FAULT DETECTION AND DIAGNOSIS IN GENE REGULATORY NETWORKS AND OPTIMAL BAYESIAN CLASSIFICATION OF METAGENOMIC DATA

A Dissertation

by

ARGHAVAN BAHADORINEJAD

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

DOCTOR OF PHILOSOPHY

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

May 2017

Major Subject: Electrical Engineering

ABSTRACT

It is well known that the molecular basis of many diseases, particularly cancer, resides in the loss of regulatory power in critical genomic pathways due to DNA mutations. We propose a methodology for model-based fault detection and diagnosis for stochastic Boolean dynamical systems indirectly observed through a single time series of transcriptomic measurements using Next Generation Sequencing (NGS) data. The fault detection consists of an innovations filter followed by a fault certification step, and requires no knowledge about the system faults. The innovations filter uses the optimal Boolean state estimator, called the Boolean Kalman Filter (BKF). We propose an additional step of fault diagnosis based on a multiple model adaptive estimation (MMAE) method consisting of a bank of BKFs running in parallel. The efficacy of the proposed methodology is demonstrated via numerical experiments using a p53-MDM2 negative feedback loop Boolean network. The results indicate the the proposed method is promising in monitoring biological changes at the transcriptomic level. Genomic applications in the life sciences experimented an explosive growth with the advent of high-throughput measurement technologies, which are capable of delivering fast and relatively inexpensive profiles of gene and protein activity on a genome-wide or proteome-wide scale. For the study of microbial classification, we propose a Bayesian method for the classification of r16S sequencing profiles of bacterial abundancies, by using a Dirichlet-Multinomial-Poisson model for microbial community samples. The proposed approach is compared to the kernel SVM, Random Forest and MetaPhyl classification rules as a function of varying sample size, classification difficulty, using synthetic data and real data sets. The proposed Bayesian classifier clearly displays the best performance over different values of between and within class variances that defines the difficulty of the classification.

DEDICATION

To my husband and my parents

ACKNOWLEDGMENTS

First and foremost, I would like to sincerely thank my academic advisor, Dr. Ulisses M. Braga-Neto, for his teaching, advising, and unconditional encouragement and support. He was more than generous with his expertise and precious time. It was definitely an honor for me to work under his supervision. I also thank my committee members, Professor E. Dougherty, Professor E. Serpedin and Dr. I. Ivanov, for their constructive comments and good-natured support. I have a special feeling of gratitude to my beloved parents for inspiring me, their sacrifice and tolerating my absence. Their words of encouragement and push for tenacity ring in my ears. Special thanks goes to my beloved husband, Pedram, for being there throughout the entire period of my doctoral studies. He has been my best cheerleader. All of my success and achievements are undoubtedly because of his endless support. It is my pleasure to appreciate my family and friends who made this dissertation a beginning of my professional journey.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

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

Funding Sources

Optimal classification work was supported by Texas A&M Engineering Strategic Initiative Seed program. Fault detection and diagnosis work was supported by the support of the National Science Foundation, through NSF award CCF-1320884.

NOMENCLATURE

BKF	Boolean Kalman Filter
NGS	Next Generation Sequencing
FDI	Fault Detection and Identification
MMAE	Multiple Model Adaptive Estimation
RNA-Seq	RNA Sequencing
c-DNA	complementary DNA
BNp	Boolean Network with perturbation
HMM	Hidden Markov Model
MMSE	Minimum Mean Square Error
PDV	Posterior Distribution Vector
FDR	False Detection Rate
ATCD	Average Time until Correct Detection
FMR	Fault Misdiagnosis Rate
ATD	Average Time until Diagnosis
POBDS	Partially-Observed Boolean Dynamical System
OBC	Optimal Bayesian Classifier
SVM	Support Vector Machin
RF	Random Forrest
OTU	Operational Taxonomic Units
CBH	Costello et al. Body Habitats
EAC	External Auditory Canal

FS	Fierer et al. Subject
FSH	Fierer et al. Subject Hand

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION AND LITERATURE REVIEW

1.1 Fault Detection and Diagnosis in Gene Regulatory Networks

Biochemical processes in the cellular environment are governed by complex cascades of molecular interactions. Of particular interest are transcriptional regulatory circuits, which govern the functioning of key cellular processes, such as the cell cycle, stress response, DNA repair, and more. Boolean networks, first introduced by Kauffman and collaborators [2, 3], have emerged as an effective model of the dynamical behavior of regulatory circuits consisting of bi-stable genes, which can be either in an activated or suppressed transcriptional state [4, 5, 6, 7, 8]. In the Boolean network model, the transcriptional state of each gene is represented by 0 (OFF) or 1 (ON), and the relationship among genes is described by logical gates updated at discrete time intervals [9, 10, 11]. This model has been successful in accurately modeling the dynamics of the cell cycle in the *Drosophila* fruit fly [6], in the *Saccharomyces cerevisiae* yeast [7], as well as the mammalian cell cycle [8], as well as the switching behavior displayed by the p53 gene in tumor-suppressing pathways [12, 13].

It is well known that the molecular basis of many diseases, particularly cancer, resides in the loss of regulatory power in critical genomic pathways due to DNA mutations. For example, mutations in the p53 gene can render it permanently inactive, with the result that proper response to DNA damage signals are not produced, leading to dangerous disturbances in the cell cycle that may eventually lead to cancerous cells [14]. In this paper, we develop a model-based *fault detection and diagnosis* methodology that can detect and classify sudden changes in the underlying Boolean regulatory network through a single time series of noisy observations of the system state, consisting of transcriptomic data from Next Generation Sequencing (NGS) experiments. The problem of fault detec-

tion and diagnosis (also known as "fault detection and identification") has been studied extensively in many diverse areas of Engineering; e.g., see [15, 16, 17, 18, 19, 20, 21]. However, most if not all of the existing model-based techniques rely on system linearity (or linearizability) assumptions. The methodology developed here applies to Boolean dynamical systems, which are highly-nonlinear, derivativeless (and thus non-linearizable) systems. The optimal state estimator for such models is called the Boolean Kalman Filter [11]. The fault detection consists of an innovations filter followed by a fault certification step, and requires no knowledge about the system faults. The innovations filter relies on the fact that the innovations of the optimal state estimator are uncorrelated under normal operation of the system, a well-known principle in linear Kalman filtering [22], which is applied here in the context of the BKF. The application of the innovations filter for fault detection previously appeared in [23]. In the presence of knowledge about the possible system faults, we propose an additional step of fault diagnosis based on a bank of BKFs running in parallel, which is known as a multiple model adaptive estimation (MMAE) procedure in the literature of linear systems [24]. Performance is assessed by means of false detection and misdiagnosis rates, as well as average times until correct detection and diagnosis. The efficacy of the proposed methodology is demonstrated via numerical experiments using the p53-MDM2 negative feedback loop network with stuck-at faults that model DNA mutation events commonly found in cancer, as described previously. This Boolean network model with stuck-at faults appears in [12], and proposes a fault detection method for deterministic systems with directly observable states. Our methodology removes these assumptions by allowing uncertainty in the state transitions and allowing indirect observation of the state through noisy NGS data.

NGS technologies are able to sequence millions of short DNA fragments in parallel; the length and number of the reads vary with the specific technology [25]. The application of NGS to transcriptional profiling is called RNA-seq (from "RNA sequencing"), which

records how frequently each transcript is represented in a sequence sample [26]. RNA-seq is a probe-free approach that can capture any relevant transcripts present in a sample, without the need of prior knowledge about the target sequence. Due to the accurate sequencing platforms available today, closely related transcripts can be easily distinguished from each other [27], making RNA-seq well-suited to dealing with splice variants, fused transcripts, and mutants. The RNA-seq experiment first randomly fragments messenger RNAs into small pieces, then converts the mRNA fragments to library complementary DNA (cDNA) fragments. The cDNA fragments are amplified and sequenced in parallel, resulting in millions of short sequences called "reads." These reads are mapped to a given region of the genome or transcriptome; the number of reads mapped to each gene determines a count, which is a discrete measure of the corresponding gene expression level [28, 25]. RNA-seq has a large dynamic range and sensitivity due to its digital nature, which is especially important for highly abundant and extremely low abundant genes.

Several tools for partially-observed Boolean dynamical system (POBDS) have been developed in recent years, such as the optimal filter and smoother based on the minimum mean square error (MMSE) criterion, called the Boolean Kalman Filter (BKF) and Boolean Kalman Smoother (BKS) [29], respectively. In addition, particle filtering implementations of these filters, as well as schemes for handling correlated noise, simultaneous state and parameter estimation, network inference, and control for POBDSs were developed [30, 31, 32, 33, 34, 35, 36, 37]. The software tool "BoolFilter" [38] is available under R library for estimation and identification of partially-observed Boolean dynamical systems.

1.2 Optimal Bayesian Classification of Metagenomic Data

The characterization of microbial communities by 16S rRNA gene amplicon sequencing has received renewed interest in the last decade, in part due to the emergence of high-

throughput sequencing technology [39]. Microbial metagenomics provides a means to determine what organisms are present without the need for isolation and culturing. Next generation sequencing, applied to microbial metagenomics, has transformed the study of microbial diversity [40]. For amplicons reads it is possible to classify sequence reads against known taxa, and determine a list of those organisms that are present and the read frequency associated with them [41]. In this case an unsupervised strategy can be used to identify proxies to traditional taxonomic units by clustering sequences, so called Operational Taxonomic Units (OTUs).

In this paper, we apply the Optimal Bayesian Classifier (OBC) [42] to sample classification based on r16S sequencing profiles of microbial abundance. Each microbial sample is represented as a set of operational taxonomic unit (OTU) frequencies. The performance of the proposed approach is compared to other classifiers such as a kernel Support Vector Machine (SVM), Random Forest (RF), which is considered to be the de-facto standard for metagenomics classification, and the MetaPhyl algorithm [43], as a function of varying sample size, classification difficulty, by means of a numerical experiment using synthetic data and real data sets.

2. FAULT DETECTION AND DIAGNOSIS IN TRANSCRIPTIONAL CIRCUITS USING NEXT-GENERATION SEQUENCING*

2.1 Boolean Networks

Let the Boolean expression state $X_{ki} \in \{0,1\}$ of each of d transcripts at a discrete time k be represented by a Boolean *state vector* $\mathbf{X}_k = (X_{k1}, \ldots, X_{kd}) \in \{0,1\}^d$. A Boolean network (with input) describes the evolution of the time series $\{\mathbf{X}_k; k = 0, 1, \ldots\}$ by

$$\mathbf{X}_k \;=\; \mathbf{f}\left(\mathbf{X}_{k-1}, \mathbf{u}_k\right) \;, \tag{2.1}$$

for $k = 1, 2, \ldots$, where $\mathbf{u}_k \in \{0,1\}^p$ is an input vector of dimension p at time k and $\mathbf{f} : \{0,1\}^{d+p} \to \{0,1\}^d$ is an arbitrary *network function* that transforms the previous state plus the current input into the current state. The network function can be written in terms of its components, $\mathbf{f} = (f_1, f_2, \ldots, f_d)$, where each component $f_i : \{0,1\}^{d+p} \to \{0,1\}$, $i = 1, \ldots, d$, is a *Boolean function*, which expresses a logical relationship among the state and input variables.

As an example, consider the *p53-MDM2* negative feedback loop transcriptional circuit in the presence of DNA double strand breaks. The *p53* transcription factor plays a key role in tumor suppression; in fact 30% to 50% of commonly occurring human cancers are associated with loss of p53 functionality through mutations [14]. The dynamics of p53 in response to DNA double strand breaks can be summarized by the activated/deactivated patterns of the genes for *p53*, *MDM2*, the serine-protein kinase *ATM* (responsible for transduction of the DNA damage signal), and the phosphatase *WIP1* [1, 13]. Hence,

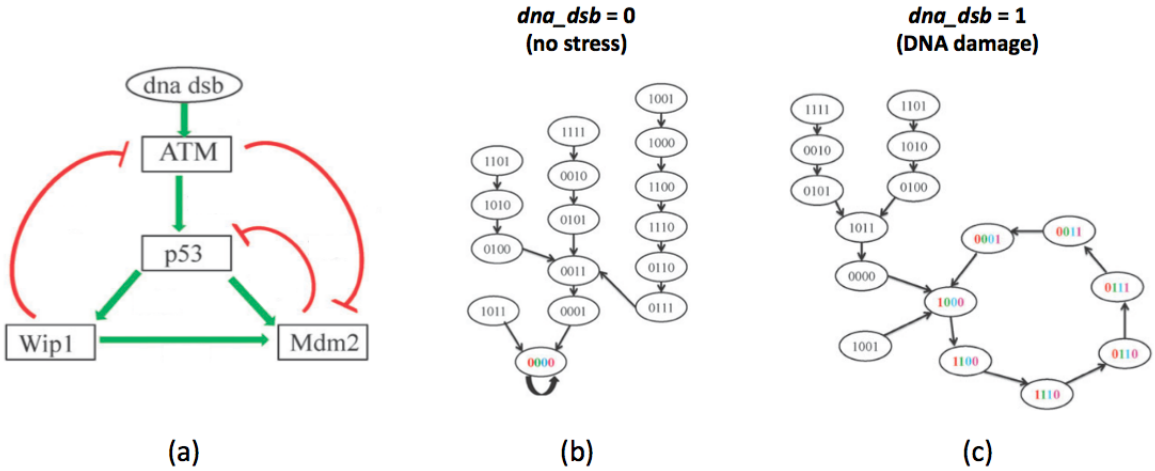

Figure 2.1: Activation/repression pathways and state transition diagrams corresponding to constant inputs $dna\_dsb_k \equiv 0$ (no stress) and $dna\_dsb_k \equiv 1$ (DNA damage) for the *p53-MDM2* negative feedback loop Boolean network model. Reproduced from [1].

$d = 4$ and the state of the system at time k can be represented by the Boolean vector $\mathbf{X}_k = (ATM_k, p53_k, WIP1_k, MDM2_k)$, where a subscript k is attached to the name of gene to indicate its expression state at time k. The input $u_k = dna\_dsb_k$ at time k is a Boolean signal that indicates the presence of DNA double strand breaks. Using the pathway information in [13], Layek and Datta obtained the following Boolean network model [1]:

$$\begin{aligned} ATM_k &= \overline{WIP1_{k-1}} \text{ AND } (ATM_{k-1} \text{ OR } dna\_dsb_k) \\ p53_k &= \overline{MDM2_{k-1}} \text{ AND } (ATM_{k-1} \text{ OR } WIP_{k-1}) \\ WIP1_k &= p53_{k-1} \\ MDM2_k &= \overline{ATM_{k-1}} \text{ AND } (p53_{k-1} \text{ OR } WIP_{k-1}) \end{aligned} \tag{2.2}$$

We can see that *MDM2* has a suppressing effect on *p53*, which in turn activates it. This is the well-known *p53-MDM2* negative-feedback regulatory loop, which, under no stress, keeps the expression of *p53* at small levels. However, we can see that *MDM2* is also

inactivated by *ATM*, which in turn is activated by the DNA damage signal. Figure 2.1(a) depicts the activation/suppression pathways corresponding to this BN. We can see that *ATM* is the transductor gene for the DNA damage signal, which eventually activates *p53*, through inactivation of *MDM2*. However, there is also a negative feedback loop between *p53* and *ATM*, through *WIP1*, so that *p53* is expected to display an oscillatory behavior under DNA damage [13]. On the other hand, under no stress, it is known that all four proteins are inactivated in the steady state [14]. These behaviors are captured nicely by the BN model. Figures 2.1(b-c) display the state transition diagram under no stress and under DNA damage, respectively. In the first case, we can see that "0000" is a singleton attractor state, while the other states are transient. In the second case, we can see that there is a cyclic attractor, corresponding to an oscillation of *p53*, along with the other proteins in its regulatory pathway. Furthermore, we can see that activation of *MDM2* and *WIP1* lags behind that of *p53*, which in turn lags behind that of *ATM*. All of these behaviors are consistent with biological knowledge [44]. Summing up, the BN displays *two basins of attraction*, corresponding to the state of the DNA damage signal. The moment the latter changes, the system jumps to the other basin of attraction, and if left undisturbed will eventually reach the restive state, in the case of no stress, or the cyclic pattern of *p53* activation, in the case of DNA damage.

2.2 Stochastic Signal Model

The Boolean network model describes a deterministic dynamical system, and assumes that the state is directly observable or that noisy non-Boolean observations can be readily thresholded into Boolean values. But, in fact, due to system noise from unmodeled variables, there is uncertainty in state transition; i.e., there is a chance that the next state of the system is not the one prescribed by the Boolean network. In addition, the Boolean states of a system are never observed directly; modern expression-based technologies, such as

RNA-seq, (1) produce noisy non-Boolean measurements, and (2) may also produce measurements of part of the state vector only. In the first case, thresholding methods can be used to binarize the expression for each gene; in the second case, there is no recovery. In this Section, we describe a stochastic signal model for Boolean Dynamical systems, first introduced in [11], that accounts for these issues.

2.2.1 Boolean State Transition Model

There have been a series of models proposed to extend the Boolean Network model to allow uncertainty in state transitions. These include Random Boolean Networks [2], Boolean Networks with perturbation (BNp) [10], and Probabilistic Boolean Networks (PBN) [45]. In [11], it is shown that PBNs are actually a special case of Boolean Network with perturbations. The BNp model is thus quite general, and is used here as the state model. The sequence of state vectors $\{\mathbf{X}_k; k = 0, 1, \ldots\}$ is a Markov stochastic process, called the *state process*, specified by

$$\mathbf{X}_k = \mathbf{f}(\mathbf{X}_{k-1}, \mathbf{u}_k) \oplus \mathbf{n}_k, \tag{2.3}$$

for $k = 1, 2, \ldots$, where "$\oplus$" indicates component-wise modulo-2 addition, $\mathbf{u}_k$ and $\mathbf{f} : \{0,1\}^{d+p} \to \{0,1\}^d$ are the input and network function, respectively (see the previous Section), whereas $\{\mathbf{n}_k; k = 1, 2, \ldots\}$ is a "white noise" process, with $\mathbf{n}_k = (N_{k1}, \ldots, N_{kd}) \in \{0,1\}^d$. The noise process is "white" in the sense that $\mathbf{n}_k$ is an independent process, which is independent from the initial state $\mathbf{X}_0$; its distribution is otherwise arbitrary. The input $\mathbf{u}_k$ is typically, but not necessarily, a deterministic signal.

The state equation (2.3) differs from (2.1) by the presence of the additive noise process. It therefore extends the previous Boolean Network model to allow for stochasticity. In the special case where the noise vector is i.i.d., with $P(N_{ki} = 1) = p$ for $i = 1, \ldots, d$, then there is a probability p that each state variable X_{ki} will be flipped from 0 to 1 or 1 to 0,

independently of the other state variables. The parameter p determines the intensity of the noise, i.e., how often a state variable is flipped. If p is very small, the system state evolves as a standard Boolean Network, settling into attractors in the long range, but the occasional flip of a variable can pull the system out of its attractor states, and even into different attractor basins altogether. On the other hand, larger p leads to much more chaotic behavior.

2.2.2 Observation Model for RNA-seq Data

The second component of the signal model is the observational model. In most real-world applications, the system state is only partially observable, and distortion is introduced in the observations by sensor noise — this is certainly the case with RNA-seq transcriptomic data.

Let $\mathbf{Y}_k = (Y_{k1}, \ldots, Y_{kd})$ be a vector containing the RNA-seq data at time k, for $k = 1, 2, \ldots$ We assume a single-lane NGS platform, so that Y_{ki} is the read count corresponding to transcript i in the single lane, for $i = 1, \ldots, d$. There are multiple methods for modeling the RNA-seq reads. Because of the discrete nature of reads, most methods are based on either the negative binomial [46] or the Poisson distribution [25]. In this study, we choose to use the Poisson model for the number of reads for each transcript:

$$P(Y_{ki} = m \mid \lambda_{ki}) \;=\; e^{-\lambda_{ki}} \, \frac{\lambda_{ki}^m}{m!}\,, \quad m = 0, 1, \ldots \tag{2.4}$$

where λ_{ki} is the mean read count of transcript i at time k. Recall that, according to the Boolean state model, there are two possible states for the abundance of transcript i at time k: high ($X_{ki} = 1$) and low ($X_{ki} = 0$). Accordingly, we model the parameter λ_{ki} as follows:

$$\begin{aligned} \log(\lambda_{ki}) \;&=\; \log(s) + \mu_b + \theta_i\,, && \text{if } X_{ki} = 0\,, \\ \log(\lambda_{ki}) \;&=\; \log(s) + \mu_b + \delta_i + \theta_i\,, && \text{if } X_{ki} = 1\,. \end{aligned} \tag{2.5}$$

The parameter s is the *sequencing depth* [28], and is assumed here to be common to all transcripts, since a single lane is being modeled. The sequencing depth s accounts for different total numbers of reads produced in the lane and plays a key role, since it determines the approximate range of read counts that is produced. The parameter $\mu_b > 0$ accounts for the baseline level of read counts produced in the single lane in the inactivated transcriptional state, which is assumed to be common to all transcripts. The differential level of expression $\delta_i > 0$ expresses the effect on the observed RNA-seq read count as transcript i goes from the inactivated to the activated state. This effect may change for different transcripts, which is modeled here by assuming δ_i to be Gaussian with mean $\mu_\delta > 0$ and variance σ_δ^2, common to all transcripts, where σ_δ is assumed to be small enough to keep δ_i positive. The Gaussianity assumption is by no means necessary to the proposed approach; any other positive-valued distribution, such as exponential or Gamma, could be substituted for the Gaussian and the methodology would apply unchanged. The term θ_i is zero-mean Gaussian noise with small variance σ_θ^2, common to all transcripts (once again, the Gaussian assumption is not essential), and models unknown and unwanted technical effects that may occur during the experiment. Typical values for all these parameters are given in Section VII when we discuss the numerical experiments performed to evaluate the proposed approach.

Note that one may rewrite the equations in (2.5) as:

$$\lambda_{ki} \;=\; s\, \exp(\mu_b + \theta_i + \delta_i X_{ki})\,. \tag{2.6}$$

As θ_i and δ_i are normally distributed and μ_b is fixed, the mean read count Y_{ki} for transcript i at time k has a *log-normal* distribution given X_{ki}.

For simplicity, we assume that, at a each time k, the read counts Y_{ki}, for $i = 1, \ldots, d$, are conditionally independent *given* $\mathbf{X}_k$. In practice, the model will still be applicable if

the read counts are only weakly correlated. However, the independence assumption can be lifted at the expense of introducing extra parameters. A model including all correlations between transcripts would require $d(d-1)$ extra parameters (the off-diagonal elements of the covariance matrix); to reduce model complexity for large d, one may assume a block covariance matrix, where clusters of transcripts are correlated (e.g., the ones with similar sequences) but the clusters are conditionally-independent of each other.

2.3 Boolean Kalman Filter

The minimum mean-square error (MMSE) state estimator for the model described in the previous sections is the *Boolean Kalman Filter* (BKF) [11]. A recursive algorithm for the exact computation of the BKF for a general signal model was given in [11]. Next, we adapt it to the signal model described in the previous sections.

The optimal *filtering* problem consists of, given the history of observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_k$ up to the present time k, find an estimator $\hat{\mathbf{X}}_k = h(\mathbf{Y}_1, \ldots, \mathbf{Y}_k)$ of the state $\mathbf{X}_k$ that optimizes a given performance criterion. The criterion considered here is the (conditional) mean-square error:

$$\mathrm{MSE} = E\left[||\hat{\mathbf{X}}_k - \mathbf{X}_k||^2 \mid \mathbf{Y}_k, \ldots, \mathbf{Y}_1\right] . \tag{2.7}$$

The solution to this problem is given next. Let $(\mathbf{x}^1, \ldots, \mathbf{x}^{2^d})$ be an arbitrary enumeration of the possible state vectors. For each time $k = 1, 2, \ldots$ define the posterior distribution vectors (PDV) $\Pi_{k|k}$ and $\Pi_{k|k-1}$ of length 2^d by:

$$\begin{aligned} (\Pi_k)_i &= P\left(\mathbf{X}_k = \mathbf{x}^i \mid \mathbf{Y}_k, \ldots, \mathbf{Y}_1\right) , \\ (\Pi_{k|k-1})_i &= P\left(\mathbf{X}_k = \mathbf{x}^i \mid \mathbf{Y}_{k-1}, \ldots, \mathbf{Y}_1\right) , \end{aligned} \tag{2.8}$$

for $i = 1, \ldots, 2^d$. These give the posterior distribution of the state given the observation

histories up to time k and $k-1$, respectively. Let the *prediction matrix* M_k of size $2^d \times 2^d$ be the transition matrix of the Markov chain defined by the state model:

$$\begin{aligned}(M_k)_{ij} &= P(\mathbf{X}_k = \mathbf{x}^i \mid \mathbf{X}_{k-1} = \mathbf{x}^j) \\ &= P\left(\mathbf{n}_k = \mathbf{x}^i \oplus \mathbf{f}(\mathbf{x}^j, \mathbf{u}_k)\right) \\ &= p^{O(\mathbf{x}^i \oplus \mathbf{f}(\mathbf{x}^j, \mathbf{u}_k))}(1-p)^{1-O(\mathbf{x}^i \oplus \mathbf{f}(\mathbf{x}^j, \mathbf{u}_k))},\end{aligned} \tag{2.9}$$

for $i, j = 1, \ldots, 2^d$, where $O(\mathbf{v})$ is defined as the number of 1's in Boolean vector $\mathbf{v}$ and we assumed that the noise vector $\mathbf{n}_k$ is i.i.d. On the other hand, let the *update matrix* T_k, also of size $2^d \times 2^d$, be a diagonal matrix defined by:

$$\begin{aligned}(T_k)_{jj} &= P(\mathbf{Y}_k \mid \mathbf{X}_k = \mathbf{x}^j) \\ &= e^{(-\sum_{i=1}^d \lambda_{ki})} \prod_{i=1}^d \frac{\lambda_{ki}^{Y_{ki}}}{Y_{ki}!}\end{aligned} \tag{2.10}$$

for $j = 1, \ldots, 2^d$, where we used (2.4) and (2.6). Finally, define the matrix A of size $d \times 2^d$ via $A = \left[\mathbf{x}^1 \cdots \mathbf{x}^{2^d}\right]$.

The following result is a specialization of Theorem 1 in [**?**] to the signal model described previously.

Theorem 1. (Boolean Kalman Filter.) *The optimal minimum MSE estimator $\hat{\mathbf{X}}_k$ of the state $\mathbf{X}_k$ given the observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_k$ up to time k is given by*

$$\hat{\mathbf{X}}_k = \overline{E\left[\mathbf{X}_k \mid \mathbf{Y}_k, \ldots, \mathbf{Y}_1\right]}, \tag{2.11}$$

where the binarization of a vector $\mathbf{v}$ is defined by $(\overline{\mathbf{v}})_i = I_{(\mathbf{v})_i > 1/2}$ for $i = 1, \ldots, d$. This estimator and its optimal conditional MSE can be computed by the following procedure:

1. *Initialization Step: The initial PDV is defined by $(\Pi_{0|0})_i = P\left(\mathbf{X}_0 = \mathbf{x}^i\right)$, for $i =$*

$1, \ldots, 2^d$.

For $k = 1, 2, \ldots$, *do:*

2. *Prediction Step: Given the previous PDV* $\Pi_{k-1|k-1}$*, the* predicted *PDV* $\Pi_{k|k-1}$ *is given by* $\Pi_{k|k-1} = M_k \, \Pi_{k-1|k-1}$.

3. *Update Step: Let* $\boldsymbol{\beta}_k = T_k \, \Pi_{k|k-1}$. *The* updated *PDV* $\Pi_{k|k}$ *is obtained by normalizing* $\boldsymbol{\beta}_k$ *to obtain a probability measure:* $\Pi_{k|k} = \boldsymbol{\beta}_k / ||\boldsymbol{\beta}_k||_1$.

4. *MMSE Estimator Computation Step: The MMSE estimator is given by*

$$\hat{\mathbf{X}}_k = \overline{A\Pi_{k|k}} \tag{2.12}$$

with optimal conditional MSE

$$\mathrm{MSE}(\mathbf{Y}_1, \ldots, \mathbf{Y}_k) = || \min\{A\Pi_{k|k}, (A\Pi_{k|k})^c\} ||_1 \,, \tag{2.13}$$

where the minimum is applied component-wise, and the complement of a vector $\mathbf{v}$ *is defined by* $(\mathbf{v}^c)_i = 1 - (\mathbf{v})_i$*, for* $i = 1, \ldots, d$.

Notice that, due to the normalization in the update step, the matrix T_k can be scaled at will. In particular, the constant $s^{(\sum_{i=1}^d Y_{ki})} / \Pi_{i=1}^d Y_{ki}!$ in (2.10) can be dropped, which results in significant computational savings.

2.4 Fault Detection and Diagnosis System

The proposed methodology for model-based fault detection and diagnosis is based on the Boolean Kalman Filter, introduced in the previous section. The fault detection step consists of an innovations filter followed by a fault certification step, and requires no

knowledge about the system faults, followed by a fault certification step that is aimed at reducing the false detection rate. The fault diagnosis step is based on a bank of BKFs running in parallel, which is known as a multiple model adaptive estimation (MMAE) procedure. The proposed fault detection and diagnosis system is represented as a block diagram in Figure 2.2.

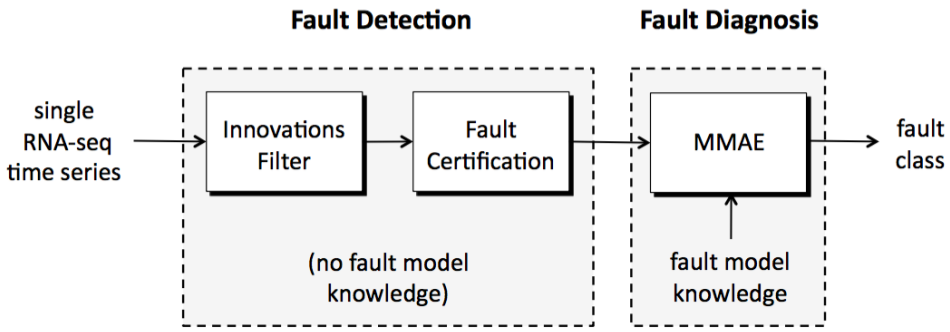

Figure 2.2: Block diagram of the proposed fault detection and diagnosis system.

2.4.1 Innovations Filter

It is a well-known fact in the theory of linear Kalman filters [22] that the innovations of the optimal state estimator constitute a "white noise" sequence. This fact is applied here in the context of the BKF. For completeness, we give the main result below. Given the history of observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_k$ up to the present time k, let $\mathbf{Y}_k^{\mathrm{MS}} = E\left[\mathbf{Y}_k \mid \mathbf{Y}_{k-1}, \ldots, \mathbf{Y}_1\right]$ is the *MS-predictable* component of $\mathbf{Y}_k$, with $\mathbf{Y}_1^{\mathrm{MS}} = E[\mathbf{Y}_1]$. The *innovation* $\mathbf{V}_k$ at time k is defined as the MS-unpredictable component of $\mathbf{Y}_k$:

$$\begin{aligned} \mathbf{V}_k &= \mathbf{Y}_k - \mathbf{Y}_k^{\mathrm{MS}} \\ &= \mathbf{Y}_k - E\left[\mathbf{Y}_k \mid \mathbf{Y}_{k-1}, \ldots, \mathbf{Y}_1\right] , \end{aligned} \tag{2.14}$$

for $k = 1, 2, \ldots$ The next theorem establishes the "white noise" property of the innovations sequence.

Theorem 2. *The innovations sequence* $\{\mathbf{V}_k; k = 0, 1, \ldots\}$ *is zero-mean and uncorrelated; i.e.,* $E[\mathbf{V}_k] = \mathbf{0}$ *and* $E[\mathbf{V}_k \mathbf{V}_l^T] = 0_{d\times d}$*, for* $k, l = 1, 2, \ldots$ *and* $k \neq l$*.*

Proof. We have

$$\begin{aligned} E[\mathbf{V}_k] \;&=\; E[\mathbf{Y}_k] - E[\mathbf{Y}_k^{\mathrm{MS}}] \\ &=\; E[\mathbf{Y}_k] - E[E[\mathbf{Y}_k \mid \mathbf{Y}_{k-1}, \ldots, \mathbf{Y}_1]] \\ &=\; E[\mathbf{Y}_k] - E[\mathbf{Y}_k] \;=\; \mathbf{0}\,, \end{aligned} \tag{2.15}$$

for $k = 1, 2, \ldots$, showing that the innovations sequence is zero mean. We now show that $E[\mathbf{V}_k \mathbf{V}_l^T] = 0_{d\times d}$, where we can assume that $k > l$, without loss of generality. First note that

$$\begin{aligned} E[\mathbf{Y}_k^{\mathrm{MS}} \mathbf{Y}_l^T] \;&=\; E[E[\mathbf{Y}_k \mid \mathbf{Y}_{k-1}, \ldots, \mathbf{Y}_1]\mathbf{Y}_l^T] \\ &=\; E[E[\mathbf{Y}_k \mathbf{Y}_l^T \mid \mathbf{Y}_{k-1}, \ldots, \mathbf{Y}_1]] \;=\; E[\mathbf{Y}_k \mathbf{Y}_l^T]\,. \end{aligned} \tag{2.16}$$

Similarly, we can show that $E[\mathbf{Y}_k^{\mathrm{MS}}(\mathbf{Y}_l^{\mathrm{MS}})^T] \;=\; E[\mathbf{Y}_k(\mathbf{Y}_l^{\mathrm{MS}})^T]$. Now,

$$\begin{aligned} E[\mathbf{V}_k \mathbf{V}_l^T] \;&=\; E[(\mathbf{Y}_k - \mathbf{Y}_k^{\mathrm{MS}})(\mathbf{Y}_l - \mathbf{Y}_l^{\mathrm{MS}})^T] \\ &=\; E[\mathbf{Y}_k \mathbf{Y}_l^T] - E[\mathbf{Y}_k(\mathbf{Y}_l^{\mathrm{MS}})^T] \\ &\quad - E[\mathbf{Y}_k^{\mathrm{MS}} \mathbf{Y}_l^T] + E[\mathbf{Y}_k^{\mathrm{MS}}(\mathbf{Y}_l^{\mathrm{MS}})^T] \;=\; 0_{d\times d}\,, \end{aligned} \tag{2.17}$$

as required. □

Now, note that $\mathbf{Y}_k^{\mathrm{MS}}$ can be written as

$$\begin{aligned}
\mathbf{Y}_k^{\mathrm{MS}} &= E[E[\mathbf{Y}_k \,|\, \mathbf{X}_k, \mathbf{Y}_{k-1}, \ldots, \mathbf{Y}_1] \,|\, \mathbf{Y}_{k-1}, \ldots, \mathbf{Y}_1] \\
&= E[\, E[\mathbf{Y}_k \mid \mathbf{X}_k] \mid \mathbf{Y}_{k-1}, \ldots, \mathbf{Y}_1] \\
&= \sum_{i=1}^{2^d} E[\mathbf{Y}_k \mid \mathbf{X}_k = \mathbf{x}^i] \\
&\qquad\qquad \times\ P\left(\mathbf{X}_k = \mathbf{x}_k \mid \mathbf{Y}_{k-1}, \ldots, \mathbf{Y}_1\right) \\
&= D_k\, \Pi_{k|k-1}
\end{aligned} \tag{2.18}$$

where $\Pi_{k|k-1}$ is the vector of posterior probabilities defined in (2.8), and D_k is a $d \times 2^d$ matrix defined by

$$\begin{aligned}
(D_k)_{ij} &= E[Y_{ki} \mid \mathbf{X}_k = \mathbf{x}^j] \;=\; \lambda_{ki} \\
&= s\, \exp(\mu_b + \theta_i + \delta_i(\mathbf{x}^j)_i)\,,
\end{aligned} \tag{2.19}$$

for $i = 1, \ldots, d$ and $j = 1, \ldots, 2^d$, where we used (2.6).

Assuming normal operation of the system, the BKF is run, and the residue at time k,

$$\mathbf{e}_k \;=\; \mathbf{Y}_k - D_k\, \Pi_{k|k-1}\,, \tag{2.20}$$

is computed, for $k = 1, 2, \ldots$ From the previous results, the sequence $\{\mathbf{e}_k; k = 1, 2, \ldots\}$ is the innovations sequence, and therefore "white noise," *provided that* the normal-operation model, assumed by the BKF, matches the actual model producing the data. A fault detection method therefore is provided by testing the hypothesis that the residue sequence is white. In this paper, this is done by means of a chi-square test applied to L lags of the sample auto-correlation function based on a "window" of observations preceding k_0 of a specified length L_D.

2.4.2 Fault Certification

To improve false-positive error rates, we adopt the fault certification step described in [20]: a fault is signaled if the number of false detections over a window of specified length L_C preceding the current time k_0 exceeds a threshold t_C; for example, in [20], a fault is signaled if a fault is detected at least 3 times over a window of 10 time points preceding the current time.

2.4.3 Fault Diagnosis

Upon detection of a fault, and if all possible system faults are known and can be modeled, then an optimal Bayesian procedure for fault diagnosis can be derived, which selects the fault as the candidate with the largest posterior probability given the observations. The computation makes use of probabilities computed by a bank of BKFs running in parallel, one for each candidate fault model. This is similar to the *multiple model adaptive estimation* (MMAE) method for linear systems [24], applied here to the nonlinear signal model described previously. Let M be the number of candidate fault models, which are indexed by $C \in \{1, \ldots, M\}$. It is assumed that this set is exhaustive; that is, once a fault is (correctly) detected, then one of the fault models must be the one in operation. Our goal is to determine which one. Let k_d be the time at which a fault is detected. At time k_d, the prior probability of fault class m is given by $P(C = m)$, for $m = 1, \ldots, M$. Let p_l^m be the posterior probability of fault class m at time $k_d + l$, given the history of observations $\mathbf{Y}_{k_d+1}, \ldots, \mathbf{Y}_{k_d+l}$ between the time immediately after the fault is detected and the present time $k_d + l$, for $l = 1, 2, \ldots$ and $m = 1, \ldots, M$. A simple application of Bayes theorem yields

$$\begin{aligned} p_l^m &= P(C = m \mid \mathbf{Y}_{k_d+l}, \ldots, \mathbf{Y}_{k_d+1}) \\ &= \frac{P(\mathbf{Y}_{k_d+l} \mid C = m, \mathbf{Y}_{k_d+l-1}, \ldots, \mathbf{Y}_{k_d+1}) p_{l-1}^m}{\sum_{i=1}^{M} P(\mathbf{Y}_{k_d+l} \mid C = i, \mathbf{Y}_{k_d+l-1}, \ldots, \mathbf{Y}_{k_d+1}) p_{l-1}^i} \end{aligned} \tag{2.21}$$

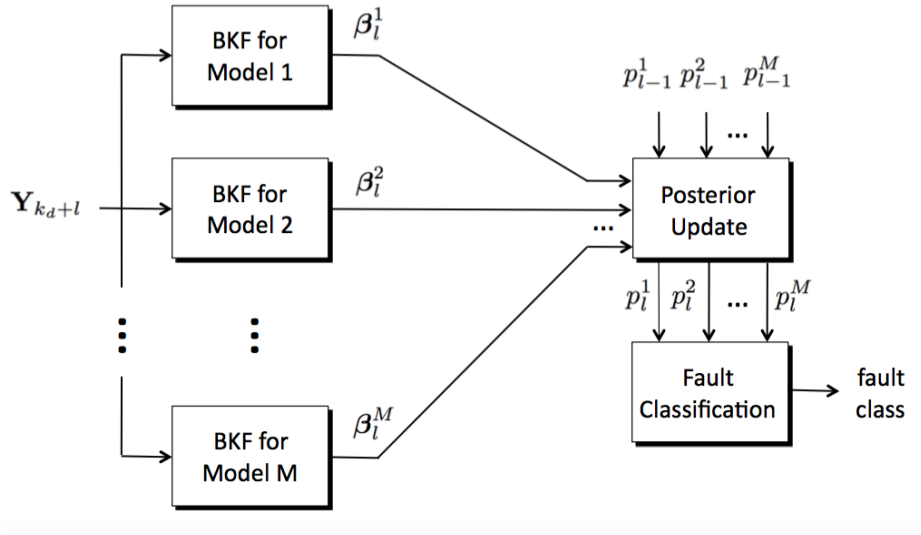

Figure 2.3: Block diagram for one iteration of the proposed fault diagnosis system.

for $l = 1, 2, \ldots$, with $p_0^m = P(C = m)$. Furthermore, one can write

$$
\begin{aligned}
& P(\mathbf{Y}_k \mid C = m, \mathbf{Y}_{k_d+l}, \ldots, \mathbf{Y}_{k_d+1}) \\
&= \sum_{i=1}^{2^d} P(\mathbf{Y}_{k_d+l}, \mathbf{X}_{k_d+l} = \mathbf{x}^i \mid C = m, \mathbf{Y}_{k_d+l-1}, \ldots, \mathbf{Y}_{k_d+1}) \\
&= \sum_{i=1}^{2^d} P(\mathbf{Y}_{k_d+l} \mid \mathbf{X}_{k_d+l} = \mathbf{x}^i, C = m, \mathbf{Y}_{k_d+l-1}, \ldots, \mathbf{Y}_{k_d+1}) \\
&\qquad \times P(\mathbf{X}_{k_d+l} = \mathbf{x}^i \mid C = m, \mathbf{Y}_{k_d+l-1}, \ldots, \mathbf{Y}_{k_d+1}) \\
&= \sum_{i=1}^{2^d} P(\mathbf{Y}_{k_d+l} \mid \mathbf{X}_{k_d+l} = \mathbf{x}^i, C = m) \\
&\qquad \times P(\mathbf{X}_{k_d+l} = \mathbf{x}^i \mid C = m, \mathbf{Y}_{k_d+l-1}, \ldots, \mathbf{Y}_{k_d+1}) \\
&= \sum_{i=1}^{2^d} (T_l^m)_{ii} \, (\Pi_{l|l-1}^m)_i \\
&= ||T_l^m \, \Pi_{l|l-1}^m||_1 \; = \; ||\boldsymbol{\beta}_l^m||_1 \,,
\end{aligned}
\tag{2.22}
$$

where $\boldsymbol{\beta}_l^m$ is the unnormalized PDV at time $k_d + l$, computed at the update step of a BKF based on model $C = m$ and started at time k_d, for $m = 1, \ldots, M$.

Hence, the posterior probability of fault class m can be obtained by running a bank of M BKFs in parallel, one for each of the fault models, where all BKFs are started at time k_d. At each time $k_d + l$ the posterior probabilities are updated by the equation:

$$p_l^m \;=\; \frac{||\boldsymbol{\beta}_l^m||_1 \, p_{l-1}^m}{\sum_{i=1}^{M} ||\boldsymbol{\beta}_l^i||_1 \, p_{l-1}^i} \tag{2.23}$$

Fault classification can be then accomplished by a maximum a-posteriori criterion. Instead of simply taking the maximum of the probabilities, we further require that the maximum probability exceed a given threshold t_s near 1. This has the property of avoiding any fluctuations in the probabilities that can occur early in the process and increasing confidence in the predicted fault class.

The entire fault diagnosis system can be summarized as follows. Upon detection of a fault at time k_d, a bank of M BKFs are started in parallel, one for each of the fault models, and the following recursion is run:

1. Initialization Step: $p_0^m = P(C = m)$, for $i = 1, \ldots, m$.

For $k = 1, 2, \ldots$, do:

2. Posterior Update: Using the outputs $||\boldsymbol{\beta}_l^i||$ of the bank of BKFs, for $i = 1, \ldots, M$, compute the current model posterior probabilities as

$$p_l^m \;=\; \frac{||\boldsymbol{\beta}_l^m||_1 \, p_{l-1}^m}{\sum_{i=1}^{M} ||\boldsymbol{\beta}_l^i||_1 \, p_{l-1}^i} \tag{2.24}$$

3. Fault Classification: If $\max_{m=1,\ldots,M} p_l^m > t_s$, diagnose

$$m^* = \arg\max_{m=1,\ldots,M} p_l^m \,. \tag{2.25}$$

as the fault class and stop. Otherwise, wait for a new observation and go back to step 2.

Figure 2.3 displays a block diagram for one iteration of the proposed fault diagnosis system.

2.5 Performance Evaluation

Several metrics can be defined to assess the performance of fault detection and diagnosis methods; e.g. see [20]. Here, we define the error rates and average lag times that will be used in the next section to evaluate the performance of the fault detection and diagnosis system described previously.

Let K be the total length of the time series observations. Let $k_0 < K$ be the time a fault occurs, and let k_d be the time when a fault is first detected. We assume that once a fault is detected, the monitoring system issues an alarm and initiates the fault diagnosis step, and no further fault detection is run. There are two possibilities: either $k_d < k_0$ or $k_d \geq k_0$. In the first case, a *false alarm* has occurred. In the second case, we are interested in how soon the fault is detected, i.e., how small $k_d - k_0$ is. In a simulation setting, average values for these quantities can be computed. Assume that T time series of length K are generated and, without loss of generality, assume that $k_0 = K/2$ is a fixed time when the fault occurs for all time series. The *false detection rate* (FDR) is defined as

$$\text{FDR} = \frac{1}{T} \sum_{t=1}^{T} I_{k_d^t < k_0} \,, \tag{2.26}$$

where k_d^t is the time where a fault is detected for time series t, for $t = 1, \ldots, T$, while the

average time until correct detection (ATCD) is defined as

$$\text{ATCD} \;=\; \frac{1}{T_d}\sum_{t=1}^{T}(k_d^t - k_0)I_{k_d^t \geq k_0}\,, \tag{2.27}$$

where $T_d = \sum_{t=1}^{T} I_{k_d^t \geq k_0}$ is the number of times when a correct detection is made.

To assess the performance of fault diagnosis, we consider only the time series for which a correct fault detection is made (i.e., $k_d^t > k_0$). Let m_0 be the true identity of the fault and m_t^* be the fault model selected by the fault diagnosis procedure for time series t, for $t = 1, \ldots, T$. The *fault misdiagnosis rate* (FMR) is defined as

$$\text{FMR} \;=\; \frac{1}{T_d}\sum_{t=1}^{T_d} I_{m_t^* \neq m_0} I_{k_d^t \geq k_0}\,. \tag{2.28}$$

As before, it is interesting to evaluate the lag between the time a fault is detected and the fault diagnosis procedure begins and the time when a diagnosis is reached. Let k_s^t denote the latter quantity. The *average time until diagnosis* (ATD) is defined as

$$\text{ATD} \;=\; \frac{1}{T_d}\sum_{t=1}^{T}(k_s^t - k_d^t)I_{k_d^t \geq k_0}\,. \tag{2.29}$$

2.6 Numerical Experiments

In this section, we conduct a numerical experiment using the p53-MDM2 Boolean network discussed in Section II. We consider the input to be either no stress, *dna_dsb* $= 0$, or DNA damage, *dna_dsb* $= 1$. In addition, we adopt the "stuck-at" system fault models described in [12]. Briefly, transcript i is stuck at 0/1 if

$$X_{ki} \;=\; 0/1\,, \quad \text{for } k \geq k_0\,, \tag{2.30}$$

is one of the equations in the state transition model, where k_0 is the time when the fault occurs. Stuck-at fault models correspond to the loss of function of a gene due to mutation that either silences (stuck-at-0) or permanently activates (stuck-at-1) the gene. In this experiment, there are a total of 16 possible cases, corresponding to the presence or absence of a DNA damage signal, and each of the four genes *p53*, *MDM2*, *ATM*, or *WIP1* is stuck at 0 or 1.

For each of these fault models, we generated $T = 100$ time series of length $K = 400$, with the fault occurring at $k = 200$. Hence, for $k < k_0$ the system is under "normal" operation, and for $k \geq k_0$, (2.30) is introduced, for the appropriate transcript i. In addition, three different settings were used for the sequencing depth parameter s in (2.5) or (2.6), which are consistent with a total number of reads (for a typical RNA sample) in the ranges 1K-50K, 50K-100K, and 250K-300K, as reported in [25]. The range of 50K-100K is a typical range of read tags in SAGE experiments [25]. Table 2.1 summarizes the values of all the parameters used in the simulation (please refer to the previous sections for the meaning of each parameter). The proposed fault detection and diagnosis methodology was applied to each of the simulated time series, and the performance metrics described in the previous section were computed. We mention that varying the noise parameters around the values given in Table 2.1 did not produce any appreciable changes, so these results are not shown.

Figure 2.4 displays typical plots of the chi-square test statistic for the innovations filter, for the DNA-damage Boolean network ($dna\_dsb = 1$), 250K-300K reads, and two fault classes: WIP1 stuck-at-0, and MDM2 stuck-at-1. Based on $L = 15$ lags of a window with $L_D = 150$ observations, the threshold for a 95%-level test ($p < 0.05$) is 24.99. This value is represented as a dashed horizontal line in the Figure 2.4. We can observe in both cases that as soon as the fault occurs at $k_0 = 200$, the test statistic rises sharply and soon exceeds the horizontal line. A fault is detected after this has occurred at 6 out of the last 10 time

Table 2.1: Parameter values used in the numerical experiments.

Parameter	**Value**
Length of time series K	400
Fault time point k_0	200
State noise probability p	0.05
Sequencing depth s	1.0175 (1K-50K reads) 2.875 (50K-100K reads) 10.75 (250K-300K reads)
Baseline expression μ_b	0.01
Differential expression mean μ_δ	3
Differential expression variance σ_δ^2	0.5
Technical effects variance σ_θ^2	0.1
Detection window length L_D	150
Auto-correlation function lags L	15
Certification window length L_C	10
Certification threshold t_C	6
Diagnosis threshold t_s	0.8

points, which is the fault certification step. The average total time between the detection time k_d and k_0 is the AVCD performance metric discussed previously.

One can examine the operation of the fault diagnosis system by means of Figures 2.5 and 2.6, which display plots of the posterior probabilities p_l^m for each of the eight fault classes, in the case of the no-stress Boolean network ($dna\_dsb = 0$) and 250K-300K reads, under two assumed fault classes. All probabilities are zero before there is a detection, which occurs a little after the fault time $k_0 = 200$. The fault prior probabilities is assumed

to be uniform, so that $p_0^m = 1/8$ for all fault classes. The posterior probability for the correct fault model rises sharply from this value to 1, while the ones for the incorrect models decrease rapidly to zero. The case of ATM stuck at zero for the no-stress Boolean system, shown in Figure 2.6, is unique in that the procedure initially experiences some confusion among a few fault classes before settling on the correct one, which is reflected in longer average times until diagnosis (ATD).

Table 2.2 contains all the performance evaluation results of the numerical experiments. Fault diagnosis was perfect in all simulated time series, so that the fault misdiagnosis rate (FMR) defined in the previous section is identically zero and is omitted from the Table. One can observe that in nearly all cases the results improve substantially as the number of NGS reads increases, as expected. Some Boolean networks and transcripts can present more difficulties than others. For example, the ATM gene under a stuck-at-0 fault presents an elevated ATD, as mentioned in the previous paragraph and seen in Figure 2.6. This is because the no-stress system tends to settle in the "0000" singleton attractor state, making stuck-at-0 faults harder to detect and discriminate among transcripts. This effect can be seen clearly in Table 2.2: the stuck-at-0 faults tend to have larger ATCD and ATD than stuck-at-one faults in the no-stress network. This effect is not visible in the DNA-damage network, where the transcripts cycle on and off in the steady state (see the discussion in Section II). The performance of fault detection and diagnosis in the no-stress network is indeed generally worse than in the DNA-damage network.

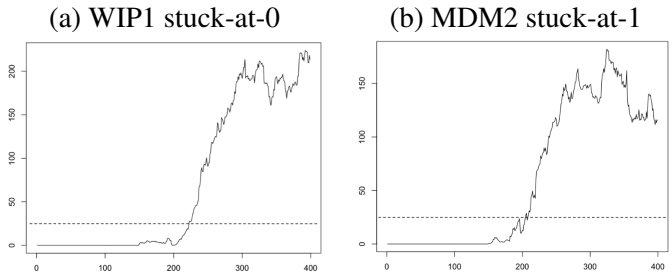

Figure 2.4: Fault detection results: chi-square test statistic for the innovations filter, for the DNA-damage Boolean network ($dna\_dsb = 1$), 250K-300K reads, and (a) WIP stuck-at-0 fault (b) MDM2 stuck-at-1 fault. The horizontal dashed line corresponds to the 95%-confidence detection threshold.

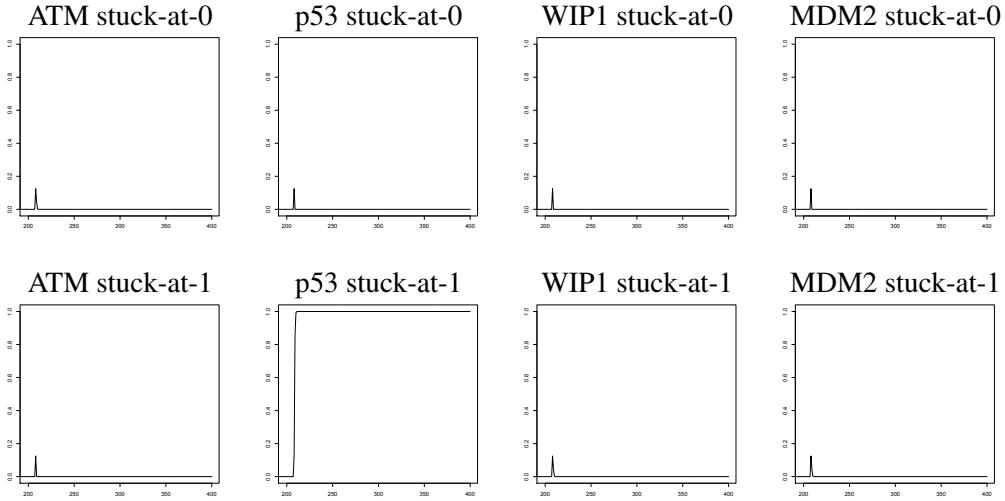

Figure 2.5: Fault diagnosis results: Posterior probabilities for each class fault when the true fault model is p53 stuck-at-1, for the no-stress Boolean network ($dna\_dsb = 0$) and 250K-300K reads.

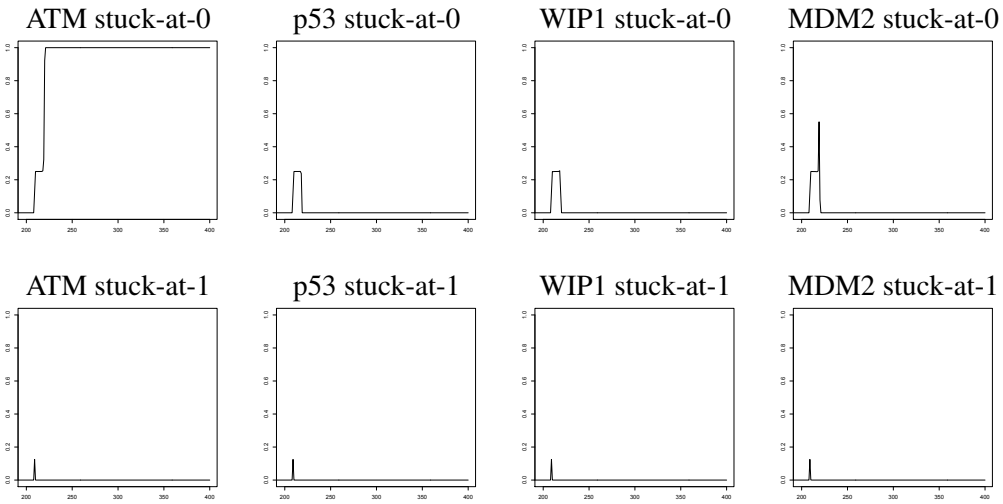

Figure 2.6: Fault diagnosis results: Posterior probabilities for each class fault when the true fault model is ATM stuck-at-0, for the no-stress Boolean network ($dna\_dsb = 0$) and 250K-300K reads.

Table 2.2: Performance evaluation results.

		No-stress			DNA-damage		
Fault class	Reads	FDR	ATCD	ATD	FDR	ATCD	ATD
ATM stuck-at-1	1K-50K	0.27	21.53	7.18	0.22	34.18	10.02
	50K-100K	0.08	13.38	2.75	0.15	15.26	2.88
	250K-300K	0.15	8.26	1.16	0.18	8.24	1.675
ATM stuck-at-0	1K-50K	0.29	28.41	31	0.26	41.33	8.82
	50K-100K	0.15	31.7	16.87	0.15	21.56	2.96
	250K-300K	0.12	18.1	10.66	0.24	16.4	1.346
P53 stuck-at-1	1K-50K	0.24	22.92	8.32	0.21	32.03	9.13
	50K-100K	0.2	13.29	2.39	0.13	15.34	2.4
	250K-300K	0.18	6.2	1.42	0.27	9.26	1.33
P53 stuck-at-0	1K-50K	0.24	53.75	11.65	0.24	27.32	9.14
	50K-100K	0.18	31.64	4.71	0.21	14.54	3.13
	250K-300K	0.15	21.81	3.75	0.26	7.87	1.04
WIP1 stuck-at-1	1K-50K	0.21	29.34	6.27	0.23	26.86	8.95
	50K-100K	0.2	15.66	2.56	0.21	16.16	3.05
	250K-300K	0.17	7.2	1.88	0.27	9.81	1.56
WIP1 stuck-at-0	1K-50K	0.16	33.92	10.08	0.21	25.01	9.29
	50K-100K	0.15	25.08	5.24	0.2	14.85	3.51
	250K-300K	0.09	19.83	2.75	0.31	9	1.32
MDM2 stuck-at-1	1K-50K	0.15	29	12.43	0.2	26.37	7.32
	50K-100K	0.16	18.47	2.72	0.16	17.43	2.59
	250K-300K	0.2	6.8	1.52	0.18	7.48	1.51
MDM2 stuck-at-0	1K-50K	0.27	35.53	7.23	0.29	38.4	7.84
	50K-100K	0.15	21.59	3.13	0.18	19.86	3.3
	250K-300K	0.13	15.78	3.01	0.29	9.98	1.32

3. A BAYESIAN APPROACH TO THE CLASSIFICATION OF MICROBIAL COMMUNITIES BASED ON R16S SEQUENCING DATA

3.1 Optimal Bayesian Classifier

The Optimal Bayesian Classifier (OBC) minimizes the expected error over the space of all classifiers under assumed forms of the class-conditional densities. Ordinary Bayes classifiers minimize the misclassification probability when the underlying distributions are known. However, Optimal Bayesian classification trains a classifier from data assuming the underlying distributions are not known exactly, but are rather part of an uncertainty class of distributions, each having a weight based on the prior and the observed data.

Let S_n be sample data consisting of metagenomic measurements on n individuals drawn independently from a mixture of two populations Π_0 and Π_1. Each measurement consists of an M-dimensional vector X of bacterial abundances, where M is the number of OTUs (features), and a label $Y \in \{0, 1\}$ identifying the population that the individual belongs to. Let c be the prior probability that an individual belongs to Π_0, and let the class conditional density $P_{\theta_y}(x|y)$ for population y be specified by a parameter vector θ_y, for $y = 0, 1$. It is assumed that c, θ_0, θ_1 are all independent prior to observing the data.

Denoting the prior for θ_y by $\pi(\theta_y)$, we have $\pi(\theta) = \pi(c)\pi(\theta_0)\pi(\theta_1)$. The posterior for c, denoted $\pi^*(c)$, is obtained from the number of sample points in each class using Bayes rule. It can be shown that the respective posterior distributions $\pi^*(c)$, $\pi^*(\theta_0)$, $\pi^*(\theta_1)$ remain independent.

$$\pi^*(\theta) = f(c, \theta_0, \theta_1|S_n) \tag{3.1}$$

$$= f(c|S_n, \theta_0, \theta_1)f(\theta_0|S_n, \theta_1)f(\theta_1|S_n). \tag{3.2}$$

As we assumed c is independent from sample values and distribution parameters and given n_y, $f(c|S_n, \theta_0, \theta_1) = f(c|n_y)$ and $f(\theta_y|S_n) = f(\theta_y|\{x_i^y\}_1^{n_y})$, hence,

$$\pi^*(\theta) = f(c|n_0)f(\theta_0|\{x_i^0\}_1^{n_0})f(\theta_1|\{x_i^1\}_1^{n_1}), \tag{3.3}$$

$$= \pi^*(c)\pi^*(\theta_0)\pi^*(\theta_1). \tag{3.4}$$

The Optimal Bayesian Classifier [42] is defined in terms of the posterior distributions as:

$$\psi_{OBC}(x) = \begin{cases} 0, & \text{if } E_{\pi^*}[c]f(x|0) \geq (1 - E_{\pi^*})f(x|1), \\ 1, & \text{otherwise,} \end{cases} \tag{3.5}$$

where

$$f(x|y) \;=\; \int_{\theta_y} P_{\theta_y}(x|y)\pi^*(\theta_y)d\theta_y \,. \tag{3.6}$$

3.1.1 Microbial Community Samples

We propose to model the r16S data for microbial abundance using a Dirichlet-Multinomial-Poisson framework, Figure 3.1. We assume that each M-dimensional abundance vector is multinomially distributed with probabilities that follow a Dirichlet distribution, in which case we have, for each class y,

$$P_{\theta_y}(x|y) = P(x|p, N_i) = \Gamma(N_i + 1)\Pi_{j=1}^{M}\frac{p_j^{x_{ij}}}{x_{ij}!}, \tag{3.7}$$

where x_{ij} is the observed abundance of OTU j in community sample i, where $i = 1, \cdots, n_y$ and $j = 1, \cdots, M$, p is the multinomial vector of probabilities, and $N_i = \sum_{j=1}^{M} x_{ij}$ is the total number of reads of community sample i, which is assumed to have a Poisson distribution with parameter λ, which in turn follows a Gamma distribution with

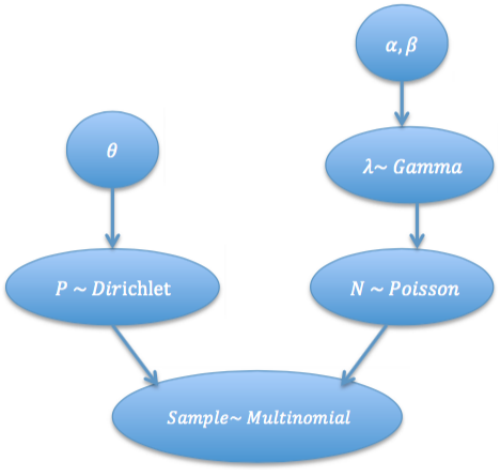

Figure 3.1: Poisson-Dirichlet Multinomial model for microbial community samples

parameters α, β. The posterior distributions for parameters of each class is defined by $\pi^*(\theta_y)$. Now,

$$\pi^*(\theta) \propto f(S_n|\theta)\pi(\theta), \tag{3.8}$$

$$= \pi(p|\alpha_d)\pi(N_i|\lambda)\pi(\lambda|\alpha,\beta)\Pi_{i=1}^{n_y} f(x_i|p_1,\cdots,p_M,N_i), \tag{3.9}$$

$$= \frac{\Gamma(\sum_{j=1}^{M}\acute{\alpha_j})}{\Pi_{j=1}^{M}\Gamma(\acute{\alpha_j})}\frac{(\beta+n_y)^{(\sum_{i=1}^{n_y}N_i+\alpha)}}{\Gamma(\sum_{i=1}^{n_y}N_i+\alpha)}e^{-\lambda(n_y+\beta)} \tag{3.10}$$

$$.\lambda^{\sum_{i=1}^{n_y}N_i+\alpha+1}\Pi_{j=1}^{M}p_j^{\acute{\alpha_j}-1}, \tag{3.11}$$

where $\acute{\alpha_j} = \alpha_{dj} + \sum_{i=1}^{n_y} x_{ij}$, $j = 1,\cdots,M$ and $i = 1,\cdots,n_y$.

Because of the discrete nature of microbial data equation (3.6) converts to $f(x|y) = \int_{\theta_y} P_{\theta_y}(x|y)\pi^*(\theta_y)d\theta_y$. We Plug the equations (3.8, 3.7) into Equation (3.6).

$$f(\mathbf{x}|y) = \int P_{\Theta_y}(\mathbf{x}|\theta_y)\pi^*(\theta_y)d\theta_y, \tag{3.12}$$

$$= \frac{1}{\Pi_{j=1}^{M}\Gamma(\mathbf{x}_j+1)} \frac{\Gamma(\sum_{j=1}^{M}\acute{\alpha_j})}{\Pi_{j=1}^{M}\Gamma(\acute{\alpha_j})} \frac{(\beta+n_y)^{(\sum_{i=1}^{n_y}N_i+\alpha)}}{\Gamma(\sum_{i=1}^{n_y}N_i+\alpha)} \tag{3.13}$$

$$\frac{\Pi_{j=1}^{M}\Gamma(\mathbf{x}_j+\acute{\alpha_j})}{\Gamma(N_{new}+\sum_{j=1}^{M}\acute{\alpha_j})} \frac{\Gamma(\sum_{i=1}^{n_y}N_i+\alpha+N_{new})}{(n_y+\beta+1)^{(\sum_{i=1}^{n_y}N_i+\alpha+N_{new})}} \tag{3.14}$$

where $N_{new} = \sum_{j=1}^{M}\mathbf{x}_j$ and $\mathbf{x}_j$ is new sample that we want to predict its label.

For simulated data because the total number of reads $N_i = \sum_{j=1}^{M}x_{ij}$ for each sample is constant so the equation (3.8) shrinks to:

$$f(\mathbf{x}|y) = \int P_{\Theta_y}(\mathbf{x}|\theta_y)\pi^*(\theta_y)d\theta_y, \tag{3.15}$$

$$= \frac{\Gamma(N_{new}+1)}{\Pi_{j=1}^{M}\Gamma(\mathbf{x}_j+1)} \frac{\Gamma(\sum_{j=1}^{M}\acute{\alpha_j})}{\Pi_{j=1}^{M}\Gamma(\acute{\alpha_j})} \frac{\Pi_{j=1}^{M}\Gamma(\mathbf{x}_j+\acute{\alpha_j})}{\Gamma(N_{new}+\sum_{j=1}^{M}\acute{\alpha_j})} \tag{3.16}$$

We assume that the parameter c is beta distributed with hyper parameters β_0, β_1 , independently of the parameters θ_y (prior to observing the data). It can be shown that the posterior distribution $\pi^*(c)$ is also beta with hyper parameters β_0+n_0 and β_1+n_1,

$$\pi^*(c) = \frac{c^{\beta^0+n_0-1}(1-c)^{\beta^1+n_1-1}}{B(\beta^0+n_0,\beta^1+n_1)} \tag{3.17}$$

in which case $E_{\pi^*}[c] = \frac{n_0+\beta_0}{n+\beta_0+\beta_1}$. In problems with more than two classes, the distribution of parameter c is assumed to be Dirichlet. This completes the specification of the Bayesian classifier in (3.5).

3.2 Prior Construction

To improve classification performance, we need to find the hyper parameter values that make the posterior function closest to the model of interest, Figure 3.2. For constructing the priors, we minimize the mean log likelihood

$$E(\log(f(x|\theta))) = \int_{x\in\chi} f(x|\theta_{true}) \log f(x|\theta)\, dx\,. \tag{3.18}$$

We split the available data, using half of it to construct the prior and half to obtain the posterior probabilities and thus the Bayesian classifier.

$$\min_{\pi(\theta)\in\Pi} -E_\theta[\ell_{n_p}(\theta)]. \tag{3.19}$$

where $\ell_{n_p}(\theta) = \frac{1}{n_p}\ell(\theta; S_{n_p}^{prior})$ is the log-likelihood of the data distribution. $\ell_{n_p}(\theta)$ can be interpreted as a measure of similarity between the true model and the model governed by θ. For the metagenomic data, we have that:

$$\ell_{n_p} = \frac{1}{n_p}\log P_{\theta_y}(x|y), \tag{3.20}$$

$$= \frac{1}{n_p}\log(\Pi_{i=1}^{n_p}\Gamma(N_i+1)\Pi_{j=1}^{M}\frac{p_j^{x_{ij}}}{x_{ij}!}) \tag{3.21}$$

$$= \frac{1}{n_p}\sum_{i=1}^{n_p}\sum_{j=1}^{M}\log(\Gamma(N_i+1)) + [x_{ij}\log(p_j) - \log(x_{ij}!)]. \tag{3.22}$$

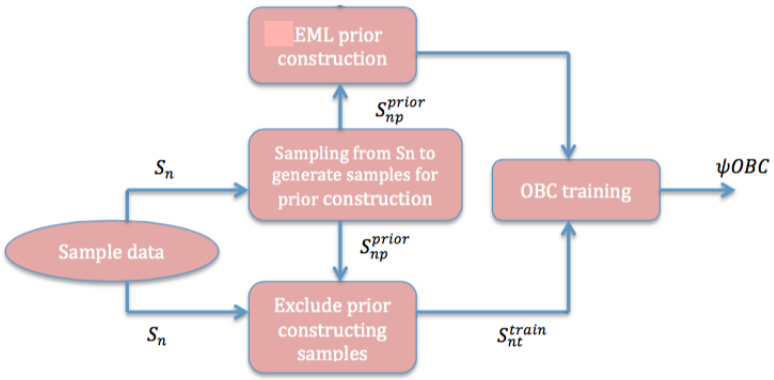

Figure 3.2: An illustrative view of general approach

where n_p denotes number of samples that we use for constructing the priors for each class. The expectation of the log-likelihood is:

$$E_\theta[\ell_{n_p}(\theta)] = \frac{1}{n_p}\sum_{i=1}^{n_p}\sum_{j=1}^{M} E_\theta(\log\Gamma(N_i+1))+ \tag{3.23}$$

$$[x_{ij}\psi(\alpha_j) - \psi(\hat{\alpha})) - \log x_{ij}] \tag{3.24}$$

where $\psi(\alpha) = \frac{d}{d\alpha}\log\Gamma(\alpha)$ is the digamma function.

3.3 Results

In this section, we demonstrate the efficacy of the proposed approach, by comparing its performance against that of the kernel SVM, Random Forest and MetaPhyl classification algorithms, as a function of varying sample size, classification difficulty, and number of OTUs (features). The numerical experiments are based on both synthetic data and real

metagenomic data sets.

3.3.1 Synthetic Data Results

Synthetic OTU abundance data and phylogeny trees were generated using the strategy proposed in [43], which considers the common phylogenetic tree T that relates OTUs in all the 16S rDNA samples. To generate samples for a class k, the tree is traversed systematically, deciding for each internal node v what fraction of species would come from each of the subtrees rooted at the child nodes of v.

Two parameters are assigned to each node v for each class k. Let μ_v^k denote the average proportion of species that correspond to the subtree rooted at the left child node of v in the k-th class, and let $(\sigma_v^k)^2$ denote the variance of this proportion within the class. New class samples are generated according the proportions of species at each node v and the normal distributions $N(\mu_v^k, (\sigma_v^k)^2)$. The parameters values μ_v^k are in turn sampled from the normal distribution $N(\tilde{\mu_v}, \tilde{\sigma_v}^2)$, where $\tilde{\sigma_v}^2$ characterizes the variance between the classes, while $\tilde{\mu_v}$ are base values that are initialized randomly.

The within- and between-class variances can be controlled by using the parameters σ_v^2 and $\tilde{\sigma_v}^2$, respectively. The exact values of $\tilde{\sigma_v}^2$ and σ_v^2 are sampled at each tree node v according to $N(0, \tilde{\lambda} d(v))$ and $N(0, \lambda d(v))$, where $d(v)$ is the distance between v and the tree root. Note that the parameters $\tilde{\lambda}$ and λ influence the difficulty of the classification problem, which is proportional to λ and inversely proportional to $\tilde{\lambda}$.

We consider three sample sizes for the training data, $n = 30, 50, 70$, with class prior probability $c = 0.5$. The sample sizes n_0 and n_1 are determined according to the class prior probability as $n_0 = n_1 = n/2$. Classifier accuracy is obtained by testing each designed classifier on a large synthetic test data set, and averaging the results over a large number of iterations using different synthetic training data sets.

Figures 3.3 and 3.4 compare the performance of our proposed Bayesian classifier,

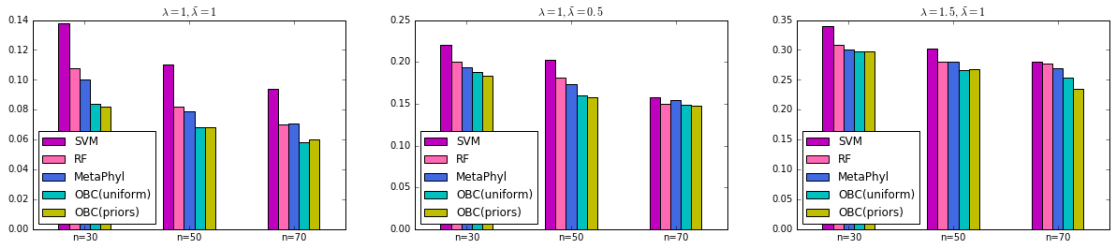

Figure 3.3: Comparison of SVM, RF, MetaPhyl and Optimal Bayesian classifier with uniform or constructed priors, on simulated datasets for varying sample size and within- and between-class variances. Each sample contains 128 OTUs.

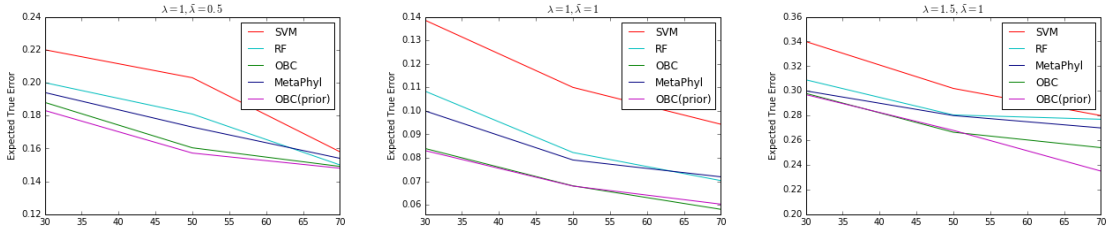

Figure 3.4: Comparison with SVM, RF, MetaPhyl and Optimal Bayesian classifier on simulated datasets for varying sample size and within- and between-class variances, each sample contains 128 OTUs.

using constructed priors, against that of kernel SVM, RF and MetaPhyl classifiers and Optimal bayesian classifier with uniform priors. We observe that the kernel SVM classifier clearly performs the worst, probably due to the highly nonlinear nature of the data, the RF and Metaphyl classifier have comparable performance overall, while the proposed OBC classifier clearly exhibits the best performance over different values for between and within class variances that defines the difficulty of the classification. As the sample size increases, classification performance improves for all classification rules, as expected.

3.3.2 Real Data Results

We considered four different data sets:

1. **Hullar et al.** [47]. Using optimized methods for fecal bacterial DNA processing and microbial community analysis, the composition of the gut microbial community as a contributor of human lignan exposure was assessed in this data set. Microbial diversity in stool was assessed using pyrosequencing of the 16S rRNA gene. Stool samples were collected from participants (n=35) in an ongoing randomized, double-blind crossover intervention of flaxseed lignan extract and placebo [48]. Each intervention lasted 60 days interspersed with washout periods of at least 60 days.

2. **Costello et al. Body Habitats (CBH)** [49]. These data included sample communities from six major categories of habitat: External Auditory Canal (EAC), Gut, Hair, Nostril, Oral cavity, and Skin. This data set is an example of a relatively easy classification task due to the generally pronounced differences between the communities.

3. **Fierer et al. Subject (FS)** [50]. This data set contains all samples from the "keyboard" data set, for which at least 397 raw sequences were recovered[51]. The class labels are the anonymized identities of the three experimental subjects. This classification task is the easiest of all four data sets because of the clear distinctions between the individuals, the fact that all of the samples come from approximately the same time point, and the large number of training samples available for each class.

4. **Fierer et al. Subject Hand (FSH)** [52]. This data set is a more challenging version of the previous ones. The class labels are the concatenation of the experimental subject identities and the label of which hand (left vs. right) the sample came from on that individual. There were three subjects, and so there are six classes in this dataset.

5. **Turnbaugh et al. Twin Gut** [53]. This data set contains gut samples from lean,

Table 3.1: Summary of data sets

Data set	Training samples	Test samples	Number of OTUs	Number of classes
CBH	415	207	2741	6
FS	68	33	565	3
FSH	68	33	565	6
Meredith	50	30	100	2
Twin Gut	170	111	462	3

Table 3.2: Performance of classifiers on the real data sets

Method	CBH	FS	FSH	Meredith	Twin Gut
RF	0.09	**0**	0.33	0.13	0.21
SVM	0.13	0.08	0.37	0.15	0.24
OBC	0.11	0.03	0.29	0.12	**0.18**
OBC(constructed priors)	**0.075**	0.01	**0.28**	**0.11**	0.19
MetaPhyl	0.1	0.05	0.328	-	0.19

obese and overweight subjects.This data set is a challenging classification task because the classes correspond to microbial communities from the same body habitat and thus are very similar.

The data sets are summarized in Table 3.1. The performance of OBC with constructed priors on real data is comparable but actually better than its performance on synthetic data, the results are summarized in Table 3.2 and Figure 3.5. It shows that our Poisson-Dirichlet-Multinomial assumption on the distribution of OTUs is valid.

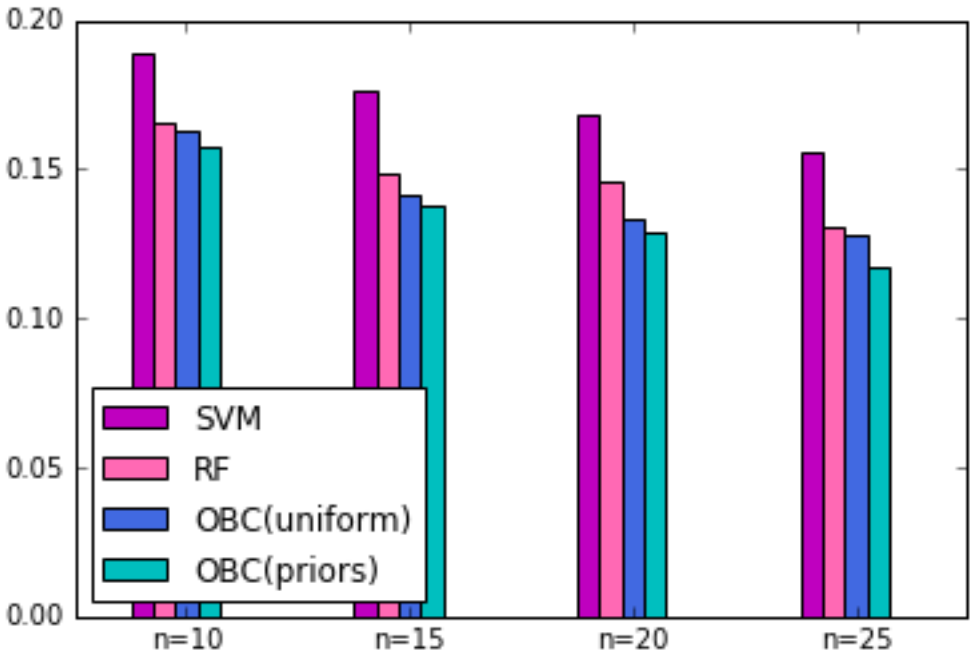

Figure 3.5: Comparison between SVM, RF, OBC with uniform priors and OBC with constructed priors for **Hullar et al.** data set

4. SUMMARY AND CONCLUSIONS

4.1 Conclusions

In this thesis, we proposed a novel methodology for model-based fault detection and diagnosis for stochastic Boolean dynamical systems based on the optimal state estimator, namely, the Boolean Kalman Filter. This model-based approach is applied here to an observation model for Next Generation Sequencing transcriptomic data. The fault detection consists of an innovations filter followed by a fault certification step, and requires no knowledge about the system faults. In the presence of knowledge about the system faults, a multiple model adaptive estimation (MMAE) procedure for fault diagnosis is proposed, which employs a bank of BKFs running in parallel. The efficacy of the proposed methodology was demonstrated via numerical experiments using the p53-MDM2 negative feedback loop network with stuck-at faults. The results indicate the the proposed method is promising in monitoring biological changes at the transcriptomic level.

In this thesis, we presented a model-based Bayesian framework for the classification of metagenomic microbial abundance data. This approach was contrasted to non-parametric classification rules such as kernel SVMs and Random Forests, the latter being considered the state of the art in metagenomics classification. We also compared performance to the recently published Metaphyl algorithm, which was designed with metagenoas an estimate of the Kullback-Leibler information quantity [54]. The proposed classifier outperformed all of those algorithms on our synthetic data sets.

4.2 Further Study

Future work on the FDI, as the number of potential fault models in gene network is finite, we will use likelihood based fault detection method along with the bank of particle filters on the gene network. Likelihood based fault detection compares the two models of

the network, before and after change by using the likelihood ratio. The typical behavior of the log-likelihood ratio shows a negative drift before change, and a positive drift after change. For large numbers of state variables in gene network, the computation of the BKF becomes impractical and computationally expensive. We will use sequential importance sampling, a basic and most popular approach for performing this approximation.

Future work on microbial classification study, we will consider the extension to informative priors using phylogeny trees and other prior information in order to boost accuracy of the classifier further, particularly in small-sample situations.

REFERENCES

[1] R. Layek and A. Datta, "From biological pathways to regulatory networks," *Molecular Biosystems*, vol. 7, pp. 843–851, 2011.

[2] S. Kauffman, "Metabolic stability and epigenesis in randomly constructed genetic nets," *Journal of Theoretical Biology*, vol. 22, pp. 437–467, 1969.

[3] S. Kauffman, "Homeostasis and differentiation in random genetic control networks," *Nature*, vol. 224, pp. 177–178, 1969.

[4] S. Kauffman, *The Origins of Order: Self-Organization and Selection in Evolution*. No. 1, Oxford University Press, 1st ed., 1993.

[5] S. Bornholdt, "Boolean network models of cellular regulation: prospects and limitations," *The Royal Society Interface*, vol. 5, pp. 85–94, 2008.

[6] R. Albert and H. Othmer, "The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in drosophila melanogaster," *Journal of Theoretical Biology*, vol. 223, pp. 1–18, 2003.

[7] F. Li, T. Long, Y. Lu, and C. Tang, "The yeast cell- cycle network is robustly designed," *Proceedings of the National Academy of Sciences*, vol. 101, no. 14, pp. 105–110, 2004.

[8] A. Faure, A. Naldi, C. Chaouiya, and D. Thieffry, "Dynamical analysis of a generic boolean model for the control of the mammalian cell cycle," *Bionformatics*, vol. 22, no. 14, pp. 245–256, 2006.

[9] I. Schmulevich, E. Dougherty, and W. Zhang, "From boolean to probabilistic boolean networks as models of genetic regulatory networks," *Proceedings of the IEEE*, vol. 90, pp. 1778–1792, 2002.

[10] I.Shmulevich and E.Dougherty, *Probabilistic Boolean Networks*. Society for Industrial and Applied Mathematics, 1st ed., 2009.

[11] U. Braga-Neto, "Optimal state estimation for boolean dynamical systems," in *45th Annual Asilomar Conference on Signals, Systems, and Computers*, vol. 10, pp. 114–119, 2011.

[12] R. Layek and A. Datta, "Fault detection and intervention in biological feedback networks," *Journal of Biological Systems*, vol. 20, pp. 441–453, 2012.

[13] E. Batchelor, A. Loewer, and G. Lahav, "The ups and downs of p53: understanding protein dynamics in single cells," *Nature Reviews Cancer,*, vol. 9, pp. 326–332, 2009.

[14] R. Weinberg, *The Biology of Cancer*. Princeton, 2nd ed., 2006.

[15] D. Himmelblau, *Fault Detection and Diagnosis in Chemical and Petrochemical Processes*. Elsevier, 1st ed., 1978.

[16] J.Gertler, *Fault Detection and Diagnosis in Engineering Systems*. Marcel Dekker, 2nd ed., 1998.

[17] J. Chen and R. Patton, "Robust model-based fault diagnosis for dynamic systems," 1999.

[18] J. Prakash, S. Patwardhan, and S. Narasimhan, "A supervisory approach to fault-tolerant control of linear multivariable systems," *Industrial and Engineering Chemical Research*, vol. 41, pp. 2270–2281, 2002.

[19] R. Isermann, *Fault-Diagnosis Systems*. Springer, 3rd ed., 2006.

[20] K. Villeza, B. Srinivasanb, R. Rengaswamyb, S. Narasimhanc, and V. Venkatasubramanian, "Kalman-based strategies for fault detection and identification (fdi): Extensions and critical evaluation for a buffer tank system," *Computers and Chemical Engineering*, vol. 35, pp. 806–816, 2010.

[21] L. Chiang, E. Russell, and R. Braatz, *Fault detection and diagnosis in industrial systems*. Springer, 1st ed., 2013.

[22] A. Gelb, *Applied optimal estimation*. MIT Press, 1st ed., 1974.

[23] A. Bahadorinejad and U. Braga-Neto, "Optimal fault detection in stochastic boolean regulatory networks," *GENSIPS*, vol. 3, pp. 115–119, 2014.

[24] P. Maybeck and P. Hanlon, "Performance enhancement of a multiple model adaptive estimator,," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 31, pp. 1240–1253, 1995.

[25] N.Ghaffari, M.Yousefi, C.Johnson, and E.Dougherty, "Modeling the next generation sequencing sample processing pipeline for the purpose of classification," *BMC Bioinformatics*, vol. 14, pp. 307–317, 2013.

[26] A.Mortazavi, B.Williams, K.McCue, and B.Wold, "Mapping and quantifying mammalian transcriptomes by rna- seq," *Nature Methods*, vol. 5, no. 7, 2008.

[27] S. Marguerat and J. Bahler, "Rna-seq: from technology to biology," 2010.

[28] J. Li, D. Witten, I. Johnstone, and R. Tibshirani, "Normalization, testing, and false discovery rate estimation for rna-sequencing data," *Cellular and Molecular Life Science*, vol. 13, no. 3, pp. 523–538, 2012.

[29] M. Imani and U. Braga-Neto, "Optimal state estimation for boolean dynamical systems using a boolean kalman smoother," in *Signal and Information Processing (GlobalSIP), 2015 IEEE Global Conference on*, pp. 972–976, IEEE, 2015.

[30] M. Imani and U. Braga-Neto, "Particle filters for partially-observed boolean dynamical systems," *arXiv preprint arXiv:1702.07269*, 2017.

[31] L. D. McClenny, M. Imani, and U. M. Braga-Neto, "Boolean kalman filter with correlated observation noise," *42nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, vol. 11, no. 426-430, 2017.

[32] M. Imani and U. M. Braga-Neto, "Maximum-likelihood adaptive filter for partially observed boolean dynamical systems," *IEEE Transactions on Signal Processing*, vol. 65, no. 2, pp. 359–371, 2017.

[33] M. Imani and U. Braga-Neto, "Optimal gene regulatory network inference using the boolean kalman filter and multiple model adaptive estimation," in *Signals, Systems and Computers, 2015 49th Asilomar Conference on*, pp. 423–427, IEEE, 2015.

[34] M. Imani and U. Braga-Neto, "Point-based value iteration for partially-observed boolean dynamical systems with finite observation space," in *Decision and Control (CDC), 2016 IEEE 55th Conference on*, pp. 4208–4213, IEEE, 2016.

[35] M. Imani and U. Braga-Neto, "State-feedback control of partially-observed boolean dynamical systems using rna-seq time series data," in *American Control Conference (ACC), 2016*, pp. 227–232, IEEE, 2016.

[36] M. Imani and U. Braga-Neto, "Control of gene regulatory networks with noisy measurements and uncertain inputs," *arXiv preprint arXiv:1702.07652*, 2017.

[37] M. Imani and U. Braga-Neto, "Multiple model adaptive controller for partially-observed boolean dynamical systems," in *2017 Ammerical Control Conference*, IEEE, 2017.

[38] L. McClenny, M. Imani, and U. Braga-Neto, "Boolfilter package vignette," 2017.

[39] S. Marguerat and J. Bahler, "Rna-seq: from technology to biology," *Cellular and Molecular Life Science*, vol. 67, no. 4, pp. 569–579, 2010.

[40] W. Streit, R. Schmitz, and T. Jiang, "Metagenomics–the key to the uncultured microbes.," *Curr Opin Microbiol*, vol. 20, no. 7, pp. 49–498, 2004.

[41] Q. Wang, G. Garrity, J. Tiedje, and J. Cole, "Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy," *Appl Environ Microbiol*, vol. 11, no. 73, pp. 5261–5267, 2007.

[42] L. Dalton and E. R. Dougherty, "Optimal classifiers with minimum expected error within a bayesian framework part i: Discrete and gaussian models," *Pattern Recognition*, vol. 46, no. 16, pp. 1301–1314, 2013.

[43] O. Tanaseichuk, J. Borneman, and T. Jiang, "Phylogeny-based classification of microbial communities," *Bioinformatics*, vol. 20, no. 6, pp. 924–930, 2013.

[44] E. Batchelor, C. Mock, I. Bhan, and G. Lahav, "Recurrent initiation: a mechanism for triggering p53 pulses in response to dna damage," *Molecular Cell,*, vol. 30, no. 3, pp. 277–289, 2008.

[45] I. Schmulevich, E. Dougherty, S. Kim, and W. Zhang, "Probabilistic boolean networks: a rule-based uncertainty model for gene-regulatory networks," *Bioinformatics*, vol. 18, no. 261–274, 2002.

[46] S. Anders and H. Huber, "Differential expression analysis for sequence count data," *Genome Biology*, vol. 11, pp. 106–115, 2010.

[47] M. Hullar, S. Lancaster, F. Li, E. Tseng, K. Beer, C. Atkinson, K. Wahala, W. Copeland, T. Randolph, K. Newton, and J. Lampe, "Enterolignan-producing phenotypes are associated with increased gut microbial diversity and altered composition in premenopausal women in the united states," *Cancer Epidemiol Biomarkers Prev*, vol. 24, no. 9, pp. 546–554, 2014.

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

[49] Ostello, Lauber, Hamady, Fierer, Jeffrey, Gordon, and Knight, "Bacterial community variation in human body habitats across space and time," *Science*, vol. 18, pp. 1694–1697, 2009.

[50] Fierer, Lauber, Zhou, McDonald, Costello, and R. Knight, "Forensic identification using skin bacterial communities," *Proceedings of the National Academy of Sciences of the United States of America, PNAS*, vol. 107, no. 14, pp. 6477–6481, 2010.

[51] D. Knights, E. K. Costello, and R. Knight, "Supervised classification of human microbiota," *Federation of European Microbiological Societies*, vol. 35, no. 2, pp. 343–359, 2010.

[52] Fierer, Hamady, Lauber, and Knight, "The influence of sex, handedness, and washing on the diversity of hand surface bacteria," *Proceedings of the National Academy of Sciences of the United States of America,PNAS*, vol. 105, no. 46, pp. 17994–17999, 2008.

[53] P. Turnbaugh, M. Hamady, T. Yatsunenko, B. Cantarel, A. Duncan, R. Ley, M. L. Sogin, W. J. Jones, B. A. Roe, J. P. Affourtit, M. Egholm, B. Henrissat, A. C. Heath, R. Knight, and J. I. Gordon, "A core gut microbiome in obese and lean twins," *Nature*, vol. 457, pp. 480–484, January 2009.

[54] H. Akaike, "Information theory and an extension of the maximum likelihood principle," *2nd International Symposium on Information*, vol. 1, pp. 199–213, 1973.