BAYESIAN MODELS FOR GENE REGULATORY NETWORKS APPLIED TO CANCER TISSUES

A Dissertation

by

ANWOY KUMAR MOHANTY

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,	Aniruddha Datta
Committee Members,	Vijayanagaram Venkatraj
	Yang Shen
	P. R. Kumar
Head of Department,	Miroslav M. Begovic

August 2015

Major Subject: Electrical Engineering

ABSTRACT

Cellular behavior is controlled through multivariate interactions between various biological molecules such as proteins and DNA. Various methods have previously been proposed to model such interactions. However many of these methods require large volumes of data to effectively estimate the associated unknown parameters. In this work we explore the use of Bayesian methods to exploit the prior knowledge about pathway information in combination with collected data in order to make accurate and useful inferences about tissue level behavior. These predictions would in turn help in the discovery of better therapeutic strategies such as the development of better combination therapies involving kinase inhibiting drugs. Various problems of modeling cancerous and healthy tissues from a Bayesian perspective have been addressed in this work. We give a short description of these problems here in this section.

An important problem in the study of cancer is the understanding of the heterogeneous nature of the cell population. The clonal evolution of the tumor cells results in the tumors being composed of multiple sub-populations. Each sub-population reacts differently to any given therapy. This calls for the development of novel (regulatory network) models, which can accommodate heterogeneity in cancerous tissues. Here we present a new approach to model heterogeneity in cancer. We model heterogeneity as an ensemble of deterministic Boolean networks based on prior pathway knowledge. We develop the model considering the use of qPCR data. By observing gene expressions when the tissue is subjected to various stimuli, the compositional breakup of the tissue under study can be determined. We demonstrate the viability of this approach by using our model on synthetic data, and real world data collected

from fibroblasts.

Another problem which is addressed in this work is the determination of locations of dysregulations in a Boolean network used to model signal transduction networks. Knowledge about which proteins/genes are dysregulated in a regulatory network, such as in the Mitogen Activated Protein Kinase (MAPK) Network, can be used not only to decide upon which therapy to use for a particular case of cancer, but also help in discovering effective targets for new drugs. The posterior inference problem is solved using a version of the message passing algorithm. We have done simulation experiments on synthetic data to verify the efficacy of the algorithm as compared to the results from the much more computationally intensive Markov Chain Monte-Carlo methods. We also applied the model to analyze data collected from fibroblasts, thereby demonstrating how this model can be used on real world data.

Another important issue in Bayesian computation is that the processing of the collected data must be done as efficiently as possible in terms of computational speed and memory requirements. The use of Markov Chain Monte Carlo methods is time consuming and hence other methods need to be used for the analysis. The use of conjugate exponential models is investigated in the modeling of the heterogeneity of cancerous tissues where variational methods could be used in a straightforward manner. Variational algorithms, which allow for the fast computations of posterior probability distributions of variables of interest, have been used in the inference of the compositional breakup of the heterogeneous tissue under study. The efficacy of these methods has been demonstrated by comparing them with other methods such as Markov chain Monte Carlo and Expectation maximization.

DEDICATION

To my family

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my parents, Pradyot Mohanty and Sulagna Mohanty (now deceased, greatly missed), for their unconditional love and support. Without them, I would never have been able to achieve my PhD degree.

I would like to express my deepest gratitude to my advisor, Dr. Aniruddha Datta, who made my dissertation possible. I thank him for his steady guidance and for generously supporting me financially through my doctorate program at Texas A&M University. The great experiences of working with him will definitely benefit the rest of my career. I thank Dr. Vijayanagaram Venkatraj, Dr. Yang Shen and Dr. P. R. Kumar, for serving on my committee and for their support and useful suggestions for the improvement of this work. I thank my fellow students , both current and graduated, especially Bibhu Prasad Mishra, Dr. Sriram Sridharan, Priyadharshini Venkat, Osama Arshad, and Dr. Ritwik Layek for friendship and memories.

I would like to thank all the faculty and staff, both current and former, at Texas A&M University, especially for the administrative support provided by Ms. Tammy Carda, Ms. Melissa Sheldon, Ms. Jeanie Marshall, Ms. Claudia Samford and Ms. Anni Brunker.

NOMENCLATURE

ATP	Adenosine triphosphate
RNA	Ribonucleic acid
mRNA	Messenger RNA
MAPK	Mitogen activated protein kinase
qPCR	Quantitative real-time polymerase chain reaction
MCMC	Markov chain monte carlo
MH	Metropolis-Hastings
FBS	Fetal bovine serum
ARACNE	Algorithm for the reconstruction of accurate cellular networks

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION *

1.1 Background

Bayesian methods are getting more and more popular in the statistics and machine learning community as the community is finding more and more use of this approach to solve various problems in science and engineering. In this work we have used Bayesian methods in conjunction with other methods (such as Boolean algebra) in the modeling of cancerous tissues. Such modeling of cancerous tissues will help in the discovery of better therapeutic strategies such as the development of better combination therapies involving kinase inhibiting drugs. Various problems of modeling cancerous and healthy tissues from a Bayesian perspective have been addressed in this work. This thesis has three primary sections. A short introduction for the following sections is given below.

1.2 Organization

In section 2, we deal with the modeling of the heterogeneity of cancer tissues. We have modeled the heterogeneity in cancerous tissues as a collection of Boolean networks. Prior knowledge about locations of various common mutations occurring in cancer tissues can be encoded as stuck-at faults in the Boolean networks. By observing gene expressions when the tissue is subjected to various stimuli, the compositional breakup of the tissue under study can be determined. A multilevel

*Parts of this section are reprinted with permission from "A Model for Cancer Tissue Heterogeneity" by A. K. Mohanty, A. Datta, and V. Venkatraj, 2013. *IEEE Transactions on Biomedical Engineering*, volume 61, no. 3, pages 966 - 974, © 2013 IEEE. doi:10.1109/TBME.2013.2294469, and "Using the message passing algorithm on discrete data to detect faults in boolean regulatory networks" by A. K. Mohanty, A. Datta, and V. Venkatraj, 2014. *BMC Algorithms for Molecular Biology*, volume 9, no. 20, 12 pages. doi:10.1186/s13015-014-0020-6, and "A Conjugate Exponential Model for Cancer Tissue Heterogeneity" by A. K. Mohanty, A. Datta, and V. Venkatraj, 2015. *IEEE Journal of Biomedical and Health Informatics*, preprint, © 2015 IEEE. doi:10.1109/JBHI.2015.2410279.

hierarchical model was used to account for the stochasticity in the observed data as well as the variations among the various gene expressions. We demonstrate the viability of this approach by using our model on synthetic data, and real world data collected from fibroblasts.

Section 3 deals with the Bayesian estimation of possible locations of dysregulations in a given Boolean network provided we have certain observed data from the tissue under study. If we have a Boolean network used to model a signal transduction network such as the Mitogen Activated Protein Kinase (MAPK) Network, estimating these possible locations of dysregulations in the network can prove to be useful in not only deciding which therapy to use for a particular case of cancer, but also help in discovering effective targets for new drugs. The posterior inference problem is solved using a version of the message passing algorithm. We have done simulation experiments on synthetic data to verify the efficacy of the algorithm as compared to the results from the much more computationally intensive Markov Chain Monte-Carlo methods. We also applied the model to analyze data collected from fibroblasts, thereby demonstrating how this model can be used on real world data.

In section 4, we have investigated the use of variational Bayesian methods in the computation of posterior marginal distributions of the unobserved variables in a probability model and applied these methods to the modeling of heterogeneity of cancer tissues. The use of conjugate exponential models is investigated in the modeling of the heterogeneity of cancerous tissues where variational methods could be used in a straightforward manner. Variational algorithms, which allow for the fast computations of posterior probability distributions of variables of interest, have been used in the inference of the compositional breakup of the heterogeneous tissue under study. The efficacy of these methods has been demonstrated by comparing them with other methods such as Markov chain Monte Carlo and Expectation maximization.

2. A BAYESIAN MODEL FOR CANCER TISSUE HETEROGENEITY *

2.1 Introduction

Cancer progression can be modeled as evolution among cells which become neoplastic due to the accumulation of mutations which give them a proliferative advantage over their normal neighbours [27]. Although there is wide spread consensus that most macroscopic tumors have a unicellular origin as described in [27, 37], stepwise accumulation of mutations as described in [27] causes the appearance of variant sublines which makes the neoplastic cell population a heterogeneous one. The heterogeneity of cancer cell populations raises certain issues in the treatment strategy to be followed because a certain treatment which may be effective on a certain subpopulation of the neoplastic cells but not on the others may show good results on a particular patient, but not on another patient where the sensitive neoplastic cell subpopulation is not a major fraction of the entire cancerous cell population. Hence estimating the proportion wise breakup of the cell subpopulations in a cancer for any given patient is a problem which needs to be addressed. Once the dominant subpopulations have been identified, the appropriate decisions regarding therapy can be taken such as which subpopulation to target and how much of therapy should be administered to the patient. Proponents of the cancer stem cell theory [30, 1] say that the growth and progression of many cancers are driven by small subpopulations of cancer stem cells and that therapies should be designed to target these stem cell subpopulations. The second popular theory is that most of the cells in the tumor are contributive to tumor maintenance [36, 5]. Such a view would imply that ther-

*Parts of this section are reprinted with permission from "A Model for Cancer Tissue Heterogeneity" by A. K. Mohanty, A. Datta, and V. Venkatraj, 2013. *IEEE Transactions on Biomedical Engineering*, volume 61, no. 3, pages 966 - 974, © 2013 IEEE. doi:10.1109/TBME.2013.2294469.

apies should be aimed to target all the major subpopulations in the cancer tissue. Whichever model may be closer to the true state of affairs, a mathematical model which incorporates heterogeneity in the cancer tissue is a vital tool in the treatment of a complex disease such as cancer.

2.2 Model description

Cellular behavior is controlled through multivariate interactions between various biological molecules such as proteins and DNA [37, 9]. Various methods have been proposed to model such interactions. These include differential equations [4], deterministic and probabilistic Boolean networks [34, 9], and Bayesian and dynamic Bayesian networks [11, 42]. For methods such as the probabilistic Boolean networks, the network parameters are very difficult to learn from real world data simply due to the huge search space for the parameters. The REVEAL algorithm [21] is a general method to learn deterministic Boolean networks from time domain data. However time domain data is difficult to collect. In addition, a lot of the previous methods rely on the discretization of real world observations such as gene expression levels, which results in the loss of valuable information. The ARACNE method [23] is a way to use continuous valued observations to determine regulatory interactions.

In the biological literature, there is a wealth of information regarding the marginal regulatory interactions, usually referred to as pathway knowledge, which has been collected by biologists over a long period of time. Unfortunately most genetic regulatory network modeling methodologies tend to ignore this information. Using this information would result in methodologies which describe cellular behavior more accurately. A method to use such prior pathway knowledge while designing networks was presented in [19]. Here Boolean networks, which are extensively used in digital logic design, were used to model signal transduction networks. Boolean networks,

which involve discrete variables, are a good choice to model protein-protein interaction networks since such reactions involve proteins changing from one state to another, usually by the addition or removal of phosphate groups, and are generally accompanied by ATP hydrolysis which pushes the reactions to completion. When such a signal transduction network contains transcription factors, then the Boolean model can be used to model the behavior of the genes whose mRNA are transcribed by these transcription factors. This is where we cross over to the domain of continuous variables. The information obtained by observing these gene expressions can be used to find out the relative effect of various sub-populations in the tumor tissue on the observables. This inferred relative effect can be interpreted as the combined effect of the proportion wise breakup of the tumor cell subpopulations as well as other random factors.

In [18] the authors present a Boolean model of the Mitogen Activated Protein Kinase (MAPK) signal transduction network, as reproduced in Fig. 2.1 and represent cancer as a stuck-at fault in the network. Such a treatment reduces the problem of the selection of kinase inhibitors for combination therapy to a simple case where the kinase inhibitors can be selected based on their effect on the variables of interest (the ones which are responsible for cell proliferation or apoptosis). Analysis in [18] has been done considering only single stuck at faults at a time which can be extended to the scenario of multiple faults. However, in either case, this approach assumes that the entire cancerous tissue can be modeled by a single faulty network. However, in reality, each faulty network models only one faulty cell type, that is models only one of the subpopulations. To model the entire cancer population, we need an ensemble of networks where the number of networks required is equal to the number of major subpopulations in the cancer tissue. This ensemble has to be deduced from expert knowledge. In our model, the subpopulations or networks in the ensemble exert their

effect on the observables in a weighted average fashion. Our objective is to find out the extent to which each network influences the behaviour of the tissue by observing the behaviour of the outputs, which is determined by the set of parameters in the model.

A survey of the existing literature can give us prior knowledge about the most likely points in a network where a stuck-at fault may occur. For instance, in 30% of human breast cancers we see an over expression of the ERBB2 gene [37]. This may cause ligand independent firing translating to a stuck-at one fault in the Boolean network. A stuck-at one fault at ERBB2 means that the variable corresponding to ERBB2 in the Boolean network shown in Fig. 2.1 is always upregulated regardless of the activity status of the proteins upstream of it. Similarly 90% of pancreatic cancer cases have a mutated Ras gene which causes it to lose its gtpase activity [37]. In other words, we have a stuck-at one fault associated with the Ras gene. Thus based on information such as the origin of the cancer tissue and prior knowledge of the most likely locations where faults can take place, we can reduce the number of networks in our ensemble.

2.2.1 A simple example

Let us consider a hypothetical cancer where we have narrowed down the number of major subpopulations to three. Let the first subpopulation be modeled by a Boolean network with a stuck-at one fault at ERK1/2, let the second subpopulation have two stuck-at-one faults at ERBB2/3 and Raf, and let the final subpopulation have a stuck-at-zero fault at PTEN. The different fault locations corresponding to the different subpopulations are shown as purple squares in the single Boolean network in Figure 2.1. Suppose we expose the cell culture to the drug U0126. This is a kinase inhibitor which targets MEK1 as shown in Figure 2.1. (All the drugs used in

this example are kinase inhibitors whose molecular targets are shown in Figure 2.1.) Let us also assume that the serum, as typically used in tissue cultures, has EGF, HBEGF, IGF, and NRG1 in it. If we observe the behavior of the transcription factor SP1 (shown at the bottom of the Boolean network in Figure 2.1 with green arrows), the first network predicts no change in the behavior of SP1 while in the second and third networks, SP1 will be downregulated. One way to observe the activity of SP1 is to measure the expression of a gene activated by the SP1 response element, for instance cMYC. In the second experiment if we expose the cell culture to a combination of AG1024 and Lapatinib, then SP1 will be upregulated in the first and second subpopulations but downregulated in the third subpopulation.

In the control experiment with no drug exposure, it is clear that all the subpopulations will have their SP1 transcription factors upregulated. The usual practice followed to calculate the normalized gene expression ratio is by the delta-delta method [22]. This involves normalizing with respect to a housekeeping gene such as GAPDH (Glyceraldehyde-3-Phosphate Dehydrogenase) followed by normalization with respect to the control experiment. The normalized gene expression ratio is the variable that we are interested in following.

A simple and realistic approach for modeling the normalized gene expression ratio utilizes the ratio of two normally distributed random variables, each with its standard deviation being directly proportional to its mean. The constant of proportionality is called the coefficient of variation, which is assumed to be constant for all the normally distributed random variables. A biological justification for this assumption of constant coefficient of variation has been provided in [7] where the gene expressions were measured using microarrays, while the observation of this phenomenon is reported in [6]. In this paper, our results will be developed specific to the above example where the observed variables are normalized gene expression ratios. However, the results

could be extended to the analysis of other observables where the relative effect of the various subpopulations on their behaviour is to be determined. This would require the use of models other than the ratio of two normal random variables, such as the gamma distribution, the log-normal distribution, or any other model which best fits the data. Though other models can be used, this model has certain advantages when it comes to determining the unknown parameters from collected data as we will demonstrate in the later sections.

Let us assume that we are observing the expression of a reporter gene of SP1, say cMYC. Let the effect of the 3 subpopulations on the normalized gene expression ratio of cMYC be in the ratio of $\alpha_{cMYC,1} : \alpha_{cMYC,2} : \alpha_{cMYC,3}$. We will call these the relative ratio parameters of cMYC which represent the extent to which each subpopulation manifests its effect on an observable (cMYC in this case). Each term in the ratio represents the net effect of a subpopulation which includes various factors such as the cell population and the concentration of the mRNA level in the cells. Thus the normalized gene expression ratio of cMYC for the first experiment, where the cell culture is exposed to U0126, is a random variable, which in turn is the ratio of two normally distributed random variables. The one in the numerator has a mean directly proportional to $\alpha_{cMYC,1}$ and standard deviation directly proportional to $\alpha_{cMYC,1} \times c$ (where c is the coefficient of variation which is considered constant for all genes). This is because the addition of U0126 shuts down the activity of the SP1 transcription factor in the other two subpopulations. The one in the denominator has a mean directly proportional to $\alpha_{cMYC,1} + \alpha_{cMYC,2} + \alpha_{cMYC,3}$ and standard deviation directly proportional to $(\alpha_{cMYC,1} + \alpha_{cMYC,2} + \alpha_{cMYC,3}) \times c$ since the control experiment has no drugs added and therefore, the activity of SP1 is not suppressed in any of the subpopulations. For the second experiment, following the same logic, the normalized gene expression ratio of cMYC is a ratio of two normally distributed

random variables. The random variable in the numerator has a mean of $\alpha_{cMYC,1} + \alpha_{cMYC,2}$ and a standard deviation of $(\alpha_{cMYC,1}+\alpha_{cMYC,2})\times c$ while the random variable for the denominator is the same as that for the first case.

For an intuitive understanding let us consider that the data points are generated by a model where the coefficient of variation is 0. In that case we will simply get the following two equations from the two experiments as shown below. If $r_{cMYC,1}$ and $r_{cMYC,2}$ denote the two measured normalized gene expression ratios of cMYC from the two experiments, we have:

$$\frac{\alpha_{cMYC,1}}{\alpha_{cMYC,1} + \alpha_{cMYC,2} + \alpha_{cMYC,3}} = r_{cMYC,1} \tag{2.1}$$

$$\frac{\alpha_{cMYC,1} + \alpha_{cMYC,2}}{\alpha_{cMYC,1} + \alpha_{cMYC,2} + \alpha_{cMYC,3}} = r_{cMYC,2}. \tag{2.2}$$

Since $\alpha_{cMYC,1} : \alpha_{cMYC,2} : \alpha_{cMYC,3}$ is a ratio, we can have the terms of the ratio sum to 1 to get another equation.

$$\alpha_{cMYC,1} + \alpha_{cMYC,2} + \alpha_{cMYC,3} = 1 \tag{2.3}$$

Equations (2.1), (2.2) and (2.3) will let us calculate the relative ratio parameters assuming the data points are drawn from a model with coefficient of variation c equal to 0.

However, in biological experiments a large sample size is hard to come by and sometimes we cannot afford to do a sufficient number of experiments to generate enough information just by observing a single observable (cMYC in our example above). For instance, if instead of the two experiments (excluding the control experiment) for the case described above, we do one experiment where we expose the cell

culture to a combination of the two drugs LY294002 and U0126, then from the faulty networks ensemble, it is apparent that the transcription factor FOS-JUN (also known as activator protein 1 or AP1) will be upregulated in the first and third subpopulations while it will be downregulated in the second one. Looking at SP1, it will be upregulated in the first subpopulation while it will be downregulated in the second and third subpopulations. Let us assume that we are observing a reporter gene of FOS-JUN. If we consider the case where c is 0 and use the same method as shown in equations (2.1) through (2.3), we will need to use the observed values of two different variables (a reporter gene of SP1 and a reporter gene of FOS-JUN) to estimate the relative influence of the subpopulations on the observables. However this method rests on the assumption that the relative effects of the different subpopulations is the same for all the observables, which in this example are the genes transcribed by FOS-JUN and SP1. This is a strong assumption since as mentioned earlier, the observed variables are affected not only by the proportions of the subpopulations, but also by individual random effects arising from many possible factors which make the assumption of equal relative ratio parameters unrealistic. However the data coming from different observables should not be ignored since all the observable data points contain information about the proportion wise breakup of the subpopulations. This calls for a model which utilizes all the information coming from various sources. Even though for each individual observable variable, the proportion wise breakup of the subpopulations is a small factor affecting its behavior, this factor affects all the observed variables. Thus taking information from all the observed variables will allow us to determine the proportion wise breakup among the subpopulations with better accuracy.

One such model is the multilevel hierarchical model. In this model, the relative ratio parameters vector $\alpha_i = \left(\alpha_{i,1} \ \alpha_{i,2} \ \alpha_{i,3} \right)^T$ for each observable variable i (the genes

transcribed by SP1 and AP1 in our examples) are drawn from a governing Dirichlet distribution having a parameter vector which needs to be estimated from the data points. Hence the relative ratio parameters for each observable variable i sum up to 1 and are all non negative. The parameter vector of the governing Dirichlet distribution is representative of the average of the information from all observed variables.

2.3 A hierarchical model for heterogeneous cancer tissue

Multilevel Hierarchical models are important new tools which are becoming increasingly popular in modern quantitative research. These models are useful in cases where the data is organized as a hierarchy of nested populations. In our case such a model is applicable since according to our requirement, the relative ratio parameters vector for gene i, $\alpha_i = \left(\alpha_{i,1} \ \alpha_{i,2} \ \alpha_{i,3} \right)^T$ determine the distribution of the gene expression ratio of the gene i and α_i will be a different vector for each gene. For the purpose of presentation, the coefficient of variation is not made to have a hierarchical structure and the same value is assumed for all the observable variables, although it is possible to develop a hierarchical structure for the coefficient of variation allowing it to vary from gene to gene. A lot of literature is available on multilevel hierarchical models [12, 13, 16]. So we will not go into an in-depth discussion about a general Hierarchical model. Instead, in this section, we will only describe the details pertaining to our model.

Figure 2.2 shows the conditional dependencies of the model. All the observations for each observable variable have a probability distribution which depends on the "relative ratio parameters" for that variable. These relative ratio parameters are drawn from an underlying Dirichlet distribution, the parameter vector K of which is to be estimated. A Dirichlet distribution generates vectors with non negative values whose elements add up to 1. With the appropriate parameter vector K, the

distribution can be made to take a variety of shapes and center around any mode (peak value of the probability distribution). This mode can be interpreted as the average effect of the subpopulations on the observables. The larger the values of the elements of the parameter vector K, the "sharper" the Dirichlet distribution is around the mode.

Another big advantage of a hierarchical model is that it allows for the sharing of information across observables. Consider the experiment discussed in the previous section where the hypothetical tissue was exposed to LY294002 and U0126. Looking at SP1 and FOS-JUN separately, there is not enough information to infer the relative ratio parameters for these two observables, but combining the data from these two observables allows us to determine the parameters of the underlying Dirichlet distribution. This will be demonstrated using synthetic data derived from the model of the MAPK signal transduction network in a simulation example and applied to real data derived from experiments on fibroblasts.

The probability distribution of the normalized gene expression ratio for the j^{th} data point collected from an experiment involving the measurement of the i^{th} gene is dependent on the relative ratio parameters vector α_i , the coefficient of variation c, and the "expression profile" $d_{i,j}$. The expression profile is simply a vector whose length is equal to the number of subpopulations in our ensemble. An element of this vector $d_{i,j}$ is 1 if the contribution to the j^{th} data point collected from an experiment involving the i^{th} gene is expected to be upregulated in the corresponding subpopulation, 0 otherwise. This will change from one experiment to the next for the same gene depending upon the behavior of the Boolean networks in the ensemble. For example the expression profile for the gene transcribed by SP1 in the first example in the previous section is $\begin{pmatrix}1&0&0\end{pmatrix}^T$ for the case where exposure to U0126 has occurred, and $\begin{pmatrix}1&1&0\end{pmatrix}^T$ for the case where exposure to AG1024 and Lapatinib

has occurred. We make the reasonable assumption that the expression profile for each observable variable is known for each experiment since it is dependent on the deterministic behavior of the Boolean networks in the ensemble. As explained in the previous section, the normalized gene expression ratio is a ratio of two normally distributed random variables. We will derive the probability density function (pdf) of the ratio of two normally distributed random variables below. Consider two normal random variables T_1 and T_2 with mean and standard deviations μ_1 and $c \times \mu_1$ and μ_2 and $c \times \mu_2$ respectively. Define

$$R := \frac{T_1}{T_2} \tag{2.4}$$

and define

$$X := T_2 \tag{2.5}$$

Following the standard procedure for computing the joint density of functions of two random variables, the Jacobian comes out to be

$$J = \left| \begin{bmatrix} X & R \\ 0 & 1 \end{bmatrix} \right| = |X| \tag{2.6}$$

Since $X = T_2$ has very thin tails in the negative region, we get

$$J \approx X \tag{2.7}$$

Thus we have

$$P_{R,X}(r, x) = P_{T_1}(t_1) \times P_{T_2}(t_2) \times J \tag{2.8}$$

or

$$P_{R,X}(r,x) \approx \frac{1}{2\pi\mu_1\mu_2c^2} \times exp\left(-\frac{1}{2c^2\mu_1^2}(rx-\mu_1)^2 - \frac{1}{2c^2\mu_2^2}(x-\mu_2)^2\right)x \quad (2.9)$$

Define $m = \frac{\mu_1}{\mu_2}$. Since we have $P_R(r) = \int P_{R,X}(r,x)dx$, integrating the joint density over all x, we obtain

$$P_R(r) = \frac{m(r+m)}{\sqrt{2\pi}c(r^2+m^2)^{\frac{3}{2}}}exp\left(-\frac{1}{2c^2}\frac{(r-m)^2}{(r^2+m^2)}\right) \quad (2.10)$$

We note that the expression in equation 2.10 above agrees with the ratio distribution derived in [7].

Define $m_{i,j} = d_{i,j}^T\alpha_i$. Thus the conditional probability distribution of the normalized gene expression ratio of the i^{th} gene in the j^{th} experiment comes out to be

$$P(r_{i,j}/\alpha_i, d_{i,j}, c) = \frac{m_{i,j}(r_{i,j}+m_{i,j})}{\sqrt{2\pi}c(r_{i,j}^2+m_{i,j}^2)^{\frac{3}{2}}} \times exp\left(-\frac{1}{2c^2}\frac{(r_{i,j}-m_{i,j})^2}{(r_{i,j}^2+m_{i,j}^2)}\right) \quad (2.11)$$

Let N be the number of networks in the ensemble. For our examples we have $N = 3$ since we have chosen to include 3 networks in the ensemble. However, it is not a hard and fast rule to include exactly three networks in the ensemble since the number of subgroups can be more or less than three. The probability distribution of

the relative ratio parameters vector α_i of the i^{th} gene is given by

$$P(\alpha_i/K) = \frac{\prod_{q=1}^{N} \alpha_{i,q}^{K_q-1}}{Beta(K)} \tag{2.12}$$

where $Beta(K)$ is the beta function defined as

$$Beta(K) = \frac{\prod_{q=1}^{N} \Gamma(K_q)}{\Gamma\left(\sum_{q=1}^{N} K_q\right)} \tag{2.13}$$

Here Γ represents the Gamma function. Let n_i be the number of data points of the i^{th} gene from all experiments combined and let V be the total number of observables (genes). Let r denote the set of all the data points $r_{i,j}$ taken together. Let d denote the set of all $d_{i,j}$ taken together. Then considering the parameters of interest K and c, we get the likelihood function of the data points to be

$$P(r/K, c, d) \quad = \quad \prod_{i=1}^{V} \int \prod_{j=1}^{n_i} P\left(r_{i,j}/\alpha_i, d_{i,j}, c\right) P(\alpha_i/K) d\alpha_i \tag{2.14}$$

This needs to be maximized over K and c in order to obtain the maximum likelihood estimate of K and c. The integrations can be difficult or impossible to perform analytically. So we will resort to Markov Chain Simulation to estimate the posterior probability distribution of the elements in the parameter vector K.

2.3.1 Estimating parameter values from observed data

Once the ensemble of networks has been chosen from biological knowledge, experimental data about gene behavior in response to kinase inhibitor drugs can be used to estimate the parameters of the model. We will use the Metropolis-Hastings (M-H) algorithm to generate samples from the posterior distributions of the unknown parameters, which are the parameter vectors K, all the α_i's and the coefficient of

variation c, conditional on all the data r. The M-H algorithm generates a Markov Chain in the unknown parameter space whose stationary distribution is the required posterior distribution of the unknown parameters. Letting this Markov Chain run to stationarity and drawing samples from the Markov chain is equivalent to drawing samples of the unknown parameters from their posterior distribution. There is a lot of available general literature on this algorithm [12, 13, 16] and so we will simply focus on the specifics for our case.

The usual Bayesian Method requires us to define priors over the parameters K and c. For c, we choose the prior such that the reciprocal of the square of c is gamma distributed with a shape parameter of $\frac{v_0}{2}$ and an inverse scale parameter of $\frac{v_0 c_0^2}{2}$.

$$\frac{1}{c^2} \sim \Gamma\left(\frac{v_0}{2}, \frac{v_0 c_0^2}{2}\right) \tag{2.15}$$

Here Γ represents the Gamma distribution and not the Gamma function.

For K we choose a proper prior where all the elements of K are independently identically exponentially distributed. The means for these exponential distributions can all be made equal and arbitrarily large so that the prior is almost flat as compared to the posterior. Choosing proper prior distributions ensures that the posterior is also proper.

To run the M-H algorithm, we need the full conditionals of the unknown variables. Define α_{-i} as the set $\{\alpha_1, \alpha_2, ..., \alpha_{i-1}, \alpha_{i+1}, ...\alpha_V\}$. Then the full conditional of α_i is

as follows

$$P(\alpha_i/K, c, r, \alpha_{-i}, d) \propto \prod_{j=1}^{n_i} P(r_{i,j}/\alpha_i, d_{i,j}, c) P(\alpha_i/K)$$
$$\propto \prod_{j=1}^{n_i} \left(\frac{m_{i,j}(r_{i,j} + m_{i,j})}{(r_{i,j}^2 + m_{i,j}^2)^{\frac{3}{2}}} exp\left(-\frac{1}{2c^2} \frac{(r_{i,j} - m_{i,j})^2}{(r_{i,j}^2 + m_{i,j}^2)} \right) \right)$$
$$\times \prod_{q=1}^{N} \alpha_{i,q}^{K_q - 1} \quad (2.16)$$

Define α as the set of all the relative ratio parameters vectors α_i's. Let $P(K)$ be the prior over K. Then the full conditional of K is as follows

$$P(K/\alpha, c, r, d) \propto P(K) \times \prod_{i=1}^{V} P(\alpha_i/K)$$
$$\propto P(K) \times \frac{1}{(Beta(K))^V} \prod_{q=1}^{N} \left(\prod_{i=1}^{V} \alpha_{i,q} \right)^{K_q - 1} \quad (2.17)$$

The full conditional of c is such that

$$\frac{1}{c^2} \sim \Gamma\left(\frac{\left(v_0 + \sum_{i=1}^{V} n_i \right)}{2}, \right.$$
$$\left. \frac{\left(v_0 c_0^2 + \sum_{i,j} \frac{(r_{i,j} - m_{i,j})^2}{(r_{i,j}^2 + m_{i,j}^2)} \right)}{2} \right) \quad (2.18)$$

The parameters are sampled from their full conditionals one after the other and after each cycle the newly generated values are stored. c can be generated from its full conditional simply by taking a sample from the standard gamma distribution with the above parameters as shown in equation 2.18 and taking the reciprocal of the square root of the sample. This convenient step is possible due to the specific form of $P(r_{i,j}/\alpha_i, d_{i,j}, c)$ which results from the ratio of two normally distributed random

variables. But K and the α_i's need to be sampled from non standard distributions. We use random walk proposal distributions to generate new values of K and α_i's from their previous values.

New values K^* are sampled from their proposal distributions in the following manner. For the q^{th} element K_q^* of K^*, do the following

- Sample t from uniform($K_q - U_K$,$K_q + U_K$), where K_q is the q^{th} element of K.
- If $t < 0$, then set $K_q^* = -t$, else set $K_q^* = t$.

U_K is a tuning parameter which can be adjusted to improve the behavior of the Markov Chain. Using the method as described above makes the proposal distribution symmetric [16]. The acceptance ratio for K is calculated as

$$R_K = \frac{P(K^*/\alpha, c, r, d)}{P(K/\alpha, c, r, d)} \tag{2.19}$$

and K is updated to K^* with a probability of $min(R_K, 1)$.

New values α_i^* are generated from a Dirichlet proposal distribution with parameter value vector given by $\frac{\alpha_i}{U_{\alpha_i}}$. U_{α_i} is a tuning parameter. Define $D(x/y)$ to be the probability distribution of x which is Dirichlet distributed with parameter y. Since the proposal distributions used for the α_i's are not symmetric, the acceptance ratio is calculated as

$$R_{\alpha_i} = \frac{P(\alpha_i^*/K, c, r, \alpha_{-i}, d) D(\alpha_i/\frac{\alpha_i^*}{U_{\alpha_i}})}{P(\alpha_i/K, c, r, \alpha_{-i}, d) D(\alpha_i^*/\frac{\alpha_i}{U_{\alpha_i}})} \tag{2.20}$$

α_i is updated to α_i^* with a probability of $min(R_{\alpha_i}, 1)$.

The series of steps described above results in a Markov Chain whose stationary distribution is the same as the posterior distribution of the unknown parameters. Letting this Markov Chain run to stationarity and drawing samples from the Markov

chain is equivalent to drawing samples of the unknown parameters from their posterior distribution.

Once the draws from the posterior distribution of the unknown parameters have been obtained, we can obtain the posterior mean, the values with the maximum posterior distribution value (the modes) and the confidence intervals of the parameters from the kernel density estimate. Such estimates of the parameter vector K can then be used to determine the proportion wise breakup of the subpopulations corresponding to the networks included in the ensemble. Such methods will be demonstrated in the coming subsections.

2.3.2 Experiments with synthetic data

To demonstrate the working of the algorithm, we ran simulations of the algorithm on synthetic data. We generated synthetic data from the example described previously which was derived from the MAPK signal transduction network, which is a well understood network. Three networks with the "stuck-at" faults as described in the previous example were taken in the ensemble. One reporter gene for each of the 4 transcription factors was considered as an observable. Thus we have 4 observables with 4 different "relative ratio parameter vectors". K was fixed to be $\begin{pmatrix} 10 & 6 & 3 \end{pmatrix}^T$. This corresponds to a Dirichlet distribution with a mode of $\begin{pmatrix} 0.5625 & 0.3125 & 0.1250 \end{pmatrix}^T$. c was fixed to be 0.1 since typical values of the coefficient of variation were reported in [7] to be close to 0.17. First the "relative ratio parameters" for the 4 observables were generated from the Dirichlet distribution with parameter vector K and then held fixed for each reporter gene for the 4 transcription factors FOS-JUN, SP1, SRF-ELK1, and SRF-ELK4. Then observations of the observables were generated for various combinations of drugs following the model of the ratio of two normally distributed random variables. 12 drug combinations were chosen out of the 63 pos-

sible combinations of the 6 drugs in the model in such a way so that the "expression profiles" for each gene cannot generate a sufficient number of equations permitting calculation of the "relative ratio parameters" for that gene, in the event that the coefficient of variation c were zero. For example all observed data-points corresponding to the reporter of FOS-JUN had their corresponding expression profiles as $(1\ 0\ 1)^T$ and all observed data-points of the rest of the observables had their corresponding expression profiles as $(1\ 0\ 0)^T$. This is done so as to demonstrate how the sharing of information from all the observables can be used to obtain an estimate of the parameter vector K of the underlying Dirichlet distribution.

For the purposes of demonstrating the algorithm, the prior for the elements of the parameter vector K were chosen to have exponential distributions with means of 1000, and the parameters for the prior of c were chosen as follows. The value v_0 was taken as 1 and c_0 was taken to be 0.

The Markov Chain was run for 3000 iterations to make it reach stationarity. The tuning parameters were adjusted to get acceptance rates of close to 30% for the unknown parameters. The Markov chain was run for 400,000 iterations and thinned 100 times (1 in 100 samples generated was stored for each parameter). This resulted in a maximum inefficiency factor of less than 4 among all the parameters. The reader is referred to [12, 13, 16] for information on Markov Chain Monte Carlo diagnostics and the inefficiency factor.

Multivariate kernel density estimation for any general N dimensional parameter vector is made using the multivariate Gaussian kernel with a diagonal covariance matrix, the j^{th} element of which is given by $C_j = \left(\frac{\sigma_j}{n^{\frac{1}{N+4}}}\right)^2$, where σ_j is the standard deviation of the j^{th} element of the parameter vector under consideration, n is the number of samples drawn from the posterior distribution, and N is the number of elements in the parameter vector (3 for K in our example). This rule of thumb is

discussed in [33].

Figure 2.3 shows the kernel density estimate of the marginal distributions of the elements of K along with their priors. The priors are far too spread out and non-informative as compared to the posteriors. Hence the value of K with the maximum posterior distribution is equivalent to the maximum likelihood estimate. This estimate comes out to be $\begin{pmatrix} 9.1367 & 5.1330 & 2.2130 \end{pmatrix}^T$ which was estimated from the kernel density estimate of the joint distribution of the 3 elements of the parameter vector K using gradient ascent with non-negativity constraints. Comparing it to the actual value of K, we can see that it is quite close. Confidence intervals can also be calculated from the kernel density estimates, although we have not shown such calculations here. The more the data fed to the model, the more accurate is the estimate and the confidence intervals are narrower.

We are more interested in the posterior distribution from which the relative ratio parameters of the observables come. That is if we know the parameter vector K, we would like to know the distribution of the relative ratio parameters, which is nothing but Dirichlet distributed with the parameter vector K. But since K has a posterior distribution, we would like to know the value of $\int P(\alpha/K) P(K/r, d) dK$, where $\alpha = \begin{pmatrix} \alpha_1 & \alpha_2 & \alpha_3 \end{pmatrix}^T$ is Dirichlet distributed with parameter vector K, and r is the set of all observed data points. This can be obtained by sampling α from Dirichlet distributions with parameters set as the samples drawn from the posterior of K. Repeating this process for all the samples of K, we get the samples of α. The posteriors of the elements of α for this example are shown in Figure 2.4. The mode is derived from the kernel density estimate using gradient ascent subject to the constraint that the elements of α sum to 1 along with non-negativity constraints. The mode obtained is $\begin{pmatrix} 0.5974 & 0.2930 & 0.1095 \end{pmatrix}^T$. Comparing it to the original mode of $\begin{pmatrix} 0.5625 & 0.3125 & 0.1250 \end{pmatrix}^T$, we can see that it is quite close.

2.3.3 Verification using experimental data

In order to test if the theory developed so far would work, we need to collect data from an experiment performed on a tissue where the dominant population or the dominant network is known. In a cancerous cell line, one cannot be sure which network is dominant. But in a normal cell line, such as adult fibroblasts, it is fair to assume that a network modeling a faultless MAPK signal transduction network would be the most dominant one, no matter what networks are included in the ensemble. Hence we performed a simple experiment on adult fibroblasts to demonstrate the approach.

Adult fibroblasts were grown in Fibroblast Basal Medium (ATCC) in 60mm tissue culture petri dishes till confluence. Following this, the cells were maintained in Dulbecco's modified Eagles medium-F12 (DMEM/F12) (Atlanta Biologicals), supplemented with 0.2% fetal bovine serum (Atlanta Biologicals) for 4 days (All concentrations of the supplements used were calculated with respect to plain DMEM/F12 medium without serum). The medium was changed every day after wash with phosphate buffer solution (PBS). All cell cultures were incubated at 37 °C in a 5% CO_2 incubator.

The cells were then exposed to DMEM/F12 supplemented with 0.2% FBS and 100μM Anisomycin for 30 minutes. Anisomycin is a protein synthesis inhibitor which activates the MAPK signal transduction network and keeps it responsive to kinase specific inhibitors [2, 10]. That is, with the addition of Anisomycin, we anticipate the MAPK signal transduction network to respond to the addition of a drug such as U0126. Anisomycin, being a protein synthesis inhibitor, would also cut of any feedback path which has a translation (protein synthesis) step in it. The tissue culture petri dishes were then grouped into 3 groups (groups 0, 1 and 2). After

the initial 30 minutes of exposure to Anisomycin, each group was then exposed to DMEF/F12 supplemented with 20% FBS,100μM Anisomycin, 50μM of LY294002, and/or 10μM of U0126 as shown in table 2.1.

Group 0 is not exposed to LY294002 or U0126, which are highly specific inhibitors of PI3 Kinase (PI3K in Figure 2.1) and MEK1 respectively. The molecular targets of LY294002 and U0126 are shown in Figure 2.1. Genes having the SP1 and SRF-ELK response elements in their promoters were quantified through real time PCR and the delta-delta method [22] with GAPDH as the reference gene and group 0 as the control.

EGR1 is measured as a reporter gene of SRF-ELK transcription factor [8]. JUN, BIRC5, and cMYC are measured as reporters of SP1 [31, 24, 20]. Other genes having the SP1 response element in their promoters are Decorin, IRF3 and VEGFA [35, 40, 32, 29]. Four different alternative transcripts of Decorin were measured. Thus we have a total of ten observables. The expression values calculated are shown in table 2.2.

For the sake of demonstration we assumed 3 networks to be in the ensemble. Network 1 has no mutations, i.e. no "stuck-at" faults. This network models the normally behaving fibroblasts. Network 2 is assumed to have a "stuck-at 1" fault at ERK1/2 and network 3 is assumed to have "stuck-at 1" faults at SRF-ELK1 and SRF-ELK4. The "expression profiles" for all the genes for the experimental conditions of groups 1 and 2 are known and depend on the behaviour of the 3 networks included in the ensemble. These are shown in table 2.2.

As described in the previous section, samples from the posterior distributions of the unknown parameters were drawn using the Metropolis-Hastings Algorithm. The number of samples were drawn until the effective sample size was atleast 300 for all the parameters. The reader is referred to [16] for information on effective sample

sizes in Markov Chain Monte Carlo analysis.

$\int P(\alpha/K)\, P(K/r,d)\, dK$ was estimated as described in the previous section. The marginals of the 3 components of α are shown in Figure 2.5. The spread of the distribution is large due to lack of enough data points. The mode is derived from the kernel density estimate using the gradient ascent subject to the constraint that the elements of α sum to 1 and non-negativity constraints. This mode comes out to be $(\,0.6453\ 0.2255\ 0.1292\,)^T$. As expected, the faultless network representing normal fibroblasts has the maximum influence on the behaviour of the observables, close to 65% . This simple experiment is a demonstration of how real world technology such as QPCR can be used to determine the composition of a heterogeneous tissue.

2.4 Summary and comments on possible future work

In this work we addressed the important problem of heterogeneity in cancer tissues and presented a model which has the ability to use prior pathway knowledge and knowledge about likely mutations in cancers to represent a heterogeneous cancer tissue as an ensemble of faulty Boolean networks. We demonstrated the general idea of our approach by considering the observed variables to be genes transcribed by key transcription factors. We modeled the gene expression ratios as ratios of normally distributed random variables whose means were affected by the networks in the ensemble to varying degrees. However, if some other observables are used, then the ratio of normally distributed random variables formulation may not hold and hence the lowest level of the hierarchical model would have to be altered. However, the overall approach of hierarchically modeling the relative ratio parameters would remain the same. We also demonstrated how the Metropolis-Hastings MCMC method can be used to estimate the relative effect that each subpopulation exerts on the observed variables. This estimate gives us an idea about which subpopulation is the

most dominant one among all the subpopulations in the ensemble. Such estimates, if obtained using data from individual patients, could help customize combination therapy design and could help improve the success rate of such cancer therapies. for more information on this work, the reader is referred to [25].

Future work could also focus on algorithms that allow the addition of networks other than the ones with which the algorithm starts or the deletion of networks so as to better fit the data. The results in this paper have been developed with qPCR data in mind. However, we do believe that similar models could be developed to integrate data from more modern technologies such as Next Generation Sequencing and flow cytometry combined with prior pathway knowledge in order to determine the compositional breakup of the tissue. The details, of course, would need to be worked out and could form the basis for future investigations. Work on speeding up the computation of the posterior marginals has also been done and is described in later sections of this thesis.

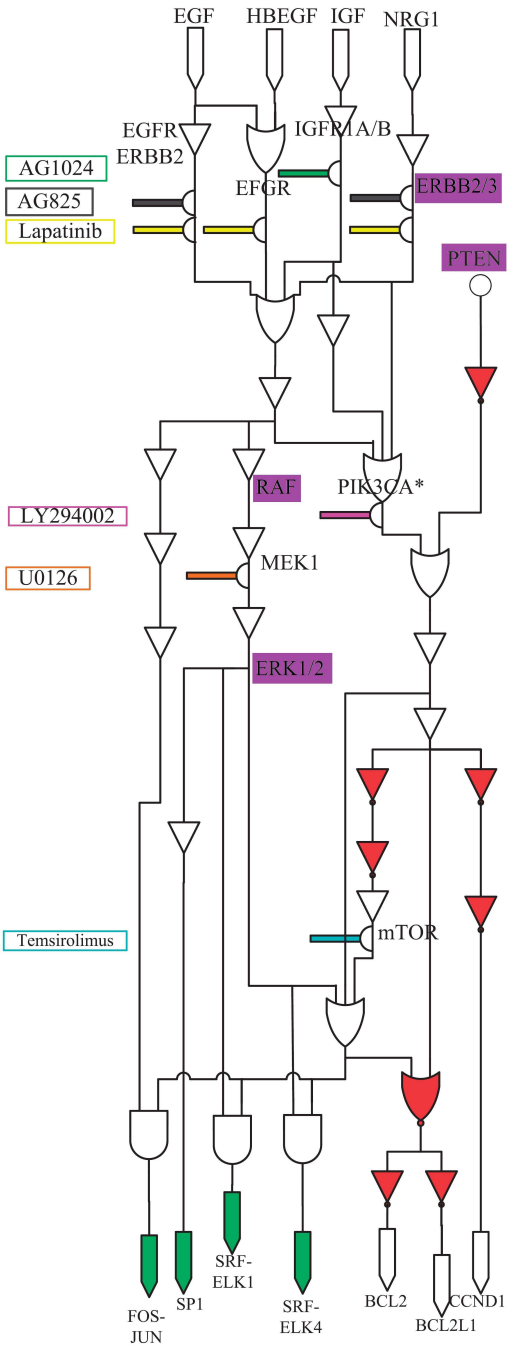

Figure 2.1: A Boolean network model of the MAPK signal transduction network with target locations of inhibitory drugs shown.

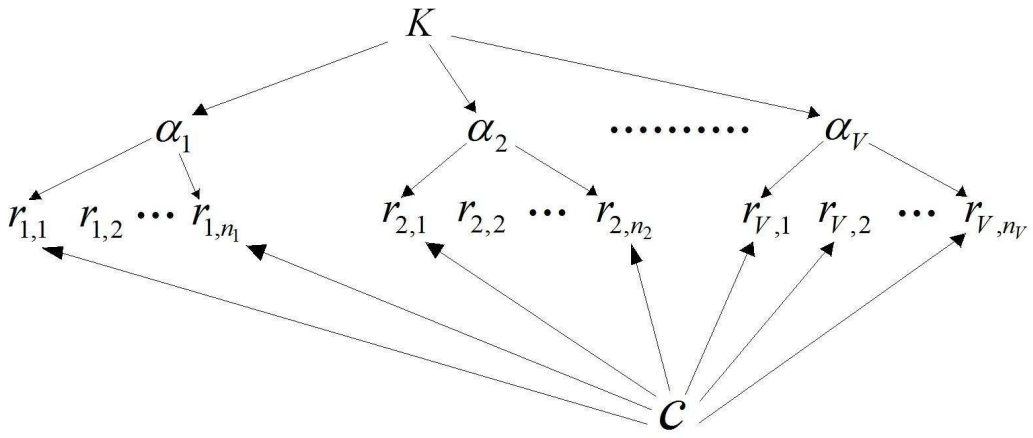

Figure 2.2: A Bayesian network representing the conditional dependencies in our model.

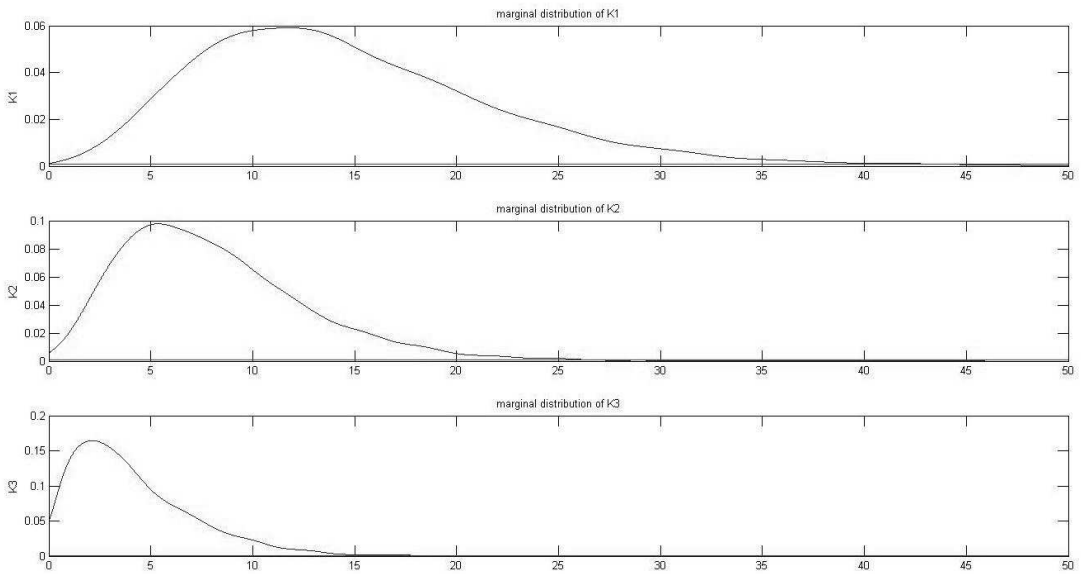

Figure 2.3: Marginal distribution of the elements of the parameter vector K.

Table 2.1: Table showing which groups were exposed to which compounds

	FBS	Anisomycin	LY294002	U0126
Group 0	X	X		
Group 1	X	X	X	
Group 2	X	X	X	X

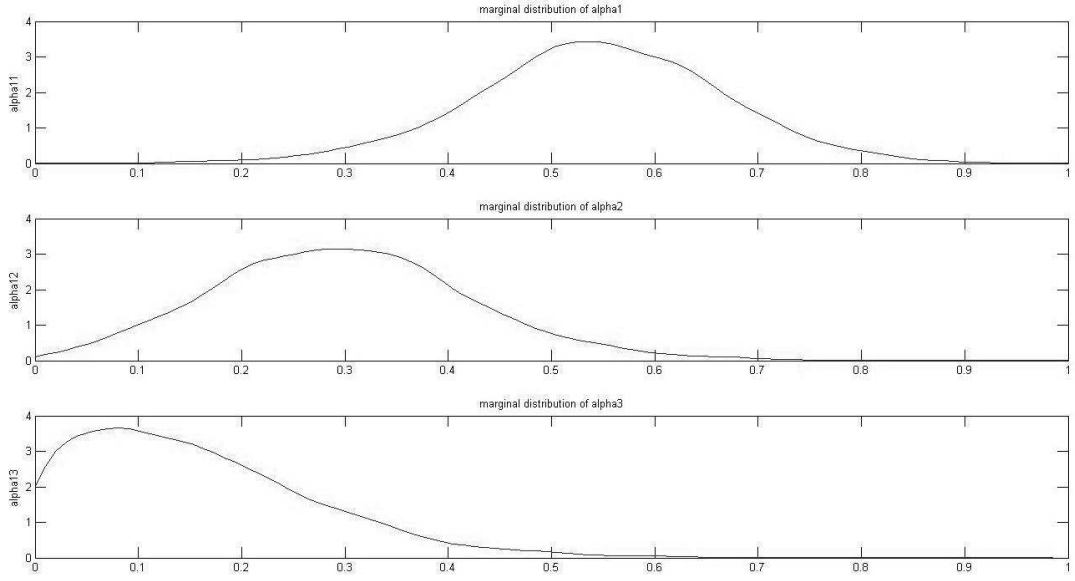

Figure 2.4: Marginal distribution of the elements of α for simulation experiments.

Table 2.2: Table showing the normalized gene expression ratios, their reference sequence (RefSeq) numbers and their "expression profiles"

gene	*RefSeq*	group 1		group 2	
		exp. profiles	*norm. gene exp.*	*exp. profiles*	*norm. gene exp.*
EGR1	NM_001964.2	1 1 1	0.598739352	0 1 1	0.47963206
JUN	NM_002228.3	1 1 1	0.493116352	0 1 0	0.154963462
BIRC5	NM_001168.2	1 1 1	0.579867973	0 1 0	0.384218795
CMYC	NM_002467.4	1 1 1	0.320856474	0 1 0	0.257028457
DNC(Decorin)	NM_133504.2	1 1 1	0.081899588	0 1 0	0.008668512
	NM_133505.2	1 1 1	0.072795849	0 1 0	0.024180703
	NM_133507.2	1 1 1	0.334481889	0 1 0	0.166085727
	NM_133503.2	1 1 1	0.435275282	0 1 0	0.279321785
IRF3	NM_001571.5	1 1 1	0.517632462	0 1 0	0.262429171
VEGFA	NM_003376.5	1 1 1	0.444421341	0 1 0	0.316439148

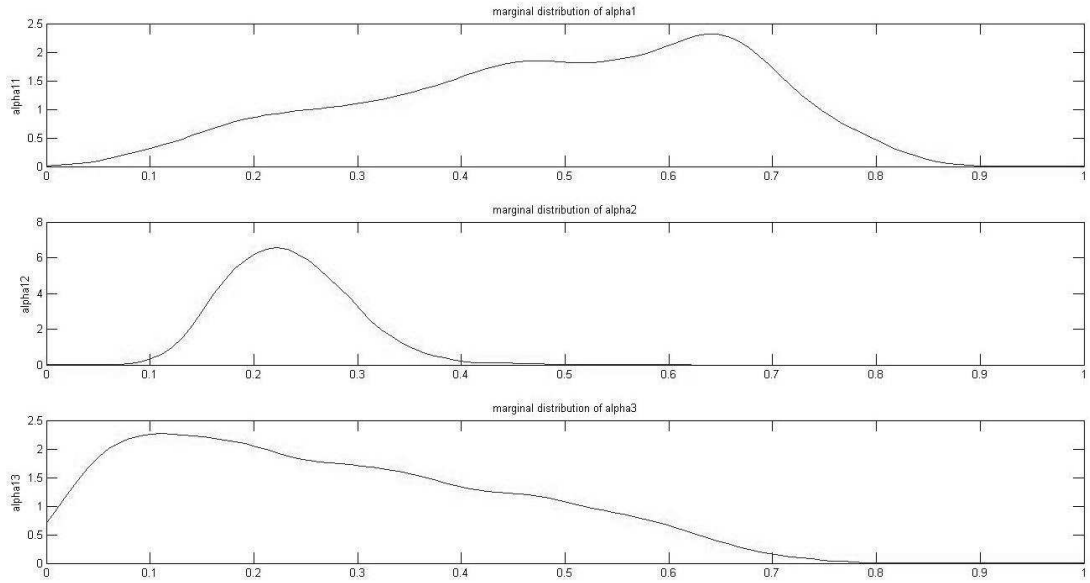

Figure 2.5: Marginal distribution of the elements of α for data derived from experiments on fibroblasts.

3. USING THE MESSAGE PASSING ALGORITHM ON DISCRETE DATA TO DETECT FAULTS IN BOOLEAN REGULATORY NETWORKS *

3.1 Introduction

Modeling cellular behavior is a first step towards the holistic understanding of the multivariate interactions among various genes. One possible approach to do that is through gene regulatory networks. These networks could also help in developing better intervention strategies in order to shift the state of the cell or the tissue to a more favorable one. Many different approaches have been proposed in the literature for modeling the behavior of genetic regulatory networks. Many of these methods have been discussed in the previous sections. These include differential equations [4], deterministic and probabilistic Boolean networks [34, 9], and Bayesian and dynamic Bayesian networks [11, 42]. Some of these methods rely on the assumption that the transition probabilities are provided beforehand. Such an assumption may not be realistic since the sheer volume of data required to effectively estimate the transition probabilities makes it a practically difficult proposition. Some methods such as the REVEAL algorithm [21] provide approaches to learn deterministic Boolean networks from discretized time course data. However time course data from biological samples itself can be difficult to come by.

One way to get around the problem of insufficient data is to use prior knowledge about the regulatory interactions between the various biological molecules in a cell. In the biological literature, a lot of information is available regarding the various regulatory interactions. This information has been collected by biologists over a long

*Parts of this section are reprinted with permission from "Using the message passing algorithm on discrete data to detect faults in boolean regulatory networks" by A. K. Mohanty, A. Datta, and V. Venkatraj, 2014. *BMC Algorithms for Molecular Biology*, volume 9, no. 20, 12 pages. doi:10.1186/s13015-014-0020-6.

period of time. These regulatory interactions, collectively referred to as pathway knowledge, are generally not incorporated into the various methods of modeling gene regulatory networks. Using this information, however, would result in models which describe cellular behavior more accurately.

A possible approach to use such prior information has been developed in [19]. In that reference, the authors use Boolean logic to model signal transduction networks. In [25], the authors have used boolean models derived from prior information to model the heterogeneity of cancerous tissues. Furthermore, in [18] Boolean logic is used to model the Mitogen Activated Protein Kinase (MAPK) signal transduction network and the result of that modeling is shown in Figure 2.1. Here, each connecting wire corresponds to a variable which represents the state of the corresponding protein/gene. In this model each variable is assumed to have two states, an activated and a deactivated one. For example the state of EGFR will be upregulated or activated when the cell is exposed to EGF. The way the various variables are dependent on each other can be modeled using standard Boolean logic functions such as AND, OR, NOT, NAND, etc.. This is shown in Figure 2.1. In [18] the authors presented a stuck-at fault model of the mutations which result in the neoplastic behavior of the tissue. A stuck-at-one fault corresponds to a variable permanently being in an activated state irrespective of the states of the variables upstream of it. Similarly a stuck-at-zero fault would mean a variable has a permanently downregulated state irrespective of the states of the other upstream variables. These "stuck" variables would however affect the variables downstream of them through the Boolean Logic gates which have these variables as inputs. To show how Boolean regulatory networks with stuck-at faults can be used to model cancerous tissue, we give the following examples. In 30% of human breast cancers there is an over expression of the ERBB2 gene [37]. This causes ligand independent firing translating to a stuck-at-one fault

in the Boolean network. A stuck-at one fault at ERBB2 means that the variable corresponding to ERBB2 in the Boolean network shown in Figure 2.1 is always up-regulated regardless of the activity status of the variables upstream of it. Similarly in 90% of the pancreatic cancer cases we see a mutated Ras gene which causes it to lose its gtpase activity [37]. In other words, we have a stuck-at-one fault associated with the Ras variable. Stuck-at faults could also be interpreted as points of dysregulation in the Boolean network brought about by certain genes irrespective of the presence of mutations.

Locating stuck-at faults in a given Boolean regulatory network could help in the identification of key dysregulated genes that have a strong impact on the observable variables. This in turn could be used to identify targets for new drugs. Knowledge about the locations of the stuck-at faults along with knowledge about the targets of the kinase inhibitory drugs can be used to come up with optimal intervention strategies. A method to devise optimal intervention strategies using such Boolean regulatory networks with stuck-at faults is described in [18]. Accordingly, the problem we pose is this: given data points, where each data point consists of a combination of drugs used as the input and the activity of the observable variables as outputs, is it possible to locate the variables where stuck-at-faults have occurred? In the following sections we represent the problem as a statistical model with unknown parameters which are estimated from the data points using the message passing algorithm. This algorithm allows for rapid computation of the posterior probabilities of the parameters. The estimates obtained are evaluated by comparison with the results given by Markov Chain Monte Carlo methods.

3.2 Model description

There are many ways to model a gene regulatory network which describes the behavior of neoplastic tissue. The general rule is that the more the number of unknown parameters, the more the amount of data that is required to get an effective estimate of those parameters. Hence the modeling must be done keeping in mind the limited amount of data available from biological experiments.

As has been pointed out before, literature survey would enable us to know the most likely locations in the Boolean network where stuck-at faults can take place. As stated in the previous sections, in 30% of human breast cancers there is an over expression of the ERBB2 gene, and in 90% of the pancreatic cancer cases we see a mutated Ras gene. These are among many examples where prior knowledge about locations of faults is available. This knowledge would allow us to limit the search space for faults in the network. For example we may provide a set of locations where we want to search for faults.

One important assumption made in the modeling of mutations is that they are random events that occur independently of each other [15, 14, 41]. We make use of this assumption in our model by assuming that the faults occur unconditionally independent from each other with certain unknown probabilities associated with them. These unknown probability parameters are to be estimated from the collected data. These estimated probabilities will indicate our confidence about where the faults have occurred in the Boolean regulatory network.

We now explain the key ideas through a simple example. Let us assume that we have narrowed down the set of locations where we want to search for faults to be composed of RAF, IRS1, and RHEB as shown in Figure 2.1 (we are assuming stuck-at-one faults). Let their probabilities of occurrences be ρ_1, ρ_2, and ρ_3 which are to

be determined. Define $\rho = (\,\rho_1\ \rho_2\ \rho_3\,)^T$ as the vector of the three parameters. Three possible locations of faults implies that there are 2^3 different fault combinations and their associated networks corresponding to the binary numbers 000, 001,....,111. The first network is one with no faults and has a probability of

$$P\,(M = 0/\rho) = (1-\rho_1)(1-\rho_2)(1-\rho_3). \tag{3.1}$$

The second network has a single stuck-at-one fault at RHEB alone, and it's probability is given by $P\,(M = 1/\rho) = (1-\rho_1)(1-\rho_2)\rho_3$. Similarly, the third network has a single stuck-at-one fault at IRS1 alone, with a probability of $P\,(M = 2/\rho) = (1-\rho_1)\rho_2(1-\rho_3)$, and so on. The variable M is the decimal equivalent of the binary number representing the different fault combinations and could equivalently represent the particular faulty Boolean network being considered. Since there are three possible locations where stuck-at-faults can take place in this example, M can take $2^3 = 8$ different values. In our convention, we use integers from 0 to $2^3 - 1$ to represent the values taken by M. For example $M = 6$ corresponds to a network with faults at RAF and IRS1 but not at RHEB and has a corresponding probability of $\rho_1\rho_2(1-\rho_3)$.

In this example the dimension of ρ is three, but it can be any integer depending on the size of the search space. Determining the entries of ρ allows us to determine the most likely faulty networks. Let V be the dimension of ρ. Then it is clear that $P\,(M = m/\rho)$ has the following form:

$$P\,(M = m/\rho) = \prod_{v=1}^{V} \rho_v^{R_{v,m}}\,(1-\rho_v)^{1-R_{v,m}} \tag{3.2}$$

where $R_{v,m}$ is either 0 or 1 and m can vary from 0 to $2^V - 1$.

Consider any one of the variables represented as arrows at the bottom of figure 2.1. Let us represent that variable by O_j. j varies from 1 to 7 in our example based on figure 2.1. The behavior or O_j is determined by the network and what faults are in it. Let $o_{i,j}$ be an observation of that variable when the combination input is I_i. $o_{i,j}$ can be either 0 or 1 since we are dealing with a boolean network here. Given that the network M is any one of the 2^V possible networks and given that the drug combination input is I_i, the probability $P\left(O_j = o_{i,j}/M = m, I_i\right)$ can be either 0 or 1. It is 1 when $o_{i,j}$ matches the output of the j^{th} output variable of the m^{th} network for the the input drug combination I_i, and is 0 otherwise. Let us represent $P\left(O_j = o_{i,j}/M = m, I_i\right)$ by $S_{m,i,j}$. The probability $P\left(M = m/\rho\right)$ is a function of ρ as described in equation (3.2). Therefore, by the theorem of total probability,

$$P\left(O_j = o_{i,j}/I_i, \rho\right) = \sum_{m=0}^{2^V-1} S_{m,i,j} P\left(M = m/\rho\right) \tag{3.3}$$

In our example, we will proceed by assuming that the observable variables (the O_j's) are independent given the faulty network and the drug combinations. This assumption can be easily relaxed for the case when the 7 observable variables represented as arrows at the bottom of figure 2.1 are observed together for each drug combination used as the input. In this case, instead of $P\left(O_j/M\right)$, we will be working with $P\left(O_1, O_2, ..., O_7/M\right)$. This however does not affect our fundamental results and is a simple extension of our example.

Let O represent all of the observed data for all the observable variables and I represent the entire set of the corresponding inputs. Let J be the number of observable variables and N be the number of observations for each observable variable. Then

we have

$$P\left(O/\rho, I\right) = \prod_{j=1}^{J} \prod_{i=1}^{N} P\left(O_j = o_{i,j}/I_i, \rho\right) \tag{3.4}$$

which is nothing but the likelihood function. In order to handle experimental repeats, we can have the the drug combinations I_i to be the same for more than one value of the index i.

An estimate of ρ can be obtained from equation 3.4, either by maximum likelihood estimation, or by calculating the posterior mean of the parameters. If the prior distributions of all the elements of ρ are assumed to be uniformly distributed between 0 and 1, the posterior distribution of ρ is directly proportional to $P\left(O/\rho, I\right)$. If $P(O/\rho, I)$ comes out to be zero for all values of ρ, then we have every reason to question the validity of the Boolean network used to model the behavior of the biological network, or the set of possible locations of faults. Various estimates of ρ, such as the posterior mean or the posterior mode (the value of ρ where the posterior distribution is maximal) can be obtained from $P\left(O/\rho, I\right)$. Now we can algebraically expand the right hand side of equation (3.4) to write $P\left(O/\rho, I\right)$ as

$$P\left(O/\rho, I\right) = \sum_{k} \prod_{v=1}^{V} \rho_v^{Q1_{v,k}} \left(1 - \rho_v\right)^{Q2_{v,k}} \tag{3.5}$$

where $Q1_{v,k}$ and $Q2_{v,k}$ are non negative integers. Calculating $P\left(\rho/O, I\right)$ from $P\left(O/\rho, I\right)$ is now trivial since it only involves calculation of a multiplicative normalization constant.

$$P\left(\rho/O, I\right) = \frac{P\left(O/\rho, I\right)}{\int P\left(O/\rho, I\right) d\rho} \tag{3.6}$$

where in the denominator there is the normalization constant which turns out to be

$$\int P\left(O/\rho, I\right) d\rho = \sum_{k} \prod_{v=1}^{V} \beta(Q1_{v,k} + 1, Q2_{v,k} + 1) \tag{3.7}$$

where $\beta(*,*)$ is the beta function. This equation is derived by considering a uniform prior on all the elements of ρ. The integrations can be done easily because of the form of equation 3.5. Each variable ρ_v is integrated from 0 to 1. Equation 3.6 shows the joint posterior distribution of all the unknown parameters ρ_1 through ρ_V considered together. In order to find the marginal distribution of any given parameter of interest, we will need to integrate out the rest of the parameters. For example $P(\rho_l/O, I)$ for any given value of l can be found out to be

$$P(\rho_l/O, I) = \frac{\sum_k \rho_l^{Q1_{l,k}} (1-\rho_l)^{Q2_{l,k}} \prod_{\substack{v=1 \\ v \neq l}}^{V} \beta(Q1_{v,k}+1, Q2_{v,k}+1)}{\sum_k \prod_{v=1}^{V} \beta(Q1_{v,k}+1, Q2_{v,k}+1)} \tag{3.8}$$

Following this the posterior means can also be calculated.

However the number of additive terms in equation 3.5 represented by the summing variable k in general rises exponentially with the number of data points collected. In the worst case, the left hand side of equation (3.3) will contain 2^V terms. Since the number of multiplicative terms in equation 3.4 is NJ (the number of data points collected), upon expanding the right hand side of equation 3.4 we get 2^{VNJ} additive terms in equations 3.5, 3.7, and 3.8. Thus the computational cost to compute the mean of any given ρ_l is $O(2^{VNJ})$. Hence the total computation cost to compute the posterior means of all the elements of ρ (ρ_1 through ρ_V) is $O\left(V \times 2^{VJN}\right)$. Therefore the straightforward approach for calculating the posterior distributions of ρ_l's and their posterior means will get intractable as the amount of data collected increases.

To get around this difficulty we will use an iterative algorithm to obtain an approximation of the marginal distributions of the elements of the parameter vector ρ. From the marginal distribution it will be straightforward to obtain the posterior means and confidence intervals of the individual elements of ρ.

3.3 Factor graph representation of the model

Factor Graphs are an important tool used in various applications such as signal processing and telecommunications. Many algorithms can be easily understood and derived using the factor graph approach. These include Kalman Filters, the Viterbi Algorithm, the Forward-Backward algorithm and Turbo Codes to name a few. The approach involves first representing the probability model as a factor graph and then applying the message passing algorithm along the edges. The reader is referred to [17] and [39] for an in-depth coverage of factor graphs and the message passing algorithm. Here we provide a short primer to the subjects and go into the details of only our particular example.

3.3.1 A simple example

Consider a simple function $g(x_1, x_2, x_3) = f_1(x_1, x_2) \times f_2(x_2, x_3) f_3(x_3)$, where x_i are discrete variables. Suppose we want to calculate $\sum_{x_1,x_3} g(x_1, x_2, x_3)$ for a particular value of x_2 (the marginal of x_2). In addition, suppose that each x_i can take A different values. Hence the straight forward approach would require us to sum $g(x_1, x_2, x_3)$ over A^2 different values. However $\sum_{x_1,x_3} g(x_1, x_2, x_3)$ can also be calculated as

$$\sum_{x_1,x_3} g(x_1, x_2, x_3) = \left(\sum_{x_1} f_1(x_1, x_2)\right)\left(\sum_{x_3} f_2(x_2, x_3) f_3(x_3)\right) \tag{3.9}$$

which sums over $2A$ different values. For continuous variables, the summation is replaced by integration. The optimal strategy for calculating the marginal of x_2 is straightforward to derive in this simple example. However a systematic approach to find the optimal strategy to calculate the marginal of any variable for any given probability function is given by the message passing algorithm which acts on the

factor graph representation of the function.

The factorization of a function can be represented by a factor graph. A factor graph is a bipartite graph with a variable node corresponding to each variable x_i and a factor node corresponding to each independent factor f_j and has an undirected edge connecting a variable node of x_i to a factor node of f_j iff x_i is an argument of f_j [17, 39]. The factor graph of $g(x_1, x_2, x_3)$ is shown in figure 3.1. Messages pass along the edges in both directions. Messages are functions of the variable whose node is associated with the edge. Let $\mu_{f_j \to x_i}(x_i)$ and $\mu_{x_i \to f_j}(x_i)$ denote the messages from f_j to x_i and vice versa. We simply write down the update equations below. For an in-depth discussion on their derivation, the reader is referred to [17] and [39]. The messages are calculated as follows:

$$\mu_{x_i \to f_j}(x_i) = \prod_{h \in n(x_i) \setminus \{f_j\}} \mu_{h \to x_i}(x_i) \tag{3.10}$$

$$\mu_{f_j \to x_i}(x_i) = \sum_{\sim\{x_i\}} \left(f_j(X) \prod_{y \in n(f_j) \setminus \{x_i\}} \mu_{y \to f_j}(y) \right) \tag{3.11}$$

where $n(x_i)$ and $n(f_j)$ denote the neighbors of x_i and f_j respectively in the factor graph. $n(x_i) \setminus \{f_j\}$ represents the set of all the neighbors of x_i except f_j. The definition of $n(f_j) \setminus \{x_i\}$ is similar. Since the factor graph is bipartite, the neighbors of a variable node can only be factor nodes, and the neighbors of a factor node can only be variable nodes. X denotes the set of arguments of f_j. $\sum_{\sim\{x_i\}}$ denotes summation over all local variables except x_i. The set of local variables will simply be the set X, since the factor node f_j is connected by undirected edges only to the variable nodes of its arguments. The message going away from a leaf variable node is the constant 1, while the message going away from a leaf factor node is the value of

that local factor. Using these rules on the simple example, we have $\mu_{x_1 \to f_1}(x_1) = 1$ and $\mu_{f_3 \to x_3}(x_3) = f_3(x_3)$.

The marginal distribution of a variable is simply the product of all the messages being received by the corresponding variable node. Hence $\sum_{x_1,x_3} g(x_1, x_2, x_3) = \mu_{f_1 \to x_2}(x_2) \times \mu_{f_2 \to x_2}(x_2)$ and thus equation (3.9) is derived using factor graphs and the message passing algorithm. Calculating the rest of the messages would allow us to calculate the marginals of x_1 and x_3 as well. The message passing algorithm would terminate when messages along both directions of all the edges in the graph have been calculated.

$$\mu_{f_1 \to x_2}(x_2) = \sum_{\sim\{x_2\}} f_1(x_1, x_2)\, \mu_{x_1 \to f_1}(x_1) = \sum_{x_1} f_1(x_1, x_2) \tag{3.12}$$

$$\mu_{x_3 \to f_2}(x_3) = \prod_{h \in n(x_3) \setminus \{f_2\}} \mu_{h \to x_3}(x_3) = \mu_{f_3 \to x_3}(x_3) = f_3(x_3) \tag{3.13}$$

$$\mu_{f_2 \to x_2}(x_2) = \sum_{\sim\{x_2\}} f_2(x_2, x_3)\, \mu_{x_3 \to f_2}(x_3) = \sum_{x_3} f_2(x_2, x_3)\, f_3(x_3) \tag{3.14}$$

The message passing algorithm terminates and gives exact marginals for the cases where the factor graph has no cycles. But the most interesting applications are for those cases where the factor graph has cycles, where the marginals are calculated by iteratively updating the messages (for example the iterative decoding of turbo codes). We similarly use an iterative version of the message passing algorithm in our model to approximate the marginal posterior distribution of the unknown parameters.

3.3.2 Using factor graphs and the message passing algorithm on the signal transduction network model

Now, $P(\rho/O, I) \propto P(O/\rho, I)$ as is evident from equation (3.6), while the expression for $P(O/\rho, I)$ is given in equation (3.4). Let $P_{i,j}$ represent the multiplicative factor $P(O_j = o_{i,j}/I_i, \rho)$ in equation (3.4). In a factor graph, each multiplicative factor is represented by a factor node and each element of ρ is represented by a variable node. Hence there are NJ number of factor nodes with each corresponding to one particular multiplicative term in equation 3.4, and there are V number of variable nodes with each corresponding to one particular unknown parameter (one out of ρ_1 through ρ_V). The purpose of this algorithm is to compute the posterior marginal distributions of the unknown parameters ρ_1 through ρ_V, which can then be used to compute their means and confidence intervals.

Figure 3.2 shows the factor graph of equation (3.4). As we can see the factor graph in figure 3.2 has cycles. In a factor graph with cycles, the message passing algorithm does not terminate and the messages are locally updated with every iteration. Every time a new message is calculated, it replaces the old message. The iterative message passing algorithm is as follows:

1. initialize all $\mu_{\rho_v \to P_{i,j}}(\rho_v) = 1$
2. calculate all $\mu_{P_{i,j} \to \rho_v}(\rho_v)$ as per equation (3.11).
3. calculate all $\mu_{\rho_v \to P_{i,j}}(\rho_v)$ as per equation (3.10).
4. repeat steps 2 and 3 in that order.

Since we are dealing with continuous variables between 0 and 1, the summations are replaced by integrations. Every time $\mu_{P_{i,j} \to \rho_v}(\rho_v)$ are computed in step 2, they come

out to be polynomials of degree one due to the multiplicatively separable nature of the integrands involved and that all the parameters ρ_v are being integrated from 0 to 1 (a rectangular integration region). Let them be represented as $b_{0,v,i,j} + b_{1,v,i,j} \times \rho_v$. Hence $\mu_{P_{i,j} \to \rho_v}(\rho_v)$ can be represented by a vector $b_{v,i,j} = \left(\begin{smallmatrix} b_{0,v,i,j} & b_{1,v,i,j} \end{smallmatrix} \right)^T$. Every time $\mu_{\rho_v \to P_{i,j}}(\rho_v)$ are computed in step 3, they will be polynomials of degree $NJ-1$ since they are simply the product of all incoming messages except one. Let them be represented as $\sum_{k=0}^{NJ-1} a_{k,v,i,j} \rho_v^k$. Hence $\mu_{\rho_v \to P_{i,j}}(\rho_v)$ can be represented by a vector $a_{v,i,j} = \left(\begin{smallmatrix} a_{0,v,i,j} & a_{1,v,i,j} & \cdots & a_{NJ-1,v,i,j} \end{smallmatrix} \right)^T$.

The values $b_{0,v,i,j}$ and $b_{1,v,i,j}$ can be updated in step 2 as follows.

$$b_{0,v,i,j} \leftarrow \sum_{m=0}^{2^V-1} S_{m,i,j}(1 - R_{v,m}) \times \prod_{\substack{l \epsilon \{1 \ldots V\} \\ l \neq v}} \left(\sum_{k=0}^{NJ-1} \frac{a_{k,l,i,j}}{k+2} \right)^{R_{l,m}} \times \left(\sum_{k=0}^{NJ-1} \frac{a_{k,l,i,j}}{(k+1)(k+2)} \right)^{1-R_{l,m}} \tag{3.15}$$

$$b_{1,v,i,j} \leftarrow \sum_{m=0}^{2^V-1} S_{m,i,j}(2R_{v,m} - 1) \times \prod_{\substack{l \epsilon \{1 \ldots V\} \\ l \neq v}} \left(\sum_{k=0}^{NJ-1} \frac{a_{k,l,i,j}}{k+2} \right)^{R_{l,m}} \times \left(\sum_{k=0}^{NJ-1} \frac{a_{k,l,i,j}}{(k+1)(k+2)} \right)^{1-R_{l,m}} \tag{3.16}$$

The $a_{v,i,j}$ can be updated in step 3 by performing polynomial multiplications of the $NJ-1$ incoming first degree polynomials to the $v'th$ variable node and comparing

coefficients. That is, the following equation must be satisfied.

$$\sum_{k=0}^{NJ-1} a_{k,v,i,j}\rho_v^k = \prod_{g\neq i, h\neq j} (b_{0,v,g,h} + b_{1,v,g,h} \times \rho_v) \tag{3.17}$$

By comparing coefficients of either side of equation (3.17), the values of the elements of the vector $a_{v,i,j}$ are updated. This is also equivalent to the convolution of the message vectors $b_{v,g,h}$ for $g \neq i, h \neq j$. At each iteration, the message vectors can be multiplied by constants so as to prevent overflow or underflow when implementing the algorithm on a digital computer with finite precision. In that case the final solutions we get are simply the required marginal distributions scaled by some unknown constant. If we are simply interested in the marginal distributions, then it is not necessary to keep track of the multiplied constants. We simply need to normalize the marginals so that their integrals from 0 to 1 give unity.

The message vectors $a_{v,i,j}$ and $b_{v,i,j}$ are iteratively updated until some convergence criteria is satisfied (for example if the Hellinger distance between the marginals of two successive iterations is below a certain threshold). In our simulations, we saw that as few as 2 iterations gave satisfactory results in terms of convergence. Hence the time complexity of the algorithm is dependent on steps 2 and 3 of the algorithm.

In order to calculate $\mu_{\rho_v \to P_{i,j}}(\rho_v)$ in step 3, first calculate the polynomial $U_v(\rho_v) = \prod_{g,h} \mu_{P_{g,h}\to\rho_v}(\rho_v)$ of degree NJ. Then find the quotient of the division operation $U_v(\rho_v) \div \mu_{P_{i,j}\to\rho_v}(\rho_v)$. This gives $\mu_{\rho_v\to P_{i,j}}(\rho_v)$. Along with that, we can also calculate and store the value of $\theta_{v,i,j,1} = \int_0^1 \rho_v \mu_{\rho_v\to P_{i,j}}(\rho_v)d\rho_v$ and $\theta_{v,i,j,0} = \int_0^1 (1-\rho_v)\mu_{\rho_v\to P_{i,j}}(\rho_v)d\rho_v$ which will be used in step 2. Note that $\theta_{v,i,j,1} = \sum_{k=0}^{NJ-1} \frac{a_{k,v,i,j}}{k+2}$ and $\theta_{v,i,j,0} = \sum_{k=0}^{NJ-1} \frac{a_{k,v,i,j}}{(k+1)(k+2)}$. Calculating the coefficients of $U_v(\rho_v)$ is of time complexity at most $O((NJ)^2)$. This is because it involves the convolution of NJ different first degree polynomials. Calculating the quotient of $U_v(\rho_v) \div \mu_{P_{i,j}\to\rho_v}$, and $\theta_{v,i,j,1}$

and $\theta_{v,i,j,0}$ are of time complexity $O(NJ)$. The last three operations of $O(NJ)$ have to be done for all NJ of the factor nodes for each variable node. Hence the time complexity of calculating the messages from one variable node to all factor nodes is of time complexity $O((NJ)^2)$. Repeating this action for all V variable nodes gives us the time complexity of step 3 of the algorithm to be $O((NJ)^2V)$.

If we look at equations (3.15) and (3.16), the computation of $b_{v,i,j}$ seems to be of $O\left(NJV2^V\right)$ time complexity. Since there are NJV of $b_{v,i,j}$ to be computed, step 2 seems to be of $O\left((NJ)^2(V)^2 2^V\right)$ time complexity. However some of the computations are repeated and storing these computations for reuse can reduce the time complexity. Let $\kappa_{m,i,j} = \prod_{l=1}^{V} \theta_{l,i,j,1}^{R_{l,m}} \theta_{l,i,j,0}^{1-R_{l,m}}$. Then $\mu_{P_{i,j}\to\rho_v}(\rho_v) = \sum_{m=0}^{2^V-1} S_{m,i,j} \rho_v^{R_{v,m}} (1-\rho_v)^{1-R_{v,m}}$ $\times \frac{\kappa_{m,i,j}}{\theta_{v,i,j,1}^{R_{v,m}} \theta_{v,i,j,0}^{1-R_{v,m}}}$. Computation of $\kappa_{m,i,j}$ for all m is of $O(V2^V)$ time complexity for a given factor node $P_{i,j}$. Computation of $\mu_{P_{i,j}\to\rho_v}(\rho_v)$ for all v is of $O(V2^V)$ time complexity for a given factor node $P_{i,j}$. Hence computation of $\mu_{P_{i,j}\to\rho_v}(\rho_v)$ from a single factor node to all variable nodes is of $O(V2^V)$ time complexity. Hence total computation for all factor nodes in step 2 comes out to be of $O(NJV2^V)$ time complexity.

Hence the complexity of each iteration of the algorithm comes out to be $O(NJV(2^V + CNJ))$, where C is a constant. This is quadratic with respect to the number of data points NJ, as opposed to the exponential complexity of the straightforward approach discussed in the previous sections section.

Once the convergence criteria is met and the algorithm is terminated, the marginal distribution of ρ_v is calculated as

$$P(\rho_v/O, I) = \gamma \prod_{i,j} \mu_{P_{i,j}\to\rho_v}(\rho_v) \tag{3.18}$$

where γ is a normalization constant which can be calculated to give $\int_0^1 P(\rho_v/O, I) d\rho_v = 1$.

3.4 Simulation experiments

We did simulations where the algorithm was tested on synthetic data as well as applied to real world data. The marginal posterior distributions estimated using the iterative message passing algorithm were compared with the marginal posteriors estimated using the time consuming and computationally intensive Markov Chain Monte Carlo (MCMC) methods and the estimates obtained using both methods came out to be close thereby verifying the iterative message passing algorithm's correctness.

Various literature on MCMC methods exist [12, 13, 16]. We will describe the details used in our simulations instead of going into a detailed discussion of MCMC methods. The Markov Chain Monte Carlo Method involves creating a Markov Chain whose stationary distribution is the required posterior distribution. The Metropolis-Hastings Algorithm will be used to generate such a Markov Chain since the samples need to be generated from a non standard probability distribution. This method will be used to generate samples from the posterior distribution of the unknown parameters of the vector ρ. These samples can then be used to get an estimate of the joint as well as the marginal posterior distributions of the unknown parameters using kernel density estimation.

Samples are drawn from the posterior distribution of ρ using the Metropolis-Hastings (MH) Algorithm in the following manner. Let the n^{th} sample drawn from the posterior distribution of ρ be $\rho^{(n)} = (\rho_1^{(n)} \ \rho_2^{(n)} \ \ldots \ \rho_V^{(n)})$.

1. Initialize all elements of $\rho^{(0)}$ to be 0.5.

2. At the n^{th} iteration of the MH algorithm, generate ρ^* from the proposal distribution $U(\rho/\rho^{(n)}, \Delta)$. The proposal distribution and the tuning parameter Δ will be discussed in the next paragraph.

3. Calculate the acceptance ratio

$$D = \frac{P(O/\rho^*, I)U(\rho^{(n)}/\rho^*, \Delta)}{P(O/\rho^{(n)}, I)U(\rho^*/\rho^{(n)}, \Delta)}$$

(Recall that the prior of the parameter vector is constant). $P(O/\rho^*, I)$ and $P(O/\rho^{(n)}, I)$ can be easily calculated for known values of ρ^* and $\rho^{(n)}$ without the expansion of $P(O/\rho, I)$ described in equation (3.5). Accept ρ^* as the next sample $\rho^{(n+1)}$ with probability $min(1, D)$, or keep $\rho^{(n+1)}$ equal to $\rho^{(n)}$ with probability $1 - min(1, D)$.

4. Repeat steps 2 and 3 to generate samples from the posterior of $P(\rho/O, I)$.

The proposal distribution $U(\rho/\rho^{(n)}, \Delta)$ is such that ρ_i is Beta distributed with parameters $\frac{\rho_i^{(n)}}{\Delta}$ and $\frac{1-\rho_i^{(n)}}{\Delta}$, that is

$$U(\rho/\rho^{(n)}, \Delta) = \prod_{i=1}^{V} \frac{\rho_i^{\frac{\rho_i^{(n)}}{\Delta}-1}(1-\rho_i)^{\frac{1-\rho_i^{(n)}}{\Delta}-1}}{Beta(\frac{\rho_i^{(n)}}{\Delta}, \frac{1-\rho_i^{(n)}}{\Delta})} \tag{3.19}$$

where $Beta(x, y)$ is the beta function with parameters x and y and Δ is a scalar tuning parameter which controls the variance of the distributions of the ρ_i's. It can be adjusted to give autocorrelation properties of the Markov Chain within acceptable ranges.

3.4.1 Experiments with synthetic data

To demonstrate the working of the algorithm, we ran simulations of the message passing algorithm as well as the MH algorithm on synthetic data. We generated synthetic data from the example described in section II which was derived from the MAPK signal transduction network, which is a well understood network.

The set of locations where faults can take place was taken to be composed of

RAF, IRS1, and RHEB. The probabilities of stuck-at-one faults at these locations (The parameters ρ_1, ρ_2, and ρ_3) were taken as 0.7, 0.4, and 0.2. Synthetic observations of the observable variable (the variables shown at the bottom of figure 2.1 as arrows) were generated for various drug combinations as inputs (the drugs being AG1024, AG825, Lapatinib, LY294002, U0126, and Temsirolimus, whose action on the Boolean network of the MAPK network is shown in figure 2.1) according to the probability model described in the previous sections. The inputs at the top of the network corresponding to growth factors (EGF, HBEGF, IGF, and NRG1) were all taken as 1 (if the cells were being grown on petridishes, then this would be equivalent to the case where all the four growth factors have been supplied in the serum). Hence the data set $\{(o_{i,1}, o_{i,2}, ..., o_{i,J}), I_i\}$ is generated. There are 6 drugs in the Boolean model. All the $2^6 - 1$ drug combinations were used to generate the data points. Hence i varies from 1 to 63.

After the synthetic data set was generated, the marginal posterior distributions of the elements of ρ (The parameters ρ_1, ρ_2, and ρ_3) were estimated using both the message passing algorithm as well as the MCMC method. For the MCMC method, the tuning parameter Δ is set to 0.04 which gives an acceptance rate of 40%. The reader is referred to [16] for information on acceptance rates. Then the Markov Chain was run to generate 50,000 samples to attain stationarity (the burn in period). Following this, the Markov chain was run long enough to generate 250,000 samples and thinned by a factor of 50 (one in 50 samples generated was stored for each parameter) resulting in 5000 samples for each ρ_v. This resulted in effective sample sizes of atleast 4000 for each of the ρ_i's. the reader is referred to [16] for information on effective sample sizes. The algorithms were implemented in MATLAB. The message passing algorithm was terminated after 2 iterations which took about 4 seconds. For our purposes, we used the Hellinger Distance between the marginals of the first pa-

rameter ρ_1 calculated at consecutive iterations of the message passing algorithm to fall below a certain threshold to signal termination of the algorithm. However other convergence criterions could also be used. The MCMC samples were generated in 30 minutes after the initial burn in period. The marginal posterior distribution of ρ_1 through ρ_3 calculated using both the message passing algorithm and the MCMC approach are shown in figure 3.3. Kernel density estimation with a Gaussian Kernel was used to estimate the marginals from the sample values generated using the MH algorithm. The estimate $\hat{P}(\rho_v/O, I)$ of $P(\rho_v/O, I)$ is calculated from the samples as follows

$$\hat{P}(\rho_v/O, I) = \frac{1}{L}\sum_{n=1}^{L} \frac{1}{\sqrt{2\pi}\sigma_v} exp\left(-\frac{(\rho_v - \rho_v^{(n)})^2}{2\sigma_v^2}\right) \tag{3.20}$$

where σ_v is the bandwidth of the Gaussian kernel which is set to $\frac{\delta_v}{L^{\frac{1}{5}}}$. L is the number of samples generated by the MH algorithm (5000 in our case) and δ_v is the standard deviation of the generated samples. This rule of thumb to calculate the bandwidth of the Gaussian kernel is discussed in [33].

As we can see in figure 3.3, there is almost no difference in the inference of the marginal posterior distributions of the unknown parameters between the message passing algorithm and the MCMC approach. The posterior mean of ρ_v is calculated from the message passing algorithm as $\int_0^1 \rho_v \gamma \prod_{i,j} \mu_{P_{i,j}\to\rho_v}(\rho_v) d\rho_v$ and from the MCMC approach as $\frac{1}{L}\sum_n \rho_v^{(n)}$. These come out to be $(\,0.7254\ 0.3891\ 0.2799\,)$ and $(\,0.7326\ 0.3961\ 0.2830\,)$ respectively. These estimates are close to each other and to the actual values of $(\,0.7\ 0.4\ 0.2\,)$.

This simulation shows that the message passing algorithm successfully calculates the posterior marginal distributions of the unknown parameters ρ_1 through ρ_3 and

gives the same inferences as the Metropolis-Hastings algorithm. We did simulations with various values of ρ and for different sets of locations of faults. The iterative message passing algorithm gave estimates of the posterior marginal distributions of the parameters same as those estimated using the MCMC approach for all the test cases considered in our simulations.

3.4.2 Applications to real data

To test our model, we performed experiments on healthy adult fibroblasts where it is fair to assume that there are no cancer causing mutations present in the tissue. Hence it is fair to assume that a Boolean regulatory network with no faults would best model this tissue.

Adult fibroblasts were grown in petri-dishes till confluence and then maintained in 0.2% FBS (Fetal Bovine Serum) for four days. It is a general assumption that FBS contains most of the important growth factors. After this, the cells were exposed to 0.2% FBS and 100μM Anisomycin for 30 minutes. Anisomycin is a protein synthesis inhibitor which activates the MAPK signal transduction network and keeps it responsive to kinase specific inhibitors [2, 10]. That is, with the addition of Anisomycin, we anticipate the MAPK signal transduction network to respond to the addition of kinase inhibitors such as U0126. Anisomycin, being a protein synthesis inhibitor, would also cut off any feedback path which has a translation (protein synthesis) step in it. The cells were then grouped into three groups (group 0, group 1, and group 2). Group 0 was the control group which was exposed to 100μM Anisomycin only. Group 1 was exposed to 100μM Anisomycin and 50μM of LY294002. Group 2 was exposed to 100μM Anisomycin, 50μM of LY294002, and 10μM of U0126. All three groups were also exposed to 20% FBS along with the other chemicals. LY294002 and U0126 are highly specific inhibitors of PI3 Kinase (PI3K in Figure 2.1) and

MEK1 respectively. The molecular targets of LY294002 and U0126 are shown in Figure 2.1. Genes having the SP1 and SRF-ELK response elements in their promoters were quantified through real time PCR and the delta-delta method [22] with GAPDH as the reference gene and group 0 as the control. The genes were measured in quadruplets for each experiment.

EGR1 is measured as a reporter gene of SRF-ELK transcription factor [8]. JUN, and cMYC are measured as reporters of SP1 [31, 20]. Other genes having the SP1 response element in their promoters are Decorin, IRF3 and VEGFA [35, 40, 32, 29]. These six genes were quantified in quadruplets for each experiment. The readings of each gene are discretized using Otsu's method [28]. As an example the readings of ERG1 and their corresponding discretized values are shown in table 3.1. The threshold level for EGR1 came out to be 0.3824. the expressions above this level are labeled as 1 and those below are labeled as 0. The measured normalized gene expression ratios are shown in table 3.2.

For demonstration purposes, we have taken the set of locations where to search for faults to be composed of ERK1/2 and IRS1 (shown in Figure2.1). The marginal posterior probability distributions of the probabilities of faults associated with these two locations are shown in figure 3.4.

As we can see in figure 3.4, the posterior marginal distribution associated with ERK1/2 comes out to be quite tightly distributed with a mean of 0.1538 while that for IRS1 comes out to be uniformly distributed between 0 and 1. This is because the data does not contain any discriminating information about the occurrence of any fault at IRS1 under this MAPK Boolean model. But it does tell us that the probability of occurrence of a fault at the variable corresponding to ERK1/2 is pretty low, judging by its mean to be having a low value of close to 15%. This is expected since the data comes from adult fibroblasts, where we can be fairly sure that no cancer

causing mutations are present. If data had been collected after exposure to other combinations of other drugs (for instance Lapatinib or Temsirolimus) then the data might have allowed the model to make meaningful inferences regarding occurrences of faults at locations besides ERK1/2 as well as give sharper confidence intervals than that shown in figure 3.4.

3.5 Summary and comments on possible future work

In this work we have described a method to estimate the probabilities with which certain faults have taken place in a given Boolean Regulatory network, provided we have the observations of the observable variables whose behavior is determined by the network. We have described the probability model and described a fast algorithm based on message passing to make the inferences about the posterior marginal probability distributions of the unknown parameters of the model (These parameters being the probabilities of the occurrences of the faults). We have compared the performance of the algorithm with Markov Chain Monte Carlo techniques (the Metropolis-Hastings Algorithm) through simulations, and we have shown that the message passing algorithm gives results comparable to those obtained using the MCMC methods with the added advantage of much smaller computation times. We also applied the model to analyze data collected from fibroblasts, thereby demonstrating how this model can be used on real world data. Such a computationally manageable approach has the potential to allow the inference of locations of faults in a Boolean regulatory network in a probabilistic setting from data, such as gene expression data. For further information on this work, the reader is referred to [26].

Locating the points of dysregulations in a deterministic Boolean signal transduction network could be used to suggest therapies as described in [18]. Since we are locating faults in a probabilistic setting, the therapy could be designed keeping

in mind the tradeoff between treating cancer and managing the side effects of the treatment. For example, consider a case where we have two possible locations of faults. Let the computed probability of the occurrence of a fault at the first location be smaller than that of the second location. Then we may only consider the second fault in our therapy design process, thereby reducing the exposure of the patient to excessive drugs which may have unwanted side effects.

Future work could focus on performing experiments on cancerous cell lines being exposed to various combinations of drugs and infer from the collected data the likely locations of dysregulations in the corresponding Boolean regulatory network. Also, algorithms could be developed to automate the process of selecting the set of locations of faults instead of having the user provide it to the algorithm.

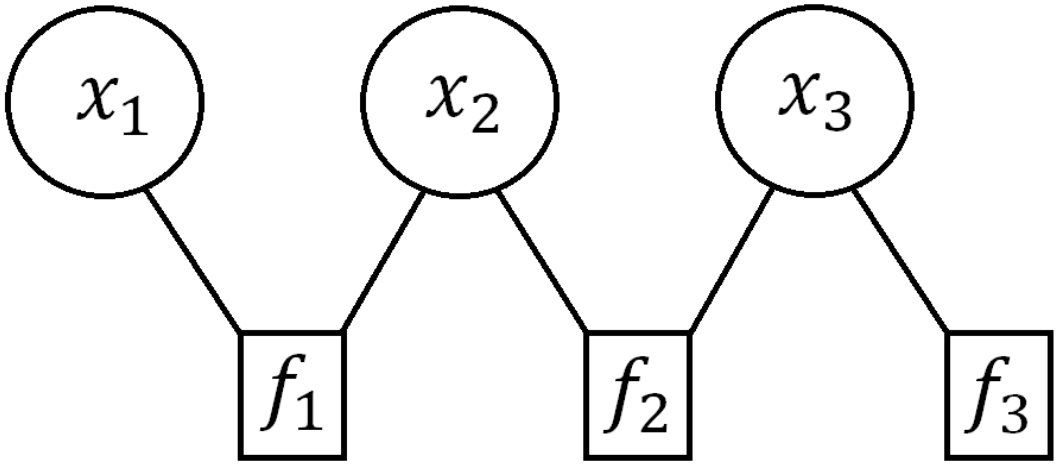

Figure 3.1: The factor graph representation of a factorizable function. The variable nodes are circular and the factor nodes are rectangular.

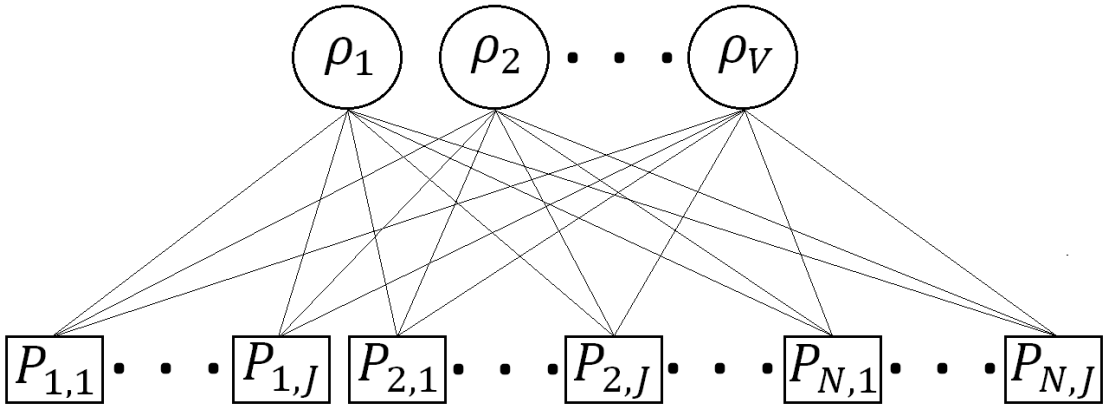

Figure 3.2: The factor graph representation of the probability model of the signal transduction network. The variable nodes are circular and the factor nodes are rectangular.

Table 3.1: Gene expression levels and their discrete values for the gene EGR1. The threshold level using Otsu's method comes out to be 0.3824 for EGR1.

group 1	*normalized gene expression*	0.5987	0.7320	0.5586	0.6199
	discrete value	1	1	1	1
group 2	*normalized gene expression*	0.4796	0.2892	0.2535	0.2698
	discrete value	1	0	0	0

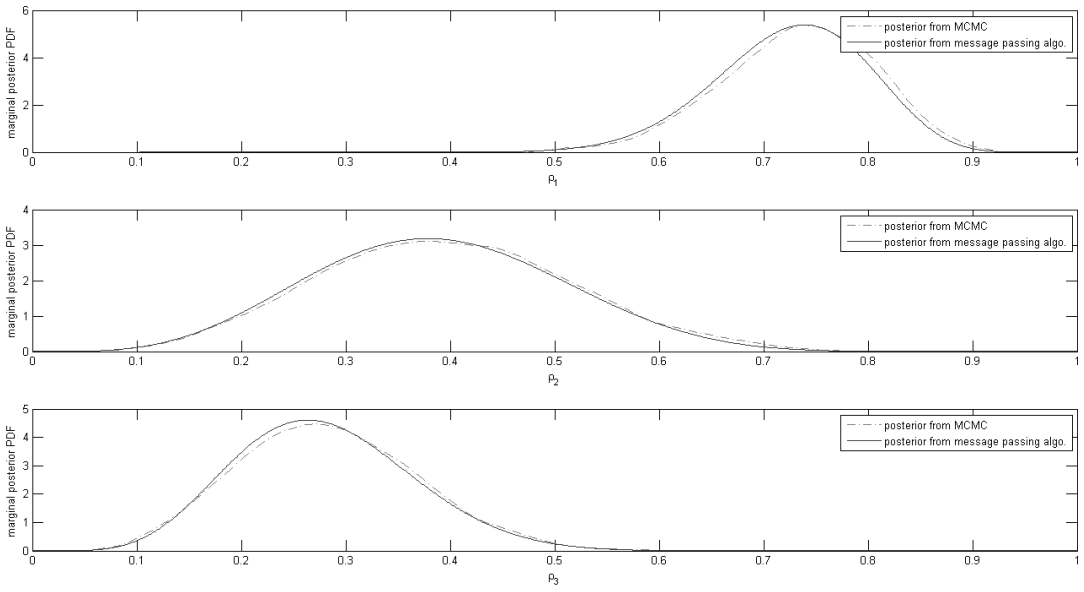

Figure 3.3: Marginal posterior distribution of ρ_1 through ρ_3 calculated using both the message passing algorithm and the MCMC approach.

Table 3.2: Table showing the normalized gene expression ratios and their reference sequence (RefSeq) numbers.

	EGR1	JUN	CMYC	DECORIN	IRF3	VEGFA
RefSeq	NM_001964.2	NM_002228.3	NM_002467.4	NM_133503.2	NM_001571.5	NM_003376.5
Group 1	0.5987	0.4931	0.3209	0.4353	0.5176	0.4444
	0.7320	0.6736	0.2852	0.4601	0.4204	0.4989
	0.5586	0.6598	0.3439	0.4147	0.3560	0.5176
	0.6199	0.7792	0.2994	0.4323	0.3345	0.5105
Group 2	0.4796	0.1550	0.2570	0.2793	0.2624	0.3164
	0.2892	0.2793	0.2059	0.3789	0.2553	0.4601
	0.2535	0.3015	0.2717	0.3737	0.2253	0.4633
	0.2698	0.3415	0.2679	0.3536	0.2031	0.3660

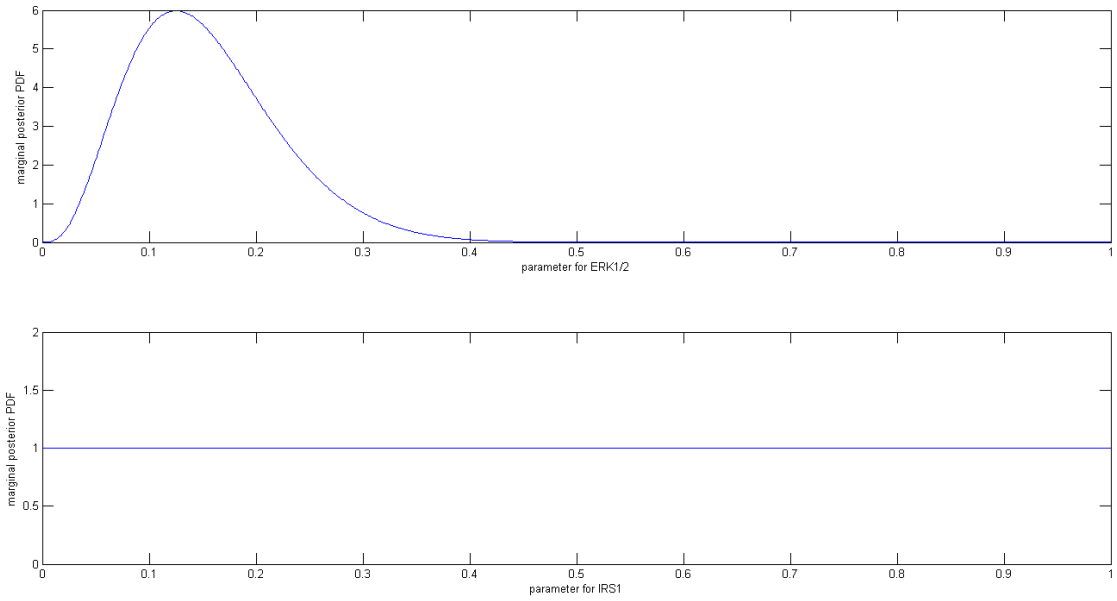

Figure 3.4: Marginal posterior distribution of the unknown parameters associated with ERK1/2 and IRS1.

4. A CONJUGATE EXPONENTIAL MODEL FOR CANCER TISSUE HETEROGENEITY *

4.1 Introduction

In the previous sections we have discussed how the clonal evolution of cells makes most neoplastic tissues heterogeneous. Hence it becomes important to incorporate heterogeneity into the modeling of gene regulatory networks which are to be used in the study of cancerous tissues, especially those oriented towards developing effective therapies for cancer treatment. An attempt was made in [25] to model cancer tissue heterogeneity. In that paper, the authors used a collection of Boolean Networks to model the various subpopulations in a given tissue. A multilevel hierarchical model was used to model the extent to which each Boolean network affects the behavior of each of the observed gene expressions. The authors demonstrated the use of this model by applying it to gene expression measurements from healthy fibroblasts when they were exposed to various stimuli. Markov Chain Monte Carlo Methods were used to estimate the posterior probability distributions of the unknown parameters which would indicate the proportion wise breakup of the tissue under study. In this paper we make certain approximations to the model which would allow us to employ faster (variational) methods to carry out the same estimation.

It has been discussed how prior knowledge can be used to model gene regulatory networks in the form of Boolean networks and how these Boolean networks can be used to design combination therapies [19, 18]. It has also been discused how an ensemble of Boolean networks can be used to represent a heterogeneous cancer tissue

*Parts of this section are reprinted with permission from "A Conjugate Exponential Model for Cancer Tissue Heterogeneity" by A. K. Mohanty, A. Datta, and V. Venkatraj, 2015. *IEEE Journal of Biomedical and Health Informatics*, preprint, © 2015 IEEE. doi:10.1109/JBHI.2015.2410279.

in previous sections and also in [25]. In this model, the subpopulations or networks in the collection of chosen networks exert their effect on the observable variables (the gene expression ratios) in a weighted average fashion. The objective was to find these weights associated with each subpopulation or network in the ensemble. This problem was solved using Markov Chain Monte Carlo (MCMC) methods. In this paper, we will address this problem of finding out these weights in a variational Bayes framework resulting in a significant speed-up of the computational time.

Prior knowledge about the qualitative location of faults in the network can be used to determine the initial model and which networks to choose in the ensemble. For instance, in 30% of human breast cancers we see an over expression of the ERBB2 gene [37]. This can be interpreted as a stuck-at one fault at the variable corresponding to ERBB2 in the Boolean network in figure 2.1. Another example is that of pancreatic cancer where 90% of the cases show a mutated Ras gene [37] translating to a stuck-at one fault in the corresponding location in the Boolean network. Using such prior knowledge, it is possible to decide which networks to include in the ensemble.

4.2 Methods

Once the networks to be included in the ensemble have been chosen based on prior knowledge, the problem is to estimate the weights associated with each of the networks from collected data. The observable variables can be anything in principle. We will develop our methods based on normalized gene expression ratios. These are real valued readings for each gene. [22] discusses the method to measure normalized gene expression ratios using QPCR. When exposed to a certain stimulus (like a particular combination of kinase inhibitory drugs), some of the output variables, as shown at the bottom of figure 2.1 using arrows, will be up-regulated or "one" for some of the Boolean networks in the ensemble, and some of the output variables will

be down-regulated or "zero" for some of the other networks in the ensemble. For example let us consider an example where we have three networks in the ensemble. This example has also been discussed in [25]. Let the first subpopulation be modeled by a Boolean network with a stuck-at one fault at ERK1/2, let the second subpopulation have two stuck-at-one faults at ERBB2/3 and Raf, and let the final subpopulation have a stuck-at-zero fault at PTEN. The different fault locations corresponding to the different subpopulations are shown as shaded squares in the single Boolean network in Figure 2.1. Suppose we expose the cell culture to the drug U0126. This is a kinase inhibitor which targets MEK1 as shown in Figure 2.1. (All the drugs used in this example are kinase inhibitors whose molecular targets are shown in Figure 1.) Let us also assume that the serum, as typically used in tissue cultures, has EGF, HBEGF, IGF, and NRG1 in it. Hence in other words, the corresponding variables represented at the top of figure 2.1 are all one or upregulated. If we observe the behavior of the transcription factor SP1 (shown at the bottom of the Boolean network in Figure 2.1 with an arrow), the first network has SP1 upregulated while in the second and third networks, SP1 will be downregulated. If we are observing the expression for a gene which has the SP1 response element (such as cMYC), then that gene will be influenced by just the first network to an extent determined by the weight assigned to that network.

We can represent the activities of the different Boolean networks with relation to the i^{th} gene (the three Boolean networks with relation to cMYC in the above example) using a vector $d_i = (\,1\;0\;0\,)^T$, where the subscript i stands for the i^{th} gene. We define this vector as the "expression profile" [25]. This expression profile will depend on the stimulus given to the tissue (the combination of the kinase inhibitor drugs for example) and the networks included in the ensemble. These expression profiles will be provided along with the data. Let the weights associated with the three networks

with relation to the i^{th} gene be represented by a vector $\alpha_i = (\ \alpha_{i,1}\ \alpha_{i,2}\ \alpha_{i,3}\ )^T$. Then if we are considering a model which combines the networks in a weighted average fashion, then the gene expression for the i^{th} gene in the overall model could be quantified by the dot product $d_i^T \alpha_i$. This approach was used in [25], where the normalized gene expression ratios were modeled as a ratio of two normal random variables. This method was an extension of the model described in [7]. Let us say that several measurements of the i^{th} gene were made. Let $d_{i,j}$ be the "expression profile" for the j^{th} measurement of the i^{th} gene. $d_{i,j}$ will depend on the drugs to which the tissue was exposed. Let $r_{i,j}$ be the corresponding measured gene expression ratio for the i^{th} gene. Then [25] derived $P\left(r_{i,j}/\alpha_i, d_{i,j}, c\right)$ as

$$P(r_{i,j}/\alpha_i, d_{i,j}, c) = \frac{m_{i,j}(r_{i,j} + m_{i,j})}{\sqrt{2\pi}c(r_{i,j}^2 + m_{i,j}^2)^{\frac{3}{2}}} \times exp\left(-\frac{1}{2c^2}\frac{(r_{i,j} - m_{i,j})^2}{(r_{i,j}^2 + m_{i,j}^2)}\right) \tag{4.1}$$

where $m_{i,j} = d_{i,j}^T \alpha_i$. The parameter c is the coefficient of variation used to account for the uncertainty in the data. For a detailed derivation of equation 4.1, the reader is referred to [25].

Such a distribution has a mode close to around $m_{i,j}$. Assuming the weights associated with each network to be the same with relation to all the genes being observed is a strong assumption. That is, assuming all the α_i weight vectors to be the same would imply that the Boolean networks affect all the genes with exactly the same ratio. Hence [25] used a multilevel hierarchical model where each of the weight vectors α_i associated with the i^{th} gene is different from the weight vectors associated with the other genes, but they all are derived from an underlying distribution which is an average of all the weight vectors. A schematic diagram of the Bayesian network of the probability model used in [25] is shown in figure 2.2. The parameter K governs the topmost level of the model.

Assuming that there are V different observable variables or genes being measured, and each gene i has n_i observations associated with it (which may come from different experiments), the variables $r_{1,1}$ through r_{V,n_V} in figure 2.2 indicate the observed gene expression data. The expression profile associated with each observation indicates how each network is affecting the output variable. These expression profiles are not shown in the diagram. The variables α_1 through α_V indicate the weight vectors associated with each of the genes being observed. In [25], the authors constrained all the components of each α_i vector to be non-negative and their sum to one. The logical choice was to make all the α_i's to be sampled from an underlying Dirichlet distribution with a parameter vector K. The larger the values of the elements of K, the closer all the α_i's are to each other. Learning this unknown parameter K from the collected data would indicate the proportion wise breakup of the tissue.

The model parameters were learned in [25] using the Metropolis-Hastings algorithm, which is a Markov Chain Monte Carlo (MCMC) based method. The problem with such a method is that it is very computationally intensive. There are problems of convergence, especially since it may be difficult to judge if the Markov Chain has reached stationarity. In addition, the mixing may be poor which will require us to use thinning to get a decent effective sample size, which further increases computation time.

To get around the use of MCMC methods, in this paper we have resorted to the use of variational methods to estimate the posterior distributions of the unknown parameters. These methods involve assuming the distribution to have a certain factorized form and iteratively refining these factors. The variational method can be conveniently applied to the conjugate exponential family of models [38]. Hence we approximate the model for heterogeneous cancer tissue presented in [25] in the form of such a conjugate exponential model and derive the corresponding iterative update

equations.

4.2.1 A description of the conjugate exponential model for cancer tissue heterogeneity

Variational methods in Bayesian inference proceed by assuming a certain factorized form of the joint posterior distribution of the unobserved variables [3]. This factorization is done by first partitioning the unobserved variables into disjoint groups. For example let us say that we have a set of unobserved variables Z and we want to find an approximation of $P\left(Z/D\right)$ which is the posterior distribution of the unobserved elements conditional on the observed data D. We approximate this by $Q\left(Z\right)$, where

$$Q\left(Z\right) = \prod_{i=1}^{M} Q_i\left(Z_i\right) \tag{4.2}$$

Z_1 through Z_M are disjointed partitions of the set Z [3]. Then the method proceeds to minimize the KullbackLeibler (KL) divergence $KL\left(Q(Z)||P(Z/D)\right)$. It should be noted that

$$\ln P(D) = KL\left(Q(Z)||P(Z/D)\right) + L_Q(D) \tag{4.3}$$

where

$$L_Q(D) = \int \ln\left(\frac{P(Z,D)}{Q(Z)}\right) Q(Z)dZ \tag{4.4}$$

A derivation of equation 4.3 can be found in [3].

As the KL divergence is minimized, the lower bound $L_Q(D)$ increases monotonically. This can be used to check if the minimization algorithm has achieved convergence. Also the maximum achieved lower bound can be used for model selection.

The KL divergence is minimized using the following update equation

$$\ln Q_j\left(Z_j\right) = \int \ln P(Z, D) \prod_{i \neq j} Q_i(Z_i) dZ_i + constant \tag{4.5}$$

for each j, from 1 through M. The *constant* term can be adjusted to make sure that $Q_j\left(Z_j\right)$ is a proper probability distribution, that is it integrates to 1. For a detailed derivation of the equation 4.5, the reader is referred to [3]. Equation 4.5 shows that the optimum $Q_j(Z_j)$ depends on the other factors $Q_i(Z_i)$ for $i \neq j$. Hence the equations are solved iteratively by first initializing the parameters which describe each distribution $Q_j(Z_j)$ to appropriate values and cycling through the equations and replacing the old values with the corresponding updates. The variational method can be applied in a straightforward manner to the class of conjugate exponential models. Conjugate exponential models are those where the conditional distributions involved in the model belong to the exponential family and are conjugate with respect to the parent variables [38]. Therefore we modeled heterogeneous cancerous tissue in the form of a conjugate exponential model which would allow us to use the variational framework to estimate the proportion wise breakup of the tissue under study.

As discussed previously, each collected gene expression reading has an "expression profile" associated with it which depends on the Boolean Networks included in the ensemble and the stimulus provided to the tissue in the form of kinase inhibitory drug combinations. Each observed variable or gene i has a weight vector α_i associated with it. The weight vectors, which determine the extent to which each chosen Boolean network in the ensemble affects the i^{th} observable gene, are different for all the genes whose expressions are measured. Hence here too we will use a Hierarchical model (which, in addition will also belong to the conjugate exponential family).

The Bayesian network of the model is shown in figure 4.1. Since we are concerned

with the ratio with which the different networks affect the observed gene expressions, we can reduce redundancy by constraining the elements of each weight vector α_i to sum to 1, as was done in [25]. That is

$$\sum_{q=1}^{N} \alpha_{i,q} = 1 \tag{4.6}$$

where N is the number of networks in the ensemble. Therefore we have $\alpha_{i,N} = 1 - \sum_{i=1}^{N-1} \alpha_{i,q}$. For convenience, we are not constraining the elements of the weight vectors to be non negative. As we will see from simulations and from applications to real data, the posterior distributions of these elements will have very little probabilities in the regions where any element is negative. The probability distribution of any gene expression reading is defined to be normally distributed with a mean of $d_{i,j}^T \alpha_i$ and a precision of ρ (inverse of variance). We are considering all the measured gene expression ratios to have the same precision, although it is possible to have a hierarchical structure for the precision too. From equation 4.6, we have $d_{i,j}^T \alpha_i = \sum_{q=1}^{N-1} \alpha_{i,q}(d_{i,j,q} - d_{i,j,N}) + d_{i,j,N}$, where $d_{i,j,q}$'s are the elements of the expression profile $d_{i,j}$. Define $\mu_{i,j} = d_{i,j,N}$, $D_{i,j} = (d_{i,j,1} - d_{i,j,N}, d_{i,j,2} - d_{i,j,N}, ..., d_{i,j,N-1} - d_{i,j,N})^T$, and $\beta_i = (\alpha_{i,1}, \alpha_{i,2}, ..., \alpha_{i,N-1})^T$. Then we have

$$P\left(r_{i,j}/\beta_i, \rho, d_{i,j}\right) = \mathcal{N}\left(r_{i,j} | D_{i,j}^T \beta_i + \mu_{i,j}, \rho^{-1}\right) \tag{4.7}$$

The probability distribution of β_i is defined to be normally distributed with a mean of K and a precision matrix of Λ, where $K = (\, K_1 \; K_2 \; ... \; K_{N-1} \,)$ and Λ is an $(N-1)\times(N-1)$ positive definite matrix. Hence we have

$$P(\beta_i/K, \Lambda) = \mathcal{N}(\beta_i | K, \Lambda^{-1}) \tag{4.8}$$

Hence the unknown parameters are ρ, K, and Λ, which in the Bayesian framework are simply unobserved variables (along with all the β_i's). K could be interpreted as the weights associated with the first $N-1$ networks. $1-\sum_{q=1}^{N-1} K_q$ could be interpreted as the weight associated with the N^{th} network.

The Bayesian approach needs us to define certain priors over the unknown parameters. Thus we define the prior over ρ to be a gamma distribution with a shape and inverse scale parameter to be a_o and b_o respectively. The prior over K and Λ was taken as the Normal-Wishart distribution. Thus we have

$$P(\rho) = Gamma(\rho|a_o, b_o) \tag{4.9}$$

$$P(K/\Lambda) = \mathcal{N}\left(K|K_o, (q_o\Lambda)^{-1}\right) \tag{4.10}$$

and

$$P(\Lambda) = Wish\left(\Lambda|n_o, \Lambda_o^{-1}\right) \tag{4.11}$$

The joint posterior distribution of the unknown variables is

$$P(\rho, \beta, K, \Lambda/r) \propto P(\Lambda)P(K/\Lambda)P(\rho)\times \prod_{i=1}^{V}\left(\left[\prod_{j=1}^{n_i} P(r_{i,j}/\beta_i, \rho, d_{i,j})\right] P(\beta_i/K, \Lambda)\right) \tag{4.12}$$

where β is the set of all the β_i's and r is the set of all the observed data. Our model belongs to the conjugate exponential family. In the following section, we will approximate the joint posterior distribution of the unknown variables using the variational approach. This would in turn simplify the derivation of the marginal distributions of the variables of interest (such as K which would indicate the proportion wise breakup of the tissue).

4.2.2 Derivation of the variational update equations

The approximation $Q(\rho, \beta, K, \Lambda)$ of the posterior $P(\rho, \beta, K, \Lambda/r)$ is assumed to factorize in the following form.

$$Q(\rho, \beta, K, \Lambda) = Q_\rho(\rho) Q_\beta(\beta) Q_{K,\Lambda}(K, \Lambda) \tag{4.13}$$

We then use equation 4.5 to derive the update equations for each of the factors. First we apply equation 4.5 to $Q_\rho(\rho)$. As per equation 4.5, we have

$$\ln Q_\rho(\rho) = \mathrm{E}_{\neq\rho}\left[\ln P(\rho, \beta, K, \Lambda/r)\right] + constants \tag{4.14}$$

where

$$\mathrm{E}_{\neq\rho}\left[\ln P(\rho, \beta, K, \Lambda/r)\right] = \int \left[\ln P(\rho, \beta, K, \Lambda/r)\right] Q_\beta(\beta) Q_{K,\Lambda}(K, \Lambda) d\beta dK d\Lambda \tag{4.15}$$

The terms which are not dependent on ρ can be absorbed into the *constants*. Thus we get

$$\ln Q_\rho(\rho) = \mathrm{E}_{\neq\rho}\left[\ln P(\rho) + \sum_{i=1}^{V}\sum_{j=1}^{n_i} \ln P(r_{i,j}/\beta_i, \rho, d_{i,j})\right] + constants \tag{4.16}$$

Upon simplifying we get

$$\ln Q_\rho(\rho) = a_\rho \ln(\rho) - b_\rho \rho + constants \tag{4.17}$$

where

$$a_\rho = a_o + \frac{1}{2}\sum_{i=1}^{V} n_i \tag{4.18}$$

and

$$b_\rho = b_o + \frac{1}{2}\sum_{i=1}^{V}\sum_{j=1}^{n_i}\{(r_{i,j} - \mu_{i,j})^2 - 2(r_{i,j} - \mu_{i,j})D_{i,j}^T \mathrm{E}[\beta_i] + D_{i,j}^T(\mathrm{E}[\beta_i\beta_i^T])D_{i,j}\} \tag{4.19}$$

and *constants* are all those terms which do not depend on ρ. Looking at the form of equation 4.19 we can deduce $Q_\rho(\rho)$ to be gamma distributed with a_ρ as the shape parameter and b_ρ to be the inverse scale parameter. That is

$$Q_\rho(\rho) = Gamma\,(\rho|a_\rho, b_\rho) \tag{4.20}$$

$\mathrm{E}[\beta_i]$ and $\mathrm{E}[\beta_i\beta_i^T]$ depend on $Q_\beta(\beta)$.

Using similar steps, we get the result $Q_\beta(\beta) = \prod_{i=1}^{V} Q_{\beta_i}(\beta_i)$. This factorization is not implicitly assumed, but comes as a result of applying equation 4.5 to derive the update equations for $Q_\beta(\beta)$. Upon inspection, $Q_{\beta_i}(\beta_i)$ comes out to be normally distributed as follows

$$Q_{\beta_i}(\beta_i) = \mathcal{N}(\beta_i|\mu_{\beta_i}, \Lambda_{\beta_i}^{-1}) \tag{4.21}$$

where Λ_{β_i} is a $(N-1)\times(N-1)$ positive semidefinite precision matrix and μ_{β_i} is the mean vector of length $N-1$ which are defined as

$$\Lambda_{\beta_i} = \mathrm{E}[\Lambda] + \mathrm{E}[\rho]\sum_{j=1}^{n_i} D_{i,j}D_{i,j}^T \tag{4.22}$$

and

$$\mu_{\beta_i} = \Lambda_{\beta_i}^{-1}\{\mathrm{E}[\Lambda K] + \mathrm{E}[\rho]\sum_{j=1}^{n_i} D_{i,j}(r_{i,j} - \mu_{i,j})\} \tag{4.23}$$

All the expectations in equations 4.22 and 4.23 are done with respect to $Q_\rho(\rho)Q_{K,\Lambda}(K,\Lambda)$. Following similar steps, $Q_{K,\Lambda}(K,\Lambda)$ comes out to be factorizable as $Q_{K,\Lambda}(K/\Lambda)Q_{K,\Lambda}(\Lambda)$ which are defined as follows:

$$Q_{K,\Lambda}(K/\Lambda) = \mathcal{N}\left(K|K_{oK}, [(q_o + V)\Lambda]^{-1}\right) \tag{4.24}$$

$$Q_{K,\Lambda}(\Lambda) = Wish\left(\Lambda|n_o + V, \Lambda_{o\Lambda}^{-1}\right) \tag{4.25}$$

where K_{oK} and $\Lambda_{o\Lambda}^{-1}$ are defined as:

$$K_{oK} = \frac{\sum_{i=1}^{V}\mathrm{E}[\beta_i] + q_o K_o}{V + q_o} \tag{4.26}$$

$$\Lambda_{o\Lambda}^{-1} = \Lambda_o^{-1} + \sum_{i=1}^{V}\mathrm{E}[\beta_i\beta_i^T] + q_o K_o K_o^T - (q_o + V)K_{oK}K_{oK}^T \tag{4.27}$$

K_{oK} and $\Lambda_{o\Lambda}$ are of length $N-1$ and of dimension $(N-1)\times(N-1)$ respectively.

Now that the optimal form of each factor in the approximation is known, the expectations can be easily computed. Thus we get:

$$\mathrm{E}[\beta_i] = \mu_{\beta_i} \tag{4.28}$$

$$\mathrm{E}[\beta_i\beta_i^T] = \mu_{\beta_i}\mu_{\beta_i}^T + \Lambda_{\beta_i}^{-1} \tag{4.29}$$

$$\mathrm{E}[\Lambda] = (n_o + V)\Lambda_{o\Lambda} \tag{4.30}$$

$$\mathrm{E}[\rho] = \frac{a_\rho}{b_\rho} \tag{4.31}$$

$$\mathrm{E}[K] = K_{oK} \tag{4.32}$$

$$\mathrm{E}[\Lambda K] = (n_o + V)\Lambda_{o\Lambda} K_{oK} \tag{4.33}$$

The constants a_ρ, b_ρ, Λ_{β_i}, μ_{β_i}, K_{oK}, $\Lambda_{o\Lambda}$ are all initialized to appropriate values and then iteratively updated by cycling through the update equations 4.18, 4.19, 4.22, 4.23, 4.26, and 4.27 using the values of the expectations shown in equations 4.28 through 4.33. Equation 4.4 is used to calculate the lower bound at each iteration. In the interest of space, the exact equation of the lower bound is not shown here. However in our simulations and applications to real world data, we will show how convergence is judged using the lower bound.

4.2.3 Simulation experiments

To demonstrate the algorithm, we ran simulations on synthetic data. First, the synthetic data was generated from the following example. Three different Boolean networks with stuck-at faults were taken in the ensemble. The first network was chosen to have a stuck-at one fault at Ras. The second network was chosen to have a stuck-at zero fault at PTEN. The third network was chosen to have a stuck-at one fault at RAF. Hence $N = 3$. The three locations are shown as shaded squares in a single Boolean network in figure 2.1. The activity of the four transcription factors shown at the bottom of figure 2.1 would be different in the three networks for any given drug combination. A total of 63 different drug combinations were chosen as the stimulus. The location of the targets of these kinase inhibitory drugs is shown in figure 2.1. Since there are six different drugs, there would be 63 different possible combinations excluding the case of no drug exposure. It was assumed that five genes per transcription factor were measured, hence a total of twenty different observable variables were assumed in the simulation. Hence $V = 20$. Each experiment was

repeated ten times in the simulation. This would result in each observable being observed 10 times. Hence $n_i = 10$ for all i's ranging from 1 through 20. Since the number of networks is three, hence the length of the vector K is two. K was set to be $\begin{pmatrix} 0.1 & 0.3 \end{pmatrix}$. Hence the first network has a weight of 0.1 associated with it, the second network has a weight of 0.3 associated with it, and the third network has a weight of 0.6 associated with it. ρ was set to be 100 and Λ was set to be

$$\Lambda = \begin{bmatrix} 0.01 & 0.005 \\ 0.005 & 0.008 \end{bmatrix}^{-1}$$

.

For the purposes of demonstration, the parameters for the prior distributions of the parameters were chosen as follows. a_o and b_o for the prior over ρ were both chosen to be 0.5. This would make the prior over ρ to have a mean of 1. K_o was chosen to be $\begin{pmatrix} 1/3 & 1/3 \end{pmatrix}^T$ and q_o was chosen to be 0.001. Hence the prior belief assigns equal weights to all the three networks. The small value of q_o means that the prior is spread out and non informative. As for the prior over Λ, n_o was chosen to be 1.1 and Λ_o was chosen as

$$\Lambda_o = \begin{bmatrix} 0.01 & 0.005 \\ 0.005 & 0.008 \end{bmatrix}^{-1}$$

.

For comparison purposes, we also did the posterior inference using Gibbs sampling and found point estimates of the unknown parameters using Expectation Maximization algorithm. The full conditionals of the unobserved variables in the Gibbs sampling algorithm are listed below. (Some of the notations used here are similar to those used in the derivation of the variational update equations. The reader is advised to keep in mind that the full conditionals are derived independent of the

derivations in the previous subsection).

The full conditional of Λ is:

$$P\left(\Lambda/...\right) = Wish\left(\Lambda|n_{\Lambda}, \Lambda_{\Lambda}^{-1}\right) \tag{4.34}$$

where

$$n_{\Lambda} = n_o + V + 1 \tag{4.35}$$

$$\Lambda_{\Lambda}^{-1} = \Lambda_o^{-1} + q_o(K - K_o)(K - K_o)^T + \sum_{i=1}^{V}(\beta_i - K)(\beta_i - K)^T \tag{4.36}$$

The full conditional of K is:

$$P\left(K/...\right) = N\left(K|\frac{q_o K_o + \sum_{i=1}^{V}\beta_i}{q_o + V}, (q_o + V)\Lambda\right) \tag{4.37}$$

The full conditional of ρ is:

$$P\left(\rho/...\right) = Gamma\left(\rho|a_{\rho}, b_{\rho}\right) \tag{4.38}$$

where

$$a_{\rho} = a_o + \frac{1}{2}\sum_{i=1}^{V} n_i \tag{4.39}$$

and

$$b_{\rho} = b_o + \frac{1}{2}\sum_{i=1}^{V}\sum_{j=1}^{n_i}\{(r_{i,j} - \mu_{i,j} - D_{i,j}^T\beta_i)^2$$
$$(r_{i,j} - \mu_{i,j})^2 - 2(r_{i,j} - \mu_{i,j})D_{i,j}^T\beta_i + (D_{i,j}^T\beta_i)^2\} \tag{4.40}$$

The full conditionals of each of the β_i's are:

$$P\left(\beta_i/...\right) = \mathcal{N}\left(\beta_i|\mu_{\beta_i}, \Lambda_{\beta_i}^{-1}\right) \tag{4.41}$$

where

$$\Lambda_{\beta_i} = \Lambda + \rho \sum_{j=1}^{n_i} D_{i,j} D_{i,j}^T \tag{4.42}$$

and

$$\mu_{\beta_i} = \Lambda_{\beta_i}^{-1} \left(\Lambda K + \rho \sum_{j=1}^{n_i} (r_{i,j} - \mu_{i,j}) D_{i,j} \right) \tag{4.43}$$

The expectation maximization algorithm can also be used to find a maximum likelihood estimate of the unknown parameters ρ, K, and Λ. The hidden variables which are not observed are simply all the β_i's for i ranging from 1 through V. The derivation is skipped in the interest of space. The Expectation Maximization Update equations are as follows. Define $\rho^{(n)}$, $K^{(n)}$, and $\Lambda^{(n)}$ to be the estimates of the parameters in the n^{th} iteration. Define $\Sigma_i^{(n)}$, $M_i^{(n)}$ and $S_i^{(n)}$ to be

$$\Sigma_i^{(n)} = \left[\Lambda^{(n)} + \rho^{(n)} \sum_{j=1}^{n_i} D_{i,j} D_{i,j}^T \right]^{-1} \tag{4.44}$$

$$M_i^{(n)} = \Sigma_i^{(n)} \left[\Lambda^{(n)} K^{(n)} + \rho^{(n)} \sum_{j=1}^{n_i} D_{i,j} (r_{i,j} - \mu_{i,j}) \right] \tag{4.45}$$

$$\begin{aligned} S_i^{(n)} = \sum_{j=1}^{n_i} \{ & (r_{i,j} - \mu_{i,j})^2 - 2(r_{i,j} - \mu_{i,j}) D_{i,j}^T M_i^{(n)} \\ & + D_{i,j}^T (M_i^{(n)} M_i^{(n)T} + \Sigma_i^{(n)}) D_{i,j} \} \end{aligned} \tag{4.46}$$

Then the update equations are:

$$\rho^{(n+1)} = \frac{\sum_{i=1}^{V} n_i}{\sum_{i=1}^{V} S_i^{(n)}} \tag{4.47}$$

$$K^{(n+1)} = \frac{\sum_{i=1}^{V} M_i^{(n)}}{V} \tag{4.48}$$

$$\left\{\Lambda^{(n+1)}\right\}^{-1} = \frac{1}{V}\sum_{i=1}^{V}\left(M_i^{(n)} M_i^{(n)T} + \Sigma_i^{(n)}\right) - K^{(n+1)} K^{(n+1)T} \tag{4.49}$$

where $\rho^{(n+1)}$, $K^{(n+1)}$, and $\Lambda^{(n+1)}$ are the updated values of the parameters.

Figure 4.2 shows the posterior marginal distributions of the elements of K derived using both the Gibbs sampling method as well as the variational Bayesian method. The third graph in figure 4.2 is simply the marginal posterior density of $1 - K_1 - K_2$. As we can see, the distributions computed using both the methods are almost identical. Same is true for the posterior distribution of ρ which is shown in figure 4.3. Figure 4.4 shows how the lower bound stops improving after 80 iterations of the variational algorithm thereby indicating convergence. The mean of the posterior distribution of K comes out to be $(0.1044, 0.3015)^T$ and $(0.1042, 0.3011)^T$ from the variational method and the Gibbs sampling method respectively. The maximum likelihood estimate of K using the expectation maximization algorithm comes out to be $(0.1042, 0.3015)^T$. Figure 4.5 shows that the log likelihood function shows no significant improvement after 100 iterations of the expectation maximization algorithm, thereby indicating convergence. All three estimates are close to the actual value of $(0.1, 0.3)^T$ thereby showing the correctness of the algorithms.

From figure 4.2, we can see that the marginal posterior distributions of the elements of K lie mostly within the interval $(0, 1)$. The marginal posterior distribution of $1 - \sum_{q=1}^{N} K_q$ (the third graph in figure 4.2) also lies within the interval $(0, 1)$.

4.2.4 Verification using experimental data

In order to test the model, we need to collect data from a tissue where the dominant population or the dominant network is already known. In a cancerous cell line, one cannot be sure which network is dominant. But in a normal cell line, such as adult fibroblasts, it is fair to assume that a network modeling a faultless MAPK signal transduction network would be the most dominant one, no matter what other networks are included in the ensemble. Hence we performed a simple experiment on adult fibroblasts to demonstrate the approach. For a detailed description of the wet lab procedures, the authors are referred to [25]. The experiments were performed on three groups of cell cultures. The first group was not exposed to any kinase inhibitory drugs and served as the control. The second and third groups were exposed to the drugs LY294002 and a combination of LY294002 and U0126 respectively. Their target locations are shown in figure 2.1. GAPDH (Glyceraldehyde-3-Phosphate Dehydrogenase) was used as the reference gene. Genes having the SP1 or the SRF-ELK response elements in their promoters were quantified through real time PCR and the delta-delta method [22]. A total of ten different genes (including alternative transcripts) were quantified [25]. Hence $V = 10$. Their measured expression values are shown in table 4.1.

For the sake of demonstration we assumed 3 networks to be in the ensemble. Hence $N = 3$. Network 1 has no mutations, i.e. no stuck-at faults. This network models the normally behaving fibroblasts. Network 2 is assumed to have a stuck-at one fault at ERK1/2 and network 3 is assumed to have stuck-at one faults at SRF-ELK1 and SRF-ELK4. The "expression profiles" for all the genes for the experimental conditions of the second and third groups are known (can be easily derived from figure 2.1) and depend on the behavior of the 3 networks included in

the ensemble. These are shown in table 4.1.

The marginal posterior distributions of the elements of K are computed using the variational approach and are shown in figure 4.6. The lower bound stops improving after 300 iterations as can be seen in figure 4.8, thereby indicating convergence. Only the results of the variational computations are shown since the Markov Chain Monte Carlo approach could not produce decent effective sample sizes. As we can see, most of the probability mass lies in the valid region. Specifically, most of the posterior marginal probabilities associated with the elements of K are within 0 and 1. Hence we get meaningful interpretations of the inferred value of K. The mean of the posterior distribution of K comes out to be $(\,0.6716\ 0.2740\,)^T$. As expected, the first faultless network representing normal fibroblasts has the maximum influence on the behavior of the observables, close to 67%. The other two networks have influences of 27% and 6% respectively. [25] reports values of $(\,0.6453\ 0.2255\ 0.1292\,)^T$ which are very close to those calculated in this paper. The Expectation maximization algorithm also gives very close values of $(\,0.6764\ 0.2745\ 0.0490\,)^T$. The log likelihood stops improving after 250 iterations of the Expectation Maximization algorithm as can be seen in figure 4.9, thereby indicating convergence. This simple experiment shows how this model can be used to determine the proportional breakup of the subpopulations in a heterogeneous tissue.

4.3 Summary and comments on possible future work

Here the problem of heterogeneity in cancer tissue cell populations was addressed and a model was developed which uses a collection of different Boolean networks to model the various sub populations in the tissue. It was demonstrated using both synthetic and real world data collected from fibroblasts, how this model can be used to find out the relative abundance of the various subpopulations in a given tissue under

study using QPCR gene expression data. This work is an extension of the previous work in [25]. The novelty of this work is in the improvement in the computation time. A hierarchical conjugate exponential model was used in this paper, which allowed the use of variational methods for Bayesian estimation of the relative abundances of the various subpopulations. The efficacy of the variational methods was verified by comparing the results obtained to those obtained using MCMC (Gibbs sampling) and Maximum likelihood (Expectation Maximization) methods. Determining the relative abundance of the various subpopulations in an individual patient could be used to come up with customized combination therapies which are tailored to the patient so as to improve the efficacy and reduce side effects (for example, we may want to target the dominant subpopulation(s) using the minimal amount of drugs so as to reduce side effects).

Variational methods are becoming increasingly important as Bayesian methods are gaining interest since these methods allow for speedy computation of posterior distributions of variables of interest. Moreover the lower bound, which is easily computed in variational methods, provides for an effective proxy for the likelihood of the data which can be used for model selection. Hence this approach can also be extended to solving the problem of determining how many Boolean networks to include in the ensemble as well as determining which Boolean networks to include in the ensemble. Besides variational methods, other methods, such as expectation propagation may also be used to solve the problem of determining the dominant subpopulations in a heterogeneous cancer tissue in a Bayesian framework with reduced computational requirements.

The model in this paper was developed keeping QPCR gene expression data in mind. However similar methods can be developed which use data from more state of the art technologies such as Next Generation Sequencing and flow cytometry.

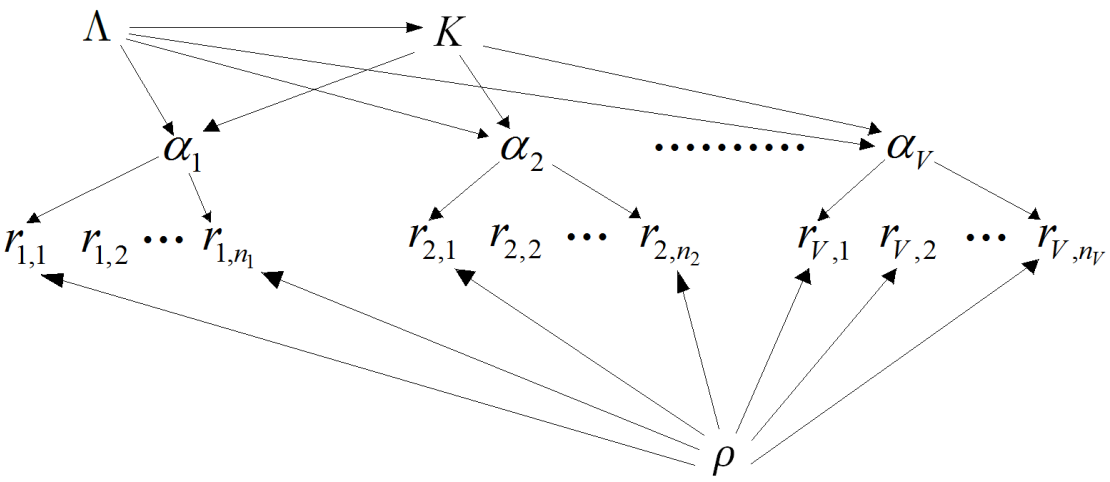

Figure 4.1: A Bayesian network representing the conditional dependencies in the conjugate exponential model.

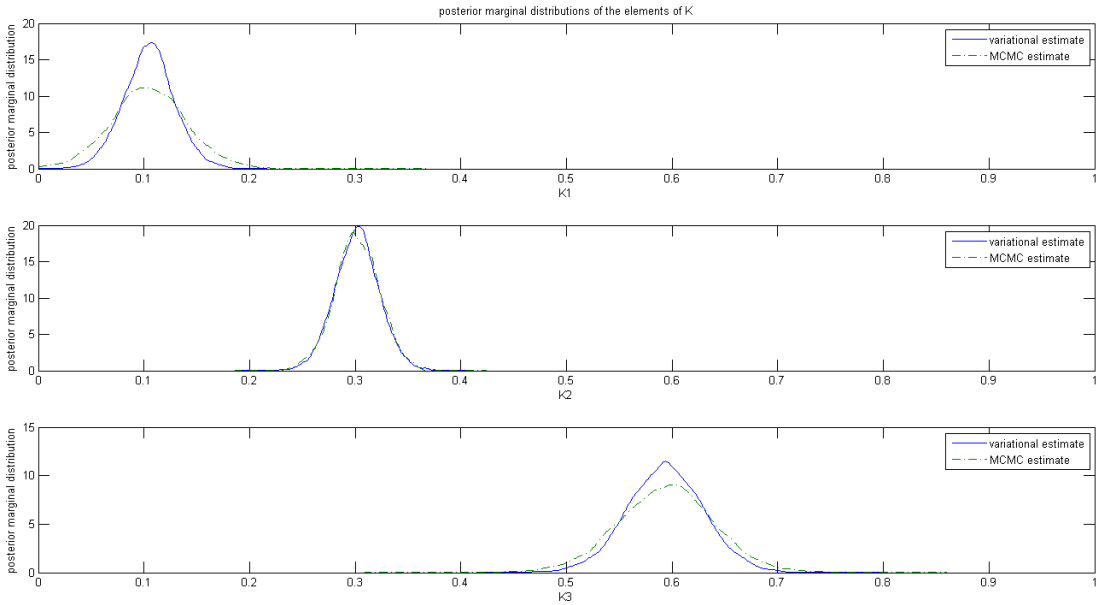

Figure 4.2: Posterior marginal distributions of the elements of K for synthetic data.

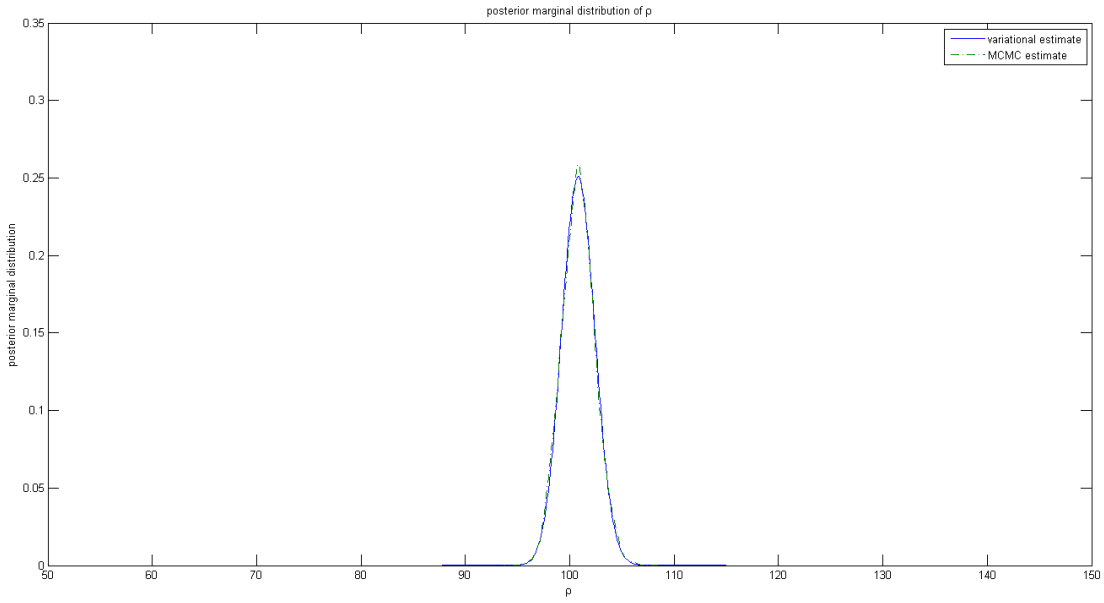

Figure 4.3: Posterior marginal distribution of ρ for synthetic data.

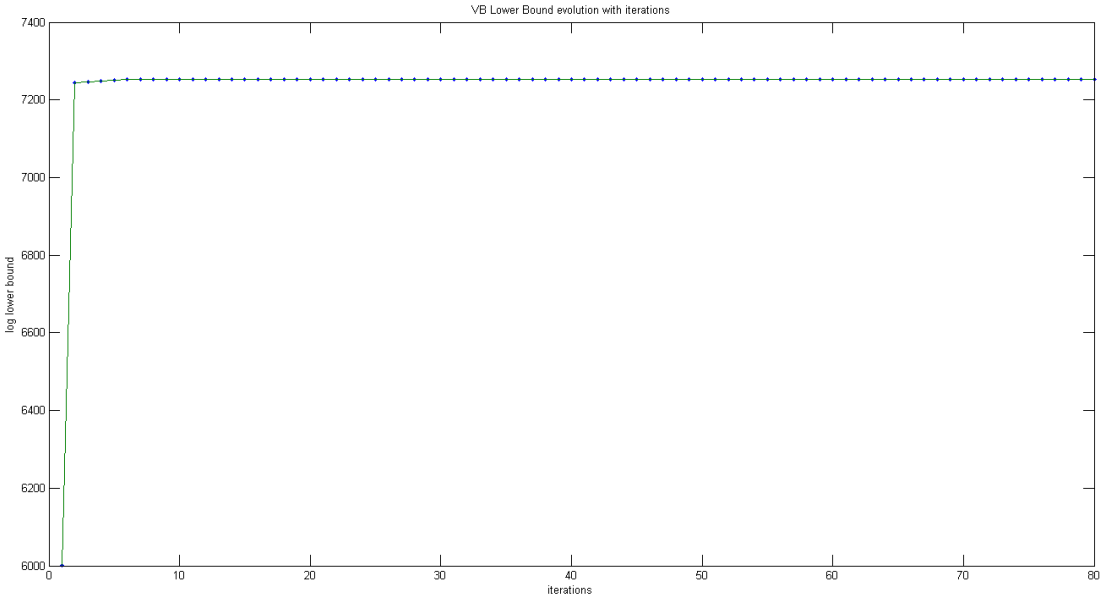

Figure 4.4: Increase of the log of the lower bound with iterations of the variational Bayes algorithm for synthetic data.

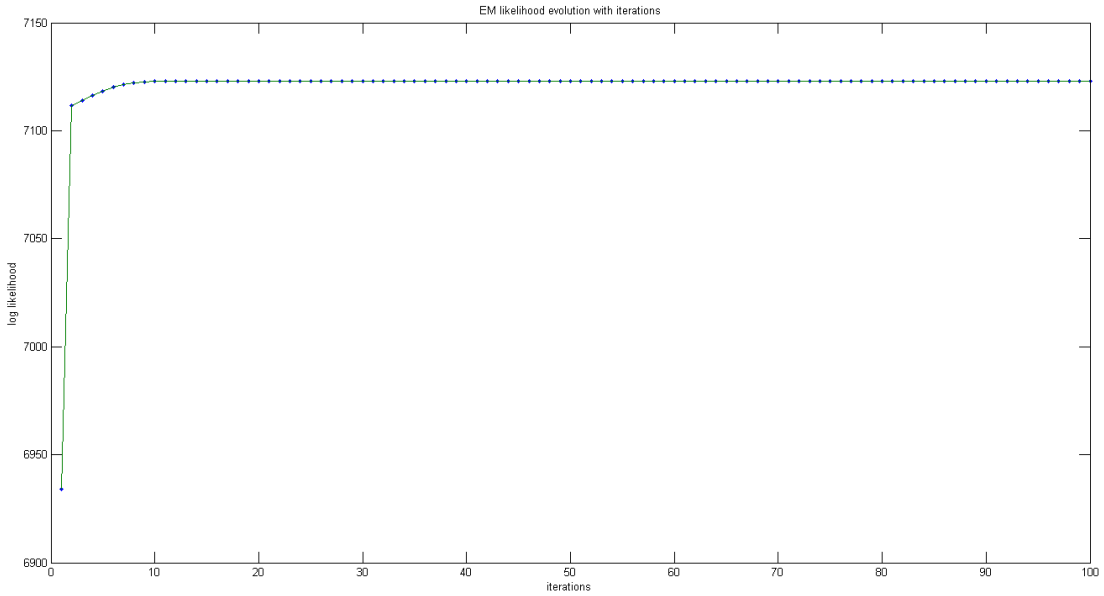

Figure 4.5: Increase of data log likelihood with the iterations of the expectation maximization algorithm for synthetic data.

Table 4.1: Table showing the gene expression measurements, their "expression profiles", and their reference sequence (RefSeq) numbers.

gene	*RefSeq*	group 1		group 2	
		expression profile	*gene expression*	*expression profile*	*gene expression*
EGR1	NM_001964.2	1 1 1	0.5987 0.7320 0.5586 0.6199	0 1 1	0.4796 0.2892 0.2535 0.2698
JUN	NM_002228.3	1 1 1	0.4931 0.6736 0.6598 0.7792	0 1 0	0.1550 0.2793 0.3015 0.3415
BIRC5	NM_001168.2	1 1 1	0.5799	0 1 0	0.3842
CMYC	NM_002467.4	1 1 1	0.3209 0.2852 0.3439 0.2994	0 1 0	0.2570 0.2059 0.2717 0.2679
Decorin	NM_133504.2 NM_133505.2 NM_133507.2 NM_133503.2	1 1 1	0.0819 0.0728 0.3345 0.4353 0.4601 0.4147 0.4323	0 1 0	0.0087 0.0242 0.1661 0.2793 0.3789 0.3737 0.3536
IRF3	NM_001571.5	1 1 1	0.5176 0.4204 0.3560 0.3345	0 1 0	0.2624 0.2553 0.2253 0.2031
VEGFA	NM_003376.5	1 1 1	0.4444 0.4989 0.5176 0.5105	0 1 0	0.3164 0.4623 0.4633 0.3660

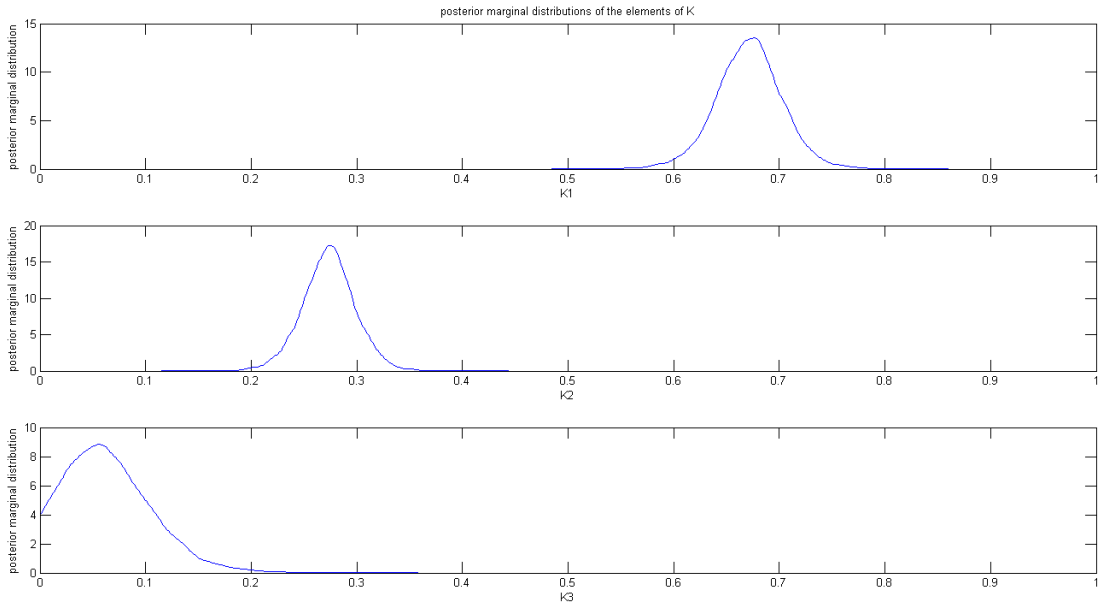

Figure 4.6: Posterior marginal distributions of the elements of K for data collected from fibroblasts.

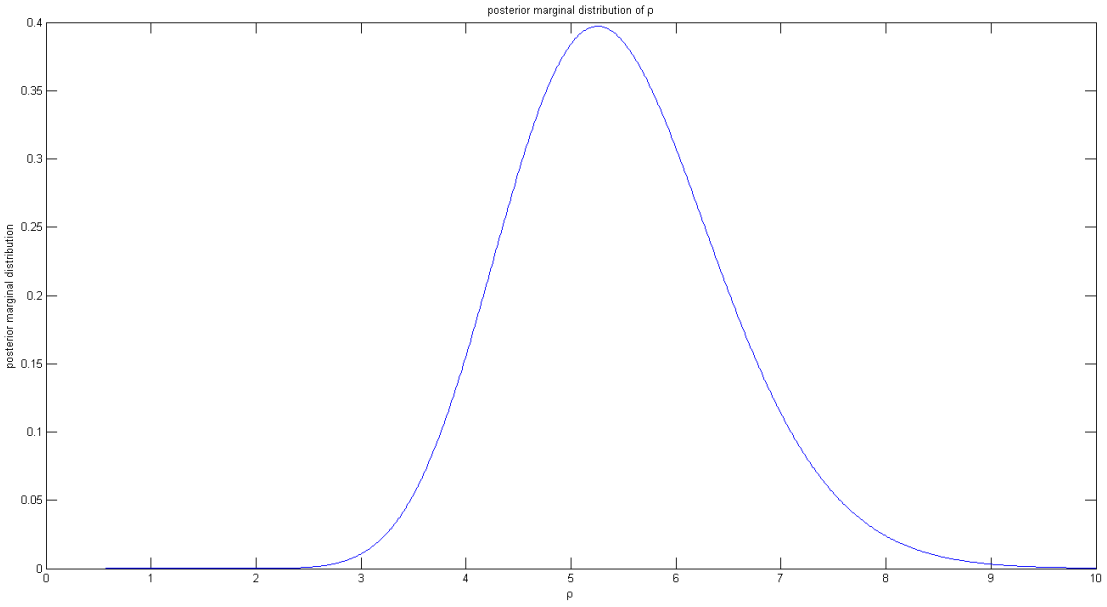

Figure 4.7: Posterior marginal distribution of ρ for data collected from fibroblasts.

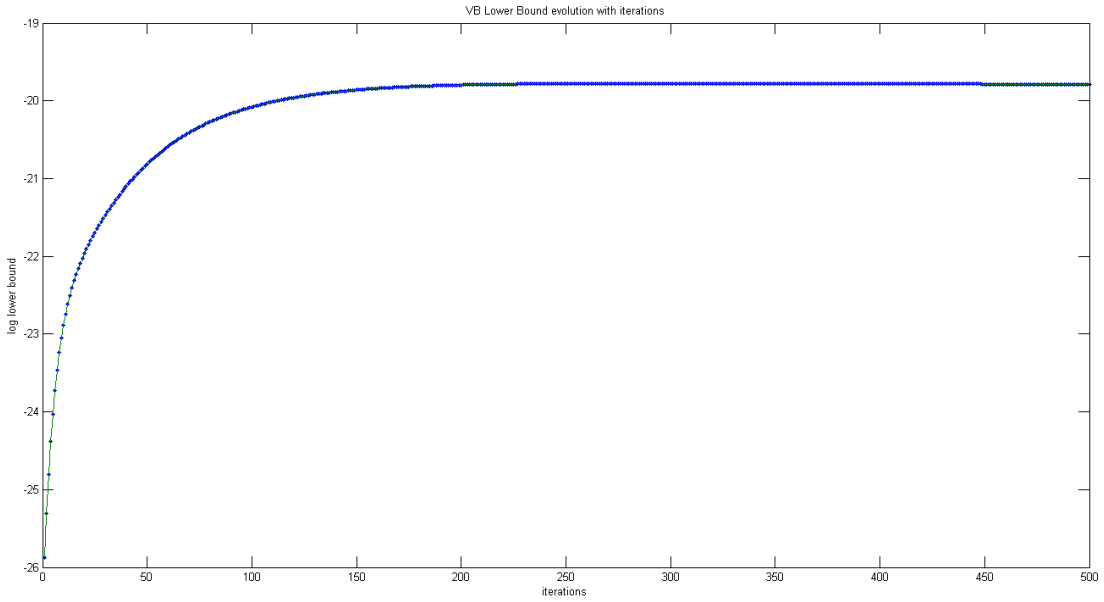

Figure 4.8: Increase of the log of the lower bound with iterations of the variational Bayes algorithm for data collected from fibroblasts.

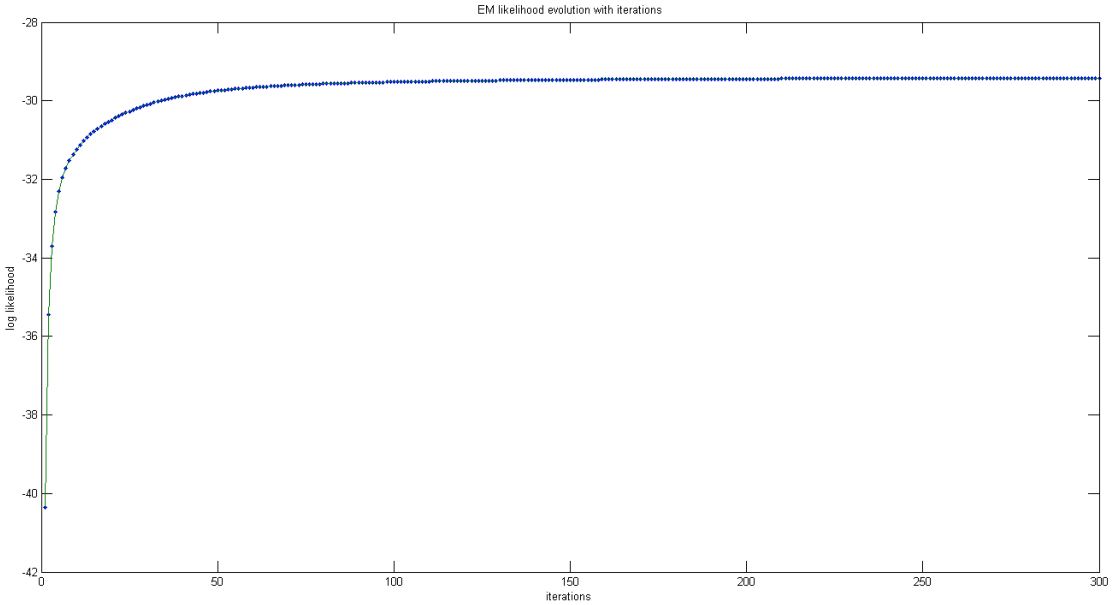

Figure 4.9: Increase of data log likelihood with the iterations of the expectation maximization algorithm for data collected from fibroblasts.

5. CONCLUSIONS *

In this dissertation, we have presented methods to model cancer tissues primarily by using Bayesian methods. These methods range from accurate modeling and inference using MCMC methods to computationally efficient methods such as belief propagation and variational Bayesian methods. The thesis was divided into three sections, each focusing on a certain sub-problem. A summary of these three sections is provided in the following paragraphs.

In section 2, we addressed the important problem of heterogeneity in cancer tissues and presented a model which has the ability to use prior pathway knowledge and knowledge about likely mutations in cancers to represent a heterogeneous cancer tissue as an ensemble of faulty Boolean networks. We demonstrated the general idea of our approach by considering the observed variables to be genes transcribed by key transcription factors. We also demonstrated how the Metropolis-Hastings MCMC method can be used to estimate the relative effect that each subpopulation exerts on the observed variables. This estimate gives us an idea about which subpopulation is the most dominant one among all the subpopulations in the ensemble. Such estimates, if obtained using data from individual patients, could help customize combination therapy design and could help improve the success rate of such cancer therapies. for more information on this work, the reader is referred to [25].

*Parts of this section are reprinted with permission from "A Model for Cancer Tissue Heterogeneity" by A. K. Mohanty, A. Datta, and V. Venkatraj, 2013. *IEEE Transactions on Biomedical Engineering*, volume 61, no. 3, pages 966 - 974, © 2013 IEEE. doi:10.1109/TBME.2013.2294469, and "Using the message passing algorithm on discrete data to detect faults in boolean regulatory networks" by A. K. Mohanty, A. Datta, and V. Venkatraj, 2014. *BMC Algorithms for Molecular Biology*, volume 9, no. 20, 12 pages. doi:10.1186/s13015-014-0020-6, and "A Conjugate Exponential Model for Cancer Tissue Heterogeneity" by A. K. Mohanty, A. Datta, and V. Venkatraj, 2015. *IEEE Journal of Biomedical and Health Informatics*, preprint, © 2015 IEEE. doi:10.1109/JBHI.2015.2410279.

In section 2, the methods suggested depended heavily on MCMC techniques. However for these methods to become practically applicable, the computational complexity needs to be reduced. Various computationally efficient methods exist to speed up the computation of posterior distributions. In section 3, an algorithm based on loopy belief propagation or message passing has been presented to estimate the most likely locations of faults in a Boolean network based on observed data. We have compared the performance of the algorithm with Markov Chain Monte Carlo techniques (the Metropolis-Hastings Algorithm) through simulations, and we have shown that the message passing algorithm gives results comparable to those obtained using the MCMC methods with the added advantage of much smaller computation times. We also applied the model to analyze data collected from fibroblasts, thereby demonstrating how this model can be used on real world data. Such a computationally manageable approach has the potential to allow the inference of locations of faults in a Boolean regulatory network in a probabilistic setting from data, such as gene expression data. For further information on this work, the reader is referred to [26].

In section 4, an approximation of model described in section 2 is presented so as to allow for the use of variational Bayesian methods in the estimation of conditional posterior distributions of the unobserved variables in the model. The novelty of this work is in the improvement in the computation time. A hierarchical conjugate exponential model was used in this section, which allowed the use of variational methods for Bayesian estimation of the relative abundances of the various subpopulations. The efficacy of the variational methods was verified by comparing the results obtained to those obtained using MCMC (Gibbs sampling) and Maximum likelihood (Expectation Maximization) methods. Variational methods are becoming increasingly important as Bayesian methods are gaining interest since these methods allow for speedy computation of posterior distributions of variables of interest. Moreover

the lower bound, which is easily computed in variational methods, provides for an effective proxy for the likelihood of the data which can be used for model selection. Hence this approach can also be extended to solving the problem of determining how many Boolean networks to include in the ensemble as well as determining which Boolean networks to include in the ensemble. Besides variational methods, other methods, such as expectation propagation may also be used to solve the problem of determining the dominant subpopulations in a heterogeneous cancer tissue in a Bayesian framework with reduced computational requirements. This could be a possible direction for future research.

REFERENCES

[1] F Barabe, J A Kennady, K J Hope, and J E Dick. Modeling the initiation and progression of human acute leukemia in mice. *Science*, 316(5824):600–604, 2007.

[2] M Bébien, S Salinas, C Becamel, V Richard, L Linares, and R A Hipskind. Immediate-early gene induction by the stresses anisomycin and arsenite in human osteosarcoma cells involves mapk cascade signaling to elk-1, creb and srf. *Oncogene*, 22(12):1836–1847, March 2003.

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, Singapore, 2006.

[4] J M Bower and H Bolouri. *Computational Modeling of Genetic and Biochemical Networks*. MIT Press, Boston, 1st edition, 2001.

[5] L L Campbell and K Polyak. Breast tumor heterogeneity:cancer stem cells or clonal evolution? *Cell Cycle*, 6(19):2332 – 2338, 2007.

[6] Y Chen, V Kamat, E R Dougherty, M L Bittner, P S Meltzer, and J M Trent. Ratio statistics of gene expression levels and applications to microarray data analysis. *Bioinformatics*, 18(9):1207–1215, 2002.

[7] Yidong Chen, E R Dougherty, and M L Bittner. Ratio-based decisions and the quantitative analysis of cdna microarray images. *J. Biomed. Opt.*, 2(4):364–374, 1997.

[8] R W Clarkson, C A Shang, L K Levitt, T Howard, and M J Waters. Ternary complex factors elk-1 and sap-1a mediate growth hormone induced transcrip-

tion of egr-1 (early growth response factor-1) in 3t3-f442a preadipocytes. *Mol. Endocrinol.*, 13(4):619–631, 1999.

[9] A Datta and E Dougherty. *Introduction to Genomic Signal Processing with control*. CRC Press, New York, 2007.

[10] P Dhawan, A Bell, A Kumar, C Golden, and K D Mehta. Critical role of p42/44(mapk) activation in anisomycin and hepatocyte growth factor-induced ldl receptor expression: activation of raf-1/mek-1/p42/44(mapk) cascade alone is sufficient to induce ldl receptor expression. *J. Lipid Res.*, 40(10):1911–1919, Oct 1999.

[11] N Friedman, M Linial, I Nachman, and D Pe'er. Using bayesian networks to analyze expression data. *J. Comput. Biol.*, 7(3-4):601–620, 2000.

[12] A Gelman, J B Carlin, H S Stern, and D B Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton, London, New York, Washington D.C., 2nd edition, 2004.

[13] A Gelman and J Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York, 2007.

[14] N Goldman and Z Yang. A codon-based model of nucleotide substitution for protein-coding dna sequences. *Mol. Biol. Evol.*, 11(5):725736, 1994.

[15] C Greenman, R Wooster, P A Futreal, M R Stratton, and D F Easton. Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics*, 173(4):2187–2198, 2006.

[16] P D Hoff. *A First Course in Bayesian Statistical Methods*. Springer Texts in Statistics, Dordrecht, Heidelberg, London, New York, 2009.

[17] F R Kschischang, B J Frey, and H A Loeliger. Factor graphs and the sum-product algorithm. *IEEE t. inform. theory*, 47(2):498–519, 2001.

[18] R K Layek, A Datta, M Bittner, and E R Dougherty. Cancer therapy design based on pathway logic. *Bioinformatics*, 27(4):548–555, 2011.

[19] R K Layek, A Datta, and E R Dougherty. From biological pathways to regulatory networks. *Mol. BioSyst.*, 7:843–851, 2011.

[20] D Levens. How the c-myc promoter works and why it sometimes does not. *J. Natl. Cancer I. Monographs*, 39:41–43, 2008.

[21] S Liang, S Fuhrman, and R Somogyi. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In *Pac. Symp. Biocomput.*, volume 3, pages 18–29, 1998.

[22] K J Livak and T D Schmittgen. Analysis of relative gene expression data using real-time quantitative pcr and the $2^{-\Delta\Delta C_t}$ method. *Methods*, 25(4):402–408, 2001.

[23] A A margolin, I Nemenman, K Basso, C Wiggins, G Stolovitzky, R D Favera, and A Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7, 2006.

[24] M V Mityaev, E P Kopantzev, A A Buzdin, T V Vinogradova, and E D Sverdlov. Functional significance of a putative sp1 transcription factor binding site in the survivin gene promoter. *Biochemistry (Moscow)*, 73(11):1183–1191, 2008.

[25] A K Mohanty, A Datta, and V Venkatraj. A model for cancer tissue heterogeneity. *IEEE t. Bio-Med. eng.*, 61(3):966 – 974, 2013.

[26] A K Mohanty, A Datta, and V Venkatraj. Using the message passing algorithm on discrete data to detect faults in boolean regulatory networks. *BMC algorithm mol. boil.*, 9(20), 2014.

[27] P C Nowel. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, 1976.

[28] N Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975.

[29] G Pagés and J Pouysségur. Transcriptional regulation of the vascular endothelial growth factor gene-a concert of activating factors. *Cardiovasc. Res.*, 65(3):564–573, 2005.

[30] T Reya, S J Morrison, M F Clarke, and I L Weissman. Stem cells, cancer, and cancer stem cells. *Nature*, 414(6859):105–111, 2001.

[31] D Rozek and G P Pfeifer. In vivo protein-dna interactions at the c jun promoter: preformed complexes mediate the uv response. *Mol. Cell. Biol.*, 13(9):5490–5499, 1993.

[32] S L Samson and N C Wong. Role of sp1 in insulin regulation of gene expression. *J. Mol. Endocrinol.*, 29(3):265–279, 2002.

[33] D. W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization.* John Wiley & Sons, New York, Chichester, Brisbane, Toronto, Singapore, 1992.

[34] I Shmulevich, E R Dougherty, S Kim, and W Zhang. Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18(2):261–274, 2002.

[35] F Verrecchia, J Rossert, and A Mauviel. Blocking sp1 transcription factor broadly inhibits extracellular matrix gene expression in vitro and in vivo: implications for the treatment of tissue fibrosis. *J. Invest. Dermatol.*, 116(5):755–763, May 2001.

[36] J C Wang and J E Dick. Cancer stem cells: lessons from leukemia. *Trends Cell. Biol.*, 15(9):494–501, 2005.

[37] R A Weinberg. *The Biology of Cancer*. Garland Science, Princeton, 1st edition, 2006.

[38] John M Winn and Christopher M. Bishop. Variational message passing. *J. Mach. Learn. Res.*, pages 661–694, 2005.

[39] H Wymeersch. *Iterative Receiver Design*. Cambridge University Press, New York, 2007.

[40] H G Xu, R Jin, W Ren, L Zou, Y Wang, and G P Zhou. Transcription factors sp1 and sp3 regulate basal transcription of the human irf-3 gene. *Biochimie*, 94(6):1390–1397, June 2012.

[41] Z Yang, S Ro, and B Rannala. Likelihood models of somatic mutation and codon substitution in cancer genes. *Genetics*, 165:695705, 2003.

[42] M Zou and S D Conzen. A new dynamic bayesian network (dbn) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics*, 21(1):71–79, 2005.