A NOVEL BAYESIAN RANK-BASED FRAMEWORK FOR THE CLASSIFICATION OF HIGH-DIMENSIONAL BIOLOGICAL DATA

A Dissertation

by

EMRE ARSLAN

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Ulisses M. Braga-Neto
Committee Members,	Edward Dougherty
	Erchin Serpedin
	Alan Dabney
Head of Department,	Miroslav M. Begovic

December 2018

Major Subject: Electrical Engineering

ABSTRACT

Statistical analysis of high-dimensional biological data is the central component of "personalized medicine" and "translational bioinformatics." Two major barriers limit the application of the extracted information in clinical studies. These barriers are small sample size and lack of biological interpretability due to the complex classification boundaries of current algorithms.

Motivated in removing these barriers, we focus in this dissertation to introduce novel statistical analysis algorithms of high-dimensional biological data. We first introduce a novel predictive model. In particular, we extend the top-scoring pair algorithm to a Bayesian setting. We test the performance on several real datasets and various simulated data scenarios and show the proposed method has the best overall performance. Besides having high accuracy rates on real and simulated data sets, the proposed algorithm has the potential to discover gene markers that may be missed via other algorithms.

We also suggested the Bayesian Top-Scoring Pair (BTSP) as a feature selection method. We compared the proposed algorithm with many well-known feature selection methods by combining the feature selection methods with different well-known classifiers. We checked the performance of all feature selection methods for different data sets and for different numbers of genes. The proposed BTSP algorithm has the best overall accuracy rates.

Finally, we introduce a novel biological pathway data-based algorithm (BTSPP). This algorithm uses all pairwise interactions in the gene level and pathway level. We apply the proposed method and well-known pathway data-based algorithms to different real data sets and check performances in terms of accurately classifying independent test sets and show the proposed algorithm superiority. We also checked the ability to find the biologically validated pathways related with diseases of these pathway data-based algorithms, over-representation analysis (ORA), and gene set enrichment analysis (GSEA). The proposed pathway analysis method has the potential to find the biologically validated pathways, whereas the others cannot detect the biologically validated pathways.

DEDICATION

To the Founder of Modern Turkey

Mustafa Kemal Ataturk

“Science is the most reliable guide in life.”
M. Kemal Ataturk

ACKNOWLEDGMENTS

First, I would like to express my gratefulness to my advisor Dr. Ulisses M. Braga-Neto for his invaluable mentorship throughout my doctoral program. Over the last four years, he taught me not only how to be a scientist, but also a decent human being.

I am grateful to Dr. Dougherty, Dr. Serpedin, and Dr. Dabney for being in my committee. I learned a lot from their lectures.

It was my luck to be surrounded by great folks. I had the pleasure of knowing Ahmet, Meltem, Arif, Sajad, Shahin, Shuilian, and Mahdi. Eunji, thank you for your computational power support.

Mom, Dad, Ulku, Fatih, and Mustafa thank you. Your sacrifices did not go to waste and I have made myself a useful person.

My wife, best friend, and companion. Thank you for being there all the time and for your endless love. You make me a better person every day.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Professor Braga- Neto and Professors Dougherty and Serpedin of the Department of Electrical & Computer Engineering and Professor Dabney of Department of Statistics. All work for the dissertation was completed by the student, under the advisement of Professor Braga-Neto of the Department of Electrical & Computer Engineering.

Funding Sources

Graduate study was supported by a research assistantship from the Center of Bioinformatics and Genomics Systems Engineering and teaching assistanship from Department of Electrical & Computer Engineering.

NOMENCLATURE

TSP	Top Scoring Pair
SVM	Support Vector Machine
k-NN	k-Nearest Neighbors
DT	Decision Trees
NB	Naive Bayes
BTSP	Bayesian Top Scoring Pair
ORA	Over-Representation Analysis
GSEA	Gene Set Enrichment Analysis
KEGG	Kyoto Encyclopedia of Genes and Genomes
MSigDB	Molecular Signatures Database
TSPP	Top-Scoring Pathway Pair
LLR	Log-Likelihood Ratio
R-LLR	Ranking Log-Likelihood Ratio
PathVar	Pathway Variance
GED	Gene Expression Deviation
RFE	Recursive Feature Elemination
BT	Bradley-Terry
IG	Information Gain
GR	Gain Ratio
NTC	Normal Tissue Centroit

TABLE OF CONTENTS

Page

ABSTRACT ii

DEDICATION iii

ACKNOWLEDGMENTS iv

CONTRIBUTORS AND FUNDING SOURCES v

NOMENCLATURE vi

TABLE OF CONTENTS vii

LIST OF FIGURES ix

LIST OF TABLES xi

1. INTRODUCTION 1

    1.1 Motivation 1
    1.2 Biological Background 3
    1.3 Dissertation Outline 5
    1.4 Summary of Main Contributions 6

2. TOP SCORING PAIRS ALGORITHMS 8

    2.1 TSP 8
    2.2 k-TSP 10
        2.2.1 Selecting k in k-TSP 11
            2.2.1.1 Cross Validation 12
            2.2.1.2 Variance Optimization 12

3. BAYESIAN TOP SCORING PAIRS CLASSIFIER 13

    3.1 Bradley-Terry Model 14
    3.2 Bayesian Approach for BT Model 15
    3.3 Bayesian Top Scoring Pairs 17
    3.4 k-BTSP 19
    3.5 Results 20
        3.5.1 Real Data 20
        3.5.2 Synthetic Data 23

3.5.3 Conclusion 27

4. BTSP AS A FEATURE SELECTION METHOD 30

4.1 Feature Selection 30
4.1.1 Recursive Feature Elimination (RFE) 31
4.1.2 χ^2-Statistics 32
4.1.3 Relief-F 32
4.1.4 Information Gain 33
4.1.5 Gain Ratio 33
4.2 BTSP as a Feature Selection Method 33
4.3 Results 34

5. A BAYESIAN PATHWAY DATA BASED CLASSIFIER 37

5.1 Bayesian Top Scoring Pathway Pairs 38
5.1.1 Skill Parameter Estimation of Genes 39
5.1.2 Pathway Mapping 39
5.1.3 Scoring of Pathway Pairs 39
5.1.4 Identification of Pathway Pairs & Classification Rule 40
5.2 Alternative Pathway-Based Algorithms 41
5.2.1 LLR 41
5.2.2 R-LLR 42
5.2.3 TSPP 44
5.2.4 Gene Expression Deviation 44
5.2.5 PathVar 45
5.3 Results 46
5.4 Conclusion 49

6. SUMMARY AND FUTURE DIRECTIONS 51

REFERENCES 53

LIST OF FIGURES

FIGURE Page

2.1 Gene expression values of a pair selected by the TSP algorithm. The training set has 15 normal samples and 15 cancer samples. 9

2.2 The number of votes for each class among k = 9 pairs selected by *k*-TSP algorithm. The training set has 15 normal samples and 15 cancer samples. 11

3.1 Overview of the BTSP algorithm. Gene expression values are used to estimate skill parameters for each gene in different biological phenotypes. BTSP scores are calculated for each pair of genes via estimated skill parameters. The pair having the highest BTSP score is used to estimate the given test sample. 19

3.2 Classification accuracy rates for colon data. The box plots show the distribution of the 100 runs for each classifier. Red starts are mean of the 100 runs. 24

3.3 Classification accuracy rates for leukemia$_1$ data. The box plots show the distribution of the 100 runs for each classifier. Red starts are mean of the 100 runs. 24

3.4 Classification accuracy rates for breast$_1$ data. The box plots show the distribution of the 100 runs for each classifier. Red starts are mean of the 100 runs. 25

3.5 Classification accuracy rates for leukemia$_2$ data. The box plots show the distribution of the 100 runs for each classifier. Red starts are mean of the 100 runs. 25

3.6 Classification accuracy rates for breast$_2$ data. The box plots show the distribution of the 100 runs for each classifier. Red starts are mean of the 100 runs. 26

3.7 A demonstration of four gene types' distribution in constructing the synthetic data. . . 26

4.1 Comparison between accuracy rates of the feature selection methods on colon data . . 35

4.2 Comparison between accuracy rates of the feature selection methods on leukemia data 35

4.3 Comparison between accuracy rates of the feature selection methods on lung data . . . 35

4.4 Comparison between accuracy rates of the feature selection methods on breast data . 36

5.1 Gene set enrichment analysis of *hsa05012* on GSE8671 data. 49

5.2 Gene set enrichment analysis of *hsa05012* on GSE9348 data. 50

LIST OF TABLES

TABLE Page

2.1 Contingency table for two genes ranking among two classes. 10

3.1 The datasets: Ten datasets involving two disease-related phenotypes, illustrating the "*small n, large d*" situation. The samples sizes for the two classes and the number of features/genes are shown in the table. 20

3.2 Comparison classification accuracy rates on real data. The highest accuracy for each dataset is highlighted in boldface. 21

3.3 P-values of three pairs of genes, having highest BTSP scores. 22

3.4 The comparison of BTSP and TSP scores for top nine disjoint pairs. The resolution of TSP scores is low, so k-TSP has to use the second score as a tie breaker, which lowers the performance. 23

3.5 Estimated classification accuracy for simulated datasets. Training sample size is 10. Distributions show different variance pair distributions. It is harder to differentiate higher variance pairs. The highest accuracy for each dataset is highlighted in boldface. 27

3.6 Estimated classification accuracy for simulated datasets. Training sample size is 20. Distributions show different variance pair distributions. It is harder to differentiate higher variance pairs. The highest accuracy for each dataset is highlighted in boldface. 28

3.7 Estimated classification accuracy for simulated datasets. Training sample size is 30. Distributions show different variance pair distributions. It is harder to differentiate higher variance pairs. The highest accuracy for each dataset is highlighted in boldface. 28

3.8 Estimated classification accuracy for simulated datasets. Training sample size is 40. Distributions show different variance pair distributions. It is harder to differentiate higher variance pairs. The highest accuracy for each dataset is highlighted in boldface. 29

4.1 Gene expression data sets 32

5.1 Estimated classification accuracy for eight real datasets. (KEGG database) 47

5.2 Estimated classification accuracy for eight real datasets. (MsigDB database) 47
5.3 Top ten-ranked pathways for the GSE8671 data according to BTSPP score. The boldface pathway is the target pathway..... 48
5.4 Top ten-ranked pathways for the GSE9348 data according to BTSPP score. The boldface pathway is the target pathway..... 50

1. INTRODUCTION

A new medicine trend is approaching for reducing the treatment costs and improving individual health via personalized medicine [1]. With this new perspective, two individuals may receive different treatments for the same disease regarding their different backgrounds. In order to play a proactive role, this new medicine approach needs to have predictive and preventive skills. These skills have been fueled by the high-throughput biological data and computational biological algorithms that extract meaningful patterns from these datasets.

New algorithms for analyzing high-dimensional biological data sets in a more accurate and meaningful way are expected to play an important role in personalized medicine. This dissertation focuses on the main topics in the analysis of high-dimensional biological data sets, namely: classification, feature selection, and pathway analysis. More specifically, we focus on gene expression analysis by using Bayesian approaches of ranking models. In this chapter, we first present the motivations of this dissertation, then introduce some biological background, provide an outline of the dissertation, and finally summarize the main contributions.

1.1 Motivation

Statistical methods that analyze high-dimensional biological datasets help discover a great deal of information in listing "marker genes" associated with cancers [2, 3, 4, 5] and explain biological processes. Although there are many proposed algorithms to find patterns in high-dimensional biological data, they do not offer a high degree of interpretability to medical researchers [6]. Authors [7] claim there are three barriers that should be handled before discovered biomarkers can be used in clinical research: technological, mathematical, and translational barriers.

The first problem concerning the technological side is the quality of high-dimensional data acquired by high-throughput technologies. There are two main problems that affect the quality of biological datasets: batch effects and lab effects. One can determine which lab created the data by finding patterns from previous datasets. It turns out that samples are not identically distributed

from different labs [8]. Since there is fast-paced progress in technology and getting data with higher quality, data quality of high throughput technologies is expected to become less problematic; however, batch and lab effects are still considerable problems. Most machine learning methods exhibit big performance changes due to pre-processing techniques and batch effects. Along with these problems, one should keep in mind that rank-based approaches, analyzing genomic data, are invariant under a monotonic transformation and reduce the batch effect.

Small sample size and a large number of features create the mathematical barrier. In contrast to many other applications of statistical learning, in most applications in medicine there are plenty of features/genes (~ 10.000) but very small sample sizes (~ 10). One cannot solve the problem by just lowering the cost of high-throughput technologies as there are many diseases with very low frequencies. This troublesome ratio between the number of features and samples is known as "*small n large d*" in statistics [9]. The proposed statistical method to analyze high-dimensional biological data should take this barrier into consideration in order to extract useful information. There is an obvious need for a reductionist approach to investigate high-dimensional datasets. While we are trying to extract the small informative portion of the full data, like in the definition of reductionism, we should not neglect the interactions between different pieces of the system.

The third and very important barrier is translational. In applications involving images or sounds, there is no need for transparency; however, in genomic studies, we should interpret the computational results from a biological perspective. Establishing a solid link with potential mechanisms is a must for drug development and clinical diagnosis [10]. Many causes of cancers, inherited or nonhereditary, can be discovered through biologically interpretable results of modern machine learning techniques. Unfortunately, the complex classification boundaries of the off-the-shelf algorithms, often used in genomic studies, make it hard to interpret the results. Examples of such predictive algorithms are: support vector machines [11, 12], neural networks[13, 3, 2], multiple decision trees [14, 15], and boosting [16, 17]. Most bioinformatics research areas suffer from the black-box approach of these predictive algorithms.

There are many ongoing research studies for the analysis of different kinds of high-dimensional

biological datasets, but this analysis is not an easy task as all of the different platforms may have different data types (continuous, discrete) and different scales. Rank-based methods have the potential to analyze different datasets, coming from different platforms regardless of the data type and range. Rank-based methods are also robust to deal with outliers.

Regarding the problems mentioned above, the proposed algorithm to analyze high-dimensional data should:

- Reduce the batch and lab effects.
- Lead to biological interpretations.
- Demonstrate that the proposed method can compete with other well-known statistical methods in terms of accuracy rates.
- Deal with the curse of dimensionality.
- Have the ability to hard-wire prior biological knowledge.

The proposed algorithms in this dissertation address these items. This dissertation considers the lack of interactions between genes or biological pathways and limitations of biological interpretations. We use *biological switches* between two genes or biological pathways in the simplest form, which allow for biological interpretation and use interactions to find the marker genes/pathways. The statistical methods that we propose deal with lab and batch effects, the curse of dimensionality, and transparency.

1.2 Biological Background

Deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and protein are essential for eukaryotic cells. DNA has the genetic instructions to regulate protein synthesis. DNA has its own alphabet to encode these genetic instructions: Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). All eukaryotic cells have almost the same DNA in the same living entity, which may be non-trivial to understand as different cells at different locations in the body exhibit different functionality.

RNA plays an important role in functional differentiation among cells. One can think of DNA as the whole instruction book where different cells use different chapters of this book. Encodable parts in this book are called genes. The amount of times that these encodable parts are used in a cell is called gene expression. RNA copies a small part of the whole DNA and leaves the nucleus. This process is called transcription and the copied unstable single-stranded RNA is called messenger RNA (mRNA). The alphabet of mRNA is Adenine (A), Cytosine (C), Guanine (G), and Uracil (U).

After leaving the nucleus, these messenger RNAs are translated into proteins. The alphabet for proteins is amino acids. Proteins are crucial for life as they are part of every process in cells. The whole process, starting with DNA and ending with proteins, is called *Central Dogma*. Proteins interact with each other and sometimes bind and create more complex structures. Analyzing the complex structures of the proteins is a whole different research area, which uses many optimization algorithms to find the optimum structure. The graphs that show interactions between proteins, are also well used in computational biology and are called protein-protein interaction (PPI) networks.

To understand the factors contributing to a disease or to diagnose an existing disease: blood tests, imaging techniques (MRI, CT, and PET), and body temperature readings/records are widely used. The most effective use of these biomarkers is to detect a severe disease at an early stage for an effective treatment/prevention. It is vital to be aware of risks before a tumor spreads, especially for cancers. Under the preventive medicine umbrella, biomarkers are great tools for risk indicators. They can help to prevent the onset of a disease.

Along with these conventional biomarkers, we can count gene expression levels, protein interactions with DNA or enzymes as molecular biomarkers. Analyzing molecular biomarkers presents great potential to tailor the treatment plan from a personalized medicine perspective. Analyzing molecular biomarkers is the only way for an early detection of some disease, like rheumatoid arthritis and Alzheimer's, as they have no symptom-phase in the early stage [18].

Besides using existing indicators of diseases, we have to discover new biomarker genes for many diseases and develop the drugs targeting these biomarker genes. There is no doubt high

dimensional biological data analysis will be the best choice for this task. Statistical bioinformatics methods, analyzing genomics, proteomics and transcriptomics have the potential to accelerate the discovery of new biomarkers. With the technological improvements, measuring tens of thousands of genes expression levels has raised hopes to obtain a more comprehensive understanding of the molecular mechanism, particularly under cancer conditions.

There are two main groups of genetic disorders: single gene disorders and polygenic disorders. For the first one, a mutation in a single gene causes the disease and for the latter one, a combined action of multiple genes causes the diseases. We need algorithms to analyze the genomic datasets and reveal which gene or genes are responsible for different diseases.

1.3 Dissertation Outline

This chapter covers the main motivations for analyzing high-dimensional biological data sets, biological backgrounds, and the main contributions of proposed algorithms. This dissertation is structured as follows:

In Chapter 2, we examine rank-based gene expression classifiers. The Top-Scoring Pair (TSP) algorithm and its well-known k-TSP extension. Mathematical equations and visual representations for TSP and k-TSP algorithms are provided.

In Chapter 3, we introduce a novel Bayesian rank-based gene expression algorithm. We provide a generalization of the proposed algorithm. We illustrate the interpretability and higher accuracy rates of the proposed algorithm with real and simulated datasets. We also discuss its consistency and ability to find cancer-related genes that cannot be detected with traditional approaches.

In Chapter 4, we covered the feature selection topic in high-dimensional biological datasets. We propose an algorithm as a feature selection method and compared its performance with well-known feature selection methods used in computational biology. We combined all feature selection methods with three well-known classifiers for a different number of genes, selected by those algorithms. We demonstrate the consistency, efficiency, and higher accuracy rates in the proposed algorithm.

Chapter 5 introduces a novel biological pathway data-based algorithm to predict a given test

sample. This novel method has the potential to discover new pathways related to diseases. We study its performance on real data sets and compare it with the proposed pathway based data analysis methods and show its superiority.

Finally, Chapter 6 discusses the dissertation's overall findings and concludes with future research directions.

1.4 Summary of Main Contributions

In brief, this dissertation brings forth the following novel research advancements:

- We propose the Bayesian Top-Scoring Pair (BTSP) Algorithm and define the BTSP score.
- The proposed algorithm takes all pairwise interactions of genes into consideration and creates interpretable results.
- Despite its simplicity in our experiments, the proposed algorithm obtains higher accuracy rates than the standard k-TSP, support vector machine, decision tree, k-nearest neighbor, and naive Bayes algorithms.
- We propose the application of the BTSP algorithm and its generalization k-BTSP as a feature selection method. We compared the performance of k-BTSP with other well-known feature selection methods, such as SVM-RFE, Gain Ratio, ReliefF, Information Gain, χ^2, combined with several classification algorithms. Most of the time the highest accuracy rates are provided by the k-BTSP algorithm.
- We propose a biological pathway data-based algorithm. This algorithm takes all interactions in the gene level and pathway level into consideration. We compare the proposed pathway data-based analysis method with well-known Log-Likelihood Ratio (LLR), Top-Scoring Pathways Pairs (TSPP), Pathway Variance (PathVar), Gene Expression Deviation (GED), and Ranking Log-Likelihood Ratio (R-LLR) algorithms, showing the introduced algorithm has higher performance and is more consistent.

- Future work considers extensions of the Bayesian approach to other rank-based classifiers, such as TST [19] and TSM [20].

2. TOP SCORING PAIRS ALGORITHMS

One of the core challenges in bioinformatics is to predict the phenotype (i.e., detecting cancer from normal samples) from a gene expression data. There are many statistical methods, proposed as predictive models in computational biology; however, very few of them are being used for clinical utility. The main problem for most of these decision rules is that they are too complex to allow for biological interpretations.

In order to solve these problems, Dr. Geman and his collaborators [21] proposed the Top-Scoring Pair (TSP) algorithm that tries to explain a disease with low complexity mechanistically. The algorithm takes rank into consideration and ignores raw expression values. It seeks rank switches across the phenotypes/labels. Its main advantage is to produce interpretable results by using few genes, yet resulting in powerful classification accuracy rates. This algorithm was used in many research studies, trying to find informative genes related to phenotypes [5, 22, 23, 24, 25].

2.1 TSP

Let **X** be a matrix representation of the gene expression profile and denote dimension by P x N, which stands for P genes and N profiles. X_i represents the i^{th} gene expression values. The probability of i^{th} gene expression values be less than j^{th} gene expression values in a class, $p_{ij}(c) = P(X_i < X_j \mid C = c)$, is the quantity of interest that we calculate for each pair. We will assume the class label set as c = $\{0, 1\}$, for the sake of simplicity. The score that the TSP classifier seeks to maximize is:

$$\widehat{\Delta}_{ij} = |\hat{p}_{ij}(0) - \hat{p}_{ij}(1)|, \tag{2.1}$$

where $\hat{p}_{ij}$ is the sample estimate of p_{ij}. TSP classifier calculates scores (2.1) for each pair of genes and takes the pair with the highest score to classify the given new sample. If there are several pairs, having the same best score, there is a second rank score to break the tie and get only one pair. The second score is:

$$\widehat{\Gamma}_{ij} = |\hat{\mu}_{ij}(0) - \hat{\mu}_{ij}(1)|, \tag{2.2}$$

$$\hat{\mu}_{ij}(c) = \frac{\sum_{n \in c} (R_{i,n} - R_{j,n})}{|c|}, c = 0, 1, \tag{2.3}$$

where $R_{i,n}$ is the rank of the i^{th} gene in sample n and $|c|$ is the number of samples of that class. Eventually, TSP classifier uses only one pair of genes to classify new samples. Let's assume that i^{th} and j^{th} gene present the highest TSP score and $\hat{p}_{ij}(0) > \hat{p}_{ij}(1)$; the TSP algorithm predicts a given test set according to the following criteria:

$$h_{TSP}(\mathbf{x}_{new}) = \begin{cases} C_0, & \text{if} \quad \mathbf{x}_{new,i^*} < \mathbf{x}_{new,j^*}, \\ C_1, & \text{otherwise.} \end{cases} \tag{2.4}$$

C_0 and C_1 are flipped for the decision rule if we have $\hat{p}_{i^*j^*}(0) \leq \hat{p}_{i^*j^*}(1)$.

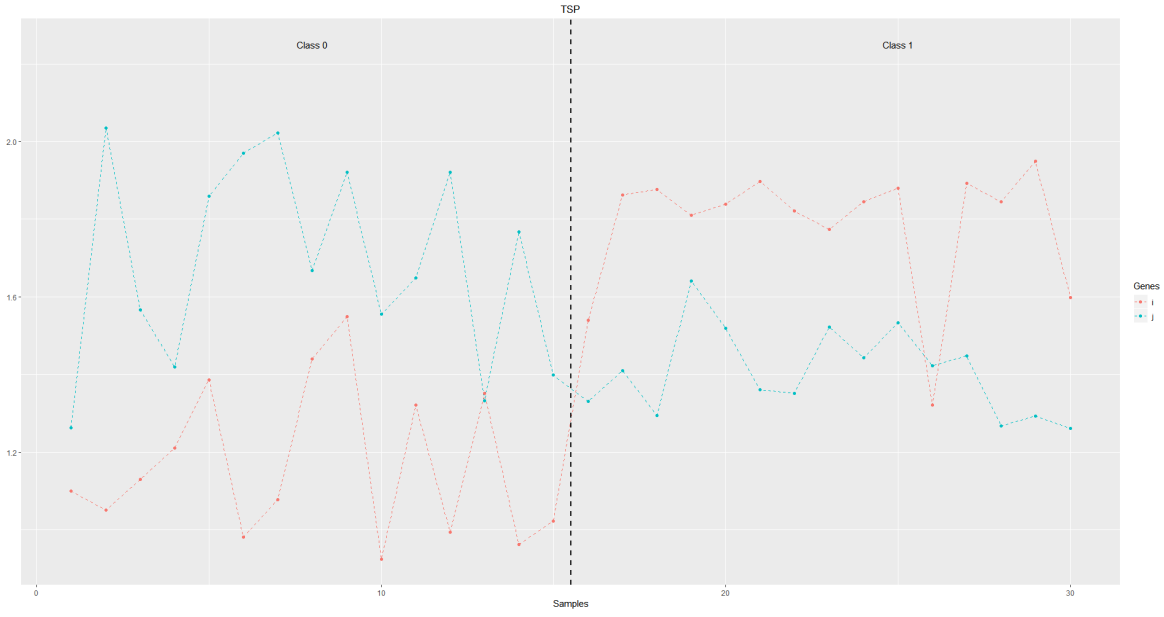

Figure 2.1: Gene expression values of a pair selected by the TSP algorithm. The training set has 15 normal samples and 15 cancer samples.

For a better understanding of the TSP algorithm, an illustration is provided. Let's assume we have 100 samples - 40 from class 0 (C_0) and 60 from class 1 (C_1). The contingency table (Table 2.1), showing the i^{th} gene expression is less than the j^{th} gene expression or vice versa is:

	$X_i \leq X_j$	$X_i > X_j$	Total
C_0	36	4	40
C_1	3	57	60

Table 2.1: Contingency table for two genes ranking among two classes.

The discrimination score $\widehat{\Delta}_{ij}$ can be calculated as:

$$\begin{aligned} \widehat{\Delta}_{ij} &= |\hat{p}_{ij}(0) - \hat{p}_{ij}(1)|, \\ &= |\frac{36}{40} - \frac{3}{60}| = 0.85 \end{aligned} \tag{2.5}$$

Let's assume we applied this process for all pairs and this is the highest score. The TSP classifier uses i^{th} and j^{th} genes for classification. Since $\hat{p}_{ij}(0)$ is greater than $\hat{p}_{ij}(1)$, the TSP classifier labels the given test sample as C_0 if $X_i < X_j$ or C_1 if $X_i \geq X_j$ for the given test sample.

An example of a TSP pair was illustrated in Figure 2.1. Each point on the x-axis represents a sample and gene expression values on the y-axis. It is clear we have a "reversal" ordering for different phenotypes.

2.2 k-TSP

The *k*-TSP classifier [26] is the most cited and applied extension of the TSP algorithm. The *k*-TSP classifier is a trivial generalization of TSP classifier, it uses k disjoint pairs, selected by TSP scores. Usually, an odd number k is selected and *k*-TSP performs the unweighted majority voting to classify the given data point:

$$h_{kTSP}(\mathbf{x}_{new}) = \underset{C\in\{C_0,C_1\}}{\arg\max} \sum_{r=1}^{k} I(h_r(\mathbf{x}_{new}) = C)\,, \tag{2.6}$$

where $h_r(\cdot)$ is the TSP classifier based on the pair r, for $r = 1, \ldots, k$.

If we take the previous example, after choosing i^{th} and j^{th} genes the *k*-TSP classifier searches for the second highest score, which doesn't include either i^{th} or j^{th} genes and uses the pair as the second one and keeps adding till it reaches k pairs.

A demonstration is provided in Figure 2.2. After applying TSP score (2.1) for each pair of genes, 9 disjoint pairs were selected (*k* = 9). Let $I = i_1, i_2, ...i_9$ and $J = j_1, j_2, ..., j_9$ represent these 9 pairs. The number of times $I > J$ and $J > I$ present a reversal attitude between phenotypes.

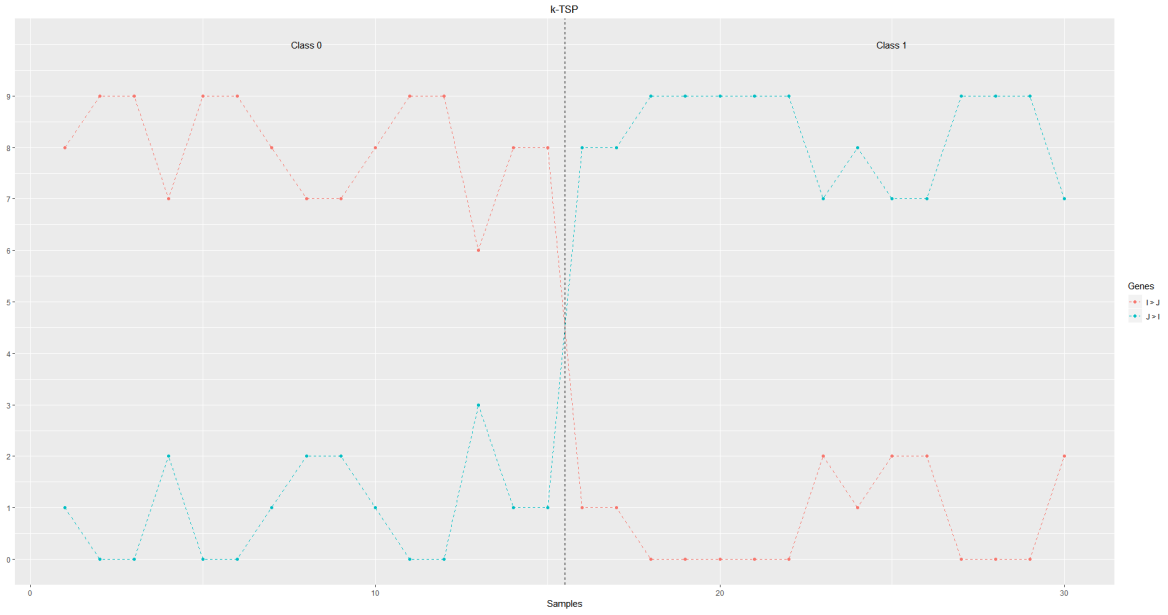

Figure 2.2: The number of votes for each class among k = 9 pairs selected by *k*-TSP algorithm. The training set has 15 normal samples and 15 cancer samples.

2.2.1 Selecting k in k-TSP

Selecting the number of pairs used in the *k*-TSP algorithm is an important problem. The number of informative genes for each training set is unknown and needs to be well defined. If we use fewer

genes than the true number, we may miss the pattern or if we use more genes than the classifier needs we may add noisy genes, causing worse performance. Two different approaches have been proposed to select *k* in *k*-TSP algorithm.

2.2.1.1 Cross Validation

The optimum k is defined by cross-validation [26]. If there are N samples with the given training set, the authors use m fold cross-validation where m = N / 3. *k*'s with the given range are used in each time and the *k* with best-averaged accuracy is chosen.

2.2.1.2 Variance Optimization

In this approach, the optimum k is defined by the one which maximizes the τ score. The score was defined in [20].

$$\begin{aligned}\widehat{\tau}_{kTSP} &= \frac{\widehat{\delta}_{kTSP}((i_1^*, j_1^*), \ldots, (i_k^*, j_k^*))}{\widehat{\sigma}_{kTSP}((i_1^*, j_1^*), \ldots, (i_k^*, j_k^*))} \\ &= \frac{\sum_{r=1}^{k} \widehat{\Delta}_{i_r^* j_r^*}}{\sqrt{\widehat{Var}(\sum_{r=1}^{k} [I(X_{i_r^*} < X_{j_r^*})] | Y = 0) + \widehat{Var}(\sum_{r=1}^{k} [I(X_{i_r^*} < X_{j_r^*})] | Y = 1)}},\end{aligned} \tag{2.7}$$

where $\widehat{\delta}_{kTSP}$ is the sum of top k-TSP scores and $\widehat{\sigma}_{kTSP}$ is the measure of the variances of these top k-TSP scores. This approach is much faster than the cross-validation approach.

3. BAYESIAN TOP SCORING PAIRS CLASSIFIER

With the ability of high-throughput technologies, high-dimensional biological data generation has been a common practice. Statistical analyzing methods, which deal with high-dimensional data, permeate the literature. Many studies analyzing high-dimensional gene expression data, helped to build predictive models by extracting compelling information from the data [27, 28, 29, 30, 31, 32, 23, 33]. From an expert point of view, it is crucial to identify informative/discriminative genes among thousands of candidates. Finding the optimal gene set is an N-P hard problem as there are 2^p possibilities for the given p genes, which are mostly thousands or even tens of thousands.

Most of the conventional classification algorithms are designed to analyze image/speech data or text mining but not for analyzing genomic datasets. These off-the-shelf methods (e.g., neural networks [13, 3], support vector machines (SVM) [11, 12], k-nearest neighbors (kNN) [34, 35], boosting [16], and naive Bayes (NB) [34]) when applied to biological datasets use non-linear functions to analyze the given data. These algorithms and their complex classification boundaries obstruct biological interpretations. There is not much hope in terms of biological interpretability and additional validations by experts due to the "black-box" approach employed by these non-linear functions.

The main problem with genomic datasets is that the number of features/genes greatly exceeds the number of samples. This problem is also known as the curse of dimensionality. Using all or many features of the data will cause overfitting, which results in worse classification accuracy for independent test data. Because of a large number of genes, it is not an easy task to remove irrelevant and redundant genes.

Ignoring the interactions between genes may cause the loss of inheritability [36]. In recent years, several methods have been proposed to find informative genes and classify cancers according to interactions between genes, such as Top-Scoring Pair classifier [21], doublets [37], binary matrix shuffling filter (BMSF) [38], etc.

The aim of this study is to design a novel classification method that presents high classification

accuracy rates for small samples, yet the decision rule allows us to make biological interpretations. We extend the TSP classification approach to a Bayesian setting. We use the Bradley-Terry model's [39] Bayesian extension proposed by Caron and Doucet [40] to rank data. The skill parameters that we get via the Bayesian Bradley-Terry model are used in the Bayesian TSP score that we define. We select an informative pair of genes according to this Bayesian TSP score and define the Bayesian TSP classifier.

We compared the accuracy of the proposed algorithm with TSP family classifiers and well-known machine learning methods used to analyze genomic datasets. 10 gene-expression datasets from various studies and synthetic data with different scenarios were used. The algorithm that we proposed shows the best overall accuracy rates, both for real and simulated data.

3.1 Bradley-Terry Model

Ranked data analysis is being used in many different research areas, such as marketing, psychology, behavioral studies, evaluating players' and teams' performances in different sports, political surveys, etc. A wide variety of methods have been proposed in paired comparisons to rank data. A bibliography about paired comparisons mentions more than a hundred entries as early as 1976 [41]. The Bradley-Terry (BT) Model is one of the most well-known and fundamental methods to rank data via paired comparisons. It has been introduced by Ralph Bradley and Milton Terry [39]. The worth or merit of a unit is measured through paired comparisons. Let i and j be 2 individuals from a group, Bradley and Terry [39] introduced the following model:

$$\pi_{ij} \;=\; Pr(i \; beats \; j \mid \vartheta_i, \vartheta_j) \;=\; \frac{\vartheta_i}{\vartheta_i + \vartheta_j}\,, \tag{3.1}$$

where parameter ϑ_i is a positive-valued skill/strength of individual i. It is trivial that π_{ij} + π_{ji} = 1. A nonlinear representation of the skill parameters, $\vartheta_i = e^{\beta_i}$, allows for a logistic representation of the BT Model:

$$\pi_{ij} \;=\; \frac{1}{1 + e^{-(\beta_i - \beta_j)}} \;=\; \eta(\beta_i - \beta_j) \tag{3.2}$$

where $\eta(.)$ represents the inverse-logit function [42]. Various applications leverage the BT model. For instance, Chess and GO players ranking by international federations, multiclass classification based on binary classifiers [43], influence rankings of journals and many more use the BT model.

Let w_{ij} represent number i ranks higher than j, w_i is the total number of "wins" of i against all other individuals, and n_{ij} denotes total number of times individual i compared with individual j, $n_{ij} = w_{ij} + w_{ji}$. We can leverage defined statistics to estimate unobserved individual skill parameters, $\{\vartheta_i\}$. The log-likelihood function of the BT model is:

$$\begin{aligned} \ell(\vartheta) &= \sum_{1\le i\neq j\le M} [w_{ij}\log\vartheta_i - w_{ij}\log(\vartheta_i+\vartheta_j)] \\ &= \sum_{i=1}^{M} w_i\log\vartheta_i - \sum_{1\le i<j\le M} n_{ij}\log(\vartheta_i+\vartheta_j)\,. \end{aligned} \tag{3.3}$$

Hunter [44] proposed Minorization-Maximization (MM) algorithms to perform ML estimator for the BT Model:

$$\vartheta_i^{(k+1)} = w_i\left(\sum_{i\neq j}\frac{n_{ij}}{\vartheta_i^{(k)}+\vartheta_j^{(k)}}\right)^{-1}, \quad i=1,\ldots,M, \tag{3.4}$$

which is repeated until convergence. Caron and Doucet [40] showed that the MM algorithm is indeed a special case of Expectation-Maximization (EM) algorithms. The main problem with the frequentist approach is its strong assumption that no player or group of players may win or lose all the times against others.

3.2 Bayesian Approach for BT Model

There have been some studies, performing Bayesian inference for generalized BT Models, in order to overcome the difficulties of the frequentist approach to estimate skill parameters. The posterior density is not tractable most of the time, and it needs to be approximated. Hussain and Aslam [45] offered a Bayesian approach for the log-linear version of the BT Model via Jeffreys Prior. An Expectation-Propagation (EP) algorithm was proposed by Guiver and Snelson [46]. It

uses a functional approximation of the posterior; however, it is not clear about its convergence properties. Metropolis-Hasting (MH) algorithms have been proposed by Adams [47] and Gormley and Murphy [48].

Caron and Doucet [40] suggested a Gibbs sampler method to estimate the parameters $\{\vartheta_i\}$. They assumed the prior for $\boldsymbol{\vartheta}$ as independent Gamma distribution, like in [49, 46]:

$$p(\boldsymbol{\vartheta}) \;=\; \prod_{i=1}^{M} \text{Gamma}(\vartheta_i; b, d)\,, \tag{3.5}$$

where b and d are the shape and rate hyperparameters.

The BT model enjoys the following "Thustonian" interpretation. Consider two independent exponential variables, $U_{li} \sim \text{Exp}(\vartheta_i)$ and $U_{lj} \sim \text{Exp}(\vartheta_j)$, for each individual pair, i and j, where $l = 1, 2, ..., M$ represents every single comparison between these two individuals. The probability of U_{li} is less than U_{lj} depends only on rate parameters:

$$P(U_{li} < U_{lj}) = \frac{\vartheta_i}{\vartheta_i + \vartheta_j}\,. \tag{3.6}$$

Instead of using these latent variables, Caron and Doucet introduced a new latent variable for each individual pair, i and j, to infer a simpler log-likelihood:

$$Z_{ij} \;=\; \sum_{l=1}^{n_{ij}} \min\{U_{lj}, U_{li}\}\,. \tag{3.7}$$

These latent variables make it easy to obtain a complete log-likelihood. Since $\min\{U_{lj}, U_{li}\} \sim \text{Exp}(\vartheta_i + \vartheta_j)$ and an exponential family property claims that summation of identically distributed exponential variables are Gamma distributed, we conclude that $Z_{ij} \sim \text{Gamma}(n_{ij}, \vartheta_i + \vartheta_j)$. Shape and rate parameters of this Gamma distribution are n_{ij} and $\vartheta_i + \vartheta_j$, respectively. The resulting density of the latent variables, $\mathbf{Z}$, conditional upon the data, $\mathcal{D}$, and skill parameters, $\boldsymbol{\vartheta}$, is:

$$p(\mathbf{z} \mid \mathcal{D}, \boldsymbol{\vartheta}) \;=\; \prod_{\substack{1 \le i < j \le M \\ \text{s.t. } n_{ij} > 0}} \text{Gamma}(z_{ij}; n_{ij}, \vartheta_i + \vartheta_j) \tag{3.8}$$

which we can use to obtain the complete log-likelihood:

$$\ell_c(\boldsymbol{\vartheta}) = \sum_{\substack{1\le i\neq j\le M \\ \text{s.t. } w_{ij}>0}} w_i \log(\vartheta_i) - \sum_{\substack{1\le i<j\le M \\ \text{s.t. } n_{ij}>0}} [(\vartheta_i+\vartheta_j)z_{ij} - (n_{ij}-1)\log z_{ij} + \log\Gamma(n_{ij})] \tag{3.9}$$

where $\Gamma(.)$ is the gamma function.

A Gibbs sampling strategy was proposed in [40] for sampling from the posterior $p(\boldsymbol{\vartheta}, \mathbf{Z} \mid \mathcal{D})$. Updating latent variables, $\mathbf{Z}$, conditional upon skill parameters, $\boldsymbol{\vartheta}$, can be carried out by using (3.8):

$$Z_{ij}^{t+1} \mid \mathcal{D}, \boldsymbol{\vartheta}^t \sim \text{Gamma}(n_{ij}, \vartheta_i^t + \vartheta_j^t) \tag{3.10}$$

It is easy to sample the skill parameters, $\boldsymbol{\vartheta}$, from the conditional probability $p(\boldsymbol{\vartheta} \mid \mathbf{Z}, \mathcal{D}) \propto p(\boldsymbol{\vartheta})\ell(\boldsymbol{\vartheta}, \mathbf{z})$, as it can be expressed easily with the conjugate prior for the complete likelihood:

$$\vartheta_i^{t+1} \mid \mathcal{D}, \mathbf{Z}^{t+1} \sim \text{Gamma}\left(b + w_i,\ d + \sum_{\substack{i<j \\ \text{s.t. } n_{ij}>0}} z_{ij}^{t+1} + \sum_{\substack{i>j \\ \text{s.t. } n_{ij}>0}} z_{ji}^{t+1}\right). \tag{3.11}$$

There are other generalizations of the Bayesian BT Model in the paper [40] such as: models taking home advantages into consideration, and models with ties and group comparisons to rank individuals. We will not mention about these generalizations as we are using the previous Gibbs sampler model. We also tried the model with ties; however, we did not see much improvement in the classification performance.

3.3 Bayesian Top Scoring Pairs

We use the Bayesian BT model to design a top scoring pairs classifier. $\boldsymbol{\vartheta}^0$ represent skills sampled from the posterior distribution of genes in class 0, and $\boldsymbol{\vartheta}^1$ denote skills sampled from the posterior distribution of genes in class 1. The probabilities of the i^{th} gene expression being greater than the j^{th} gene expression in class 0 and class 1 are:

$$\pi_{ij}^0 \;=\; \frac{\vartheta_i^0}{\vartheta_i^0+\vartheta_j^0} \quad \text{and} \quad \pi_{ij}^1 \;=\; \frac{\vartheta_i^1}{\vartheta_i^1+\vartheta_j^1} \tag{3.12}$$

The main goal is to find the most informative/discriminative genes among different phenotypes. For this purpose, one should find the biggest flip of the probability of the i^{th} gene expression being greater than the j^{th} between classes. We introduce the *Bayesian Top-Scoring Pair (BTSP) score* to answer this question:

$$\Omega_{ij} = \left|\pi_{ij}^0 < \pi_{ij}^1\right| \;=\; \left|\frac{\vartheta_i^0}{\vartheta_i^0+\vartheta_j^0} - \frac{\vartheta_i^1}{\vartheta_i^1+\vartheta_j^1}\right|, \tag{3.13}$$

and choose the pair maximizing the BTSP score,

$$(i^*, j^*) \;=\; \underset{(i,j)\in P}{\arg\max}\, \Omega_{ij} \tag{3.14}$$

where P represents the whole gene space. If $\pi_{i^*j^*}^0 < \pi_{i^*j^*}^1$, the (*BTSP*) classifier predicts a given test sample as:

$$h_{BTSP}(\mathbf{x}_{new}) = \begin{cases} C_0, & \text{if } \mathbf{x}_{new,i^*} < \mathbf{x}_{new,j^*}, \\ C_1, & otw. \end{cases} \tag{3.15}$$

If $\pi_{i^*j^*}^0 \geq \pi_{i^*j^*}^1$, the (*BTSP*) classifier predicts a given test sample as:

$$h_{BTSP}(\mathbf{x}_{new}) = \begin{cases} C_0, & \text{if } \mathbf{x}_{new,i^*} > \mathbf{x}_{new,j^*}, \\ C_1, & otw. \end{cases} \tag{3.16}$$

Figure 3.1 summarizes the BTSP algorithm.

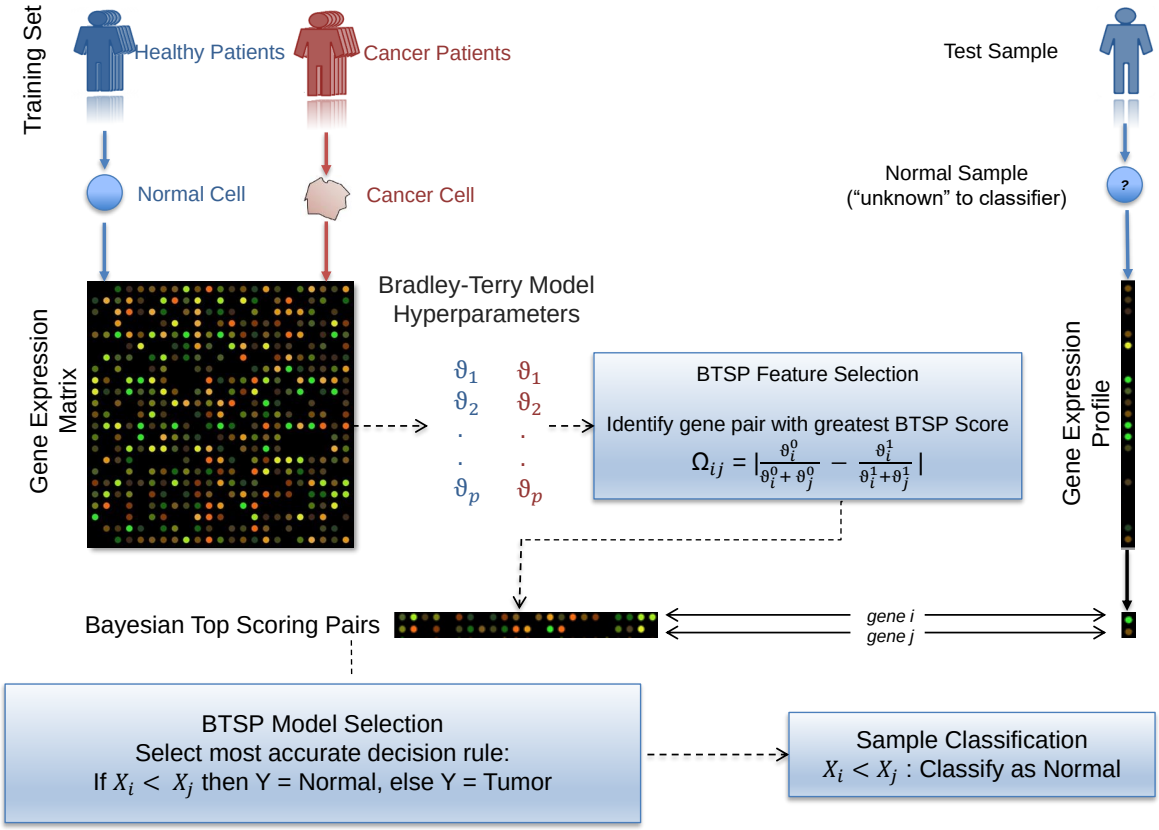

Figure 3.1: Overview of the BTSP algorithm. Gene expression values are used to estimate skill parameters for each gene in different biological phenotypes. BTSP scores are calculated for each pair of genes via estimated skill parameters. The pair having the highest BTSP score is used to estimate the given test sample.

3.4 k-BTSP

The BTSP classification rule can be extended by choosing more than one pair. We choose an odd number k of disjoint top scoring pairs according to Ω_{ij} and define a classifier as in (2.6). $h_r(\cdot)$ denotes the BTSP classifier based on pair r, for $r = 1, \ldots, k$. We call this the *k-BTSP classifier*.

Dataset	Number of Genes	Class I Size	Class II Size	References
Colon	2000	22	40	Alon et al. (1999) [50]
Leukemia_1	7129	25	47	Golub et al. (1999) [27]
DLBCL	7129	58	19	Shipp et al. (2002) [28]
Lung	12,533	150	31	Gordon et al. (2002) [29]
Breast_1	22,283	62	42	Chowdary et al. (2006) [30]
Leukemia_2	12,564	24	24	Armstrong et al. (2002) [31]
Squamous	12,625	22	22	Kuriakose, Chen et al. (2004) [32]
CNS	7129	25	9	Pomeroy et al. (2002) [51]
Myeloma	12,625	137	36	Tian et al. (2003) [52]
Breast_2	22215	43	75	Chin et al. [53]

Table 3.1: The datasets: Ten datasets involving two disease-related phenotypes, illustrating the "*small n, large d*" situation. The samples sizes for the two classes and the number of features/genes are shown in the table.

3.5 Results

3.5.1 Real Data

10 genomic datasets (Table 4.1) from different studies are used to compare the proposed classifier's performance against the conventional TSP family classifiers and other off-the-shelf machine learning methods, which are used in analyzing genomic datasets. We first employ the variance filter in order to reduce the number of genes to 2000 for all datasets.

Since we investigate the small sample size performances of all classifiers, a small part of the data was used as a training set while the rest was used as the test set. 20% of each data is used to train classifiers and performances of these classifiers are tested in the remaining 80% of each data. We repeat this procedure 100 times and record the average accuracy for all classifiers.

Estimated classification accuracy rates are presented in Table 3.2. The proposed Bayesian-based Top Scoring Pair approach and its generalization are shown in BTSP and *k*-BTSP columns accordingly. TSP and *k*-TSP represent the Top Scoring Pair and its generalization to top disjoint k pairs. We applied Decision Tree (DT), Naive Bayes (NB), Support Vector Machine (SVM) and k Nearest Neighbor (kNN) algorithms as predictive models.

One should compute the skill parameters of all genes before applying BTSP and *k*-BTSP algo-

Dataset	TSP	k-TSP	DT	BTSP	k-BTSP	NB	SVM	k-NN
Leukemia_1	0.881	0.8968	0.7862	0.8713	**0.9222**	0.8305	0.89	0.7989
Colon	0.714	0.746	0.6586	0.7446	**0.8072**	0.657	0.7506	0.7366
DLBCL	0.7783	0.8429	0.7508	0.7714	0.8682	0.7798	**0.8771**	0.7601
Lung	0.9504	0.971	0.9299	0.9368	0.979	0.9626	**0.985**	0.9518
Breast_1	0.9297	0.9457	0.8692	0.9359	**0.974**	0.9192	0.9301	0.9256
Leukemia_2	0.8884	0.9315	0.8328	0.8711	**0.9582**	0.8744	0.9531	0.9468
Squamous	0.8042	0.8085	0.7122	0.7505	0.8454	0.6362	**0.8768**	0.7977
CNS	0.6992	0.7251	0.5966	0.6796	**0.7522**	0.7215	0.7114	0.7070
Myeloma	0.6695	0.7318	0.6903	0.6793	0.749	0.7878	0.7592	**0.7885**
Breast_2	0.7829	0.8313	0.7468	0.7731	**0.8672**	0.8495	0.8442	0.8524
Average	0.8098	0.8431	0.7573	0.8014	**0.8722**	0.8019	0.8559	0.8265

Table 3.2: Comparison classification accuracy rates on real data. The highest accuracy for each dataset is highlighted in boldface.

rithms. We run 1700 iterations, with 400 burn-in iterations, in (3.10, 3.11) . To construct the prior (3.5) for skills, we tried fixed hyperparameters and a Metropolis-Hasting algorithm; however, in both cases, the accuracies did not change much.

Choosing k for the k-TSP algorithm is another important point. We applied both variance optimization [20] and cross validation [54] methods to choose optimum k in the k-TSP algorithm. The k-TSP column is created by choosing the maximum accuracy of variance optimization, cross-validation, and fixed 9 pairs methods in Table 3.2. Even though the k-TSP classifier has the freedom to choose optimum k up to 9, we fixed k to 9 in the k-BTSP classifier. If one fixes the k to 9 in k-TSP the average accuracy rate for the k-TSP column will be 2% less. The SVM [55] classifier uses the top 100 genes selected by the recursive feature elimination (RFE) method. If we use around 20 genes as we do for the k-BTSP classifier, the average accuracy rate for SVM drops around 2-3%. kNN uses three neigbours and NB uses all genes. The algorithm we proposed, k-BTSP, has the highest average accuracy rate. k-BTSP has the highest accuracy rate six times out of ten data sets and the second highest accuracy rate three times out of ten data sets.

Using individual gene ranking [56] does not guarantee to have the most informative genes as it does not take redundancy and interactions into consideration. As a consequence of these

approaches, selected features may have low power. An example is shown in Table 3.3. The k-BTSP classifier is trained by 20% of the Colon data [50]. Three pairs of genes, having highest BTSP scores, are represented in Table 3.3. Four out of six genes fail to reject the null hypothesis that different class samples' means are the same at the default 5% significance level. So an individual gene ranking method does not take these 4 genes into consideration as there is not much change between classes. On the other hand, the k-BTSP classifier detects these genes and gets a high accuracy rate for the given new samples. The gene with a 0.8752 p-value is GH1 and it is shown that this gene is related to colon cancer [57, 58].

Entrez ID	p-value
493	0.0457
576	0.2198
1727	0.8752
1102	0.3321
1791	0.3986
251	0.0042

Table 3.3: P-values of three pairs of genes, having highest BTSP scores.

The k-BTSP algorithm preserves the k-TSP's advantages. A remarkable advantage is to prevent experimental system differences, such as batch effects and background deviations. One of the main advantage of the k-BTSP algorithm compared to k-TSP is that k-TSP checks just pairwise interactions to calculate TSP scores while k-BTSP checks all pairwise interactions to get skill parameters for each gene and calculate BTSP scores.

If we have few samples, the TSP score (2.1) to select pairs in the k-TSP classifier does not have high resolution where the BTSP score (3.13) has better resolutions. Table 3.4 shows the highest 9 TSP and BTSP scores of disjoint pairs. Since there are many pairs having the same TSP score the k-TSP algorithm checks for the second score for many pairs. One may think to use the second

BTSP Score	TSP Score
0.9013	1
0.8742	1
0.8711	1
0.8707	1
0.8628	1
0.8362	1
0.8327	1
0.8269	1
0.8241	1

Table 3.4: The comparison of BTSP and TSP scores for top nine disjoint pairs. The resolution of TSP scores is low, so k-TSP has to use the second score as a tie breaker, which lowers the performance.

TSP score (2.2) first instead of the first TSP score (2.1). This practice will lower the accuracy rates of classification of the independent test sets. So deciding pairs based on the second score is not a good choice compared to the first score; however, the TSP algorithm has to use the second score most of the time as the resolution of the score is low.

We also compared the consistency of k-TSP and k-BTSP algorithms. It is important to check if the proposed algorithm uses very different genes for every different training set selected from the same data. For this reason, we checked the total number of different genes that k-TSP and k-BTSP are using to classify the Colon data [50]. We again used 20% of the data as the training set and repeated the procedure 100 times. The k-BTSP algorithm uses 332 different genes out of 2000 genes while the k-TSP algorithm uses 488 different genes out of 2000 genes. So the k-TSP algorithm uses 50% more genes than k-BTSP.

3.5.2 Synthetic Data

In order to understand the accuracy trends in synthetic data, we follow the settings in [59] that considers four feature/gene types in constructing the synthetic data. As Figure 3.7 shows, global markers differ between classes, heterogeneous markers change within a class and a bigger part of

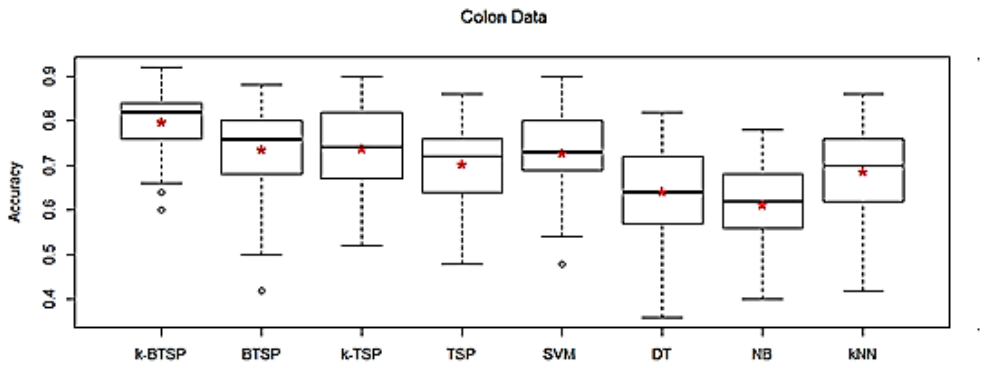

Figure 3.2: Classification accuracy rates for colon data. The box plots show the distribution of the 100 runs for each classifier. Red starts are mean of the 100 runs.

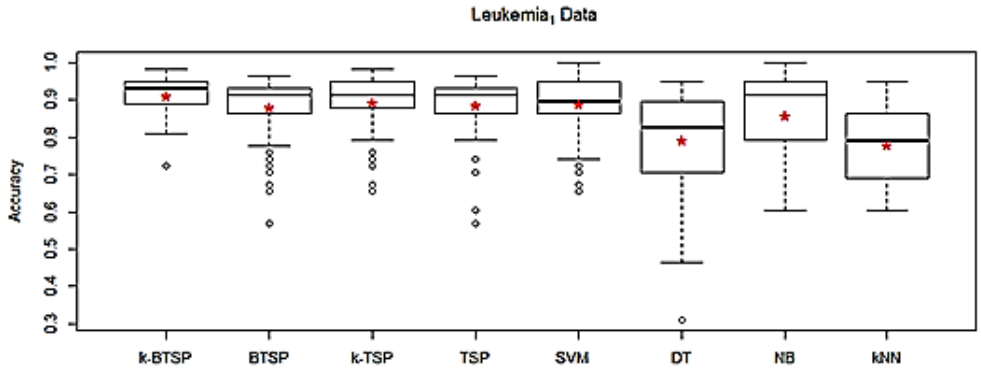

Figure 3.3: Classification accuracy rates for leukemia$_1$ data. The box plots show the distribution of the 100 runs for each classifier. Red starts are mean of the 100 runs.

the data contains high-variance and low variance non-markers. As it can be seen from the Figure 3.7, the majority of genes are low-variance and high-variance non-markers in the simulated data setup. We used 3 different scenarios to choose the numbers of global and heterogenous markers:

- $D_{gm} = 10$ / $D_{hm} = 45$
- $D_{gm} = 20$ / $D_{hm} = 40$
- $D_{gm} = 30$ / $D_{hm} = 35$

where D_{gm} represents the number of global marker genes and D_{hm} represents the number of

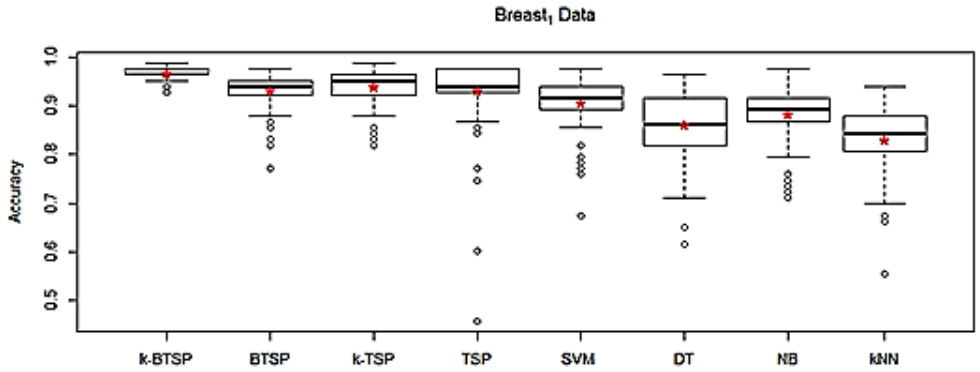

Figure 3.4: Classification accuracy rates for breast$_1$ data. The box plots show the distribution of the 100 runs for each classifier. Red starts are mean of the 100 runs.

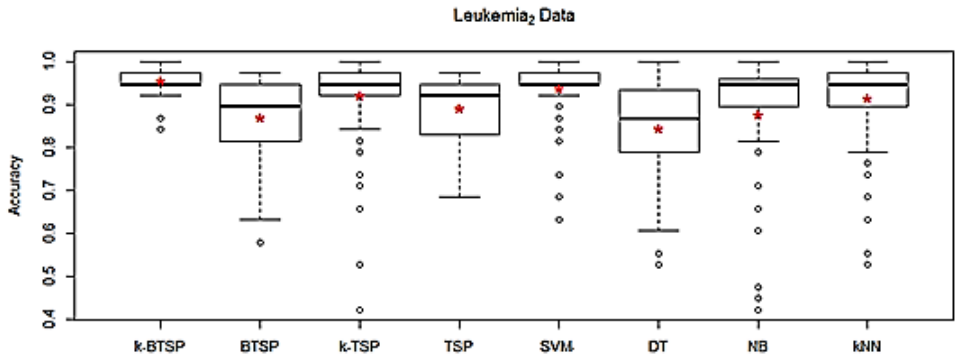

Figure 3.5: Classification accuracy rates for leukemia$_2$ data. The box plots show the distribution of the 100 runs for each classifier. Red starts are mean of the 100 runs.

heterogenous markers. We selected 3900 high-variance non-markers and 16000 low-variance non-markers.

We applied the same procedure that we followed for the real data. A variance filter is applied first and we used top 2000 genes for all classifiers. The k-TSP algorithm has the freedom to choose up to nine pairs. The k-TSP accuracy rates represent the highest value of variance optimization, cross-validation and fixed $k = 9$. If we use the fixed $k = 9$ like we do in the k-BTSP, the accuracy rates of k-TSP will drop around 2%. SVM uses top 20 genes selected by recursive feature elimination method. k-NN uses three nearest neighbors and Naive Bayes uses all genes.

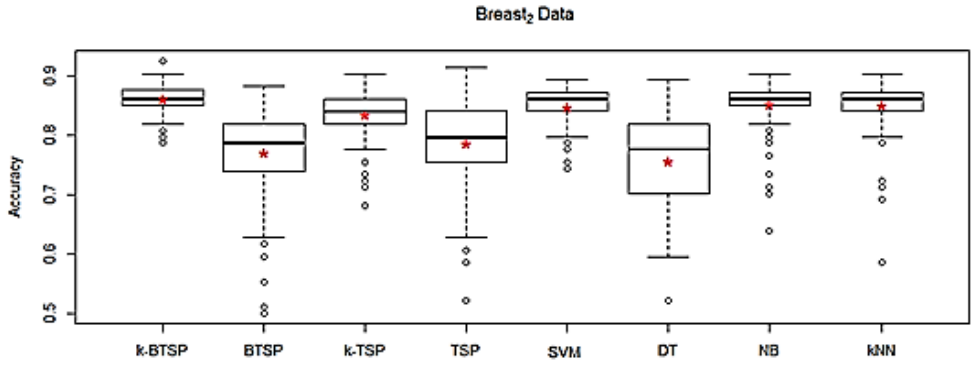

Figure 3.6: Classification accuracy rates for breast$_2$ data. The box plots show the distribution of the 100 runs for each classifier. Red starts are mean of the 100 runs.

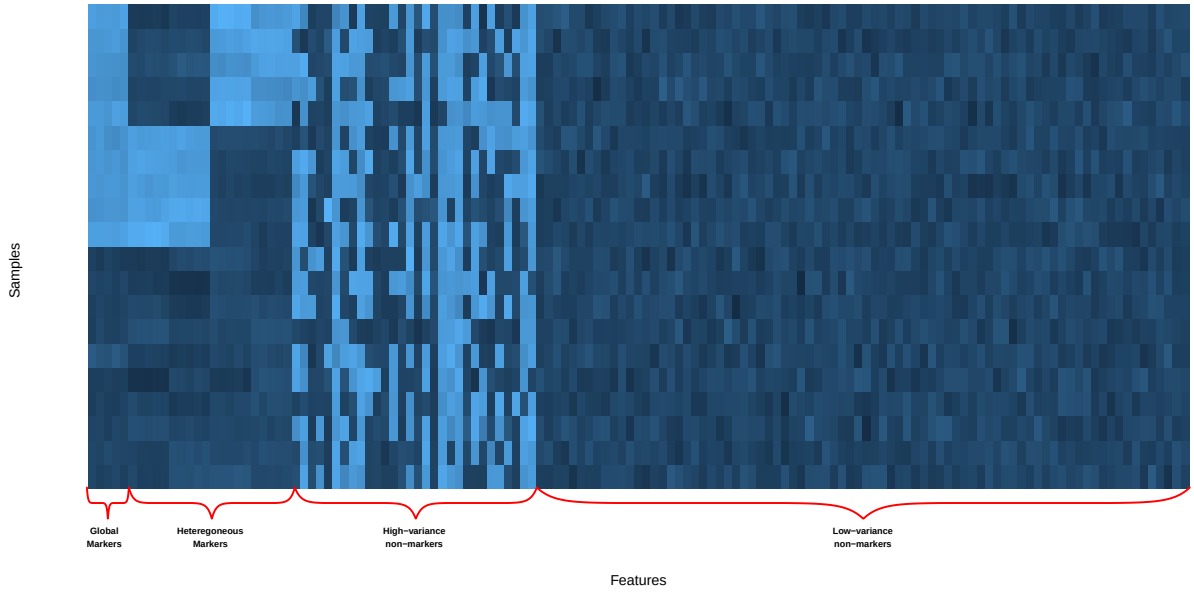

Figure 3.7: A demonstration of four gene types' distribution in constructing the synthetic data.

Variance pairs, σ_1, σ_2, define how hard it is to differentiate two phenotypes. If they are small, it is easier to differentiate two phenotypes, whereas it becomes more difficult to find the true label for bigger variance pairs. We combine four different variance pair distributions, σ_1/σ_2, with three different global marker choices for different training sample sizes. 100 test samples generated with the same parameters of training sets and accuracy rates calculated on these test sets. The procedure

Distributions	Global Markers	TSP	k-TSP	DT	BTSP	k-BTSP	NB	SVM	k-NN
$\sigma_1 = 0.3$ $\sigma_2 = 0.5$	10	0.5979	0.7062	0.6166	0.6039	**0.7194**	0.5230	0.6160	0.6135
	20	0.6199	0.7560	0.6635	0.6111	**0.7618**	0.5531	0.7130	0.6750
	30	0.6327	0.8034	0.6918	0.6298	**0.8174**	0.5614	0.7539	0.7561
$\sigma_1 = 0.5$ $\sigma_2 = 0.5$	10	0.5614	0.6426	0.5729	0.5580	**0.6472**	0.5157	0.5948	0.6307
	20	0.6004	0.6945	0.6043	0.5998	**0.7191**	0.5250	0.6704	0.6795
	30	0.6652	0.7434	0.6451	0.6428	**0.7714**	0.5477	0.6976	0.7370
$\sigma_1 = 0.7$ $\sigma_2 = 0.7$	10	0.5516	0.5776	0.5233	0.5362	**0.5880**	0.5058	0.5494	0.5645
	20	0.5682	0.6136	0.5377	0.5594	**0.6214**	0.5106	0.5768	0.5993
	30	0.5688	0.6458	0.5612	0.5667	**0.6489**	0.5257	0.5902	0.6272
$\sigma_1 = 0.5$ $\sigma_2 = 0.8$	10	0.5341	0.5696	0.5244	0.5437	**0.5838**	0.5109	0.5300	0.5482
	20	0.5640	0.6008	0.5550	0.5662	**0.6234**	0.5261	0.5633	0.5717
	30	0.5993	0.6686	0.5847	0.5929	**0.6821**	0.5400	0.5937	0.6098

Table 3.5: Estimated classification accuracy for simulated datasets. Training sample size is 10. Distributions show different variance pair distributions. It is harder to differentiate higher variance pairs. The highest accuracy for each dataset is highlighted in boldface.

was repeated 100 times and the average accuracy rates are presented in Tables 3.5, 3.6, 3.7 and 3.8.

The results show for all different global marker numbers, variance pair distributions and training sample sizes, the average accuracy rates of k-BTSP algorithm beat all the others. The larger variance we use, it becomes harder to differentiate between the different phenotypes. Results show the algorithm we proposed is a clear winner most of the times, especially for larger variance pairs.

3.5.3 Conclusion

Main problems with analyzing high-dimensional biological datasets are their small sample sizes, compared to feature/gene set and interpretable results of predictive models. We proposed a novel Bayesian rank based classifier to overcome these problems. We compared the performance of the proposed algorithm with popular rank based algorithms and off-the-shelf methods. The results on real datasets and simulated datasets show that the algorithm we propose presents the highest overall performance. Besides its high performance, the simplicity of the proposed decision rule provides links with biological understanding. The k-BTSP algorithm promises to investigate important mutations and marker genes.

Distributions	Global Markers	TSP	k-TSP	DT	BTSP	k-BTSP	NB	SVM	k-NN
$\sigma_1 = 0.3$ $\sigma_2 = 0.5$	10	0.6952	0.7382	0.7516	0.6938	**0.8315**	0.6460	0.8140	0.6760
	20	0.7130	0.8070	0.8364	0.7092	**0.8899**	0.6946	0.8823	0.7460
	30	0.7218	0.8174	0.8593	0.7201	**0.9069**	0.7239	0.8957	0.8167
$\sigma_1 = 0.5$ $\sigma_2 = 0.5$	10	0.6657	0.7012	0.7397	0.6554	**0.7823**	0.6291	0.7431	0.7272
	20	0.6817	0.7604	0.7834	0.6747	**0.8380**	0.6306	0.8212	0.7778
	30	0.6958	0.8122	0.7946	0.7033	**0.8500**	0.6636	0.8442	0.8230
$\sigma_1 = 0.7$ $\sigma_2 = 0.7$	10	0.6018	0.6254	0.6162	0.5909	**0.6835**	0.5353	0.6175	0.6445
	20	0.6085	0.666	0.6543	0.6361	**0.7286**	0.5718	0.6740	0.6672
	30	0.6410	0.6815	0.6953	0.6393	**0.7586**	0.5740	0.7240	0.6982
$\sigma_1 = 0.5$ $\sigma_2 = 0.8$	10	0.5936	0.6137	0.6178	0.6201	**0.6950**	0.5321	0.6019	0.5716
	20	0.6279	0.6554	0.6795	0.6259	**0.7422**	0.5572	0.6810	0.5993
	30	0.6366	0.7059	0.7154	0.6428	**0.7671**	0.5969	0.7343	0.6570

Table 3.6: Estimated classification accuracy for simulated datasets. Training sample size is 20. Distributions show different variance pair distributions. It is harder to differentiate higher variance pairs. The highest accuracy for each dataset is highlighted in boldface.

Distributions	Global Markers	TSP	k-TSP	DT	BTSP	k-BTSP	NB	SVM	k-NN
$\sigma_1 = 0.3$ $\sigma_2 = 0.5$	10	0.7163	0.8312	0.7128	0.8453	**0.8810**	0.7402	0.8780	0.7050
	20	0.7209	0.8543	0.8526	0.7244	0.9062	0.8310	**0.9215**	0.7704
	30	0.7265	0.8579	0.8572	0.7408	0.9330	0.8516	**0.9438**	0.8382
$\sigma_1 = 0.5$ $\sigma_2 = 0.5$	10	0.6797	0.7971	0.7837	0.6946	**0.8320**	0.7271	0.8260	0.7629
	20	0.7093	0.8226	0.8107	0.7025	0.8708	0.7696	**0.8890**	0.8101
	30	0.7122	0.8361	0.8185	0.7181	0.9031	0.8370	**0.9230**	0.8480
$\sigma_1 = 0.7$ $\sigma_2 = 0.7$	10	0.6279	0.7025	0.6535	0.6186	**0.7162**	0.6044	0.6870	0.6618
	20	0.6296	0.7375	0.6782	0.6503	**0.7564**	0.6425	0.7515	0.6950
	30	0.6580	0.7635	0.7000	0.6597	**0.8043**	0.6809	0.7908	0.7387
$\sigma_1 = 0.5$ $\sigma_2 = 0.8$	10	0.6348	0.6930	0.6930	0.6481	**0.7059**	0.5741	0.6683	0.5770
	20	0.6493	0.7524	0.7403	0.6635	**0.7721**	0.6332	0.7570	0.6247
	30	0.6660	0.7766	0.7588	0.6826	**0.8387**	0.6907	0.8141	0.6679

Table 3.7: Estimated classification accuracy for simulated datasets. Training sample size is 30. Distributions show different variance pair distributions. It is harder to differentiate higher variance pairs. The highest accuracy for each dataset is highlighted in boldface.

Distributions	Global Markers	TSP	k-TSP	DT	BTSP	k-BTSP	NB	SVM	k-NN
$\sigma_1 = 0.3$ $\sigma_2 = 0.5$	10	0.7182	0.8601	0.8687	0.7304	0.8890	0.8481	**0.9040**	0.7049
	20	0.7291	0.8831	0.8815	0.7404	0.9109	0.8838	**0.9326**	0.8050
	30	0.7273	0.8884	0.8913	0.7573	0.9426	0.891	**0.9547**	0.8660
$\sigma_1 = 0.5$ $\sigma_2 = 0.5$	10	0.7008	0.8232	0.8118	0.7061	0.8523	0.846	**0.8617**	0.7387
	20	0.7096	0.8470	0.8243	0.7246	0.882	0.8567	**0.9038**	0.835
	30	0.7143	0.8689	0.8298	0.7321	0.9096	0.8677	**0.9382**	0.8632
$\sigma_1 = 0.7$ $\sigma_2 = 0.7$	10	0.6437	0.7183	0.6737	0.6499	**0.7398**	0.6626	0.7066	0.6786
	20	0.6638	0.7811	0.7077	0.6724	**0.8081**	0.7184	0.7865	0.7030
	30	0.6714	0.7960	0.7234	0.6833	**0.8195**	0.7520	0.7978	0.7479
$\sigma_1 = 0.5$ $\sigma_2 = 0.8$	10	0.6511	0.7283	0.7156	0.6605	**0.7471**	0.6235	0.7027	0.5859
	20	0.6727	0.7768	0.7572	0.6889	**0.7983**	0.6886	0.7922	0.6364
	30	0.6715	0.8020	0.7737	0.6905	**0.8350**	0.7386	0.8302	0.6815

Table 3.8: Estimated classification accuracy for simulated datasets. Training sample size is 40. Distributions show different variance pair distributions. It is harder to differentiate higher variance pairs. The highest accuracy for each dataset is highlighted in boldface.

4. BTSP AS A FEATURE SELECTION METHOD *

Stephens [60] compared domains of big data in terms of storage and showed that Twitter requires 1-17 petabytes, YouTube needs 1-2 exabyte, whereas Genomics requires 2-40 exabytes of space to store its yearly data. From an expert point of view, it is crucial to identify informative genes among thousands of them. It is, therefore, no surprise that feature selection has become an indispensable tool for researchers and scientists in computational biology.

Furthermore, the number of genes (on the order of tens of thousands) in high-dimensional biological datasets greatly exceeds the number of samples (on the order of tens), and there are many irrelevant or redundant genes in a given set. Irrelevant or redundant genes can make it very difficult to identify relevant genes [61]. In addition, without a dimensionality reduction step, most classification rules will overfit the data.

We propose a novel feature selection approach based on the Bayesian Top Scoring Pair criterion [62], which is itself based on the TSP algorithm proposed by Geman and collaborators [21]. We conduct a detailed numerical experiment comparing its performance against that of well-known state-of-the-art feature selection algorithms using different classification rules and real gene-expression datasets. One may categorize approaches to reduce data dimensionality into two groups: feature extraction and feature selection. Feature extraction algorithms create new features by using existing ones; these approaches restrain biological interpretations as they do not preserve the original features (genes).

4.1 Feature Selection

Feature selection methods simply choose a subset of genes without any transformation, discarding redundant, noisy genes. Feature selection methods play a critical role in enhancing the performance of classification and facilitating biological interpretation by producing small panels

*Part of this section is reprinted with permission from "Bayesian Top Scoring Pairs for Feature Selection," by Emre Arslan and Ulisses Braga-Neto, 2017, 51th Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, October-November 2017.

of *biomarkers*. However, exhaustive feature selection (i.e., searching all possible feature subsets of a given size) is impossible for high-dimensional biological data as the computational load increases exponentially. Indeed, exhaustive feature selection is an NP-hard problem [63].

One can further categorize feature selection methods into three main groups:

- Filter methods: A score is calculated for each gene and genes with higher scores are selected. These approaches are independent of the classification rule used on the selected feature set, which minimizes overfitting. Due to the high-dimensionality of the data, the computational efficiency of these methods makes them popular. Filter methods are divided into univariate and multivariate methods. Univariate methods consider each gene as independent, whereas multivariate methods take relationships among genes into consideration.

- Wrapper methods: These approaches use a classification rule to validate selected genes performance and then employ the same rule on the final selected feature set. Wrapper methods tend to have better performance as they "match" selected genes to the classification rule. However, the major disadvantage of wrapper methods is their computational complexity.

- Embedded methods: These approaches combine feature selection and classifier construction.

4.1.1 Recursive Feature Elimination (RFE)

Guyon et al. [55] proposed a Support Vector Machine (SVM)-based recursive feature elimination method. The following weighting factor is calculated for all genes:

$$|w_j| = \left| \sum_{i=1}^{n} \alpha_i y_i x_{ij} \right| \tag{4.1}$$

where x_{ij} is the j^{th} gene expression of the i^{th} sample point, y_i is the label of i^{th} sample point and α_i is the Lagrange multiplier for the i-th sample point, which is obtained by quadratic programming optimization in SVM training. After calculating expression (4.1) for all genes, one selects the genes with the largest values. This method has displayed very good performance in different fields, for instance, proteomics [64], metabolomics [65], genomics [66], and more.

4.1.2 χ^2-Statistics

Another widely used feature selection method is the χ^2 method. The χ^2 statistic is used to test whether two events are independent. First, gene expression values are discretized into several intervals and then we calculate χ^2 values for each gene:

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=0}^{1} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{4.2}$$

where m is the number of intervals, O_{ij} is the frequency of patterns in the i^{th} interval and j^{th} class, and E_{ij} is the expected frequency of O_{ij}. We use the FSelector package [67] in R to perform χ^2 feature selection by ranking the genes according to their χ^2 values.

Study	Number of Genes	Class 1 Size	Class 2 Size	Reference
Colon	2000	22	40	Alon et al. [50]
Leukemia	7129	25	47	Golub et al. [27]
Breast	22,283	62	42	Chowdary et al. [30]
Lung	12,533	150	31	Gordon et al. [29]

Table 4.1: Gene expression data sets

4.1.3 Relief-F

Relief-F [68] is an extension of the Relief algorithm [69]. A sample of the data is chosen randomly. The algorithm searches for k nearest neighbors from the same class called "nearHits," and k nearest neighbor from the other class called "nearMisses." ReliefF evaluates the quality for each gene using the following equation:

$$W(i) \;=\; W(i) - \overline{\mathrm{diff}(i, n, \mathrm{nearHit})} + \overline{\mathrm{diff}(i, n, \mathrm{nearMiss})}, \tag{4.3}$$

where $W(i)$ represents the quality of estimation of the i^{th} gene, $\overline{\mathrm{diff}(i, n, \mathrm{nearHit})}$ is the average difference between the i^{th} gene expression and among the selected sample points, and the nearHit

sample and $\overline{\mathrm{diff}(i, n, \mathrm{nearMiss})}$ is the average difference of the i^{th} gene expression among the selected sample points and nearMiss sample points. $\overline{\mathrm{diff}(i, n, \mathrm{nearMiss})}$ rewards the distance between different classes and penalizes the distance within the same class. This process is repeated M times, where M is defined by the user.

4.1.4 Information Gain

The information Gain (IG) of the i^{th} gene over class Y is the uncertainty reduction if we know the value of gene expression X_i of the i^{th} gene,

$$\mathrm{IG}(X_i) = H(Y) - H(Y \mid X_i) \tag{4.4}$$

where H denotes the information-theoretic entropy. Entropy-based discretization is the first step to calculate IG for gene expression as IG can be calculated for discrete values. IG is a filter feature selection method. We use the FSelector package [67] in R to perform IG feature selection method.

4.1.5 Gain Ratio

Gain Ratio (GR) feature selection methods were proposed to deal with the symmetric measurement bias of IG. GR normalizes IG by the entropy of X_i:

$$\mathrm{GR} = \frac{IG}{H(X_i)} \tag{4.5}$$

where $\mathrm{GR} = 1$ indicates that the knowledge of X_i completely predicts class Y.

4.2 BTSP as a Feature Selection Method

We use the Bayesian interpretation of the BT Model in [40] to get skill parameters of each gene by using **W** matrix. A cell in the **W** represents how many times a gene expression value is greater than the other one, i.e. W_{ij} represents how many times i^{th} gene is greater than j^{th} gene. It is trivial that all diagonal elements are 0 and W_{ij} + W_{ji} represent how many times i^{th} and j^{th} genes are compared, which is the sample size if we don't have a missing value in either of these genes. After we get the skill parameters of each gene we calculate the BTSP score (eq.4.7) for each pair

of genes with the probability of i^{th} gene expression is greater than j^{th} gene expression within a given class.

$$\pi_{ij}^0 = \frac{\lambda_i^0}{\lambda_i^0 + \lambda_j^0} \quad \text{and} \quad \pi_{ij}^1 = \frac{\lambda_i^1}{\lambda_i^1 + \lambda_j^1} \tag{4.6}$$

$$\Omega_{ij} = \left|\pi_{ij}^0 - \pi_{ij}^1\right| \tag{4.7}$$

One can easily generalize the BTSP algorithm by selecting the top k disjoint pairs. This generalization is called k-BTSP [62].

4.3 Results

We investigated the well-known feature selection methods defined in the previous section under the Linear Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), with a cross-validated choice of k, and Naive Bayes (NB) classification rules, in order to assess their performance on real genomic data sets, listed in Table 4.1.

We employed a variance filter in order to reduce the number of genes to 2000 for all datasets. We used 20% of the data as the training data and the rest as the test data in order to mimic the small sample size problem. We repeated this procedure 100 times and computed the average accuracy rates. To investigate the feature selection performance, we selected 20, 40, 60, 80 and 100 genes.

We can observe in Figures 4.1 to 4.4 that the k-BTSP algorithm produced better classification accuracy rates with different classifiers, different number of features/genes and in different datasets, most of the time.

Accurate feature selection is needed to analyze high-dimensional biological data sets. We proposed here a novel feature selection method, based on the Bayesian TSP method, and compared its performance against well-known feature selection methods under SVM, k-NN and NB classification rules, using real genomic data sets. The results indicate the promise of the k-BTSP in the analysis of high-dimensional biological data.

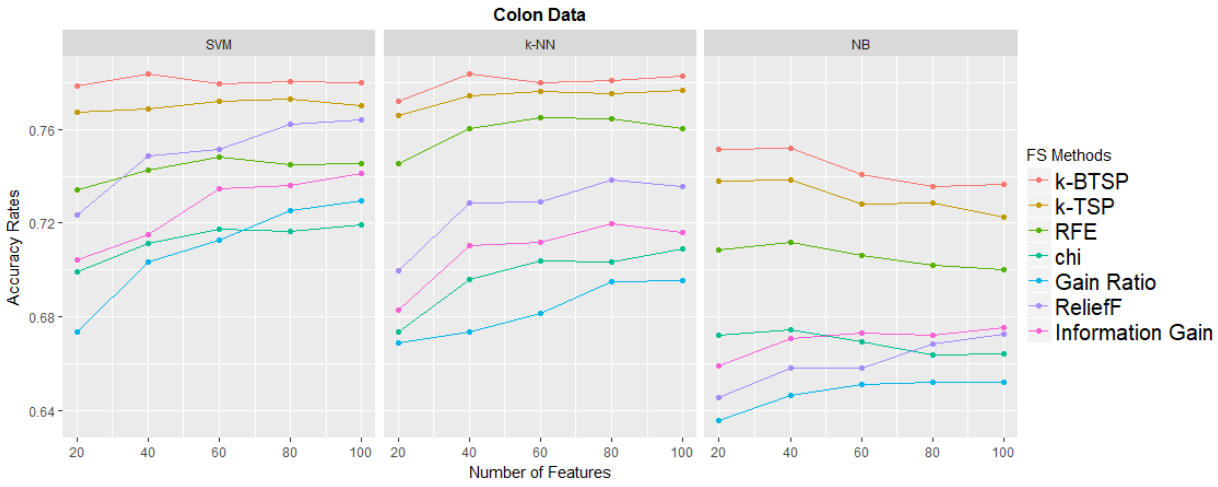

Figure 4.1: Comparison between accuracy rates of the feature selection methods on colon data

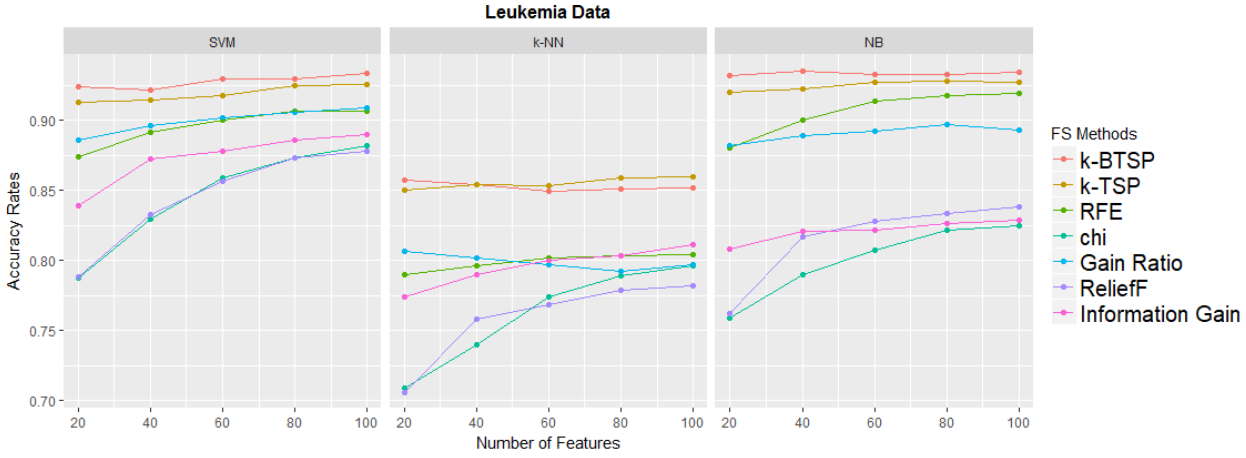

Figure 4.2: Comparison between accuracy rates of the feature selection methods on leukemia data

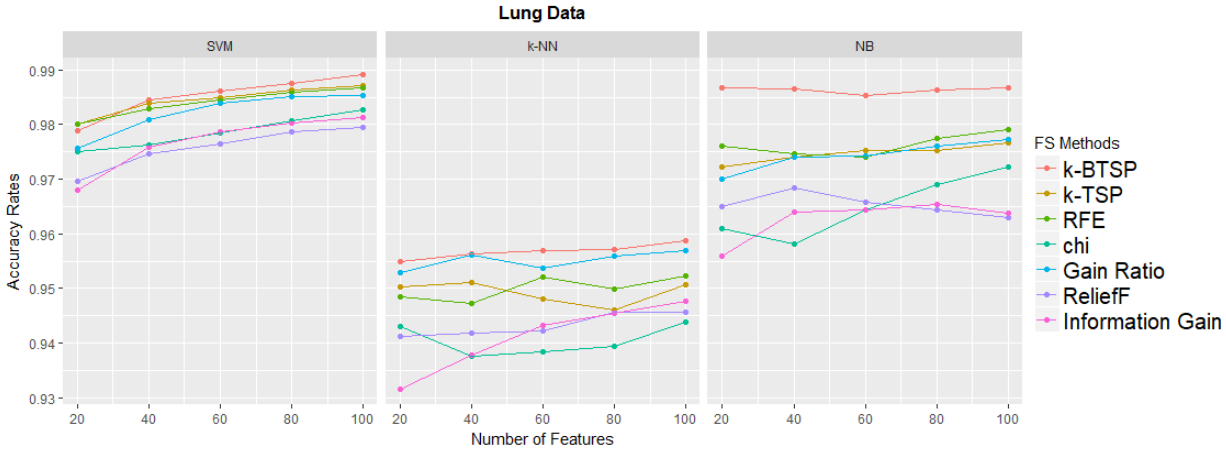

Figure 4.3: Comparison between accuracy rates of the feature selection methods on lung data

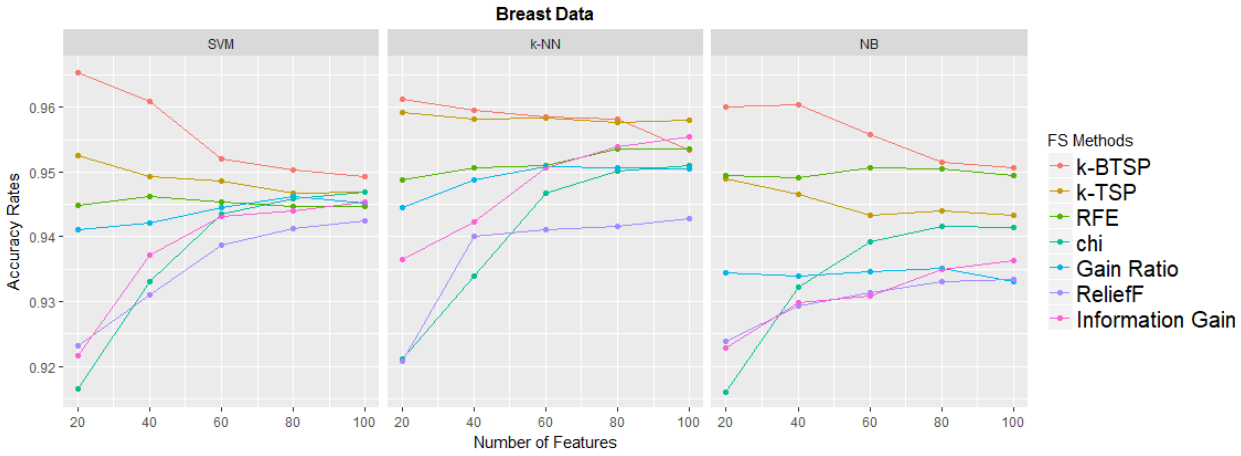

Figure 4.4: Comparison between accuracy rates of the feature selection methods on breast data

5. A BAYESIAN PATHWAY DATA BASED CLASSIFIER

Pathway analysis can help us understand the signaling mechanisms which are responsible for initiating and advancing a particular disease. Pathway analysis helps automate the traditional approach that checks each gene manually, look up each gene in the literature, and try to learn about genes' involvement in a particular biological process.

Some benefits of pathway analysis are:

- Easy biological interpretation
- Prediction of new roles of genes
- Finding possible mechanisms related to diseases
- Integration of multiple data types
- Comparibility

There are two factors that make analyzing genomic dataset at the functional level very valuable. First, using sets/pathways instead of genes reduce the dimension significantly. Second, identifying pathways related to the phenotypes may have more explanatory power. Pathway analysis also increases the statistical power in two ways: one, it collects small count into a big count which makes it easier to define a pattern related to a phenotype and two, it reduces the number of tests. It has been shown that analyzing the genomic data set in a "modular" level creates more reproducible results [70, 71].

One of the problems with pathway analysis methods is that the marker pathways are often selected via their stand-alone score and no interaction is taken into consideration. Therefore, selected pathways may have redundant information which affects the performance of the analysis. We can alleviate the aforementioned problem by using a method that doesn't analyze each pathway independently.

There have been significant amount of studies on finding differentially expressed genes, but there are not that many studies for pathway analysis. It is a challenging problem to take interactions and sets of genes and analyze the high-dimensional genomic data.

In this chapter, we propose a novel algorithm, Bayesian Top Scoring Pathway Pair (BTSPP), which analyzes genomic data at the level of pathways instead of gene level. Briefly, we find skill parameters for each gene by using (5.1, 5.2), define scores for each pathway in different phenotypes, and find the pair of pathways that have the biggest Bradley-Terry probability flips between phenotypes.

The algorithm we proposed identifies discriminative pairs of pathways by using summarized statistics of pathways. The BTSPP is designed with the following benefits:

- Revealing new biological insights in the pathway level
- Offering interpretable decision rules for predictive models
- Minimizing the technological differences, such as different platforms, to compare/combine different datasets
- Reducing the dimensionality of the data from thousands to tens by summarizing the expression values

We apply the proposed algorithm to eight real data sets and demonstrate it achieves higher accuracy rates compared to several pathway data-based algorithms. We also compared the proposed algorithm ability to find biologically validated pathways for different phenotypes with other algorithms, Over-Representation Analysis (ORA) and Gene Set Enrichment Analysis (GSEA).

5.1 Bayesian Top Scoring Pathway Pairs

The aim of this approach is to identify pathway markers by taking ranking and pairwise interactions into consideration, and use these pathway markers in a predictive model. The proposed algorithm is motivated by the great performance of BTSP algorithm as a classifier [72], [62] and

feature selection [73] method. The Bayesian Top Scoring Pathway Pair (BTSPP) algorithm classifies a given test sample according to the following four-step procedure:

5.1.1 Skill Parameter Estimation of Genes

Let X represent the gene expression matrix with p x n, where p represents number of genes and n represents number of samples. For the sake of simplicity, we will assume there are two classes/phenotypes. We are treating every gene as individuals, competing with each other and trying to estimate their skill parameters. We will leverage the following Gibbs sampling approach proposed in [40].

$$Z_{ij}^{t+1} \mid \mathcal{D}, \boldsymbol{\lambda}^t \sim \text{Gamma}(n_{ij}, \lambda_i^t + \lambda_j^t)\,; \tag{5.1}$$

$$\lambda_i^{t+1} \mid \mathcal{D}, \mathbf{Z}^{t+1} \sim \text{Gamma}\left(a + w_i,\ b + \sum_{\substack{i<j \\ \text{s.t. } n_{ij}>0}} z_{ij}^{t+1} + \sum_{\substack{i>j \\ \text{s.t. } n_{ij}>0}} z_{ji}^{t+1}\right). \tag{5.2}$$

where $\boldsymbol{\lambda}$ represents the skill parameters of each gene, $\mathcal{D}$ is data and $\mathbf{Z}$ is a latent variable. Detailed explanation of the sampling approach was given in Chapter 3.

5.1.2 Pathway Mapping

Public databases (KEGG, MSigDB) were used to extract the gene set information representing cellular pathways. Assignments to the pathways are computed for all genes in the data. Some genes are removed if there is no pathway to cover them. There are some genes, having the same Entrez ID, we selected the gene having the highest variance between phenotypes. We will have skill parameters for all genes in class 0 (normal/drug treatment) as $\boldsymbol{\lambda}^0$ and class 1 (cancer/no treatment) as $\boldsymbol{\lambda}^1$.

5.1.3 Scoring of Pathway Pairs

In order to score each pathway, we are using the individual skill parameters of genes. Let's assume we have g genes in a pathway. The pathway score is calculated as the average skill of these

g genes in that pathway.

$$\lambda_{p_i} = \frac{\sum_{i \in G_i} \lambda_i}{|G_i|} \tag{5.3}$$

where G_i represents genes in the i^{th} pathway and $|G_i|$ is the number of genes in this pathway. We have to find skills of pathways for each class, so we will have ${\lambda_{p_i}}^0$ and ${\lambda_{p_i}}^1$.

5.1.4 Identification of Pathway Pairs & Classification Rule

According to the Bradley-Terry Model, the probability of individual i beats individual j is

$$P(i > j | \lambda_i, \lambda_j) = \frac{\lambda_i}{\lambda_i + \lambda_j} \tag{5.4}$$

where λ_i represents the skill of individual i. We will use calculated skill parameters for each pathway in phenotypes to find the biggest flip in terms of the probability of one pathway beating the other between classes. We define the following score with this aim:

$$\Psi_{i,j} = \left| \frac{{\lambda_{p_i}}^0}{{\lambda_{p_i}}^0 + {\lambda_{p_j}}^0} - \frac{{\lambda_{p_i}}^1}{{\lambda_{p_i}}^1 + {\lambda_{p_j}}^1} \right| \tag{5.5}$$

We calculate the score for each pair of pathways and select the pair with the highest score. An exhaustive search to check for all pairs should be feasible as the number of pathways is not that many when compared with the number of genes.

After we select the pair, having the highest pathway score, we compare the mean of these 2 pathways for a given test sample. Let's assume we choose i^{th} and j^{th} pathways and ${\lambda_{p_i}}^0 > {\lambda_{p_i}}^1$; the classifier is defined for a test sample as:

$$h_{BTSPP}(\mathbf{x}_{new}) = \begin{cases} C_0, & \text{if } \frac{\sum_{\ell \in G_i} \mathbf{x}_\ell}{|G_i|} < \frac{\sum_{\ell \in G_j} \mathbf{x}_\ell}{|G_j|}, \\ C_1, & otw. \end{cases} \tag{5.6}$$

5.2 Alternative Pathway-Based Algorithms

Pathway-based algorithms have become more popular to analyze the high-dimensional biological data sets. For further comparison, we implemente six pathway based algorithms: Log-Likelihood Ratio (LLR) [70], Ranking Log-Likelihood Ratio (R-LLR) [74], Top Scoring Pathway Pair (TSPP) [75], Gene Expression Deviation (GED) [76], and Pathway Variance (PathVar) [77].

5.2.1 LLR

Su et al. [70] proposed a method, log-likelihood ratio (LLR), that estimates pathway activities based on the log-likelihood ratio between different phenotypes. This method assumes each gene has a Gaussian distribution. Mean and standard deviation of each gene is estimated and the conditional probability density function (PDF) is calculated by using these statistics.

Gene expression matrices for each pathway are transformed to LLR matrices by using estimated conditional PDFs. The LLR of the j^{th} sample and i^{th} gene is:

$$\lambda_i(x_j^i) = log\left(\frac{f_i^1(x_j^i)}{f_i^2(x_j^i)}\right) \tag{5.7}$$

where $f_i^1(x)$ represents the conditional probability function of gene i in the first class and $f_i^2(x)$ is the conditional probability density function in class 2. The LLR is a matrix with the number of genes in a pathway by the number of samples. LLR matrices are calculated for every pathway. The next step is to normalize the LLR matrix according to the following formula:

$$\widehat{\lambda}(x_j^i) = \frac{\lambda_i(x_j^i) - \mu(\lambda_i)}{\sigma(\lambda_i)} \tag{5.8}$$

where $\lambda_i = \lambda_{i1}, \lambda_{i2}, ..., \lambda_{iN}$ is the i^{th} row of the LLR matrix with N samples, $\mu(\lambda_i)$ and $\sigma(\lambda_i)$ denote the mean and standard deviation of that row.

The defined pathway activity score combines every gene for a sample in a pathway. Authors estimate the pathway activity, a_j, as:

$$a_j = \sum_{i=1}^{n} \lambda_i \left(x_j^i\right) \tag{5.9}$$

Authors apply t-test statistics on pathway activity vector for each pathway and sort pathways according to p-values. They used some part of the training set as the marker-evaluation set and remaining as feature-selection to decide the number of pathways to be used when constructing the LDA classifier.

5.2.2 R-LLR

In [74], Khunlertgit and Yoon proposed pathway activity inference method based on the ranking of genes within a sample. First, they create a pathway matrix with each row represents genes and columns as samples. Let's assume we have n genes in a given pathway with m_0 and m_1 samples representing two classes' sample sizes. So the total sample size is $m = m_0 + m_1$. This n x m pathway matrix is converted into a gene ranking binary matrix. Each row of this gene ranking matrix represents a pairwise comparison beginning with (1,2) and ends with (n-1,n). So the size of this matrix is $n(n-1)/2$ x m. The comparison between i^{th} and j^{th} gene in the k^{th} sample of this ranking matrix is $r_k^{i,j}$. By using the following equation, the binary gene ranking matrix is filled for each pathway.

$$r_k^{i,j} = \begin{cases} 1, & if\ x_k^i < x_k^j \\ 0, & otw. \end{cases} \tag{5.10}$$

The ranking profile of a sample is:

$$r_k = \left(r_k^{i,j} | 1 \leq i < j \leq n\right). \tag{5.11}$$

Unlike the LLR algorithm, this method tries to estimate the conditional probability mass function (PMF). $f_{i'j}^c(r)$ represents the conditional PMF of the ranking of the expression level of gene i and gene j under phenotype c. The ranking log-likelihood ratio, $\lambda_{i,j}(r_k^{i,j})$, is calculated by using the estimated PMFs for each phenotypes. The ranking LLR is defined as:

$$\lambda_{i,j}(r_k^{i,j}) = log\left(\frac{f_{i,j}^1(r_k^{i,j})}{f_{i,j}^2(r_k^{i,j})}\right). \tag{5.12}$$

The ranking LLR matrix has the same dimension as the gene ranking matrix. It is not easy to estimate the PMFs in a small sample size area. In order to overcome this problem authors used the normalization idea, like in the LLR algorithm.

$$\widehat{\lambda}_{i,j}(r_k^{i,j}) = \frac{\lambda_{i,j}(r_k^{i,j} - \mu(\lambda_{i,j})}{\sigma(\lambda_{i,j})}, \tag{5.13}$$

where $\mu(\lambda_{i,j})$ and $\sigma(\lambda_{i,j})$ are the mean and standard deviation of i^{th} and j^{th} genes row in the R-LLR matrix.

They followed the same strategy to assess the pathway activity level like in LLR. The pathway activity level, a_k, is the sum of the column of R-LLR matrix:

$$a_j = \sum_{i=1}^{n} \lambda_{i,j}\left(r_k^{i,j}\right). \tag{5.14}$$

The discriminative power of a pathway is checked with the t-test statistic score pathway activity levels:

$$t(\mathbf{a}) = \frac{\mu_{\mathbf{1}} - \mu_{\mathbf{2}}}{\sqrt{\frac{\sigma_{\mathbf{1}}}{\mathbf{K_1}} + \frac{\sigma_{\mathbf{2}}}{\mathbf{K_2}}}}, \tag{5.15}$$

where $\mathbf{a}$ is the vector of inferred pathway activity levels. μ_c is the mean of class c samples activity levels mean, σ_c is the standard deviation of pathway activity levels of class c, and K_c is the number of samples in class c.

After applying the t-test for all pathways, pathway markers were sorted according to their t-scores. R-LLR algorithm followed the similar setup to choose the number of pathways used for the classification and LDA as the classifier.

The main difference between R-LLR and LLR is the previous approach takes the ranking in order to estimate the probabilistic evidence by each gene whereas the second one uses the expression

values.

5.2.3 TSPP

Glaab et al. proposed Top Scoring Pathway Pair, TSPP, in [75]. The algorithm is a methodological extension of TSP algorithm.

The "Rank Matrix" R is created from the gene expression matrix by sorting genes within each sample. R matrix contains each genes position index within a sample. Each row in this R matrix represents genes and columns represent samples. If a gene cannot be assigned to any pathway, the row corresponding to that gene is removed. A pathway submatrix is extracted from the big R rank matrix for each phenotype. For sake of the simplicity, let's assume we have two classes. Each pathway has R_1 and R_2 matrices representing two classes. These matrices are reduced to r_1 and r_2 vectors by taking the median rank for each sample/column.

$$Partial\ Score_1 = \sum_{i \in S_1} \mathbb{1}(r_{1i} \geq r_{2i}) + \sum_{i \in S_2} \mathbb{1}(r_{1i} < r_{2i}) \tag{5.16}$$

$$Partial\ Score_2 = \sum_{i \in S_1} \mathbb{1}(r_{1i} < r_{2i}) + \sum_{i \in S_2} \mathbb{1}(r_{1i} \geq r_{2i}) \tag{5.17}$$

$$score = \frac{max(Partial\ Score_1,\ Partial\ Score_2)}{|S_1| + |S_2|} \tag{5.18}$$

where $\mathbb{1}$ denotes the indicator function. An exhaustive search is applied for all pairs to get the score. The TSPP has a simple decision to predict a given test sample by checking the median rank of the selected pair of pathways.

5.2.4 Gene Expression Deviation

Gene Expression Deviation algorithm proposed by Young and Craft [76]. This algorithm uses over and under-expression of genes within a pathway. There are two parameters estimated for each pathway and each sample, namely: overexpression score and underexpression score. A preselection method is applied according to the Kolmogorov-Smirnov test. If a gene in a pathway passes

that test, which means there is a significant difference between phenotypes and it will be used for further steps.

Let e_{pg} represent the expression level of gene g in sample p, the differential-score is defined as:

$$\Delta_{pg} = \frac{e_{pg} - \mu_{Ng}}{\sigma_{Ng}}, \tag{5.19}$$

where μ_{Ng} is the mean of gene g in normal/reference class and σ_{Ng} is the standard deviation across normal/reference samples. The positive differential-scores and the negative ones are summed for each sample. Pathway scores are aggregated by sample under-representation and over-representation scores.

Authors proposed another method called normal tissue centroid (NTC) [76]. This method represents each sample in a g (number of genes in that pathway) dimensional space and aggregates these locations in normal class to find the normal tissue centroid. We implemented this method as well; however, results are not competitive at all, so we decided not to use the NTC as a benchmark method.

The original implementations of NTC and GED have several other steps like weighting samples via silhouette; however, these steps are not related to the pathway analysis. So to be consistent with all other methods, we did not use these tuning steps.

5.2.5 PathVar

Unlike the common approach that compares the mean or median across the difference between biological phenotypes, the Pathway Variance (PathVar) [77] compares the change of the variance across classes.

PathVar has a freely available website *http://pathvar.embl.de* which provides easy access to analyze the genomic data in terms of pathway variance difference. It has a tutorial and sample datasets to introduce the use of the algorithm.

5.3 Results

In order to compare the predictive performance of the proposed BTSPP algorithm with other pathway data-based algorithms, we use eight real datasets: GSE19728 [78], GSE16515 [79], GSE781 [80], GSE21354 [78], GSE8671 [81], GSE9348 [82], GSE15852 [83], and GSE9574 [84].

We use Kyoto Encyclopedia of Genes and Genomes (KEGG) and Molecular Signatures Database (MSigDB) databases to obtain the set of biological pathways. For each data set, we identify genes in pathways and use these sets by removing the genes that are not represented in that dataset. We use pathways having at least ten genes to aggregate meaningful statistics for all methods.

The small sample size is a major problem to analyze high-dimensional biological datasets. It has been argued in [85] and showed that over 57% of GEO datasets' sample size is less than 12. In order to simulate the small sample size problem, we use 50% of each dataset as a training set and evaluate the predictive accuracy on the other 50% of the data. We repeat this procedure 100 times for each dataset and take the average as the predictive performance.

The BTSPP algorithm uses the same procedure to estimate the skill parameters mentioned in [62]. We apply 1500 iterations with 300 burn-in part to estimate the skill parameters of each gene. We apply a variance filter and use the top 2000 genes to reduce the computational complexity of evaluating genes' skill parameters, whereas we use all genes for other pathway based algorithms. If we apply the same procedure and use the top 2000 genes selected via variance filter for other algorithms as well, these pathway based algorithms will have lower accuracy rates.

We implemented LLR, R-LLR, and GED as we summarized in the previous section. LLR and R-LLR choose the number of pathways by using some part of the training data as the marker-evaluation set and the rest of the training data as the feature-selection set, whereas GED uses all pathways. We implement the TSPP with an exhaustive search and find the pair having the highest TSPP score. The TSPP algorithm uses one pair of pathways like BTSPP.

The classification accuracy rates for each algorithm applied to different datasets are presented in Tables 5.1 and 5.2. The proposed algorithm, BTSPP, has the best average accuracy rate and for

Datasets	BTSPP	TSPP	R-LLR	LLR	GED	PathVar
GSE 19728	**0.9**	0.731	0.891	0.854	0.879	0.887
GSE 21354	**0.963**	0.7934	0.9362	0.9387	0.8457	0.95
GSE 16515	**0.7556**	0.7191	0.6972	0.7306	0.6976	0.7461
GSE 781	**0.7569**	0.6075	0.65	0.6325	0.685	0.69
GSE 8671	**0.9966**	0.9225	0.9916	0.9912	0.9944	0.9875
GSE 9348	**1**	0.9461	0.9361	0.9765	0.9753	0.9789
GSE 15852	0.838	0.6744	0.7542	**0.8416**	0.8414	0.7749
GSE 9574	**0.6777**	0.5	0.5236	0.5378	0.57	0.5936
Average	**0.8610**	0.7367	0.7974	0.8129	0.8111	0.8259

Table 5.1: Estimated classification accuracy for eight real datasets. (KEGG database)

Datasets	BTSPP	TSPP	R-LLR	LLR	GED	PathVar
GSE 19728	**0.9395**	0.7775	0.8693	0.84	0.8816	0.905
GSE 21354	0.9146	0.67	0.8654	0.9038	0.875	**0.9231**
GSE 16515	0.7746	0.7705	0.7229	**0.8**	0.7729	0.7538
GSE 781	**0.7771**	0.61	0.69	0.665	0.685	0.615
GSE 8671	0.9967	0.914	0.9981	0.9856	**1**	**1**
GSE 9348	0.9836	0.8913	0.9683	0.9791	0.994	**0.995**
GSE 15852	**0.8695**	0.699	0.7958	0.8553	0.8609	0.8623
GSE 9574	0.7264	0.5714	0.5271	0.6035	0.7021	**0.7857**
Average	**0.8727**	0.7376	0.8082	0.8290	0.8471	0.8555

Table 5.2: Estimated classification accuracy for eight real datasets. (MsigDB database)

most of the datasets, it provides the highest accuracy rates, despite it using only two pathways to classify the given test sample.

Guo et al. [86] use the mean and median expression values to derive the pathway activity score. Lee et al. [87] use a subset of a pathway, namely condition-responsive genes (CORGs), to find informative pathways. It has been shown in [70] and [74] that LLR and R-LLR algorithms outperform CORG, mean, and median pathway analysis methods. So we didn't implement these methods as we were already having better performances than LLR and R-LLR.

It seems BTSPP is a good candidate as a predictive method based on biological pathway data.

$Pathway_1$	$Pathway_2$	Score
hsa04931 Insulin resistance	hsa04657 IL-17 signaling pathway	0.9712
hsa04910 Insulin signaling pathway	hsa04668 TNF signaling pathway	0.9498
hsa04152 AMPK signaling pathway	hsa05225 Hepatocellular carcinoma	0.9424
hsa00980 Drug metabolism - cytochrome P450	hsa05213 Endometrial cancer	0.9185
hsa00982 Drug metabolism - cytochrome P450	**hsa05210 Colorectal cancer**	0.9006

Table 5.3: Top ten-ranked pathways for the GSE8671 data according to BTSPP score. The boldface pathway is the target pathway.

Further, we check if the proposed algorithm can detect biologically validated pathway for a disease. We choose GSE8671 and GSE9348 datasets, as classification accuracies for both datasets are high for most algorithms. We use whole data as the training set and check for top ten pathways for each algorithm.

GSE8671 has 64 samples with two phenotypes (Normal/Colorectal Cancer). We apply LLR, R-LLR, PathVar, GED, and TSPP on GSE8671. None of them have the target *hsa05012* pathway in the top ten pathway lists. We also check the proposed BTSPP algorithm and top ten pathways are listed in Table 5.3. BTSPP detects the target pathway in top ten lists. We use clusterProfiler package [88] to apply ORA and GSEA.

We apply the Over-Representation Analysis (ORA) on GSE8671 and investigate its ability to find the target *hsa05012* pathway. This pathway is not included in the top 10 pathways found by ORA. The p-value of the *hsa05012* according to ORA is 0.7966 and the q-value is 0.7833. So ORA cannot detect the target pathway by using whole GSE8671 data.

Gene Set Enrichment Analysis (GSEA) is another way to find the informative pathway related to the phenotypes. We apply the log ratio ranking and check if GSEA can detect the *hsa05012*. The GSEA cannot detect the target pathway as well. Figure 5.1 shows enrichment analysis for the *hsa05012*. The enrichment score is around 0.3 whereas top pathways found by GSEA are around 0.6. The target pathway is not in the top ten of the list of GSEA.

GSE9348 has 82 samples with two phenotypes (Normal/Colorectal Cancer). We apply LLR, R-LLR, PathVar, GED, and TSPP on GSE9348. None of them has the target *hsa05012* pathway in

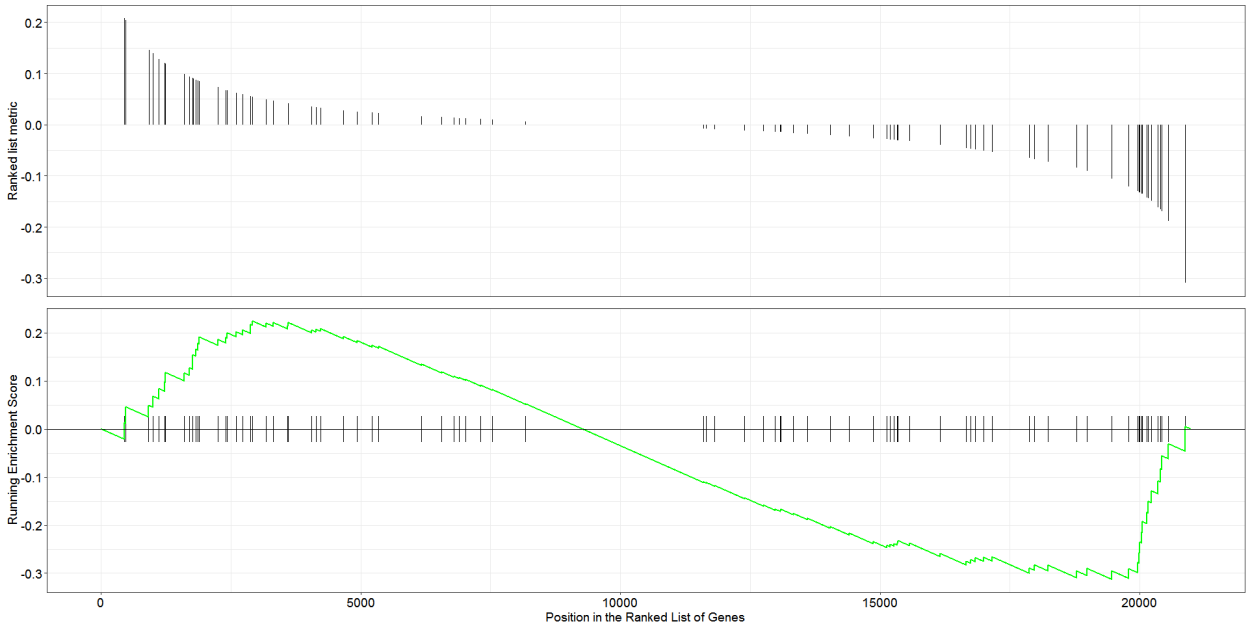

Figure 5.1: Gene set enrichment analysis of *hsa05012* on GSE8671 data.

top ten pathway list. We also check the proposed BTSPP algorithm and top ten pathways are listed in Table 5.4. BTSPP detects the target pathway in top ten list.

We applied the Over-Representation Analysis (ORA) on GSE9348 and investigated its ability to find the target *hsa05012* pathway. This pathway is not included in the top ten pathways found by the ORA. The p-value of the *hsa05012* according to the ORA is 0.2478 and the q-value is 0.3411. So the ORA can not detect the target pathway by using whole GSE9348 data.

Gene Set Enrichment Analysis (GSEA) is another way to find the informative pathway related to the phenotypes. We apply the log ratio ranking and check if the GSEA can detect the *hsa05012*. The GSEA cannot detect the target pathway as well. Figure 5.2 shows enrichment analysis for the *hsa05012*. The enrichment score is around 0.3 whereas top pathways found by GSEA are around 0.6. The target pathway is not in the top ten of the list of the GSEA.

5.4 Conclusion

We proposed a pathway analysis method which can be used to find robust, reproducible and interpretable results. The proposed method leverages from two effective strategies namely, the

Pathway_1	Pathway_2	Score
hsa00830 Retinol metabolism	**hsa05210 Colorectal cancer**	0.9569
hsa00980 Metabolism of xenobiotics by cytochrome P450	hsa05224 Breast cancer	0.9347
hsa00040 Pentose and glucuronate interconversions	hsa05219 Bladder cancer	0.9269
hsa00982 Drug metabolism - cytochrome P450	hsa04512 ECM-receptor interaction	0.9184
hsa05032 Morphine addiction	hsa04120 Apoptosis	0.9079

Table 5.4: Top ten-ranked pathways for the GSE9348 data according to BTSPP score. The boldface pathway is the target pathway.

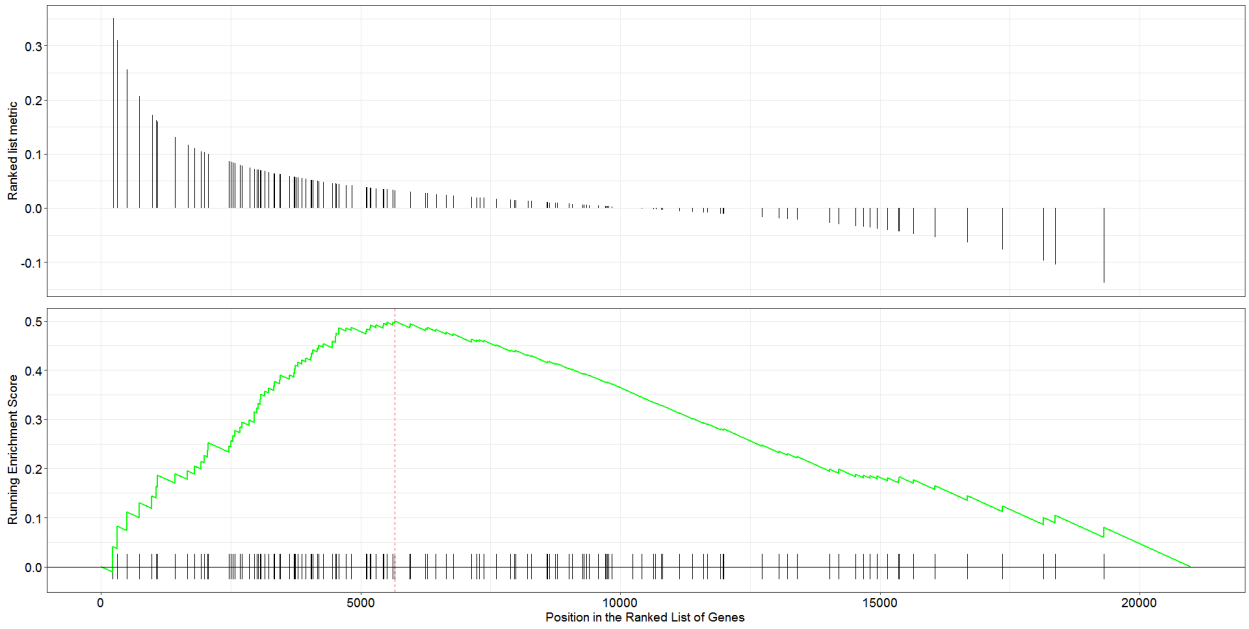

Figure 5.2: Gene set enrichment analysis of *hsa05012* on GSE9348 data.

pairwise interactions between genes and ranking-based gene expression analysis. Experimental results on real datasets show our method finds more informative pathway markers to predict a test sample.

6. SUMMARY AND FUTURE DIRECTIONS

The small sample size of training sets and lack of biological interpretations due to the complex classification boundaries of algorithms are two major barriers before computational biology. It has been shown in [85] that the small sample size problem remains for studies of recent years. 34% of GEO datasets' sample size is less than 6, and over 57% of GEO datasets' sample size is less than 12. The complex classification boundaries make it impossible for biological interpretations. In this dissertation, we offered a classifier, a feature selection method, and a pathway analysis to overcome these two barriers.

The proposed classifier is a Bayesian extension of the TSP algorithm. We have investigated the classification accuracy rates on independent test sets. We use real datasets and simulated datasets with many different scenarios. Results show that the proposed algorithm is a clear winner over current algorithms in the overall accuracy rate and for most of the individual cases.

The BTSP algorithm is introduced as a feature selection method. We compare the performance of the BTSP algorithm with many well-known feature selection algorithms. We combine each feature selection method with SVM, kNN, and NB, and check their performances for different numbers of features/genes on different datasets. Results show that the BTSP algorithm has the best accuracy rates for many scenarios.

We also proposed a novel biological pathway data-based algorithm which takes all pairwise interactions in the gene level and pathway level. Discriminative performance of the proposed algorithm is compared with Log-Likelihood Ratio (LLR) [70], Ranking Log-Likelihood Ratio (R-LLR) [74], Top Scoring Pathway Pair (TSPP) [75], Gene Expression Deviation (GED) [76], and Pathway Variance (PathVar) [77]. It has been shown that the proposed algorithm has the best overall performance. We also check the ability to find the biologically validated pathway of the proposed algorithms, along with all algorithms mentioned above, Over-Representation Analysis (ORA), and Gene Set Enrichment Analysis (GSEA). The proposed algorithm was able to find the target pathway within the top ten pathways it selected; however, any other method could not.

We believe the proposed algorithm will have great contributions to find informative/discriminative genes and pathways related to different phenotypes/diseases.

There are many extensions of the TSP algorithm such as TST [22], TSM [20], TSN [89], TSG [19], etc. The BTSP algorithm can assume each of these extensions, and we will try these extensions with the BTSP algorithm.

We can also extend the proposed pathway analysis method to check the variance change of skill parameters within a pathway, like in PathVar. We are also planning to extend the R-LLR algorithm from a Bayesian ranking perspective.

REFERENCES

[1] A. M. Hosey, J. J. Gorski, M. M. Murray, J. E. Quinn, W. Y. Chung, G. E. Stewart, C. R. James, S. M. Farragher, J. M. Mulligan, A. N. Scott, *et al.*, "Molecular basis for estrogen receptor α deficiency in brca1-linked breast cancer," *JNCI: Journal of the National Cancer Institute*, vol. 99, no. 22, pp. 1683–1694, 2007.

[2] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, *et al.*, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Medicine*, vol. 7, no. 6, p. 673, 2001.

[3] G. Bloom, I. V. Yang, D. Boulware, K. Y. Kwong, D. Coppola, S. Eschrich, J. Quackenbush, and T. J. Yeatman, "Multi-platform, multi-site, microarray-based human tumor classification," *The American Journal of Pathology*, vol. 164, no. 1, pp. 9–16, 2004.

[4] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Diagnosis of multiple cancer types by shrunken centroids of gene expression," *Proceedings of the National Academy of Sciences*, vol. 99, no. 10, pp. 6567–6572, 2002.

[5] L. B. Edelman, G. Toia, D. Geman, W. Zhang, and N. D. Price, "Two-transcript gene expression classifiers in the diagnosis and prognosis of human diseases," *BMC Genomics*, vol. 10, no. 1, p. 583, 2009.

[6] R. B. Altman, H. K. Kroemer, C. A. McCarty, M. J. Ratain, and D. Roden, "Pharmacogenomics: will the promise be fulfilled?," *Nature Reviews Genetics*, vol. 12, no. 1, p. 69, 2011.

[7] R. L. Winslow, N. Trayanova, D. Geman, and M. I. Miller, "Computational medicine: translating models to clinical care," *Science Translational Medicine*, vol. 4, no. 158, pp. 158rv11–158rv11, 2012.

[8] D. M. Simcha, *Statistical learning applied to transcriptional regulation in small N, large D domains*. The Johns Hopkins University, 2013.

[9] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*, vol. 1. Springer series in statistics, New York, 2001.

[10] R. Winslow, N. Trayanova, D. Geman, and M. Miller, "The emerging discipline of computational medicine," *Science Translational Medicine*, vol. 4, no. 158, p. 158rv11, 2012.

[11] S. Peng, Q. Xu, X. B. Ling, X. Peng, W. Du, and L. Chen, "Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines," *FEBS Letters*, vol. 555, no. 2, pp. 358–362, 2003.

[12] C.-H. Yeang, S. Ramaswamy, P. Tamayo, S. Mukherjee, R. M. Rifkin, M. Angelo, M. Reich, E. Lander, J. Mesirov, and T. Golub, "Molecular classification of multiple tumor types," *Bioinformatics*, vol. 17, no. suppl_1, pp. S316–S322, 2001.

[13] S. Bicciato, M. Pandin, G. Didone, and C. Di Bello, "Pattern identification and classification in gene expression data using an autoassociative neural network model," *Biotechnology and Bioengineering*, vol. 81, no. 5, pp. 594–606, 2003.

[14] A.-L. Boulesteix, G. Tutz, and K. Strimmer, "A cart-based approach to discover emerging patterns in microarray data," *Bioinformatics*, vol. 19, no. 18, pp. 2465–2472, 2003.

[15] H. Zhang, C.-Y. Yu, and B. Singer, "Cell and tumor classification using gene expression data: construction of forests," *Proceedings of the National Academy of Sciences*, vol. 100, no. 7, pp. 4168–4172, 2003.

[16] M. Dettling and P. Bühlmann, "Boosting for tumor classification with gene expression data," *Bioinformatics*, vol. 19, no. 9, pp. 1061–1069, 2003.

[17] Y. Qu, B.-L. Adam, Y. Yasui, M. D. Ward, L. H. Cazares, P. F. Schellhammer, Z. Feng, O. J. Semmes, and G. L. Wright, "Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients," *Clinical Chemistry*, vol. 48, no. 10, pp. 1835–1843, 2002.

[18] P. Chen, G. Ratcliff, S. Belle, J. Cauley, S. DeKosky, and M. Ganguli, "Cognitive tests that best discriminate between presymptomatic ad and those who remain nondemented," *Neurology*, vol. 55, no. 12, pp. 1847–1853, 2000.

[19] H. Wang, H. Zhang, Z. Dai, M.-s. Chen, and Z. Yuan, "Tsg: a new algorithm for binary and multi-class cancer classification and informative genes selection," *BMC Medical Genomics*, vol. 6, no. 1, p. S3, 2013.

[20] B. Afsari, U. M. Braga-Neto, D. Geman, *et al.*, "Rank discriminants for predicting phenotypes from rna expression," *The Annals of Applied Statistics*, vol. 8, no. 3, pp. 1469–1491, 2014.

[21] D. Geman, C. d'Avignon, D. Q. Naiman, and R. L. Winslow, "Classifying gene expression profiles from pairwise mrna comparisons," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, pp. 1–19, 2004.

[22] X. Lin, B. Afsari, L. Marchionni, L. Cope, G. Parmigiani, D. Naiman, and D. Geman, "The ordering of expression among a few genes can provide simple cancer biomarkers and signal brca1 mutations," *BMC Bioinformatics*, vol. 10, no. 1, p. 256, 2009.

[23] N. D. Price, J. Trent, A. K. El-Naggar, D. Cogdell, E. Taylor, K. K. Hunt, R. E. Pollock, L. Hood, I. Shmulevich, and W. Zhang, "Highly accurate two-gene classifier for differentiating gastrointestinal stromal tumors and leiomyosarcomas," *Proceedings of the National Academy of Sciences*, vol. 104, no. 9, pp. 3414–3419, 2007.

[24] M. Raponi, J. E. Lancet, H. Fan, L. Dossey, G. Lee, I. Gojo, E. J. Feldman, J. Gotlib, L. E. Morris, P. L. Greenberg, *et al.*, "A 2-gene classifier for predicting response to the farnesyltransferase inhibitor tipifarnib in acute myeloid leukemia," *Blood*, vol. 111, no. 5, pp. 2589–2596, 2008.

[25] R. R. Weichselbaum, H. Ishwaran, T. Yoon, D. S. Nuyten, S. W. Baker, N. Khodarev, A. W. Su, A. Y. Shaikh, P. Roach, B. Kreike, *et al.*, "An interferon-related gene signature for dna

damage resistance is a predictive marker for chemotherapy and radiation for breast cancer," *Proceedings of the National Academy of Sciences*, vol. 105, no. 47, pp. 18490–18495, 2008.

[26] A. C. Tan, D. Q. Naiman, L. Xu, R. L. Winslow, and D. Geman, "Simple decision rules for classifying human cancers from gene expression profiles," *Bioinformatics*, vol. 21, no. 20, pp. 3896–3904, 2005.

[27] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, *et al.*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *science*, vol. 286, no. 5439, pp. 531–537, 1999.

[28] M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus, *et al.*, "Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," *Nature Medicine*, vol. 8, no. 1, pp. 68–74, 2002.

[29] G. J. Gordon, R. V. Jensen, L.-L. Hsiao, S. R. Gullans, J. E. Blumenstock, S. Ramaswamy, W. G. Richards, D. J. Sugarbaker, and R. Bueno, "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma," *Cancer Research*, vol. 62, no. 17, pp. 4963–4967, 2002.

[30] D. Chowdary, J. Lathrop, J. Skelton, K. Curtin, T. Briggs, Y. Zhang, J. Yu, Y. Wang, and A. Mazumder, "Prognostic gene expression signatures can be measured in tissues collected in rnalater preservative," *The Journal of Molecular Diagnostics*, vol. 8, no. 1, pp. 31–39, 2006.

[31] S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer, "Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia," *Nature Genetics*, vol. 30, no. 1, pp. 41–47, 2002.

[32] M. Kuriakose, W. Chen, Z. He, A. Sikora, P. Zhang, Z. Zhang, W. Qiu, D. Hsu, C. McMunn-Coffran, S. Brown, *et al.*, "Selection and validation of differentially expressed genes in head and neck cancer," *Cellular and Molecular Life Sciences CMLS*, vol. 61, no. 11, pp. 1372–1383, 2004.

[33] F. Borovecki, L. Lovrecic, J. Zhou, H. Jeong, F. Then, H. Rosas, S. Hersch, P. Hogarth, B. Bouzou, R. Jensen, *et al.*, "Genome-wide expression profiling of human blood reveals biomarkers for huntington's disease," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 31, pp. 11023–11028, 2005.

[34] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. Wiley, New York, 1973.

[35] L. Li, C. R. Weinberg, T. A. Darden, and L. G. Pedersen, "Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the ga/knn method," *Bioinformatics*, vol. 17, no. 12, pp. 1131–1142, 2001.

[36] C. Kooperberg, M. LeBlanc, J. Y. Dai, and I. Rajapakse, "Structures and assumptions: strategies to harness gene× gene and gene× environment interactions in gwas," *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, vol. 24, no. 4, p. 472, 2009.

[37] P. Chopra, J. Lee, J. Kang, and S. Lee, "Improving cancer classification accuracy using gene pairs," *PLoS One*, vol. 5, no. 12, p. e14305, 2010.

[38] H. Zhang, H. Wang, Z. Dai, M.-s. Chen, and Z. Yuan, "Improving accuracy for cancer classification with a new algorithm for genes selection," *BMC Bioinformatics*, vol. 13, no. 1, p. 298, 2012.

[39] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[40] F. Caron and A. Doucet, "Efficient bayesian inference for generalized bradley-terry models," *Journal of Computational and Graphical Statistics*, vol. 21, no. 1, pp. 174–196, 2012.

[41] R. R. Davidson and P. H. Farquhar, "A bibliography on the method of paired comparisons," *Biometrics*, pp. 241–252, 1976.

[42] P. McCullagh and J. A. Nelder, *Generalized linear models*, vol. 37. CRC Press, 1989.

[43] T. Hastie and R. Tibshirani, "Classification by pairwise coupling," in *Advances in Neural Information Processing Systems*, pp. 507–513, 1998.

[44] D. R. Hunter, "Mm algorithms for generalized bradley-terry models," *Annals of Statistics*, pp. 384–406, 2004.

[45] S. Hussain and M. Aslam, "Bayesian inference of log-linear version of the bradley-terry model for paired comparisons using uninformative prior," *International Journal of Probability and Statistics*, vol. 2, no. 3, pp. 43–49, 2013.

[46] J. Guiver and E. Snelson, "Bayesian inference for plackett-luce ranking models," in *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 377–384, ACM, 2009.

[47] E. S. Adams, "Bayesian analysis of linear dominance hierarchies," *Animal Behaviour*, vol. 69, no. 5, pp. 1191–1201, 2005.

[48] I. C. Gormley, T. B. Murphy, *et al.*, "A grade of membership model for rank data," *Bayesian Analysis*, vol. 4, no. 2, pp. 265–295, 2009.

[49] I. C. Gormley and T. B. Murphy, "Exploring voting blocs within the irish electorate: A mixture modeling approach," *Journal of the American Statistical Association*, vol. 103, no. 483, pp. 1014–1027, 2008.

[50] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences*, vol. 96, no. 12, pp. 6745–6750, 1999.

[51] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. Kim, L. C. Goumnerova, P. M. Black, C. Lau, *et al.*, "Prediction of central nervous system embryonal tumour outcome based on gene expression," *Nature*, vol. 415, no. 6870, pp. 436–442, 2002.

[52] E. Tian, F. Zhan, R. Walker, E. Rasmussen, Y. Ma, B. Barlogie, and J. D. Shaughnessy Jr, "The role of the wnt-signaling antagonist dkk1 in the development of osteolytic lesions in multiple myeloma," *New England Journal of Medicine*, vol. 349, no. 26, pp. 2483–2494, 2003.

[53] K. Chin, S. DeVries, J. Fridlyand, P. T. Spellman, R. Roydasgupta, W.-L. Kuo, A. Lapuk, R. M. Neve, Z. Qian, T. Ryder, *et al.*, "Genomic and transcriptional aberrations linked to breast cancer pathophysiologies," *Cancer Cell*, vol. 10, no. 6, pp. 529–541, 2006.

[54] J. Damond, "ktspair: k-top scoring pairs for microarray classification, 2011," *R package version*, vol. 1.

[55] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1, pp. 389–422, 2002.

[56] T. Li, C. Zhang, and M. Ogihara, "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression," *Bioinformatics*, vol. 20, no. 15, pp. 2429–2437, 2004.

[57] L. Le Marchand, T. Donlon, A. Seifried, R. Kaaks, S. Rinaldi, and L. R. Wilkens, "Association of a common polymorphism in the human gh1 gene with colorectal neoplasia," *Journal of the National Cancer Institute*, vol. 94, no. 6, pp. 454–460, 2002.

[58] S. Khoury-Shakour, S. B. Gruber, F. Lejbkowicz, H. S. Rennert, L. Raskin, M. Pinchev, and G. Rennert, "Recreational physical activity modifies the association between a common gh1 polymorphism and colorectal cancer risk," *Cancer Epidemiology and Prevention Biomarkers*, vol. 17, no. 12, pp. 3314–3318, 2008.

[59] J. Hua, W. D. Tembe, and E. R. Dougherty, "Performance of feature-selection methods in the classification of high-dimension data," *Pattern Recognition*, vol. 42, no. 3, pp. 409–424, 2009.

[60] Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha, and G. E. Robinson, "Big data: astronomical or genomical?," *PLoS Biol*, vol. 13, no. 7, p. e1002195, 2015.

[61] D. Koller and M. Sahami, "Toward optimal feature selection," tech. rep., Stanford InfoLab, 1996.

[62] E. Arslan and U. M. Braga-Neto, "A bayesian approach to top-scoring pairs classification," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pp. 871–875, IEEE, 2017.

[63] J. Bins and B. A. Draper, "Feature selection from huge feature sets," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 2, pp. 159–165, IEEE, 2001.

[64] M. West-Nielsen, E. V. Høgdall, E. Marchiori, C. K. Høgdall, C. Schou, and N. H. Heegaard, "Sample handling for mass spectrometric proteomic investigations of human sera," *Analytical Chemistry*, vol. 77, no. 16, pp. 5114–5123, 2005.

[65] S. Mahadevan, S. L. Shah, T. J. Marrie, and C. M. Slupsky, "Analysis of metabolomic data using support vector machines," *Analytical Chemistry*, vol. 80, no. 19, pp. 7562–7570, 2008.

[66] Y. Tang, Y.-Q. Zhang, and Z. Huang, "Development of two-stage svm-rfe gene selection strategy for microarray expression data analysis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 4, no. 3, pp. 365–381, 2007.

[67] P. Romanski and L. Kotthoff, "Fselector: selecting attributes," *Vienna: R Foundation for Statistical Computing*, 2009.

[68] I. Kononenko, "Estimating attributes: analysis and extensions of relief," in *European Conference on Machine Learning*, pp. 171–182, Springer, 1994.

[69] K. Kira and L. A. Rendell, "The feature selection problem: Traditional methods and a new algorithm," in *AAAI*, vol. 2, pp. 129–134, 1992.

[70] J. Su, B.-J. Yoon, and E. R. Dougherty, "Accurate and reliable cancer classification based on probabilistic inference of pathway activity," *PloS one*, vol. 4, no. 12, p. e8161, 2009.

[71] J. Su, B.-J. Yoon, and E. R. Dougherty, "Identification of diagnostic subnetwork markers for cancer in human protein-protein interaction network," in *BMC Bioinformatics*, vol. 11, p. S8, BioMed Central, 2010.

[72] U. M. Braga-Neto, E. Arslan, U. Banerjee, and A. Bahadorinejad, "Bayesian classification of genomic big data," in *Signal Processing and Machine Learning for Biomedical Big Data*, pp. 411–427, CRC Press, 2018.

[73] E. Arslan and U. M. Braga-Neto, "Bayesian top scoring pairs for feature selection," in *2017 51st Asilomar Conference on Signals, Systems, and Computers*, pp. 387–391, IEEE, 2017.

[74] N. Khunlertgit and B.-J. Yoon, "Identification of robust pathway markers for cancer through rank-based pathway activity inference," *Advances in Bioinformatics*, vol. 2013, 2013.

[75] E. Glaab, J. M. Garibaldi, and N. Krasnogor, "Learning pathway-based decision rules to classify microarray cancer samples," 2010.

[76] M. R. Young and D. L. Craft, "Pathway-informed classification system (pics) for cancer analysis using gene expression data," *Cancer Informatics*, vol. 15, pp. CIN–S40088, 2016.

[77] E. Glaab and R. Schneider, "Pathvar: analysis of gene and protein expression variance in cellular pathways using microarray data," *Bioinformatics*, vol. 28, no. 3, pp. 446–447, 2011.

[78] Z. Liu, Z. Yao, C. Li, Y. Lu, and C. Gao, "Gene expression profiling in human high-grade astrocytomas," *Comparative and Functional Genomics*, vol. 2011, 2011.

[79] H. Pei, L. Li, B. L. Fridley, G. D. Jenkins, K. R. Kalari, W. Lingle, G. Petersen, Z. Lou, and L. Wang, "Fkbp51 affects cancer cell response to chemotherapy by negatively regulating akt," *Cancer Cell*, vol. 16, no. 3, pp. 259–266, 2009.

[80] M. E. Lenburg, L. S. Liou, N. P. Gerry, G. M. Frampton, H. T. Cohen, and M. F. Christman, "Previously unidentified changes in renal cell carcinoma gene expression identified by parametric analysis of microarray data," *BMC Cancer*, vol. 3, no. 1, p. 31, 2003.

[81] J. Sabates-Bellver, L. G. Van der Flier, M. de Palo, E. Cattaneo, C. Maake, H. Rehrauer, E. Laczko, M. A. Kurowski, J. M. Bujnicki, M. Menigatti, *et al.*, "Transcriptome profile of human colorectal adenomas," *Molecular Cancer Research*, vol. 5, no. 12, pp. 1263–1275, 2007.

[82] Y. Hong, T. Downey, K. W. Eu, P. K. Koh, and P. Y. Cheah, "A 'metastasis-prone'signature for early-stage mismatch-repair proficient sporadic colorectal cancer patients and its implications for possible therapeutics," *Clinical & Experimental Metastasis*, vol. 27, no. 2, pp. 83–90, 2010.

[83] I. B. P. Ni, Z. Zakaria, R. Muhammad, N. Abdullah, N. Ibrahim, N. A. Emran, N. H. Abdullah, and S. N. A. S. Hussain, "Gene expression patterns distinguish breast carcinomas from normal breast tissues: the malaysian context," *Pathology-Research and Practice*, vol. 206, no. 4, pp. 223–228, 2010.

[84] A. Tripathi, C. King, A. De la Morenas, V. K. Perry, B. Burke, G. A. Antoine, E. F. Hirsch, M. Kavanah, J. Mendez, M. Stone, *et al.*, "Gene expression abnormalities in histologically normal breast epithelium of breast cancer patients," *International Journal of Cancer*, vol. 122, no. 7, pp. 1557–1566, 2008.

[85] C. Yu, H. J. Woo, X. Yu, T. Oyama, A. Wallqvist, and J. Reifman, "A strategy for evaluating pathway analysis methods," *BMC Bioinformatics*, vol. 18, no. 1, p. 453, 2017.

[86] Z. Guo, T. Zhang, X. Li, Q. Wang, J. Xu, H. Yu, J. Zhu, H. Wang, C. Wang, E. J. Topol, *et al.*, "Towards precise classification of cancers based on robust gene functional expression profiles," *BMC Bioinformatics*, vol. 6, no. 1, p. 58, 2005.

[87] E. Lee, H.-Y. Chuang, J.-W. Kim, T. Ideker, and D. Lee, "Inferring pathway activity toward precise disease classification," *PLoS Computational Biology*, vol. 4, no. 11, p. e1000217,

2008.

[88] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He, "clusterprofiler: an r package for comparing biological themes among gene clusters," *Omics: a journal of integrative biology*, vol. 16, no. 5, pp. 284–287, 2012.

[89] A. T. Magis and N. D. Price, "The top-scoring 'n'algorithm: a generalized relative expression classification method from small numbers of biomolecules," *BMC Bioinformatics*, vol. 13, no. 1, p. 227, 2012.