However, Figure 4.1(a) shows small overlapping clusters, while Figure 4.1(b) shows more CFS patients concentrated to one side of the origin. In spite of the fact that previous research finds blood evaluation attributes useless in CFS diagnosis, we include this data in subsequent analysis hoping to catch any indications that could possibly help to assign a patient to a particular class.



(a) 106 patients (b) 75 patients

Figure 4.2: Scree plots of the singular values from the microarray data

4.1.2 SVD analysis results for the microarray data

The singular values in the matrix S from SVD indicate the relative importance of each new axis. To capture as much structure as possible, we look for a suitable cutoff in the scree plot of the S matrix. Figure 4.2 shows the first 50 singular value scree plots from 106-patient and 75-patient data matrices. The best cutoff in all the scree plots is 3 to 5. The sharp knee at 3 in ICA-cleaned data matrices suggests that the forth and the fifth dimensions might contain as much important structure as the third dimension. Truncating U and V matrices after three dimensions provide a good estimate of the correlation between objects and attributes and has the advantage that the human can visualize the position of the points. All that we need is to decide on how many dimensions are needed to capture as much structure as possible.

Figure 4.3 shows the 3-dimensional plots of rows of the U for our six data matrices. In Figure 4.3(a), although clusters of patients and controls are not very clear, it can be noticed that some CFS patients form a more compact cluster downwards. The cluster of NF controls is less noticeable and less compact upwards and a small CFS

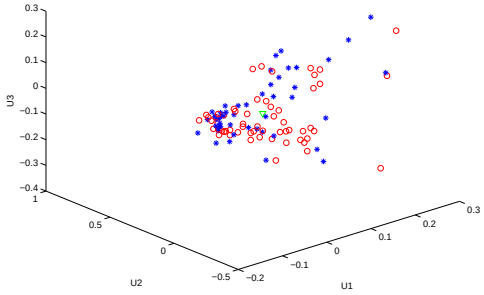
patient cluster interferes with it. In Figure 4.3(c), the "v-shape" of the data points distribution is no longer clear. However, the compact cluster of some CFS patients is still clear in the upper side of the plot and a small cluster of controls below it. In the ICA case, the main strip is a CFS-patient cluster surrounded by small clusters of controls.

A closer look at the U-plots for the original data and the log-transformed data contrasting the 106 patients with the 75 patients, shows that many of the removed patients with medical or psychological exclusions (most of them classified as CFS) appear to lie in the middle area between the potential clusters of CFS and NF. Therefore, further analysis on the 75 patients might lead to a better separation for the CFS patients and controls. In each of the above experiments, we removed a single CFS outlying patient that was detected in all the plots.

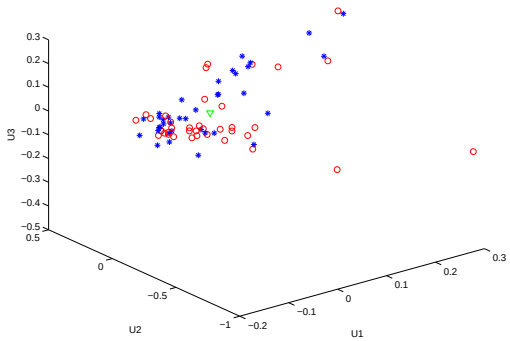
Plotting the first dimensions of V-plot does not seem to be helpful since the number of genes is large. Thus, observing useful structure requires filtering the genes first. In previous research, the biological knowledge about the genes and their functions have been assumed such that only genes which proved to have a relation to CFS or any of its symptoms are considered in the analysis. In our analysis, we assume no previous biological knowledge about CFS. In Section 4.3, we use the random forests importance measurement to identify significantly differentially expressed genes.

4.1.3 SVD analysis results for the SNP data

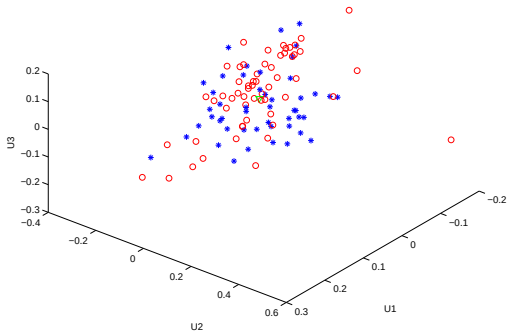
The analysis of SNP data, as shown in Figure 4.4, does not indicate any strong correlation that can differentiate between the objects of the two classes. However, taking a closer look at Figure 4.4(a), one might recognize small 3- to 4-patient groups



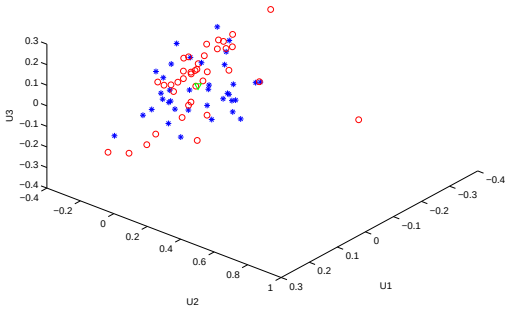
(a) 106 patients data matrix



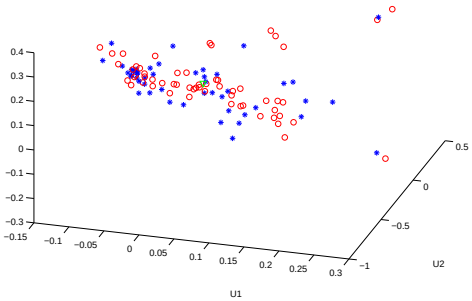
(b) 75 patients data matrix



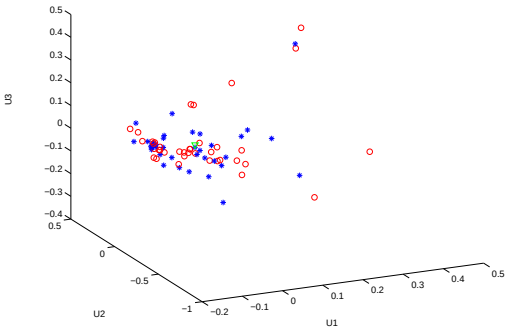
(c) 106 patients log-transformed data matrix



(d) 75 patients log-transformed data matrix

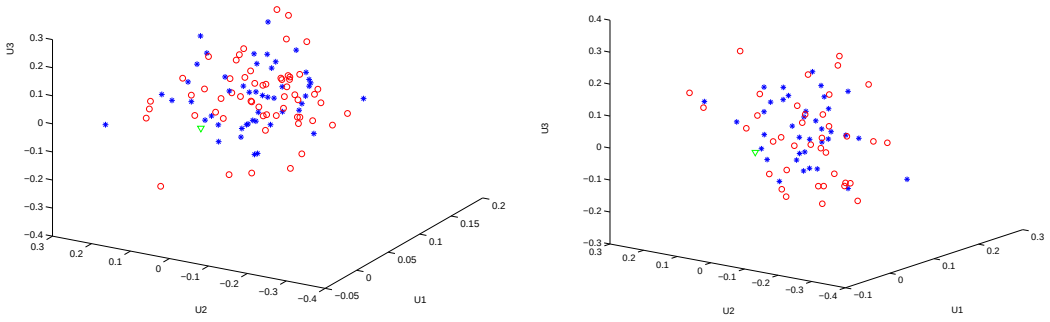


(e) 106 patients data matrix after cleaning with ICA



(f) 75 patients data matrix after cleaning with ICA

Figure 4.3: 3-dimensional plots of rows of the U matrix from the microarray data (red = CFS, blue = NF, green = origin)



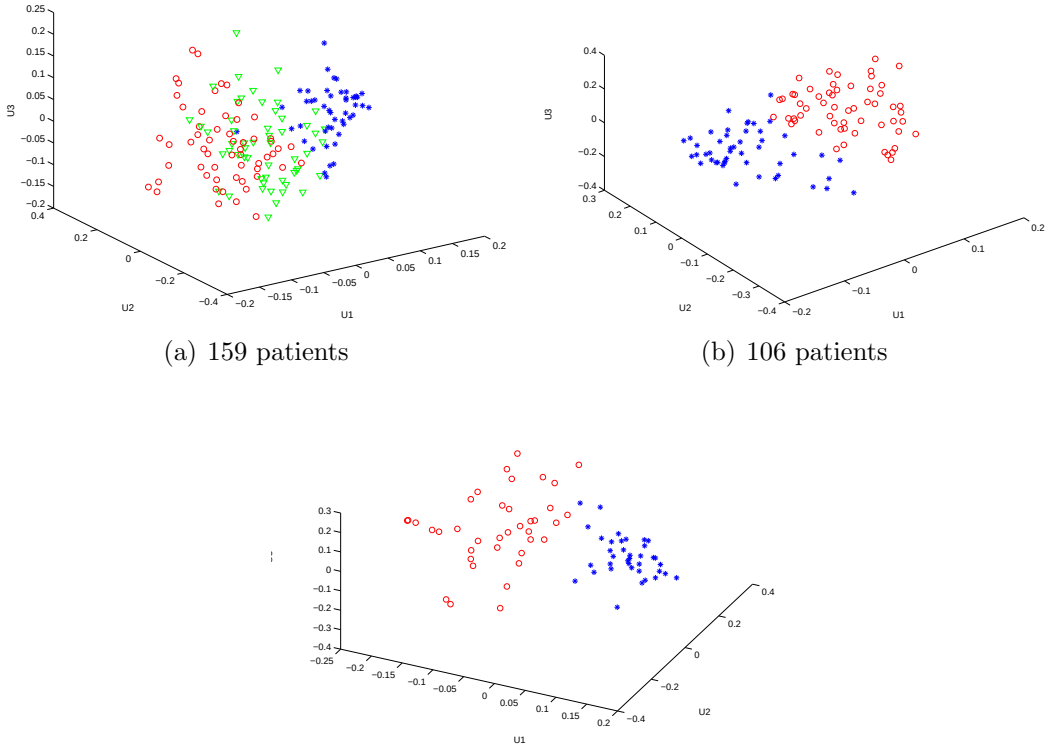
(a) 106 patients (b) 75 patients

Figure 4.4: 3-dimensional plots of rows of the U matrix from the SNP data (red = CFS, blue = NF, green = origin)

scattered in the 3-dimensional space of the U-plot. For this reason and the previously reported results in the literature, as discussed in Sections 2.3.3 and 2.3.5, we decided to include the SNP data in further analysis.

4.1.4 SVD analysis results for the clinical data

Figure 4.5 shows clear clusters of people classified empirically as CFS, ISF and NF. We first consider the original 159 patients, with ISF patients included. Figure 4.5(a) shows that the ISF cluster lies exactly between the CFS and NF clusters. This figure supports the exclusion of ISF patients from further analysis in order to have a clearer separation of the two classes (CFS vs. NF), as illustrated in Figures 4.5(b) and 4.5(c). It is also interesting to look at the 3-dimensional plot of columns of the V matrix. Figure 4.6 shows that SVD is capable of capturing the correlations between attributes that describe the three major domains of CFS.



(a) 159 patients

(b) 106 patients

(c) 75 patients

Figure 4.5: 3-dimensional plots of rows of the U matrix from the clinical data (red = CFS, blue = NF, green = ISF)

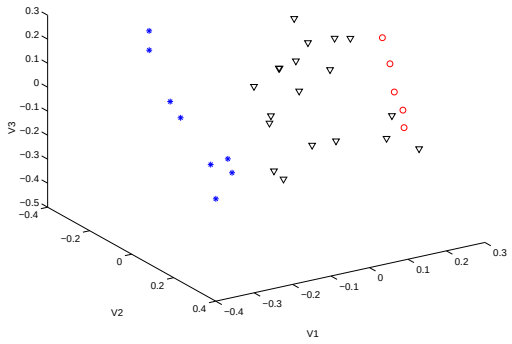
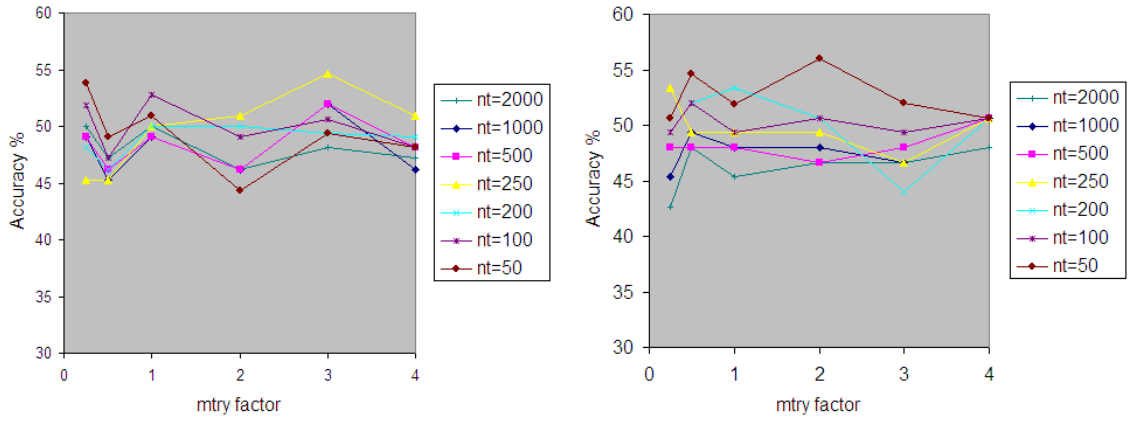


Figure 4.6: 3-dimensional plot of columns of the V matrix from the clinical data (blue = impairment, red = fatigue, black = accompanying symptoms)

4.2 Random forests classification results

To further investigate the discriminative power of the different types of data, we utilize the random forests algorithm using a wide range of the parameters *mtry* and *nt* on each type of data as described in Section 3.4.2. The main purpose of these experiments is to evaluate the effect of different settings of *mtry* and *nt* on random forests performance.

In the following sections, we evaluate changing the parameters and report the classification accuracy for each experiment. Using the full set of 106 patients, we have 58 CFS patients and 48 controls. Thus, a base accuracy of approximately 55% results from always guessing the larger class. The subset of 75 patients consists of almost the same number of CFS patients and controls in each class so that the base accuracy is roughly 50%.



(a) 106 patients

(b) 75 patients

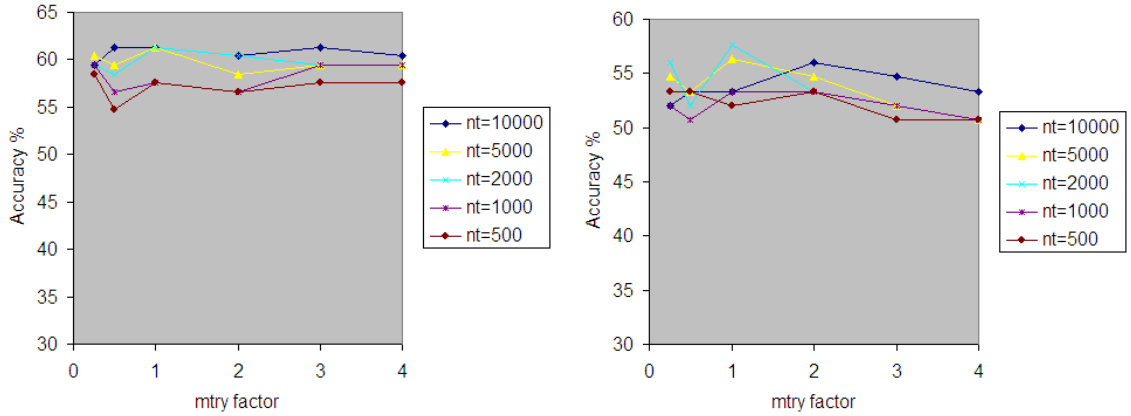
Figure 4.7: Accuracy vs. *mtry factor* of the blood data using different *nt* values

4.2.1 Random forests analysis results for the blood data

For both 106 and 75-patient data matrices, Figures 4.7(a) and 4.7(b) show that *mtry factor* and *nt* do not have a noticeable effect on the classification rate. The best achieved accuracy using 106 patients is 54.67% (using *mtry factor*=3 and *nt*=250), which is approximately same as the base accuracy. The best achieved accuracy using 75 patients is 56% (using *mtry factor*=2 and *nt*=50), which is higher than the base accuracy.

4.2.2 Random forests analysis results for the microarray data

Our results from analyzing the microarray data show that a biological basis of CFS etiology is detectable in gene expression levels. The best accuracy achieved using the full set and the subset of 75 patients are 62.26% (using *mtry factor*=1 and *nt*=7000) and 57.67% (using *mtry factor*=1 and *nt*=2000), respectively. Both results are around 7% above the base accuracy. These and other results using different combinations of



(a) 106 patients (b) 75 patients

Figure 4.8: Accuracy vs. *mtry factor* of the microarray data using different *nt* values

mtry factor and *nt* are shown in Figures 4.8(a) and 4.8(b). In some cases, random forests perform better when using larger *nt* values (*nt* ranges between 2000 and 10000). Increasing *mtry factor* keeps the results stable with no significant improvement on the overall performance. Throughout the rest of our analysis, we use *nt* values of 2000 and 5000 since further increase have insignificant effects on performance and noticeably increase the execution time. We use *mtry factor* of 1 and 2 since the best accuracies are mostly found within these values.

Table 4.1 shows the average accuracy of performing the same experiments using the same settings on the log-transformed and ICA-cleaned data. For each of these experiments, we obtain the average accuracy of 10 runs. The accuracy of the results is described in terms of 95% confidence intervals placed on these results. We notice that the results lie in close intervals and there is no significant improvement using the two versions of the microarray data. While we do not see any strong reasons to consider the log-transformed data in subsequent random forests analysis, we believe that the

Number of patients	*nt*	*mtry factor*	Original data	Log-transformed data	ICA-cleaned data
106	5000	2	59.91±0.84	59.44±1.05	58.49±0.91
		1	61.32±0.55	60.38±0.99	58.49±0.66
	2000	2	60.38±1.66	59.33±0.64	61.33±0.84
		1	61.32±0.75	61.32±0.52	58.49±2.3
75	5000	2	50.93±1.05	54.67±0.76	54.8±1.06
		1	56.33±0.75	52.00±0.99	54.67±0.84
	2000	2	53.34±0.41	53.34±0.62	52.00±0.41
		1	57.67±0.64	56.33±1.31	54.67±0.96

Table 4.1: Random forests accuracy(%) of the original, the log-transformed and the ICA-cleaned microarray data

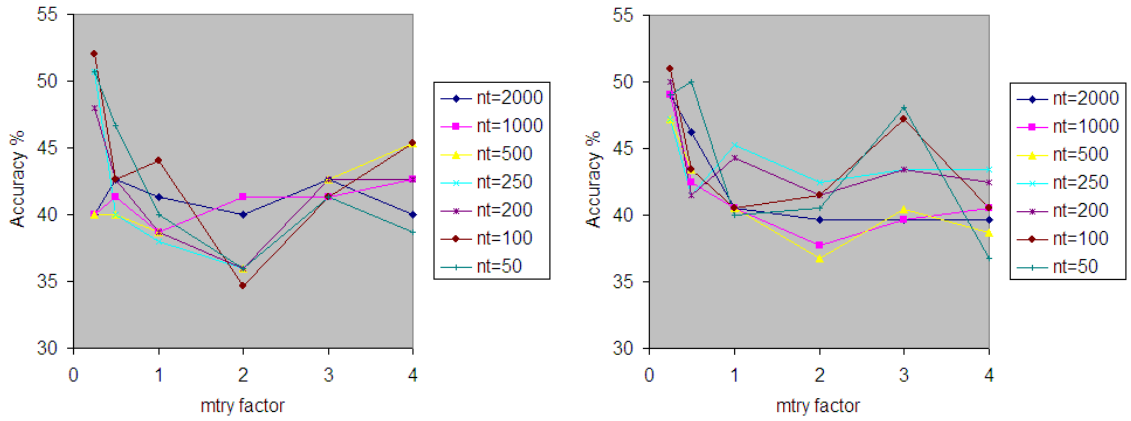
ICA-cleaned data is more reliable and might perform better in further analysis.

4.2.3 Random forests analysis results for the SNP data

Figures 4.9(a) and 4.9(b) for both data matrices show no clear effect for *mtry factor* and *nt* on the classification rate. While, in many cases, using small values of *mtry factor* leads to higher accuracies, it is hard to decide what the best settings for *nt* is. The best accuracies are 50.94% and 52% for 106 and 75 patients (using *mtry factor*=0.25 and *nt*=100), respectively. Both results are roughly equal to the base accuracy.

4.2.4 Random forests analysis results for the clinical data

Using the clinical data, random forests achieves an accuracy of 100% for both 106-patient and 75-patient data matrices. Table 4.2 shows the accuracies obtained from the attributes that summarize each dimension of CFS separately. These accuracies



(a) 106 patients (b) 75 patients

Figure 4.9: Accuracy vs. *mtry factor* of the SNP data using different *nt* values

	Impairment	Fatigue	Symptoms
106 patients	94.34	100	89.62
75 patients	97.33	98.67	88

Table 4.2: Random forests accuracy(%) using attributes that describe each dimension of CFS in the clinical data

are obtained using the default value of *mtry factor* and *nt* of 500.

4.3 Attribute selection

Our previous results suggest the need for attribute-selection techniques to improve the clustering of SVD and the classification performance of random forests. It is also more plausible to integrate the data after selecting the most distinguishable attributes from each type of the data. By doing so, we discard big amount of irrelevant information from the microarray data and the number of attributes of each type of the data

becomes roughly the same. In this section we make use of two attribute-selection techniques, SVD and the permutation importance measurement of the random forests algorithm. We employ both techniques to select the most relevant attributes to CFS.

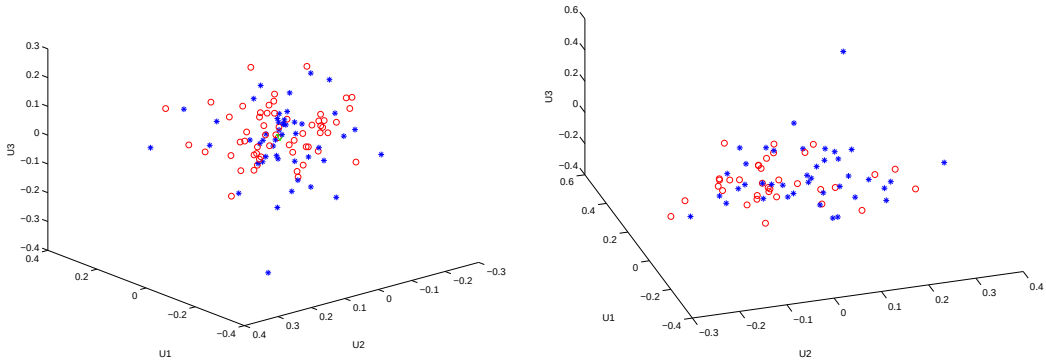
4.3.1 Attribute selection using SVD

SVD analysis results of the selected blood features

Considering those attributes whose distance from the origin is greater than the mean distance did not show notable clustering. Thus, we rank the attributes based on their distance from the origin and then remove one by one from the bottom of the list. Figures 4.10(a) and 4.10(b) show the results of applying SVD to the top 10 blood features of the ranked list. For 106 patients, a cluster of NF controls can be detected between two clusters of CFS patients. The 75 patients have different structure in which a CFS-patient cluster can be observed next to a NF-control cluster. Table 4.3 shows the lists of most separable blood features ordered according to their distance from the origin. We notice that the 106-patient and 75-patient lists contain the same blood features in different order.

SVD analysis results of the selected genes

In the following, we show the results of applying SVD to the top 500 genes of the ranked list. Figure 4.11 shows that for the original and log-transformed data, a slight improvement from the results before attribute selection can be noticed (see Figure 4.3 for comparison). We include lists of the top 50 genes of the ranked list in Appendix A for the sake of comparison with other lists of genes.

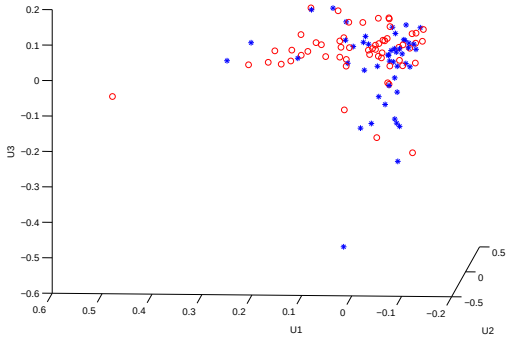


(a) 106 patients (b) 75 patients

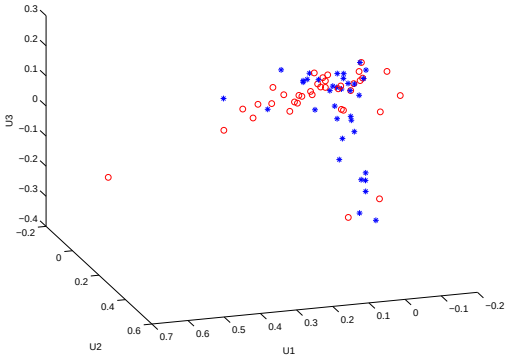
Figure 4.10: 3-dimensional plots of rows of the U matrix from the top SVD-ranked blood features (red = CFS, blue = NF)

106 patients	75 patients
# granulocytes	# granulocytes
MCH	MCH
MCV	MCV
% granulocytes	WBC
WBC	% granulocytes
HGB	HGB
% lymphocytes	HCT
RBC	RBC
HCT	% lymphocytes
RDW	RDW

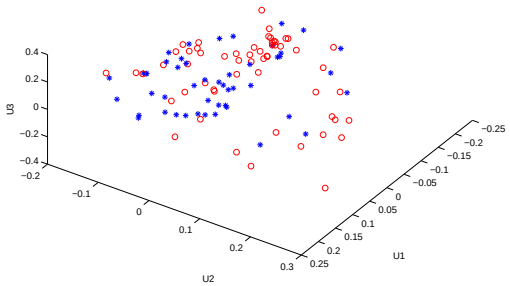
Table 4.3: The most important blood features using SVD (granulocytes: a category of white blood cells, MCH: Mean corpuscular hemoglobin, MCV: Mean corpuscular volume, WBC: White blood cells, HGB: Hemoglobin, lymphocytes: single nucleus white blood cells , RBC: Red blood cells, HCT: Hematocrit, RDW: Red blood cell distribution width)



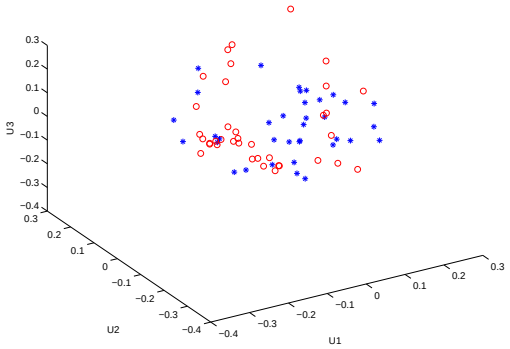
(a) 106 patients data matrix



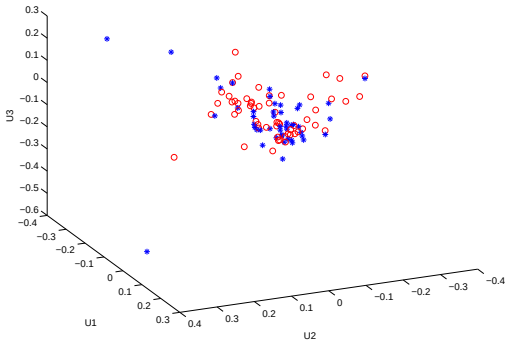
(b) 75 patients data matrix



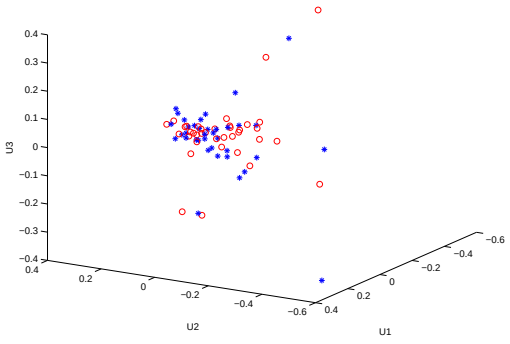
(c) 106 patients log-transformed data matrix



(d) 75 patients log-transformed data matrix

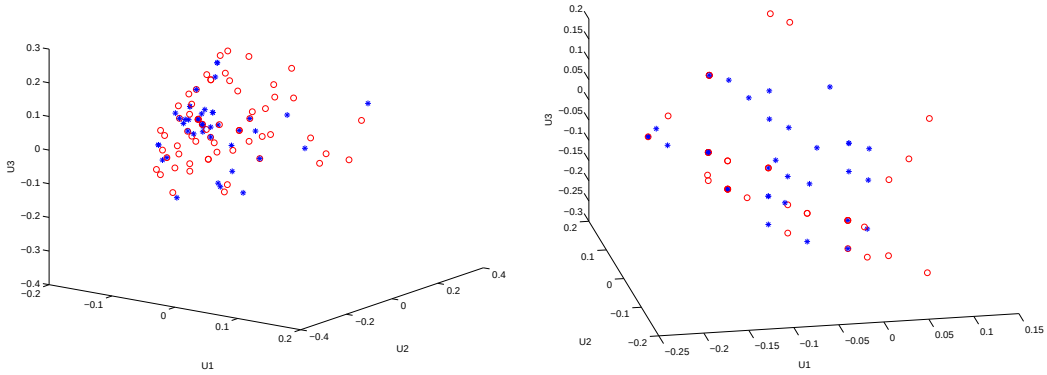


(e) 106 patients data matrix after cleaning with ICA



(f) 75 patients data matrix after cleaning with ICA

Figure 4.11: 3-dimensional plots of rows of the U matrix from the top SVD-ranked genes (red = CFS, blue = NF)



(a) 106 patients (b) 75 patients

Figure 4.12: 3-dimensional plots of rows of the U matrix from the top SVD-ranked SNPs (red = CFS, blue = NF)

SVD analysis results of the selected SNPs

Using the top 14 and 10 attributes of the ranked list obtained from the 106-patient and 75-patient data matrices, respectively, Figures 4.12(a) and 4.12(b) show the most noticeable clusters of CFS patients and NF controls. Table 4.4 shows the lists of the most separable SNPs ordered according to their distance from the origin. The lists has five SNPs in common. Out of the 14 SNPs from the 106-patient list, 11 of them can be found in three genes: COMT (Catechol-O-methyltransferase), NR3C1 (Nuclear receptor subfamily 3, group C,member 1) and CRHR1 (Corticotropin-releasing hormone receptor 1). The genes containing all the SNPs in the 75-patient list are MAOA (Monoamine oxidase A), TPH2 (Tryptophan hyroxylase 2) and NR3C1 .

4.3.2 Attribute selection using random forests

In this section, we show the effect of different settings of *mtry factor* and *nt* on the accuracy when we select different numbers of attributes. Using reasonable choices of

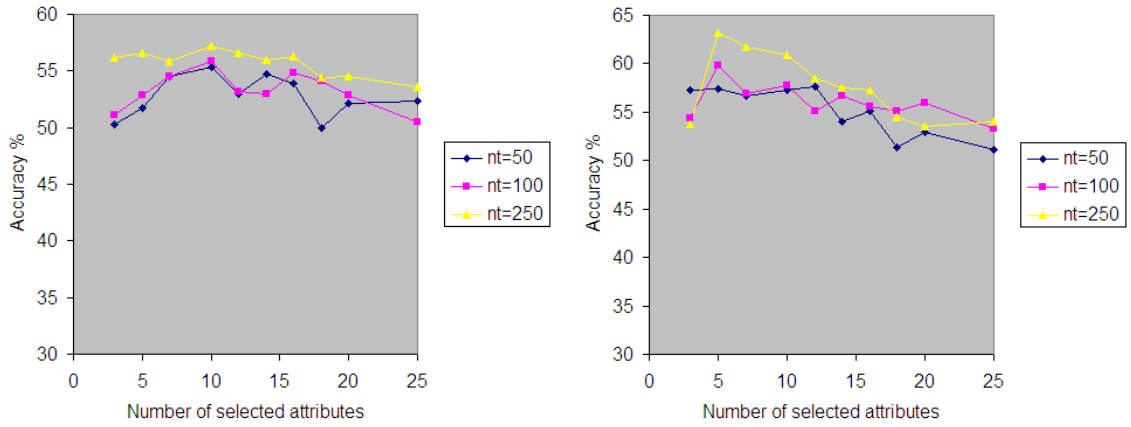
106 patients	75 patients
hCV2538747	hCV8878819
hCV2539306	hCV8878818
hCV2538746	hCV8878813
hCV11804650	hCV15836061
hCV8950988	hCV8376173
hCV11837659	hCV8376146
hCV1046361	hCV1046361
hCV1570087	hCV8376042
hCV2257689	hCV8950988
hCV2544843	hCV11837659
hCV8376042	
hCV2544836	
hCV8878819	
hCV8872233	

Table 4.4: The most important SNPs using SVD

mtry factor and *nt*, we obtain the average accuracy of 10 runs per experiment. The accuracy of the results is described in terms of 95% confidence intervals placed on the average values of these results. For each type of the data, we identify small sets of attributes that appear in more than one of the most important lists obtained from each experiment.

Random forests analysis results for the selected blood features

Figures 4.13(a) and 4.13(b) show the effect of changing the value of *nt* on the average accuracy of the experiments performed using different values of *mtry factor* for each number of selected attributes. The accuracy of the results in most cases increases as the number of trees increases and the number of selected important attributes decreases. Figures 4.14(a) and 4.14(b) show the effect of changing the value of *mtry*

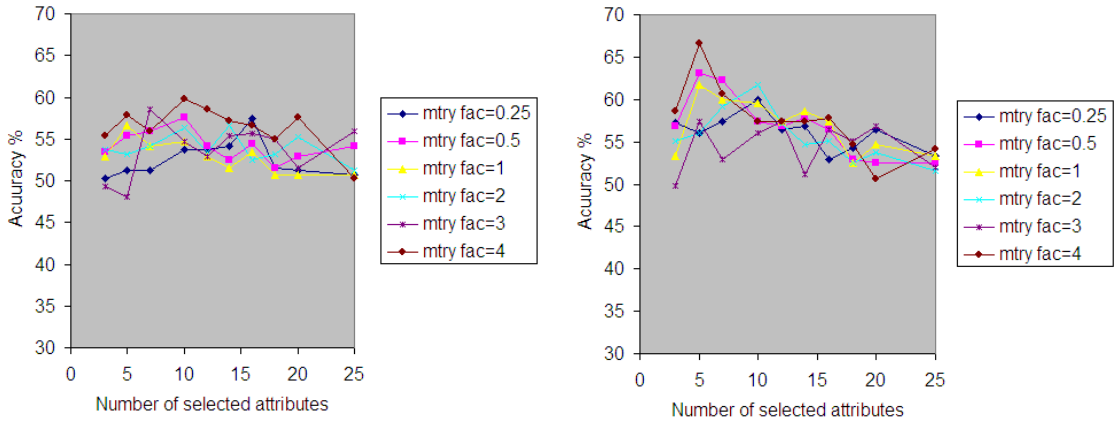


(a) 106 patients (b) 75 patients

Figure 4.13: Accuracy vs. number of selected blood features using different values of *nt*

factor on the average accuracy of the experiments performed using different values of *nt* for each number of selected attributes. Although there is no clear relation between changing the value of *mtry factor* and the classification rate, experiments using *mtry factor*=4 seems to achieve good results in many cases.

Next, using *mtry factor*=4 and *nt*=250, we average the results of 10 runs per experiment. Each experiment is performed using different number of selected attributes. Figure 4.15 shows the confidence intervals of the average accuracy versus the number of selected attributes. For the 106 patients, the average accuracy tend to decrease after selecting 14 attributes. The experiments on the 75 patients achieve higher accuracy for smaller numbers of selected attributes. Selecting seven attributes seem to achieve the best average accuracy. Each run produces a list of the most important 14 and 7 attributes using 106 and 75 patients, respectively. Table 4.5 presents the lists of the most important attributes that appear in at least five of the ten importance lists. The listed attributes are ordered according to their number of appearance in the



(a) 106 patients (b) 75 patients

Figure 4.14: Accuracy vs. number of selected blood features using different values of *mtry factor*

most important attributes lists. We notice that the most differentiable blood features for the 75 patients are a subset of those for the 106 patients.

SVD analysis results for the selected blood features

We further analyze the most important blood features selected by the random forests algorithm using SVD. Figure 4.16 shows the clusters of 106 and 75 patients in a three dimensional U-plot. A number of clusters can be recognized in Figure 4.16(a). For the 75 patients, the six blood features separate the two classes fairly well as shown in Figure 4.16(b).

Random forests analysis results for the selected genes

The results in Figures 4.17(a) and 4.17(b) show the effect of changing the value of nt on the average accuracy of the experiments performed using *mtry factor*=1 and 2. For different numbers of selected attributes, the experiments using nt=5000 seem

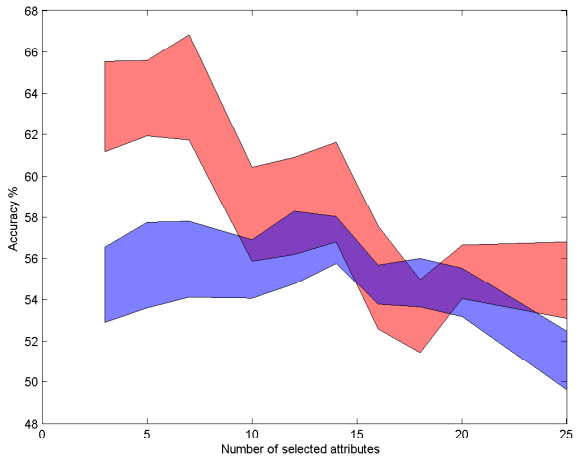
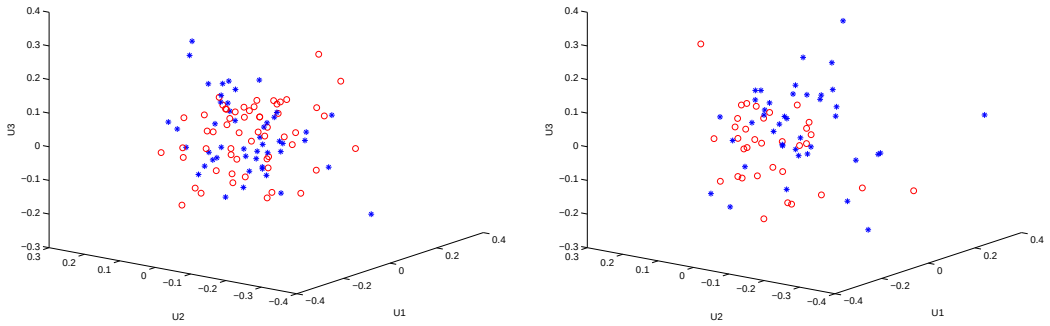


Figure 4.15: Accuracy % vs. number of selected blood features (blue = 106 patients, red = 75 patients)

106 patients	75 patients
HCT	sodium
sodium	potassium
potassium	% granulocytes
RBC	RBC
RDW	glucose
HGB	calcium
% granulocytes	
# granulocytes	
glucose	
albumin	
ALT/SGPT	
BUN	
calcium	
chloride	

Table 4.5: The most important blood attributes using random forests (HCT: Hematocrit, granulocytes: a category of white blood cells, RBC: Red blood cells, RDW: Red blood cell distribution width, HGB: Hemoglobin, BUN: Blood urea nitrogen, albumin: proteins, ALT/SGPT: alanine aminotransferase/serum glutamic pyruvic transaminase)

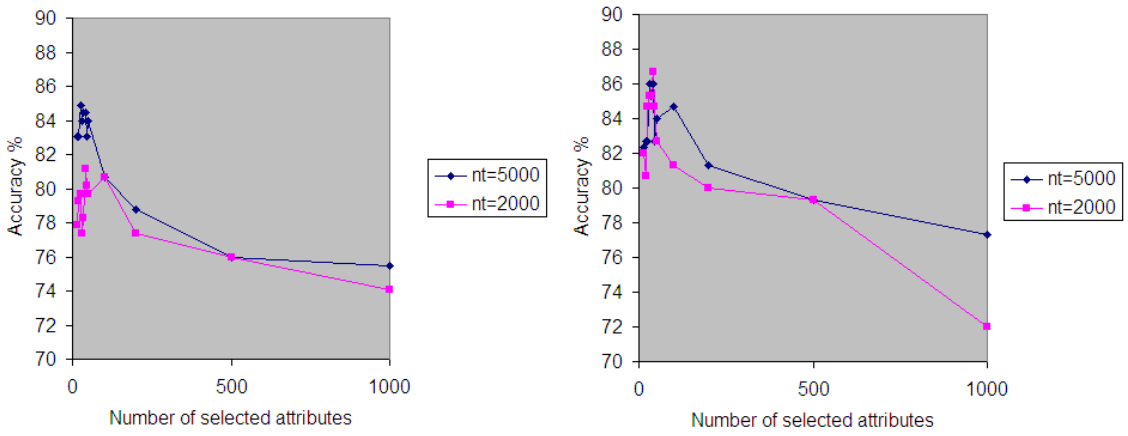


(a) 106 patients data matrix (b) 75 patients data matrix

Figure 4.16: 3-dimensional plots of rows of the U matrix from the selected blood features (red = CFS, blue = NF)

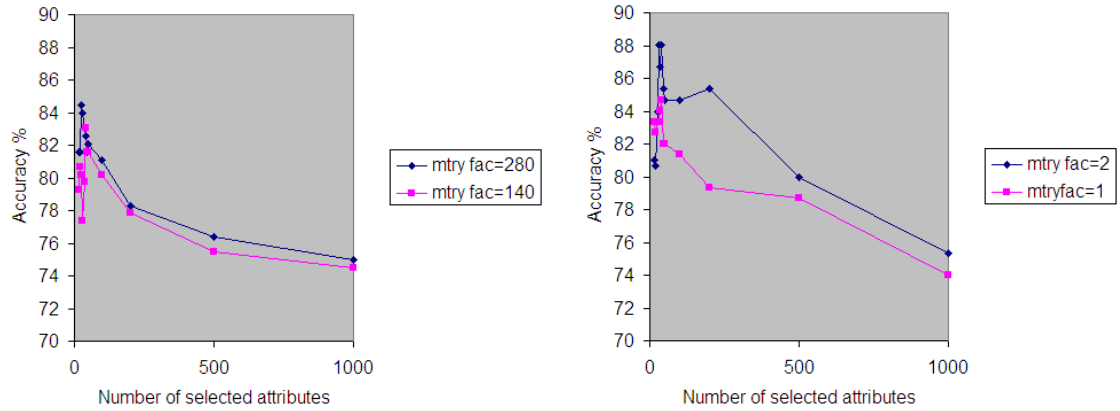
to achieve higher average accuracy in most of cases. Similarly, Figures 4.18(a) and 4.18(b) show the effect of changing the value of *mtry factor* on the average accuracy of the experiments performed using nt=5000 and 2000 for different number of selected attributes. In general, we notice that the average accuracy of the experiments using *mtry factor*=2 is higher.

Using *mtry factor*=2 and nt=5000, we repeat each experiment performed using different numbers of selected attributes, 10 times. Figure 4.19 shows the confidence intervals of the average accuracy versus the number of selected attributes. In general, the average accuracy using the 75 patients is higher, except for small overlapping with the 106-patient confidence intervals between 35 and 45 selected genes. Although it is hard to decide how many genes should be selected, we choose to report the results of selecting 40 and 50 genes of the 106-patient and 75-patient data matrices, respectively. Selecting these numbers of genes seem to achieve one of the best average accuracies as shown in Figure 4.19. Each run produces a list of the most important 40 and 50 attributes using 106 and 75 patients, respectively. Tables 4.6 and 4.7 include the lists



(a) 106 patients (b) 75 patients

Figure 4.17: Accuracy vs. number of selected genes using different values of *nt*



(a) 106 patients (b) 75 patients

Figure 4.18: Accuracy vs. number of selected genes using different values of *mtry factor*

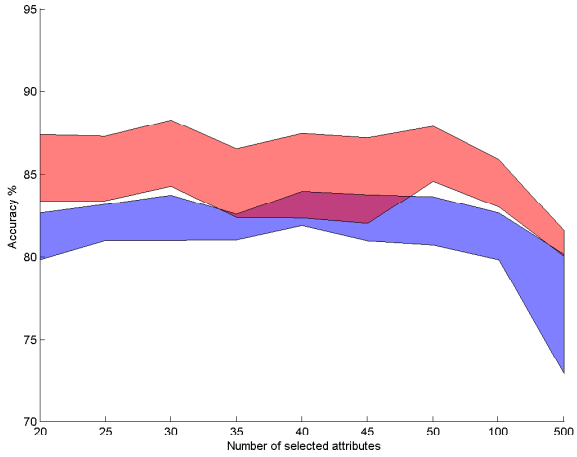


Figure 4.19: Accuracy % vs. number of selected genes from the microarray data (blue = 106 patients, red = 75 patients)

of the most important genes ordered according to their number of appearance in the importance lists of each experiment. We list only those genes that appear in at least five of the ten importance lists. The two lists have five genes in common.

We also perform the same experiments on the ICA-cleaned microarray data. Figure 4.20 shows that the average accuracy for the classification of 75 patients is higher except for the overlapping at 500 selected genes. The best average accuracy is achieved by selecting 35 and 45 genes of the 106-patient and 75-patient data matrices, respectively. Tables 4.8 and 4.9 show the lists of most important genes ordered according to their number of appearance in the importance lists of each experiment. Again, the listed genes are those that appear in at least five of the ten importance lists obtained from each experiment. The two lists have six genes in common.

Gene ID[1]	Description
AK074045(4)	flj00103 protein; flj00103
BC001528	unknown (protein for mgc:3298)
BC011636	unknown (protein for mgc:12913)
AF151022(2)	hspc188
BC022334	5' nucleotidase, deoxy (pyrimidine), cytosolic type c
BC005327(2)	similar to nucleoporin-like protein 1
AL035691	dj249f5.2 (mas1 oncogene); mas1
AF349452	claudin-18a2.1
BC001047(2)	nucleolar autoantigen (55kd) similar to rat synaptonemal complex protein
AB007774	cystatin a
AK026618	homo sapiens cdna: flj22965 fis, clone kat10418; unnamed protein product.
AF112210	heat shock protein hsp70-related protein
XM_062367	similar to aspartate aminotransferase, mitochondrial precursor (transaminase a) (glutamate oxaloacetate transaminase-2); loc120943
AF412028	williams-beuren syndrome critical region protein 20 copy a; wbscr20a
AK093186	cdna clone weakly similar to zinc finger protein 29; unnamed protein product.
AF051325	sh3 domain containing adaptor protein; scap
AF178072	rgs9 isoform 2; rgs9
AF338109(2)	proapoptotic caspase adaptor protein
AF281049	testis fascin; fscn3

Table 4.6: The most important 19 genes obtained from random forests analysis of the 106-patient microarray data

[1]The numbers appearing next to boldfaced Gene IDs indicate how many times each gene appears in the following four lists: the most important genes obtained from the 106-patient microarray data, the 75-patient microarray data, the 106-patient ICA-cleaned microarray data and the 75-patient ICA-cleaned microarray data

Gene ID	Description
NM_048368(3)	ctd (carboxy-terminal domain, rna polymerase ii, polypeptide a) phosphatase, subunit 1, isoform fcp1b; ctdp1
AF118275(2)	atrophin-related protein arp
XM_084429	similar to 60s ribosomal protein l21; loc143125
AF255650	dc27
AK074045(4)	flj00103 protein; flj00103
AK027561	cdna clone unnamed protein product.
AF100928	apoptosis-inducing factor aif
AK002198	cdna clone weakly similar to testis-specific protein pbs13; unnamed protein product.
AK025470	homosapiens cdna: flj21817, clone hep01141; unnamed protein product.
L25851	integrin alpha e precursor
AB057723(2)	phosphatidyl inositol glycan class s; pig-s
NM_015974	lambda-crystallin; cryl1
XM72647	similar to meningioma-expressed antigen 6/11 (mea6) (mea11); loc253288
L05515	camp response element-binding protein; creb1
AK026486	homosapiens cdna: flj22833, unnamed protein product.
NM_003608	g protein-coupled receptor 65; gpr65
AL023653	dj753p9.2 (novel protein); dj753p9.2
AK026308	homosapiens cdna: flj22655, unnamed protein product.
AF151022(2)	hspc188
AF035839(2)	nadh:ubiquinone oxidoreductase b12 subunit
BC001327	interferon-related developmental regulator 2
BC001047(2)	nucleolar autoantigen similar to rat synaptonemal complex protein
AF212848	ets domain transcription factor; esej
NM34263	solute carrier family 26, member 6, isoform b; slc26a6
AK000010	cdna clone unnamed protein product.
BC005327(2)	similar to nucleoporin-like protein 1
AB007447(3)	fln29; fln29
AK021784	cdna clone unnamed protein product.
XM_098013	hypothetical protein xp_098013; loc151174
AK002136	cdna clone weakly similar to metal homeostasis factor atx2; unnamed protein product.
BC014097	similar to old astrocyte specifically induced substance
BC036014	unknown (protein for mgc:32916)
AK000183	cdna clone unnamed protein product.
AF338109(2)	proapoptotic caspase adaptor protein
AF030424	histone acetyltransferase 1
AB044343	udp-glucuronic acid; hugtrel7
AB023421	apg-1

Table 4.7: The most important 37 genes obtained from random forests analysis of the 75-patient microarray data

Gene ID	Description
AK001563	cdna clone unnamed protein product.
AK074045(4)	flj00103 protein; flj00103
AF152513(2)	protocadherin gamma a6 short form protein; pcdh-gamma-a6
AB007447(3)	fln29; fln29
NM_006986	melanoma antigen, family d, 1; maged1
NM_006400(2)	dynactin 2; dctn2
AF263462	cingulin
NM_019624	atp-binding cassette, sub-family b, member 9, isoform 2; abcb9
AF419332	tls-associated protein tasr-2
AB075819	kiaa1939 protein; kiaa1939
M74089	tb1
NM_048368(3)	ctd (carboxy-terminal domain, rna polymerase ii, polypeptide a) phosphatase, subunit 1, isoform fcp1b; ctdp1
AK026773	homo sapiens cdna: flj23120 fis, clone lng07989; unnamed protein product.
NM_004939	dead/h (asp-glu-ala-asp/his) box polypeptide 1; ddx1
NM_014158	hspc067 protein; hspc067
AB005038	25-hydroxyvitamin d3 1-alpha-hydroxylase
NM_018288	phd zinc finger protein xap135, isoform a; xap135
AF083249	rb binding protein homolog
BC015130	cytochrome c
NM_025226(2)	mstp032 protein; mstp032
AL031774	dek (putative oncogene); dek
XM_085827	similar to zinc finger protein 347; zinc finger 1111; loc147658

Table 4.8: The most important 22 genes obtained from random forests analysis of the 106-patient ICA-cleaned microarray data

Gene ID	Description
NM_048368(3)	ctd (carboxy-terminal domain, rna polymerase ii, polypeptide a) phosphatase, subunit 1, isoform fcp1b; ctdp1
AF035839(2)	nadh:ubiquinone oxidoreductase b12 subunit
AF152513(2)	protocadherin gamma a6 short form protein; pcdh-gamma-a6
AB007447(3)	fln29; fln29
AF047445	ly-49l; ly49l
NM_006400(2)	dynactin 2; dctn2
NM_021832	a disintegrin and metalloproteinase domain 17, isoform 2 prepro-protein; adam17
AK027028	homo sapiens cdna: flj23375 fis, clone hep16206; unnamed protein product.
AB057723(2)	phosphatidyl inositol glycan class s; pig-s
AF421380	anthrax toxin receptor
AF118275(2)	atrophin-related protein arp
NM30388	ankyrin repeat and socs box-containing protein 12; asb12
AF127090	oxysterol 7alpha-hydroxylase; cyp7b1
AK024077	cdna clone unnamed protein product.
BC007950	similar to heterogeneous nuclear ribonucleoprotein u (scaffold attachment factor a)
BC028721	similar to solute carrier family 1 (high affinity aspartate/glutamate transporter), member 6
AF052205	phd finger protein 2; phf2
AF365931	kruppel-type zinc-finger protein zim3
BC028101	bruno-like 5, rna binding protein (drosophila)
AK074045(4)	flj00103 protein; flj00103
AF177941	collagen type v alpha 3 chain; col5a3
D38076	ran-bp1(ran-binding protein 1)
AF111706	sltp003; lst003
AF210317	facilitative glucose transporter family member glut9; slc2a9
BC004230	triosephosphate isomerase 1
AF016004	m6b1
NM_025226(2)	mstp032 protein; mstp032

Table 4.9: The most important 27 genes obtained from random forests analysis of the 75-patient ICA-cleaned microarray data

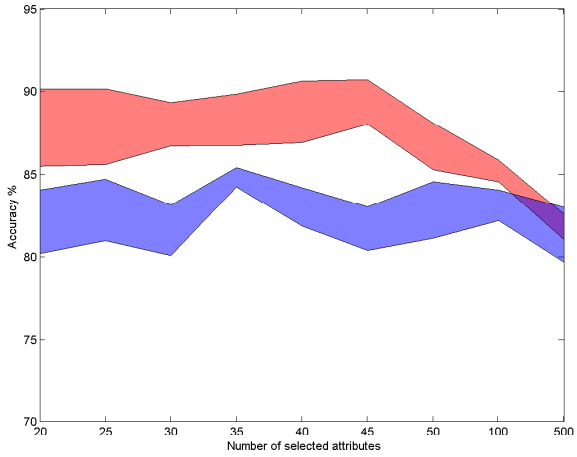


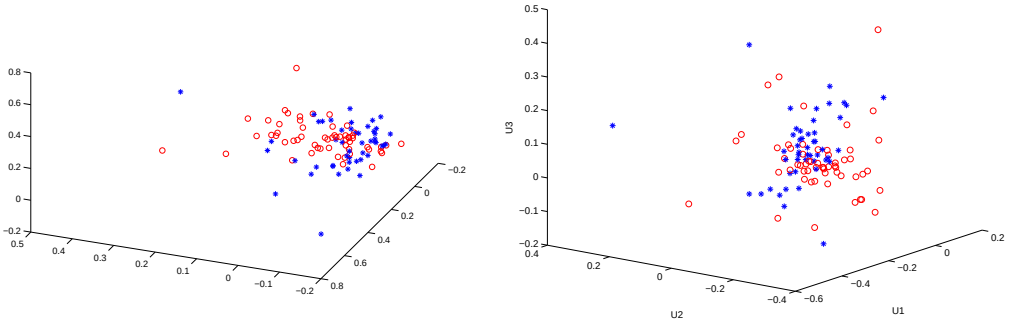
Figure 4.20: Accuracy % vs. number of selected genes from the ICA-cleaned microarray data (blue = 106 patients, red = 75 patients)

SVD analysis results for the selected genes

We further analyze the significantly differentially expressed genes using SVD. We apply SVD to the most important 19, 22, 37 and 27 genes identified in at least five of the ten runs from the 106-patient original data, 106-patient ICA-transformed data, 75-patient original data and 75-patient ICA-transformed data matrices, respectively. Figure 4.21 shows the clustering of the 106 patients. SVD clusters the two classes fairly well in both figures. Moreover, Figure 4.22 shows that SVD can separate the two classes of the 75 patients better than those of 106 patients. The two classes appear in two adjacent clusters.

Random forests analysis results for the selected SNPs

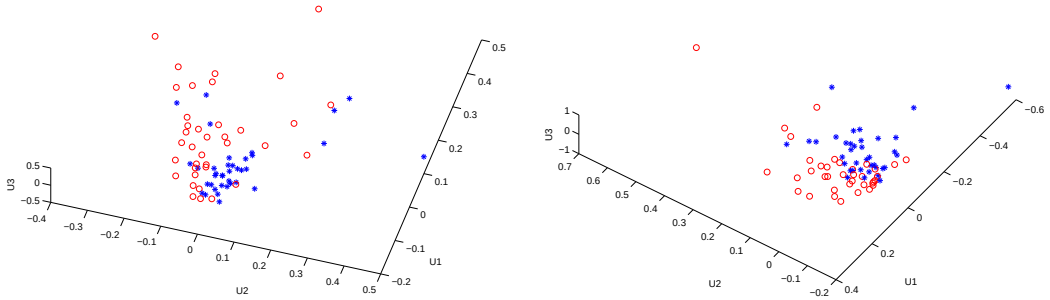
Figures 4.23(a) and 4.23(b) show the effect of changing the value of *nt* on the average accuracy of the experiments performed using different values of *mtry factor* for



(a) The original microarray data matrix

(b) The ICA-cleaned microarray data matrix

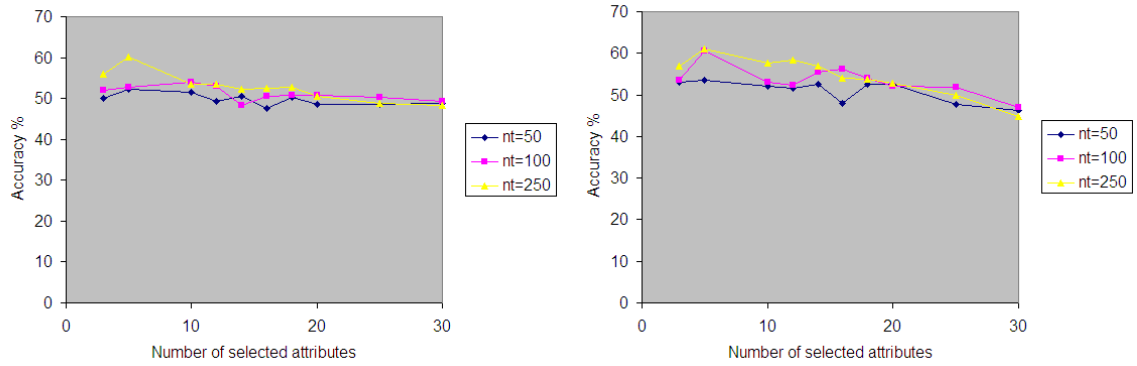
Figure 4.21: 3-dimensional plots of rows of the U matrix of the selected genes of 106 patients (red = CFS, blue = NF)



(a) The original microarray data matrix

(b) The ICA-cleaned microarray data matrix

Figure 4.22: 3-dimensional plots of rows of the U matrix of the selected genes of 75 patients (red = CFS, blue = NF)

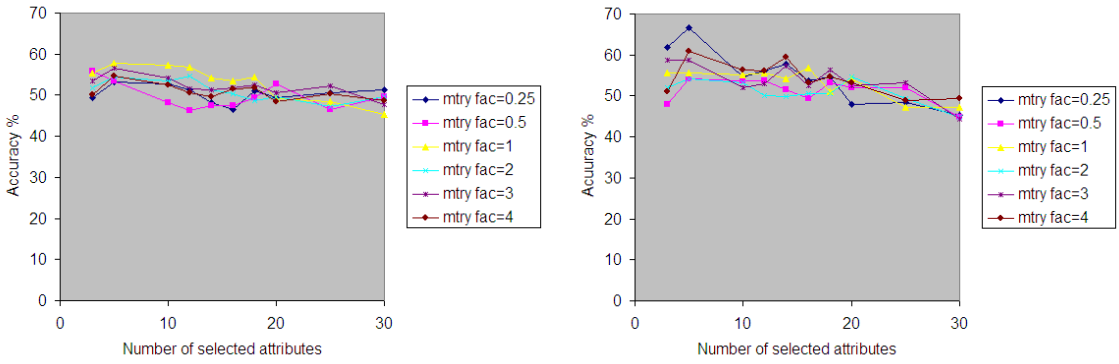


(a) 106 patients (b) 75 patients

Figure 4.23: Accuracy vs. number of selected SNPs using different values of *nt*

different numbers of selected attributes. The accuracy of the results in many cases increases as the number of trees increases and the selected important attributes decreases. Figures 4.24(a) and 4.24(b) show the effect of changing the value of *mtry factor* on the average accuracy of the experiments performed using different *nt* values for different numbers of selected attributes. In general, the accuracy increases when the number of selected important attributes decreases with no particular value of *mtry factor* outperforms other values. However, setting *mtry factor* to 4 seems to be a reasonable choice.

To estimate the overall performance of the random forests algorithm on the SNP data, we perform 10 runs for each experiment, setting *mtry factor* to 4 and *nt* to 250. Figure 4.25 shows the confidence intervals of the average accuracy obtained using difference number of selected attributes. For both sets of patients, the average accuracy tend to increase as the number of selected attributes decreases. Selecting the most important five attributes achieves the best average accuracy. Table 4.10 shows the lists of most important SNPs ordered according to their number of appearance in



(a) 106 patients (b) 75 patients

Figure 4.24: Accuracy vs. number of selected SNPs using different values of *mtry factor*

the importance lists of each run. We list only those attributes that appear at least twice in the 10 importance lists obtained from the 10 runs. The two lists has five SNPs in common. MAOA and SLC6A4 (Solute carrier family 6, member 4) have six out of the 10 SNPs of the 106-patient list. Seven out of the 10 SNPs of the 75-patient list belong to the genes MAOA and NR3C1.

SVD analysis results for the selected SNPs

We apply SVD to the the most important 10 SNPs obtained from the random forest analysis. Figure 4.26 demonstrates the clusters of 106 and 75 patients in three dimensional U-plots. For 106 patients, Figure 4.26(a) shows a number of clusters, one of them is clearly a cluster of NF controls. In Figure 4.26(b), we notice that the 10 selected SNPs separate the two classes fairly well.

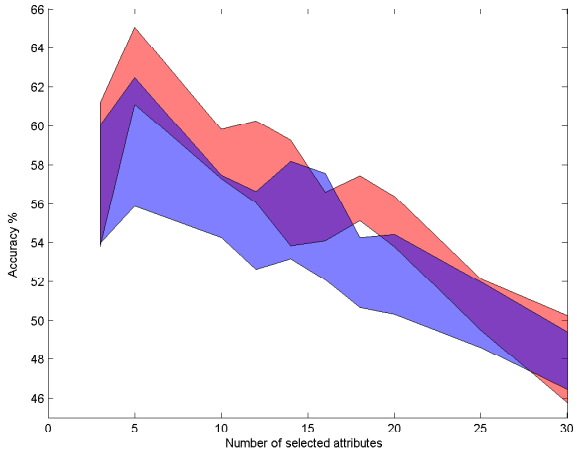
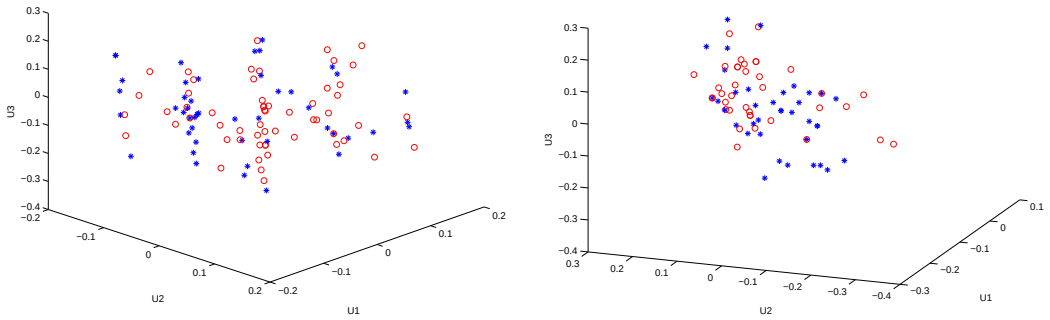


Figure 4.25: Accuracy % vs. number of selected SNPs (blue = 106 patients, red = 75 patients)

106 patients	75 patients
hCV1046360	hCV15836061
CRHR1-7450777	hCV2539273
hCV8878819	hCV8878813
hCV1841702	hCV1046361
hCV8950998	hCV3227244
hCV7911132	hCV8878818
hCV8376146	hCV8950998
hCV8878813	hCV1841702
hCV8878818	hCV8878819
hCV11159943	hCV11837659

Table 4.10: The most important SNPs using random forests



(a) 106 patients data matrix (b) 75 patients data matrix

Figure 4.26: 3-dimensional plots of rows of the U matrix from the selected SNPs (red = CFS, blue = NF)

	Blood data	microarray data	SNP data	Voting
106 patients	59.43	85.85	68.87	76.42
75 patients	68	89.33	72	85.33

Table 4.11: A comparison of the best accuracies(%) obtained from the individual data and voting using the original microarray data

4.3.3 Analysis results for the integration

In the following, we present the results of integrating different types of data using voting and combination as explained in Section 3.4.3. Tables 4.11 and 4.12 shows the best results that could be obtained from each type of data separately and from voting. The results show a decrease in the accuracy. The reason is that the blood and SNP data seem to agree on the misclassified patients. Moreover, misclassified patients using the microarray data are also misclassified using either the blood or SNP data. As a result, voting does not improve the classification.

To examine the effectiveness of combining the different types of data, we apply

	Blood data	ICA-cleaned microarray data	SNP data	Voting
106 patients	59.43	85.85	68.87	76.42
75 patients	68	90.67	72	85.33

Table 4.12: A comparison of the best accuracies(%) obtained from the individual data and voting using the ICA-cleaned microarray data

	Blood features	Genes	SNPs	Total
106-patient original data	14	19	10	43
106-patient ICA-cleaned data	14	22	10	46
75-patient original data	6	37	10	53
75-patient ICA-cleaned data	6	27	10	43

Table 4.13: The number of the most important attributes selected from each type of the data obtained using the random forest permutation importance measurement

the random forests algorithm to a matrix that contains the 106 and 75 patients, each patient is described by the most important attributes selected from each type of the data. Table 4.13 is a summary of the number of those selected attributes. A comparison of the average accuracies obtained from each type of the data separately in previous sections and for the combined data is shown in Tables 4.14 and 4.15.

	Blood data	microarray data	SNP data	Combination
106 patients	56.89 ± 1.14 from Fig. 4.15	82.93 ± 1.05 from Fig. 4.19	59.15 ± 3.3 from Fig. 4.25	83.96 ± 0.83
75 patients	64.4 ± 1.7 from Fig. 4.15	86.22 ± 1.65 from Fig. 4.19	63.6 ± 1.98 from Fig. 4.25	88.31 ± 0.98

Table 4.14: A comparison of the best accuracies(%) obtained from the individual data and random forests combination using the original microarray data

	Blood data	microarray data	SNP data	Combination
106 patients	56.89 ± 1.14 from Fig. 4.15	84.81 ± 0.58 from Fig. 4.20	59.15 ± 3.3 from Fig. 4.25	84.03 ± 0.73
75 patients	64.4 ± 1.7 from Fig. 4.15	89.33 ± 1.34 from Fig. 4.20	63.6 ± 1.98 from Fig. 4.25	91.18 ± 0.69

Table 4.15: A comparison of the best accuracies(%) obtained from the individual data and random forests combination using the ICA-cleaned microarray data

The results suggest an improvement in the average classification accuracy obtained from the integrated 106-patient data using the original microarray data and the integrated 75-patient data using the original and the ICA-cleaned microarray data. Integrating the data of 106 patients achieves an increase of approximately 1% of the average accuracy with tighter confidence intervals. For 75 patients, an improvement of approximately 2% is achieved by integrating the three types of data using both versions of the microarray data. In general, we notice that the confidence intervals of the accuracies obtained from the integrated data are tighter than those obtained from each data individually .

4.4 Discussion

This study clearly supports the hypothesis that Chronic Fatigue Syndrome (CFS) has a biological basis. In this section, we first compare the results obtained from the two attribute-selection techniques we employed in this study, SVD and the random forest permutation importance measurement. Second, we discuss the effect of changing the parameters of the random forests algorithm on the performance. Finally, we compare the best results of our analysis with those obtained previously in the literature.

SVD and random forests show that the irrelevant features reduce the ability to distinguish between clusters and degrade the classification accuracy of patients. As we mentioned earlier, ranking attributes according to their distance from the origin in the V-space of SVD gives a sensible measurement of attribute importance. In many cases, selecting the top-ranked attributes improves the clustering of CFS patients and controls (for example, Figure 4.4 and Figure 4.12). However, this technique seems to be unsuitable for the microarray data since the structure of the clusters for small numbers of selected genes is unclear. We also noticed that the change in the U-plot before and after selecting the most important 500 attributes is considered small in comparison with the number of discarded attributes. Using the random forests permutation importance measurement, a more significant improvement in the classification accuracy can be observed by selecting different numbers of attributes. We also show that SVD performs well on these selected attributes. The results indicate that SVD is not the right attribute selection technique for the microarray data. However, SVD is considered robust against the number of irrelevant attributes in comparison to random forests which is strongly misled by those irrelevant features.

We have thoroughly evaluated the effect of different settings of random forests parameters, *mtry factor* and *nt*, before and after selecting different numbers of attributes. The initial results before attribute selection show that the random forest achieves higher accuracies for large values of *mtry factor* and *nt* using the microarray data. Other types of the data did not indicate a notable relation between the changes of those parameters and the performance of random forests. The subsequent analysis shows that setting *mtry factor* and *nt* to large values leads to higher accuracies using different numbers of selected attributes.

	Previous best results			Our best results	
Data	Ref.	Accuracy(%)	#	Accuracy(%)	#
Blood data	[8]	50	191	56.89 ± 1.14 64.4 ± 1.7	106 75
Microarray data	[8]	77	130	84.81 ± 0.58 89.33 ± 1.34	106 75
SNP data	[27]	76	101	59.15 ± 3.3 61.6 ± 1.98	106 75
Integrated data	[35]	73	164	84.03 ± 0.73 91.18 ± 0.69	106 75

Table 4.16: A comparison of the best accuracies(%) obtained from previous research and our results

In order to compare our analysis with other studies in the literature, Table 4.16 shows the best results obtained using different classification techniques and the number of patients on which these studies were based. Although there is no evidence in the literature that the blood data is useful in predicting CFS, however, our analysis shows that some blood features might correlate with the presence of CFS. A number of those features are common between the two lists of the important features obtained by SVD and random forests. Several features such as RBC, RDW HGB and HCT which are related, indicate possible abnormalities in red blood cells. They might also indicate the redundancy of the attributes of the blood data. In all cases, these features are not sufficient by themselves to indicate any biological basis for CFS. Kennedy and colleagues [33] also find imbalances in the attributes associated with red blood cells. However, the imbalances are found in the normal ranges and described as "insufficient to characterize fatigued patients". Further analysis of the selected blood features is required in order to draw more reliable conclusions about the efficiency of the blood evaluation in CFS diagnosis.

Our analysis yields strong indications that CFS has a biological basis that is better detectable in the microarray data than any other type of the data. As shown in Table 4.16, our prediction accuracies obtained using random forests are significantly higher than the best accuracy reported so far. We notice that the classification accuracy using the ICA-cleaned microarray data is higher than the classification accuracy using the raw data. This indicates that ICA was able to remove some of the spatial artifacts which led to this improvement.

To provide extra validation of the results, we randomly labeled the patients as CFS or control. The average accuracy using the shuffled set of the ICA-cleaned microarray data is 72% and 74% for 106 and 75 patients respectively. We also performed the same shuffling experiment on the integrated data. The average accuracy using the shuffled set of the integrated data is approximately 50% and 60% using the ICA-cleaned microarray data of 106 and 75 patients respectively. The difference between the results on the shuffled and unshuffled dataset is significant indicating that our results are significant. The improvement of the accuracy we get is due to the fact that each tree of the forests is constructed using a subset of randomly selected objects and each node is split using m randomly selected attributes from the original M attributes. As a result, the chance of spurious correlations with the target attribute being included in the model is small.

The random forests analysis produces four lists of differentially expressed genes that have some genes in common. In this study, we do not focus on the functional annotation of the genes. However, we compare our lists with others that have appeared in the related studies to find any common genes. Also, our collaborators are working on verifying the biological plausibility of our lists of genes. Carmel and

colleagues [16] find SLC1A6 as one of the most interesting genes that distinguish between most classes of fatigued and healthy subjects. This gene is also one of our 27 most important genes obtained from the 75-patient ICA-cleaned microarray data. The gene is a high affinity aspartate/glutamate transporter and it has a role in regulating excitatory neurotransmission. Bassetti and colleagues [8] assessed their differentially expressed genes with the GO structures and find a number of their genes annotated with the general GO term "metabolism". One of these genes is hspc188 (Homo sapiens Hematopoietic stem and progenitor cell-188 mRNA). This gene is also one of the top genes in the list of the most important 19 genes obtained from the 106-patient microarray data. Another list of 24 differentially expressed genes is identified in [22]. Two of these genes, SCN4A and BLP1, are identified from the original and the log-transformed microarray data, respectively. SCN4A function involves metal ion binding, ion transport and ion channel activity. BLP1 is involved in signal transduction, cell-cell signaling and regulation of cell growth.

Although the accuracy obtained from our analysis of the SNP data is not significant, the best result achieved among the 10 runs is 72% which is comparable to the best result achieved earlier by Goertzel and colleagues [27]. However, there are several issues that are worth mentioning in their work (a summary of Goertzel's and colleagues method is provided in Section 2.3.3). First, the threshold to which the rules are compared is selected for each rule to allow the SNP set to achieve the maximum accuracy, which makes the method impractical. Also, the authors use random shuffled datasets to validate their results and the frequency of the accuracies obtained using these shuffled datasets concentrates between 68 to 72% achieving the best accuracy of 75%. In addition to this, the enumerative search is not always an attractive

technique since the SNP data is usually much larger than the CAMDA provided set. Comparing our lists of SNPs to those obtained in Bassetti and colleagues study [8], we find that 10 out of 17 SNPs identified in their work as predictive SNPs, exists in our lists. In particular, the SNP hCV8878819 can be found in all the lists, hCV11837659 and hCV1046361 in three lists and hCV8376042 in two lists. The genes associated with these SNPs are MAOA, TPH2 and NR3C1. Our results also agree with those discussed by Smith and colleagues [47]. MAOA (MonoAmine Oxidase A) has been found to correlate with mood abnormalities and stress response. It has been reported also that SNPs in TPH2 (Tryptophan hydroxylase 2) may indicate dysregulation of the sertoronergic system. Thus, the association of both genes indicates that mood abnormalities might be mediated by the serotonergic system [47]. NR3C1 (Nuclear receptor subfamily 3 group C memeber1) is the glucocorticoid receptor which binds with high affinity to cortisol and other glucocorticoids. Cleare and colleagues have detected lower cortisol responses in CFS patients [19]. There results indicate that further analysis of the SNP data might contribute to the diagnosis of CFS.

The prediction accuracy obtained from integrating the blood, microarray and SNP data indicates that studying more than one type of data to investigate the characteristics of CFS is likely to help further in CFS classification. Each of the selected blood features and SNPs slightly helps to improve CFS prediction. However, none of them clearly reflects the biological basis of CFS. Analyzing larger sets of blood and SNP data might be more informative to draw more precise conclusions about the nature of CFS. So far, the integrative model achieves a higher accuracy than the best accuracy achieved previously as shown in Table 4.16.

4.5 Summary

In summary, the first part of the experiments showed that clustering analysis using SVD resulted in an initial clustering of CFS patients and NF controls. SVD also confirmed the empirical classification using the clinical data. Moreover, the classification analysis using random forests led to accuracies close to the base accuracy. In general, the initial results from Sections 4.1 and 4.2 suggested the need to select subsets of attributes that are more relevant to CFS. Therefore, we utilized SVD and the random forests permutation importance measurement to select the most important attributes from each type of the data. Using the most important attributes, the accuracy of the classification improved significantly. We also showed that combining different types of data slightly improved the average accuracy obtained from different experiments. Finally, we discussed the outcomes of our analysis.

Chapter 5

Conclusions and limitations

5.1 Conclusions

In the absence of laboratory and clinical markers for CFS, research to find a biological basis for it is still open. Multiple studies have sought to reach conclusions about the nature of CFS; however, they are inconsistent. One direction of these studies is to examine the clinical data to improve the definition of the syndrome. Another direction is to study the biological data to find biological markers for CFS. Moreover, studies have been conducted to explore the unexplained disorder by constructing integrative models that combine multiple types of data. For all these purposes, CAMDA has provided a dataset that describes a population of CFS patients and controls. The dataset consists of clinical data, blood profiles, SNPs and the expressions of thousand of genes for different numbers of patients and controls.

Many data-mining techniques have been widely employed to analyze biomedical data. However, standard techniques often lack the ability to handle high-dimensional datasets, especially when the number of patients is small. SVD is a powerful technique

for handling high-dimensional datasets. This is due to fact that SVD captures the most important variations detected in the data in early dimensions. Random forests is another effective data-mining technique for classification problems. The fact that the random forests algorithm uses both bagging and random variable selection for constructing the trees makes it suitable for handling situations where there are more attributes than patients.

In this thesis, we studied each type of the CAMDA data individually, and in an integrated dataset. We showed that CFS has a biological basis that is more detectable in the gene expression data than the blood and the SNP data. The random forests permutation importance measurement proved to be more effective than SVD in selecting the most distinguishable attributes between CFS patients and controls. The wide examination of the effect of changing the parameters of the random forests algorithm showed that assigning *mtry factor* and *ntree* to relatively large values leads to higher accuracies. We also showed that irrelevant features reduce the ability to distinguish the clusters of patients and degrade the classification accuracy. The classification of CFS patients and controls excluding those patients who suffer from exclusionary medical or psychiatric conditions achieved a higher accuracy than the classification of the full set of patients and controls. Thus, this study also support the previous recommendations to consider these conditions as exclusions for CFS. Finally, integrating the most distinguishable blood features, genes and SNPs showed that the integration of relevant attributes from different sources improves the diagnostic classification accuracy for CFS.

5.2 Limitations

Analyzing the CAMDA data has several limitations summarized as follows.

First, the number of patients provided is relatively small for training and testing the random forests classifier. Integrating different types of the data required matching patients and controls which decreased the sample size even more. Although bagging is argued to enhance the accuracy and gives a good estimate of the prediction error, a larger number of patients is required to achieve a more reliable accuracy.

Moreover, the noise introduced to the microarray data during the experiments might be crucial and have great effect on the most relevant selected genes. Although we have taken a step toward cleaning spatial artifacts from the data, we believe that the data still needs further careful cleaning. One possible solution to avoid considering irrelevant genes, is to construct a model based on prior biological knowledge of candidate genes. However, this solution presents a new limitation for building models based on inaccurate prior knowledge.

Another limitation is that the number of SNPs for candidate genes provided by CAMDA is extremely small in comparison to several millions of SNPs available in the public domain [4]. For such small number of SNPs, even high accuracies may do not reflect the real correlation of SNP data to CFS.

Each type of data has different type of noise and degree of reliability. Integrating different types of data measured by different technologies may introduce new challenges to the analysis. One clear challenge is the additional noise in the resulting dataset. To overcome such challenges, we need more effective methods that can handle such integrated data from various sources with different qualities and noise features.

Bibliography

[1] CAMDA Conference, July 2006, retrieved November 5, 2007, from http://www.camda.duke.edu/camda06.html.

[2] K. Aas, "Microarray data mining: A survey," Norsk Regnesentral Note, Norwegian Computing Center, Tech. Rep., 2001, SAMBA/02/01.

[3] H. Abdulsalam, D. Skillicorn, and P. Martin, "Streaming random forests," in *Eleventh International Database Engineering and Applications Symposium (IDEAS07)*, 2007, pp. 225–232.

[4] J. Aerts, Y. Wetzels, N. Cohen, and J. Aerssens, "Data mining of public SNP databases for the selection of intragenic SNPs," *Human Mutation*, vol. 20, no. 3, pp. 162–173, 2002.

[5] N. Afari and D. Buchwald, "Chronic fatigue syndrome: a review," *Psychiatry*, vol. 60, pp. 221–226, 2003.

[6] O. Alter, P. Brown, and D. Botstein, "Singular value decomposition for genome-wide expression data processing," *Proceedings of the National Academy of Sciences*, vol. 97, no. 18, pp. 10 101–10 106, August 2000.

[7] "Taqman SNP genotyping assays," Applied Biosystems, 2007, retrieved November 5, 2007, from http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_040597.pdf.

[8] M. Bassetti, M. Bernabe, M. Borile, C. Desilvestro, T. Fedrizzi, A. Giordani, R. Larcher, A. Palmisano, A. Salteri, S. Schivo, N. Segata, L. Tambosi, R. Valentini, P. Andritsos, P. Fontana, A. Malossini, and E. Blanzieri, "Validation of CFS classification with different data sources," in *Proceedings of Workshop: Critical Assessment of Microarray Data Analysis (CAMDA)*, 2006.

[9] C. Bierl, R. Nisenbaum, D. Hoaglin, B. Randall, A. B. Jones, E. R. Unger, and W. C. Reeves, "Regional distribution of fatiguing illnesses in the united states: a pilot study," *Cost Effectiveness Resource Allocation*, vol. 2, no. 1, 2004.

[10] D. Boswell, "Introduction to Support Vector Machines," 2002, retrieved November 5, 2007, from http://www.work.caltech.edu/~boswell/IntroToSVM.pdf.

[11] K. A. Boyden, "Blood count health article," 2002, retrieved November 5, 2007, from http://www.healthline.com/galecontent/blood-count.

[12] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[13] L. Breiman and A. Cutler, "Random forests," Jan. 2001, retrieved November 5, 2007, from http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.

[14] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and regression trees*. Wadsworth, 1984.

[15] A. Brookes, "Review: The essence of SNPs," *GENE*, vol. 234, no. 2, pp. 177–186, 1999.

[16] L. Carmel, S. Efroni, P. D. White, E. Aslakson, U. Vollmer-Conna, and M. S. Rajeevan, "Gene expression profile of empirically delineated classes of unexplained chronic fatigue," *Pharmacogenomics*, vol. 7, no. 3, pp. 375–386, Apr 2006.

[17] "Chronic fatigue syndrome: Possible causes," Centers for Disease Control and Preventation, May 2006, retrieved November 5, 2007, from http://www.cdc.gov/cfs/cfscauses.htm.

[18] G. Chu, B. Narasimhan, R. Tibshirani, and V. Tusher, "Significance analysis of microarrays, users guide and technical document," 2001, retrieved November 5, 2007, from http://www-stat.stanford.edu/~tibs/SAM/sam.pdf.

[19] A. J. Cleare, J. Miell, E. Heap, S. Sookdeo, L. Young, G. S. Malhi, and V. O'Keane, "Hypothalamo-pituitary-adrenal axis dysfunction in chronic fatigue syndrome, and the effects of low-dose hydrocortisone therapy," *The Journal of Clinical Endocrinology and Metabolism*, vol. 86, no. 8, pp. 3545–3554, 2001.

[20] R. Daz-Uriarte and S. A. de Andrs, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol. 7, no. 3, 2006.

[21] F. Emmert-Strieb, E. Glynn, C. Seidel, C. Bausch, and A. Mushegian, "Detecting pathological pathways of the chronic fatigue syndrome by the comparison of networks," in *Proceedings of Workshop: Critical Assessment of Microarray Data Analysis (*CAMDA*)*, 2006.

[22] H. Fang, Q. Xie, R. Boneva, J. Fostel, R. Perkins, and W. Tong, "Gene expression profile exploration of a large dataset on chronic fatigue syndrome," *Pharmacogenomics*, vol. 7, no. 3, pp. 429–440, Apr 2006.

[23] K. Fukuda, S. Straus, I. Hickie, M. Sharpe, J. Dobbins, and A. Komaroff, "The chronic fatigue syndrome: a comprehensive approach to its definition and study," *International Chronic Fatigue Syndrome Study Group*, vol. 121, no. 12, pp. 953–959, 1994.

[24] "An introduction to the gene ontology," The Gene Ontology, 2007, retrieved November 5, 2007, from http://www.geneontology.org/GO.doc.shtml.

[25] E. F. Glynn, "Understanding CAMDA '06 data: SNP data," January 2006, retrieved November 5, 2007, from http://research.stowers-institute.org/efg/2006/CAMDA/SNP.htm.

[26] E. Glynn, "Understanding CAMDA '06 data: Clinical data," Jan. 2006, retrieved November 5, 2007, from http://research.stowers-institute.org/efg/2006/CAMDA/Clinical.htm.

[27] B. Goertzel, C. Pennachin, L. Coelho, B. Gurbaxani, E. Maloney, and J. Jones, "Combinations of single nucleotide polymorphisms in neuroendocrine effector and receptor genes predict chronic fatigue syndrome," *Pharmacogenomics*, vol. 7, no. 3, pp. 475–483, Apr 2006.

[28] E. Gunther, D. Stone, R. Gerwien, P. Bento, and M. Heyes, "Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles in vitro," *Proc Natl Acad Sci*, vol. 100, no. 16, pp. 9608–9613, 2003.

[29] M. Hall and G. Holmes, "Benchmarking attribute selection techniques for discrete class data mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 6, pp. 1437–1447, 2003.

[30] G. Izmirlian, "Application of the random forest classification algorithm to a SELDI-TOF proteomics study in the setting of a cancer prevention trial," *Annals of the New York Academy of Science*, vol. 1020, pp. 154–174, 2004.

[31] T.-P. Jung, S. Makeig, T.-W. Lee, M. McKeown, G. Brown, A. Bell, and T. Sejnowski, "Independent component analysis of biomedical signals," *The 2nd Int'l Workshop on Indeppendent Component Analysis and Signal Separation*, pp. 633–644, 2000.

[32] M. Kantardzic, *Data Mining Concept, Models, Methods, and Algorithms.* IEEE Press, 2003.

[33] P. Kennedy, S. Simoff, D. Catchpoole, F. Ubaudi, A. Al-Oqaily, S. Yildiz, Y. Du, and D. Skillicorn, "Does CFS have a biological basis? - a constructionist approach," in *Proceedings of Workshop: Critical Assessment of Microarray Data Analysis (*CAMDA*)*, 2006.

[34] E. Lee, S. Cho, and T. Park, "Integration of expression data and genotype data: Application of chronic fatigue syndrome data," in *Proceedings of Workshop: Critical Assessment of Microarray Data Analysis (*CAMDA*)*, 2006.

[35] S. Lim, W. Le, P. Hu, B. Xing, C. Greenwood, and J. Beyene, "Integration of clinical, SNP, and microarray gene expression measurements in prediction of

chronic fatigue syndrome," in *Proceedings of Workshop: Critical Assessment of Microarray Data Analysis (*CAMDA*)*, 2006.

[36] S. Lin, J. Devakumar, and W. Kibbe, "Improved prediction of treatment response using microarrays and existing biological knowledge," *Pharmacogenomics*, vol. 7, no. 3, pp. 495–501, Apr 2006.

[37] E. Masood, "As consortium plans free SNP map of human genome," *Nature*, vol. 398, pp. 545–546, 1999.

[38] "Understanding your complete blood count," National institute of health, March 2006, retrieved November 5, 2007, from http://clinicalcenter.nih.gov/ccc/patient_education/pepubs/cbc97.pdf.

[39] A. Nicholson, E. Unger, R. Mangalathu, H. Ojaniemi, and S. Vernon, "Exploration of neuroendocrine and immune gene expression in peripheral blood mononuclear cells," in *Molecular Brain Research*, vol. 129, 2004, pp. 193–197.

[40] H. Pang, A. Lin, M. Holford, B. Enerson, B. Lu, M. Lawton, E. Floyd, and H. Zhao, "Pathway analysis using random forests classification and regression," *Bioinformatics*, vol. 22, no. 16, pp. 2028–2036, 2006.

[41] A. Presson, E. Sobel, J. Papp, A. J. Lusis, and S. Horvath, "Integration of genetic and genomic approaches for the analysis of chronic fatigue syndrome implicates Forkhead Box N1," in *Proceedings of Workshop: Critical Assessment of Microarray Data Analysis (*CAMDA*)*, 2006.

[42] I. Price, "Two-way analysis of variance," 2000, retrieved November 5, 2007, from http://www.une.edu.au/WebStat/unit_materials/c7_anova/twoway_anova.htm.

[43] C. Reeves, D. Wagner, R. Nisenbaum, J. Jones, B. Gurbaxani, L. Solomon, D. Papanicolaou, E. Unger, S. Vernon, and C. Heim, "Chronic fatigue syndrome - a clinically empirical approach to its definition and study," *BMC Medicine*, vol. 3, no. 19, 2005.

[44] M. Reimers and J. N. Weinstein, "Quality assessment of microarrays: Visualization of spatial artifacts and quantitation of regional biases," *BMC Bioinformatics*, vol. 1, no. 166, 2005.

[45] K. J. Reynolds, S. D. Vernon, E. Bouchery, and W. C. Reeves, "The economic impact of chronic fatigue syndrome," *Cost Effectiveness Resource Allocation*, vol. 2, no. 4, 2004.

[46] D. Skillicorn, *Understanding Complex Datasets: Data Mining using Matrix Decompositions*. CRC Press, June 2007.

[47] A. K. Smith, P. D. White, E. Aslakson, U. Vollmer-Conna, and M. S. Rajeevan, "Polymorphisms in genes regulating the HPA axis associated with empirically delineated classes of unexplained chronic fatigue," *Pharmacogenomics*, vol. 7, no. 3, p. 387394, 2006.

[48] L. I. Smith, "A tutorial on Principal Components Analysis," 2002, retrieved November 5, 2007, from http://kybele.psych.cornell.edu/~edelman/Psych-465-Spring-2003/PCA-tutorial.pdf.

[49] "Latent gold 4.0: About lc modeling," Statistical Innovations, 2003, retrieved November 5, 2007, from http://www.statisticalinnovations.com/products/latentgold_v4_aboutlc.html.

[50] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: a classification and regression tool for compound classification and QSAR modeling," *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 6, p. 19471958, 2003.

[51] S. Vernon and W. Reeves, "The challenge of integrating disparate high-content data: epidemiological, clinical and laboratory data collected during an in-hospital study of chronic fatigue syndrome," *Pharmacogenomics*, vol. 7, no. 3, pp. 345–354, Apr 2006.

[52] U. Vollmer-Conna, E. Aslakson, and P. White, "An empirical delineation of the heterogeneity of chronic unexplained fatigue in women," *Pharmacogenomics*, vol. 7, no. 3, pp. 355–364, Apr 2006.

[53] Z. Wang and J. Moult, "SNPs, protein structure, and disease," *Human Mutation*, vol. 17, pp. 263–270, 2001.

[54] P. D. Wentzell, T. K. Karakach, S. Roy, M. J. Martinez, C. P. Allen, and M. Werner-Washburne, "Multivariate curve resolution of time course microarray data," *BMC Bioinformatics*, vol. 7, no. 3, 2006.

[55] T. Whistler, R. Taylor, R. Craddock, G. Broderick, N. Klimas, and E. Unger, "Gene expression correlates of unexplained fatigue," *Pharmacogenomics*, vol. 7, no. 3, pp. 395–405, Apr 2006.

[56] B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, and H. Zhao, "Comparison of statistical methods for classification

of ovarian cancer using mass spectrometry data," *Bioinformatics*, vol. 19, no. 13, pp. 1636–1643, 2003.

Appendix A

The first 50 genes of the ranked-SVD lists

A.1 The first 50 genes of the ranked-SVD lists for the original microarray data

Table A.1: 106 patients

Gene ID [2]	Description
AK026288(3)	homo sapiens cdna: flj22635, clone hsi06649; unnamed protein product.
AF519531	small inducible cytokine a2 (monocyte chemotactic protein 1); scya2
BC013117	unknown (protein for mgc:8711)
BC009475(2)	chromosome 2 open reading frame 1
AJ133352	znf237 protein; znf237
AK021892(2)	cdna clone moderately similar to musculus praja1 (praja1) mrna; unnamed protein product.
XM66112	similar to ldl receptor precursor - rabbit (fragment); dkfzp761l0424
NM_002202(2)	islet-1; isl1
NM39289(2)	a kinase (prka) anchor protein 4, isoform 2; akap4
AF090897	pro0132

[2]The numbers appearing next to boldfaced Gene IDs indicate how many times each gene appears in the following six lists in Appendices A.1, A.2, A.3

D83077	tprd
NM_005910(3)	microtubule-associated protein tau, isoform 2; mapt
L01983(2)	sodium channel alpha subunit; scn4a
BC013951	unknown (protein for mgc:2815)
M61900	prostaglandin d synthase
NM_080760	dachshund homolog, isoform b; dach
XM_045472	similar to kiaa1271 protein; kiaa1271
AC006540	rna-binding protein nova-2 [aa 29-492]
NM_001981(2)	epidermal growth factor receptor pathway substrate 15; eps15
NM_001814(3)	cathepsin c; ctsc
XM71176	similar to calcium-promoted ras inactivator; loc257408
AL049704(4)	hypothetical protein
AK025611(3)	homo sapiens cdna: flj21958, clone hep05355; unnamed protein product.
AK001748(4)	cdna clone unnamed protein product.
NM34264(2)	socs box-containing wd protein swip-1 isoform 3; wsb1
AJ276171	aspic; aspic1
AK000642	cdna clone unnamed protein product.
AK026011(4)	homo sapiens cdna: flj22358, clone hrc06415; unnamed protein product.
AF023268	glucocerebrosidase; gba
AF015913(2)	skb1hs; skb1hs
NM_002403(2)	microfibrillar-associated protein 2 precursor; mfap2
AF145439(2)	cc chemokine receptor 9a; ccr9
AL096765	ba422a16.1 (e1a binding protein p300); ep300
BC000211(2)	eukaryotic translation elongation factor 1 beta 2
XM_043976(4)	similar to pi-3-kinase-related kinase smg-1, isoform 1; lambda/iota protein kinase c-interacting protein; phosphatidylinositol 3-kinase-related protein kinase; lat1-3tm
NM_007375(4)	tar dna binding protein; tardbp
NM_005893	calicin; ccin
AF083930(2)	es18
AF161526(2)	hspc178
NM_033666(3)	integrin beta 1 isoform 1b precursor; itgb1
AF494061(4)	adp-ribosylation-like factor 8
D90228(2)	mitochondrial acetoacetyl-coa thiolase precursor
AK027115(4)	homo sapiens cdna: flj23462, clone hsi08475; unnamed protein product.
BC000509(4)	proteasome (prosome, macropain) subunit, beta type, 7
AF010144(2)	neuronal thread protein ad7c-ntp
AF1297563(2)	lymphotoxin alpha
AK001842(3)	cdna clone unnamed protein product.
AK027558(4)	cdna clone unnamed protein product.
AF233395(2)	sir2-related protein type 7; sirt7
AK096396(3)	cdna clone highly similar to tbx15 protein; unnamed protein product.

Table A.2: 75 patients

Gene ID	Description
AF078776(3)	p53 tumor suppressor-binding protein 1
NM_002753	mitogen-activated protein kinase 10, isoform 1; mapk10
XM_038864	similar to kiaa1632 protein; kiaa1632
NM39289(2)	a kinase (prka) anchor protein 4, isoform 2; akap4
AC003109	xrcc2; xrcc2
AF015913(2)	skb1hs; skb1hs
AK026288(3)	homo sapiens cdna: flj22635 fis, clone hsi06649; unnamed protein product.
AL049704(4)	hypothetical protein
AF145439(2)	cc chemokine receptor 9a; ccr9
BC001528(3)	unknown (protein for mgc:3298)
AK002149	cdna clone weakly similar to vegetatible incompatibility protein; unnamed protein product.
AK091309	cdna clone unnamed protein product.
NM_001981(2)	epidermal growth factor receptor pathway substrate 15; eps15
AF1297563(2)	lymphotoxin alpha
NM38635	histone h2a.f/z variant, isoform 2; h2av
BC000211(2)	eukaryotic translation elongation factor 1 beta 2
AK021892(2)	cdna clone moderately similar to mus musculus praja1 mrna; unnamed protein product.
AC004472(2)	tera_human
AL109963	dj1188j21.1 (fsh primary response (lrpr1, rat) homolog 1); fshprh1
AF045937	deoxyribonuclease ii precursor
AF261689(2)	dna polymerase epsilon p17 subunit
AK000891(2)	cdna clone unnamed protein product.
NM_002403(2)	microfibrillar-associated protein 2 precursor; mfap2
D90228(2)	mitochondrial acetoacetyl-coa thiolase precursor
XM_091095	similar to sprouty protein with evh-1 domain 1, related sequence; sprouty protein with evh-1 domain 1; loc161742
AK023772	cdna clone weakly similar to f-spondin precursor; unnamed protein product.
NM_033666(3)	integrin beta 1 isoform 1b precursor; itgb1
AK025611(3)	homo sapiens cdna: flj21958 fis, clone hep05355; unnamed protein product.
NM_001814(3)	cathepsin c; ctsc
AK026011(4)	homo sapiens cdna: flj22358 fis, clone hrc06415; unnamed protein product.
XM_043976(4)	similar to pi-3-kinase-related kinase smg-1, isoform 1; lambda/iota protein kinase c-interacting protein; phosphatidylinositol 3-kinase-related protein kinase; lat1-3tm

U37194	unc-104- and kif1a-related protein
NM_032236(2)	flj23277 protein; flj23277
BC009475(2)	chromosome 2 open reading frame 1
NM_002202(2)	islet-1; isl1
L01983(2)	sodium channel alpha subunit; scn4a
AF083930(2)	es18
AK001748(4)	cdna clone unnamed protein product.
NM_007375(4)	tar dna binding protein; tardbp
AF161526(2)	hspc178
AL133353	ba483f11.3 (cgi-32 protein); ba483f11.3
AK001842(3)	cdna clone unnamed protein product.
AK027115(4)	homo sapiens cdna: flj23462, clone hsi08475; unnamed protein product.
AF233395(2)	sir2-related protein type 7; sirt7
AK027558(4)	cdna clone unnamed protein product.
AF010144(2)	neuronal thread protein ad7c-ntp
AF494061(4)	adp-ribosylation-like factor 8
NM_015072(2)	kiaa0998 protein; kiaa0998
BC000509(4)	proteasome (prosome, macropain) subunit, beta type, 7
AK096396(3)	cdna clone highly similar to tbx15 protein; unnamed protein product.

A.2 The first 50 genes of the ranked-SVD lists for the log-transformed microarray data

Table A.3: 106 patients

Gene ID	Description
AB040968	kiaa1535 protein; kiaa1535
NM_004949	desmocollin 2, isoform dsc2b preproprotein; dsc2
AE0064670	kiaa0683
AK021784	cdna clone unnamed protein product.
BC011958	unknown (protein for image:4158010)
NM_005932	mitochondrial intermediate peptidase; mipep
XM_035608	similar to ring finger protein 12 (lim domain interacting ring finger protein) (ring finger lim domain-binding protein) (r-lim) (ny-ren-43 antigen); loc145644
AF101044(2)	snrpn upstream reading frame protein; snurf
BC009849	unknown (protein for mgc:15400)
BC001841	unknown (protein for mgc:4473)
BC038239(2)	tyrosine kinase with immunoglobulin and epidermal growth factor homology domains
AK027115(4)	homo sapiens cdna: flj23462, clone hsi08475; unnamed protein product.
NM44490(2)	a kinase (prka) anchor protein 11, isoform 2; akap11
AF261689(2)	dna polymerase epsilon p17 subunit
AK027558(4)	cdna clone unnamed protein product.
NM_004420	dual specificity phosphatase 8; dusp8
AL032821(2)	dj55c23.1 (vanin 1); vnn1
NM_012404(2)	acidic nuclear phosphoprotein 32d; anp32d
AL049704(4)	hypothetical protein
XM71283	similar to heat shock 70kd protein 8; heat shock cognate protein 70; loc253590
AF016371(2)	u-snrnp-associated cyclophilin; usa-cyp
NM_003317(2)	thyroid transcription factor 1; titf1
AF353991(2)	bbp-like protein 1; blp1
AF229068	hspc170 protein
BC000035(2)	cgi-89 protein
AK026011(4)	homo sapiens cdna: flj22358, clone hrc06415; unnamed protein product.
AB037938	nonclathrin coat protein zeta-cop; copz2
NM_017482(2)	adducin 2, isoform b; add2
BC000651(2)	similar to solute carrier family 1 (glutamate transporter), member 7
AF316829(2)	fibronectin type 3 and spry domain-containing protein 1; fsd1
AB037757(2)	kiaa1336 protein; kiaa1336
NM_005910(3)	microtubule-associated protein tau, isoform 2; mapt

AK023794(2)	cdna clone moderately similar to tensin; unnamed protein product.
BC001528(3)	unknown (protein for mgc:3298)
AF078776(3)	p53 tumor suppressor-binding protein 1
AF316824(2)	asporin precursor; aspn
AF095906	transcription elongation factor a sii-like 1; tceal1
XM_098013(2)	hypothetical protein xp_098013; loc151174
BC000509(4)	proteasome (prosome, macropain) subunit, beta type, 7
AK092590(2)	cdna clone unnamed protein product.
XM_043976(4)	similar to pi-3-kinase-related kinase smg-1, isoform 1; lambda/iota protein kinase c-interacting protein; phosphatidylinositol 3-kinase-related protein kinase; lat1-3tm
NM_022791(2)	matrix metalloproteinase 19, isoform rasi-6; mmp19
BC010421(2)	dkfzp434f091 protein
AF494061(4)	adp-ribosylation-like factor 8
AK000764(2)	cdna clone unnamed protein product.
M74089(2)	tb1
AB002533(2)	qip1; qip1
BC036458(2)	similar to vesicular inhibitory amino acid transporter
AK001748(4)	cdna clone unnamed protein product.
NM_007375(4)	tar dna binding protein; tardbp

Table A.4: 75 patients

Gene ID	Description
AK027115(4)	homo sapiens cdna: flj23462, clone hsi08475; unnamed protein product.
AK023794(2)	cdna clone moderately similar to tensin; unnamed protein product.
AL049704(4)	hypothetical protein
BC014392	similar to s100 calcium-binding protein a1
U51007	26s protease subunit s5a
AK027558(4)	cdna clone unnamed protein product.
BC000035(2)	cgi-89 protein
BC000651(2)	similar to solute carrier family 1 (glutamate transporter), member 7
NM_032236(2)	flj23277 protein; flj23277
S77154	tinur
BC010047	protein inhibitor of activated stat protein piasy
NM_004524	lethal giant larvae (drosophila) homolog 2; llgl2
AK026011(4)	homo sapiens cdna: flj22358, clone hrc06415; unnamed protein product.
NM_012404(2)	acidic nuclear phosphoprotein 32d; anp32d
AF316829(2)	fibronectin type 3 and spry domain-containing protein 1; fsd1
NM_015072(2)	kiaa0998 protein; kiaa0998
AF101044(2)	snrpn upstream reading frame protein; snurf
AB037757(2)	kiaa1336 protein; kiaa1336
AF353991(2)	bbp-like protein 1; blp1

BC002616	transgelin 2
AF116608	pro0907
BC001528(3)	unknown (protein for mgc:3298)
AK000891(2)	cdna clone unnamed protein product.
AC004472(2)	tera_human
AB083586	putative g-protein coupled receptor; gpcr
AL032821(2)	dj55c23.1 (vanin 1); vnn1
AF494061(4)	adp-ribosylation-like factor 8
NM_003317(2)	thyroid transcription factor 1; titf1
XM_043976(4)	similar to pi-3-kinase-related kinase smg-1, isoform 1; lambda/iota protein kinase c-interacting protein; phosphatidylinositol 3-kinase-related protein kinase; lat1-3tm
NM_005910(3)	microtubule-associated protein tau, isoform 2; mapt
BC010421(2)	dkfzp434f091 protein
AF095906	transcription elongation factor a sii-like 1; tceal1
NM_001814(3)	cathepsin c; ctsc
AK000764(2)	cdna clone unnamed protein product.
BC038239(2)	tyrosine kinase with immunoglobulin and epidermal growth factor homology domains
BC036458(2)	similar to vesicular inhibitory amino acid transporter
NM_022791(2)	matrix metalloproteinase 19, isoform rasi-6; mmp19
NM44490(2)	a kinase (prka) anchor protein 11, isoform 2; akap11
XM_098013(2)	hypothetical protein xp_098013; loc151174
M74089(2)	tb1
AF016371(2)	u-snrnp-associated cyclophilin; usa-cyp
BC000509(4)	proteasome (prosome, macropain) subunit, beta type, 7
AK092590(2)	cdna clone unnamed protein product.
AK001748(4)	cdna clone unnamed protein product.
NM_001246	ectonucleoside triphosphate diphosphohydrolase 2; entpd2
NM_017482(2)	adducin 2, isoform b; add2
AB002533(2)	qip1; qip1
AF316824(2)	asporin precursor; aspn
AF078776(3)	p53 tumor suppressor-binding protein 1
NM_007375(4)	tar dna binding protein; tardbp

A.3 The first 50 genes of the ranked-SVD lists for the ICA-cleaned microarray data

Table A.5: 106 patients

Gene ID	Description
BC001000	unknown (protein for mgc:5363)
NM_004304	anaplastic lymphoma kinase ki-1; alk
XM_036558	similar to titin_ isoform n2-a; connectin; cmh9_ included; loc91156
AL359555	bk2308n23.1 (bactericidal/permeability-increasing protein); bpi
AF125182	single-strand selective monofunctional uracil dna glycosylase
AF058332	titin; ttn
AF092565(2)	splicing factor prp8
Y15572	r51h3
AK096396(3)	cdna clone highly similar to tbx15 protein; unnamed protein product.
AK000231	cdna clone unnamed protein product.
AF067147	uroplakin 1b; upk1b
XM66968	similar to data source:mgd_ source key:mgi:98158_ evidence:iss putative ribosomal protein s4_ x-linked; loc220433
BC007308	similar to riken cdna 1110033j19 gene
BC015374	similar to dolichyl-phosphate mannosyltransferase polypeptide 2_ regulatory subunit
AF016270	thyroid hormone receptor coactivating protein
BC037567	unknown (protein for mgc:45484)
NM_021114	serine protease inhibitor_ kazal type_ 2 (acrosin-trypsin inhibitor); spink2
M61156	activator protein 2b; ap-2b
XM70835	similar to syntaxin 18; loc254890
AK026288(3)	homo sapiens cdna: flj22635 fis_ clone hsi06649; unnamed protein product.
NM_033666(3)	integrin beta 1 isoform 1b precursor; itgb1
D87686	kiaa0017 protein; kiaa0017
AK023133	cdna clone moderately similar to heterogeneous nuclear ribonucleoprotein m; unnamed protein product.
NM34264(2)	socs box-containing wd protein swip-1 isoform 3; wsb1
S75762	fus-chop fusion protein; fus-chop
BC022090	nucleosome assembly protein 1-like 4
AF229179(2)	collectrin
XM_004009	similar to threonyl-trna synthetase_ cytoplasmic (threonine-trna ligase) (thrrs); loc94887
AB037808	kiaa1387 protein; kiaa1387
AF324241	epsin 3

BC035723(2)	complement component 6
NM47174	heparan sulfate 6-o-sulfotransferase 2; hs6st2
AK001842(3)	cdna clone unnamed protein product.
BC003551	similar to transglutaminase 2 (c polypeptide_ protein-glutamine-gamma-glutamyltransferase)
BC036828	unknown (protein for image:5247772)
AK025611(3)	homo sapiens cdna: flj21958, clone hep05355; unnamed protein product.
AB006590	estrogen receptor beta
S53268	ras-related protein mel; mel
AY122472(2)	defensin beta 119; defb119
BC000220(2)	unknown (protein for image:3353019)
AF449428	srrp35
NM_033100	mt-protocadherin precursor; kiaa1775
AY118228	nephrocystin-4; nphp4
AB021660(2)	carbonic anhydrase vb; ca 5b
NM_002312(2)	dna ligase iv; lig4
AK001998(2)	cdna clone unnamed protein product.
AK097525(2)	cdna clone unnamed protein product.
BC012082(2)	unknown (protein for mgc:19988)
NM_003162(2)	striatin_ calmodulin binding protein; strn
U66619(2)	swi/snf complex 60 kda subunit; baf60c

Table A.6: 75 patients

Gene ID	Description
BC018906	b-cell linker
AB083584	putative g-protein coupled receptor; gpcr
NM_002905	retinol dehydrogenase 5 (11-cisand 9-cis); rdh5
XM_043563	similar to insulin receptor-related protein precursor(ir-related receptor); insrr
AB026893	vascular cadherin-2
NM_013267	glutaminase ga, isoform a; ga
AK023447	cdna clone unnamed protein product.
AB024057	vascular rab-gap/tbc-containing protein; vrp
NM_006724	map/erk kinase kinase 4, isoform b; map3k4
BC002349	cd63 antigen (melanoma 1 antigen)
BC001191	unknown (protein for mgc:3162)
AB021660(2)	carbonic anhydrase vb; ca 5b
NM_005859	purine-rich element binding protein a; pura
XM_059923	similar to nuclear receptor binding factor-2; loc137829
NM_002312(2)	dna ligase iv; lig4
NM_006952	uroplakin 1b; upk1b
AY122472(2)	defensin beta 119; defb119

XM_030834	similar to f08f8.7.p; rpe
AC006538	small glutamine-rich tetratricopeptide (sgt)
XM_084866	similar to alpha tubulin; k-alpha-1
AB046614	nonmuscle myosin light chain 2; mlc-2
AK001998(2)	cdna clone unnamed protein product.
AF005482	histone deacetylase-3c
AK001358	cdna clone unnamed protein product.
BC032430	digeorge syndrome critical region gene 2
BC009901	unknown (protein for mgc:2242)
BC000220(2)	unknown (protein for image:3353019)
AF349446	urea transporter ut-a1; slc14a1
AF229179(2)	collectrin
K01566	mhc serum complement factor b
NM_015838	ficolin 2 isoform c precursor; fcn2
XM72829	hypothetical protein xp_172829; loc256112
NM_003162(2)	striatin, calmodulin binding protein; strn
BC035723(2)	complement component 6
AF034970	docking protein; dok-2
AB009249	fgf-17
AF092565(2)	splicing factor prp8
M96747	type l voltage-gated k+ channel of lymphocytes; kcnc1
BC017110	unknown (protein for mgc:16010)
NC_0018071	nadh dehydrogenase subunit 5; nd5
BC022816	unknown (protein for mgc:39248)
BC012082(2)	unknown (protein for mgc:19988)
BC014984	small nuclear rna activating complex, polypeptide 1, 43kdv
D38548	kiaa0076
U66619(2)	swi/snf complex 60 kda subunit; baf60c
AK097525(2)	cdna clone unnamed protein product.
BC014601	similar to riken cdna 1700017i11 gene
AF006386	axonemal dynein light chain; hp28
AC004774	dlx-6; dlx6
AB000114	osteomodulin