DETECTION OF IMPORTANT GENES IN A GENE REGULATORY NETWORK AND EFFECTS OF GENE INTERACTION IN PHENOTYPE CLASSIFICATION

A Dissertation

by

EUNJI KIM

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,	Edward R. Dougherty
Co-Chair of Committee,	Ivan Ivanov
Committee Members,	Ulisses Braga-Neto
	Xiaoning Qian
	Philip Hemmer
Head of Department,	Miroslav M. Begovic

December 2019

Major Subject: Electrical Engineering

ABSTRACT

The recent advancements in high-throughput technologies provide a wealth of information on gene expression patterns and gene-regulatory pathways. As a result, researchers in life sciences have an unprecedented opportunity for more sophisticated, integrative and holistic approaches to identify phenotype-associated (signaling) molecular markers. Biomarker discovery is one of the most important goals in bioinformatics; however, achieving this objective requires comprehensive analysis of gene expression profiles and gene-gene interactions that exist in high-dimensional data spaces. In this dissertation, we are concerned with critical issues that hinder biomarker discovery.

In the first part, we focus on the effects of measurement platforms on ranking of genes. Analyzing gene expression patterns and selecting informative genes amounts to a supervised classification and feature selection. However, if the sample is small, error estimation is problematic and the performance of the feature-selection algorithm will be impacted by the performance of the error estimator. The problem is compounded by the fact that the accuracy of classification depends on the manner in which the phenomena are transformed into data by the measurement technology. Therfore, the first part of this dissertation is devoted to the study of the effects of the nonlinear transformation of the actual gene concentrations introduced by a sequencing machine on the feature-set ranking.

The second part of this dissertation is devoted to *canalizing genes* which possess an ability to correct abnormal cellular processes for the purpose of biological robustness under genetic mutations or environmental perturbations. Despite their central role in gene regulatory networks (GRNs), the observation/detection of canalizing genes is often impeded because of their particular behavior. Therefore, we focus on inherent characteristics of canalizing genes and develop a quantitative framework that allows for the estimation of the power of canalizing genes in the context of Boolean Networks with perturbations (BN$_p$s). We also consider the problem of reducing the network complexity while preserving the distribution of the canalizing power of genes. We evaluate the stability of canalizing power under network reduction and proceed with the problem of selecting the relevant network features that allow for discriminating reducible networks which are determined by the degree of preservation of canalizing power.

DEDICATION

To my parents, sister, and grandmother Park

ACKNOWLEDGEMENTS

This thesis would not have been possible without the inspiration and support of a number of wonderful individuals — my thanks and appreciation to all of them for being part of this journey and making this thesis possible. I owe my deepest gratitude to my advisors, Prof. Edward R. Dougherty and Prof. Ivan Ivanov. I feel honored to have had a chance to work with Dr. Dougherty and learn from him. His continuous dedication to his students and high-quality research is admirable and inspirational. Dr. Ivanov was always generous with his time despite his overwhelming schedule. My appreciation for his guidance and continuous support is immeasurable.

I express my warmest gratitude to Dr. Jianping Hua. His expertise in modeling was a siginificant influence in shaping the simulations presented in this thesis. I would also like to thank to Dr. Ulisses M. Braga-Neto, Dr. Xiaoning Qian and Dr. Philip Hemmer for serving on my committee and for their constructive advice.

Finally, I would like to thank my parents Jungrae Kim and Milim Chang. Without their love and ultimate support in all its forms, I would not be where I am today. I will always appreciate all they have done to encourage me throughout the entire doctorate program. I dedicate my dissertation work to my beloved sister Hyunji who is the reason for my existence. I owe every achievement in my life to my grandmother Guirye Park who gave me her unconditional love and has never left my side. A heartfelt thank goes out to all students in the Genomic Signal Processing Lab who provided a generous source of support, inspiration, and motivation along the way.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Professor Edward R. Dougherty, Professor Ulisses Braga-Neto, Professor Xiaoning Qian and Professor Philip Hemmer of the Department of Electrical and Computer Engineering and Professor Ivan Ivanov of the Department of Veterinary Physiology and Pharmacology.

Funding Sources

This work was funded in part by grants from the National Institutes of Health, U01 CA-162077, R35CA-197707, and P30ES-023512.

Graduate study was supported by a Chevron Scholarship from the Department of Electrical and Computer Engineering and a scholarship from Korean American Scholarship Foundation.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

1. INTRODUCTION

Biomarkers discovery is one of the most important topics in bioinformatics because it provides targets for diagnosis, prognosis and therapeutic interventions [1]-[5]. The development of high-throughput technologies enables researchers to measure thousands of genes simultaneously in one single experiment and has fueled the growth of the field of computational biology and bioinformatics [6]-[9]. While obtaining large-scale gene expression profiles of cells should theoretically accelerate the identification of uncovered biomarkers, high dimesnsionality has become the bottleneck of data analysis [10], [11]. The plethora of complex omics data has complicated the problem of extracting meaningful molecular signatures. Moreover, the great number of features is often accompanied by a small number of available samples [12], [13]. The main challenge is that the amount of data required to provide a reliable analysis grows exponentially as the dimensionality of the data rises. Bellman referred to this phenomenon as the *curse of dimensionality* which describes the inherent limitation of high dimensionality ($p \gg n$ where p is the number of dimensions and n is the sample size) [14]. The curse of dimensionality leads to a peaking phenomenon – adding more features will degrade the performance of a classifier [15]. Therefore, feature selection is imperative to alleviate the problem of high dimensionality.

Feature selection not only serves as a strategy to tackle the high-dimensionality problem, but also provides a framework for biomarker discovery [16]-[19]. A large number of methods have been proposed for the identification of molecular markers and they can be categorized into two main approaches: (i) classical univariate statistical methods, where each biomarker is considered as independent from the others; (ii) multivariate methods that take into account the correlation structure of the data and interactions existing among the genes.

In the univariate scheme, statistical tests are applied to each biomarker candidate individually to evaluate the statistically significant differences between two groups of samples, e.g. normal cells vs. cancer cells [17], [20]. There are numerous univariate statistcal methods such as t-test, ANOVA, and non parametric tests, like the Mann-Whitney test, Kruskal-Wallils test, and Chi-squre test [18]-[21]. Although these methods provide useful information regarding differences in gene expression between conditions, they have a major drawback: the potential correlation and the synergic or antagonistic effects between groups of genes are not considered [19], [22], [23]. There are some differential expression analysis methods implemented in software packages such as edgeR and DESeq2 that model relationships between genes; however, the statistical test for assessing significance is still a univariate test performed independently for each gene [24], [25]. Genes are known to play key roles as a group in the cell but not when each gene is considered independently [26]-[28]. Clearly, the univariate testing for significance is not reflecting the gene-gene interactions that are present in the cell. Therefore, biological considerations imply that biomarker discovery should be performed in a multivariate setting.

Multivariate methods compare two or more groups of samples considering the relationships existing between the candidate molecules, i.e. concordant or discordant effects of different factors [20]. These methods usually belong to one of the following categories: (i) unsupervised pattern recognition methods, e.g. clustering methods; (ii) supervised classification methods, based on a priori information about the membership of each sample to a specific class. Clustering algorithms in unsupervised learning allow the identification of groups of samples or features in a dataset [29]. The samples are grouped on the basis of a measure of their similarity and attributes selected from a clustering algorithm are thus the representative attributes of the same cluster [30], [31]. However, ground truth class labels that can evaluate the performance of selection are not usually available and thus, one cannot evaluate the accuracy of the clustering. On the other hand, samples are given with known lables in supervised learning and this allows outcome assessment with misclassification error. Therefore, the discovery of biomarkers is often modeled as feature selection based on the use of supervised learning methods [16]-[18], [20]. For this reason, we deploy classification approaches throughout this dissertation.

The accurate selection of biomarker candidates is crucial, because it determines the outcome of further validation studies and the ultimate success of efforts to develop diagnostic and prognostic assays with high specificity and sensitivity. During this process, three important issues have to be addressed: (i) the small number of available samples, (ii) the effects of the measurement technology on the obtained data, and (iii) the appropriate mathematical and statistical framework for modeling the activity of important genes. The following sections discuss these issues in greater detail.

1.1 Feature selection for high dimensional and small-sample size datasets

When feature selection is based on classification error, the goodness of feature sets is determined by their error estimates. The performance of feature selection or feature-set ranking concerns the relationship between the true ordering and ranking based on error estimates. The most critical aspect of evaluating the performance is the degree of preserving the true ordering. In many studies [16], [32], [33], it has been experimentally verified that the relatively small number of samples in high-dimensional data is one of the main sources of the problem of feature selection. Moreover, Ein-Dor *et al*. [34] showed that at least thousands of samples are needed to generate a robust list that can be used for predicting outcome in cancer patients. If there is a large data set, one can obtain good error estimates; however, if the sample size is small, the performance of the feature-selection is affected to a great extent by the performance of the error estimator [35]-[37]. Previous studies [35], [36] focused on the role of error estimators, feature-selection algorithms and classification rules in feature selection for small samples and compared the estimated results with the absolute/true ranking of feature sets. All feature sets of a given size were used to design classifiers and the feature sets were ranked based on their true and estimated errors. U. M. Braga-Neto *et al.* showed that with small samples and a large number of features, error estimators have substantial variance, especially cross-validation, and possess little correlation or regression with the true error [35], [36]. Moreover, it was shown in [38] that the performance differences among the feature-selection algorithms appear to be less significant than differences in performance among the error estimators used to implement the algorithms for small samples. Studies in [35]-[37] also indicated that error estimators suffer from different degrees of imprecision in the small-sample setting.

All of these previous studies suggest that feature selection algorithms are unreliable in the small sample setting [35], [36], [39]. Therefore, one may have poor feature sets whose corresponding classifiers possess errors far in excess of the classifier corresponding to the optimal feature set [37]. Zhao *et al*. [39] also showed that the estimated errors for the top features may be biased by a low error estimate, and thereby selecting a top scored feature set can be misleading. Therefore, Zhao *et al*. [39] suggested that rather than reporting a single feature set, providing a list of the best performing feature sets increases the likelihood of finding good features sets when samples are small. This is based on an idea that some feature sets in that list will be close to optimal.

1.2 Feature selection using gene-expression data obtained from NGS pipeline

Next-generation sequencing (NGS) refers to a class of technologies that sequence millions of short DNA fragments in parallel [40]. NGS has rapidly become the method of choice for transcriptional profiling experiments due to many advantages compared to the available microarray expression platforms [41]. In contrast to microarray technology, the high throughput sequencing allows the identification of novel transcripts and isoforms and does not require a sequenced genome [9]. Furthermore, the background correction, probe design and spot filtering, which are typical for microarray-based technology, are no longer problematic due to the different nature of NGS technology [12], [42].

The specific application of NGS for RNA sequencing is called RNA-Seq, which is a high-throughput measurement of gene-expression levels of thousands of genes simultaneously as represented by discrete expression values for regions of interest on the genome (e.g. genes) [7], [9], [43]. In particular, RNA-Seq sequences small RNA fragments (mRNA) which are produced when a gene is expressed [7]. The schematic of the key steps of RNA-Seq analysis pipeline from sample preparation to data analysis is illustrated in Figure 1.1. The RNA-Seq experiment randomly shears and converts the RNA fragments to cDNAs, sequences them, and finally outputs the results in the form of short reads [12], [42], [44]. Then, cDNA fragment reads are mapped back to a reference genome to determine the gene-expression levels [6], [12]. The technology assumes that the cDNA cleavage is random and so the read start position is independent of the genomic sequence [45]. Therefore, it allows to use the number of reads mapping to certain regions of the genome as a quantitative measurement. The number of reads mapped to a gene on the reference genome defines the count data, which is a discrete measure of the gene-expression levels [12]. RNA-seq experiments are subject to some systematic variations such as library size (i.e., sequencing depth) differences between samples, as well as transcript length bias and GC content within a specific sample [46], [47]. Therefore, it should be noted that it is essential to normalize data in order to adjust for such biases. After correcting systematic variations within and between samples, downstream analysis such as differential expression analysis, multivariate statistical analysis and visualization is performed [47].

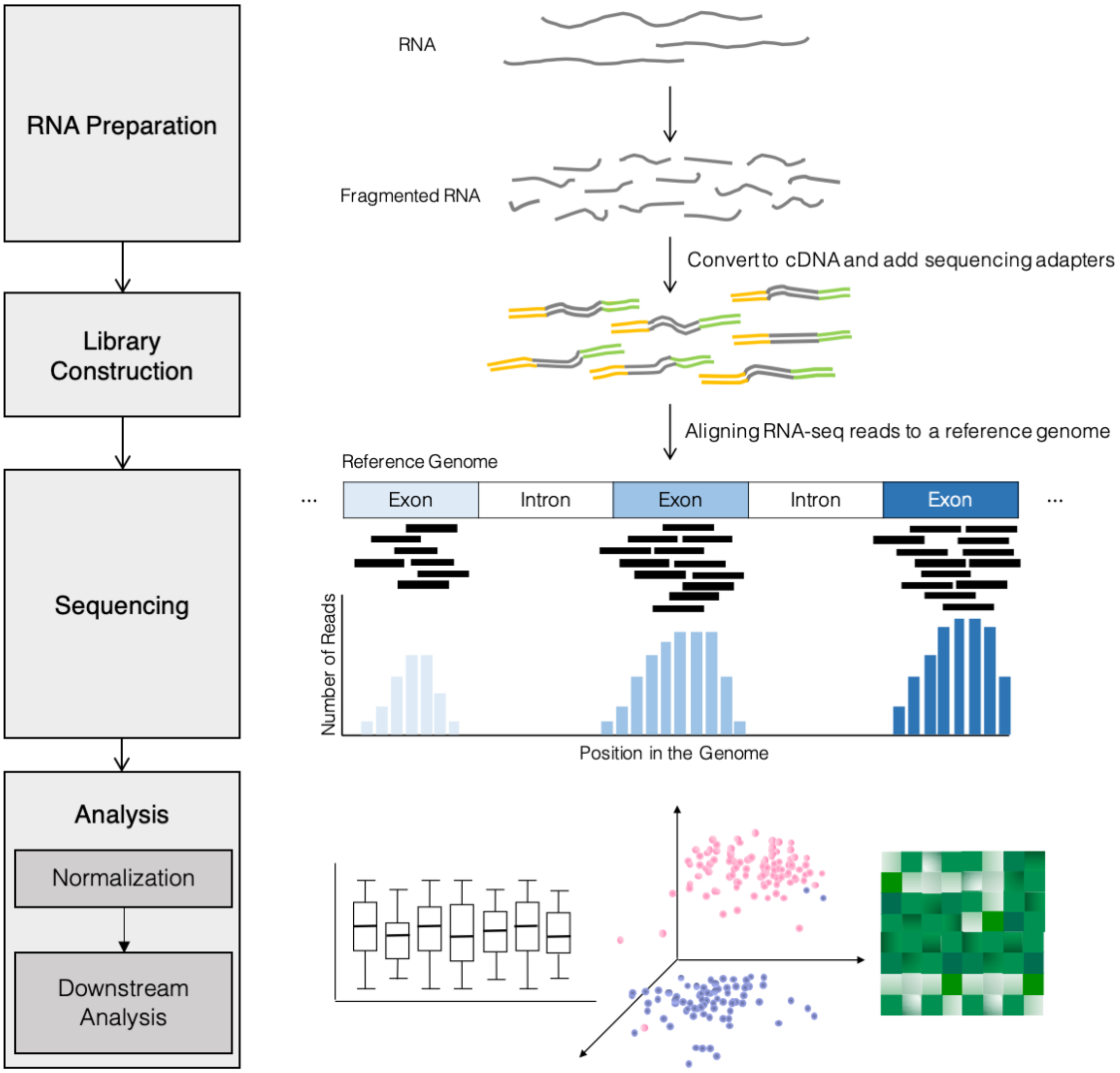


Figure 1.1 A typical RNA-seq experiment consists of the following steps: RNA extraction, library preparation, sequencing of the samples, and data analysis. Purified RNA samples are sent for library preparation and sequencing. RNA-seq reads are aligned to a reference genome and the expression level of each gene is estimated by counting the number of reads that align to each exon or full-length transcript. Downstream analyses with RNA-Seq data include differential expression analysis, multivariate statistical analysis and visualization, etc.

Many studies have focused on modeling the discrete NGS data obtained from the sequencing instrument. Two popular models for statistical representation of the discrete

NGS data are the negative binomial [24], [48] and Poisson [43]. The Poisson model is completely parameterized by its mean and thus is known to exhibit problems in fitting RNA-Seq data because RNA-Seq generates gene-expression data with overdispersion where the variance exceeds the mean [49], [50]. Therefore, the counts are frequently modeled by the negative binomial distribution [12], [24], [48]. However, with the relatively small number of samples available in most current NGS experiments, it is difficult to accurately estimate the dispersion parameter of the negative binomial model [12]. Therefore, Noushin *et al*. [12] modeled the NGS data using a hierarchical, multivariate Poisson (MP) model. Specifically, gene concentration levels are modeled using a log-normal distribution [51], [52] and the sequencing instrument sampling of these is modeled via a Poisson process [12], [49]. Therefore, the read counts are not marginally Poisson distributed, but they are modeled as conditionally Poisson where the RNA-Seq data overdispersion is obtained by marginal variance calculations.

With the widespread use of the NGS techniques, many studies have been conducted to examine classification approaches for sequencing data [53]-[55]. Noushin *et al*. [12] studied how the NGS processing pipeline affects classification performance. This work shows that the NGS pipeline transforms the original Gaussian data and produces less discriminative data relative to the actual gene expression levels which diminishes classification accuracy. Those results show that the accuracy of classification and feature selection depends on the manner in which the phenomena are transformed into data by the measurement technology in addition to the choice of error estimators and the availability of enough samples.

1.3 Important genes in a gene regulatory network

In Biology, genes that occupy the top of a regulatory hierarchy and control multiple downstream genes either directly or by initiating a cascade of gene responses are called *master* or *canalizing genes* [56]-[59]. Genes that are under the regulatory influence of such commanders are called *slave genes*. It is important to note that these concepts are both relative and local. A gene considered as a master for a given portion of a regulatory network could be a slave from another local perspective or in a different context. Both master and canalizing genes play a key role in the variety of processes such as cellular development and differentiation and exert a strong control over many downstream gene pathways.

The conceptual difference between master and canalizing genes is that canalizing genes have an additional ability of taking over the control and overriding other regulatory instructions [56], [57], [60]. One example is a pathway involving DUSP1 and Ras genes which are important in melanoma tumors. Figure 1.2 shows a DUSP1 network and it illustrates the canalizing characteristic of DUSP1. When DUSP1 is OFF, or down-regulated, the downstream genes are controlled by the Ras oncogene through phosphorylation and transcriptional activation. However, when DUPS1 is activated, or up-regulated, it dephosphorylates ERK1/2, thereby overriding the signal sent by Ras [56]. The biological role of DUSP1 indicates that it is likely a canalizing gene that can take over control of downstream genes when it is ON.

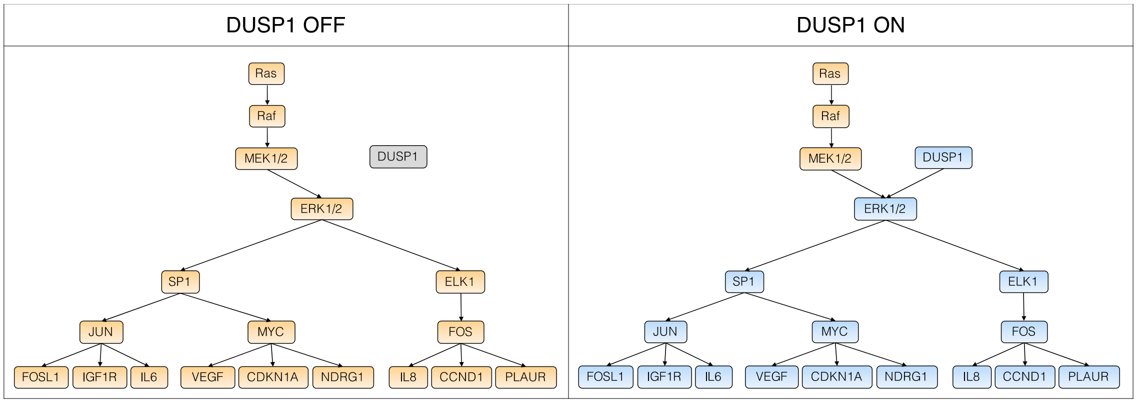
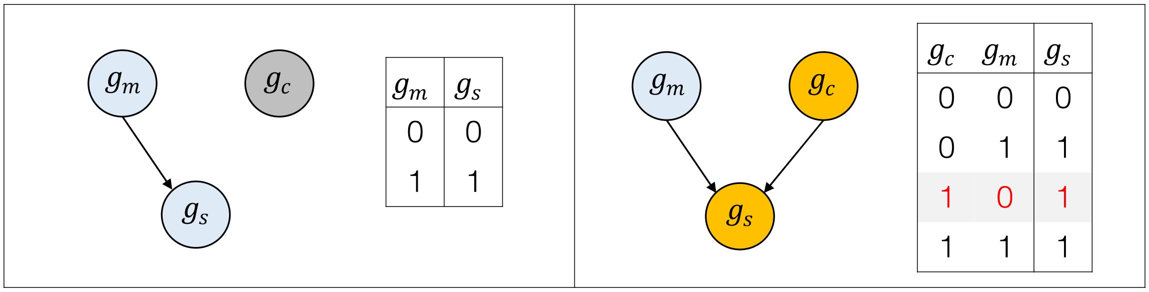


Figure 1.2 The regulatory pathway including DUSP1 and Ras constructed from canonical pathway knowledge presented in [56]. Orange color represents the regulatory influence of Ras while blue color indicates the effects of DUSP1 activation.

Moreover, canalizing genes produce adaptive and optimal reactions for the purpose of biological robustness when there are external stimuli or genetic perturbations [60]-[62]. They enforce corrective actions on cellular processes to maintain homeostasis and buffer itself from the effects of random alterations or operating errors [56], [57], [60]. Despite their central role in biological systems, the observation/detection of canalizing genes is often impeded because the behavior of affected genes is highly varied relative to the inactive canalizer [57], [60]. For example, cellular p53 is expressed at low levels under normal physiological conditions, thereby turning off the activity of p53 network [63]. Most of the time, canalizing gene such as p53 is turned off or exists in the cell at a very low level in a relatively inactive mode. Therefore, the activity of canalizing genes is difficult to predict to any significant degree by their subject genes under normal cell conditions.



g_m	g_s
0	0
1	1

g_c	g_m	g_s
0	0	0
0	1	1
1	0	1
1	1	1

Figure 1.3 A pictorial representation of the activity of a canalizing gene g_c. Corresponding wiring diagram and truth table are presented. A left box shows the case where the slave gene g_s is completely regulated by g_m when g_c is inactive. The box in the right shows that g_c overrides the instructions sent by g_m, so the expression of g_s is determined by g_c regardless of g_m. Blue and yellow color indicates the influence of g_m and g_c, respectively.

If a strict complete control is applied in a closed environment, it is obvious to identify a canalizing gene and measure its effects on other genes. Figure 1.3 illustrates the situation where the effects of canalizing genes are easily found when all the genes are deterministically controlled. Assume that a master gene g_m is known to turn on g_s when g_m is ON and turn off g_s when it is OFF, i.e. $g_m \mapsto \text{ON} \Longrightarrow g_s \mapsto \text{ON}$, $g_m \mapsto \text{OFF} \Longrightarrow g_s \mapsto \text{OFF}$. With the full knowledge of this rule, the effects of canalizing genes are easily estimated if change of slave genes' expression under the same condition of master gene is observed as shown in Figure 1.3. By looking at the case where the gene g_s is turned on even with inactive g_m, we can infer that g_c turns on g_s. When no latent variable and perturbation are assumed, the changes of slave genes' expression are entirely due to the activity of a canalizing gene and one can conclude that g_c determines the state of slave genes regardless of the activity of g_m upon its activation. However, such a strict complete control is extremely unlikely in practice and it is hard to tell whether the changes observed in slave genes under the same conditions of master genes are due to latent genes, noise or activation of a canalizing gene. This suggests that it is hard to detect canalizing genes from gene-expression data only which in turn requires a mathematical framework that can capture the activity of the genes.

1.4 Contributions

In this dissertation, we consider aforementioned factors and conduct three distinct research projects. First, we study the effects of a next-generation sequencing measurement platform on the ranking of feature sets. We then formulate a mathematical framework that can measure the power of important genes which reflects their intrinsic characteristics. Finally, we apply the proposed framework to characterize Boolean networks that are reducible under the constraint of preserving the the canalizing power of genes and identify relevant network attributes that can detect/discriminate such networks.

The work in Section 2 builds on previous work by Zhao *et al*. [39] where they studied the effectiveness of reporting list of feature sets for multivariate Gaussian model in the small-sample setting. We extend this study by applying methodologies suggested in [39] to sequencing data because it is a widespread biological measurement technique which does not conform to Gaussian distributional assumptions. We use a hierarchical Poisson model [12], [49] to represent the NGS sample processing pipeline: from the biological sample to the gene counts. We investigate the performance of feature-set ranking and compare the list of selected features derived from a model of RNA-Seq data with top ranked features from a multivariate normal model of gene concentrations. Three measures are employed for comparison: (i) ranking power, (ii) length of extensions, and (iii) Bayes features. We perform the model-based study to examine the effectiveness of reporting lists of small feature sets using RNA-Seq data and the effects of different model parameters and error estimators on the ranked list.

In Section 3, we present a quantitative framework that reflects inherent characteristics of canalizing genes and allows for the estimation of the power of canalizing genes in the context of Boolean Networks with perturbation (BN$_p$s). We define the canalizing power (CP) using two terms: regulation power (RP) and incapacitating power (IP). We base this assumption on the idea that canalizing power of a gene should be quantified by the extent of its regulation on the overall network and the extent of control that such gene takes away from other master genes when it is activated. Following this, the CP concept is demonstrated on synthetic and real data to provide preliminary evidence that CP can be used to characterize the ability of canalizing genes.

In Section 4, we study the problem of reducing BNs with a perturbation by consecutively removing genes with the smallest canalizing power. A systematic empirical study demonstrates that there are two classes of networks, reducible and irreducible with respect to the preservation of canalizing power of the genes. With these observations in mind, we introduce the definition of reducible networks based on two criteria: (i) Spearman’s rank-order correlation coefficient, and (ii) weighted Euclidean distance between canalizing power vectors of the original and reduced networks. From the perspective of optimal control, investigating attributes of the network in a certain class is viewed as obtaining the best estimates of conditions which are most likely to elicit a particular behavior of the network. Therefore, to understand inherent properties of the networks in the two different classes, we proceed with the problem of selecting their relevant network features that allow for discriminating reducible from irreducible networks. Discriminant features are obtained from simulated networks of 12 genes and the efficacy of the selected features is demonstrated on synthetic networks of 13 genes and a real 16-gene p53 regulatory network.

In Section 5, we summarize the main contributions of the work and discuss some future directions of research.

2. THE MODEL-BASED STUDY OF THE EFFECTIVENESS OF REPORTING LISTS OF SMALL FEATURE SETS USING RNA-SEQ DATA[1]

2.1 Introduction

Ranking feature sets for phenotype classification based on gene expression can be viewed as gene selection and is a key issue for cancer informatics. Because ranking feature sets is often based on error estimates of the designed classifiers and error estimators based on training data from small samples tend to perform poorly, exhibiting optimistic bias or high variance, a feature set with a low error estimate cannot be automatically declared to be credible. Also, it is important to choose an error estimator which yields a reliable ranking for the feature sets [35]. Furthermore, when confronted with a small sample, feature-selection algorithms often fail to find good feature sets. The problem is exacerbated for high-dimensional data, i.e., data sets with feature sets of high cardinality. It is difficult to find a good feature set in the small-sample setting even when one uses a mathematically favorable gene concentration/expression model [39]. These observations suggest that it is prudent to report a list of potential feature sets rather than attempting to find the best feature set. In addition to the unreliability of feature selection and error estimation, the accuracy of classification depends on the manner in which the phenomena

[1] Part of this section is reprinted with permission (CC by 4.0) from “The Model-Based Study of the Effectiveness of Reporting Lists of Small Feature Sets Using RNA-Seq Data” by E. Kim, I. Ivanov, J. Hua, J. W. Lampe, M. A. Hullar, R. S. Chapkin, and E. R. Dougherty, *Cancer Informatics*, vol. 16, Feb. 2017.

are transformed into data by the measurement technology. High-throughput sequencing technologies such as NGS have recently emerged as popular tools to quantify gene transcripts. However, NGS technologies pose new computational and statistical challenges because their applications result in nonlinear transformations of the underlying gene-concentration distributions. A recent study showed that a NGS pipeline could lead to transformation degradation in classification performance [12]. In this section, we address the effects of the nonlinear transformation induced by the sequencing machine and the choice of error estimators on feature-set ranking.

The development of NGS technologies enables simultaneous measurements of the abundance of mRNA transcripts and such information can be utilized to detect differential gene expression and design gene-expression-based classifiers for phenotypic discrimination and medical diagnosis or prognosis. RNA-Seq provides discrete counting measurements for the gene-expression levels [44]. All RNA-Seq data generation follows a similar protocol, starting with shearing samples to generate millions of small RNA fragments. These fragments are then converted to cDNA and the adapter sequences are ligated to their ends. This collection, referred to as a library, is then sequenced, which produces millions of short sequence reads that correspond to individual cDNA fragments. Finally, those reads are mapped to a reference genome. The number of reads mapped to a gene on the reference genome defines the count data, which is a discrete measure of the respective gene expression levels.

Much of the literature concerning the statistical representation of RNA-Seq data models it via a negative binomial [24], [48] or Poisson distribution [43]. The Poisson model is parameterized by its mean and it is already known that RNA-Seq data may exhibit more variability than the single Poisson distribution parameter. The negative binomial distribution can mitigate this over-dispersion problem, allowing the variance to exceed the mean; however, when dealing with a relatively small number of samples, it is difficult to accurately estimate the dispersion parameter of the negative binomial model. Therefore, in this dissertation we focus on a hierarchical multivariate Poisson model [12]. Specifically, gene concentration levels are extracted from a log-normal distribution and their subsequent processing by the sequencing instrument is modeled via a Poisson process. The hierarchical model is not as restrictive as the simple Poisson model, and can be considered as a compromise between the Poisson and negative binomial models in the small-sample setting [49]. The simulated NGS data follow a conditionally Poisson distribution and the marginal distribution of the data is a mixture of Poisson and Gaussian distributions.

Although multivariate data offer the potential for finding features for phenotypic discrimination, large-scale and high dimensionality classification problems with small sample sizes can result in overfitting of the data. A variety of feature-selection algorithms for classification have been proposed over the past decades [64], [65]. Feature selection has inherent problems due to its combinatorial nature and sampling procedures. To select a subset of k features out of n potential features and be assured that it provides an optimal classifier with minimum error among all optimal classifiers for subsets of size k, all $\binom{n}{k}$ possible sets must be checked to guarantee that the best one is selected [66]. In other words, nothing but an exhaustive search can assure finding the best feature set. In practice, feature selection must proceed from sample data, which leads to the well-known peaking phenomenon, i.e., the tendency of achieving improved classification performance with an increasing number of features only to a point, beyond which more features lead to degradation of the classification accuracy [15], [67]-[70]. Therefore, employing too many features in a small-sample setting yields poorer classification accuracy, thereby leading to the need for feature selection. This raises a critical question: can one expect a feature-selection algorithm to yield a feature set whose error is close to that of an optimal feature set?

A good feature selector is expected to report a list of feature sets without missing the true target. Thus, ranking of feature sets becomes a key issue for classification. Unfortunately, for small samples, error estimators deployed to perform the ranking of the feature sets suffer from different degrees of imprecision. Moreover, there is little correlation between the errors of the selected feature set and a close-to-optimal feature set [37]. When the number of samples is small, using re-sampling-based classifier error estimators such as cross-validation and bootstrap is risky owing to the substantial variance [36] and lack of regression with the true error [36], [71]-[73], which is exacerbated in the presence of feature selection [74], [75]. Hence, it is important to choose a computationally feasible error estimator that yields rankings that better correspond to rankings produced by the true errors.

Often, when ordering a list of feature sets based on the estimated errors, the smaller estimates tend to be biased optimistically and the larger estimates tend to be biased pessimistically [39]. Thus, reporting a list of feature sets is preferred compared to providing a single good feature set, the idea being that some in the list of top-performing feature sets will be close to optimal [39]. This approach assures that there is at least one feature set on the list whose true classification error is within some given tolerance of the best feature set with high probability. Given the list, one can either focus on the feature sets in the list for further sampling or take a classical wet-lab approach to determine which ones are predictive of the phenotype of interest [39].

In this chapter, we investigate the effects of the nonlinear transformation induced by NGS technology and the choice of error estimators on feature-set ranking. Quantification of changes in feature-set lists due to a measurement technology requires a baseline to compare, i.e., underlying gene-concentration as the biological ground truth. This can be accomplished via simulated data experiments. For this purpose, we utilize a model-based approach and provided a distribution from which the synthetic data arise. We also consider an application of the proposed methodology to real RNA-Seq data as an example of one possible way to derive power curves that estimate the goodness of the feature-set ranking under user-defined settings.

We focus on the LDA classification rule and our work is neither a comparison study of different pattern classifiers nor a model selection study. The rationale for focusing on LDA classifiers is based on our previous studies [35], [76]. The performance of seven different classification rules on real patient data was compared in terms of the expected classification error, for different sample sizes and dimensionality [76]. Classification rules considered were linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), nearest-mean classification (NMC), 1-nearest neighbor (1NN), 3-nearest neighbor

(3NN), CART with a stopping rule that ends splitting when there are six or fewer sample points in a node, and a neural network (NNET) with 4 nodes in the hidden layer. As a result, LDA has proved to be a very robust classification rule, which is effective for a wide range of sample sizes and therefore, we focus on the LDA classification rule.

2.2 Methods

2.2.1 Ranking Power

The *ranking power* is a measure of the goodness of a ranked list of classification feature sets and is defined by [39]

$$\Delta_{D,d}^{n,r}(m) = P(\varepsilon_1 - \varepsilon_0 \ < \ r), \tag{2.1}$$

where ε_1 is the lowest test error for the feature sets in a ranked list of length m sorted by their estimated errors, and ε_0 is the test error of the classifier computed for the Bayes features. Specifically, to compute the ranking power consider all of the possible feature sets of size d among the number D of total features. Then, rank them according to their estimated errors and obtain the top m feature sets, $\mathcal{F}_1, \mathcal{F}_2, \cdots \ , \mathcal{F}_m$. Define ε_1 as the lowest test error of the classifier among the m feature sets considered. The ith lowest estimated error $\hat{\varepsilon}_{(i)}$ corresponds to the feature set $\mathcal{F}_{(i)}$, but the ith lowest test error $\varepsilon_{(i)}$ will likely not correspond to the feature set $\mathcal{F}_{(\mathrm{i})}$.

The ranking power provides the probability that given a ranked list of m feature sets there is at least one feature set in that list with an error that is close to that of the best feature set. The ranking power depends on the list length m, the total number D of features, the number d of selected features, and sample size n. The original ranking power definition takes into account the difference between the smallest test error of the classifier in a given list of feature sets and the respective test error of the Bayes feature set. However, it is often desirable to consider the magnitude of the Bayes error. Thus, we propose the following modification to the ranking power definition:

$$\Delta_{D,d}^{n,c}(m) = P(\ \varepsilon_1 - \varepsilon_0 \ < \ c \cdot \varepsilon_0) \tag{2.2}$$

This modification allows for an explicit comparison of the difference between ε_1 and ε_0 to the magnitude of the test error of the Bayes feature set as represented by the parameter c. For example, c = 0.01 indicates that we are only interested in ranked lists of features sets where the feature set with the smallest test error differs from the test error of the Bayes feature set by less than 1% of the test error of the Bayes feature set. For any given ε_0, there is a clear relationship between the value of r in the original definition of the ranking power and the parameter c in the modified version above. Thus, for the purpose of comparing our simulation results to those from the previous study by Zhao *et al.* [39], we report the values of the parameter r.

Ranking power of the gene-expression concentration generated from the Multivariate Normal (MVN) distribution [77], [78] is computed by the probability of the following inequality

$$\varepsilon_{1,MVN} - \varepsilon_{0,MVN} \ < c \cdot \varepsilon_{0,MVN}, \tag{2.3}$$

where $\varepsilon_{1,MVN}$ is the lowest test error for the feature sets in the MVN ranked list and $\varepsilon_{0,MVN}$ is the test error of the Bayes feature set in the MVN model. In the same way, ranking power of the NGS data is calculated by the probability of

$$\varepsilon_{1,NGS} - \varepsilon_{0,NGS} < c \cdot \varepsilon_{0,NGS.} \qquad (2.4)$$

The same Bayes feature set in the MVN model is used as the Bayes feature set of the NGS model and $\varepsilon_{0,NGS}$ is the respective test error of the Bayes feature set in the NGS data. The smallest test error for the feature sets in the NGS ranked list is $\varepsilon_{1,NGS}$.

2.2.2 Length of Extensions

Gene-expression concentration is the biological ground truth and has often been modeled by the multivariate normal distribution. We use the MVN model to assess the effects of the NGS transformation on the ranking power and the composition of the ranked lists of feature sets. In general, when one desires to compare two ranked lists of feature sets, one is interested how a particular feature set is ranked in each one of the two lists. While there are several possible ways to measure this difference in the ranking we focus on the ranking of a top-performing feature set from one of the two lists in the other list. To achieve the desired comparison we introduce the following notation: $\mathcal{F}_{MVN}$ denotes the feature set ranked at the top in the list of feature sets obtained using the MVN model of gene concentrations; the rank of $\mathcal{F}_{MVN}$ in the respective NGS list is denoted as $\tau_{NGS.}$ Similarly, τ_{MVN} is the rank of the top feature set $\mathcal{F}_{NGS}$ from the NGS list in the respective MVN ranked list of feature sets.

2.2.3 Bayes Features

The Mahalanobis distance provides a way to calculate the Bayes error. If class densities are Gaussian, the Bayes error can be simply calculated using only sample mean vectors μ_i and sample covariance matrices Σ_{i} of class i. The Mahalanobis distance Δ is given by

$$\Delta = \sqrt{(\mu_1 - \mu_2)^{\mathrm{T}} \Sigma^{-1} (\mu_1 - \mu_2)}, \tag{2.5}$$

where Σ denotes the average covariance matrix given by $\Sigma = P(c_1) \cdot \Sigma_1 + P(c_2) \cdot \Sigma_2$ and $P(c_i)$ is *a priori* class probability of class i = 1, 2. Equal prior probabilities for the classes and equal covariance matrices are assumed in our model. Therefore the Bayes error for any feature set $\mathcal{F}$ of size d is $\Phi(-\Delta/2)$, where Φ is the standard normal cumulative distribution function. $\mathcal{F}_{bayes}$ denotes the feature set having the largest Mahalanobis distance and, accordingly, the minimum Bayes error.

Bayes features of a hierarchical model cannot be easily found as in the Gaussian case. Simulated NGS data are the mixed form of Poisson and Gaussian distributions, so there is no analytical formula for the Bayes error. The Bayes error of the hierarchical model can be estimated using Monte-Carlo sampling. In this dissertation, Bayes features of the MVN are used as the Bayes features of the NGS data in the biological context. Although Bayes features of the MVN are not equal to those of the transformed data, MVN Bayes features reflect the biological ground-truth markers.

2.3 The Models for Gene Concentrations and NGS data

Two different types of synthetic data are generated for simulation experiments: (i) actual gene-expression concentration, called MVN and (ii) Poisson-transformed MVN data, denoted as NGS, which emulate NGS-reads.

2.3.1 Multivariate Gaussian Model

Gene concentration levels can be modeled using a log-normal distribution [51], [79], [80] and the hybrid multivariate Gaussian model proposed in Zhao *et al.* [39] is adopted in this dissertation. Genes/features are categorized into two groups: markers and non-markers. There is a total of $D = \upsilon + \eta$ features and υ and η represent the number of markers and non-markers in the model respectively. Markers resemble genes associated with diseases and they have two class-conditional Gaussian distributions with equally likely classes and common covariance matrix Σ. The mean vectors for the markers are $\mu_0 = m_0 \times (0,0,\cdots,0)^{\mathrm{T}}$ and $\mu_1 = m_1 \times (a_1, a_2, \cdots, a_\upsilon)^{\mathrm{T}}$ for class 0 and class 1, respectively, where m_0 and m_1 are scalars and υ denotes the total number of marker features generated. In order to mimic real experimental situations, where every marker performs well but not exactly the same, all elements of vector μ_1 are not equal to one another. μ_1 is an equally spaced vector with a_1 = 1 and a_υ = 0.8. The covariance matrix Σ is blocked and each block Σ_ρ has variance ${\sigma_\mu}^2$ along the diagonal and correlation coefficient ρ off the diagonal:

$$\Sigma = \begin{bmatrix} \Sigma_\rho & 0 & \cdots & 0 & 0 \\ 0 & \Sigma_\rho & & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \cdots & \Sigma_\rho & 0 \\ 0 & 0 & & 0 & \Sigma_\rho \end{bmatrix} \tag{2.6}$$

where

$$\Sigma_\rho = {\sigma_\mu}^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & & \rho \\ \vdots & & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}. \tag{2.7}$$

Different blocks correspond to different gene regulatory pathways [77], [78] and model the assumption that groups of genes in the same pathway are biologically or functionally correlated and interacting with each other, while genes in different pathways are uncorrelated. Non-markers are uncorrelated and modeled as one-dimensional zero-mean random Gaussian noise, with a total of η features.

2.3.2 The Hierarchical Multivariate Poisson Model

The gene-expression levels in NGS data are measured by the number of reads that are mapped to the corresponding gene in the reference genome. Thus, NGS-type data values are discrete with non-negative integers. Several statistical models for NGS data based on the negative binomial model or Poisson distribution have been proposed [24], [43], [48]. In this dissertation, the hierarchical multivariate Poisson model [12] is adopted. It assumes that the sequencing facility samples mRNA concentration through a Poisson process, and the expected number of reads is the mean of the Poisson distribution. Read count for a sample point i and the jth gene is $X_{i,j}$. It is obtained by the generalized linear model [81] for a given s_i:

$$p\left(X_{i,j} \middle| s_i\right) \sim Poisson\left(s_i \exp\left(\lambda_{i,j} + \theta_{i,j}\right)\right), \tag{2.8}$$

where s_i denotes the sequencing depth for the i-th sample point in the model and is randomly generated from a uniform distribution, *U(α, β)*, where *α > 0* and *β > α*. To generate count data for RNA-Seq reads, the hybrid Gaussian model is fed to the pipeline as $\lambda_{i,j}$, the *j*-th gene expression level in a sample point *i*. The value is perturbed by $\theta_{i,j}$, which reflects technical effects associated with the experiment and is drawn from a

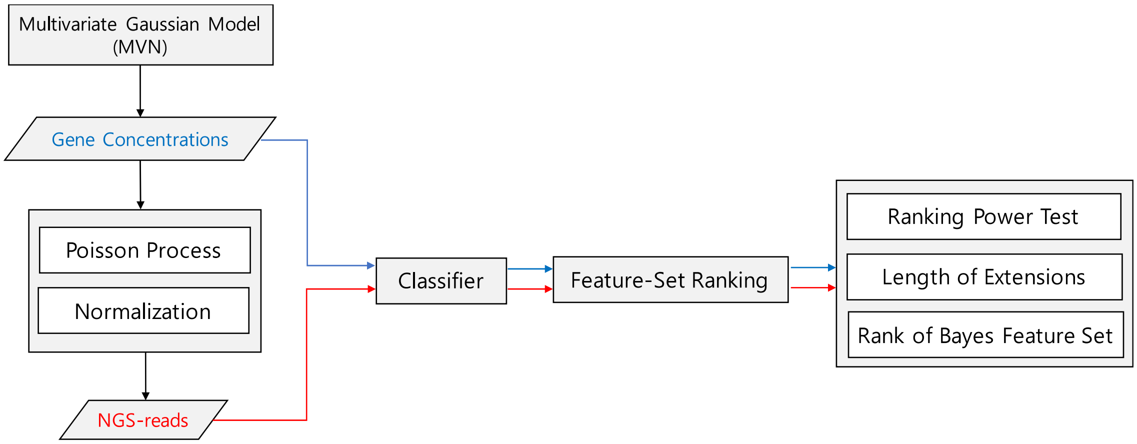


Figure 2.1 An overview of the simulation. Two different types of synthetic data are generated: (1) MVN; (2) NGS. Datasets are generated from a multivariate Gaussian model and a hierarchical multivariate Poisson model. Subsequently, the datasets are fed to the same test modules: classification, error estimation, and feature-set ranking.

Gaussian distribution,

$$\theta_{i,j} \sim N(0, |m_1 - m_0|COV), \tag{2.9}$$

where COV is the coefficient of variation. Once the NGS data are generated, the features are normalized in a way that each feature is zero mean and unit standard deviation across all the sample points.

2.4 Implementation

2.4.1 Simulation Procedure

Figure 2.1 presents a general overview of the simulation employed herein. General implementation follows a similar simulation procedure proposed in Zhao *et al* [39].

(1) Set up a hybrid Gaussian model with υ marker features and η non-markers to yield $D = \upsilon + \eta$ features. Find the Bayes feature set $\mathcal{F}_{bayes}$ of size d.

(2) Generate a large test set of independent data using the MVN model.

(3) For every feature set of size d, design an LDA classifier and compute its estimated and test errors. Compute the test error $\varepsilon_{0,MVN}$ for $\mathcal{F}_{bayes.}$

(4) Rank all feature sets by their estimated errors based on the training data and select the top m of them to form the MVN ranked list.

(5) Let $\varepsilon_{1,MVN}$ be the lowest test error in the top m list.

If $\varepsilon_{1,MVN} - \varepsilon_{0,MVN} < c \cdot \varepsilon_{0,MVN}$, set $count_{MVN} := count_{MVN} + 1$.

(6) The MVN data generated from (1) and (2) are fed to the Poisson transformation pipeline to obtain the NGS data.

(7) Repeat steps (3) through (5) for the NGS data. Use the same Bayes feature set $\mathcal{F}_{bayes}$ to compute the test error $\varepsilon_{0,NGS}$. If $\varepsilon_{1,NGS} - \varepsilon_{0,NGS} < c \cdot \varepsilon_{0,NGS}$, set $count_{NGS} := count_{NGS} + 1$.

(8) Repeat steps (1) through (7) N times to get ${\Delta_{D,d}^{n,c}}_{MVN}(m) = count_{MVN}/N$ and ${\Delta_{D,d}^{n,c}}_{NGS}(m) = count_{NGS}/N$.

(9) Compare MVN and NGS lists and obtain τ_{MVN} and $\tau_{NGS.}$

(10) Find the ranks of Bayes feature sets in the MVN and NGS lists. Denote them as B_{MVN} and B_{NGS}, respectively.

2.4.2 Simulation Parameters

RNA-Seq technology can provide different numbers of reads per sample, depending on many factors, such as quality of the sample, the desired coverage, sample multiplexing etc. In order to deal with this issue, a previous study [12] examined a variety

Table 2.1 Model parameters for generating synthetic data.

Exp No.	D	v	n	${\sigma_\mu}^2$	ρ	B	d
1	{50,100,150}	{5,10,20}	40	1	0.8	5	{2,3}
2	150	10	{40,80,120}	1	0.8	5	2
3	150	10	40	{0.5,1,2}	0.8	5	2
4	150	10	40	1	{0.1,0.5,0.8}	5	2
5	150	10	40	1	0.8	{2,5,10}	2
6	100 200 300	5 10 15	40	1	0.8	5	2

of ranges of the sequencing depth and NGS-read counts for real RNA-Seq experiments and the parameters α and β are chosen accordingly. Therefore, our selections for the model parameters reflect how real data behaves because they take into account a range of NGS-read counts one can expect from real data. Our study is model-based and we do not focus on the problems of inference or parameter estimation from data. Thus, we adopt the parameters' ranges/values from the work by Ghaffari *et al* [12]. Parameters for the sequencing depth $s_i \sim U(\alpha, \beta)$ are set to $\alpha = 9,\ \beta = 11$, and $COV = 0.05$; $m_0 = 0$, $m_1 = 1$ are used for the distribution of technical effects, $\theta_{i,j}$. Simulation setups and the list of parameters used for the multivariate Gaussian model are provided in Table 2.1. Experiment numbers in Table 2.1 correspond to the parameter setting of each experiment in Table 2.2, Table A.1 and Table A.2 in Appendix A. Absolute bound r is used for comparisons between our results and those in Zhao *et al* [39]. Corresponding values

Table 2.2. Mean of ε_0 in the MVN and NGS list and relative differences between ε_0 and ε_1 with respect to ε_0.

Exp No.	Parameters		$\mathrm{E}[\varepsilon_{0,MVN}]$	c_{MVN} (r = 0.03)	$\mathrm{E}[\varepsilon_{0,NGS}]$	c_{NGS} (r = 0.03)
1 (d = 2)	υ = 5	D = 50	0.2557	0.1173	0.2993	0.1002
		D = 100	0.2559	0.1172	0.2990	0.1003
		D = 150	0.2558	0.1173	0.2994	0.1002
	υ = 10	D = 50	0.2558	0.1173	0.2990	0.1003
		D = 100	0.2557	0.1173	0.2988	0.1004
		D = 150	0.2558	0.1173	0.2992	0.1003
	υ = 20	D = 50	0.2559	0.1172	0.2992	0.1003
		D = 100	0.2556	0.1174	0.2988	0.1004
		D = 150	0.2559	0.1173	0.2992	0.1003
1 (d = 3)	υ = 5	D = 50	0.2557	0.1173	0.2993	0.1002
		D = 100	0.2559	0.1172	0.2990	0.1003
		D = 150	0.2558	0.1173	0.2994	0.1002
	υ = 10	D = 50	0.2558	0.1173	0.2990	0.1003
		D = 100	0.2557	0.1173	0.2988	0.1004
		D = 150	0.2558	0.1173	0.2992	0.1003
	υ = 20	D = 50	0.2559	0.1172	0.2992	0.1003
		D = 100	0.2556	0.1174	0.2988	0.1004
		D = 150	0.2559	0.1173	0.2992	0.1003
2	n, bresub	40	0.2558	0.1173	0.2992	0.1003
		80	0.2498	0.1201	0.2956	0.1015
		120	0.2479	0.1210	0.2947	0.1018
	n, loo	40	0.2558	0.1173	0.2994	0.1002
		80	0.2497	0.1201	0.2958	0.1014
		120	0.2479	0.1210	0.2946	0.1018
3	$\sigma_\mu{}^2$	0.5	0.1734	0.1731	0.2190	0.1370
		1	0.2558	0.1173	0.2992	0.1003

Table 2.2. Continued.

Exp No.	Parameters		$E[\varepsilon_{0,MVN}]$	c_{MVN} (r = 0.03)	$E[\varepsilon_{0,NGS}]$	c_{NGS} (r = 0.03)
3	$\sigma_{\mu}{}^{2}$	2	0.3266	0.0919	0.3737	0.0803
4	ρ	0.1	0.2557	0.1173	0.2995	0.1002
		0.5	0.2559	0.1172	0.2991	0.1003
		0.8	0.2558	0.1173	0.2992	0.1003
5	B	2	0.2626	0.1142	0.3042	0.0986
		5	0.2558	0.1173	0.2992	0.1003
		10	0.2535	0.1184	0.2972	0.1009
6	D=100, υ=5		0.2559	0.1172	0.2990	0.1003
	D=200, υ=10		0.2557	0.1173	0.2991	0.1003
	D=300, υ=15		0.2558	0.1173	0.2996	0.1001

for the relative significance of the difference c are provided in Table 2.2. Because there is no closed form to calculate the true errors of designed classifiers, large independent test sets are generated. When using independent test data, the Root Mean Square (RMS) between the true and estimated error is bounded above by $\frac{1}{2\sqrt{n_{test}}}$ [12]. Test sample of size n_{test} = 10,000 are generated and samples are divided equally between the two classes.

2.5 Results

2.5.1 Synthetic data

Effects of $D, \upsilon, n, \sigma_{\mu}{}^{2}, \rho, B, d$ and proportion of υ to D are studied.

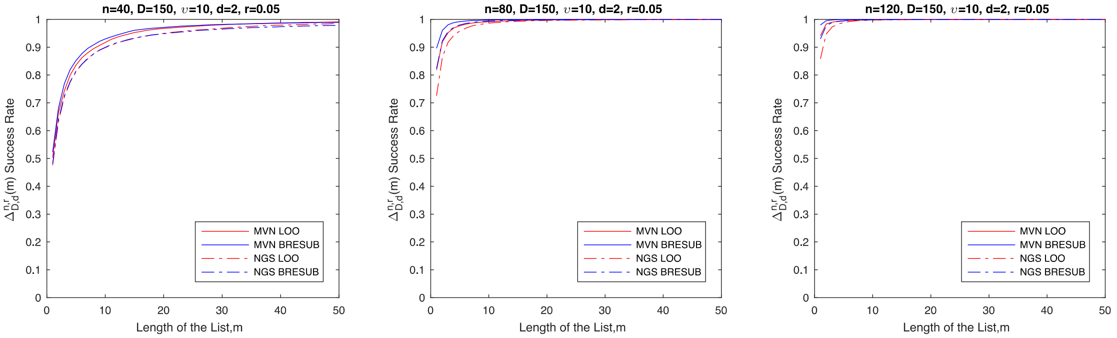


Figure 2.2. Power curves for different error estimators and sample size n. solid: MVN, dashed: NGS, red: leave-one-out (LOO), blue: bolstered resubstitution (BRESUB).

2.5.1.1 Effects of Error Estimators

Previous literature shows that cross-validation methods, especially leave-one-out estimators, display large variance [82], [83]. This variance results in a widely dispersed deviation between the true and estimated errors of a classifier, thereby making cross-validation unreliable for ranking feature sets in the small-sample setting. It has been shown that bolstering and resubstitution-based feature ranking outperform leave-one-out cross-validation-based feature ranking for discovering top-performing feature sets for classification when using small samples [35]. Previous studies [2], [35] are based on a Gaussian mixture model and microarray-based patient data. In this dissertation, we examine the effects of error estimators on the ranking of feature sets of RNA-Seq data. Two different error estimators, bolstered resubstitution (BRESUB) and leave-one-out (LOO), are used to sort the lists. Figure 2.2 indicates that the hit rate of finding a good feature set in the list sorted by bolstered resubstitution error estimators is higher than the

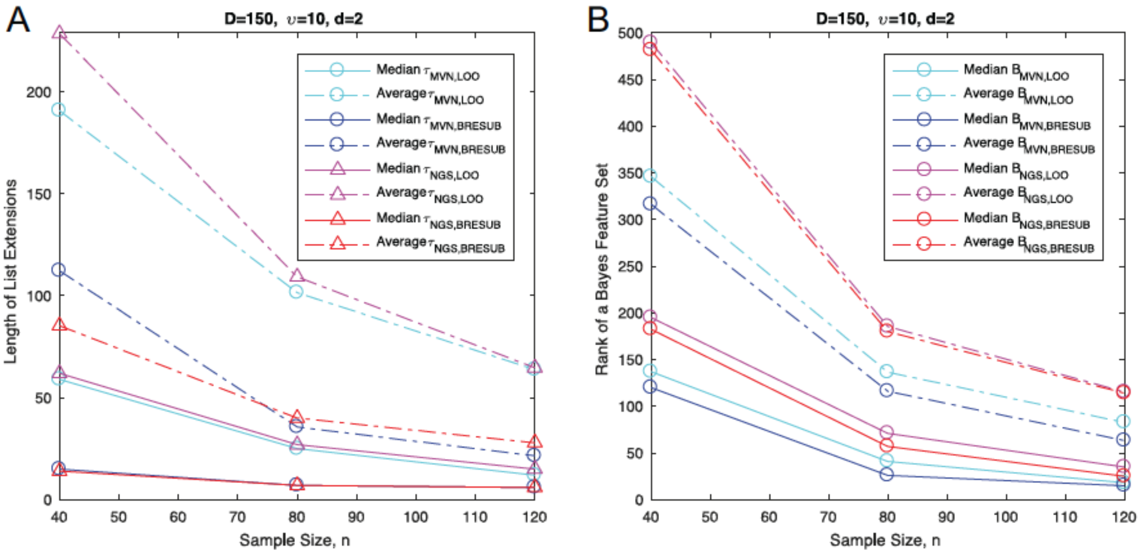


Figure 2.3. Effects of different error estimators and sample size n on (A) length of list extensions and (B) rank of a Bayes feature set. solid: median, dashed: average, cyan: MVN, LOO, blue: MVN, BRESUB, pink: NGS, LOO, red: NGS, BRESUB.

success rate of the leave-one-out-based list. Figure 2.3(A) shows that both $\tau_{MVN,LOO}$ and $\tau_{NGS,LOO}$ are larger than $\tau_{MVN,BRESUB}$ and $\tau_{NGS,BRESUB}$, respectively, which implies that leave-one-out mixes up the orders more harshly than bolstered resubstitution. Moreover, Figure 2.3(B) shows that the ranks of Bayes feature sets in the leave-one-out-based list are larger than that of the bolstered resubstituion-based list. All of these results suggest that leave-one-out estimators perform poorly with RNA-Seq data, producing less accurate ranking orders compared to the list sorted by BRESUB error estimators.

2.5.1.2 Effects of the Sample Size, n

A larger sample size generally leads to better performance of classification and ranking feature sets. The results of NGS shown in Figure 2.2 and Figure 2.3 are in accord with this expectation. As sample size increases, ranking power curves for NGS are also

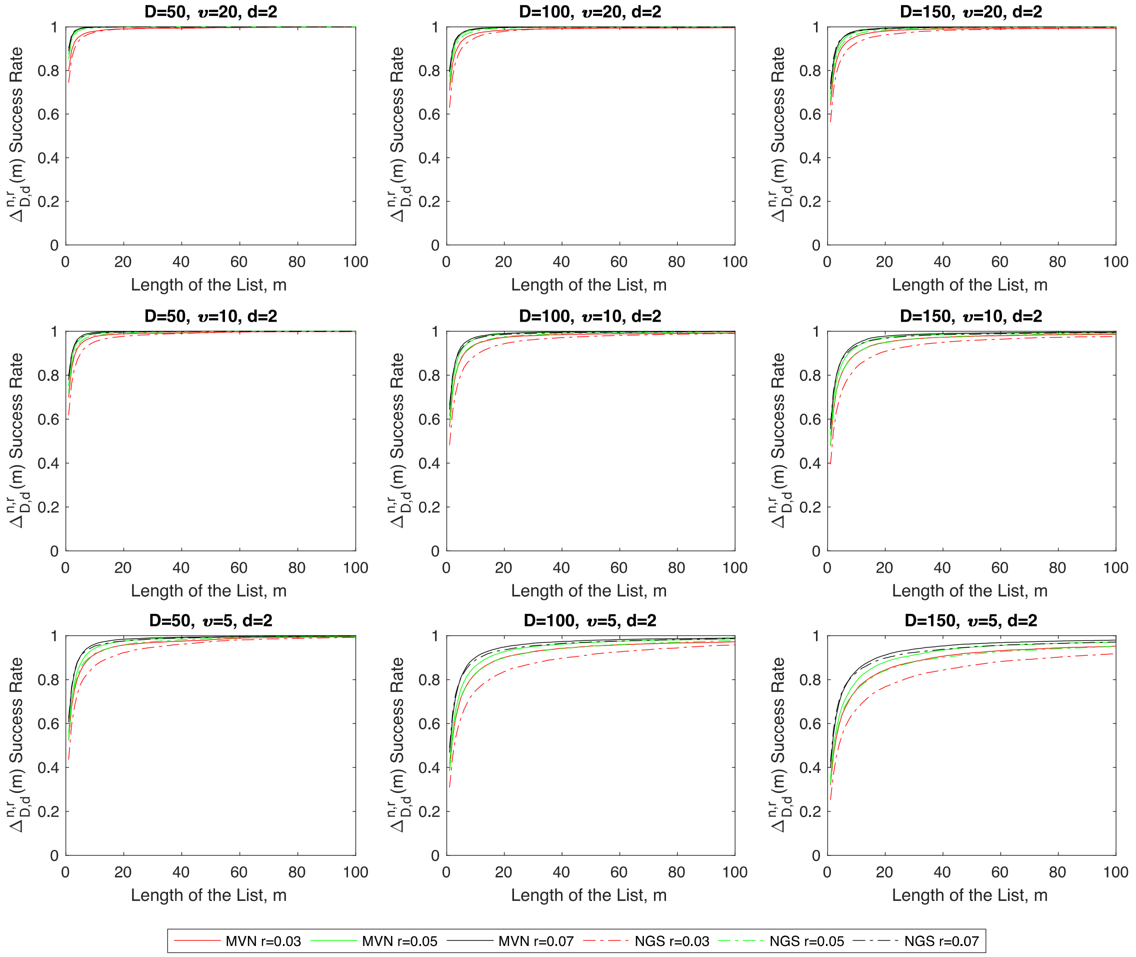


Figure 2.4 Power curves for different D and υ when d=2.

elevated. For both types of data, monotonic decrease of extension length and Bayes rank in median are observed as sample size gets larger.

2.5.1.3 Effects of the Total Number of Features D and the Number of Marker Features υ

Figure 2.4 represents the effects of the total number D of features and the number υ of marker features on the ranking power curves when the final number d of selected features is 2. The ranking power curves for d = 3 are provided in Figure 2.5. Zhao et al.

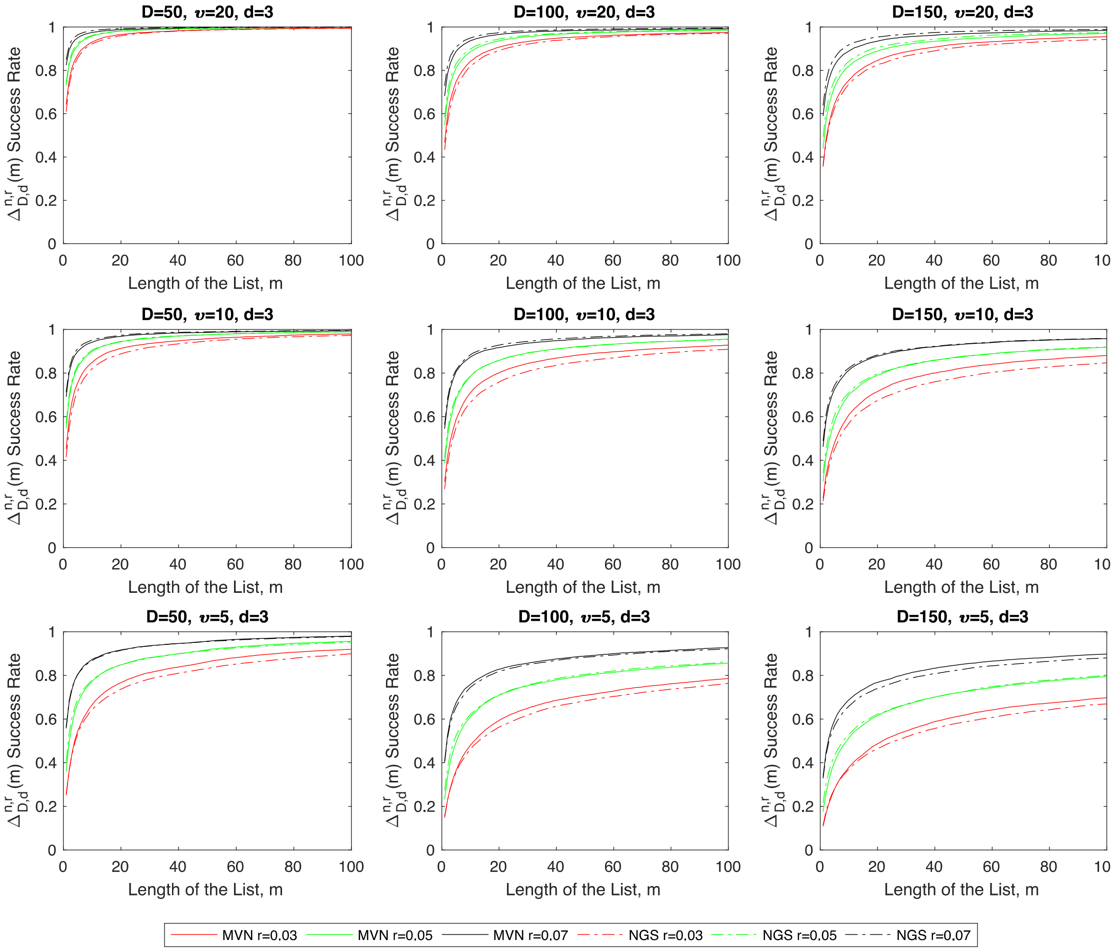


Figure 2.5 Power curves for different D and υ when d=3.

[39] have shown that the power curves are lowered in the MVN model as the total number of features increases. Figure 2.4 and 2.5 show analogous results in the RNA-Seq model. The plots also indicate that for a fixed value of D, the power increases as υ increases. This is not surprising because the prior information provided by the biologist becomes richer, containing more markers. Figure 2.6 illustrates the effects of increasing D and υ on τ_{MVN} and τ_{NGS}. As D gets larger, a monotonic increase in median and average extension length is observed in both models. In Figure 2.6(B) and 2.6(D), no obvious trend can be discerned

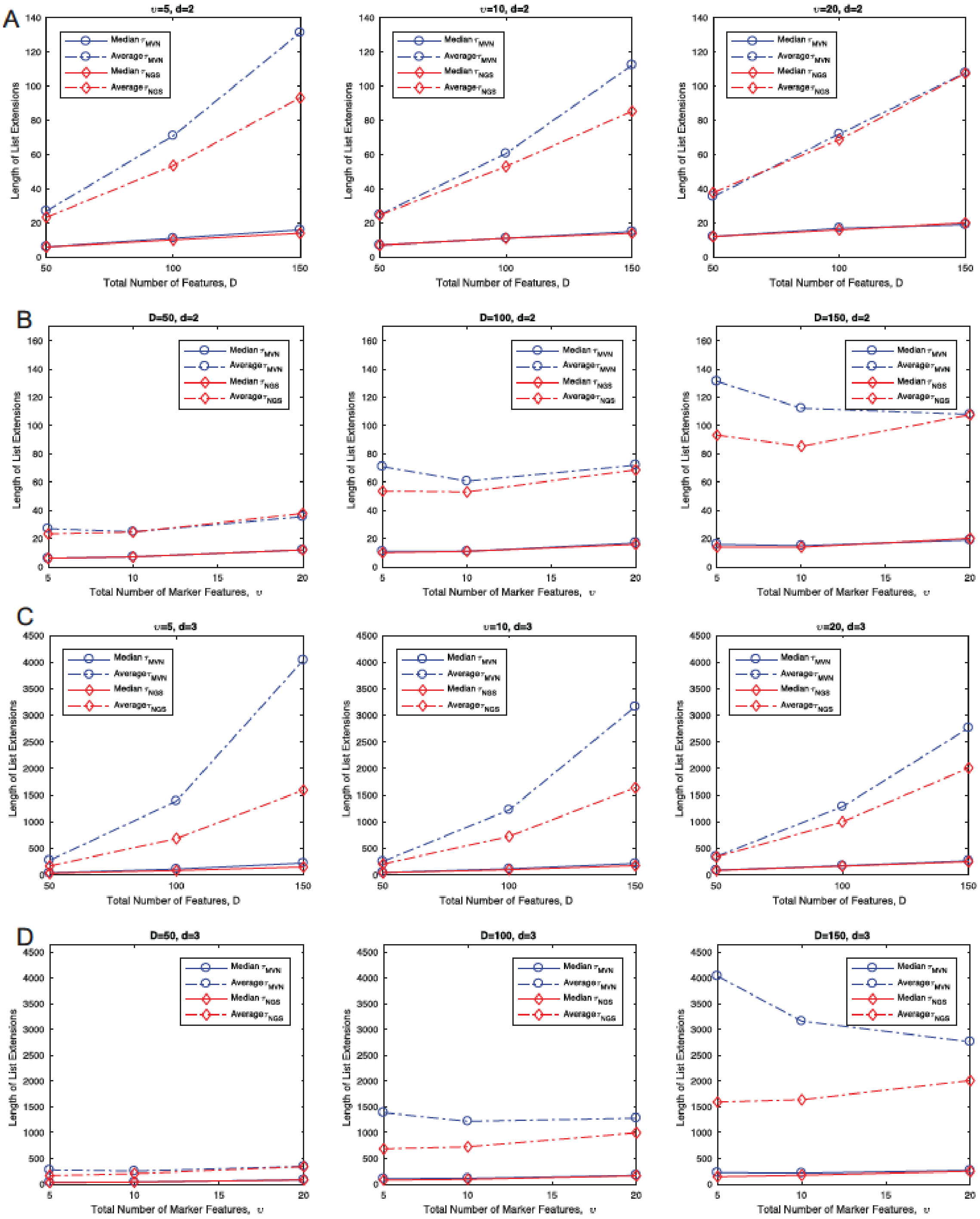


Figure 2.6 Effects of different D and υ on length of list extensions for $d = 2$ are presented in (A) and (B). Graphs for $d = 3$ are presented in (C) and (D). solid: median, dashed: average, blue: MVN, red: NGS.

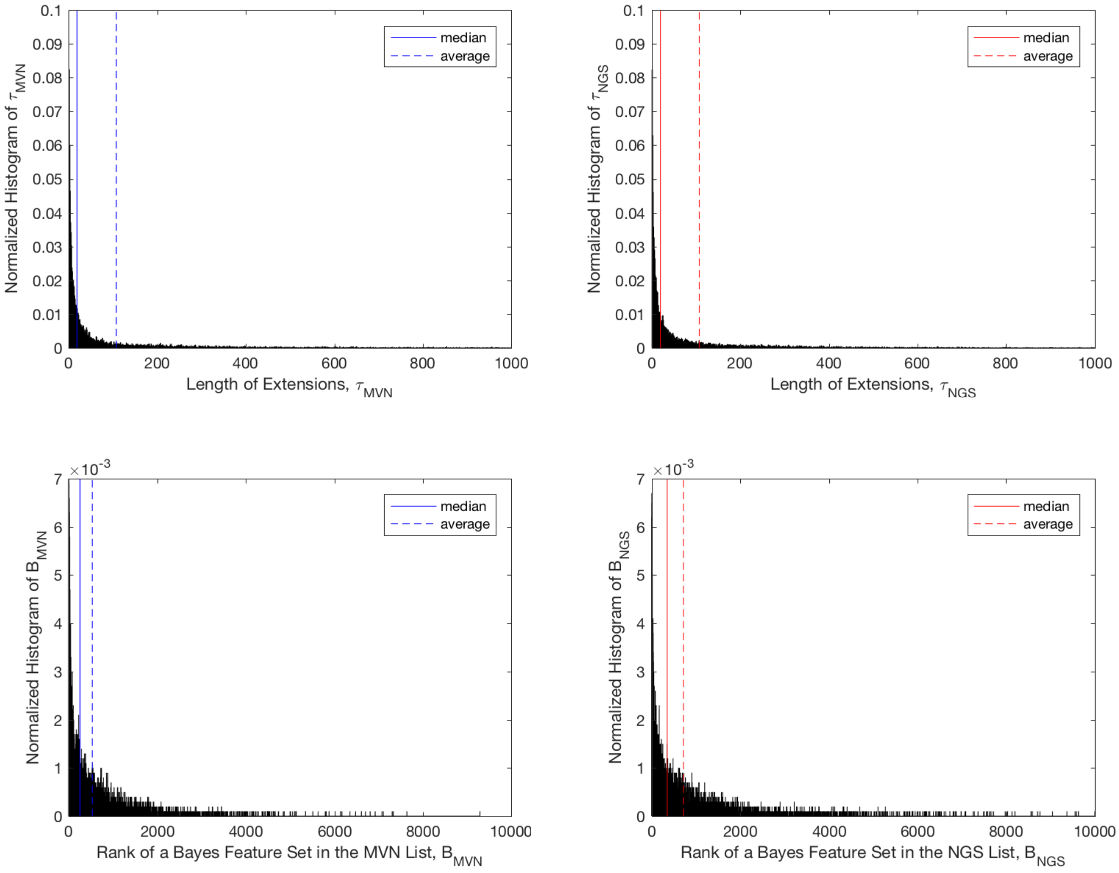


Figure 2.7 Histogram of length of list extensions and rank of a Bayes feature set. solid: median, dashed: average, blue: MVN, red: NGS.

in terms of the mean, nor is there any consistency. However, the median extension length exhibits a slight increasing trend.

Histograms of length of extensions and the rank of the Bayes feature set are illustrated in Figure 2.7. It is a skewed heavy-tailed distribution with the mean farther out in the long tail than the median. Because the mean is highly vulnerable to outliers, it should be interpreted with caution when extreme values are present. Focusing on the median values, which are less affected by outliers, an increasing trend of median extension length is exhibited as υ gets larger. This is because it becomes more competitive to rank at the top

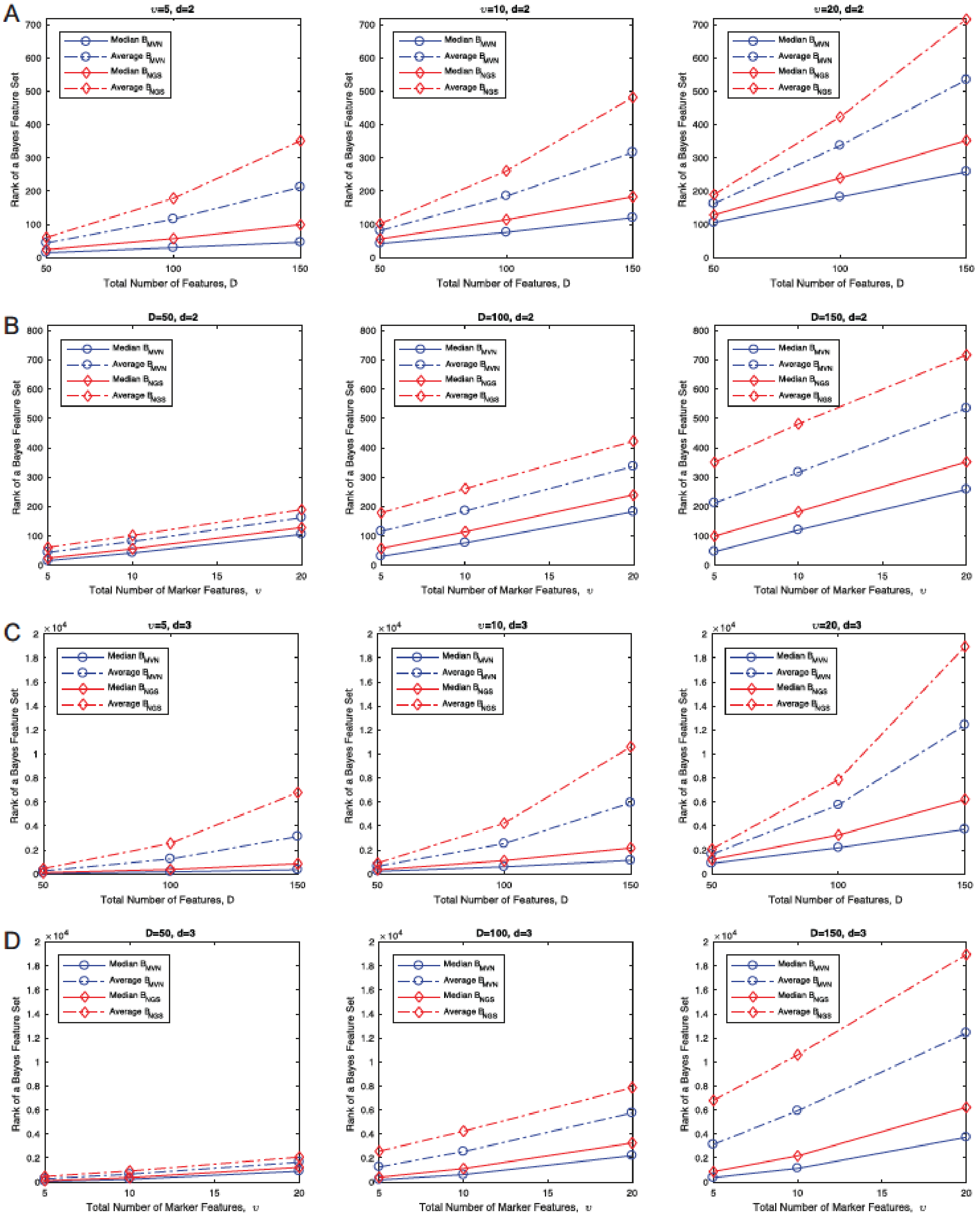


Figure 2.8 Effects of different D and υ on rank of a Bayes feature set for $d = 2$ are shown in (A) and (B). Graphs for $d = 3$ are presented in (C) and (D). solid: median, dashed: average, blue: MVN, red: NGS.

as more markers enter into the data and the one which occupies the top becomes more variable, thereby resulting in the increase of extension length to match two lists. The monotonic increase of median rank of Bayes feature pair is presented in Figure 2.8(B) and 8(D), as υ increases. As more marker features are included, there are more feature pairs which perform as well as a Bayes feature pair. Therefore, the Bayes feature set is no longer a unique and distinguishing feature pair, and the multitude of marker features obscures the Bayes feature pairs.

2.5.1.4 Effects of the Variance $\sigma_{\mu}{}^{2}$ in the Marker Model

Figure 2.9(A) shows the effect of the variance in the marker model. Higher variance results in larger overlaps of the two distributions, which leads to degradation of classification performance and increasing difficulty of finding top-performing feature sets. Therefore, the success rates of both models decrease as variance increases. When $\sigma_{\mu}{}^{2}$ = 2.0, the power curve of the NGS model is higher than that of MVN. This does not necessarily mean that it is better to use the RNA-Seq model to detect a good feature set when the problem is difficult. A better interpretation is that mixing is so extensive that even the underlying gene concentrations are useless for finding a good feature set. Figure 2.9(A) also shows that both extension length and rank of Bayes feature sets increase as variance increases.

2.5.1.5 Effects of the Correlation ρ in the Covariance Matrix

Zhao *et al.* [39] have shown that a higher correlation makes it slightly harder to

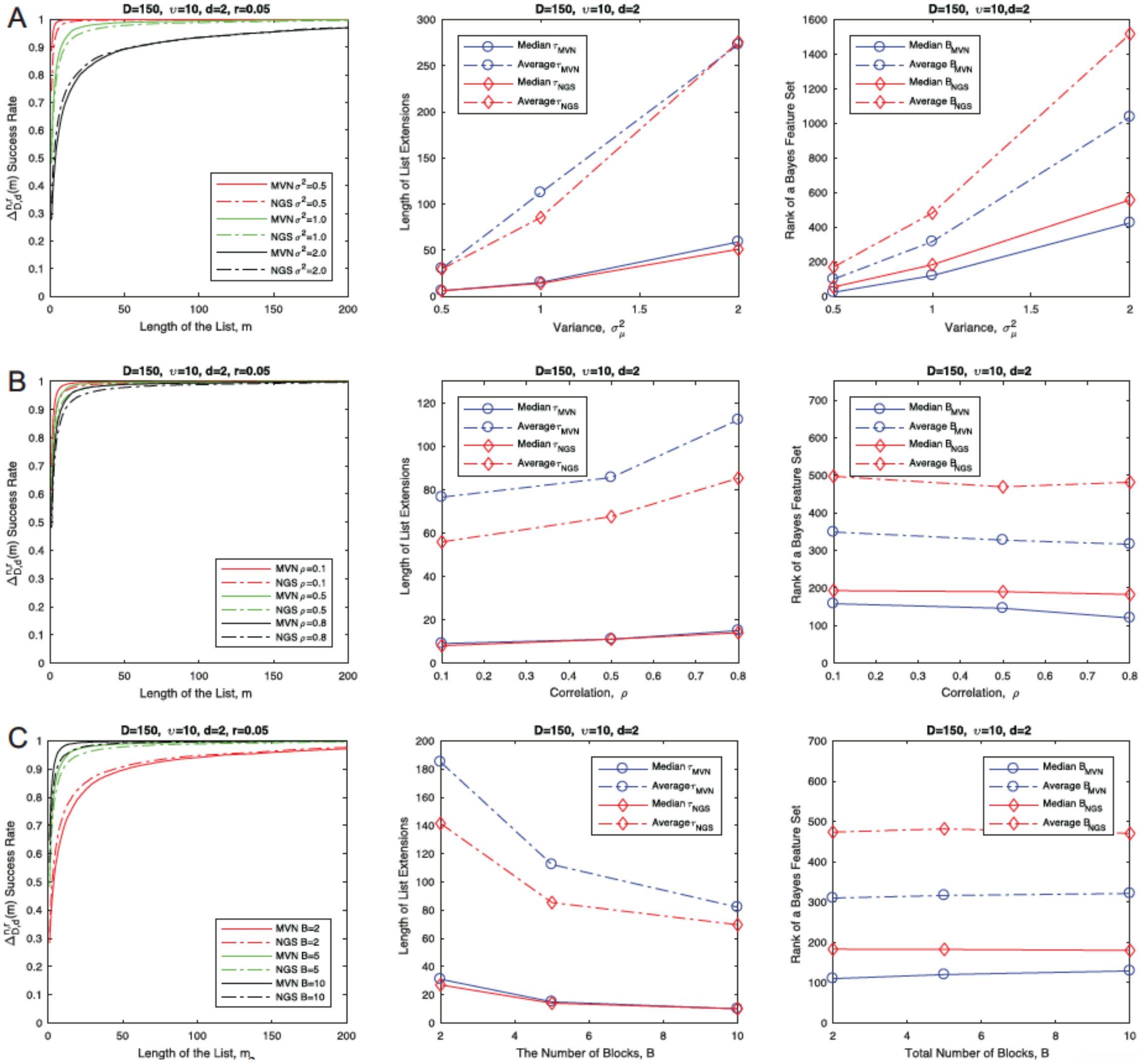


Figure 2.9 Effects of (A) variance, σ_μ^2 (B) correlation, ρ and (C) the number of blocks, B on the ranking power, length of list extensions, and rank of a Bayes feature set.

find good features in the MVN model. Figure 2.9(B) indicates that the same applies to the RNA-Seq model. As ρ increases, ranking power of both MVN and RNA-Seq models decreases. Curves for median extension length and the rank of Bayes feature sets are almost flat with respect to the correlation.

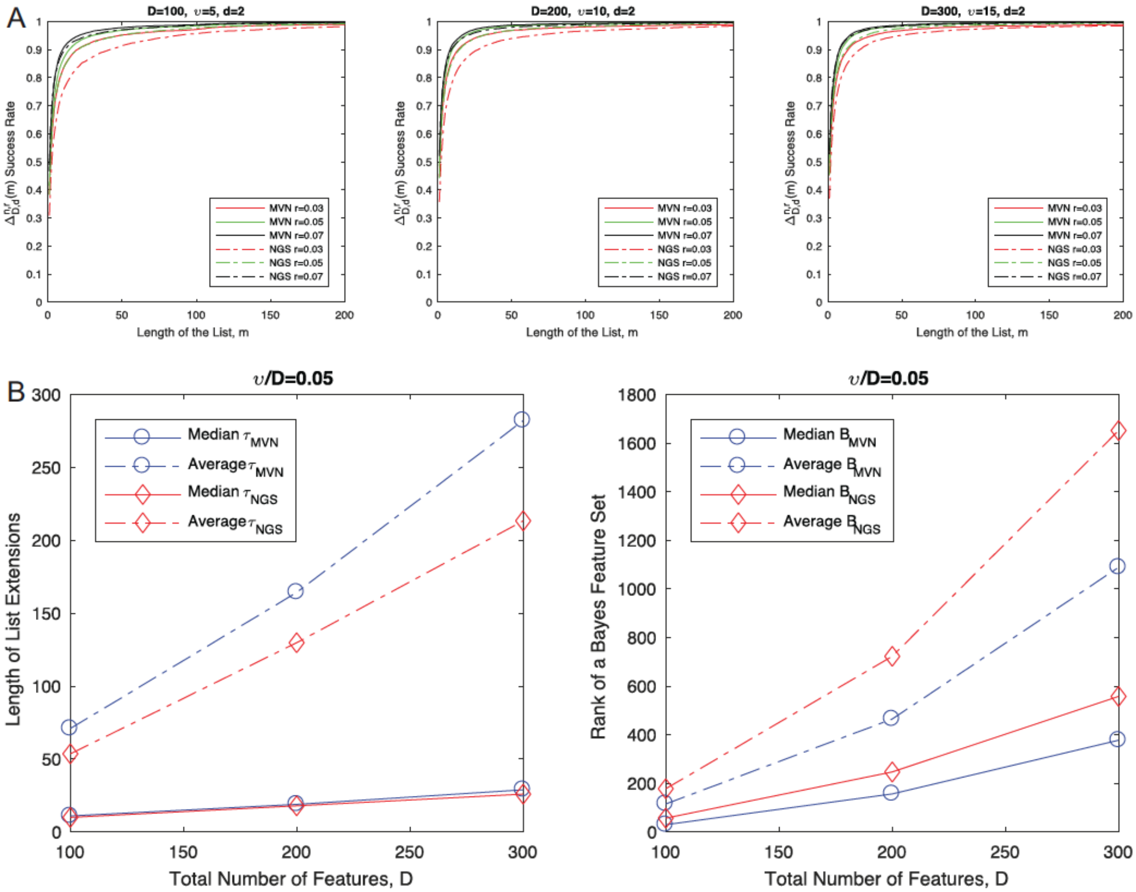


Figure 2.10 Effects of increasing D at the same rate υ increases on (A) the ranking power (B) length of list extensions and (C) rank of a Bayes feature set. Ratio of υ to D remains the same as 0.05.

2.5.1.6 Effects of the Number of Blocks B in the Covariance Matrix

Different blocks represent different metabolic/biologic pathways and as the number of blocks increases, genes may become spread among more pathways and may increase the power to find good features. Zhao *et al*. [39] showed that it is easier to find good features with more blocks. Figure 2.9(C) demonstrates that the ranking power becomes higher as B increases in the RNA-Seq model. When there are only two blocks, RNA-Seq exhibits a higher success rate compared to the MVN model, but it is very unlikely to have only two pathways in real data. Figure 2.9(C) shows decreasing extension length with larger B, which is consistent with the power curve. No specific trend is observed in the rank of the Bayes feature set with respect to B.

2.5.1.7 Effects of Increasing D at the Same Rate υ Increases

To examine the effects of increasing D at the same rate υ increases, the proportion of marker features in the data were fixed at 0.05 with the total number of features ranging from 100 to 300. Figure 2.10(A) shows that when the proportion υ/D is kept constant, the power curves are relatively unchanged as the number of total features D increases. However, Figure 2.10(B) shows that the extension length and the rank of the Bayes feature sets increase under the same conditions, pointing to the increased difficulty of the problem as the number of total features increases.

2.5.2 An example of feature-set ranking for a real data set

We consider a real RNA-Seq dataset from a randomized, double-blind crossover intervention of flaxseed lignan extract and placebo [84]. Colonic mucosal biopsies from healthy participants are used to characterize the site-specific global gene-expression signatures associated with stromal versus epithelial tissue. The data provide insight into the gene expression landscape of the normal epithelium and stroma prior to the onset of intestinal tumorigenesis. This is noteworthy because the development of cancer is intimately linked to cross talk between cancer cells and the surrounding stromal cells [84]. The dataset consists of 29 epithelium and 30 stroma biopsies from the sigmoid colon.

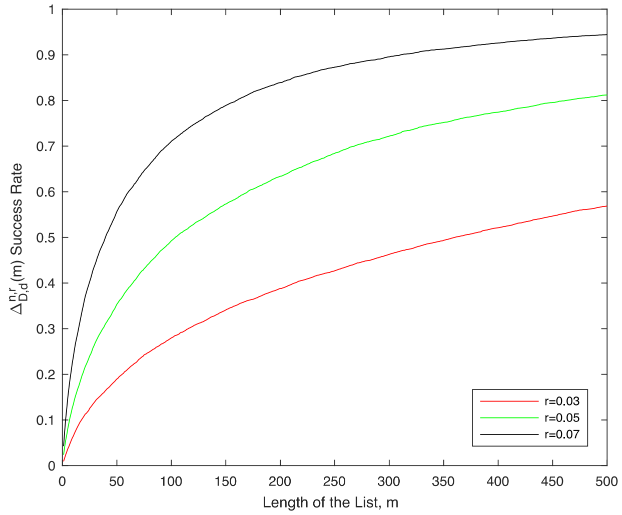


Figure 2.11 Power curves for a real dataset where n=59, D=960, d=2. red: r=0.03, green: r=0.05, black: r=0.07.

Epithelium samples belong to class 0 and stroma samples are labeled as class 1. In total, 960 intestinal genes were selected using prior biological knowledge [85]. Out of 960 genes, 259 stromal genes and 9 epithelial genes were included which were shown to be highly expressed in stroma and epithelium, respectively [86], [87]. Repeated random subsampling holdout [88] method was employed on the dataset. Twenty samples were randomly selected and used for training and the remaining data samples were assigned to the test set. This process was repeated 10,000 times with different subsamples to improve the reliability of the holdout estimate [88]. The proportion of samples from each class was kept the same in both the training and test sets. For every feature set of size two, we designed an LDA classifier and computed its estimated and test errors. Bolstered resubstitution error estimators were used to sort the feature sets. As there is no analytical way to obtain a set of Bayes features for real data, we determined ε_0 empirically. Random subsampling was repeated 10,000 times and the mean of the lowest test errors was taken as ε_0 ($\varepsilon_0 = 0.1892$).

Figure 2.11 shows the ranking power for this dataset. The parameter $r = 0.03$ ind icates that we are interested in ranked lists of feature sets where the feature set with the smallest test error differs from ε_0 less than 15.9% of the ε_0. Typically, when a smaller r is employed, a short list may miss interesting gene sets worthy of consideration. Therefore, the list should be further extended to increase the probability of the existence of candidate genes that provide a good approximation of the Bayes features. It is also important to note that a longer list does not always increase the number of candidate genes. As shown in Zhao *et al.* [39], some genes repeatedly appear in the list combined with other genes. Ranking power of the real data for large m is provided in Appendix A.

2.6 Conclusion

This section examines the ranking performance of feature sets derived from a model of RNA-Seq data and compares it to that of a multivariate normal model of gene concentrations. The results demonstrate that the general trends of the parameter effects on the ranking power of underlying gene concentrations are preserved in the RNA-Seq data; however, the power of finding a good feature set becomes weaker and the data become less discriminative when gene concentrations are transformed by the sequencing machine. Moreover, the consistency between the ranked lists of feature sets based on the MVN and the NGS data is poor, which indicates unreliable classification performance in the case of RNA-Seq data.

3. QUANTIFYING THE NOTIONS OF CANALIZING AND MASTER GENES IN A GENE REGULATORY NETWORK – A BOOLEAN NETWORK MODELING PERSPECTIVE[1]

3.1 Introduction

The concept of genes that can constrain, or canalize, a biological system to a specific behavior was first proposed by C. Waddington in 1942 [62]. Waddington proposed the existence of genes that can produce reliable developmental effects against genetic mutations or environmental changes during evolution [62], [89]. Lehner investigated Waddington's intuition and stated that canalizing genes are hub genes that present similar robustness when faced with environmental, stochastic and genetic perturbations [61]. The term *canalizing gene* has been used by Martins *et al.* [60] to refer to genes that possess broad regulatory power, and their action sweeps across a wide swath of processes for which the full set of affected genes are not highly correlated under normal conditions. Zhao *et al.* [56] made a clear distinction between master genes and canalizing genes. Both master and canalizing genes exert a strong control over many downstream gene pathways; however, canalizing genes have an additional ability of taking over the control and overriding other regulatory instructions. In this work, canalizing genes refer

[1] Part of this section is reprinted with permission from "Quantifying the notions of canalizing and master genes in a gene regulatory network—a Boolean network modeling perspective" by E. Kim, I. Ivanov, and E. R. Dougherty, *Bioinformatics*, vol. 35, no. 4, pp. 643–649, Jul. 2018. Copyright 2018 Oxford University Press.

to genes that are not highly active under normal conditions, but are capable of taking over the control of many pathways and exerting broad regulatory power upon such activation. Canalizing genes produce adaptive and optimal reactions to environmental, stochastic and genetic perturbations and they are essential in a complex system so it can achieve biological robustness and buffer itself from the effects of random alterations or operating errors. We also suggest that the currently adopted definitions of canalizing and master genes could be modified so that a particular gene does not have to be exclusively a master or a canalizing gene. It is important to emphasize that the notions of canalizing and master gene are relative. Any gene possesses some degree of canalizing power over its subnetwork. The notion of canalizing gene can only be defined relative to other genes and the notation c, m_i and s_i used in this Section for a canalizing, master and slave gene, respectively, is used with this understanding.

The principle of a canalizing gene is similar to the concept of an interrupt in computer architecture. In systems programming, an interrupt is a mechanism by which the hardware or software alerts the processor to a high-priority condition indicating an event that needs immediate attention and requests the processor to stop the normal processing or current code it is executing and perform a specific action [90]. The processor responds by suspending its current activities and jumping to a separate piece of code to deal with the event. Similar to an interrupt handler or an interrupt service routine (ISR) that is invoked by a special instruction or by an exceptional condition and puts the program into a different execution context [91], the activation of a canalizing gene occurs in response to diverse stress signals or situations where special attention is needed and results in a regulatory mode switch. There are multiple opportunities for canalizing behavior to be observed along the signal-transducing pathway that governs central cellular functions such as cell-cycle, survival, apoptosis and metabolism [60]. Early observations of canalization along the mitogenic pathway involved dual specificity protein phosphatase 1 (DUSP1) and Ras [92]. DUSP1 antagonizes the activity of the p38 mitogen activated kinase, MAPK1 (ERK), which is a central component of the pathway by which extracellular signal-regulated kinases send mitogenic signals [93]; thus, this gene is canalizing in its phosphorylated state, and DUSP1 is canalizing when it dephosphorylates MAPK1 [60]. Another important instance of canalization involves the tumor protein 53 (p53) gene with regard to stresses to the genome [63]. While canalizing genes can be extremely potent, their potency is often obscured by other features of the regulatory apparatus operating in the particular cell where control is attempted [60].

Martins *et al*. [60] proposed Intrinsically Multivariate Predictive (IMP) scores, which quantify the synergistic prediction effect of multiple genes, and provided evidence that IMP could potentially be used as a practical tool for discovery of canalizing genes. Chen *et al*. [94] developed a statistical tool for this inference problem based on the IMP score by providing a test for a nonzero IMP score between a Boolean target and its respective Boolean predictors. Rejection of the null hypothesis of zero IMP score at a given level of statistical significance gives evidence for the presence of IMP properties. Zhao *et al*. [56] defined canalizing power in a tree model in the context of Bayesian networks. The canalizing power of a gene in the study by Zhao and co-authors measures the total increase in prediction power using pairs of predictors over the maximum prediction power of the respective single predictors, which is equivalently the sum of the IMP scores from all genes in the model. The paper concludes that target genes showing large IMP scores with multiple predictor sets tend to be canalizing. However, when single predictors provide perfect predictions for a canalizing gene, the sum of the IMP scores becomes zero, leading to a paradoxical result: the canalizing power is zero. Furthermore, a key characteristic of a canalizing gene is its ability to override other regulatory instructions and none of the previously mentioned papers considers terms associated with the regulation power of other controlling genes that lose control by the activation of the canalizing gene. Although Zhao *et al*. [56] suggested a formula to measure the canalizing power of a gene, their definition fails to capture the incapacitating trait of canalizing genes. Therefore, we introduce a novel definition of the canalizing power that can quantitatively characterize the power of a canalizing gene based on two important characteristics: (i) It has to be sensitive to the strength of the influence of the canalizing gene on downstream genes; (ii) It should be able to detect how much the canalizing gene incapacitates other regulatory instructions upon its activation. The novelty of this chapter lies in the introduction of the notion of incapacitating power and development of a mathematical formula for canalizing power in terms of regulation power and incapacitating power.

This chapter is organized as follows. In Section 3.2, we present the Boolean Networks with random gene perturbations (BN_ps) as a model for gene regulatory networks and the concept of CoD. In Section 3.3, we define the regulation power, incapacitating power and canalizing power. In Section 3.4, we apply the novel definition of canalizing power to both synthetic data and real gene expression data to evaluate effectiveness of the proposed measurements in quantitatively characterizing canalizing genes. Finally, Section 3.5 gives concluding remarks.

3.2 Background

In this section, we provide the basic definitions and notations concerning Boolean Networks with random gene perturbations (BN$_p$s), and then review the notion of CoD. We restrict ourselves to the binary case and note that the methodology presented here presupposes that gene expression has been preprocessed and quantized into binary values. There are several methods that accomplish this [95], [96]. We do not address these methods in this dissertation, but they are naturally central to the accuracy of the results.

3.2.1 Boolean Networks with Gene Perturbations as a Model for Gene Regulatory Networks

A Boolean network $G(V, \boldsymbol{f})$ is defined by a set of binary-valued nodes $V = \{x_1, \cdots, x_n\}$ and a corresponding list of Boolean functions $\boldsymbol{f} = (f_1, \cdots, f_n)$. Each node x_i represents the state (expression) of gene i, where $x_i = 1$ means that gene i is expressed and $x_i = 0$ means it is not expressed. $\boldsymbol{f}$ represents the rules of regulatory interactions between genes. To every node x_i, a Boolean function $f_i : \{0,1\}^n \rightarrow \{0,1\}$ determining the value of gene x_i is assigned. It is known that genes may become either activated or inhibited due to external stimuli. Moreover, noise could also affect the Boolean relationships. To capture this uncertainty, we consider a BN with perturbation, which has been discussed in [97]. A Boolean additive-noise model with a random *perturbation vector*

$\boldsymbol{\gamma} \in \{0, 1\}^n$ is given by

$$x_i = f_i(x_1 \cdots x_n) \oplus \gamma_i \tag{3.1}$$

where f_i is a Boolean logic function of gene x_i, γ_i is the ith component of $\boldsymbol{\gamma}$ and $\oplus$ is modulo-2 addition. $\boldsymbol{\gamma}$ does not need to be independent and identically distributed (i.i.d.) and we suppose that $P\{\gamma_i = 1\} = p_i$. Then, equation (3.1) states that when $\gamma_i = 1$, the ith gene is flipped with probability p_i because of the noise, independently of other genes; otherwise it remains unperturbed. If $p_i = 0$ for all i, then the model is reduced to a deterministic Boolean Network and the standard network transition function $\boldsymbol{f}$ determines the evolution of the model. If $p_i > 0$, then with a probability $1 - \prod_{i=1}^{n}(1 - p_i)$, the current network state will change due to at least one random bit perturbation.

The randomness of this particular network model is encoded by the selection of the initial starting state of the network and also by the gene perturbation probabilities. In order to have a useful probabilistic description of this dynamical system, it is necessary to consider the joint probabilities of all of the genes over time. The dynamics of BNs can be modeled by Markov chains, consisting of 2^n states with the $2^n \times 2^n$ state transition matrix $P = (\Pr(\boldsymbol{s}, \boldsymbol{s}'))_{\boldsymbol{s},\boldsymbol{s}'}$ where $\Pr(\boldsymbol{s}, \boldsymbol{s}')$ is the probability of the chain undergoing the transition from $\boldsymbol{s}$ to $\boldsymbol{s}'$. The perturbation probability makes the chain ergodic and therefore it possesses a steady-state probability distribution π defined by $\pi^T P = \pi^T$, where T denotes transpose [98].

3.2.2 Coefficient of Determination

Let $Y \in \{0, 1\}$ be a binary target random variable and $\boldsymbol{X} \in \{0, 1\}^{\eta}$ be a vector composed of η binary predictor random variables. The CoD for $\boldsymbol{X}$ predicting Y is defined by

$$CoD_{\boldsymbol{X}}(Y) = \frac{\varepsilon_0(Y) - \boldsymbol{\varepsilon}_{\circ}(\boldsymbol{X}, Y)}{\varepsilon_0(Y)} \tag{3.2}$$

where $\varepsilon_0(Y) = min\{P(Y = 0), P(Y = 1)\}$ is the optimal error of predicting Y in the absence of observations and

$$\boldsymbol{\varepsilon}_{\circ}(\boldsymbol{X}, Y) = \sum_{\boldsymbol{x} \in \{0,1\}^{\eta}} min\{P(Y = 0, \boldsymbol{X} = \boldsymbol{x}), P(Y = 1, \boldsymbol{X} = \boldsymbol{x})\} \tag{3.3}$$

is the optimal error upon observation of $\boldsymbol{X}$ [99]. By convention, one assumes 0/0=1 in the above definition because zero prediction error indicates strong interaction between discrete predictor and target variable. The CoD measures the relative decrease in the classification/prediction error when optimally predicting a random variable Y using random vector $\boldsymbol{X}$ as opposed to optimally predicting Y based only on its own statistics. The CoD measures the inherent strength of the nonlinear interaction between a target gene and its predictors and is therefore more appropriate to genomics than the correlation coefficient, which only measures linear interaction. If $CoD_{\boldsymbol{X}}(Y) = 0$, there is no association between $\boldsymbol{X}$ and Y , whereas if $CoD_{\boldsymbol{X}}(Y) = 1$, then $\boldsymbol{X}$ and Y are deterministically related. The CoD measures nonlinear association (increase in prediction power), not causality. Moreover, the CoD is often used to measure the strength of downstream genes predicting upstream genes. The intuition behind this interpretation is that, if gene Y regulates genes X_1 and X_2, the observation of X_1 and X_2 should allow one to predict the behavior of Y. Moreover, the stronger the control by Y, the stronger is the prediction based on X_1 and X_2.

3.3 Definition of Canalizing Power

3.3.1 Regulation Power

In [56], the mean CoD value of a gene was defined to represent its regulatory importance in the model. Specifically, the mean CoD of a node Y using all single predictors of $\boldsymbol{X} = (X_1, \cdots X_\eta)$ is given by

$$\overline{CoD_{\boldsymbol{X},1}(Y)} = \frac{\sum_{i=1}^{\eta} CoD_{X_i}(Y)}{\eta} \tag{3.4}$$

Similarly, the mean CoD of a node Y using all sets of double predictors is given by

$$\overline{CoD_{\boldsymbol{X},2}(Y)} = \frac{\sum_{1 \le i < j \le \eta} CoD_{X_i, X_j}(Y)}{{}_{\eta}C_2} \tag{3.5}$$

A generalized definition for the mean CoD using d predictors is given by

$$\overline{CoD_{\boldsymbol{X},d}(Y)} = \frac{\sum_{i=1}^{{}_{\eta}C_d} CoD_{\boldsymbol{X}_d^{(i)}}(Y)}{{}_{\eta}C_d} = RP_{\boldsymbol{X},d}(Y) \tag{3.6}$$

where $\boldsymbol{X}_d^{(i)} \in \mathbb{R}^d$ is the ith d-dimensional vector composed of the elements of $\boldsymbol{X}$ when all possible combinations of size d from the array $\boldsymbol{X}$ are lexicographically ordered for $i = 1, \cdots, {}_{\eta}C_d$ (e.g., $\boldsymbol{X}_3^{(1)} = (X_1, X_2, X_3)$, $\boldsymbol{X}_3^{(2)} = (X_1, X_2, X_4), \cdots, \boldsymbol{X}_3^{(20)} = (X_4, X_5, X_6)$ when η=6 and d=3). Equation (3.6) gives the average strength of predicting Y by using all possible combinations of size d formed by the genes in the network. This general mean CoD measures the influence of the gene Y on the overall network and therefore we call it the "d-regulation power" of the gene Y in the network when d predictors are used for measurements and denote it by $RP_{\mathbf{X},d}(Y)$.

3.3.2 Incapacitating and Enhancing Power

In this section, we assume that $S = \{s_1, \cdots s_\alpha\}$ is a set of regulated/slave genes. Furthermore, suppose that there is a master gene m_i which controls the slave genes in a regular network regime and there is a canalizing gene c that is capable of overriding the instructions from the master genes. Intuitively, the regulation power of m_i could experience significant changes depending on the activation of c. The conditional CoD for S predicting m_i given that c is on is defined by

$$COD_S(m_i|c=1) = \frac{\varepsilon_0(m_i|c=1) - \boldsymbol{\varepsilon}_\circ(S, m_i|c=1)}{\varepsilon_0(m_i|c=1)} \tag{3.7}$$

where $\varepsilon_0(m_i|c=1)$ is the error of the best predictor of m_i in the absence of observations under the condition that c is turned on and $\boldsymbol{\varepsilon}_\circ(S, m_i|c=1)$ is the error of the best predictor of m_i based on the observation of S when c is on. Change in control of S by m_i relative to the activity of c is defined by

$$\Delta CoD_S(m_i|c) = CoD_S(m_i|c=0) - CoD_S(m_i|c=1). \tag{3.8}$$

A positive value of $\Delta CoD_S(m_i|c)$ indicates that c incapapcitates m_i as c is turned on. We call this value the *incapacitating power* (IP) of c relative to the regulation of S by m_i. A negative value of $\Delta CoD_S(m_i|c)$ means that there is an increase in control of m_i over S as

c is turned on and the magnitude is referred to as the *enhancing power* (EP) of c with respect to m_i upon the activation of c. This can be written as

$$|\Delta CoD_S(m_i|c)| = \begin{cases} IP_S(m_i|c) & if \;\; \Delta CoD_S(m_i|c) > 0 \\ EP_S(m_i|c) & if \;\; \Delta CoD_S(m_i|c) < 0 \end{cases} \tag{3.9}$$

Equation (3.8) can be generalized to (3.10) if one wants to consider all possible subsets of size $d \leq \alpha$ of predictors in S being used:

$$\Delta CoD_{S,d}(m_i|c) = \frac{\sum_{j=1}^{{}_\alpha C_d} CoD_{\boldsymbol{v}_d^{(j)}}(m_i|c=0) - CoD_{\boldsymbol{v}_d^{(j)}}(m_i|c=1)}{{}_\alpha C_d} \tag{3.10}$$

where $\boldsymbol{v}_d^{(j)} \in \mathbb{R}^d$ is jth d-dimensional vector consisting of the entries of $S = \{s_1, \cdots s_\alpha\}$ when all possible combinations of size d from S are lexicographically ordered for $j = 1, \cdots, {}_\alpha C_d$.

3.3.3 Canalizing Power

In this section, we define the *canalizing power* of the gene c, $CP_{S \cup M}(c)$, as a quantitative measure of canalization potential of a gene c relative to the set of genes $S \cup M = \{s_1, \cdots, s_\alpha, m_1, \cdots, m_\beta\}$, where $S = \{s_1, \cdots, s_\alpha\}$ and $M = \{m_1, \cdots, m_\beta\}$ are sets of slave genes and master genes, respectively. The canalizing power of gene c is expressed in terms of the regulation power and incapacitating power of c. This follows from the intuition that canalizing power should be quantified by the control of a gene c on the overall network and the extent of control that the gene c takes over from master genes when c is activated, which is equivalent to reduction of the control of the master genes M due to gene c. Thus, the canalizing power of gene c is given by

$$CP_{S\cup M,d}(c) = RP_{S\cup M,d}(c) + \sum_{i} IP_{S,d}(m_i|c)$$

$$= RP_{S\cup M,d}(c) + \sum_{i} \Delta CoD_{S,d}(m_i|c) \times 1_{[CoD_{S,d}(m_i|c=0)-CoD_{S,d}(m_i|c=1)>0]} \quad (3.11)$$

where $1[\cdot]$ is an indicator function. Note that the summation is over only those master genes that have been incapacitated by the activation of c.

3.3.4 Applications

Consider a network consisting of n genes and assume that it has a canalizing gene and one is interested in detecting it. One possible approach to do this is to sort out all of the controlling genes which could be either a master gene or a canalizing gene by computing the mean CoD because both master and canalizing genes should exhibit high regulation power. Hypothesis testing based on user-selectable thresholds or statistical tools presented in [94], [100] can be also used for picking out controlling genes. Suppose that we constitute a set of controlling genes $Z = \{z_1, \cdots, z_a\}$ and slave genes $S = \{s_1, \cdots, s_b\}$, where $n = a + b$. Furthermore, let $Z_{-j} = Z\backslash\{z_j\} = \{z_1, \cdots, z_{j-1}, z_{j+1} \cdots, z_a\}$ be the set Z without the element z_j. The canalizing power of z_j is

$$CP_{Z_{-j}\cup S,d}(z_j) = RP_{Z_{-j}\cup S,d}(z_j) + \sum_{k\neq j} IP_{S,d}(z_k|z_j)$$

$$= RP_{Z_{-j}\cup S,d}(z_j) + \sum_{k\neq j} \Delta CoD_{S,d}(z_k|z_j)$$

$$\times 1_{[CoD_{S,d}(z_k|z_j=0)-CoD_{S,d}(z_k|z_j=1)>0]}. \quad (3.12)$$

By taking turns, compute the canalizing power for each of the gene in the set of controlling genes Z. The gene z_{i^*} possessing the maximum canalizing power is the most likely candidate for the canalizing gene with respect to our model assumptions, where

$$\begin{aligned} i^* &= \underset{j\in 1,\cdots,a}{argmax}\, CP_{Z_{-j}\cup S,d}\left(z_j\right) \\ &= \underset{j\in 1,\cdots,a}{argmax}\left[RP_{Z_{-j}\cup S,d}\left(z_j\right) + \sum_{k\neq j} IP_{S,d}\left(z_k|z_j\right)\right] \end{aligned} \qquad (3.13)$$

Since the power of incapacitation is a key attribute of canalizing genes which can be used to distinguish canalizing genes from master genes, only the second term in (3.13) can be utilized for a fast approximate search. Thus,

$$\begin{aligned} i^* &\approx \underset{j\in 1,\cdots,a}{argmax} \sum_{k\neq j} IP_{S,d}\left(z_k|z_j\right) \\ &= \underset{j\in 1,\cdots,a}{argmax} \sum_{k\neq j} \Delta CoD_{S,d}\left(z_k|z_j\right) \times 1_{\left[CoD_{S,d}\left(z_k|z_j=0\right) - CoD_{S,d}\left(z_k|z_j=1\right)>0\right]}. \end{aligned} \qquad (3.14)$$

3.4 Results

In this section, we illustrate the application of canalizing power in a number of experiments using both synthetic data and real data sets.

3.4.1 Synthetic Data

We generate a synthetic BN with n=10 genes as shown in Figure 3.1 which is composed of one canalizing gene C, two master genes M_1 and M_2 and three levels of slave

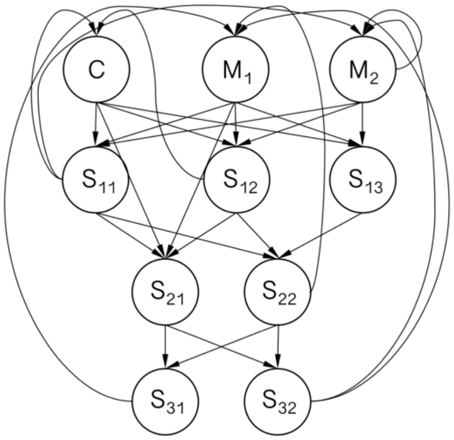


Figure 3.1 A synthetic BN with n=10 genes which is composed of one canalizing gene C, two master genes M_1 and M_2 and three levels of slave genes $S_{11}, \cdots S_{32}$. Upstream genes C, M_1 and M_2 regulate downstream genes and downstream genes provide feedback signals to the upregulators.

genes $S_{11}, \cdots S_{32}$. Regulatory influences on downstream genes are transferred between master genes and the canalizing gene depending on the activity of C. Thus, Boolean functions that govern the activity of downstream genes are designed to differ according to the expression of the canalizing gene C and therefore, the canalizing gene is embedded in the network by these Boolean rules. When there is no noise, the system transitions in accordance with its structural rules as defined by the Boolean functions listed in Table 3.1. The regulation power, incapacitating power and canalizing power of controlling genes are measured at each time point along the network evolution under various settings of the model parameters. Each target is predicted by $d = 3$ predictors. Given the network, we consider four different simulation scenarios: 1) no gene is perturbed, 2) only one specific gene is perturbed while other genes are noiseless, 3) all genes are perturbed with equal probability and 4) all of the genes are susceptible to noise where the perturbation probability for each gene is randomly generated from a beta distribution. Since the

Table 3.1 Boolean functions of genes in the synthetic BN, where the symbols $\vee$, $\wedge$ and $\oplus$ denote the Boolean disjunction, conjunction and exclusive-OR, respectively.

		Boolean Expression	C	C
			Inactivated	Activated
Controlling Genes	C	$S_{11} \oplus S_{12}$		
	M_1	$S_{22} \wedge (\overline{S_{31}} \oplus S_{32})$		
	M_2	$M_2 \wedge S_{11} \wedge S_{32}$		
Level 1	S_{11}	$C \vee M_1 \vee M_2$	$M_1 \vee M_2$	C
	S_{12}	$\bar{C} \wedge M_2$	M_2	$\bar{C}$
	S_{13}	$C \vee (M_1 \oplus M_2)$	$M_1 \oplus M_2$	C
Level 2	S_{21}	$S_{11} \wedge S_{12} \vee \bar{C} \wedge M_1$	$M_1 \vee M_2$	$\bar{C}$
	S_{22}	$S_{11} \vee S_{12} \wedge \overline{S_{13}}$	$M_1 \vee M_2$	C
Level 3	S_{31}	$S_{21} \wedge S_{22}$	$M_1 \vee M_2$	$\bar{C}$
	S_{32}	$S_{21} \oplus S_{22}$	0	C

behavior of the network depends not only on the perturbation probabilities but also on the initial state distribution, we compute the average RP, IP and CP over ten thousand random generations of its initial joint probability distribution, D_0.

In the first case, no gene is perturbed and we plot the mean RP, IP and CP measured at each time step averaged over 10,000 random starting joint probability distributions. Figure 3.2(A) shows that the mean regulation power of C is similar to or even less than that of M_1, whereas incapacitating power is exhibited only for the canalizing gene. This leads to higher CP of C, which indicates that the incapacitating power is a key attribute of canalizing genes that can be used to distinguish canalizing genes from other controlling genes.

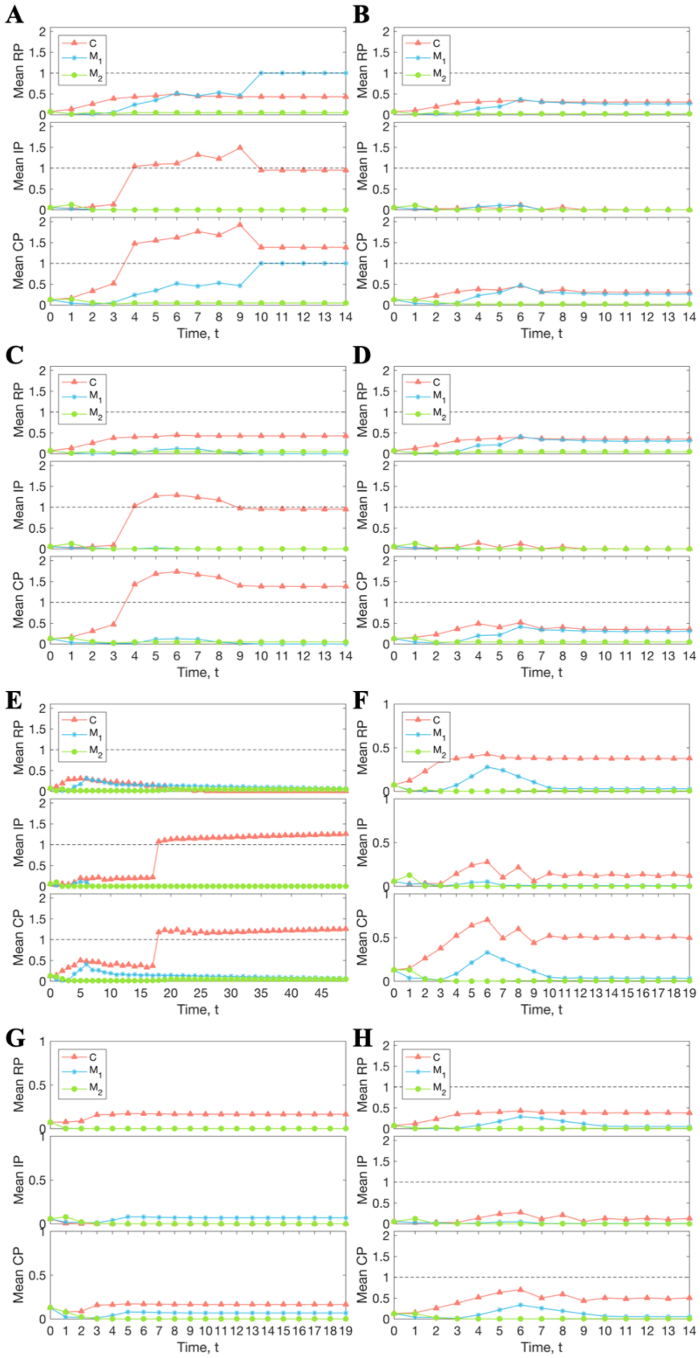


Figure 3.2 Mean regulation power, incapacitating power and canalizing power over time (A) when no gene is perturbed. (B) A particular controlling gene is perturbed with $P_C =$ 0.1 and (C) $P_{M_1} = 0.1$. (D) Effects of noise in the expression of downstream genes on mean RP, IP and CP when $P_{S_{12}} = 0.1$ and (E) $P_{S_{22}} = 0.1$. All genes are perturbed with the same probability (F) $P = 0.01$ and (G) $P = 0.1$. (H) All genes are perturbed with different probabilities which are randomly generated from $beta(2,200)$ distribution.

The results for the second case where only one specific controlling gene is perturbed are presented in Figure 3.2(B) and 3.2(C). In Figure 3.2(B), the canalizing gene is perturbed with a probability $P_C = 0.1$ and other genes are not perturbed at all. Presence of noise in the canalizing gene corrupts its gene expression resulting in its lower IP, which negatively impacts its canalizing power. A case where only M_1 is perturbed with a probability $P_{M_1} = 0.1$ is shown in Figure 3.2(C). In concordance with intuition, RP and IP of the canalizing gene is hardly impacted which does not compromise its predominance in CP. Effects of noise imposed on each of the downstream genes S_{12} and S_{22} with perturbation probabilities $P_{S_{12}} = 0.1$ and $P_{S_{22}} = 0.1$ are presented in Figure 3.2(D) and 3.2(E), respectively. S_{12} is given as an input to C and therefore, IP of C deteriorates substantially which makes CP of C contiguous to that of M_1 across the timeline when $P_{S_{12}} = 0.1$. S_{22} provides a feedback signal to the master gene M_1, thus, the RP of M_1 dwindles when $P_{S_{22}} = 0.1$ as illustrated in Figure 3.2(E). While there is little noticeable distinction in regulation power between C and M_1, IP of the canalizing gene is remarkably higher in comparison to the rest of controlling genes, resulting in CP of C being greater than M_1 and M_2.

For the next group of experiments, all of the genes are equally perturbed. Figure 3.2(F) shows that when the common perturbation probability is relatively small, $P = 0.01$, C experiences a decrease in its IP and CP. However, it still remains the gene with the highest canalizing potential in the network. When the perturbation probability is increased to $P = 0.1$, IP of C is virtually nonexistent and mean canalizing power of C has fallen

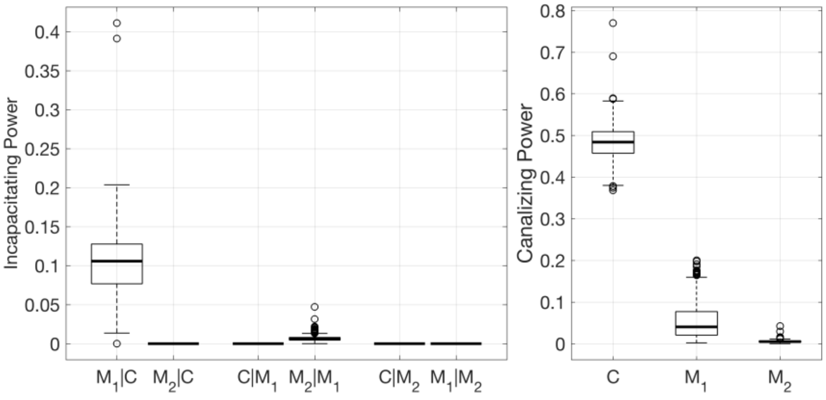


Figure 3.3 Boxplots of IP and CP of upstream genes. The label $M_1|C$ on the horizontal axis of the left panel represents that the first boxplot indicates a decrease in control of M_1 over downstream genes as C is turned on. The boxplots are based on the data measured at $t = 14$ and generated from random starting joint probability distributions.

below 0.18 as depicted in Figure 3.2(G). This suggests that the amount of noise in the regulatory network could negatively affect the canalizing potential of certain genes.

For the final group of simulation experiments, all genes are perturbed with different probabilities. The beta distribution, which is defined on the interval [0,1], can represent all the possible values of a probability and it is widely used as a probability distribution of probabilities [101]. The perturbation probability for each gene is randomly generated from a beta distribution with two parameters $\alpha = 2$, $\beta = 200$, which is a right-skewed distribution with mean 0.0099 to introduce a moderately small perturbation. The results are displayed in Figure 3.2(H). When the entire network is exposed to such type of random noise, RP of M_1 decreases over time and while the canalizing gene C remains the most potent canalizer in the network despite its diminished IP. Boxplots of incapacitating power of controlling genes measured at t =14 are shown in Figure 3.3 and the first boxplot represents a decrease in control of M_1 over downstream genes as C is turned on. The boxplots are based on the 10,000 samples which are generated from random starting joint probability distributions under the same experimental conditions as used for Figure 3.2(H). The expected canalizing power of C is clearly higher than the CP of the rest of the controllers in the network: $E[CP_C] = 0.483$, $E[CP_{M_1}] = 0.053$, $E[CP_{M_2}] = 0.005$.

3.4.2 Real Data

In this section, the proposed definition of the canalizing power is applied to a real data set from a study on ionizing radiaiton (IR) responsive genes in [102] to assess the usefulness of our quantification in characterizing a canalizing gene. Note that our goal is not to discover new canalizing genes, but rather to illustrate the potential of our measurement on well-known canalizing genes. The data set consists of 12 genes under three conditions (i.e., IR, MMS, UV) in 30 cell lines of both p53 proficient and p53 deficient cells. The data are ternary, indicating up-regulated (+1), down-regulated (-1), or no-change (0) status. Here we map this to binary expressions using the following rules: change (1), for either up-regulated or down-regulated genes, and no-change (0). Additionally, we consider the three binary conditions (IR, MMS, and UV) as possible predictive factors, for a total of 15 Boolean variables in the BN model of this data set. We then apply the definitions of regulation power, incapacitating power and canalizing power outlined in the previous section. Figure 3.4 shows a bar chart with the canalizing power of each gene when triple predictors are used (d = 3). It is stacked to display the regulation power and incapacitating power of each gene. p53 turns out to be the most powerful canalizing gene in the data set. This is in accordance with the known fact that p53 is kept

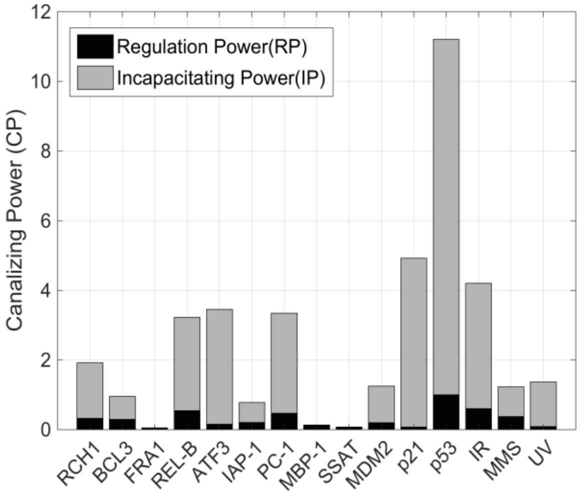


Figure 3.4 A stacked bar chart of CP for each gene in the real dataset. The height of the black and gray bar segments represent contributions of RP and IP to CP, respectively.

at a low level/dephosphorylated in unstressed cells and becomes significantly activated/phosphorylated in response to environmental stresses like UV, IR and oxidative stress, leading to a quick accumulation of p53 in stressed cells [103].

3.5 Discussion

It is a well-established notion in biology that canalizing genes possess broad regulatory power, and can enforce broad corrective actions. Canalizing genes can be extremely potent not only because they produce optimal reactions to operating errors and external stimuli, but also because they don’t act alone. Canalizing genes are more like master switches that set in motion a cascade of regulatory events that have huge impacts on downstream genes for the sake of driving the system to a desired condition. Discovering such potential drug targets that affect the disease trajectories is a strong step toward significant therapeutic benefits. From the perspective of optimal control, this is viewed as obtaining the best estimates of inputs which are most probable to elicit certain behavior of the network. However, the detection of these genes is circumscribed by their particular behavior. Under normal cell conditions, canalizing genes are not active and they are turned on only when cells encounter unfavorable situation. p53, one of the most intensively studied tumor suppressor genes, best describes this situation in which it is found at very low levels in normal cells while it is frequently observed in its phosphorylated state cancer-prone cells.

Although there have been several studies attempting to mathematically characterize canalizing genes and their power, they all missed the opportunity to characterize an important property of canalizing genes; that is, their incapacitating power. Therefore, we introduce a conditional CoD that characterizes predictive power of a set of genes with respect to a target gene under a specific condition of other genes. Our approach also suggests that the currently adopted definitions of canalizing and master genes could be modified so that a particular gene does not have to be exclusively a master or a canalizing gene. The newly introduced canalizing power resides in the continuous domain; therefore it presents a relative characterization of controlling genes. Although we have focused on BNs with perturbations to validate our ideas in a simplified environment, the same concept can be easily extended to Probabilistic Boolean Networks (PBNs), which offers more model flexibility.

4. NETWORK CLASSIFICATION BASED ON REDUCIBILITY WITH RESPECT TO THE STABILITY OF CANALIZING POWER OF GENES IN A GENE REGULATORY NETWORK – A BOOLEAN NETWORK MODELING PERSPECTIVE

In the previous chapter, we developed a quantitative framework that reflects inherent characteristics of canalizing genes and allows the estimation of the power of canalizing genes. In this chapter, we use the proposed measurement in reducing Boolean network with perturbation.

4.1 Introduction

Probabilistic Boolean networks (PBNs) form a widely accepted mathematical model for cellular systems and gene regulatory networks [98]. One of their important applications is to design intervention strategies that beneficially alter cell dynamics through studying long-run network behavior. Because the dynamics of a PBN are represented by an ergodic and irreducible finite Markov chain, the model possesses a steady-state distribution (SSD) reflecting its long-run dynamics. Various types of stochastic control policies for PBNs have been employed to change the long-run dynamics of the model, with immediate implications to practical problems such as reducing the risk of entering aberrant states and thereby altering the extant cell behavior [104]; however, owing to the inherent computational complexity of optimal control methods using Markov chain theory, it is often infeasible to design optimal control policies for large networks

[105], [106]. Several approximate and greedy algorithms [107]-[109] have been proposed to find suboptimal solutions but many of them still have complexity that increases exponentially or hyper exponentially with the number of genes in the network. Even relatively small networks can pose serious difficulties in assessing the dynamics, considering that a Boolean network of n genes has 2^n states and the transition probability matrix has size $2^n \times 2^n$. Given the exponential dependence of the state space on the number of nodes, there is a need for network reduction mappings that produce more tractable models whose stationary control policies induce suboptimal stationary control policies on the original network.

While past efforts to network reduction focused on developing reduction algorithms that maintain structural consistency or the dynamical behavior of the original network [110], [77], the major focus of our work is the preservation of the canalizing properties of genes in the original network. For this purpose, we examine what happens to canalization when genes are consecutively deleted. It is hypothesized that deleting a gene with the smallest canalizing power may help to preserve the canalizational properties of the original network under network reduction. The work in this chapter is centered around preservation of gene canalizing power under network reduction mappings. An important observation made in the course of the study leads to the hypothesis that genes in some networks retain their canalization properties after network compression, while there is a class of networks that do not possess this property. Naturally, this hypothesis leads to definitions of reducible and irreducible networks. Thus, one can formulate a related classification problem that aims to find relevant network features that can separate reducible from irreducible networks.

Compelling evidence for the existence of such network features/parameters is found in previous studies [111]-[116]. A previous study [117] showed that the dynamics of Boolean networks are mainly determined by two parameters N and K, where N is the number of nodes in the network and K is the average number of directional links between them. A change of the network dynamics from chaotic to orderly behavior was observed at some critical value of the connectivity parameter K. There have been several studies investigating the structural properties of networks in relation to their operating regime [112], [117], [118]. The most frequently studied network parameters have been the average connectivity K and the function bias which represents the probability for a Boolean function to take on value 1. In addition, it has been demonstrated that networks constructed from functions belonging to various classes, such as canalizing functions or certain Post classes, can also exhibit a tendency toward ordered behavior [119]-[121]. These results suggest that network parameter space can be partitioned, which naturally leads to the formulation of various classification problems.

This chapter extends the above considerations to parameters pertaining to network reduction and canalizational stability. Hence, we quantitatively define different classes of networks in relation to canalizing power robustness under model reduction. Our hypothesis about the existence of two classes, reducible and irreducible networks with respect to the preservation of canalizing power, requires a systematic empirical study in order to properly define the two classes. After completing the study, we introduce the definition of reducible networks in terms of canalizing robustness and proceed with the problem of selecting the relevant network features which allow discriminating reducible from irreducible networks. For the corresponding classification problem, the feature selection part aims to select a subset of highly discriminating features. The goals of this study are to (i) examine the stability of canalization under the network reduction mapping and (ii) find relevant network features that characterize different classes of networks. It is important to note that our novel approach relies only on estimates of canalizing power of the participating genes and does not assume any specific information about the Boolean rules between nodes. Thus, one can approach the network reduction problem without explicitly inferring the network model from data.

This chapter is organized as follows. In Section 4.2, we present background information about reduction mapping and discuss a tentative list of network features that might be useful when solving classification problems. In addition, the definition of reducible networks is presented and the general process of simulation is described. Section 4.3 shows experimental results and Section 4.4 provides concluding remarks.

4.2 Systems and Methods

4.2.1 Reduction Mapping

Consider a mapping ψ: $G \longrightarrow \tilde{G}$ that transforms the original network $G(V, \boldsymbol{f})$ into a new one, where a gene is deleted. A specific type of reduction mapping ψ was proposed in [122]. Assuming gene x_j is to be deleted from the network, the reduction mapping defines the transition rules for states in the network where that gene is removed, i.e. 1-

reduced network. Every predictor $f_i \in \boldsymbol{f}$ generates two predictors $\hat{f}_{i,0}$ and $\hat{f}_{i,1}$ according to the rule

$$\hat{f}_{i,g}(x_1, \cdots, x_{j-1}, x_{j+1}, \cdots, x_n) = f_i(x_1, \cdots, x_{j-1}, g, x_{j+1}, \cdots, x_n), \tag{4.1}$$

where $g \in \{0, 1\}$. The selection policy we use here is suggested in [122] and is based only on the SSD of the network. The selection of every function $\tilde{f}_i \in \tilde{\boldsymbol{f}}$ is performed pointwise, and for two states that only differ in the deleted gene x_j, the state transitions of the states possessing larger steady-state probability mass will be kept as transitions for the reduced states, excluding the gene for deletion. Therefore, a selection procedure for the function $\tilde{f}_i$ is given as follows:

(a) For all i, select numbers $-1 \leq \omega_i \leq 1$.

(b) For every state $\boldsymbol{s} = (x_1, \cdots, x_{j-1}, x_{j+1}, \cdots, x_n)$, define

$$\tilde{f}_i(\boldsymbol{s}) = \begin{cases} \hat{f}_{i,0}(\boldsymbol{s}) & if \ \Pr\{(x_1, \cdots, x_{j-1}, 0, x_{j+1}, \cdots, x_n)\} \\ & \qquad > \omega_i + \Pr\{(x_1, \cdots, x_{j-1}, 1, x_{j+1}, \cdots, x_n)\}; \\ \hat{f}_{i,1}(\boldsymbol{s}) & otherwise. \end{cases}$$

We set $\omega_i = 0$, which means that we do not assume any additional information. In addition, the perturbation probability p remains the same after applying the reduction mapping. Assuming that the original network has n genes, the m-reduced network can be defined by a set of the $n - m$ remaining genes based on m deletion-selection applications.

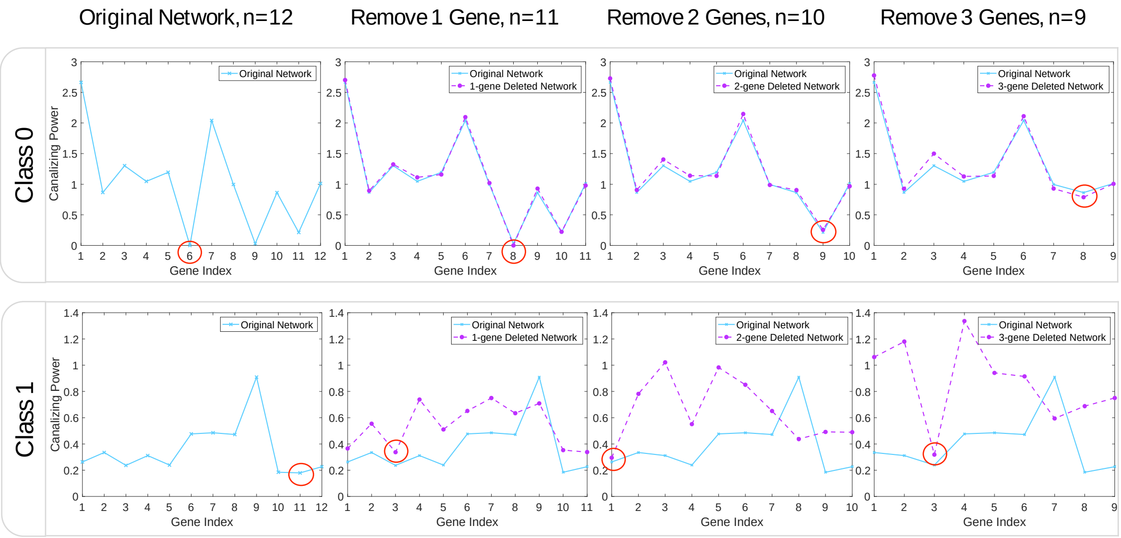


Figure 4.1 Each row corresponds to a network that belongs to the respective class. The columns represent the results of the consecutive removal of three genes from the network. X-axes represent the gene index in the reduced network and the Y-axes indicate the canalizing power of genes. A gene with the smallest canalizing power is marked with circle and is deleted by applying the reduction mapping. Note that in the case of the network from Class 0, the canalizing power trends in its CP vectors are maintained up to and including the deletion of three consecutive genes while the network from the Class 1 loses the original distribution of canalizing power even after a single gene is removed.

4.2.2 Gene Deletion

Let $V = \{x_1, \cdots, x_n\}$ be the set of genes in the original network. Supposing the jth gene is deleted from the network, the set of genes in the 1-reduced network is $\tilde{V} = \{x_1, \cdots, x_{j-1}, x_{j+1}, \cdots, x_n\}$. After a gene removal, the remaining genes are re-indexed: the indices of $\{x_{j+1}, \cdots, x_n\}$ are decreased by 1, and thus $\{\tilde{x}_1, \cdots \tilde{x}_{n-1}\} = \{x_1, \cdots, x_{j-1}, x_{j+1}, \cdots, x_n\}$. For example, consider the network that belongs to class 0, Figure 4.1. Gene 6 is removed from the original network and the remaining genes are relabeled. Therefore, gene 7 in the original network becomes gene 6 in the 1-reduced network. Consider the canalizing power of x_i with respect to the rest of the genes in the original network $CP_{V\backslash\{x_i\},d}(x_i)$ as defined in (3.11). In this study, we assume that there is no prior knowledge of controlling genes and downstream regulated genes. Thus, we compute the canalizing power of each gene x_i in the network assuming that $S \cup M = V\backslash\{x_i\}$. Let $\boldsymbol{Q}_{n-1}$ denote the $(n-1)$-dimensional CP vector consisting of the canalizing power of genes in the original network, excluding the gene marked for deletion:

$$\begin{aligned}\boldsymbol{Q}_{n-1} &= \left(CP_{V\backslash\{x_1\},d}(x_1), \cdots, CP_{V\backslash\{x_{j-1}\},d}(x_{j-1}), CP_{V\backslash\{x_{j+1}\},d}(x_{j+1}), \cdots, CP_{V\backslash\{x_n\},d}(x_n)\right)\\ &= \left(CP_{V\backslash\{\tilde{x}_1\},d}(\tilde{x}_1), \cdots, CP_{V\backslash\{\tilde{x}_{n-1}\},d}(\tilde{x}_{n-1})\right)\\ &= (q_1, \cdots, q_{n-1}). \qquad (4.2)\end{aligned}$$

The second line of (4.2) is obtained by re-indexing the remaining genes. For simplicity, we use the notation q_i in this work to denote $q_i = CP_{V\backslash\{\tilde{x}_i\},d}(\tilde{x}_i)$. Similarly, the CP vector of the canalizing power of genes in the 1-reduced network is given by

$$\widetilde{\boldsymbol{Q}}_{n-1} = \left(CP_{\tilde{V}\backslash\{\tilde{x}_1\},d}(\tilde{x}_1), \cdots, CP_{\tilde{V}\backslash\{\tilde{x}_{n-1}\},d}(\tilde{x}_{n-1})\right) = (\tilde{q}_1 \cdots, \tilde{q}_{n-1}), \qquad (4.3)$$

where $\tilde{q}_i = CP_{\tilde{V}\backslash\{\tilde{x}_i\},d}(\tilde{x}_i)$ is the canalizing power of $\tilde{x}_i$ with respect to the rest of the genes in the reduced network $\tilde{V}$. Similarly, one can obtain $\boldsymbol{Q}_{n-m}$ and $\widetilde{\boldsymbol{Q}}_{n-m}$ – the CP vectors of the original network and the m-reduced networks. The normalized Euclidean distance between them is given by

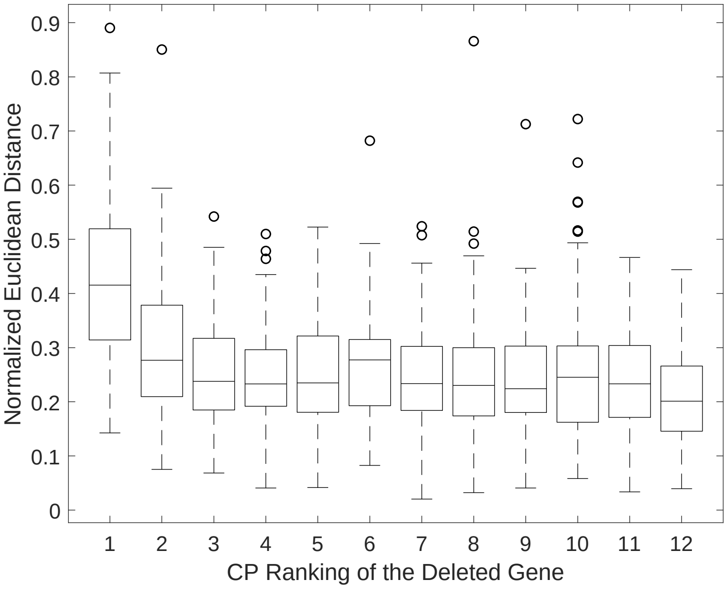


Figure 4.2 Boxplot of Euclidean distances between the CP vectors of the original network with n=12 and the respective 1-reduced network. Each number on the X-axis represents the CP ranking of the deleted gene in the original network, and the canalizing power is sorted in descending order. Therefore, the rightmost box in the graph corresponds to the case where a gene with the smallest canalizing power is deleted.

$$D_N\left(\boldsymbol{Q}_{n-m}, \widetilde{\boldsymbol{Q}}_{n-m}\right) = \sqrt{\sum_{i=1}^{n-m}\left(\frac{q_i}{\|\boldsymbol{Q}_{n-m}\|} - \frac{\tilde{q}_i}{\|\widetilde{\boldsymbol{Q}}_{n-m}\|}\right)^2}, \tag{4.4}$$

where ||·|| is the Euclidean norm.

In model reduction, we aim to preserve the canalizational properties of the original network. Our hypothesis is that the removal of the gene with the smallest CP would achieve this objective for a specific class of networks. To this end, we examine all of the genes in the network by removing one at a time by applying the reduction mapping described in Section 4.2.1 and measuring the normalized Euclidean distance between the respective CP vectors given by (4.4). Figure 4.2 shows the boxplot of normalized

Euclidean distances between $\boldsymbol{Q}_{n-1}$ and $\widetilde{\boldsymbol{Q}}_{n-1}$ when 1,000 randomly generated networks of n =12 genes are examined. The result suggests that removing the gene with the smallest CP from the original network might provide a useful heuristic to achieve network reduction with a minimal impact on the CP vectors.

4.2.3 Network Empirical Labeling

Our empirical study of network reduction based on removing the gene with the smallest CP shows that some networks retain their canalization properties, while others do not. Figure 4.1 shows examples of these two distinctive categories of networks and suggests that there might be classes of networks that are better suitable for reduction than others. This observation leads to the hypothesis that networks can be divided into two categories: (i) reducible, i.e. maintaining the existing distribution of canalizing power among genes after applying the reduction mapping based on the deletion of the gene with the smallest CP, and (ii) irreducible, i.e. where the network CP vectors are significantly altered after applying the reduction mapping. Preserving the order of genes with respect to their canalizing power and having small distance between the CP vectors of the original and reduced networks are the two criteria used in this work for evaluating the stability of canalizing power under model reduction. We empirically label the networks based on two criteria: (i) Spearman's rank-order correlation coefficient, and (ii) weighted Euclidean distance between CP vectors of the original and reduced networks.

4.2.3.1 Spearman's Rank Correlation Coefficient

The Spearman's rank correlation coefficient is defined as the Pearson correlation coefficient between ranked variables [123]. Suppose the original network has n genes and we remove one gene at a time until m genes ($0 < m < n$) are deleted. Let $\rho_k,\ k = 1, \cdots, m$, denotes the Spearman's rank correlation coefficient between $\boldsymbol{Q}_{n-k}$ and $\widetilde{\boldsymbol{Q}}_{n-k}$, where $\boldsymbol{Q}_{n-k} = (q_1, \cdots, q_{n-k})$ and $\widetilde{\boldsymbol{Q}}_{n-k} = (\tilde{q}_1, \cdots, \tilde{q}_{n-k})$ are the CP vectors of the original and the k-reduced network, respectively. Let $r(q_i, \boldsymbol{Q}_{n-k})$ denote the rank of q_i in $\boldsymbol{Q}_{n-k}$. Similarly, let $r(\tilde{q}_i, \widetilde{\boldsymbol{Q}}_{n-k})$ denote the rank of $\tilde{q}_i$ in the CP vector $\widetilde{\boldsymbol{Q}}_{n-k}$. Consider $\boldsymbol{r}_{n-k} = [r(q_1, \boldsymbol{Q}_{n-k}), \cdots, r(q_{n-k}, \boldsymbol{Q}_{n-k})]$ and $\tilde{\boldsymbol{r}}_{n-k} = [r(\tilde{q}_1, \widetilde{\boldsymbol{Q}}_{n-k}),\ \cdots, r(\tilde{q}_{n-k}, \widetilde{\boldsymbol{Q}}_{n-k})]$ and define

$$\rho_k = \frac{cov(\boldsymbol{r}_{n-k}, \tilde{\boldsymbol{r}}_{n-k})}{\sigma(\boldsymbol{r}_{n-k}) \cdot \sigma(\tilde{\boldsymbol{r}}_{n-k})}, \tag{4.5}$$

where $cov(\boldsymbol{r}_{n-k}, \tilde{\boldsymbol{r}}_{n-k})$ is the covariance of two CP rank vectors, and $\sigma(\boldsymbol{r}_{n-k})$ and $\sigma(\tilde{\boldsymbol{r}}_{n-k})$ are the standard deviations of $\boldsymbol{r}_{n-k}$ and $\tilde{\boldsymbol{r}}_{n-k}$, respectively. Next, we consider vector $\boldsymbol{\rho} = [\rho_1, \rho_2, \cdots, \rho_m]$, containing m rank correlation coefficients computed at each gene removal. Denote $\overline{\boldsymbol{\rho}}$ by the mean of the vector $\boldsymbol{\rho}$ and ρ_{min} by the minimum value of $\boldsymbol{\rho}$. Our first criterion for m-reducibility of a network is based on these two quantities $\overline{\boldsymbol{\rho}}$ and ρ_{min}: if $\overline{\boldsymbol{\rho}}$ is larger than a user-defined parameter θ_{avg} and ρ_{min} is greater than a user-defined threshold θ_{min}, the network is considered to be m-reducible from the perspective of canalizing power ranking preservation.

4.2.3.2 Weighted Euclidean Distance

Weighted Euclidean distance is used to measure the distance between the CP vectors of the original and reduced networks. We normalize the canalizing power of genes to unity to put CP on the same scale. Then, the weighted Euclidean distance [124] is given by

$$D\left(\boldsymbol{Q}_{n-k}, \widetilde{\boldsymbol{Q}}_{n-k}\right) = \sqrt{\sum_{i=1}^{n-k} w_i \left(\frac{q_i}{\|\boldsymbol{Q}_{n-k}\|} - \frac{\tilde{q}_i}{\|\widetilde{\boldsymbol{Q}}_{n-k}\|} \right)^2} \tag{4.6}$$

where $\|\cdot\|$ is the Euclidean norm and the weight for each gene is

$$w_i = \frac{\left| r(q_i, \boldsymbol{Q}_{n-k}) - r\left(\tilde{q}_i, \widetilde{\boldsymbol{Q}}_{n-k}\right) \right|}{min\left(r(q_i, \boldsymbol{Q}_{n-k}),\ r\left(\tilde{q}_i, \widetilde{\boldsymbol{Q}}_{n-k}\right) \right)} . \tag{4.7}$$

The weight w_i, assigned to the CP difference of the ith gene, is the CP ranking gap, $\left| r(q_i, \boldsymbol{Q}_{n-k}) - r\left(\tilde{q}_i, \widetilde{\boldsymbol{Q}}_{n-k}\right) \right|$, divided by the minimum of the two rankings. The denominator of w_i is used to penalize for changes in the upper ranks of the CP. Thus, this weighting scheme puts larger weights to bigger fluctuations in CP ranking and changes in higher ranks. The distance measure $D\left(\boldsymbol{Q}_{n-k}, \widetilde{\boldsymbol{Q}}_{n-k}\right)$ not only represents the differences in CP, but also involves the ranking of genes to penalize ranking discrepancies. Given $D\left(\boldsymbol{Q}_{n-k}, \widetilde{\boldsymbol{Q}}_{n-k}\right)$ for $k = 1, \cdots, m$, we consider

$$D_{max} = max\left(D\left(\boldsymbol{Q}_{n-1}, \widetilde{\boldsymbol{Q}}_{n-1}\right), \cdots, D\left(\boldsymbol{Q}_{n-m}, \widetilde{\boldsymbol{Q}}_{n-m}\right) \right). \tag{4.8}$$

Our second criterion for m-reducibility of a network is based on (4.8): if D_{max} is smaller than a threshold θ_{dist}, the network is considered m-reducible with respect to preserving the distance between the respective CP vectors. Combining the criteria 1 and 2 together, one arrives at the following definition:

Definition 1. *Given the parameters $\theta_{avg}, \theta_{min}$ and θ_{dist}, the Boolean network G is said to be m-gene reducible $(0 < m < n)$ if and only if the following conditions are satisfied:*

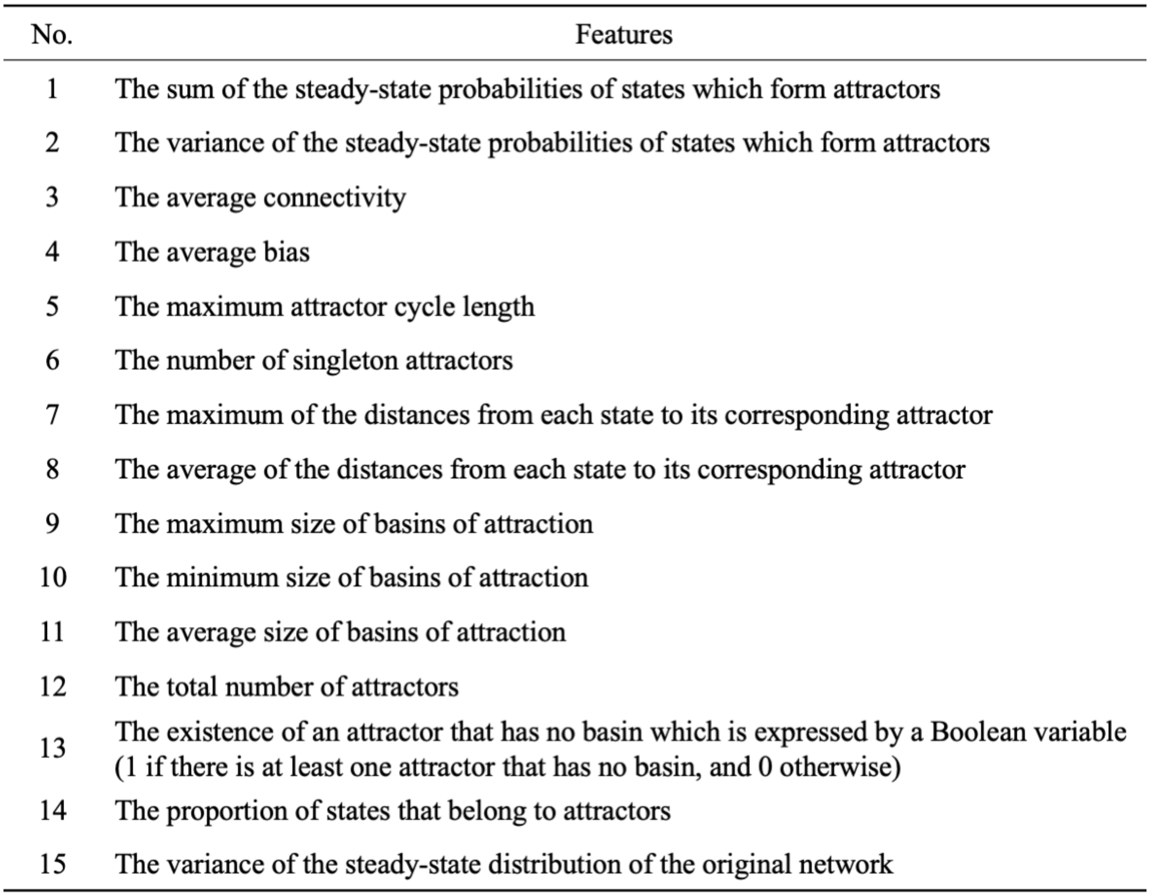
i. $\bar{\rho} \geq \theta_{avg}$ *and* $\rho_{min} \geq \theta_{min}$

ii. $D_{max} \leq \theta_{dist}$

4.2.4 Network Features

There are a number of parameters of a Boolean network that could be used to characterize different classes of networks. It is well known that the bias and average connectivity are important because they can modulate the order-disorder transition in the network dynamics [121]. Attractor structure is another important characteristic of the dynamical behavior of a Boolean network. Starting from any initial state, a BN will eventually enter a fixed state called a *singleton* or *fixed-point attractor,* or a set of states, called an *attractor cycle*, through which it will cycle endlessly. The attractors capture the model's essential long-term behavior and have been widely recognized to correspond to meaningful cellular phenotypes [125], [126]. Hence, we include values associated with the attractor structure, such as the *cycle length*, i.e., the number of states an attractor comprises, the number of singleton attractors, the proportion of states that belong to attractors, and mass of the attractors in the SSD of the network. Furthermore, each network

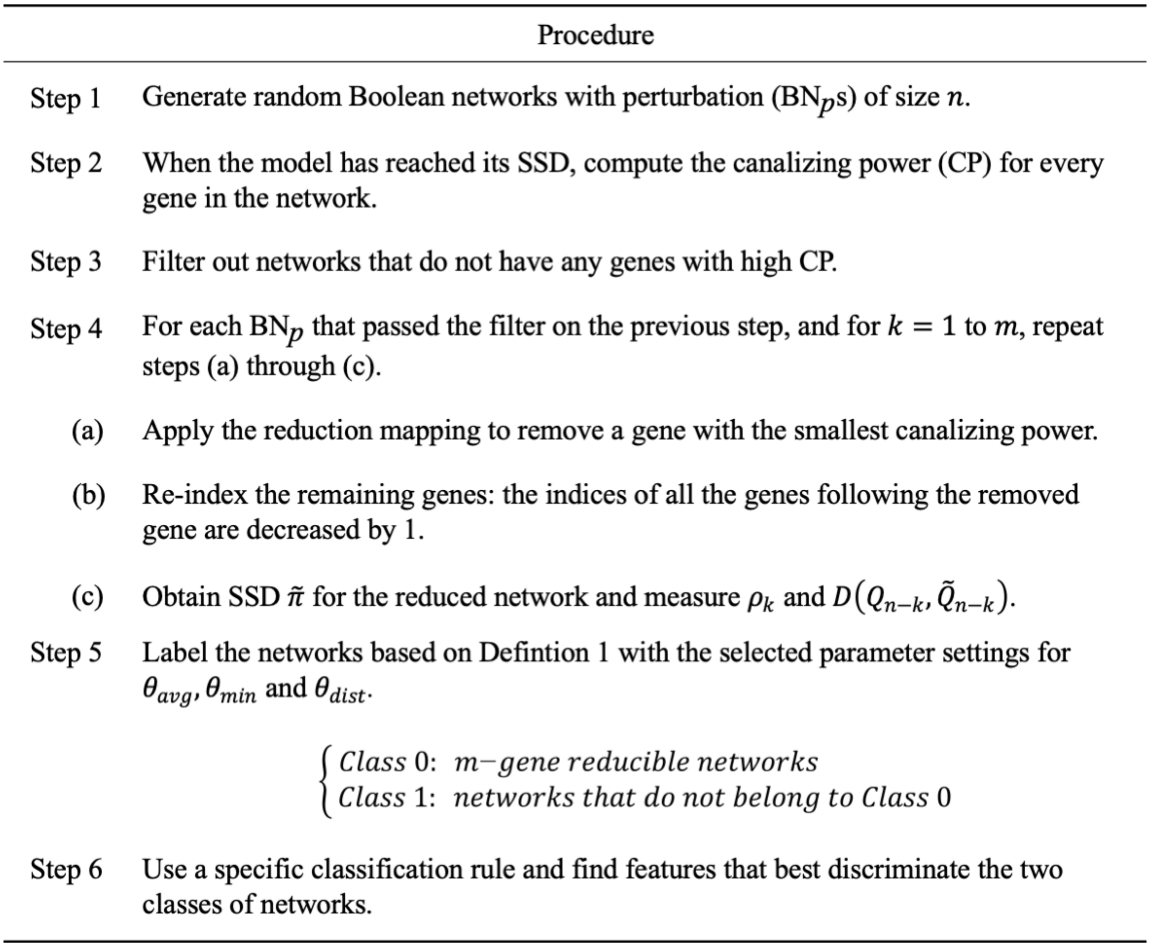
Table 4.1 The list of Boolean network attributes



No.	Features
1	The sum of the steady-state probabilities of states which form attractors
2	The variance of the steady-state probabilities of states which form attractors
3	The average connectivity
4	The average bias
5	The maximum attractor cycle length
6	The number of singleton attractors
7	The maximum of the distances from each state to its corresponding attractor
8	The average of the distances from each state to its corresponding attractor
9	The maximum size of basins of attraction
10	The minimum size of basins of attraction
11	The average size of basins of attraction
12	The total number of attractors
13	The existence of an attractor that has no basin which is expressed by a Boolean variable (1 if there is at least one attractor that has no basin, and 0 otherwise)
14	The proportion of states that belong to attractors
15	The variance of the steady-state distribution of the original network

state flows into a unique attractor cycle, and the set of states leading to that cycle is known as its *basin of attraction* (BOA). The *level* of a state is defined as the number of transitions required for that state to flow into an attractor cycle. It has been shown that the steady-state probabilities for attractors are dependent on the size and structure of BOAs [127]. After considering all these factors that characterize network behavior, we outline a list of network features for our classification task in Table 4.1. We restrict ourselves to 15 features using descriptive statistics to avoid the ‘curse of dimensionality’ [14].

Table 4.2 The general procedure of network reduction followed by network classification and feature selection



	Procedure
Step 1	Generate random Boolean networks with perturbation (BN$_p$s) of size n.
Step 2	When the model has reached its SSD, compute the canalizing power (CP) for every gene in the network.
Step 3	Filter out networks that do not have any genes with high CP.
Step 4	For each BN$_p$ that passed the filter on the previous step, and for $k = 1$ to m, repeat steps (a) through (c).
(a)	Apply the reduction mapping to remove a gene with the smallest canalizing power.
(b)	Re-index the remaining genes: the indices of all the genes following the removed gene are decreased by 1.
(c)	Obtain SSD $\tilde{\pi}$ for the reduced network and measure ρ_k and $D\left(Q_{n-k}, \tilde{Q}_{n-k}\right)$.
Step 5	Label the networks based on Defintion 1 with the selected parameter settings for $\theta_{avg}, \theta_{min}$ and θ_{dist}. $\begin{cases} Class\ 0\text{:}\ \ m\text{−}gene\ reducible\ networks \\ Class\ 1\text{:}\ \ networks\ that\ do\ not\ belong\ to\ Class\ 0 \end{cases}$
Step 6	Use a specific classification rule and find features that best discriminate the two classes of networks.

4.2.5 Simulation Procedures

To study canalization stability in network reduction and find network features that contribute to the classification of networks as either m-gene reducible or not, we carried out simulations on synthetic data. The entire procedure is given in Table 4.2. As the network SSD provides insight into long-run dynamics and allows one to compute the long-term influence of a gene on another gene, we use it in (3.2) and (3.11) for computing the

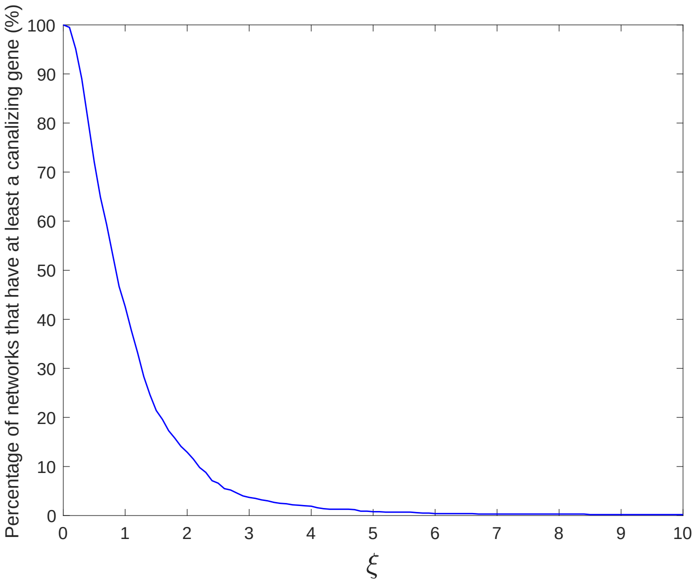


Figure 4.3 The percentage of networks that have at least one canalizing gene depending on the parameter ξ.

canalizing power of genes. Since networks with strong canalizing genes are the central focus of our study, we filter out networks that do not have any sets of genes with relatively high CP at step 3. We use the interquartile range (IQR) method for outlier detection and the presence of outliers indicates the existence of genes with comparatively high canalizing power. Thus, any gene with CP greater than Q3+ξ ×IQR is considered to be a gene with high canalizing power, where ξ is the parameter for the determination of a canalizing gene. Figure 4.3 shows the effects of the parameter ξ on the percentage of networks with a canalizing gene. We set ξ=1.5 as the standard IQR method and filter out networks that have no canalizing genes. As to the remaining networks, we remove the gene with the smallest canalizing power and obtain the reduced network by applying the reduction mapping described in Section 4.2.1. We repeat the steps 4(a) through 4(c) until m genes are removed and compute the CP vectors of the reduced network after each gene removal. Then, we compute the CP ranking correlation coefficients and distances between the CP vectors of the original and reduced networks. According to Definition 1 with the selected parameter settings for $\theta_{avg}, \theta_{min}$ and θ_{dist}, networks are labeled as either m-gene reducible or not. Finally, classification and feature selection are performed at step 6.

4.3 Results

4.3.1 Synthetic Data

In our simulation study, we randomly generated 10,000 Boolean Neteworks with perturbation (BN$_p$s) with n=12 genes and a perturbation parameter p=0.01. There are several computational challenges that have to be addressed to accomplish the goals of the study. Owing to the large number of repetitions in the simulation, the exponential increase of the network state space with the number of participating genes makes it computationally prohibitive to deal with networks having more than 12 genes. Owing to our desire to use a brute-force feature selection method to exhaustively evaluate all possible combinations of the input features listed in Table 4.1, we use 12-gene networks for the feature selection step of our procedure. The computation of the network's SSD usually includes construction of the transition probability matrix; however, matrix-based methods quickly become prohibitive for large sizes of networks, and therefore, we use Monte Carlo methods. We derive the SSD by running the network for a long time from a randomly selected initial state. To test the convergence to its SSD, we apply the Kolmogorov-Smirnov statistic.

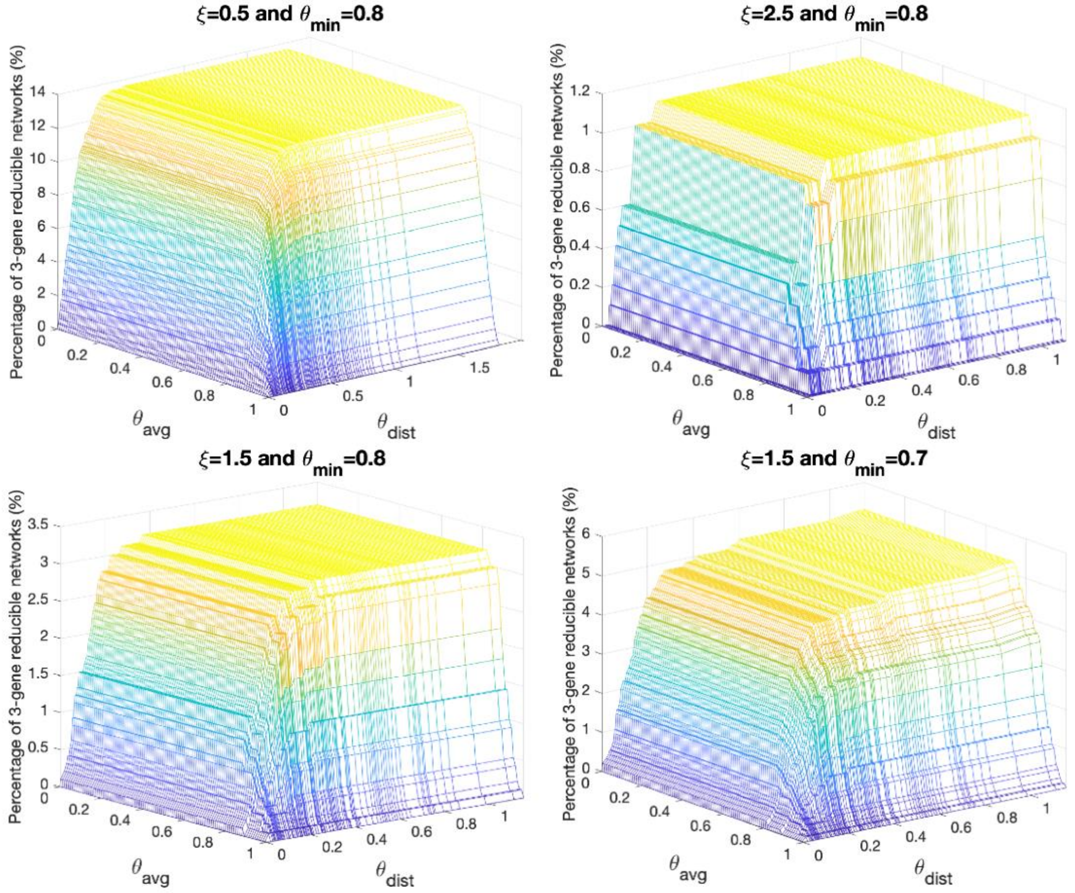


Figure 4.4 Effects of the parameters ξ, θ_{avg}, θ_{min}, and θ_{dist} on the percentage of 3-gene reducible networks out of 10,000 generated networks with n=12 genes.

Our first objective is to examine the proportion of reducible networks among the set of randomly generated BN$_p$ s. While consecutively removing 3 genes from the networks, we investigate the effects of the parameters ξ, θ_{avg}, θ_{min} and θ_{dist} on the percentage of 3-gene reducible networks as shown in Figure 4.4. Based on these results, we empirically tuned up the parameters ξ=1.5, θ_{avg}=0.9 and θ_{min}=0.8 for the rest of our simulation studies. The threshold θ_{dist} is set to 0.0938 which corresponds to the 20th percentile of the mean of the vector $\left(D\left(Q_{n-1},\tilde{Q}_{n-1}\right),\cdots,D\left(Q_{n-3},\tilde{Q}_{n-3}\right)\right)$ of 10,000

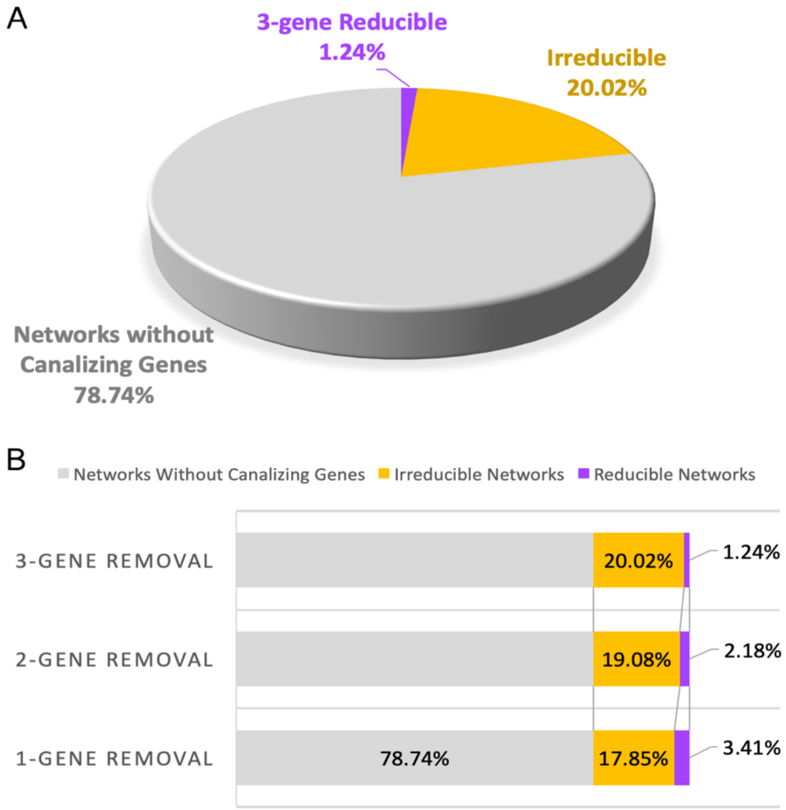


Figure 4.5 (A) The pie chart shows the percentage of networks without canalizing genes which are not considered for model reduction in our simulation. The remaining networks have been tested for reducibility and only 1.24% of the total networks are 3-gene reducible. (B) The bar graph represents that the proportion of reducible networks decreases as more genes are deleted.

networks when n=12. As shown in Figure 4.5(A), 2,126 out of 10,000 generated networks (21.26%) possess at least one canalizing gene and only 1.24% are 3-gene reducible based on Definition 1. Figure 4.5(B) shows that the percentage of reducible networks drops from 3.41% to 1.24% as more genes are removed from the model. These figures confirm that it is unusual to find networks with a canalizing gene, and it is even more rare to have a model that can be reduced without unduly altering canalizational properties of the participating genes.

After labeling networks based on Definition 1, we proceed to address classification in the particular setting of severely imbalanced data: approximately 5% of the networks belong to the class 0 (i.e. 3-gene reducible networks). Thus, one could simply classify all observations into the majority class and be correct 95% of the time. There are several available approaches to address the issue of such imbalanced data [128]. For example, one can resample the dataset to balance the skewed distribution [129]. Alternatively, other performance metrics such as precision and recall can be applied to evaluate the model [128]. We use both balanced and imbalanced data to build classifiers. Previous studies [130], [131] have shown that imbalanced data sets introduce a significant reduction in classifiers' performance because most of standard classifiers tacitly assume or expect balanced class distribution or equal misclassification cost. Rather than directly applying classification algorithms to severely imbalanced data, it is a common practice to preprocess the data to balance the distribution of the classes before the learning stage of each classifier [132]-[134]. Therefore, results obtained from the balanced data are presented in the main body of this chapter while the results for the case of imbalanced data are provided in the Appendix B. Importantly, we adopt the downsampling method for resampling thus, randomly selecting samples without replacement from the majority class [130].

Consider a binary classifier $\Psi: \mathbb{R}^{\lambda} \to \{0, 1\}$ which assigns a network on λ-dimensional feature space $\mathbb{R}^{\lambda}$ to either class 0 or 1. Let class 0 (i.e. m-gene reducible networks) indicate the positive class and class 1 (i.e. not m-gene reducible networks) be the negative class. Given a network G, there are four possibilities when comparing the predicted class $\Psi(G)$ to its true class y: true-positive $y = 0$ and $\Psi(G) = 0$; false-negative $y = 0$ and $\Psi(G) = 1$; false-positive $y = 1$ and $\Psi(G) = 0$; true-negative $y = 1$ and $\Psi(G) = 1$. We denote by TP and TN the number of true-positive and true-negative networks, respectively. Similarly, FP and FN denote the number of false-positive and false-negative networks, respectively. Using these notations, we define the *positive predictive value* of a classifier by $TP/(TP + FP)$, the *negative predictive value* by $TN/(TN + FN)$, the *sensitivity* by $TP/(TP + FN)$, and the *specificity* by $TN/(TN + FP)$.

We evaluate the performance of the features listed in Table 4.1 with six different classification rules: linear discriminant analysis (LDA), 3-nearest neighbors (3NN), support vector machine (SVM), decision tree (DT), Naïve Bayes (NB), and a neural network (NN) with 10 hidden layers. A training set of size 10,000 is used for classifier construction, and an independent test set of size 3,000 is used for evaluation. We exhaustively perform all of the λ-feature (λ=1,⋯,15) classifications and Figure 4.6 shows the respective performance metrics. Given a specific classification rule, each data point in Figure 4.6 corresponds to the best performance metric among all of the results from λ-tuples of features. For example, sensitivity at $\lambda = 1$ in Figure 4.6 represents the highest of all 15 sensitivities when the classification is based on a single feature. Figure 4.6 shows that a nonlinear support vector machine with a radial basis function (RBF-SVM) achieves the lowest test error 3.48% across all of the considered classifiers when using the following couple of features: the variance of the steady-state probabilities of attractor states and the variance of the SSD of the original network. In general, the decision tree performs better

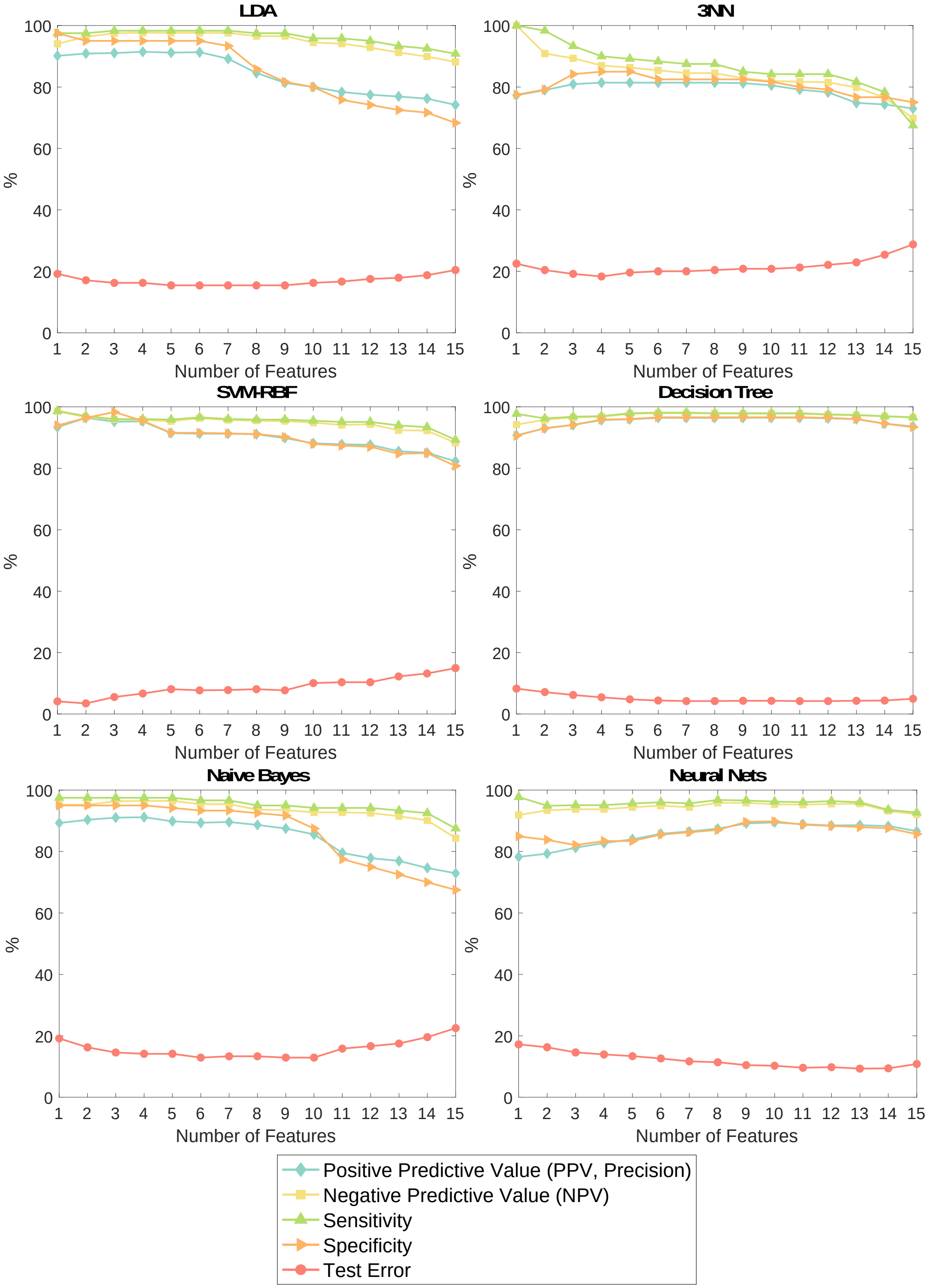


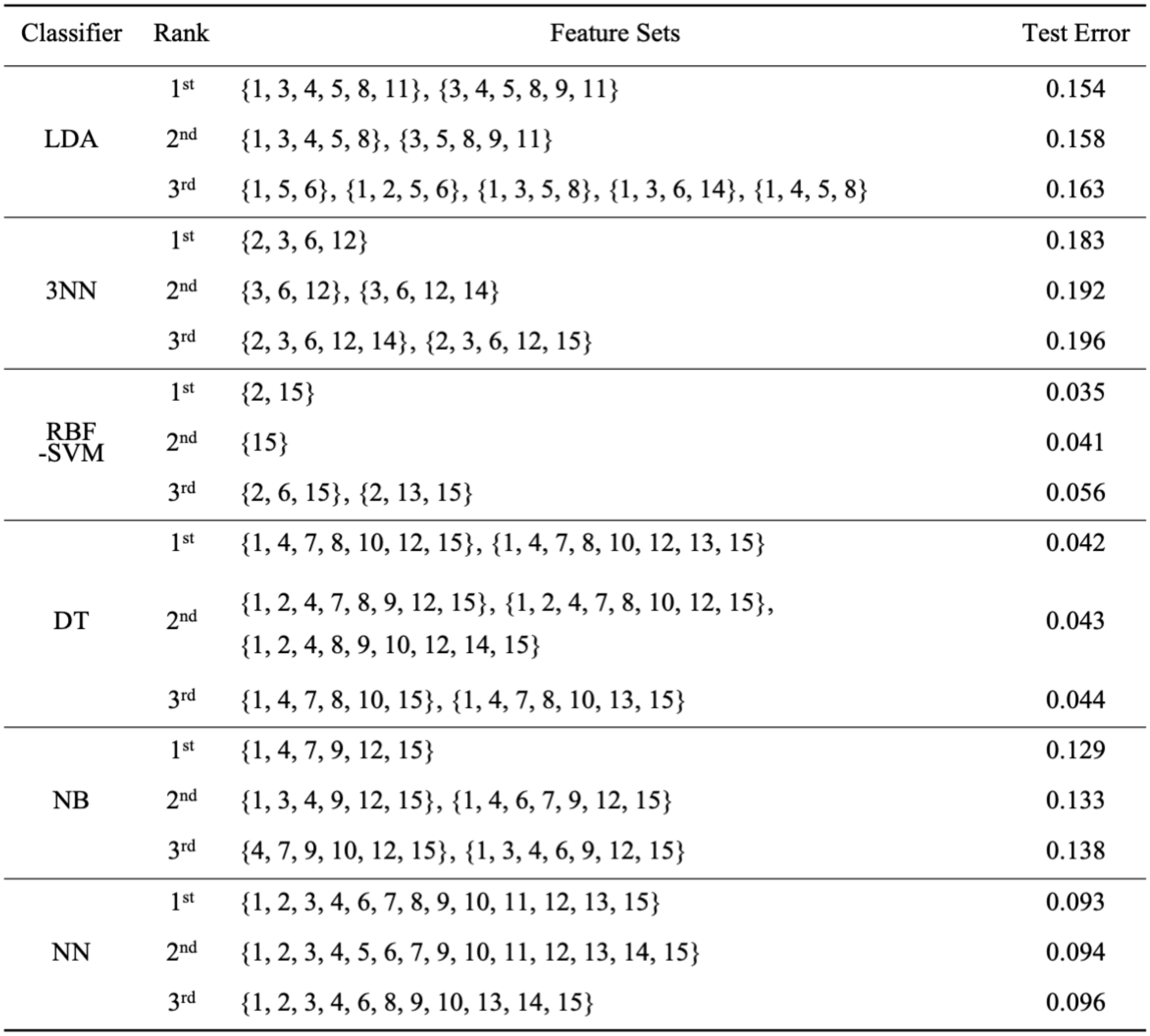
Figure 4.6 Based on 15 features listed in Table 4.1, all of the λ-feature (λ =1,⋯,15) classifications are performed using six different classification rules: LDA, 3NN, RBF-SVM, DT, NB and NN. Each data point on the graphs represents the best value among the respective performance measures.

than the other classification rules while 3NN shows the worst performance. Figure 4.6 also shows that in most cases the designed classifiers have high sensitivity with moderate specificity which indicates that the classifiers are capable of detecting reducible networks; however, there is also a presence of false positives. If one desires to decrease the number of false positives, one can increase the values of the parameters θ_{avg} and θ_{min} while simultaneously decreasing the value of θ_{dist}.

After performing all of the λ-feature (λ=1,⋯,15) classifications, we sort feature sets according to their test error and list the top 3 best performing feature sets with respect to that error. Feature sets that achieve the 3 lowest misclassification errors for each classification rule and the corresponding test errors are listed in Table 4.3. Note that the two top ranked LDA feature sets share the features: the average connectivity, the average bias, the maximum attractor cycle length, the average of the distances from each state to its corresponding attractor, and the average size of basins of attraction. Table 4.3 also shows that feature sets resulting in the lowest test error for decision tree and Naïve Bayes are similar: the sum of the steady-state probabilities of attractor states, the average bias, the maximum of the distances from each state to its corresponding attractor, the total number of attractors, and the variance of the steady-state distribution of the original network.

The normalized empirical histogram of the appearance of each individual feature in the top 3 performing feature sets is shown in Figure 4.7. It shows that the variance of the steady-state probabilities of attractor states and the number of singleton attractors appear in most of six classification rules' top 3 performing feature sets. One can also see

Table 4.3 Feature sets that achieve the 3 lowest misclassification errors for each classification rule and their respective test error estimates. Note that numerical identifiers are provided in Table 4.1.



Classifier	Rank	Feature Sets	Test Error
LDA	1st	{1, 3, 4, 5, 8, 11}, {3, 4, 5, 8, 9, 11}	0.154
	2nd	{1, 3, 4, 5, 8}, {3, 5, 8, 9, 11}	0.158
	3rd	{1, 5, 6}, {1, 2, 5, 6}, {1, 3, 5, 8}, {1, 3, 6, 14}, {1, 4, 5, 8}	0.163
3NN	1st	{2, 3, 6, 12}	0.183
	2nd	{3, 6, 12}, {3, 6, 12, 14}	0.192
	3rd	{2, 3, 6, 12, 14}, {2, 3, 6, 12, 15}	0.196
RBF -SVM	1st	{2, 15}	0.035
	2nd	{15}	0.041
	3rd	{2, 6, 15}, {2, 13, 15}	0.056
DT	1st	{1, 4, 7, 8, 10, 12, 15}, {1, 4, 7, 8, 10, 12, 13, 15}	0.042
	2nd	{1, 2, 4, 7, 8, 9, 12, 15}, {1, 2, 4, 7, 8, 10, 12, 15}, {1, 2, 4, 8, 9, 10, 12, 14, 15}	0.043
	3rd	{1, 4, 7, 8, 10, 15}, {1, 4, 7, 8, 10, 13, 15}	0.044
NB	1st	{1, 4, 7, 9, 12, 15}	0.129
	2nd	{1, 3, 4, 9, 12, 15}, {1, 4, 6, 7, 9, 12, 15}	0.133
	3rd	{4, 7, 9, 10, 12, 15}, {1, 3, 4, 6, 9, 12, 15}	0.138
NN	1st	{1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 15}	0.093
	2nd	{1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15}	0.094
	3rd	{1, 2, 3, 4, 6, 8, 9, 10, 13, 14, 15}	0.096

that the variance of the SSD of the original network is the most frequently appearing feature.

Finally, we train classifiers using 3,000 networks with 13 genes using the top performing feature sets for each classification rule given in Table 4.3 and test the designed classifiers on 1,000 independently generated networks. Table 4.4 summarizes the

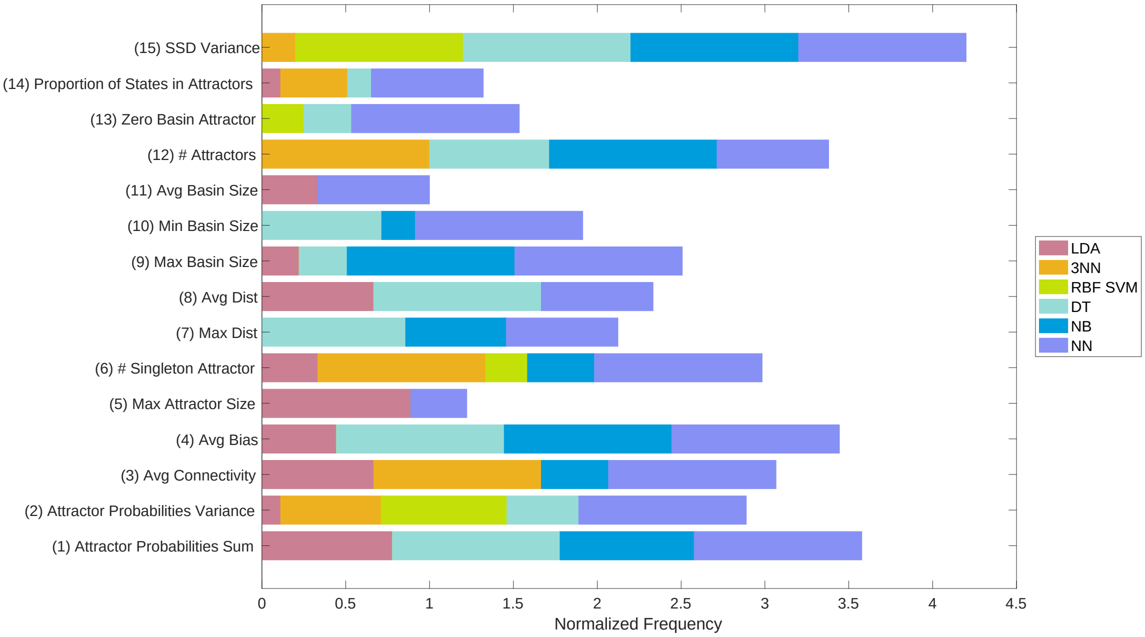


Figure 4.7 The stacked bar represents how often each feature appears in the top 3 performing feature sets for each classification rule. Each segment of the stacked bar represents the normalized count which is the frequency of individual feature's appearance in Table 4.3 divided by the total number of feature sets shown in Table 4.3 for each classification rule. For example, there are 9 feature sets in Table 4.3 for LDA, and therefore, the occurences of each feature in Table 4.3 for LDA are divided by 9.

Table 4.4 Test errors of classifiers on 13-gene networks when the best performing feature sets given in Table 4.3 are used.

Classifier	Feature Sets	Test Error
LDA	{1, 3, 4, 5, 8, 11}	0.189
	{3, 4, 5, 8, 9, 11}	0.171
3NN	{2, 3, 6, 12}	0.149
RBF-SVM	{2, 15}	0.027
DT	{1, 4, 7, 8, 10, 12, 15}	0.081
	{1, 4, 7, 8, 10, 12, 13, 15}	0.081
NB	{1, 4, 7, 9, 12, 15}	0.169
NN	{1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 15}	0.059

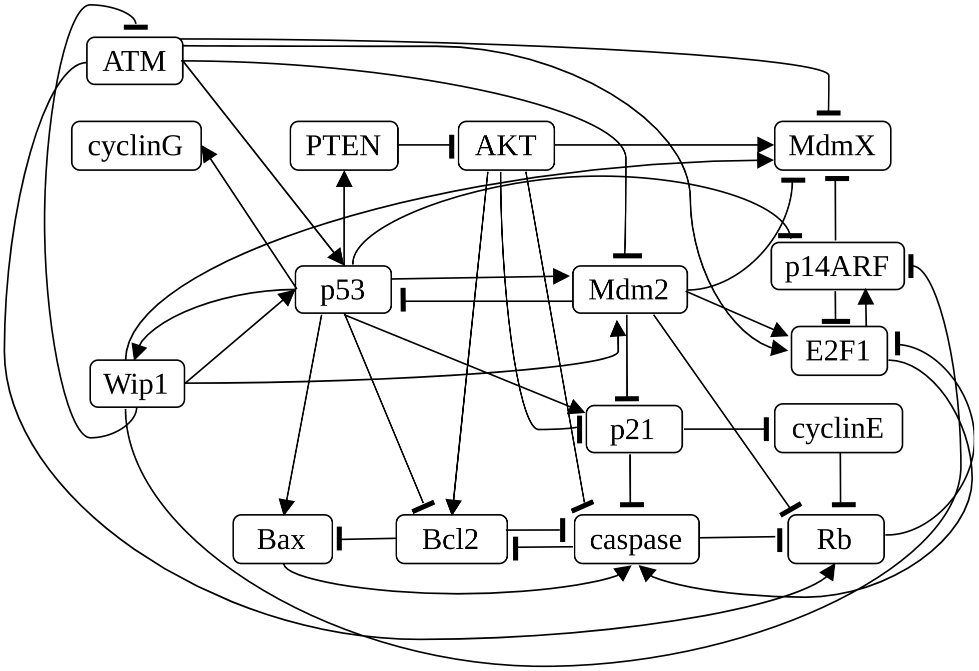


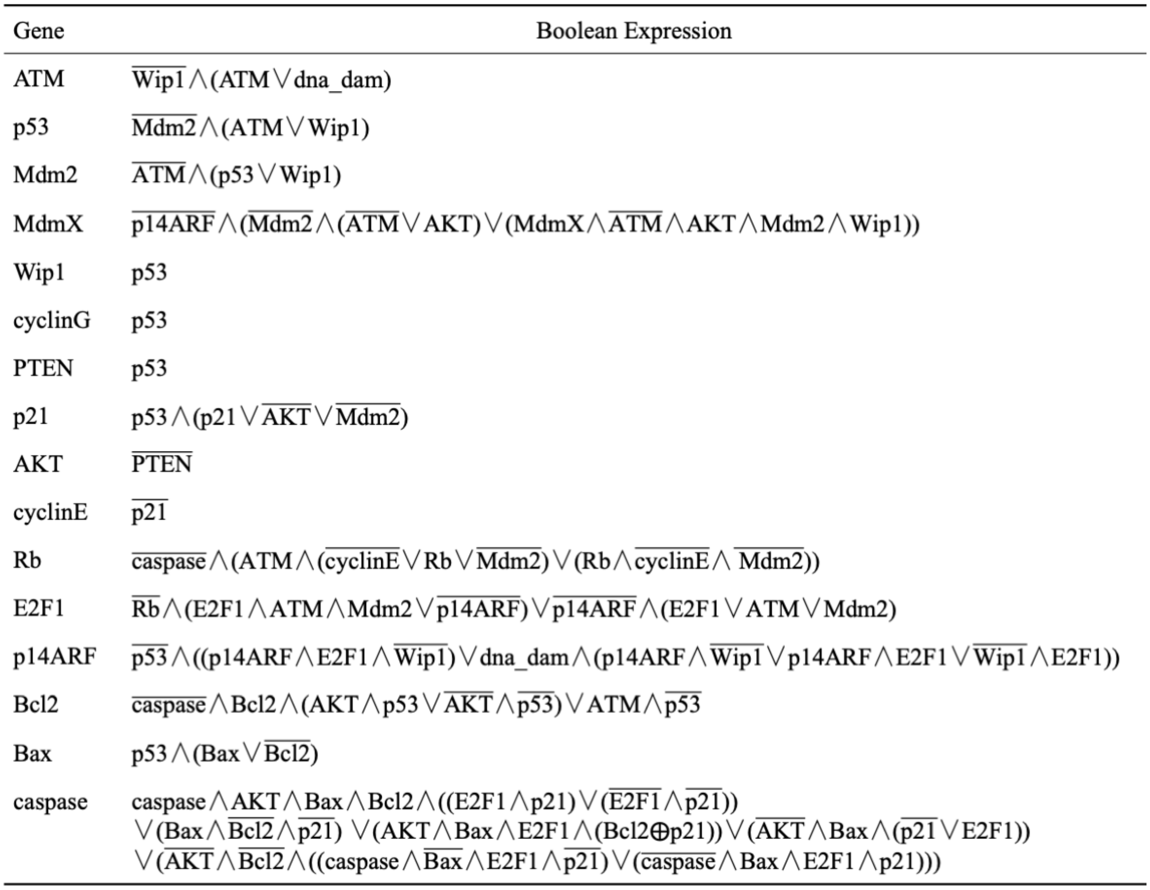
Figure 4.8 The p53 regulatory network adapted from [8]. Blunt arrows represent inhibition while normal arrows represent activation.

respective test errors for the six classification rules considered in our study. The result demonstrates that RBF-SVM achieves the lowest error among all of the classifiers when a pair of features {2, 15} is used. It is also shown that test errors of decision tree and neural nets are less than 0.1 while LDA, 3NN and NB result in high test error rates of 0.14 or greater.

4.3.2 p53 Network

We now apply our algorithm to a 16-gene p53 regulatory network presented in [8]. M. Choi et al. [8] modeled the signaling response of DNA damage in a p53 network using a Boolean network with a set of state transition rules defined on the basis of biological evidence. The 16-gene network includes the following genes: ATM, p53, Mdm2, MdmX,

Table 4.5 Boolean functions for the genes in the p53 regulatory network, where the symbols $\vee$, $\wedge$ and $\oplus$ denote the Boolean disjunction, conjunction and exclusive-OR, respectively. The DNA damage input is denoted by 'dna_dam'.



Gene	Boolean Expression
ATM	$\overline{\mathrm{Wip1}}\wedge(\mathrm{ATM}\vee\mathrm{dna\_dam})$
p53	$\overline{\mathrm{Mdm2}}\wedge(\mathrm{ATM}\vee\mathrm{Wip1})$
Mdm2	$\overline{\mathrm{ATM}}\wedge(\mathrm{p53}\vee\mathrm{Wip1})$
MdmX	$\overline{\mathrm{p14ARF}}\wedge(\overline{\mathrm{Mdm2}}\wedge(\overline{\mathrm{ATM}}\vee\mathrm{AKT})\vee(\mathrm{MdmX}\wedge\overline{\mathrm{ATM}}\wedge\mathrm{AKT}\wedge\mathrm{Mdm2}\wedge\mathrm{Wip1}))$
Wip1	p53
cyclinG	p53
PTEN	p53
p21	$\mathrm{p53}\wedge(\mathrm{p21}\vee\overline{\mathrm{AKT}}\vee\overline{\mathrm{Mdm2}})$
AKT	$\overline{\mathrm{PTEN}}$
cyclinE	$\overline{\mathrm{p21}}$
Rb	$\overline{\mathrm{caspase}}\wedge(\mathrm{ATM}\wedge(\overline{\mathrm{cyclinE}}\vee\mathrm{Rb}\vee\overline{\mathrm{Mdm2}})\vee(\mathrm{Rb}\wedge\overline{\mathrm{cyclinE}}\wedge\overline{\mathrm{Mdm2}}))$
E2F1	$\overline{\mathrm{Rb}}\wedge(\mathrm{E2F1}\wedge\mathrm{ATM}\wedge\mathrm{Mdm2}\vee\overline{\mathrm{p14ARF}})\vee\overline{\mathrm{p14ARF}}\wedge(\mathrm{E2F1}\vee\mathrm{ATM}\vee\mathrm{Mdm2})$
p14ARF	$\overline{\mathrm{p53}}\wedge((\mathrm{p14ARF}\wedge\mathrm{E2F1}\wedge\overline{\mathrm{Wip1}})\vee\mathrm{dna\_dam}\wedge(\mathrm{p14ARF}\wedge\overline{\mathrm{Wip1}}\vee\mathrm{p14ARF}\wedge\mathrm{E2F1}\vee\overline{\mathrm{Wip1}}\wedge\mathrm{E2F1}))$
Bcl2	$\overline{\mathrm{caspase}}\wedge\mathrm{Bcl2}\wedge(\mathrm{AKT}\wedge\mathrm{p53}\vee\overline{\mathrm{AKT}}\wedge\overline{\mathrm{p53}})\vee\mathrm{ATM}\wedge\overline{\mathrm{p53}}$
Bax	$\mathrm{p53}\wedge(\mathrm{Bax}\vee\overline{\mathrm{Bcl2}})$
caspase	$\mathrm{caspase}\wedge\mathrm{AKT}\wedge\mathrm{Bax}\wedge\mathrm{Bcl2}\wedge((\mathrm{E2F1}\wedge\mathrm{p21})\vee(\overline{\mathrm{E2F1}}\wedge\overline{\mathrm{p21}}))$ $\vee(\mathrm{Bax}\wedge\overline{\mathrm{Bcl2}}\wedge\overline{\mathrm{p21}})\vee(\mathrm{AKT}\wedge\mathrm{Bax}\wedge\mathrm{E2F1}\wedge(\mathrm{Bcl2}\oplus\mathrm{p21}))\vee(\overline{\mathrm{AKT}}\wedge\mathrm{Bax}\wedge(\overline{\mathrm{p21}}\vee\mathrm{E2F1}))$ $\vee(\overline{\mathrm{AKT}}\wedge\overline{\mathrm{Bcl2}}\wedge((\mathrm{caspase}\wedge\overline{\mathrm{Bax}}\wedge\mathrm{E2F1}\wedge\overline{\mathrm{p21}})\vee(\overline{\mathrm{caspase}}\wedge\mathrm{Bax}\wedge\mathrm{E2F1}\wedge\mathrm{p21})))$

Wip1, cyclinG, PTEN, p21, AKT, cyclinE, Rb, E2F1, p14ARF, Bcl2, Bax, and caspase. The network consists of 160 negative and 218 positive feedback loops and several crosstalk pathways, such as pathways involved in survival signaling and the cell cycle regulatory pathway involving retinoblastoma (Rb). The wiring diagram of the model is given in Figure 4.8 and a logic table that determines the response of the output nodes for a given set of inputs is shown in Table 4.5.

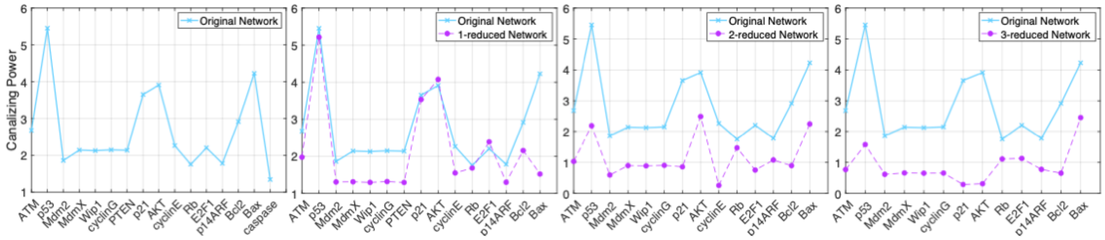


Figure 4.9 Canalizing power of genes in the 16-gene p53 regulatory network. A gene with the smallest canalizing power is removed at each consecutive step (caspase, PTEN and cyclinE).

We use a perturbation parameter p=0.01 and the SSD is estimated by running the network for a long time from a randomly selected initial state. The Kolmogorov-Smirnov test is used to decide if the network has reached its steady state. We compute the CP of 16 genes in the network and observe that p53 has the largest canalizing power 5.454. It agrees with the known biological fact that p53 is a tumor suppressor gene with strong canalizing ability. It is kept at a low level/dephosphorylated in unstressed cells but becomes significantly activated/phosphorylated in response to DNA damage [103]. Our computations show that when the parameter ξ is less than 1.68, the network possesses a canalizing gene which makes the experimental result consistent with the known fact.

Next, we assess if the model could be reduced while preserving its CP vector. At each iterative step, the gene having the smallest CP is removed. For 3-gene reduction, caspase, PTEN and cyclinE are removed consecutively. The CP ranking correlation coefficients computed at each gene removal are ρ_1=0.721, ρ_2=0.314 and ρ_3=0.093 which gives $\overline{\boldsymbol{\rho}}$=0.376 and ρ_{min}=0.093. The distances between CP vectors of the original and k-reduced networks are 0.371, 0.343 and 0.618 for k=1, 2 and 3, respectively. Based on Definition 1, this network is not k-gene, k=1, 2, 3 reducible with the parameter setting

θ_{avg}=0.9 and θ_{min}=0.8. It could be considered 1-gene reducible with lower correlation coefficient values of θ_{min} < 0.721, but it requires far less stringent value for distance parameter θ_{dist} > 0.371. The network is 2-gene reducible when θ_{avg}< 0.518, θ_{min} < 0.314 and θ_{dist} > 0.371, but a correlation coefficient less than 0.5 is not considered to indicate a strong association between CP vectors of the original and reduced network. As shown in Figure 4.9, the network does not maintain the existing distribution of canalizing power among genes and even with less stringent thresholds of θ_{avg} > 0.376, θ_{min} > 0.093 and θ_{dist} < 0.618, this network is classified to class 1, i.e., not 3-gene reducible.

4.4 Conclusion

Canalization is the tendency of a biological process to follow particular trajectories despite external or internal perturbation, and it plays a pivotal role for phenotypic robustness [60]-[62]. Therefore, the preservation of the canalizational properties of the genes participating in a network model of gene regulation should be one of the main objectives in network reduction. Our empirical study suggests that it is rare to find networks with a canalizing gene. Moreover, network reduction could easily destroy the canalizational properties of the genes in the original network. In this article, we show that removing genes with weak canalizing power seems to be a good heuristic for network compression. Furthermore, our study indicates that there are two major classes of BN$_p$ model which are determined by the degree of preservation of the CP vectors after applying the reduction mapping. Based on this observation, we introduce the definition of m-gene reducible networks using the CP ranking correlation coefficients and distances between the CP vectors of the original and reduced networks. Subsequently, we proceed with the problem of selecting the relevant network features which help to discriminate m-gene reducible from other networks.

Our comprehensive simulation study leads to several important observations: (i) RBF-SVM achieves the lowest error among the six classification rules considered. The couple of features that provides the best separation between the two classes of networks is composed of the variance of attractors' stationary mass and the variance of the steady-state probabilities of the network. The SSD reflects the long-run behavior of a given network, and stationary masses of attractors reflect the structure of BOAs. Importantly, many studies [122], [135] have utilized SSD and stationary masses of critical states in network reduction and intervention. Our result also confirms that the SSD and the stationary masses of attractors encode important properties of the network in the context of network compression with the objective to preserve canalization properties of the genes; (ii) Neural nets achieve the lowest test error when employing nearly all of the features. Although neural network can have test error rates as low as 0.0586, it is very often the case that a neural net requires thousands of labeled samples [136]. Therefore, it might be impractical to use this classification rule. Furthermore, the typical small sample size in experiments related to gene regulation could lead to overfitting when it is coupled with the requirement for a large number of features as in the case of NN classification; (iii) Our study shows that the number of singleton attractors in the networks is an important feature as it appears in the top 3 performing feature sets across all of the six classification rules; (iv) Decision trees give the second lowest test error in 12-gene network classification while NN is the second best performer in 13-gene network classification. This implies that the total number of genes in the network or the proportion of the removed genes influence the network reducibility and classification performance. Our results also suggest that the classifier that performs the best in identifying 1-gene reducible networks might be different from the best classifier in 2-gene reducibility classification. Thus, the attributes that determine the network reducibility could change depending on how many genes are already deleted from the model.

It is practically impossible to have complete knowledge about the structure of the network. Therefore, one may not be able to exploit the best combination of features and classification rules as outlined by our simulation study. To facilitate the practical selection of combinations of feature sets and classification rules, we present not only the best result but also all possible feature sets and corresponding error rates in the following website: https://sites.google.com/view/canalizingpower/. Thus, our simulation study can serve as a guideline in selecting a suitable classifier depending on the feasibility and availability of features, as it also provides an estimate for the performance of the selected feature sets and classification rules.

5. CONCLUSION

In this dissertation, we have focused on the problem of identifying important genes that exist in high-dimensional data spaces. In Chapter 2, we performed a model-based study to examine the effectiveness of reporting lists of gene sets using RNA-Seq data. We compared the performance of ranking of feature sets derived from a model of RNA-Seq data with that of a multivariate normal model of gene concentrations. The results of our study show that the performance of feature selection using RNA-Seq data deteriorates which is most likely due to a nonlinear transformation of the actual gene or RNA concentrations by next-generation sequencing technologies. We also examined the effects of different model parameters and error estimators on the ranked lists of features. The results demonstrate that the general trends of the parameter effects on the ranking power of the underlying gene concentrations are preserved in the RNA-Seq data, whereas the power of finding a good feature set becomes weaker when gene concentrations are transformed by the sequencing machine.

In Chapter 3, we developed a quantitative framework that allows to compute the canalizing power of genes in the context of Boolean Networks with perturbation (BN$_p$s). This framework borrows tools from the Pattern Recognition theory and uses the CoD to capture the capacity of the canalizing genes. The canalizing power of a gene is quantitatively characterized by two terms: regulation power and incapacitating power. Following this, the CP concept was illustrated with examples to verify how CP can be used to characterize the ability of canalizing genes.

In Chapter 4, we study the problem of reducing BN$_p$s by deleting genes with the smallest canalizing power consecutively and evaluated the stability of canalizing power. Our systematic empirical study demonstrates that there are two classes of networks with respect to the preservation of canalizing power of the genes when such reduction mapping is applied. Subsequently, we proceeded with the problem of selecting the relevant network features that allow for discriminating these two different classes of networks.

In our opinion, there are several potential avenues for further development and application of the definition of the canalizing power: (i) One could extend the study by investigating the relationship between network attributes and the magnitude of the canalizing power of genes. There might be some factors that potentially determine the canalizing power of genes such as Boolean rules, connectivity of the corresponding gene, attractor structures of the network, existence of feedback loops, etc; (ii) To our knowledge, there is no explicit representation of canalizing genes in Boolean expressions. The Boolean disjunction and conjuction can represent canalizing functions in which at least one of the input variables, called a canalizing variable, is able to determine the function ouput regardless of the values of the other variables. However, the canalizing function does not explicitly show which one is a canalizing variable. Mathematically, any variable associated with Boolean disjuction or conjunction operator could be a canalizing variable. However, in biology, the main driver is preordained and sometimes it has been even decided who the next will be when the main driver is deactivated. Therefore, in order to reflect this hierarchical architecture inherently embedded in gene regulatory networks, there is a need to explicitly represent commanders and their respective order.

REFERENCES

[1] I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, M. Raffeld, Z. Yakhini, A. Ben-Dor, E. Dougherty, J. Kononen, L. Bubendorf, W. Fehrle, S. Pittaluga, S. Gruvberger, N. Loman, O. Johannsson, H. Olsson, B. Wilfond, G. Sauter, O.-P. Kallioniemi, Å. Borg, and J. Trent, "Gene-Expression Profiles in Hereditary Breast Cancer," *N Engl J Med*, vol. 344, no. 8, pp. 539–548, Feb. 2001.

[2] M. J. van de Vijver, Y. D. He, L. J. van 't Veer, H. Dai, A. A. M. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards, "A Gene-Expression Signature as a Predictor of Survival in Breast Cancer," *N Engl J Med*, vol. 347, no. 25, pp. 1999–2009, Dec. 2002.

[3] A. Barrier, A. Lemoine, P.-Y. Boelle, C. Tse, D. Brault, F. Chiappini, J. Breittschneider, F. Lacaine, S. Houry, M. Huguier, M. J. Van der Laan, T. Speed, B. Debuire, A. Flahault, and S. Dudoit, "Colon cancer prognosis prediction by gene expression profiling," *Oncogene*, vol. 24, no. 40, pp. 6155–6164, Sep. 2005.

[4] J. S. Wei, B. T. Greer, F. Westermann, S. M. Steinberg, C.-G. Son, Q.-R. Chen, C. C. Whiteford, S. Bilke, A. L. Krasnoselsky, N. Cenacchi, D. Catchpoole, F. Berthold, M. Schwab, and J. Khan, "Prediction of Clinical Outcome Using Gene Expression Profiling and Artificial Neural Networks for Patients with Neuroblastoma," *Cancer Res*, vol. 64, no. 19, pp. 6883–6891, Oct. 2004.

[5] M. Reedijk, S. Odorcic, L. Chang, H. Zhang, N. Miller, D. R. McCready, G. Lockwood, and S. E. Egan, "High-level Coexpression of JAG1 and NOTCH1 Is Observed in Human Breast Cancer and Is Associated with Poor Overall Survival," *Cancer Res*, vol. 65, no. 18, pp. 8530–8537, Sep. 2005.

[6] M. S. Poptsova, I. A. Il'icheva, D. Y. Nechipurenko, L. A. Panchenko, M. V. Khodikov, N. Y. Oparina, R. V. Polozov, Y. D. Nechipurenko, and S. L. Grokhovsky, "Non-random DNA fragmentation in next-generation sequencing," *srep*, vol. 4, no. 1, p. 4532, Mar. 2014.

[7] K. R. Kukurba and S. B. Montgomery, "RNA Sequencing and Analysis," *Cold Spring Harb Protoc*, vol. 2015, no. 11, Nov. 2015.

[8] M. Choi, J. Shi, S. H. Jung, X. Chen, and K. H. Cho, "Attractor Landscape

Analysis Reveals Feedback Loops in the p53 Network That Control the Cellular Response to DNA Damage," *Sci. Signal.*, vol. 5, no. 251, pp. ra83–ra83, Nov. 2012.

[9] P. L. Auer and R. W. Doerge, "Statistical Design and Analysis of RNA Sequencing Data," *Genetics*, vol. 185, no. 2, pp. 405–416, Jun. 2010.

[10] N. Goossens, S. Nakagawa, X. Sun, and Y. Hoshida, "Cancer biomarker discovery and validation," *Translational cancer research*, vol. 4, no. 3, pp. 256-269, Jun. 2015.

[11] B. Wooden, N. Goossens, Y. Hoshida, and S. L. Friedman, "Using Big Data to Discover Diagnostics and Therapeutics for Gastrointestinal and Liver Diseases," *Gastroenterology*, vol. 152, no. 1, pp. 53–67.e3, Jan. 2017.

[12] N. Ghaffari, M. R. Yousefi, C. D. Johnson, I. Ivanov, and E. R. Dougherty, "Modeling the next generation sequencing sample processing pipeline for the purposes of classification," *BMC Bioinformatics*, vol. 14, no. 1, p. 307, Oct. 2013.

[13] S. R. Chowdhuri, S. Roy, S. E. Monaco, M. J. Routbort, and L. Pantanowitz, "Big data from small samples: Informatics of next-generation sequencing in cytopathology," *Cancer Cytopathology*, vol. 125, no. 4, pp. 236–244, Apr. 2017.

[14] R. Ernest and E. Bellman, *Adaptive control processes: a guided tour*. Princeton University Press, 1961.

[15] C. Sima and E. R. Dougherty, "The peaking phenomenon in the presence of feature-selection," *Pattern Recognition Letters*, vol. 29, no. 11, pp. 1667–1674, Aug. 2008.

[16] Z. He and W. Yu, "Stable Feature Selection for Biomarker Discovery," *arXiv*, Jan. 2010.

[17] D. Grissa, M. Pétéra, M. Brandolini, A. Napoli, B. Comte, and E. Pujos-Guillot, "Feature Selection Methods for Early Predictive Biomarker Discovery Using Untargeted Metabolomic Data," *Front. Mol. Biosci.*, vol. 3, p. 1, Jul. 2016.

[18] C. Christin, H. C. J. Hoefsloot, A. K. Smilde, B. Hoekman, F. Suits, R. Bischoff, and P. Horvatovich, "A Critical Assessment of Feature Selection Methods for Biomarker Discovery in Clinical Proteomics," *Molecular & Cellular Proteomics*, vol. 12, no. 1, pp. 263–276, Jan. 2013.

[19] I. H. Lee, G. H. Lushington, and M. Visvanathan, “A filter-based feature selection approach for identifying potential biomarkers for lung cancer,” *J Clin Bioinform*, vol. 1, no. 1, pp. 1–8, Dec. 2011.

[20] E. Robotti, “Biomarkers Discovery through Multivariate Statistical Methods: A Review of Recently Developed Methods and Applications in Proteomics,” *Journal of Proteomics & Bioinformatics*, vol. 3, Jan. 2013.

[21] J. de Winter, “Using the Student’s t-test with extremely small sample sizes,” *Practical Assessment, Research & Evaluation*, vol. 18, no.10, Aug. 2013.

[22] M. A. Hayat, *Methods of Cancer Diagnosis, Therapy and Prognosis: Breast Carcinoma*. Springer Netherlands, 2008.

[23] S. Stoppelkamp, K. Veseli, K. Stang, C. Schlensak, H. P. Wendel, and T. Walker, “Identification of Predictive Early Biomarkers for Sterile-SIRS after Cardiovascular Surgery,” *PLOS ONE*, vol. 10, no. 8, p. e0135527, Aug. 2015.

[24] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data,” *Bioinformatics*, vol. 26, no. 1, pp. 139-140, 2010.

[25] M. I. Love, W. Huber, and S. Anders, “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2,” *Genome Biology*, vol. 15, no. 12, pp. 1–21, Dec. 2014.

[26] M. Kanehisa, S. Goto, “KEGG: kyoto encyclopedia of genes and genomes,” *Nucleic Acids Res*, vol. 28, no. 1, pp. 27-30, Jan. 2000.

[27] P. C. Phillips, “Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems,” *Nature Reviews Genetics 2008 9:11*, vol. 9, no. 11, pp. 855–867, Nov. 2008.

[28] M. Vidal, M. E. Cusick, and A. L. Barabási, “Interactome Networks and Human Disease,” *Cell*, vol. 144, no. 6, pp. 986–998, Mar. 2011.

[29] J. G. Dy and C. E. Brodley, “Feature Selection for Unsupervised Learning,” *Journal of Machine Learning Research*, vol. 5, no. Aug, pp. 845–889, 2004.

[30] M. Yousef, N. Najami, L. Abdallah, and W. Khalifa, “Computational Approaches for Biomarker Discovery,” *Journal of Intelligent Learning Systems and Applications*, vol. 6, pp. 153–161, Oct. 2014.

[31] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. Prentice Hall PTR, 1988.

[32] S. Loscalzo, L. Yu, and C. Ding, “Consensus Group Stable Feature Selection,” *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2009, pp. 567–576.

[33] S. Y. Kim, “Effects of sample size on robustness and prediction accuracy of a prognostic gene signature,” *BMC Bioinformatics*, vol. 10, no. 1, pp. 1999–10, May 2009.

[34] L. Ein-Dor, O. Zuk, and E. Domany, “Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer.,” *Proceedings of the National Academy of Sciences*, vol. 103, no. 15, pp. 5923–5928, Apr. 2006.

[35] C. Sima, U. Braga-Neto, and E. R. Dougherty, “Superior feature-set ranking for small samples using bolstered error estimation.,” *Bioinformatics*, vol. 21, no. 7, pp. 1046–1054, Apr. 2005.

[36] U. M. Braga-Neto and E. R. Dougherty, “Is cross-validation valid for small-sample microarray classification?,” *Bioinformatics*, vol. 20, no. 3, pp. 374–380, Feb. 2004.

[37] C. Sima and E. R. Dougherty, “What should be expected from feature selection in small-sample settings.,” *Bioinformatics*, vol. 22, no. 19, pp. 2430–2436, Oct. 2006.

[38] E. R. Dougherty, *Genomic Signal Processing and Statistics*. Hindawi Publishing Corporation, 2005.

[39] C. Zhao, M. L. Bittner, R. S. Chapkin, and E. R. Dougherty, “Characterization of the Effectiveness of Reporting Lists of Small Feature Sets Relative to the Accuracy of the Prior Biological Knowledge,” *Cancer Informatics*, vol. 9, p. 49, 2010.

[40] S. Behjati and P. S. Tarpey, “What is next generation sequencing?,” *Arch Dis Child Educ Pract Ed*, vol. 98, no. 6, pp. 236–238, Dec. 2013.

[41] M. Jain, “Next-generation sequencing technologies for gene expression profiling in plants,” *bfg*, vol. 11, no. 1, pp. 63–70, Dec. 2011.

[42] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton,

C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. K. Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. J. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. D. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. A. Bridgham, R. C. Brown, A. A. Brown, D. H. Buermann, A. A. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, M. C. E. Catenazzi, S. Chang, R. N. Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. F. Fajardo, W. S. Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, K. Hoschler, S. Hurwitz, D. V. Ivanov, M. Q. Johnson, T. James, T. A. H. Jones, G.-D. Kang, T. H. Kerelska, A. D. Kersey, I. Khrebtukova, A. P. Kindwall, Z. Kingsbury, P. I. Kokko-Gonzales, A. Kumar, M. A. Laurent, C. T. Lawley, S. E. Lee, X. Lee, A. K. Liao, J. A. Loch, M. Lok, S. Luo, R. M. Mammen, J. W. Martin, P. G. McCauley, P. McNitt, P. Mehta, K. W. Moon, J. W. Mullens, T. Newington, Z. Ning, B. L. Ng, S. M. Novo, M. J. O'Neill, M. A. Osborne, A. Osnowski, O. Ostadan, L. L. Paraschos, L. Pickering, A. C. Pike, A. C. Pike, D. C. Pinkard, D. P. Pliskin, J. Podhasky, V. J. Quijano, C. Raczy, V. H. Rae, S. R. Rawlings, A. C. Rodriguez, P. M. Roe, J. Rogers, M. C. R. Bacigalupo, N. Romanov, A. Romieu, R. K. Roth, N. J. Rourke, S. T. Ruediger, E. Rusman, R. M. Sanches-Kuiper, M. R. Schenker, J. M. Seoane, R. J. Shaw, M. K. Shiver, S. W. Short, N. L. Sizto, J. P. Sluis, M. A. Smith, J. E. S. Sohna, E. J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. L. Tregidgo, G. Turcatti, S. vandeVondele, Y. Verhovsky, S. M. Virk, S. Wakelin, G. C. Walcott, J. Wang, G. J. Worsley, J. Yan, L. Yau, M. Zuerlein, J. Rogers, J. C. Mullikin, M. E. Hurles, N. J. McCooke, J. S. West, F. L. Oaks, P. L. Lundberg, D. Klenerman, R. Durbin, and A. J. Smith, "Accurate whole human genome sequencing using reversible terminator chemistry," *Nature*, vol. 456, no. 7218, pp. 53–59, Nov. 2008.

[43] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, "RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays," *Genome Res.*, vol. 18, no. 9, pp. 1509–1517, Sep. 2008.

[44] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq.," *Nat*

Methods, vol. 5, no. 7, pp. 621–628, Jul. 2008.

[45] B. Byeon and I. Kovalchuk, “Pattern Recognition on Read Positioning in Next Generation Sequencing,” *PLOS ONE*, vol. 11, no. 6, p. e0157033, Jun. 2016.

[46] A. Oshlack and M. J. Wakefield, “Transcript length bias in RNA-seq data confounds systems biology,” *Biol Direct*, vol. 4, no. 1, pp. 1–10, Dec. 2009.

[47] S. Ranganathan, K. Nakai, and C. Schonbach, *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, 1st ed. Amsterdam, The Netherlands, Elsevier, 2018.

[48] S. Anders and W. Huber, “Differential expression analysis for sequence count data,” *Genome Biology*, vol. 11, no. 10, p. R106, Oct. 2010.

[49] J. M. Knight, I. Ivanov, and E. R. Dougherty, “MCMC implementation of the optimal Bayesian classifier for non-Gaussian models: model-based RNA-Seq classification,” *BMC Bioinformatics*, vol. 15, no. 1, p. 401, Dec. 2014.

[50] D. M. Witten, “Classification and clustering of sequencing data using a Poisson model,” *The Annals of Applied Statistics*, vol. 5, no. 4, pp. 2493–2518, Dec. 2011.

[51] S. Attoor, E. R Dougherty, Y. Chen, M. L Bittner, and J. M Trent, “Which is better for cDNA-microarray-based classification: Ratios or direct intensities,” *Bioinformatics*, vol. 20, no. 16, pp. 2513–20, Dec. 2004.

[52] D. C. Hoyle, M. Rattray, R. Jupp, and A. Brass, “Making sense of microarray data distributions,” *Bioinformatics*, vol. 18, no. 4, pp. 576–584, Apr. 2002.

[53] G. Zararsız, D. Goksuluk, S. Korkmaz, V. Eldem, G. E. Zararsiz, I. P. Duru, and A. Ozturk, “A comprehensive simulation study on classification of RNA-Seq data,” *PLOS ONE*, vol. 12, no. 8, p. e0182507, Aug. 2017.

[54] D. Goksuluk, G. Zararsiz, S. Korkmaz, V. Eldem, G. E. Zararsiz, E. Ozcetin, A. Ozturk, and A. E. Karaagaoglu, “MLSeq: Machine learning interface for RNA-sequencing data,” *Computer Methods and Programs in Biomedicine*, vol. 175, pp. 223–231, Jul. 2019.

[55] N. Iqbal and P. Kumar, “A Framework for the RNA-Seq Based Classification and Prediction of Disease,” *Preprints*, Jan. 2019.

[56] C. Zhao, I. Ivanov, M. L. Bittner, and E. R. Dougherty, “Pathway regulatory

analysis in the context of Bayesian networks using the coefficient of determination," *Journal of Biological Systems*, vol. 19, no. 4, pp. 651–682, Apr. 2012.

[57] E. Kim, I. Ivanov, and E. R. Dougherty, "Quantifying the notions of canalizing and master genes in a gene regulatory network—a Boolean network modeling perspective," *Bioinformatics*, vol. 35, no. 4, pp. 643–649, Jul. 2018.

[58] M. K. Sunny Sun-Kin Chan, "What is a Master Regulator?," *Journal of stem cell research & therapy*, vol. 3, no. 2, 2013.

[59] C. Lefebvre, P. Rajbhandari, M. J. Alvarez, P. Bandaru, W. K. Lim, M. Sato, K. Wang, P. Sumazin, M. Kustagi, B. C. Bisikirska, K. Basso, P. Beltrao, N. Krogan, J. Gautier, R. D. Favera, and A. Califano, "A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers," *Molecular Systems Biology*, vol. 6, no. 1, p. 377, Jan. 2010.

[60] D. C. Martins, U. M. Braga-Neto, R. F. Hashimoto, M. L. Bittner, and E. R. Dougherty, "Intrinsically Multivariate Predictive Genes," *IEEE J. Sel. Top. Signal Process.*, vol. 2, no. 3, pp. 424–439, 2008.

[61] Ben Lehner, "Genes Confer Similar Robustness to Environmental, Stochastic, and Genetic Perturbations in Yeast," *PLOS ONE*, vol. 5, no. 2, p. e9035, Feb. 2010.

[62] C. H. Waddington, "Canalization of development and the inheritance of acquired characters," *Nature*, vol. 150, no. 3811, pp. 563–565, 1942.

[63] M. Gomez-Lazaro, F. J. Fernandez-Gomez, and J. Jordán, "p53: Twenty five years understanding the mechanism of genome protection," *J. Physiol. Biochem.*, vol. 60, no. 4, pp. 287–307, 2004.

[64] M. Kudo and J. Sklansky, "Comparison of algorithms that select features for pattern classifiers," *Pattern Recognition*, vol. 33, no. 1, pp. 25–41, Jan. 2000.

[65] A. Jain and D. Zongker, "Feature selection: evaluation, application, and small sample performance," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, no. 2, pp. 153–158, 1997.

[66] T. M. Cover and J. M. Van Campenhout, "On the Possible Orderings in the Measurement Selection Problem," *IEEE Trans. Syst., Man, Cybern.*, vol. 7, no. 9, pp. 657–661, 1977.

[67] A. K. Jain and W. G. Waller, “On the optimal number of features in the classification of multivariate Gaussian data,” *Pattern Recognition,* vol. 10, no. 5, pp. 365–374, Jan. 1978.

[68] G. Hughes, “On the mean accuracy of statistical pattern recognizers,” *IEEE Trans. Inform. Theory,* vol. 14, no. 1, pp. 55–63, Jan. 1968.

[69] J. Hua, Z. Xiong, and E. R. Dougherty, “Determination of the optimal number of features for quadratic discriminant analysis via the normal approximation to the discriminant distribution,” *Pattern Recognition,* vol. 38, no. 3, pp. 403–421, Mar. 2005.

[70] J. Hua, Z. Xiong, J. Lowey, E. Suh, and E. R. Dougherty, “Optimal number of features as a function of sample size for various classification rules,” *Bioinformatics,* vol. 21, no. 8, pp. 1509–1515, Apr. 2005.

[71] B. Hanczar, J. Hua, and E. R. Dougherty, “Decorrelation of the true and estimated classifier errors in high-dimensional settings,” *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2007, pp. 2–12, Jan. 2007.

[72] B. Hanczar and E. R Dougherty, “On the Comparison of Classifiers for Microarray Data,” *Current Bioinformatics,* vol. 5, pp. 29–39, Mar. 2010.

[73] B. Hanczar and E. R. Dougherty, “The reliability of estimated confidence intervals for classification error rates when only a single sample is available,” *Pattern Recognition*, vol. 46, no. 3, pp. 1067–1077, Mar. 2013.

[74] A. M. Molinaro, R. Simon, and R. M. Pfeiffer, “Prediction error estimation: a comparison of resampling methods,” *Bioinformatics,* vol. 21, no. 15, pp. 3301–3307, Aug. 2005.

[75] Y. Xiao, J. Hua, and E. R. Dougherty, “Quantification of the impact of feature selection on the variance of cross-validation error estimation,” *EURASIP Journal on Bioinformatics and Systems Biology*, p. 16354, Jan. 2007.

[76] U. Braga-Neto and E. R. Dougherty, “Classification,” *Eurasip Book Series on Signal Processing and Communications*, vol. 2, pp. 93–128, Jan. 2005.

[77] I. Shmulevich and E. Dougherty, *Genomic Signal Processing*. Princeton University Press, 2007.

[78] E. R. Doughtery, H. Jianping, and M. L. Bittner, “Validation of computational methods in genomics.,” *CG*, vol. 8, no. 1, pp. 1–19, Mar. 2007.

[79] 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, Mar. 2009.

[80] L. A. Dalton and E. R. Dougherty, “Application of the Bayesian MMSE estimator for classification error to gene expression microarray data,” *Bioinformatics*, vol. 27, no. 13, pp. 1822–1831, Jul. 2011.

[81] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit, “Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments,” *BMC Bioinformatics*, vol. 11, no. 1, p. 94, Feb. 2010.

[82] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. Springer New York, 2013.

[83] U. M. B. Neto and E. R. Dougherty, *Error estimation for pattern recognition*. Wiley, 2015.

[84] J. M. Knight, E. Kim, I. Ivanov, L. A. Davidson, J. S. Goldsby, M. A. J. Hullar, T. W. Randolph, A. M. Kaz, L. Levy, J. W. Lampe, and R. S. Chapkin, “Comprehensive site-specific whole genome profiling of stromal and epithelial colonic gene signatures in human sigmoid colon and rectal tissue,” *Physiological Genomics*, vol. 48, no. 9, pp. 651–659, Sep. 2016.

[85] C. Zhao, I. Ivanov, E. R. Dougherty, T. J. Hartman, E. Lanza, G. Bobe, N. H. Colburn, J. R. Lupton, L. A. Davidson, and R. S. Chapkin, “Noninvasive Detection of Candidate Molecular Biomarkers in Subjects with a History of Insulin Resistance and Colorectal Adenomas,” *Cancer Prev Res*, vol. 2, no. 6, pp. 590–597, Jun. 2009.

[86] A. Mo, S. Jackson, K. Varma, A. Carpino, C. Giardina, T. J. Devers, and D. W. Rosenberg, “Distinct Transcriptional Changes and Epithelial–Stromal Interactions Are Altered in Early-Stage Colon Cancer Development,” *Mol Cancer Res*, vol. 14, no. 9, pp. 795–804, Sep. 2016.

[87] A. Calon, E. Lonardo, A. Berenguer-Llergo, E. Espinet, X. Hernando-Momblona, M. Iglesias, M. Sevillano, S. Palomo-Ponce, D. V. F. Tauriello, D. Byrom, C. Cortina, C. Morral, C. Barcelo, S. Tosi, A. Riera, C. S. O. Attolini, D. Rossell, E. Sancho, and E. Batlle, “Stromal gene expression defines poor-prognosis subtypes in colorectal cancer.,” *Nat Genet*, vol. 47, no. 4, pp. 320–329, Apr. 2015.

[88] A. K. Jain, P. W. Duin, and Jianchang Mao, “Statistical pattern recognition: a review,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 1, pp. 4–

37, 2000.

[89] A. Wagner, *Robustness and evolvability in living systems*. Princeton University Press, 2007.

[90] A. Rubini and J. Corbet, *Linux Device Drivers*. O'Reilly & Associates, 2001.

[91] B. Govindarajalu, *Comp Arch And Org, 2E*. McGraw-Hill Education, 2010.

[92] C. J. Tabin and R. A. Weinberg, “Analysis of viral and somatic activations of the cHa-ras gene.,” *J. Virol.*, vol. 53, no. 1, pp. 260–265, Jan. 1985.

[93] L. Chang and M. Karin, “Mammalian MAP kinase signalling cascades.,” *Nature*, vol. 410, no. 6824, pp. 37–40, Mar. 2001.

[94] T. Chen and U. M. Braga-Neto, “Statistical detection of intrinsically multivariate predictive genes,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 12, no. 4, pp. 951–963, Jul. 2015.

[95] X. Zhou, X. Wang, and E. R. Dougherty, “Binarization of Microarray Data on the Basis of a Mixture Model1,” *Mol Cancer Ther*, vol. 2, no. 7, pp. 679–684, Jul. 2003.

[96] I. Shmulevich and W. Zhang, “Binary analysis and optimization-based normalization of gene expression data,” *Bioinformatics*, vol. 18, no. 4, pp. 555–565, Apr. 2002.

[97] I. Shmulevich, E. R. Dougherty, and W. Zhang, “Gene perturbation and intervention in probabilistic Boolean networks,” *Bioinformatics*, vol. 18, no. 10, pp. 1319–1331, Oct. 2002.

[98] I. Shmulevich, E. R. Dougherty, S. Kim, and W. Zhang, “Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks,” *Bioinformatics*, vol. 18, no. 2, pp. 261–274, Feb. 2002.

[99] E. R. Dougherty, S. Kim, and Y. Chen, “Coefficient of determination in nonlinear signal processing,” *Signal Processing*, vol. 80, no. 10, pp. 2219–2235, Oct. 2000.

[100] T. Chen and U. M. Braga-Neto, “Statistical Detection of Boolean Regulatory Relationships,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 10, no. 5, p. 1, 2013.

[101] J. Mun, *Advanced Analytical Models: Over 800 Models and 300 Applications*

from the Basel II Accord to Wall Street and Beyond. Wiley, 2008.

[102] S. Kim, E. R. Dougherty, M. L. Bittner, Y. Chen, K. Sivakumar, P. Meltzer, and J. M. Trent, "General nonlinear framework for the analysis of gene interaction via multivariate expression arrays," *Journal of Biomedical Optics*, vol. 5, no. 4, pp. 411–425, 2000.

[103] S. Collot-Teixeira, J. Bass, F. Denis, and S. Ranger-Rogez, "Human tumor suppressor p53 and DNA viruses," *Reviews in Medical Virology*, vol. 14, no. 5, pp. 301–319, Aug. 2004.

[104] A. Datta, R. Pal, and E. Dougherty, "Intervention in Probabilistic Gene Regulatory Networks," *Current Bioinformatics*, vol. 1, no. 2, pp. 167–184, May 2006.

[105] T. Akutsu, M. Hayashida, W. K. Ching, and M. K. Ng, "Control of Boolean networks: Hardness results and algorithms for tree structured networks," *Journal of Theoretical Biology*, vol. 244, no. 4, pp. 670–679, Feb. 2007.

[106] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, no. 2. Athena Scientific, 2012.

[107] M. K. Ng, S. Q. Zhang, W. K. Ching, and T. Akutsu, "A Control Model for Markovian Genetic Regulatory Networks," *Transactions on Computational Systems Biology V*, Berlin, Heidelberg, 2006, pp. 36–48.

[108] X. Qian, I. Ivanov, N. Ghaffari, and E. R. Dougherty, "Intervention in gene regulatory networks via greedy control policies based on long-run behavior," *BMC Systems Biology*, vol. 3, no. 1, p. 177, Jun. 2009.

[109] G. Vahedi, B. Faryabi, J. F. Chamberland, A. Datta, and E. R. Dougherty, "Intervention in Gene Regulatory Networks via a Stationary Mean-First-Passage-Time Control Policy," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 10, pp. 2319–2331, 2008.

[110] I. Ivanov, R. Pal, and E. R. Dougherty, "Dynamics Preserving Size Reduction Mappings for Probabilistic Boolean Networks," *IEEE Transactions on Signal Processing*, vol. 55, no. 5, pp. 2310–2322, 2007.

[111] S. A. Kauffman, "Emergent properties in random complex automata," *Physica D: Nonlinear Phenomena*, vol. 10, no. 1, pp. 145–156, Jan. 1984.

[112] B. Derrida and Y. Pomeau, "Random Networks of Automata: A Simple Annealed Approximation," *Europhysics Letters (EPL)*, vol. 1, no. 2, pp. 45–

49, Jul. 2007.

[113] H. J. Hilhorst and M. Nijmeijer, “On the approach of the stationary state in Kauffman's random Boolean network,” *Journal de Physique*, vol. 48, no. 2, pp. 185–191, 1987.

[114] G. Weisbuch and D. Stauffer, “Phase transition in cellular random Boolean nets,” *Journal de Physique*, vol. 48, no. 1, pp. 11–18, 1987.

[115] D. Stauffer, “Random Boolean networks: Analogy with percolation,” *Philosophical Magazine B*, vol. 56, no. 6, pp. 901–916, Dec. 2006.

[116] S. A. Kauffman, “Requirements for evolvability in complex systems: Orderly dynamics and frozen components,” *Physica D: Nonlinear Phenomena*, vol. 42, no. 1, pp. 135–152, Jun. 1990.

[117] M. Aldana, S. Coppersmith, and L. P. Kadanoff, “Boolean Dynamics with Random Couplings,” in *Perspectives and Problems in Nolinear Science: A Celebratory Volume in Honor of Lawrence Sirovich*, E. Kaplan, J. E. Marsden, and K. R. Sreenivasan, Eds. New York, NY: Springer New York, 2003, pp. 23–89.

[118] S. A. Kauffman, “Metabolic stability and epigenesis in randomly constructed genetic nets,” *Journal of Theoretical Biology*, vol. 22, no. 3, pp. 437–467, Mar. 1969.

[119] S. A. Kauffman, “Origins of Order in Evolution: Self-Organization and Selection,” in *Understanding Origins*, no. 8, Dordrecht: Springer, Dordrecht, 1992, pp. 153–181.

[120] I. Shmulevich, H. Lähdesmäki, E. R. Dougherty, J. Astola, and W. Zhang, “The role of certain Post classes in Boolean network models of genetic networks,” *Proceedings of the National Academy of Sciences*, vol. 100, no. 19, pp. 10734–10739, Sep. 2003.

[121] I. Shmulevich and S. A. Kauffman, “Activities and Sensitivities in Boolean Network Models,” *Physical Review Letters*, vol. 93, no. 4, p. 437, Jul. 2004.

[122] I. Ivanov and E. R. Dougherty, “Reduction Mappings between Probabilistic Boolean Networks,” *EURASIP Journal on Advances in Signal Processing*, vol. 2004, no. 1, p. 314935, Jan. 2004.

[123] J. L. Myers and A. D. Well, *Research Design & Statistical Analysis*, 2nd ed. Psychology Press, 2003.

[124] A. Howard, *Elementary linear algebra with applications : applications version*. Wiley, 2000.

[125] F. Li, T. Long, Y. Lu, Q. Ouyang, and C. Tang, “The yeast cell-cycle network is robustly designed,” *Proceedings of the National Academy of Sciences*, vol. 101, no. 14, pp. 4781–4786, Apr. 2004.

[126] S. A. Kauffman, “Homeostasis and Differentiation in Random Genetic Control Networks,” *Nature*, vol. 224, no. 5215, pp. 177–178, Oct. 1969.

[127] M. Brun, E. R. Dougherty, and I. Shmulevich, “Steady-state probabilities for attractors in probabilistic Boolean networks,” *Signal Processing*, vol. 85, no. 10, pp. 1993–2013, Oct. 2005.

[128] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, “Handling imbalanced datasets: A review,” *GESTS International Transactions on Computer Science and Engineering*, vol. 30, 2006.

[129] S. B. Kotsiantis and P. E. Pintelas, “Mixture of expert agents for handling imbalanced data sets,” *Annals of Mathematics, Computing & Teleinformatics*, vol. 1, pp. 46–55, 2003.

[130] V. G. Ganganwar, “An overview of classification algorithms for imbalanced datasets,” *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, no. 4, Apr. 2012.

[131] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, “An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics,” *Information Sciences*, vol. 250, pp. 113–141, Nov. 2013.

[132] J. Błaszczyński, M. Deckert, J. Stefanowski, and S. Wilk, “Integrating Selective Pre-processing of Imbalanced Data with Ivotes Ensemble,” in *Rough Sets and Current Trends in Computing*, vol. 6086, no. 9, Berlin, Heidelberg: Springer, Berlin, Heidelberg, 2010, pp. 148–157.

[133] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, “RUSBoost: A Hybrid Approach to Alleviating Class Imbalance,” *IEEE Trans. Syst., Man, Cybern. A*, vol. 40, no. 1, pp. 185–197, 2010.

[134] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, “SMOTEBoost: Improving Prediction of the Minority Class in Boosting,” in *Knowledge Discovery in Databases: PKDD 2003*, vol. 2838, no. 1, Berlin, Heidelberg:

Springer, Berlin, Heidelberg, 2003, pp. 107–119.

[135] X. Qian, N. Ghaffari, I. Ivanov, and E. R. Dougherty, "State reduction for network intervention in probabilistic Boolean networks," *Bioinformatics*, vol. 26, no. 24, pp. 3098–3104, Oct. 2010.

[136] M. A. Nielsen, *Neural Networks and Deep Learning*. Determination Press, 2015.

APPENDIX A

ADDITIONAL TABLES AND FIGURES FOR CHAPTER 2

Table A.1 Median and average length of list extensions to find the top set of genes in the other list. τ_{MVN} is the rank of $\mathcal{F}_{\mathrm{NGS}}$ in the MVN list and τ_{NGS} is the rank of $\mathcal{F}_{\mathrm{MVN}}$ in the NGS list where $\mathcal{F}_{\mathrm{NGS}}$ and $\mathcal{F}_{\mathrm{MVN}}$ denote the top set of features in the NGS and MVN list, respectively. Detailed parameter information for each experiment is provided in Table 2.1 in accordance with the experiment number.

Exp No.	Parameters		Median		Average	
			τ_{MVN}	τ_{NGS}	τ_{MVN}	τ_{NGS}
1 (d = 2)	υ = 5	D = 50	6	6	27.0	23.3
		D = 100	11	10	71.0	53.6
		D = 150	16	14	131.3	93.1
	υ = 10	D = 50	7	7	24.8	24.6
		D = 100	11	11	60.6	53.0
		D = 150	15	14	112.3	85.4
	υ = 20	D = 50	12	12	35.4	37.8
		D = 100	17	16	72.0	68.6
		D = 150	19	20	107.8	107.5
1 (d = 3)	υ = 5	D = 50	37	31	273.8	167.6
		D = 100	109	82	1387.9	683.8
		D = 150	224	148	4037.0	1593.7
	υ = 10	D = 50	43.5	42	253.5	204.7
		D = 100	115	96	1220.9	723.9
		D = 150	215	175	3162.4	1639.6
	υ = 20	D = 50	87	84	344.6	338.6
		D = 100	174	165	1282.2	997.3
		D = 150	266	250	2762.9	2010.7
2	n, bresub	40	15	14	112.3	85.4
		80	7	7	35.7	40.0
		120	6	6	21.5	27.9
	n, loo	40	59	62	190.8	228.6
		80	25	27	101.5	109.3
		120	12	15	63.8	64.7
3	$\sigma_{\mu}{}^{2}$	0.5	6	6	30.1	29.6

Table A.1 Continued.

Exp No.	Parameters		Median		Average	
			τ_{MVN}	τ_{NGS}	τ_{MVN}	τ_{NGS}
3	$\sigma_{\mu}{}^{2}$	1	15	14	112.3	85.4
		2	59	51	272.9	275.6
4	ρ	0.1	9	8	76.6	55.9
		0.5	11	11	85.6	67.6
		0.8	15	14	112.3	85.4
5	B	2	31	27	185.1	141.4
		5	15	14	112.3	85.4
		10	10	10	82.1	69.6
6	D=100, υ=5		11	10	71.0	53.6
	D=200, υ=10		19	18	164.3	129.8
	D=300, υ=15		29	26	282.1	213.4

Table A.2 Median and average rank of the Bayes feature set in the MVN and NGS lists.

Exp.	Parameters		Median		Average	
			B_{MVN}	B_{NGS}	B_{MVN}	B_{NGS}
1 ($d = 2$)	$\upsilon = 5$	$D = 50$	15	24	43.9	60.6
		$D = 100$	30	57	115.6	178.3
		$D = 150$	46	99	211.7	350.8
	$\upsilon = 10$	$D = 50$	42	56	81.0	101.7
		$D = 100$	76	114	184.9	260.7
		$D = 150$	120	182.5	316.3	481.5
	$\upsilon = 20$	$D = 50$	105	128	162.0	189.1
		$D = 100$	182	239	336.8	422.6
		$D = 150$	258	352	534.2	716.4
1 ($d = 3$)	$\upsilon = 5$	$D = 50$	59.5	112	263.0	460.7
		$D = 100$	182	387	1258.1	2566.6
		$D = 150$	361	866	3130.1	6794.1
	$\upsilon = 10$	$D = 50$	227	362	648.3	914.0
		$D = 100$	606	117.5	2539.3	4238.0
		$D = 150$	1139	2176	5941.2	10601.0
	$\upsilon = 20$	$D = 50$	892.5	1222.5	1650.0	2074.0
		$D = 100$	2198	3254	5762.0	7867.0
		$D = 150$	3727.5	6222	12426.0	18960.0
2	n, bresub	40	120	182.5	316.3	481.5
		80	26	57	116.2	180.1
		120	15	25	63.3	114.1
	n, loo	40	137	195	346.0	489.5
		80	41	71	136.5	185.5
		120	18	35	82.8	115.4
3	$\sigma_{\mu}{}^{2}$	0.5	23	54	98.8	169.1
		1	120	182.5	316.3	481.5
		2	425	557	1036.3	1516.7
4	ρ	0.1	158	193	349.0	497.3

Table A.2 Continued.

Exp No.	Parameters		Median		Average	
			τ_{MVN}	τ_{NGS}	τ_{MVN}	τ_{NGS}
4	ρ	0.5	146	190	327.4	469.3
		0.8	120	182.5	316.3	481.5
5	B	2	110	183	309.8	473.6
		5	120	182.5	316.3	481.5
		10	129	180	321.3	470.3
6	D=100, υ=5		30	57	115.6	178.3
	D=200, υ=10		157	247	464.9	723.1
	D=300, υ=15		377.5	558	1089.3	1651.9

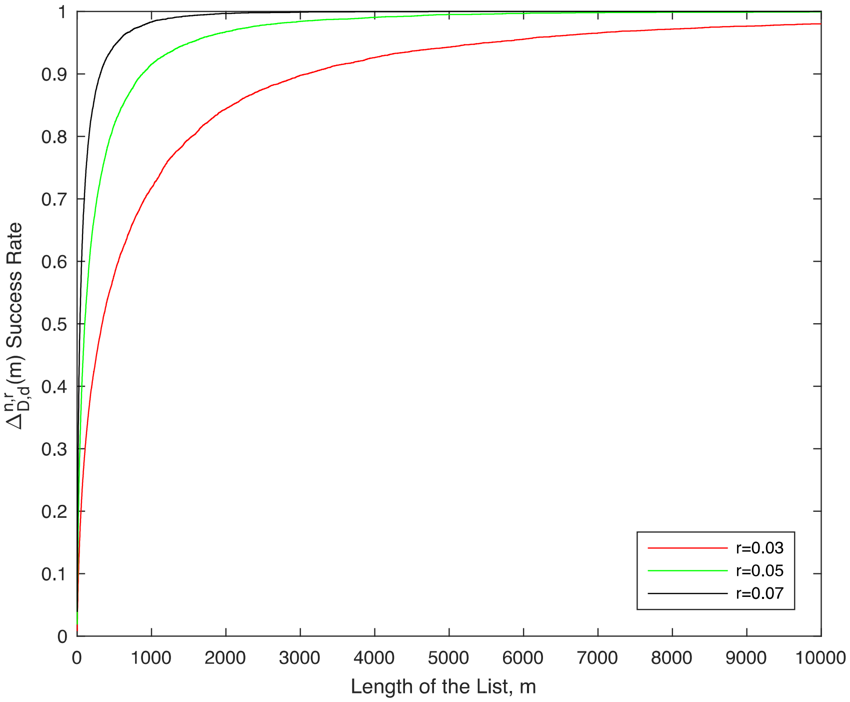


Figure A.1 Power curves for a real dataset where n=59, D=960, d=2. red: r=0.03, green: r=0.05, black: r=0.07.

APPENDIX B

ADDITIONAL FIGURES FOR CHAPTER 4

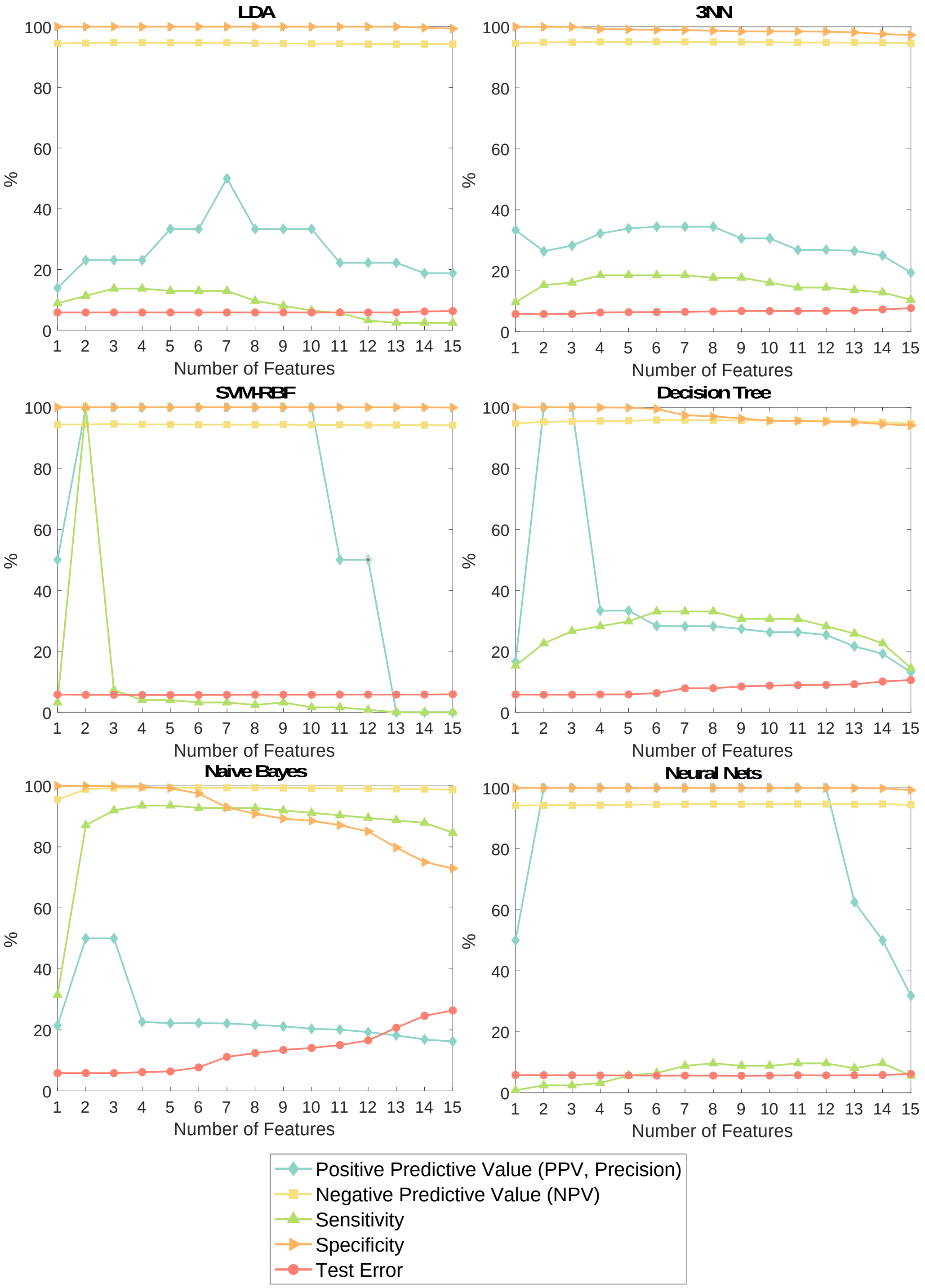


Figure B.1 Based on 15 features listed in Table 4.1, all of the λ-feature ($\lambda = 1,\cdots,15$) classifications are performed on imbalanced data set using six different classification.