UNIVERSITY OF CALIFORNIA, SAN DIEGO

Statistical Approaches for Big Data Analytics and Machine Learning: Data-Driven Network Reconstruction and Predictive Modeling of Time Series Biological Systems

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Engineering Science (Mechanical Engineering)

by

Farzaneh Farhangmehr

Committee in charge:

Professor Daniel M. Tartakovsky, Chair
Professor Jan Kleissl
Professor Ratneshwar Lal
Professor Francesco Lanza Di Scalea
Professor Alison Marsden
Professor Padmini Rangamani

2014

The dissertation of Farzaneh Farhangmehr is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

_____
Chair

University of California, San Diego

2014

DEDICATION

To whoever is reading this!

EPIGRAPH

We are just an advanced breed of monkeys on a minor planet of a very average star. But we can understand the Universe. That makes us something very special.

—Stephen Hawking

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This dissertation owes its existence to the help, support, and inspiration of many people. In the first place, I wish to express my deepest sense of gratitude and sincere appreciation to my advisor, Professor Daniel Tartakovsky, for his encouragement and great revisions throughout the last four years. His support and flexibility provided me with the great opportunity to implement my ideas in a peaceful and intellectual working environment. Thanks Daniel for being a great professor, advisor, supervisor, and friend!

I would also like to thank my committee members Professors Lal, Professor Marsden, Professor Rangamani, Professor Kleissl and Professor Lanza Di Scalea for taking the time to be on my committee and providing their inputs to this work. I am also very grateful to have had the opportunity to collaborate with Professor Subramaniam. In addition, a sincere thank you to Professor Bahadori for all his help and support.

The illuminating discussions with various participants in the NSF CSoI (Center for Science of Information) program are gratefully acknowledged. CSoI truly advances the next generation of information theory through collaborative research. It has been a great pleasure for me to attend their events and serve as their council member during the last four years.

Special thanks also to all my UCSD friends and lab-mates, especially, Parastou − for opening my eyes to real-world challenges in bioinformatics, Francesca and Delphine − for being amazing officemates, friends and co-drinkers! I am really grateful for the excellent working atmosphere.

Lastly, I would like to thank my family for their love and support. I owe special gratitude to my parents for continuous and unconditional support in all my life's endeavors. I am also greatly indebted to my brother, Alex, for his unique support, unlimited encouragement and

sharing valuable experience. Thanks for making me proud! I also extend my appreciation to my lovely grandmother, Anis. I have no words to express my full gratitude to my family as I would not be where I find myself today without them.

I dedicate this dissertation to whoever goes through it! I hope that the direction of this research can inspire other PhD students to work on this area and implement my algorithms!

Chapter 2 is a reprint of: F. Farhangmehr, M. R. Maurya, D. M. Tartakovsky, and S. Subramaniam. Information theoretic approach to complex biological network reconstruction: Application to cytokine release in raw 264.7 macrophages. BMC Syst. Biol., 8:77, 2014.

Chapter 3 is a reprint of: F. Farhangmehr, and D. M. Tartakovsky. A Bayesian and information theoretic approach for predictive modeling of large-scale networks. Statistical Analysis and Data Mining Journal, 2014. *In review*

Chapter 4 is a reprint of: F. Farhangmehr, D.M. Tartakovsky, P. Sadatmousavi, M.R. Maurya, and S. Subramaniam. An information-theoretic algorithm to data-driven genetic pathway interaction network reconstruction of dynamic systems. In Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on, pages 214-217, Dec 2013.

Chapter 5 is currently being prepared for submission for peer review and publication: F. Farhangmehr, and D. M. Tartakovsky. Statistical approach to data analysis and reverse engineering of dynamic networks from time-course microarray data sets. 2014. *Under preparation*.

Chapter 6 is currently being prepared for submission for peer review and publication: F. Farhangmehr, P. Rangamani and D. M. Tartakovsky. Reverse Engineering of Gene Expression Data from Multiple Sclerosis Patients Undergoing Interferon-β Therapy. 2014. *Under preparation*.

VITA

2014	Doctor of Philosophy in Engineering Science (Mechanical Engineering), University of California, San Diego.
2009	Master of Science in Mechanical Engineering, Oregon state University.

JOURNAL PUBLICATIONS

F. Farhangmehr, M. R. Maurya, D. M. Tartakovsky, and S. Subramaniam. Information theoretic approach to complex biological network reconstruction: application to cytokine release in raw 264.7 macrophages. BMC Syst. Biol., 8:77, 2014.

F. Farhangmehr, D.M. Tartakovsky, P. Sadatmousavi, M.R. Maurya, and S. Subramaniam. An information-theoretic algorithm to data-driven genetic pathway interaction network reconstruction of dynamic systems. In Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on, pages 214-217, Dec 2013.

F. Farhangmehr, and D. M. Tartakovsky. A Bayesian and information-theoretic approach for predictive modeling of large-scale networks. Statistical Analysis and Data Mining Journal, 2014. *In review*

F. Farhangmehr, and D. M. Tartakovsky. Statistical approach to data analytics and reverse engineering of dynamic networks from time-course microarray data sets. 2014. *Under preparation*.

F. Farhangmehr, P. Rangamani and D. M. Tartakovsky. Reverse Engineering of Gene Expression Data from Multiple Sclerosis Patients Undergoing Interferon-β Therapy. 2014. *Under preparation*.

ABSTRACT OF THE DISSERTATION

Statistical Approaches for Big Data Analytics and Machine Learning: Data-Driven Network Reconstruction and Predictive Modeling of Time Series Biological Systems

by

Farzaneh Farhangmehr

Doctor of Philosophy in Engineering Science (Mechanical Engineering)

University of California, San Diego, 2014

Professor Daniel M. Tartakovsky, Chair

Ever-increasing quantity of data generated by modern technologies necessitates the development of advanced approaches for big data analytics. The ultimate goal of such approaches is to capture insightful patterns and turn them into actionable information. This information not

only reveals the hidden patterns underlying complex systems but also facilitates the design and development of new mechanisms to overcome multidisciplinary challenges. The data mining process can be divided into two steps: network reconstruction - to determine the structure and details of interactions, and predictive modeling - to represent constructed networks as predictive models capable of predicting the performance of systems under new conditions.

The main goal of this research is to develop algorithms and methodologies to overcome challenges in big data analytics. Statistical approaches for data-driven network reconstruction and predictive modeling developed in this research have several advantages: First, unlike most data-mining methods, they do not make any assumptions about the linearity, functional or parametric forms of variables. Second, they decrease the complexity of computations for time-series data sets. Finally, these algorithms are applicable to multiple systems, ranging from social networks to complex biological systems which are the main focus of this research.

We propose a Bayesian and information-theoretic approach for data-driven network reconstruction and predictive modeling of phosphoprotein-cytokine signaling networks in RAW 264.7 macrophages. To decrease computational complexities associated with dynamic networks, an algorithm is presented for network reconstruction of large-scale systems from time-course microarray data sets. The applicability of this algorithm is demonstrated by constructing the network of pathway interactions in yeast cell-cycle. This algorithm is implemented to also capture predictive models of dynamic networks and applied to reverse engineer E. coli under Ampicillin. Finally, we demonstrate a data-mining methodology for linking changes in gene expressions and health over time by reverse engineering a GEO dataset in which gene expressions of Multiple Sclerosis (MS) patients under Interferon-β therapy have been measured over a 10-year time interval.

Chapter 1

Introduction

1.1 Motivation

Ever-increasing quantity of data generated by modern technologies necessitates the development of advanced approaches for big data analytics to help understand the underlying mechanisms and networks. The main goal of all data mining techniques is to dig into data, extract the insightful patterns and turn them into actionable information. The process of data mining and reverse engineering of complex networks can be categorized into two main tasks: network reconstruction - to determine the structure and details of interactions, and predictive modeling - to represent the constructed network as a predictive model capable of predicting outputs for a given input.

Data mining techniques are typically grouped into three categories: optimization-based methods, regression analysis and statistical approaches. Optimization-based methods minimize the objective function on a feasible set. Regression techniques focus on the relationship between dependent variables and one or more independent variable(s). Finally, statistical approaches

analyze statistical dependencies of interactions by using correlation measurements as metrics to identify significant connections.

Since most of the above-mentioned techniques have been developed to analyze steady-state data sets, only a few of them work efficient for dynamic networks. In addition, most of these methods require assumptions of the linearity or functional and parametric forms of the variables. These shortcomings necessitate development of new algorithms and methodologies for reverse engineering of large-scale networks. A proper data-mining approach should not only accurately extract the hidden patterns behind massive amounts of data but also minimize the computational time and complexity. Our efforts are focused on developing such methodologies and applying them to several problems in computational systems biology and bioinformatics.

1.2 Challenges

The research described in the subsequent chapters is motivated by the following challenges.

Simplifying assumptions: As mentioned before, most of the current approaches for reverse engineering of big data make assumptions about the linearity or functional and parametric forms of the system. The majority of these methods often fail to reach their objectives when these assumptions are violated. We develop probabilistic methods capable of dealing with both linear and nonlinear systems, without resorting to simplifying assumption about their structure.

Computational cost: Analysis of large-scale data sets is a complex endeavor typically associated with significant computational cost. Developing specific methodologies to decrease these complexities is one of the most important challenges in data science. The proposed algorithms

decrease the computational cost and complexity by identifying potentially inter-related components and by using copula entropy as an estimator of mutual information.

Time-course data sets: Dealing with dynamic systems and time-course data sets is one of the most challenging task in data mining, primarily, due to their (often prohibitive) computational cost. Since most of data mining approaches are developed to deal with steady-state data, only a few of them can accurately and efficiently address challenges in extracting hidden patterns of time-course data sets. We develop algorithms specifically designed to address challenges in reverse engineering of dynamic networks from time-course data sets.

1.3 Objectives

The research described in the subsequent chapters aims to

1. Develop an information-theoretic method to data-driven network reconstruction, which does not require assumptions about the nature of an underlying complex system;

2. Develop a probabilistic method to infer a network of interactions in dynamic systems from time-course data sets, which is computationally tractable; and

3. Develop a probabilistic method for predictive modeling and use it to reconstruct predictive network models of biological systems.

1.4 Dissertation Outline

Information-theoretic approaches to data-driven network reconstruction. Chapter 2 of this dissertation presents an information-theoretic-based model to data-driven network recon-

struction of complex systems. Our approach provides a statistical method to reconstruct the networks of interactions without the necessity of taking the linearity, functional and parametric forms of variables into account. We use this method to overcome a significant challenge in systems biology which is the reconstruction of biological networks from measured data of different components. We demonstrate the applicability of the proposed approach by constructing phosphoprotein-cytokine signaling networks in RAW 264.7 macrophage cells. Since cytokines are secreted upon activation of a wide range of regulatory signals transduced by the phosphoprotein network, identifying these components can help identify regulatory modules responsible for the inflammatory responses.

Our approach to capture phosphoprotein-cytokine signaling patterns is based on estimation of mutual information of interactions by using kernel density estimators. Mutual information provides a measure of statistical dependencies between interacting components. Then, using the topology of the derived network, we develop a linear data-driven parsimonious input-output model of the phosphoprotein-cytokine network. For the phosphoprotein-cytokine network, this approach not only captures most of the known signaling components involved in cytokine release but also predicts new signaling components involved in the release of cytokines. The results of this study are important for gaining a clear understanding of macrophage activation during the inflammation process.

A Bayesian and information-theoretic approach to data-driven predictive modeling and network reconstruction. In chapter 3, we present a probabilistic algorithm, which employs information-theoretic and Bayesian approaches for predictive modeling and network reconstruction of complex systems. This algorithm does not assume the linearity of an underlying system and can be extended to dynamic systems.

We deploy this method to reconstruct the network model of signaling phosphoprotein-cytokine in Raw 264.5 macrophage (initially constructed in Chapter 2) from a data set including the concentrations of signaling phosphoproteins and to predict probability distributions of the released cytokines. To quantify the accuracy and robustness of this methodology, we compute the values for accuracy and F-measure of the predicted model and compare the reconstructed network with the initial and predicted networks obtained by other methods relying the linearity assumptions. The results of this study provide a probabilistic and systematic framework for nonlinear predictive modeling of complex networks.

Development of a statistical framework for network reconstruction from time-course data. In chapter 4, we address a challenging task of building network models from time-course data sets. A properly developed algorithm for time-course data sets should reduce the computation complexity and computational cost by identifying and avoiding unnecessary calculations. We introduce a probabilistic algorithm, which is designed to overcome these challenges without relying on assumptions about the nature (linearity, functional and parametric forms, etc.) of a system.

We demonstrate the applicability of this algorithm by employing it in systems biology. Specifically, we build a network model of genetic pathway interactions in yeast cell-cycle from time-course microarray data. Grouping genes into KEGG pathways, identifying potentially dependent pathways, and using copula entropy to measure the maximum mutual information of candidate pathways over all possible time intervals, helps us to identify the functional behavior and topology of significant pathway interactions in yeast cell cycle while significantly decreasing the computational cost and complexity.

Statistical approach to reverse engineering of dynamic networks from time-course microarray data. In Chapter 5, we combine the two algorithms described in Chapters 3 and 4 to reconstruct a network of gene interactions from a time-course data set of *E. coli*. This allows us to both capture the structure of parameters behind a biological system and predict the system's response to a certain condition.

The proposed method doesn't make any assumptions about the system and decreases the computational complexity by identifying the potentially related interactions. It uses copula entropy to measure statistical dependencies and then builds a network model using maximum mutual information over all possible time intervals. Using this network model and the predictive process suggested by this study, we estimate the performance of our system for any input. The developed methodology enables one to capture, predict and design biological mechanisms and responses.

Reverse Engineering of Gene Expression Data from Multiple Sclerosis Patients Undergoing Interferon-β Therapy. In Chapter 6 we apply the algorithm developed in Chapter 5 to a GEO data set, which consists of the gene expressions of multiple sclerosis (MS) patients undergoing Interferon-β treatment over a time interval of ten years.

Reconstruction of the network of gene interactions for MS patients undergoing Interferon-β therapy helps us understand the impact of Interferon-β therapy on humans. It also provides a potential diagnostic tool and therapeutic solutions for problems caused by changes in gene expressions for these patients. The suggested algorithm can be used in pharmaceutical industry for design and development processes, as well as for making informed decisions and developing mechanisms to forecast and overcome potential pitfalls.

Chapter 2

Information Theoretic Approach to Complex Biological Network Reconstruction: Application to Cytokine Release in RAW 264.7 Macrophages

Important Symbols used

f	Probability Density Function
I	Mutual Information
h	Bandwidth
f_h	Kernel Density Estimator Using Bandwidth h
$MISE$	Mean Integrated Squared Error
I_0	Threshold
p	p-Value
R^2	Coefficient of Determination
$RMSE$	Root Mean Squared Error

2.1 Introduction

Cellular functions and biological processes are the result of and are regulated by complex biochemical reactions within and between the cells [124, 74]. Bimolecular techniques can be used to measure concentrations of various molecular components, such as proteins and metabolites, allowing a partial reconstruction of the networks involving these components. A goal of systems biology is to reconstruct these underlying networks and to infer associated biological phenomena from large scale measurements [15]. More specifically, reconstruction of biological networks yields a framework for understanding the relationship between molecular measurements and higher-level phenotypes [4, 54].

Analyses of diverse read-outs from cells allow one to map an input onto responses associated with a given phenotype, i.e., to reconstruct the underlying biological network that results in the phenotype. Current computational approaches for network reconstruction include principal component regression (PCR) [58], partial least squares (PLS) regression [69], linear matrix inequalities (LMI) [30], and Bayesian Networks (BNs) [95]. These approaches are briefly described below.

PCR is a regression procedure that uses a principal component analysis to estimate re-

gression coefficients [58]. Usually, principal components with the highest variance are selected in three steps. First, a principal component analysis is performed on the data matrix of explanatory variables. Second, a least-squares regression is applied between the selected components (latent variables) and the output/response variables. Finally, the model's parameters are calculated for the selected explanatory variables by combining the two steps [111]. In contrast to PCR, PLS regression captures the maximum variance in the output variables while capturing sufficient variance in the input variables [69, 129]. PLS makes a linear model by projecting the input and output variables onto a new space [138, 32]. LMI converts nonlinear convex optimization problems into linear optimization problems [19]. The basic idea of the LMI is to approximate a given input/output modeling problem posed as a quadratic optimization problem with a linear objective and so-called LMI constraint [30]. Approaches such as PCR and PLS essentially work based on a linear model template. Bayesian networks are graphical models that describe causal or pseudo-causal interactions between variables [95, 43]. Nodes of a BN represent random variables in the Bayesian sense and edges represent conditional dependencies among the random variables [50]. BNs have a number of drawbacks related to the so-called representation problem: they require one to choose between discrete or continuous variables and parametric or non-parametric forms of the conditional probability distribution, and to decompose the joint probability distribution into conditional probability distributions among the relevant variables [105, 22].

Information-theoretic approaches provide a non-parametric alternative to Bayesian networks. They construct parsimonious models of biological networks by establishing statistical dependencies of interactions based on their uncertainty reductions [20, 70, 52]. Unlike PCR/PLS, this approach does not make any assumptions about the linearity of the system and the functional form of the statistical distribution of the variables [126, 97]. We describe our information-

theoretic approach to the reconstruction of biological networks in 2.2. This method is used in 2.3 to develop a parsimonious model of phosphoprotein-cytokine network in RAW 264.7 macrophages. In 2.4 and 2.5, we compare the regulatory components captured by our approach with those identified by previous approaches and the knowledge available in scientific literature.

2.1.1 Shannon's Information Theory

Building upon Hartley's conceptual framework [47], which relates the information of a random variable with its probability, Shannon [119] defined entropy of a random variable in terms of its probability distribution. For a random variable X given a random sample $\{x_1,...,x_n\}$ with probabilities $P(x_i)$, entropy H is defined as

$$H(X) = \sum_{i=1}^{n} P(x_i) \ln[P(x_i)]. \tag{2.1}$$

Shannon's information theory defines mutual information as the amount of information about a random variable X that can be obtained by observing another random variable Y. This definition implies that the information that Y provides about X reduces uncertainty about X due to the knowledge of Y. Intuitively, mutual information infers the information that Y and X share by measuring how much knowing one of the variables can reduce the uncertainty about the other [64]. Then, the mutual information of Y relative to X, or X relative to Y, is given by

$$I(X,Y) = H(X) + H(Y) - H(X,Y) = I(Y,X). \tag{2.2}$$

Mutual information provides a metric for measuring statistical dependencies of interactions. It has several advantages over other methods [20, 70, 52]: It does not make any assumption about

the functional form of the statistical distribution of variables [97]; and, information theoretic approaches are not dependent on the linearity assumption of the model for the ease of computation [126].

2.1.2 Threshold Selection on Mutual Information

A parsimonious model of a complex system has to capture a necessary and sufficient model of the entire system, while minimizing the number of interacting components, from the measured data for the system. The ultimate goal of data-driven network reconstruction methods is to achieve such a necessary and sufficient model. Information theoretic approaches analyze the statistical dependencies of interacting components by measuring the mutual information coefficients of interactions. A mutual information network of a complex system is obtained by computing the mutual information matrix (MIM) and selecting the threshold of mutual information (TMI). MIM is a square matrix, whose elements $MIM_{ij} = I(X_i, Y_j)$ are the mutual information between the variables X_i and Y_j. TMI defines the threshold of statistical dependencies of interactions. Choosing an appropriate TMI is a nontrivial problem. A straightforward but computationally demanding approach is to perform permutations of measurements several times and to recalculate a distribution of the mutual information for each permutation. Then permuted distributions are averaged and the largest mutual information in the averaged permuted distribution represents the threshold [13]. Some of the algorithms for network reconstruction and threshold selection in biological networks are discussed below.

The Relevance Network (RelNet) constructs a network in which a pair of random variables X_i and Y_j is linked by an edge if the mutual information $I(X_i, Y_j)$ is larger than a given threshold [14]. The Context Likelihood of Relatedness (CLR) algorithm derives a score from

the empirical distribution of the mutual information for each pair of random variables X_i and Y_j [35]. CLR estimates a score

$$Z_{ij} = \sqrt{Z_i^2 + Z_j^2} \tag{2.3}$$

where

$$Z_i = \max\left[0, \frac{I(X_i, X_j) - \mu_i}{\sigma_i}\right]. \tag{2.4}$$

Here μ and σ are the mean and standard deviation of the distribution of the mutual information of X_i and all other variables Y_j, $(j = 1, \ldots, n)$.

The Minimum Redundancy Network (MRNet) relies on the conditional mutual information to make inference. MRNet is applied to determine regulatory targets and pathways. If two random variables X and Y have a large mutual information but are conditionally independent given a third random variable Z, MRNet considers no statistical dependency between them [104]. ARACNE (Algorithm for the Reconstruction of Accurate Cellular NEtworks) assigns to each pair of nodes a weight equal to their mutual information and removes the weakest edges by applying a proper threshold [83]. ARACNE applies Kernel Density Estimation (KDE) approaches to measure mutual information between nodes and selects the bandwidth of kernels by minimizing the Kullback-Leibler distances between kernel density distributions of variables before and after removing the i-th observation. It also applies an information-theoretic property called the Data Processing Inequality (DPI) to remove statistically weak connections. DPI states that, if X_i interacts with X_j through a random variable X_k then

$$I(X_i, X_j) < \min[I(X_i, X_k), I(X_j, X_k)]. \tag{2.5}$$

We employ an information-theoretic approach both to reconstruct complex biological networks and to establish a parsimonious model of the entire system. Our strategy is to determine mutual information of interactions using kernel density estimators based on an unbiased cross-validation [121] estimation of kernel bandwidths and to analyze statistical dependencies of nodes by selecting a threshold obtained by applying the large deviation theory [20] employed by ARACNE [83].

2.2 Information-Theoretic Approach for Biological Network Reconstruction

As mentioned before, MI measures the information that X and Y share by measuring how much knowing one of these variables will reduce the uncertainty of the other and reflects the statistical dependencies of two variables. Hence, higher MI between an input and an output indicates a larger reduction in uncertainty and suggests a stronger input-output connection. Small (statistically zero) MI between two random variables indicates that variables are independent.

Measuring mutual information with a kernel density estimator (KDE)—a non-parametric method for estimating probability densities of variables—is more advantageous than histogram-based methods in terms of a better mean square error rate of convergence of the estimate to the underlying density [109]. A disadvantage of KDE is the need to specify an optimal kernel bandwidth [90]. Once the optimal kernel bandwidth is obtained and the MI coefficients of the network are measured using KDE, the next step is to select a proper threshold to determine the boundary of statistically significant connections and the weak connections to be removed. Following these three steps, information theoretic model of the network is obtained. It provides a

parsimonious network in which the number of false connections are reduced considerably.

The following subsections present a description of the above-mentioned steps to create a data-driven model of complex networks. These steps are applied to decipher, in a lumped manner, regulatory mechanisms involved in the release of 7 cytokines by activation of 22 signaling proteins in RAW 264.7 macrophage. The Alliance for Cellular Signaling (AfCS) has generated a systematic profiling of signaling responses and cytokine releases in RAW 264.7 macrophage [44, 1]. This dataset consists of data from stimulation of macrophages by both Toll and non-Toll receptor ligands. The objective is to create an input-output model, in which signaling responses (22 inputs) are used to predict cytokine release (7 outputs).

2.2.1 Nonparametric Estimations of Mutual Information

Kernel Density Estimation (KDE) is a non-parametric method to determine the Probability Density Function (PDF) of a random variable. Given a sample $\{x_1,...,x_n\}$ of a univariate random variable X with an unknown PDF $f_X(x)$, a kernel density estimator (KDE) estimates the shape of this function as [109]

$$f(x) = \frac{1}{nh^2\sqrt{2\pi}} \sum_{i=1}^{n} \exp\left[-\frac{(x-x_i)^2}{2h^2}\right], \tag{2.6}$$

where h is the kernel bandwidth and n is the size of the sample. A bivariate kernel density function of two random variables X and Y given their samples $\{x_1,...,x_n\}$ and $\{y_1,...,y_n\}$ is defined as

$$f_{XY}(x,y) = \frac{1}{2nh^2\pi} \sum_{i=1}^{n} \exp\left[-\frac{(x-x_i)^2+(y-y_i)^2}{2h^2}\right]. \tag{2.7}$$

The mutual information of X and Y is computed as [87]

$$I(X,Y) = \frac{1}{n}\sum_{i=1}^{n} \ln\left[-\frac{f_{XY}(x_i,y_j)}{f_X(x_i)f_Y(y_j)}\right]. \tag{2.8}$$

2.2.2 Selection of Optimal Kernel Bandwidth

The use of KDEs to evaluate the MI coefficients requires the optimal selection of the kernel bandwidth h. The main criterion used to determine the optimal kernel width is the minimization of the expected risk function, defined as the mean integrated squared error (MISE) between the computed and true (unknown) distributions [109, 90],

$$\mathrm{MISE}(h) = E\int [f_h(x) - f(x)]^2\,\mathrm{d}x, \tag{2.9}$$

where $f_h(x)$ is the kernel density estimate of the PDF $f_X(x)$ with bandwidth h. MISE cannot be used directly since it involves the unknown density function $f_X(x)$. To address this issue, several algorithms have been developed to get an estimate of the optimal bandwidth. One of the most commonly used algorithms employs a cross-validation type approach [132]. Based on this approach, if $f_h(x)$ is the kernel density estimation at x for a bandwidth of h using all of the data to fit the KDE, then a cross-validated estimate of the bandwidth is the value for h that minimizes [122, 121]

$$\int f_h^2(x)\,\mathrm{d}x - \frac{2}{n}\sum_{i=1}^{n} f_{(-i),h}(x_i). \tag{2.10}$$

Here $f_{(-i),h}(x_i)$ is the kernel density estimator using the bandwidth h at x_i obtained after removing ith observation. For two vectors X and Y, the cross-validation method determines the optimal kernel width for each pair of randomly selected set of n pairs of variables and the mean of

optimal kernel widths for these n pairs is used as an approximated kernel width for the entire dataset [133].

2.2.3 Network Reconstruction and Threshold Selection

Once the optimal kernel width has been selected and the MI matrix has been computed, the next step is to find an appropriate threshold of MI, I_0. Based on large deviation theory used by ARACNe algorithm [83], the probability that an empirical value of mutual information I is greater than I_0, provided that its true value is $\bar{I} = 0$, is

$$\mathbb{P}(I > I_0 \mid \bar{I} = 0) \equiv p \sim \mathrm{e}^{-cNI_0} \tag{2.11}$$

where c is a constant. Taking the natural logarithm of both sides yields

$$\ln p = a + bI_0 \tag{2.12}$$

where b is proportional to the sample size N. Therefore, $\ln p$ is a linear function of I_0 with the slope b. Using these results, for any dataset with sample size N and a desired p-value, the corresponding threshold can be obtained where a and b are fitted from the data. This threshold is used to remove statistically weak edges. Since each cytokine is explicitly an output we do not employ any further analysis such as DPI [20] to identify and remove indirect connections.

Using the network thus obtained, a predictive model is developed (see Appendix A for detail).

2.3 Application to Phosphoprotein-Cytokine Signaling Network

We employ this information theoretic approach to to reconstruct the phosphoprotein-cytokine network in RAW 264.7 macrophages. To achieve this goal, the first step is the creation of the MI matrix (MIM) interactions for each Toll and non-Toll data set separately and then finding a proper threshold for each network.

Macrophages play key roles in both innate and adaptive immunity, regulating the immune responses and the development of acute and chronic inflammations by producing a wide array of powerful chemical substances and regulatory factors such as cytokines [89]. Cytokines are a group of proteins and act as mediators between cells. Cytokines locate and interact with the target immune cells by binding to their receptor [114, 40]. The release of immune-regulatory cytokines is regulated by a complex signaling network [125, 107]. Multiple stimuli generate different signals and these signals generate different cytokine responses. Clear delineation of these signaling pathways is a prerequisite for understanding the causes of cytokine releases.

In order to determine the signaling components involved in the cytokine release, we used the AfCS data on the phosphoproteins and cytokines under Toll and non-Toll conditions. The information theoretic approach was employed to construct a reduced model that predicts the responses of seven cytokines (Tumor Necrosis Factor alpha or TNFα; Interleukin-1α or IL-1α; Interleukin-6 or IL-6; Interleukin-10 or IL-10; Granulocyte Macrophage Colony Stimulating Factor or GM-CSF; Regulated on Activation, Normal T Expressed and Secreted or RANTES; and Macrophage Inflammatory Protein-1α or MIP-1α) from the activation of 22 signaling proteins in RAW 264.7 macrophages. The latter include Signal Transducers and Activator of Transcription (STAT) 1α (STAT1α), STAT1β, STAT3, STAT5, Ribosomal Protein S6 (Rps6), Ribosomal S6 kinase (RSK), Glycogen Synthase Kinase (GSK) 3A (GSK3A), GSK3B,

Extracellular-signal Regulated Kinases (ERK) 1 (ERK1), ERK2, cyclic Adenosine Monophosphate (cAMP), c-Jun N-terminal Kinases (JNK) long (JNK lg), JNK short (JNK sh), AKT, p40 Phagocyte Oxidase (p40Phox), Ezrin [Ezr]/Radixin [Rdx](Ezr/Rdx), Membrane-organizing Extension Spike Protein (Moesin or MSN), P38, Sma and Mad related proteins 2 (SMAD2), Nuclear Factor Kappa-light-chain-enhancer of activated B cells p65 (NF-$\kappa\beta$ p65), Protein Kinase C Delta (PKCD) and Protein kinase Cμ2 (PKC μ2).

Our data consist of Toll and non-Toll sets. A reduced model of each set was obtained by applying the principles of information theory described above. Combining these two models, we obtained the network model based on the entire data set. The resulting network provides a parsimonious phosphoprotein-cytokine model, in which the number of signaling components involved in cytokine releases is minimized considerably. This model not only successfully captures most of the known signaling components involved in cytokine releases, but also predicts new signaling components involved in releases of cytokines.

2.4 Results

The proper kernel bandwidth has been estimated by applying the above-mentioned cross-validation approach (Eq. (2.10)). For the Toll data set, the bandwidth $h = 0.14$ and for the non-Toll data set, $h = 0.17$. Figure 2.1 shows the probability density functions of 7 released cytokines, as inferred by the KDE in Eq. (2.6) computed through the MATLAB function ksdensity [2] using the estimated value of h. All of the estimated densities are highly non-Gaussian. In this figure, the x-axis shows the measured values of cytokines after being normalized and the y-axis demonstrates their densities by applying KDE.

Using these kernel density estimators, we used Eq. (2.6) to compute the MI coefficients

of all protein-cytokine connections for the Toll and non-Toll datasets. Figure 2.2 shows these coefficients as a bar-graph, with the corresponding thresholds shown by the dashed lines (I_0 = 0.19 for Toll data and I_0 = 0.17 for non-Toll data). The MI coefficients below these thresholds are considered to be statistically insignificant and discarded without any significant impact.

Figure 2.3 shows the reconstructed networks obtained from the non-Toll (left panel and orange nodes) and Toll (right panel and pink nodes) data for 22 signaling phosphoproteins and 7 cytokines. These two networks are combined to yield the network of the entire system, which is shown in Figure 2.4. Blue nodes in Figure 2.4 show phosphoproteins involved in both datasets. This network captures most of the known signaling components involved in cytokine releases and confirms the potentially important novel signaling components that have been suggested recently by other methods, such as PCR [107]. Our approach also identifies new signaling components involved in the release of cytokines, including Ribosomal S6 kinase on TNFα.

Since phosphoproteins may also have regulatory impacts on other phosphoproteins, the above mentioned process is applied again to capture all the significant phosphoprotein-phosphoprotein and phosphoprotein-cytokine connections in one network. The mutual information matrix of all interactions is built again and the proper kernel bandwidth and threshold is selected ($h = 0.14$ and $I_0 = 0.20$ for Toll data and $h = 0.17$ and $I_0 = 0.17$ for non-Toll data). Figure 2.5 shows the reconstructed networks obtained from the non-Toll (left panel and orange nodes) and Toll (right panel and pink nodes) and Figure 2.6 is the final network obtained by combining the two networks in Figure 2.5 containing significant phosphoprotein-phosphoprotein and phosphoprotein-cytokine connections in the entire system.

To demonstrate the robustness of our results, this network is built again by capturing the networks of each cytokine individually and combining the seven reconstructed networks.

Figure 2.7 shows the networks obtained from node-by-node analysis for TNFα (left panel) and IL-6 (right panel). In comparison with the network of Figure 2.6, such a network does not capture the regulatory effect of PKCμ2 on G-CSF for Toll-data and cAMP on IL-6 and AKT on TNFα from non-Toll data. As the lower panel in Figure 2.2 shows, the mutual information of these interactions are very close to the selected threshold. All other connections present in Figure 2.6 are also included in such a network.

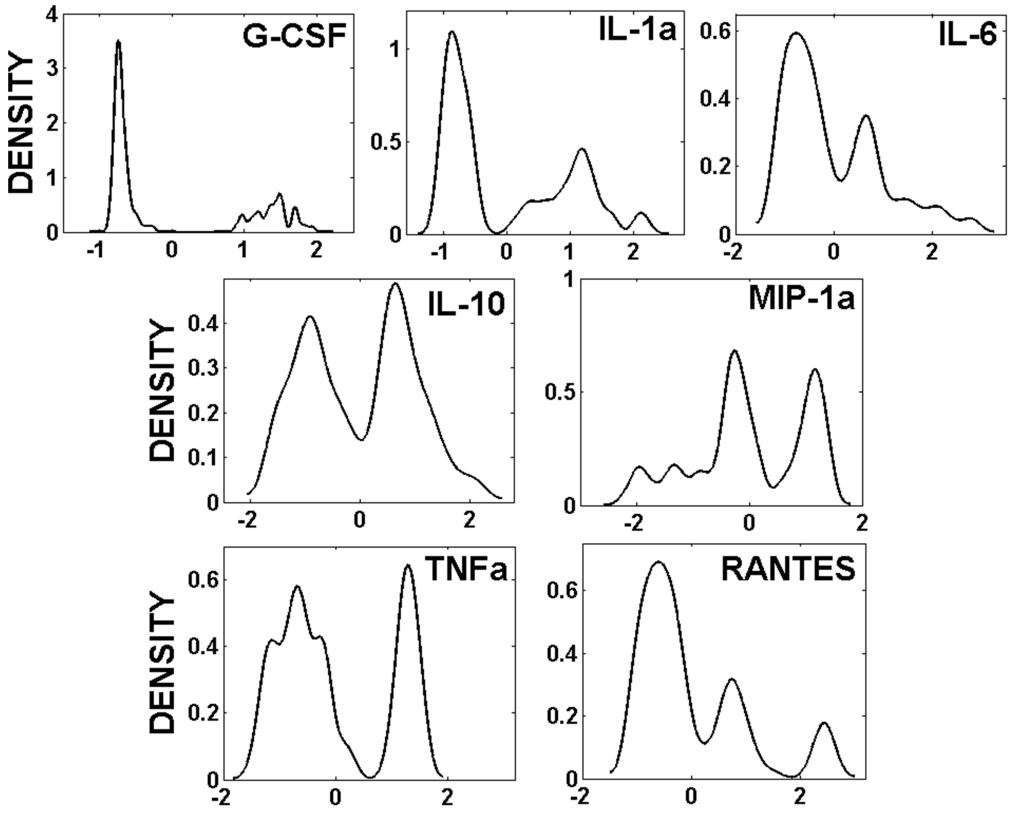

Figure 2.1: Kernel density estimations (y-axis) of seven released cytokines (x-axis) in RAW264.7 macrophage cells upon stimulation with ligands, using kernel bandwidth $h = 0.14$ (Toll data).

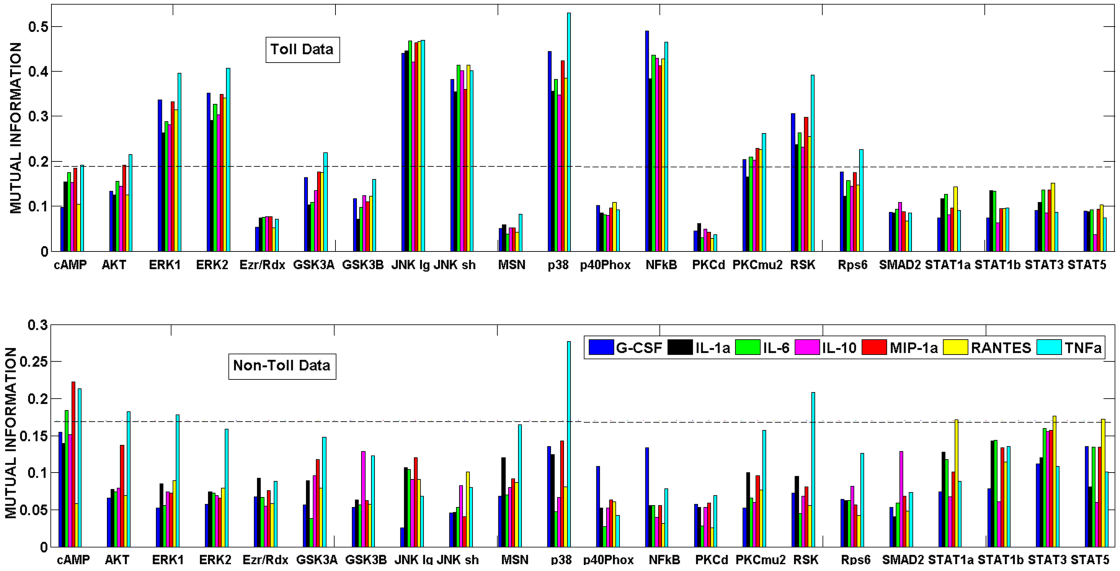

Figure 2.2: Mutual information of all phosphoprotein-cytokine pairs from Toll (the upper bar) and non-Toll (the lower bar) datasets. Thresholds ($I_0 = 0.19$ for Toll data and $I_0 = 0.17$ for non-Toll data) are shown by dashed lines.

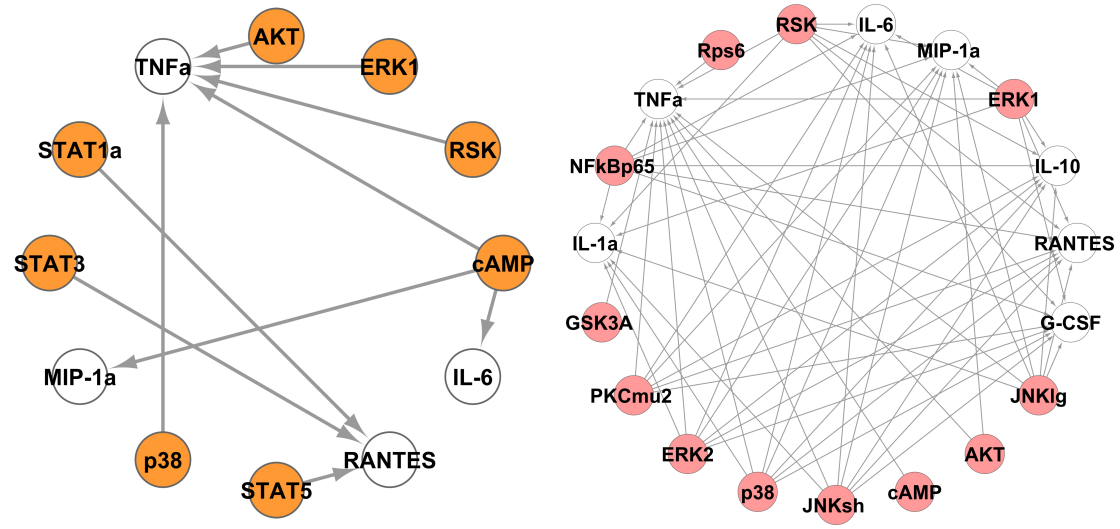

Figure 2.3: Reconstructed networks of signaling phosphoproteins-cytokines obtained from the non-Toll (left panel with orange nodes for the phosphoproteins) and Toll (right panel with pink nodes for the phosphoproteins) data.

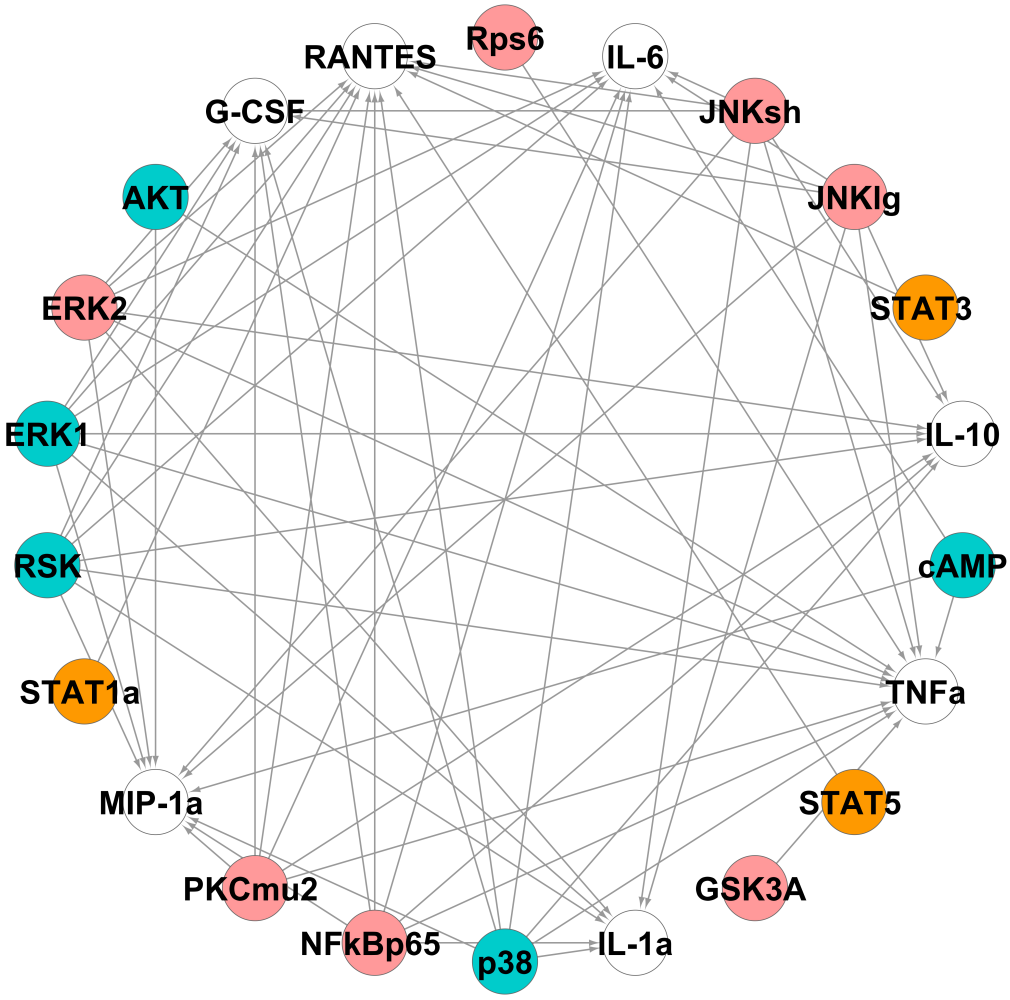

Figure 2.4: The reconstructed phosphoprotein-cytokine network obtained by combining networks from non-Toll dataset (orange nodes) and Toll da-taset (pink nodes). Blue nodes are phosphoproteins involved in both datasets and white nodes represent the cytokines (outputs).

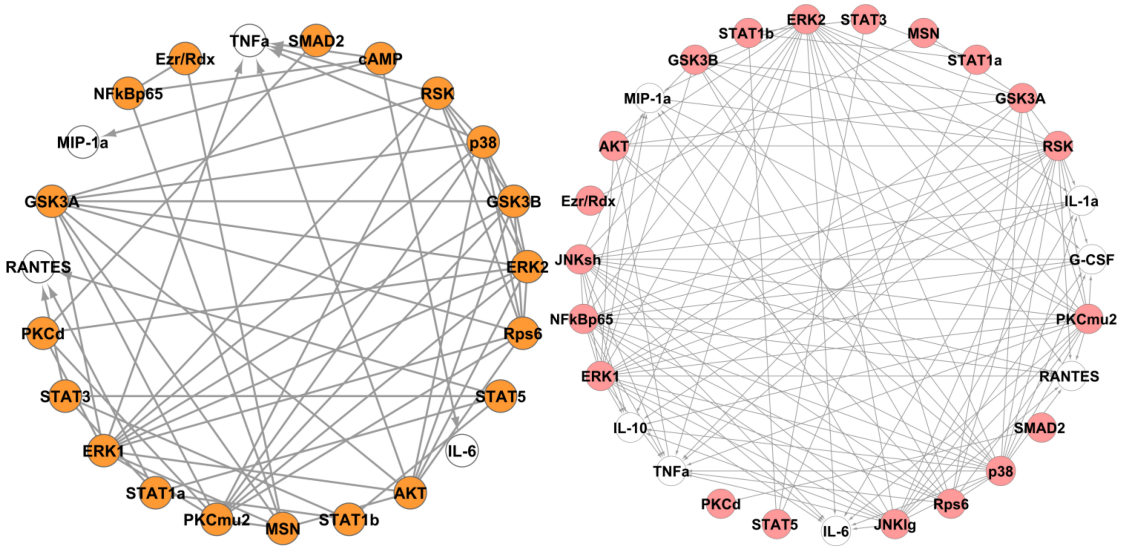

Figure 2.5: Reconstructed networks of phosphoprotein-phosphoprotein / phosphoprotein-cytokine obtained from the non-Toll (left panel and orange nodes) and Toll (right panel and pink nodes) data.

2.5 Discussion

The information theoretic approach accurately identifies the main signaling phosphoproteins involved in cytokine release (Figure 2.4) and the corresponding linear model predicts the quantitative levels of cytokine releases (Figure A.1) reasonably well. We analyzed both Toll and non-Toll datasets. Non-Toll data is required to identify the regulatory effects of STAT1α, STAT1β, STAT3, STAT5 and cAMP and Toll-data provides information about PKCμ2, JNK lg, JNK sh and NF-$\kappa\beta$ P65 and ERK2. ERK1, AKT, P38 and RSK are identified as significant in both datasets. We provide a comparison of the regulatory components necessary for cytokine releases identified by the information theoretic approach and other computational methods and biochemical knowledge available in literature such as PCR with statistical significance testing [107]. The results of this comparison are summarized in Table 2.1. Activated macrophages secrete cytokines [78]. Various pathways transmit the signals that initiate cytokine production

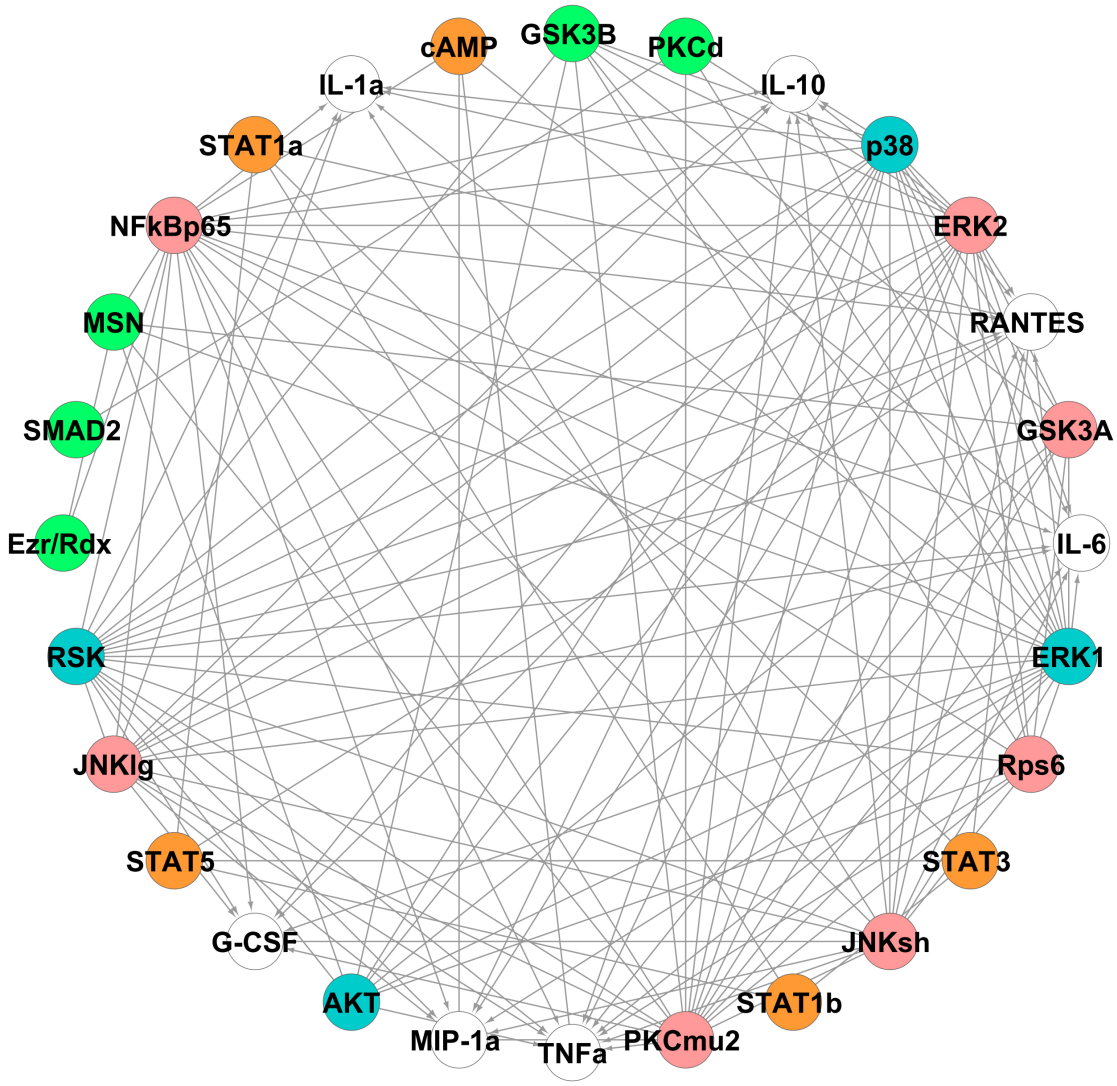

Figure 2.6: The reconstructed phosphoprotein-phosphoprotein/cytokine network from combining networks from non-Toll dataset (orange) and Toll dataset (pink).

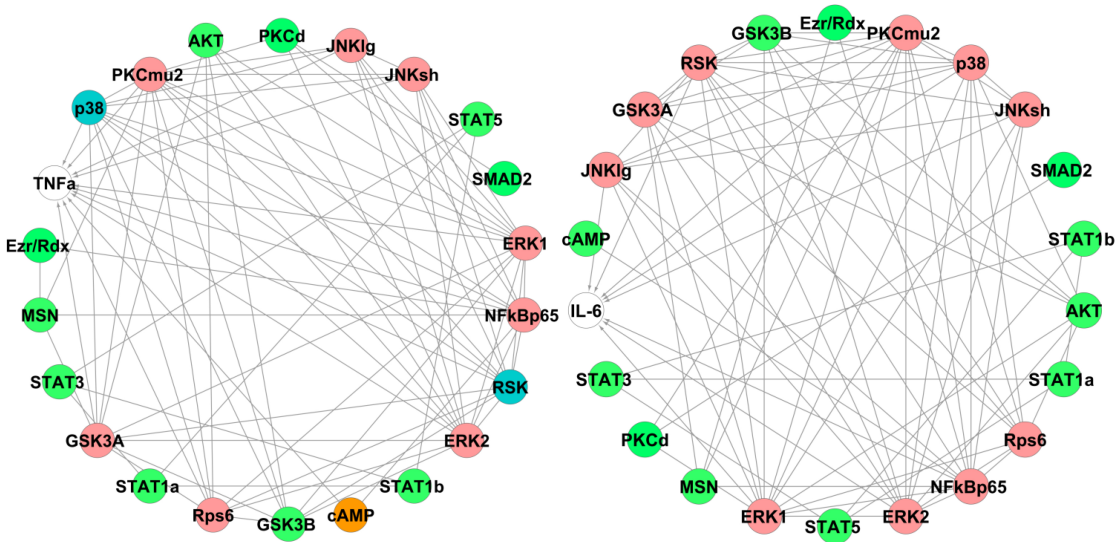

Figure 2.7: Node-by-node reconstructed networks of TNFα(left panel) and IL-6 (right panel) after combining non-Toll dataset (orange nodes) and Toll dataset (pink nodes). Blue nodes are involved in cytokine regulation from both datasets and green nodes are not directly involved in cytokine regulation.

[120, 66]. Cytokines are classified based on their functions or their sources [78, 101]. They can be grouped into anti-inflammatory and pro-inflammatory cytokines based on their functional role in inflammatory responses. Pro-inflammatory cytokines such as TNFα, IL-1α and GM-CSF induce both acute and chronic inflammatory responses. Anti-inflammatory cytokines, such as IL-10 limit the magnitude of inflammation and chemokines, such as MIP and RANTES are involved in chemotaxis of leukocytes.

Pro-inflammatory Cytokines. Granulocyte/macrophage Colony Stimulating Factor (G-CSF) regulates the production of neutrophil G granulocytes and stimulates the function of mature neutro-phils [98]. We identify the phosphoproteins PKCμ2 [67], NF-$\kappa\beta$ p65 [[140], JNK lg/sh [16], P38, RSK [60] and ERK1/2 [128] as the main regulators for the production and release of G-CSF. Tumor Necrosis Factor alpha (TNFα) is involved in normal host defense in both mediating inflammatory and immune responses [8]. Our study captures the largest network of

regulatory components for TNFα which consists of twelve signaling phosphoproteins: RSK, AKT, RPS6, PKCμ2, GSK3A, cAMP, ERK1/2, JNK sh/lg, NF-$\kappa\beta$ p65 and P38. Some studies suggest the regulatory impact of STAT1α and STAT1β, on TNFα [136]. Both our network and the network from PCR minimal model [107] missed these connection. Interleukin-1alpha (IL-1α) is produced by activated macrophages and is responsible for inflammation [28]. The information theoretic approach identifies cAMP, JNK lg/sh, ERK1/2, P38 and NF-$\kappa\beta$ p65 as the main regulators of production/release of IL-1α .

As Table 2.1 shows, this study identifies most of signaling components of pro-inflammatory cytokines captured by other computational methods and strongly confirms the regulatory effect of P38 which has been proposed by the PCR minimal model in 2006 [107]. Unlike the PCR minimal model [107], our approach successfully captures the regulatory effects of ERK1 and ERK2 on GCS-F [128] and TNFα [84]. It confirms the regulatory effect of GSK3A on TNFα [33] which have been suggested by studies. NF-$\kappa\beta$, ERK, JNK (targets c-Jun [27]) and Sp1 are the transcriptional activators of TNFα [71, 131]. In this light, our results show good agreement with other studies by capturing all signaling components identified by the PCR minimal model, in addition to predicting the known regulatory effects of ERK1/2, GSK3A (regulated by c-Jun which is affected by JNK) [140, 33, 131]. The information theoretic approach also identifies RSK, a substrate of ERK [42], as a potentially novel regulatory component involved in the release of TNFα.

P38 (from Toll data) has the strongest and ERK1 (from non-Toll data) has the weakest regulatory impact on TNFα. As Figure A.1 shows, TNFα yields the best linear fit in terms of the coefficient of determination (R^2 = 0.62), which is in good agreement with other models obtained by PCR [107] and PLS [140] methods. NF-$\kappa\beta$ p65 represents the highest statistical

dependency while PKCμ2 has the lowest mutual information coefficient among the captured regulatory network components of GCS-F. JNK lg (from Toll data) shows the highest regulatory effects on IL-1α.

Anti-inflammatory cytokines. Interleukin-10 (IL-10) is an anti-inflammatory cytokine that has important roles in immune regulation and inflammation [112]. Our approach shows the regulatory effects of PKCμ2 [130], P38 [29], RSK [94], ERK1/2 [108], NF-$\kappa\beta$ p65 [137] and JNK sh/lg, on IL-10. Macrophage Inflammatory Protein-1α (MIP-1α) belongs to the group of CC chemokines that regulate several inflammatory responses including trafficking and activation of leukocytes, as well as the fever response [23]. We capture the regulatory effects of cAMP [5], AKT [72], RSK, ERK1/2 [11], P38 [24], JNK sh/lg [73] and NF-$\kappa\beta$ p65 [12] on MIP-1α. One study suggests the regulatory effects of STAT1α/β and STAT3 on MIP-1α [46]. The PCR minimal model [107] only identifies STAT1α as a significant component of MIP-1α. Regulated on Activation, Normal T Expressed and Secreted (RANTES), is a CC chemokine and has a key role in recruiting leukocytes into inflammatory sites [113]. The information theoretic approach suggests that STAT3, STAT5, STAT1α, NF-$\kappa\beta$ p65, PKCμ2, P38 JNK lg/sh, ERK1/2 and RSK regulate RANTES and unlike the PCR minimal model [107], it is in good agreement with the cytokine literature.

As indicated in Table 2.1, the network identified by our study includes most of known identified signaling components of anti-inflammatory cytokines described in the literature and unlike the PCR minimal model [107], captures the regulatory effects of NF-$\kappa\beta$ p65, ERK1/2 on MIP-1α. Some studies suggest that the TLR ligand pathways that activate IL-10 are P38 dependent and NF-$\kappa\beta$ signaling pathway has no contribution on the activation of IL-10 [10, 80]. However, our study and the PCR model [107] identify the regulatory effects of JNK lg/sh which

are activated through NF-$\kappa\beta$ p65.

The information theoretic approach and PCR [107] models both yield low coefficient of determination for cytokines ($R^2 < 0.8$) possibly due to their regulations by unmeasured pathways and/or a nonlinear relationship between the levels of cytokines and the phosphoproteins. In comparison to the PCR approach, information theoretic approach shows a better agreement with known regulatory components in the literature. The high variability of data (low coefficient of determination) might explain this by considering the fact that when noise or variability is high, the threshold used in the PCR approach is high so that it identifies a relatively lesser number of components as being significant. The non-linear nature of the biological processes might be an explanation for the failure of PCR to identify the regulatory effects of ERK1/2, cAMP and RSK on cytokines. JNK lg (from Toll data) has the strongest effect and AKT (from non-Toll data) has the weakest effect on MIP-1α. Our network shows the highest mutual information (from non-Toll data) for NF-$\kappa\beta$ and IL-10. PKCμ2 has the weakest regulatory effects on IL-10. JNK lg has the strongest regulatory effect on RANTES and STAT3 shows the lowest statistical dependencies to it.

Interleukin-6. Interleukin-6 (IL-6) is secreted by macrophages and T cells and acts as both a pro-inflammatory and anti-inflammatory cytokine [117]. Our model identifies the regulatory effects of phosphoproteins RSK, PKCμ2, ERK1/2, JNK sh/lg, P38, NF-$\kappa\beta$ and cAMP. The regulatory roles of cAMP [26] and P38 [3] which could not be captured by the PCR minimal model [107], are identified by the information theoretic approach. JNK lg (from Toll data) yields the strongest regulatory effect and cAMP (from non-Toll data) yields the weakest regulatory effect on IL-6.

Overall, our network model and quantitative predictions are in good agreement with

other studies available in literature and captures most of known regulatory components involved in cytokine release. Our model confirms the regulatory effect of P38 on G-CSF that has been suggested by the PCR minimal model several years ago [107] and captures one potentially novel regulatory effect of RSK on TNFα. The advantages of the information theoretic method has been demonstrated by comparing the accuracy of its parsimonious model to the models obtained by other computational methods such as PCR minimal models in predicting the regulatory components for cytokines with high variability and low coefficient of determination.

2.6 Conclusion

Identifying the regulatory components for cytokines is critical for understanding the mechanisms that control their production and release in immune cells. In recent years, several computational methods have been applied to develop networks which have led to an improved understanding of cytokine releases in macrophages. In this work, we developed a parsimonious input-output model of regulatory phosphoprotein-cytokine network based on an information theoretic approach. Our model demonstrated the applicability of this approach to the data-driven reconstruction of biological network. The data, which consisted of a systematic profiling of signaling responses in RAW 264.7 macrophage cells upon treatment with Toll- and non-Toll receptor ligands, was provided by the Alliance for Cellular Signaling (AfCS). Information theoretic approach as a non-parametric method identified the regulatory components (phosphoproteins) on which specific cytokines showed significant statistical dependence (measured in terms of mutual information). The reconstructed network also was able to capture the regulatory network of phosphoproteins interactions. We calculated mutual information of interactions by using kernel density estimator (KDE) and discarded weak connections using proper thresholds. Using such

a parsimonious list of significant inputs, a predictive model was also developed for each of the cytokines which predicted a separate test data well. Most of the significant connections are validated against the known literature. Some novel connections, such as Ribosomal S6 kinase for Tumor Necrosis Factor are also identified by the mutual information approach, which were not detected by the PCR approach. These novel regulatory components serve as testable hypotheses.

2.7 Acknowledgements

This research was supported by the National Science Foundation (NSF) collaborative grants DBI-0835541 and STC-0939370, and National Institutes of Health (NIH) collaborative grants U54GM69338, R01HL106579 and R01HL108735 to SS.

This Chapter is a reprint of the material as it appeared in:

F. Farhangmehr, M. R. Maurya, D. M. Tartakovsky, and S. Subramaniam. Information theoretic approach to complex biological network reconstruction: Application to cytokine release in raw 264.7 macrophages. BMC Syst Biol, 8:77, 2014.

The dissertation author was the primary investigator and author on this paper.

Table 2.1: A Comparison of phosphoprotein-cytokine regulatory connections identified by information-theoretic approach ('MI'), PCR ('PCR') and the literature knowledge ('Lit').

Interactions		MI	PCR	Lit.	Interactions		MI	PCR	Lit.
G-CSF (Pro-infl.)	NF-KB	Y	Y	Y [140]	IL-6 (Anti-infl. and Pro-infl.)	RSK	Y	N	Y [110]
	JNK lg	Y	Y	Y [16]		JNK lg	Y	Y	Y [141]
	JNK sh	Y	Y	Y [16]		JNK sh	Y	Y	Y [141]
	P38	Y	Y	N		P38	Y	N	Y [3]
	PKCμ2	Y	N	Y [67]		PKCμ2	Y	N	Y [115]
	ERK1	Y	N	Y [128]		NF-KB	Y	Y	Y [34]
	ERK2	Y	N	Y [128]		ERK1	Y	N	Y [110]
	RSK	Y	N	Y [60]		ERK2	Y	N	Y [110]
						cAMP	Y	Y	Y [26]
TNFα (Pro-infl.)	RSK	Y	N	N	IL-10 (Anti-infl. and Pro-infl.)	JNK lg	Y	Y	N
	AKT	Y	Y	Y [102]		P38	Y	N	Y [29]
	P38	Y	Y	Y [24]		ERK1	Y	N	Y [108]
	RPS6	Y	N	Y [71]		ERK2	Y	N	Y [108]
	GSK3A	Y	N	Y [33]		JNK sh	Y	Y	N
	GSK3B	N	N	Y [33]		NF-KB	Y	Y	Y [137]
	PKCμ2	Y	N	Y [57]		PKCμ2	Y	N	Y [130]
	cAMP	Y	Y	Y [100]		RSK	Y	N	Y [94]
	NF-KB	Y	Y	Y [33]					
	JNK lg	Y	Y	Y [27]					
	JNK sh	Y	Y	Y [27]					
	ERK2	Y	N	Y [84]					
	ERK1	Y	N	Y [84]					
MIP-α (Anti-infl.)	P38	Y	Y	Y [24]	RANTES (Anti-infl.)	STAT3	Y	N	Y [68]
	NF-KB	Y	Y	Y [12]		STAT5	Y	N	Y [63]
	cAMP	Y	Y	Y [5]		STAT-1α	Y	Y	Y [99]
	RSK	Y	N	Y [11]		NF-KB	Y	Y	Y [51]
	JNK lg	Y	Y	Y [73]		P38	Y	N	Y [21]
	JNK sh	Y	Y	Y [73]		PKCμ2	Y	N	Y [59]
	AKT	Y	N	Y [72]		JNK sh	Y	Y	Y [51]
	ERK1	Y	N	Y [11]		JNK lg	Y	Y	Y [51]
	ERK2	Y	N	Y [11]		RSK	Y	N	Y [142]
	STAT1α	N	Y	Y [46]		ERK2	Y	N	Y [142]
	STAT1β	Y	N	Y [46]		ERK1	Y	N	Y [142]
	STAT3	Y	N	Y [46]					
IL-1α (Pro-infl.)	ERK2	Y	N	Y [53]					
	ERK1	Y	N	Y [53]					
	RSK	Y	N	Y [9]					
	P38	Y	N	Y [48]					
	JNK lg	Y	Y	Y [6]					
	JNK sh	Y	Y	Y [6]					
	NF-KB	Y	Y	Y [88]					

Chapter 3

A Bayesian and Information-Theoretic Approach to Data-driven Predictive Modeling and Network Reconstruction of Complex Networks

Important Symbols used

f	Probability Density Function
I	Mutual Information
h	Bandwidth
f_h	Kernel Density Estimator of f Using Bandwidth h
$MISE$	Mean Integrated Squared Error
I_0	Threshold
p	p-Value

3.1 Introduction

A central challenge in information science is the reverse engineering of complex networks to capture the underlying structure of interactions and then, by representing these interactions as a network, build a predictive model that best predicts the performance of this network (output) under a specific situation (a given dataset as input). Therefore, data-mining tasks can be broken up into two tasks: descriptive tasks and predictive tasks.

A predictive modeling process builds a model that expresses the target variable as a function of the explanatory variables [65]. Hence, the goal of this model would be to minimize the difference between the predicted values and the real values. Predictive modeling and analytics techniques can itself be categorized in three general approaches [85]: traditional, data-adaptive and model dependent. The traditional approach involves estimation of parameters for linear interactions. Data-adaptive approaches are data-driven and adapt to the available data to represent non-linear relationships among variables. Hence, data-adaptive approaches search through data to find the useful non-linear relationships. Model-dependent approaches, such a simulation methods, specify a model first and then use this model to generate data and make predictions.

Predictive modeling techniques are applied in various areas such as marketing - to find uncertainties (risks/opportunities) of a given customer or a specific decision to be made, model-

based design and decision making − to predict the behavior of a complex engineered systems when the physical phenomenon is not accessible or measurable, and climate science − to capture the dynamics in climate data. Accordingly, this paper is presenting a data-adaptive approach for data-driven predictive modeling of complex systems in application of systems biology.

In this paper, we provide a novel algorithm providing a Bayesian and information-theoretic framework for predictive modeling of complex networks to capture a model capable of predicting output for a given dataset and reconstructing the network again using the predicted output. To show the applicability of our approach, we apply this framework to predict the probability densities of seven released cytokines from the activation of 22 signaling phosphoproteins in a constructed network model of Raw 264.5 macrophage. Then, using information-theoretic approaches, we reconstruct this signaling phosphoprotein-cytokine network again and compare the obtained network with the original network model regarding of the number of false positive and false negative interactions, which is an indication of the accuracy and F-measure of the obtained network. This methodology is applied to develop predictive models of two different phosphoprotein-cytokine networks constructed under two different p-values (p=0.0001 and 0.005). A comparison of these predicted models with a non-probabilistic predictive model obtained from least square method will be presented in discussion section. The proposed methodology and its application in phosphoprotein-cytokine signaling network are presented in 3.2 and 3.3 In this section, a brief description of the methods and materials used by our proposed methodology is presented.

3.1.1 Kernel Density Estimation (KDE)

Suppose we are given a sample $\{x_1, \ldots, x_n\}$ of a random variable X, whose probability density function (PDF), $f_X(x)$, is unknown. A kernel density estimation (KDE) of $f_X(x)$ is computed as [109]

$$f_X(x) = \frac{1}{nh^2\sqrt{2\pi}} \sum_{i=1}^{n} \exp\left[-\frac{(x-x_i)^2}{2h^2}\right] \tag{3.1}$$

where h is the kernel bandwidth and n is the sample size. Suppose further that we are also given a sample $\{y_1, \ldots, y_n\}$ of a random variable Y. A KDE of the joint PDF, $f_{XY}(x,y)$, of random variables X and Y is

$$f_{XY}(x,y) = \frac{1}{2nh^2\pi} \sum_{i=1}^{n} \exp\left[-\frac{(x-x_i)^2 + (y-y_i)^2}{2h^2}\right]. \tag{3.2}$$

KDE has several advantages over other estimation methods. It is a nonparametric method, and has a better mean square error rate of convergence of the estimate to the underlying PDF [109]. However, the use of KDEs requires the selection of an optimal kernel bandwidth h [90]. The process of selecting the optimal kernel bandwidth is described below.

3.1.2 Bayesian Network

A Bayesian network, as a statistical approach, is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). In a Bayesian network, nodes V are random variables and edges E between the nodes represent probabilistic dependencies among the corresponding nodes (or random variables). A network is considered a Bayesian network if its joint PDF can be written as the product of the individual (marginal) PDFs of its nodes, conditional on their parents variables [103, 95]. Hence,

if $G = (V, E)$ is a Bayesian network and $X = \{X_v : v \in V\}$ is a set of n random variables indexed by V, then

$$f_X(x) = \prod_{v=1}^{n} f_{X_v|X_j}(x) \tag{3.3}$$

for each X_j that is a parent of X_v. Equation (3.3) implies that variables X_v are conditionally independent from their non-descendants, given the values of their parent variables.

Bayesian models play a significant role in many fields, such as risk and reliability engineering, machine learning and bioinformatics. We propose a methodology, which combines the principles underlying Bayesian networks and information theory. It results in a framework for predicting PDFs of variables and reconstructing predictive models of (biological) networks. The proposed methodology is described below.

3.2 A Probabilistic Approach for Predictive Modeling of Complex Networks

The proposed methodology develops a strategy capable of predicting output values from any given input dataset; then, using these values, it builds a predictive network model. This algorithm relies on two probabilistic frameworks, Bayesian nets and an information-theoretic approach. It is free of any assumptions about the functional form and linearity of the system. If the input has the same PDF as that of the original network, regardless of the method used for initial reconstruction of the original network, then the predicted PDFs of the output should ideally be the same as the PDFs of the output in the original network. Therefore, a predictive model should ideally be able to predict a network's performance with 100% accuracy and reconstruct the same network for a test dataset (assumed to have the PDF of training data).

Our goal is predict the PDFs of a system from the input test data. The original network used for this prediction is a constructed network model of signaling phosphoprotein-cytokines in Raw 264.5 macrophage cells [37]. We deploy an information-theoretic approach to reconstruct the initial network using these predicted values. This reconstructed network shows a good agreement with the original network. High values for accuracy and F-measure of this model demonstrate the applicability of this approach.

The proposed algorithm consists of the following six steps, the first three of which are predictive (i.e., used to predict a system's performance) and the remaining three are descriptive (i.e., used to reconstruct the predicted network model).

1. From an initial network (or a training dataset), detect sets of random variables responsible for regulating each node of output. In our examples, these are significant signaling phosphoproteins and released cytokines in the original network identified in [37].

2. Measure prior distributions and likelihood functions.

3. Predict posterior density functions for outputs and build the prediction matrix.

4. Find mutual information (MI) of interactions for new measurements.

5. Select a proper MI threshold and remove connections below this threshold.

6. Reconstruct the network by considering only remaining interactions (connections, whose MI is above the threshold).

A detailed description of these steps is presented below, followed by a discussion of the results in Sections 3.3 and 3.4.

3.2.1 Predictive Module (Steps 1-3)

Suppose that a probabilistic network reconstruction procedure (e.g., the one described in the previous Chapter) has identified a network in which a node (random variable) Y_j is connected to (regulated by) a set of n nodes (random variables) $\mathbf{X} = \{X_i\}_{i=1}^n$. Among other information, this procedure has yielded a PDF $f_{Y_j}(y)$ of Y_j. The predictive component of our algorithm aims to probabilistically forecast the behavior of Y_j from a new set of k measurements $\{x_{i1}, \ldots, x_{ik}\}$ of each of the random variables X_i. We accomplish this by employing a Bayesian data assimilation technique [139], which treats $f_{Y_j}(y)$ as a *prior* PDF of Y_j.

To assimilate a set of measurements $\mathbf{x}_{\text{test}}$ (an $n \times k$ matrix), we treat them as independent random variables $\mathbf{X}_{\text{test}}$ (an $n \times k$ matrix) that follow a Gaussian (conditioned on y) data model [139]

$$f_{X_{i,\text{test}}|Y_j}(\mathbf{x}_{i,\text{test}}|y) = \prod_{v=1}^{k} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(x_{iv}-y)^2}{2\sigma_i^2}\right], \qquad i = 1, \ldots, n \tag{3.4}$$

where σ_i^2 is the variance of X_i. This is a *likelihood function* to be used in Bayes' rule,

$$f_{Y_j|\mathbf{X}_{\text{test}}}(y|\mathbf{x}_{\text{test}}) = \frac{f_{\mathbf{X}_{\text{test}}|Y_j}(\mathbf{x}_{\text{test}}|y) f_{Y_j}(y)}{f_{\mathbf{X}_{\text{test}}}(\mathbf{x}_{\text{test}})}, \qquad f_{\mathbf{X}_{\text{test}}}(\mathbf{x}_{\text{test}}) = \int f_{\mathbf{X}_{\text{test}}|Y_j}(\mathbf{x}_{\text{test}}|y) f_{Y_j}(y) \mathrm{d}y \tag{3.5}$$

The *posterior* PDF $f_{Y_j|\mathbf{X}_{\text{test}}}(y|\mathbf{x}_{\text{test}})$ represents an updated distribution of Y_j obtained from new data $\mathbf{x}_{\text{test}}$. Assuming the independence of $\{X_i\}_{i=1}^n$, this yields

$$f_{Y_j|\mathbf{X}_{\text{test}}}(y|\mathbf{x}_{\text{test}}) = \frac{f_{Y_j}(y) \prod_{i=1}^n f_{X_{i,\text{test}}|Y_j}(\mathbf{x}_{i,\text{test}}|y)}{\int f_{Y_j}(y) \prod_{i=1}^n f_{X_{i,\text{test}}|Y_j}(\mathbf{x}_{i,\text{test}}|y) \, \mathrm{d}y}. \tag{3.6}$$

This procedure is repeated for all nodes Y_j ($j = 1, \ldots, m$) of the network.

This approach provides a probabilistic prediction of the system's response to any input $\mathbf{x}_{\text{test}}$. The predicted PDFs $f_{Y_j|\mathbf{X}_{\text{test}}}(y|\mathbf{x}_{\text{test}})$ $(j = 1,\ldots,m)$ are then used to reconstruct an updated network model. Within a mutual information (MI) framework, this is accomplished by following the steps described below.

3.2.2 Descriptive Module (Steps 4-6)

A typical MI-based network reconstruction procedure consists of two stages: building an MI matrix, and selecting a proper threshold. The MI matrix of a system is a square matrix, whose elements are MI of each pair of components. The threshold determines which interactions can be considered statistically insignificant and, hence, discarded. The interactions above this threshold are considered significant and used to reconstruct the network.

Calculation of Mutual Information Matrix. If X and Y are two random variables with random samples $\{x_1,\ldots,x_n\}$ and $\{y_1,\ldots,y_m\}$, MI between X and Y is computed as [87]

$$I(X,Y) = \sum_{j=1}^{m}\sum_{i=1}^{n} f_{XY}(x_i,y_j)\ln\frac{f_{XY}(x_i,y_j)}{f_X(x_i)f_Y(y_j)} \tag{3.7}$$

where $f_X(x_i)$ and $f_Y(y_j)$ are the marginal PDFs of X_i and Y_j, and $f_{XY}(x_i,y_j)$ denotes the joint PDF of X_i and Y_j. The MI matrix of the predicted network, $I(\mathbf{X}_{\text{test}},\mathbf{Y}_{\text{pred}})$, is evaluated from (3.7).

After evaluating the MI matrix, a proper boundary to determine whether or not a connection should be considered significant is determined by selecting a proper threshold. Interactions, whose MI is above this threshold, are considered statistically significant. The methodology to select the threshold is described below.

3.2.3 Threshold selection

Selecting a proper threshold is a non-trivial problem. A traditional approach is to perform permutations of measurements several times to calculate the MI for each permutation and average these MIs to select the largest MI as the threshold [13].

To select the appropriate threshold as a metric to identify the boundary of statistically weak and significant connections, we apply the large deviation theory. The latter can be traced back to Laplace and has formally defined by Varadhan [134]. Consider a set of N outcomes $\{x_1, \ldots, x_N\}$ of a random variable X, whose sample mean is $M_N = N^{-1} \sum_{i=1}^{N} x_i$. The large deviation theory states the probability $\mathbb{P}(M_N > x)$ decays exponentially as $N \to \infty$ at a rate depending on x [135], i.e.,

$$\mathbb{P}(M_N > x) \approx \mathrm{e}^{-NR(x)}, \qquad M_N = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{3.8}$$

where $R(x)$ is called a rate function. The large deviation theory was first used in 1937 in the insurance business [36], and introduced by Thomas M. Cover [20] to information science. It has since been used to provide a systematic methodology for the threshold selection in MI-based networks.

Bioinformatics applications of the large deviation theory, e.g., in ARACNE algorithm [83], compute the probability of an empirical value of mutual information I exceeding a value I_0, provided that its true value is $\bar{I} = 0$, as

$$\mathbb{P}(I > I_0 \mid \bar{I} = 0) \equiv p \sim e^{-cNI_0} \tag{3.9}$$

where c is a constant. Taking the natural logarithm of both sides yields

$$\ln p = a + bI_0 \tag{3.10}$$

where b is proportional to the sample size N. Therefore, for any dataset of sample size N and a desired p-value, the corresponding threshold is obtained by fitting a and b to the data. Using this methodology, we calculate a proper threshold for the MI matrix obtained above. The final predicted network is built by removing the statistically weak connections. In the next section, we demonstrate the applicability of the proposed algorithm by applying it in a systems biology case study, which deals with the signaling phosphoprotein-cytokines network in RAW 264.5 macrophages.

3.3 Application to Systems Biology: Phosphoprotein-Cytokine Signaling Network

Ever-increasing quantity of biological data in systems biology necessitates the development of various high-through methodologies for the measurement and analysis of data to extract the underlying mechanisms. We demonstrate the applicability of our approach in systems biology by predicting the performance and reconstructing the predicted network model of phosphoprotein-cytokine signaling networks in RAW 264.7 macrophage cells.

Macrophage cells produce a wide variety of regulatory substances, such as cytokines, to regulate both acute and chronic inflammations [89]. Cytokines, which are a group of proteins, bind to a target immune cell's receptor and hence, play a critical role in interacting with the immune cells as mediators [114, 40]. A complex signaling network transduced by the signaling

phosphoprotein network regulates cytokine releases [125, 107]. Understanding the underlying structure of this network can help identify regulatory modules that are responsible for the inflammatory responses during the activation of macrophage.

Following steps 1-3 in Section 3.2, we use the constructed cytokine-phosphoprotein network model [37] (for p values of 0.0001 and 0.005) to predict the behavior of cytokines for the given dataset. The predicted network model is then reconstructed by following steps 4-6. Since we use the test dataset, as the given dataset, for signaling phosphoproteins, we expect the predicted output values for released cytokines to show the same behavior as the original model. To evaluate the accuracy of the obtained network model, we compute the accuracy and F-measure of this model.

The original network model has also been constructed by using information-theoretic approaches and developed for two different p-values of $p = 0.0001$ and 0.005 for the threshold selection in Eq. (3.10). Alternatively, this network could have been obtained by any other methodology. We chose this network since it has a higher accuracy than other network models available in the literature [58, 140]. The dataset is borrowed from the AfCS data, which include data on the phosphoprotein and cytokine in RAW 264.7 macrophage under Toll and non-Toll conditions [44, 1]. The Toll dataset represents the data in which one of the ligands activates Toll-like receptors (TLRs). The non-Toll dataset refers to the data in which the ligands do not activate one or more of the TLRs [107]. The network models obtained from the Toll and non-Toll dataset are combined to provide the final reconstructed network.

We deal with two different datasets obtained from AfCS [1]: Train and test datasets. The train data are used to build an initial network model; we use it to estimate prior densities. The test data are used as input of the predicted model to predict the performance of the system.

Ideally, the test dataset has the same PDF as that of the train dataset. This would indicate that the predicted PDF of output under the test dataset as input should be the same as that of the train dataset. In this light, using the test dataset as input enables us to evaluate the strength of our methodology. In an ideal situation, where the PDFs of test data are the same as that of the train data, the two predicted models must be exactly the same.

Both the train and test dataset include 22 phosphoproteins and 7 cytokines. The phosphoproteins include: Signal Transducers and Activator of Transcription (STAT) 1α (STAT1α), STAT1β, STAT3, STAT5, Ribosomal Protein S6 (Rps6), Ribosomal S6 kinase (RSK), Glycogen Synthase Kinase (GSK) 3A (GSK3A), GSK3B, Extracellular-signal Regulated Kinases (ERK) 1 (ERK1), ERK2, cyclic Adenosine Monophosphate (cAMP), c-Jun N-terminal Kinases (JNK) long (JNK lg), JNK short (JNK sh), AKT, p40 Phagocyte Oxidase (p40Phox), Ezrin [Ezr]/Radixin [Rdx](Ezr/Rdx), Membrane-organizing Extension Spike Protein (Moesin or MSN), P38, Sma and Mad related proteins 2 (SMAD2), Nuclear Factor Kappa-light-chain-enhancer of activated B cells p65 (NF-$\kappa\beta$ p65), Protein Kinase C Delta (PKCD) and Protein kinase Cμ2 (PKC μ2. The cytokines include: $-$ Tumor Necrosis Factor alpha (TNFα); Interleukin-1α (IL-1α); Interleukin-6 (IL-6); Interleukin-10 (IL-10); Granulocyte Macrophage Colony Stimulating Factor (GM-CSF); Regulated on Activation, Normal T Expressed and Secreted (RANTES) and Macrophage Inflammatory Protein- 1alpha (MIP-1α.

Figure 3.1 shows the original phosphoprotein-cytokine network borrowed from [37] using the p-value of 0.0001. Figure 3.2 demonstrates this network using p-value of 0.005. In these figures, each node (circle) represents a protein or a cytokine and the edges, shown by solid lines, represent significant interactions between nodes. White nodes in both figures demonstrate the released cytokines. A pink color-coded node states that the phosphoprotein is connected to

the associated cytokines from the information obtained from the Toll-data and an orange node indicates that the phosphoprotein is considered to have statistically significant regulation effect on the cytokine from the information obtained from the non-Toll data. Blue nodes indicate that both the Toll and non-Toll data show regulatory effects of the phosphoprotein on the associated cytokine.

To find the marginal PDFs of these data, we first calculate the optimal kernel bandwidth that minimizes the risk functions. The bandwidths are selected by using Eq. (2.10), and then the marginal and joint PDFs are measured using Eqs. (3.1) and (3.2). Knowing both the input and output from the train dataset, we measure the joint PDFs of each interacting phosphoproteins and cytokines in the initial network model (Figs. 3.1 and 3.2).

To understand the robustness of our test data (how similar its PDFs are to those of the train dataset), we compare the PDFs of the inputs of both data sets computed with Eq. (3.1). Figure 3.3 shows a visual color-coded comparison of the PDFs of signaling phosphoproteins from the train and test dataset. The PDFs of the test dataset (blue) and train dataset (red) are not exactly the same, indicating that the predicted PDF of released cytokines and, hence, the final network model might not be exactly the same as the initial network due to inaccuracy in the input model.

Figures 3.1 and 3.2 show that, for each cytokine (white nodes), there is a set of related signaling phosphoproteins that are responsible for its release. For example, in Figure 3.1, IL-1α, which is represented by Y_j in Eq. (3.6), is regulated by a set of five regulating phophoproteins JNK sh,JNK lg,ERK1, ERK2, P38, and NF-$\kappa\beta$ p65, which are represented by $\mathbf{X} = \{X_i\}_{i=1}^{5}$ in Eq. (3.6). After finding the prior PDFs and likelihood functions of the interactions, we use these measurements to develop the predicted model using Eq. (3.6).

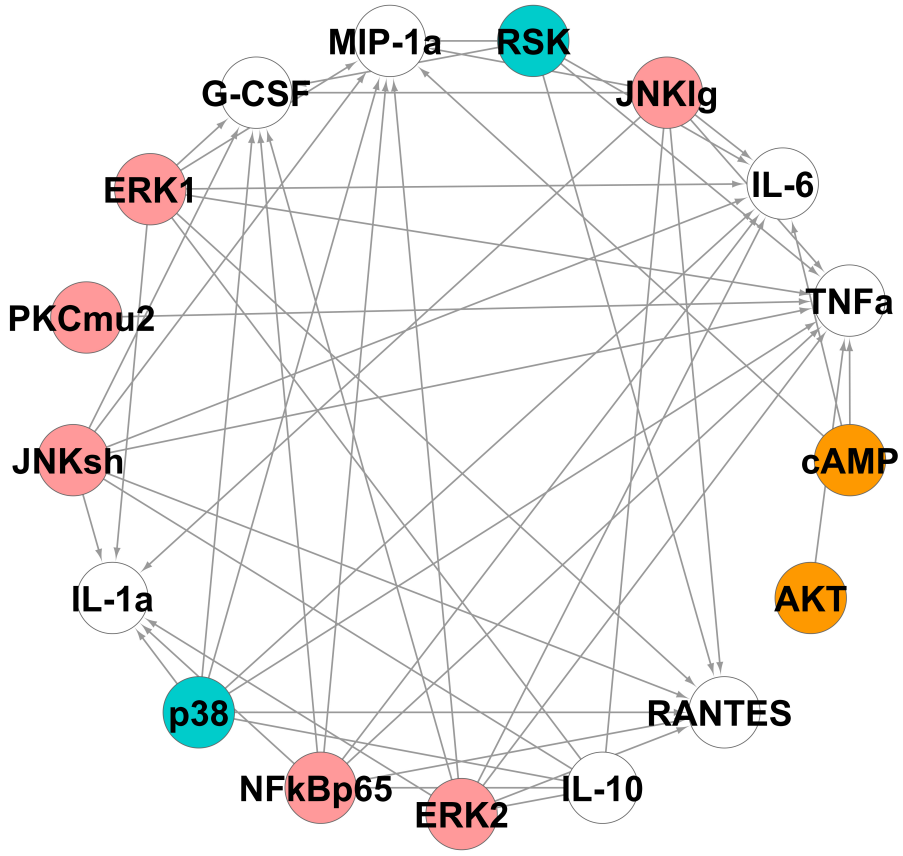

Figure 3.1: The initial network model of phosphoprotein-cytokine built with p-value of 0.0001.

Applying Eq. (3.6) provides the predicted PDFs that indicates the behavior of the released cytokines under the test dataset as input. Each of the seven rows in this matrix ($j = 1, \ldots, 7$) represents the predicted PDF of each of the released cytokines.

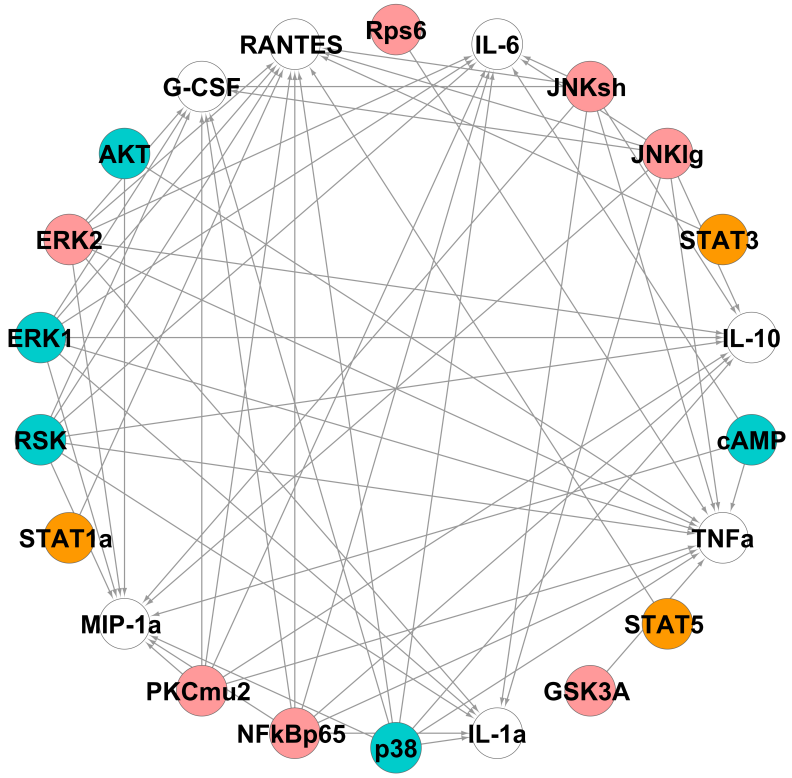

Figure 3.2: The initial network model of phosphoprotein-cytokine built with p-value of 0.005.

Figures 3.4 and 3.5 provide a comparison of the predicted PDFs of released cytokines obtained from the test dataset as input (blue) and the PDFs of cytokines directly measured from the train dataset (red) for p-values of 0.0001 and 0.005. As these two figures indicate, for the p-value of 0.0001 the predicted PDFs of released cytokines are more similar to the PDFs of cytokines in the train data (indicating a more reliable model).

The prediction of the PDFs of cytokines, using the test dataset as input, enables us to measure MI of interactions and to use these values to reconstruct the final predicted network models. Next, we obtain the MI matrix by using Eq. (3.7). The following step is to select of a

proper threshold by using Eqs. (3.9) and (3.10), a proper threshold for this MI matrix is selected for p-values of 0.0001 and 0.005. Interactions whose MI is below these thresholds are removed and the final network models are reconstructed for both p-values by considering the interactions with MI above the threshold.

Figure 3.6 shows the predicted network model from the initial network built with $p = 0.0001$. This model has a higher accuracy than the model built with $p = 0.005$. The number of false positive and false negatives is measured for the two constructed network models to determine the accuracies and F-measures. In the next section, a comparison of these values is presented. A non-probabilistic predictive methodology also will be applied to develop linear predictive models of phosphoproteins-cytokines signaling networks.

3.4 Discussion

The applicability of the proposed probabilistic methodology was demonstrated in Section 3.2. The network models from these predictions were captured for both p-values after predicting the probability density of outputs following the six steps described in Sections 3.2 and 3.3. In comparison with the predicted PDFs of output (Fig. 3.5) measured from the initial network built upon the p-value of 0.005 (Fig. 3.2), the predicted PDFs (Fig. 3.4) built using the initial network with p-value 0f 0.0001 (Fig. 3.1) provides a more accurate prediction of the performance of the system. As Figure 3.4 indicates, it makes a better estimation for the predicted PDFs of cytokines G-CSF, IL-1α, IL-10 and TNFα. For IL-6, MIP-1α and RANTES, both models contains some sort of inaccuracy in comparison with the measured PDF of the train dataset.

The inaccuracy in the predictions of the output PDFs is due to the inaccuracy in the input

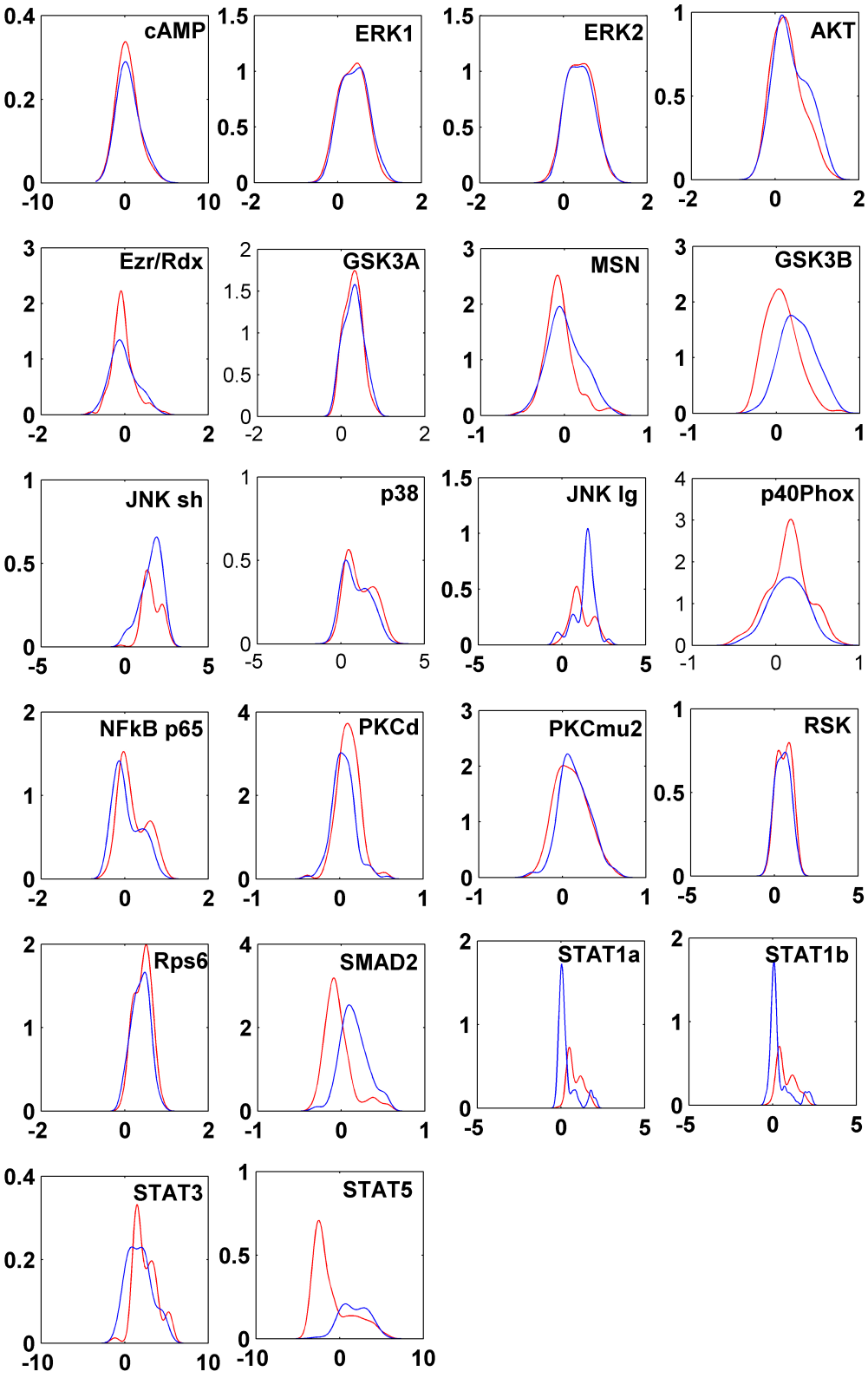

Figure 3.3: Comparison of phosphoproteins' probability densities measured from train dataset (red) and test dataset (blue)

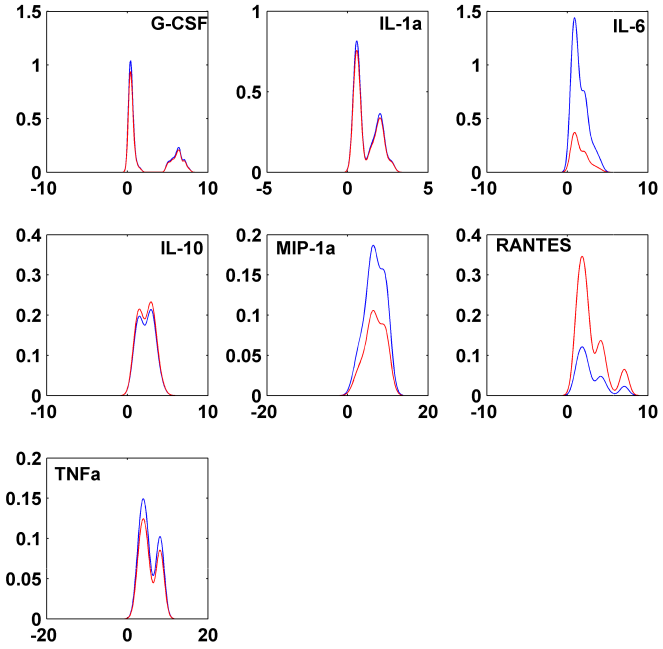

Figure 3.4: Comparison of probability densities of released cytokines measured from train dataset (red) and predicted dataset (blue) for p-value=0.0001

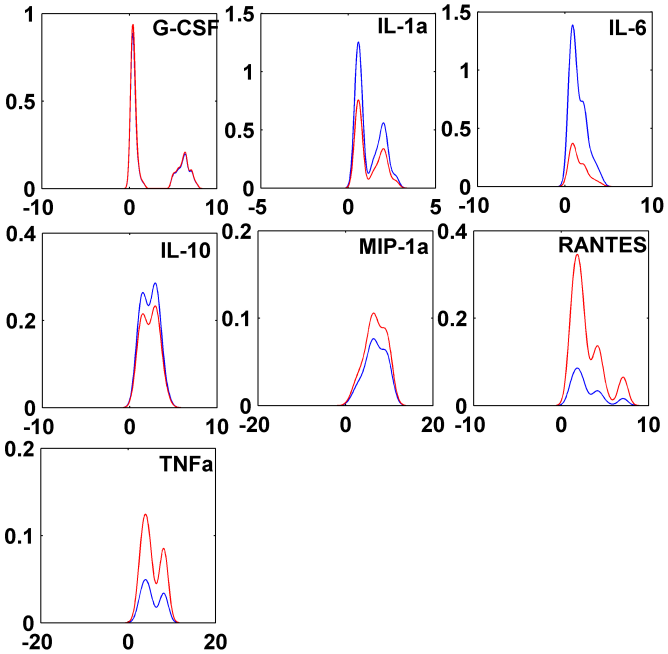

Figure 3.5: Comparison of probability densities of released cytokines measured from train dataset (red) and predicted dataset (blue) for p-value=0.005

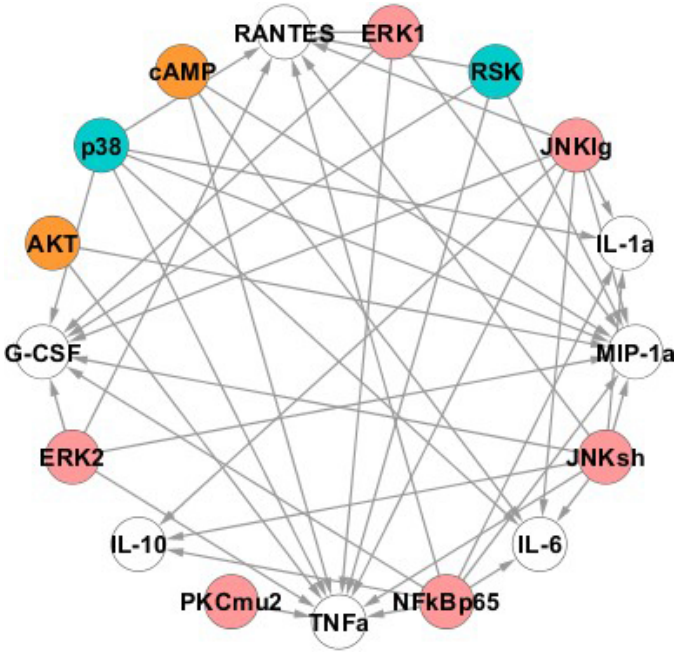

Figure 3.6: The predicted network model of signaling phosphoprotein-cytokine in RAW 264.5 macrophage cell for p-value=0.0001.

model. Intuitively, since the PDFs of the input test data are not exactly the same as those of the train data, there is some inaccuracy in the predicted PDFs. In addition, phosphoproteins may also have regulatory effects on each other, which were considered negligible in the initial network. Using the predicted PDFs in Figs. 3.4 and 3.5, the final network models for both p-values are obtained. The number of true positives and true negatives for both networks is measured and the accuracy of the models are obtained as

$$\text{Accuracy} = \frac{TP+TN}{n} \tag{3.11}$$

where TP denotes the number of true positives, TN denotes the number true negatives, and n represents the total number of significant interactions. The accuracies of the two models are then compared to quantitatively determine the relative accuracy of their predictions. To have a better

metric for understanding the accuracy of the proposed probabilistic approach, the accuracies are then compared with the accuracy of a predictive model constructed using another approach (Table 3.1).

We develop a linear predictive model from a non-probabilistic approach, called least square method (LSM). Consider a sample $X = (x_1, \ldots, x_n)$ that represents the set of significant inputs interacting with Y_j. A linear model of this network is

$$Y_j = \widehat{b}X + \varepsilon \tag{3.12}$$

where b is the coefficient matrix and ε represents the white noise. In the least square method, X is mean-centered and normalized by the standard deviation, and Y_j is mean-centered. Based on the least square method [11], the matrix b is obtained as

$$\widehat{b} = (X^T X)^{-1}(X^T Y_j). \tag{3.13}$$

If the coefficient matrix b is determined by using the training dataset, the values of output from the test data are predicted by substituting the X_{test} with

$$Y_{j,\text{pred}} = \widehat{b}X_{\text{test}}. \tag{3.14}$$

A commonly used metric to determine the accuracy of $Y_{j,\text{pred}}$ is called coefficient of determination [25],

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_{j,i} - y_{i,\text{pred}})^2}{\sum_{i=1}^{n}(y_{j,i} - \bar{y}_j)^2} \tag{3.15}$$

where n is the number of data points, and $\bar{y}_j$ is the mean of all data points in the selected Y_j.

The coefficient of determination lies between 0 and 1 and indicates the accuracy of the predicted values for the Y_j. A higher coefficient of determination represents a higher accuracy in the predicted data points.

Following Eq. (3.14), $Y_{j,\text{pred}}$ for all outputs (seven cytokines) are obtained by applying the LSM methodology. Using these values, coefficients of determinations for all output are measured. Figure 3.7 shows the scatter-platter of the predicted data points (from the test data) versus measured data points (from the train data). The coefficients of determination for all seven cytokines have low values ranging from 0.33 to 0.59. The low coefficients of determination indicate the inability of this approach to predict the performance of the system.

To be able to measure the accuracy of LSM in prediction, we reconstruct its predicted network model from the predicted values of output data points. Optimal kernel bandwidths of data points are obtained using Eq. (2.10) and KDE estimations of marginal and joint PDFs are measured. The mutual information is then calculated using Eq. (3.7) and the threshold is selected using Eq. (3.7).

Figure 3.8 shows the PDFs obtained from the predicted data points using Eq. (3.14). Comparing this figure with Figs. 3.4 and 3.5 provides a simple visual way to compare the ability of the methods to predict the PDF of released cytokines. To obtain a quantitative metric for this comparison, the network model of this predictive model is obtained and its accuracy is evaluated against the proposed probabilistic methodology for both p-values.

Table 3.1: Comparison of the accuracy of the predicted network models obtained using the proposed methodology (under two p-values) and least square method.

Method	p-Value	False Positive	False Negative	Accuracy
Proposed Methodology	0.0001	2	6	88%
Proposed Methodology	0.005	5	7	83%
Least Square Method	NA	8	11	61 %

As Table 3.1 indicates, in comparison with the least square model, the proposed approach significantly increases the accuracy of the prediction (about 27%) when the most robust model with p-value of 0.0001 is chosen. For $p = 0.005$ using the proposed methodology increases the accuracy 22%. To investigate the built models further, we also measure the precisions, recalls and hence, F-measures of the models.

Precision is referred to the probability that a positive prediction is correct and recall or sensitivity represents the probability that a prediction is correct [76]. Therefore, recall and precision are defined by

$$\text{Recall} = \frac{TP}{TP+FN}, \qquad \text{Precision} = \frac{TP}{TP+FP} \tag{3.16}$$

Measuring recall and precision enables us to predict F-measures of models. F-measure is the harmonic mean of precision of recall and can be used as a single measure of performance [106]

$$\text{F-measure} = \frac{(2 \times \text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}. \tag{3.17}$$

Table 3.2 lists a comparison of recalls, precisions and F-measures of the three constructed predictive models. Similar to accuracy table (Table 3.1), the F-measure of the predicted model with $p = 0.0001$ is the highest (17% higher than the least square method) and the least square method has the lowest F-measure. The F-measure of the predictive model obtained with $p = 0.005$ is 15% higher than the least square method.

The advantage of the proposed approach is its ability to develop predictive models of systems regardless of their applications, functional or parametric forms, and linearity of the system. The very high accuracies and F-measures of the predicted networks (listed in Table 3.1

and Table 3.2) captured by the proposed methodology confirms this claim. As expected, the most reliable predictive model is obtained when the p-value of 0.0001 is selected for the selection of the threshold.

Table 3.2: Comparison of the F-measure of the predicted network models obtained from using the proposed methodology (under two p-values) and least square method.

Method	p-Value	Recall	Precision	F-measure
Proposed Methodology	0.0001	0.88	10.95	92 %
Proposed Methodology	0.005	0.89	0.92	90 %
Least Square Method	NA	0.78	0.73	75 %

3.5 Conclusion

Reverse engineering is the process of discovering a complex system through analysis of its structure or performance. In recent years, several computational methods in different areas have attempted to systematize this process and develop methodologies capable of capturing the most accurate models. These methodologies vary depending on different assumptions that are made and various applications in different areas. Most of these approaches make some assumptions about the functional or parametric forms of systems or rely on the linearity of the system.

We attempted to develop a systematic framework, which is applicable in all kinds of systems and areas ranging from economy to systems biology. The proposed probabilistic methodology first predicts the PDFs of output for any given input dataset and then, constructs the predictive network models of systems using Bayesian and information-theoretic approaches. To show the applicability of this framework, we used it to build predictive models of phosphoprotein-cytokine signaling networks in RAW 264.7 macrophage cells. The values of accuracy and F-measure of this model were calculated and compared with values of a predictive model obtained

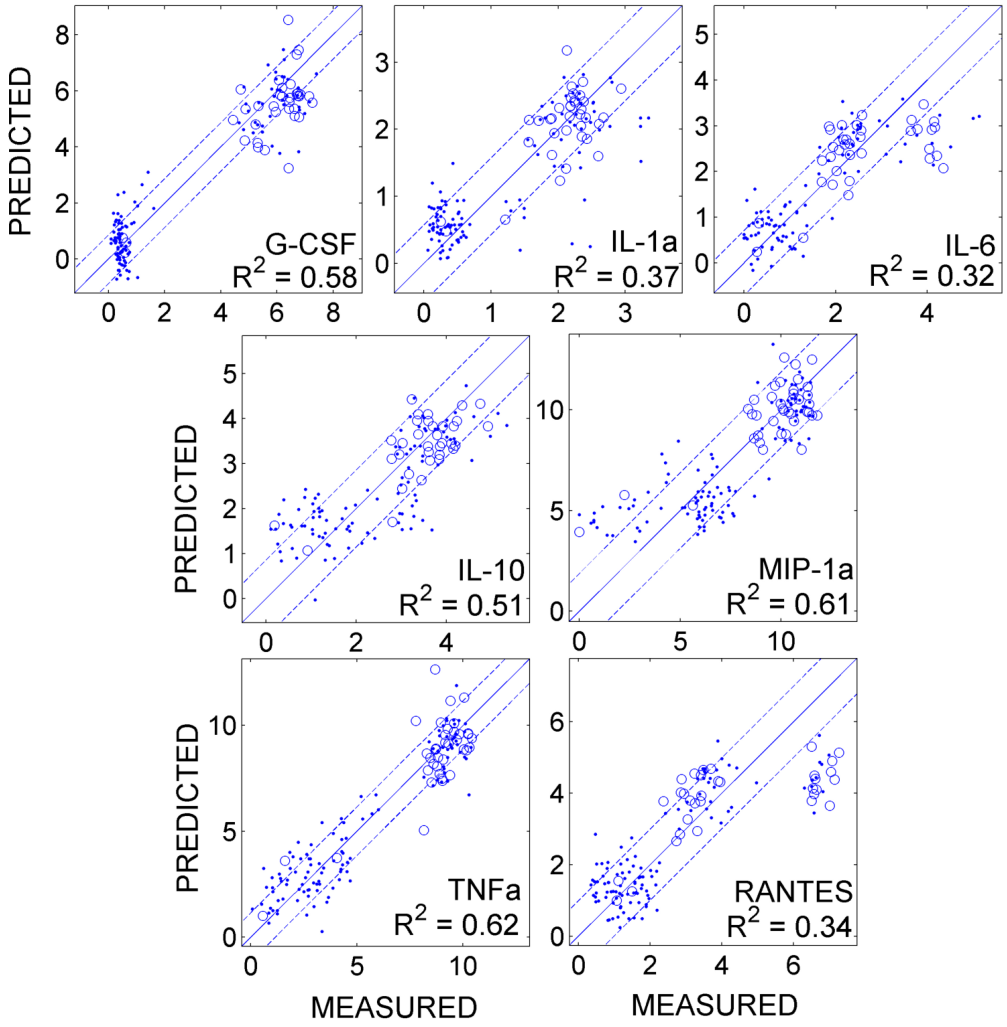

Figure 3.7: The predicted values (y-axis) for seven released cytokines vs. measured values (train data) for signaling phosphoprotein-cytokine in RAW 264.5 obtained by least square method.

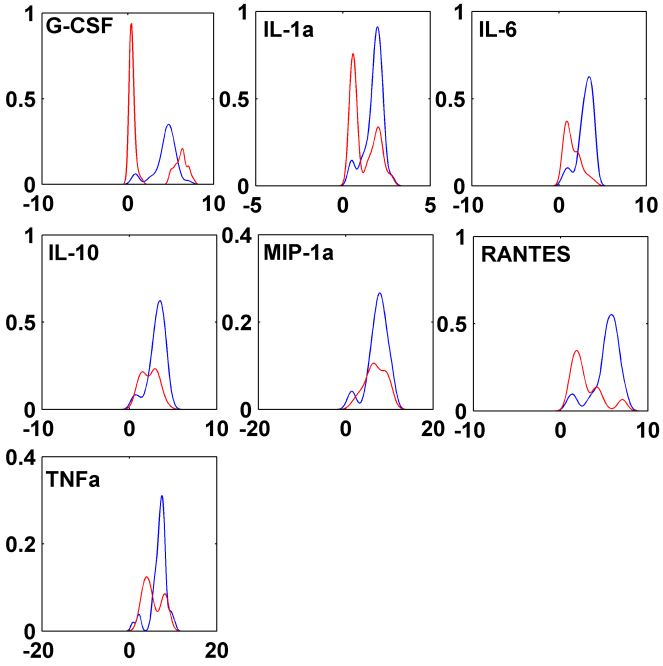

Figure 3.8: Comparison of probability densities of released cytokines measured from train dataset (red) and predicted densities of cytokines using least square method (blue)

by applying another methodology (least square method). Significantly high values for accuracy and F-measure of the predicted network models obtained by this methodology indicate the ability of the proposed approach to develop predictive models of all kinds of systems making no assumptions about functional and parametric forms and the linearity of the system.

3.6 Acknowledgements

This Chapter is a reprint of: F. Farhangmehr, and D. M. Tartakovsky. A Bayesian and Information Theoretic Approach for Predictive Modeling of Large-scale Networks. Submitted for review to Statistical Analysis and Data Mining Journal, 2014. The dissertation author was the primary investigator and author on this paper.

Chapter 4

An Information-Theoretic Algorithm to Data-driven Genetic Pathway Interaction Network Reconstruction of Dynamic Systems

Important Symbols used

f	Probability Density Function
I	Mutual Information
h	Bandwidth
f_h	Kernel Density Estimator Using Bandwidth h
$MISE$	Mean Integrated Squared Error
I_0	Threshold
p	p-Value
S	Subnetwork
e	Subnetwork Activity
$ETCA$	Earliest Time of Change in Activity
T_{up}	Up Threshold
T_{down}	Down Threshold
H_C	Copula Entropy
C	Copula Density
R	Rank of Observation
u	Pseudo Copula Sample

4.1 Introduction

Most (if not all) network reconstruction approaches have been developed to deal with static systems and data. Generalizations and extensions necessary to adapt them for analysis of time-dependent networks are challenging due to the complexity of biological systems and a large number of interactions among components. For example, time-delay is a common phenomenon in gene regulatory networks since the expression-level of a gene at a certain time may depend on the activation of another protein (gene-product) at a previous time [75]. Reconstructing such a complex network may require analysis of the behavior of all gene interactions during entire time-course.

We propose an information-theoretic algorithm to reconstruct networks of pathway interactions from microarray time-course data. The proposed methodology employs mutual information as a metric for capturing the causality of pathways from microarray data during time periods. This approach avoids unnecessary computations by grouping genes into pathways they

belong to and then identifying the pathways that may potentially regulate each other. Another feature of our approach is the use of copula entropy to detect the mutual information between pathways, which significantly reduces computational cost. Finally, we build the mutual information matrix from the maximum values of mutual information between each pair of flagged pathways over all possible time delays. This facilitates identification of the significant interactions among subnetworks to reconstruct the final pathway interaction network. We apply our approach to reconstruct the pathway interaction network of yeast-cell cycle using microarray time-course data by identifying the significant interactions of underlying pathways.

4.2 Background

4.2.1 Mutual-Information Networks

The amount of information about a random variable X that can be obtained by observing a random variable Y is often referred to as mutual information (MI) [119]. The higher the mutual information $I(X,Y)$, the higher the statistical dependence between X and Y. Consider two sets of measurements $\{x_1,\ldots,x_n\}$ and $\{y_1,\ldots,y_m\}$ of random variables X and Y, respectively. Using these data, we use the kernel density estimators (see previous Chapter) to compute the joint PDF $f_{XY}(x,y)$ and the marginal PDFs $f_X(x)$ and $f_Y(y)$. Then mutual information $I(X,Y)$ of X and Y is defined as

$$I(X,Y) = \sum_{j=1}^{m}\sum_{i=1}^{n} f_{XY}(x_i,y_j)\ln\frac{f_{XY}(x_i,y_j)}{f_X(x_i)f_Y(y_j)}. \tag{4.1}$$

A typical MI-based network reconstruction procedure consists of two stages: building the mutual information matrix, and selecting a proper threshold. The MI matrix of a system is a square matrix whose elements are mutual information of each pair of components. The thresh-

old determines which elements can be considered statistically negligible and be discarded. The interactions above this threshold are considered significant and used to reconstruct the network. Selecting a proper threshold is a nontrivial problem. A traditional approach is to perform permutations of measurements several times to calculate the distribution of the mutual information for each permutation and average these distributions to select the largest mutual information in the averaged permuted distribution as the threshold [13].

4.2.2 ARACNE

ARACNE (Algorithm for the Reconstruction of Accurate Cellular NEtworks) [83] relies on MI to reconstruct gene regulatory networks. It assigns to each pair of interactions a weight equal to their MI and eliminates the weak edges by measuring and applying a proper threshold. It calculates probability density function (PDF) of variables using non-parametric kernel density estimators (KDEs). ARACNE employs large deviation theory [20] to determine a proper threshold by approximating the PDF of mutual information by an exponential function. Through such an approximation, for any dataset with sample size N and a desired p-value, the corresponding threshold can be obtained. ARACNE also applies an information-theoretic property called the data processing inequality (DPI) to identify indirect connections. The DPI states that if X_i interacts with X_j through a random variable X_k then

$$I(X_i, X_j) < \min[I(X_i, X_k), I(X_j, X_k)]. \tag{4.2}$$

For each gene triplet, whose MI exceeds the threshold, ARACNE applies the DPI to identify and eliminate indirect connections. TimeDelay-ARACNE [5] allows ARACNE to capture time-delay gene regulatory networks. This framework first detects the time point of the initial changes

in the expression for all genes and then carries out network pruning and construction steps. We present an ARACNE-based algorithm to reconstruct networks from time-course data by detecting significant connections among pathways.

4.2.3 KEGG Pathways

A biological pathway is defined as a set of processes or biochemical transformations that accomplishes specific functions and leads to a change in a specific set of products [118]. Various types of biological pathways range from metabolic pathways to genetic and information processing pathways. The complexity of a biological network can be reduced by first decomposing it into pathways and then constructing a network of interactions from pathways [56]. Analysis of the structure of such networks facilitates understanding of key aspects of functionality of and causality in complex biological processes [127]. It enables one to model functional behavior of biological systems.

Our algorithm for reconstruction of interaction networks treats each pathway as a sub-network. Each sub-network represents a set of interconnected genes belonging to the same pathway with functional similarity. These genes have a higher probability of co-evolution than unrelated ones [31]. An information-theoretic approach is used to identify significant interactions between these sub-networks, enabling reconstruction of networks from time-course microarray data. The topology of underlying sub-networks is captured by assigning to each pathway a score that depends on the pathway's activity. This score, called sub-network activity, enables treatment of each pathway as a sub-network of a biological process and provides a metric to determine the activity of each sub-network, i.e., its role in the associated biological process.

Alternative ways to calculate sub-network activities include their treatment as an ag-

gregate expression profile of genes belonging to that sub-network [17, 45] or as the mean of expression levels of genes belonging to that sub-network [41]. To group genes into pathways, we use the information from KEGG (Kyoto Encyclopedia of Genes and Genomes) database [62]. The latter is a collection of manually curated pathway maps representing our knowledge of molecular interactions and reaction networks [61]. A gene may belong to more than one pathway and it may or may not actively participate in the same pathway(s) at any given time. Grouping genes into pathways enables us to measure the activity of pathways at each time point after normalizing the gene expression data. These scores are used to build mutual information network of pathways. In Section4.4 we use our algorithm to reconstruct a network of pathways using data from yeast cell cycle progression.

4.3 Methodology

Our algorithm adapts an ARACNE-like information-theoretic framework to network reconstruction from time-course microarray data. This is done as follows. First, the genes are grouped into pathways, and then a sub-network activity is assigned to each pathway at each time point. The pathways that may potentially regulate each other are identified by measuring the minimum amount of time necessary to observe a significant change in the activity of a sub-network. It is assumed that sub-network A can potentially regulate sub-network B only if the time necessary to observe a significant change in the activity of sub-network A does not exceed its counterpart for sub-network B. For two sub-networks that may potentially have regulatory effects, the algorithm identifies the maximum MI among all possible time points and then constructs the mutual information network by removing mutual information below the measured threshold and applying the DPI. The proposed algorithm consists of the following four steps.

1. Group genes into n sub-networks S_i ($i = 1, \ldots, n$) based on the pathways they belong to. As defined in [41] for static data, activity of sub-network S_i at time point t, $e^t_{s_i}$, is computed as the mean of expression of genes belonging to S_i at time t,

$$e^t_{s_i} = \frac{1}{m} \sum_{j=1}^{m} e^t_j. \tag{4.3}$$

   Here m is the number of genes in pathway S_i and e^t_j is the expression of jth gene in pathway S_i, after being normalized. Next, activities of all sub-networks during all time points are calculated. This step provides a good measure of the activity and significance of a sub-network at a given time point.

2. Find the minimum amount of time necessary for a significant change in the activity of a sub-network to occur. We refer to this time as Earliest Time of Change in Activity (ETCA) of sub-networks. This property is specified by using two up and down thresholds, $T_{\text{down}} = 1.18$ and $T_{\text{up}} = 0.85$. The values of these thresholds may vary (for example, TimeDelay-ARACNE [143] uses $T_{\text{down}} = 1.2$ and $T_{\text{up}} = 0.83$). ETCA is calculated as

$$ETCA(S_i) = \text{argmin}_t \left\{ \frac{e^0_{s_i}}{e^t_{s_i}} \geq T_{\text{up}} \quad \text{or} \quad \frac{e^t_{s_i}}{e^0_{s_i}} \leq T_{\text{down}} \right\}. \tag{4.4}$$

   Then, we postulate that sub-network S_i can regulate sub-network S_j only if

$$ETCA(S_i) \leq ETCA(S_j). \tag{4.5}$$

   Interactions between the sub-networks, which do not satisfy Eq. (4.5), are ignored, eliminating unnecessary calculations. This step speeds up the algorithm by identifying the

pathways that potentially have regulatory impacts.

3. For each pair of sub-networks detected by Eq. (4.5), obtain mutual information matrix considering all possible time delays and select the maximum mutual information. A challenge in extending methods of steady-state mutual information measurements to time-course data is to find the joint PDF of two variables. we rely on copula distributions of marginal PDFs of the two variables.

   A copula entropy, $H_C(X,Y)$, of two random variables X and Y is defined as [79]

$$H_C(X,Y) = -\int c(\mathbf{u}) \ln[c(\mathbf{u})]\, \mathrm{d}\mathbf{u} \tag{4.6}$$

   where $\mathbf{u} = [F_X, F_Y]$; F_X and F_Y are marginal densities of random variables X and Y given their random samples $\{x_1,\ldots,x_n\}$ and $\{y_1,\ldots,y_m\}$;

$$c(\mathbf{u}) = \frac{\partial^2 C(\mathbf{u})}{\partial u_1 \partial u_2} \tag{4.7}$$

   and $C(\mathbf{u})$ denotes the copula density of $\mathbf{u}$. In d dimensions, a d-dimensional empirical copula for random variables $\{x_1^i,...,x_d^i\}$ is given by [96]

$$C^d(\mathbf{u}) = \frac{1}{n}\sum_{j=1}^{n} \mathscr{I}(\widetilde{u}_1^i \le u_1,\ldots,\widetilde{u}_d^i \le u_d) \tag{4.8}$$

   where $\mathscr{I}$ is the indicator function, $\widetilde{u}_k^i = R_k^i/n$ is called a pseudo copula sample, R_k^i is the rank of observation x_k^i. In the case study presented in 4.4, we used two-dimensional empirical copulas to find mutual information between pathways. Finally, the copula entropy

$H_C(X,Y)$ in (6.1) is related to $I(X,Y)$ by [79]

$$I(X,Y) = -H_C(X,Y). \tag{4.9}$$

This approach avoids unnecessary computations made by other information-theoretic algorithms (such as [143]), which apply copula measurements to Eq. (4.1). To the best of our knowledge, the use of Eq. (6.4) is new in the context of information-theoretic reconstruction of biological networks from time-course data.

4. Select maximum values of mutual information for each pair of potentially dependent pathways over all possible time delays, and use these values to build a mutual information matrix. Similar to TimeDelay-ARACNE [143], this matrix is then used to reconstruct a network by employing the ARACNE framework. The weak connections are removed by measuring a proper threshold, and the DPI is used to eliminate indirect connections.

Figure 4.1 shows a flow diagram of our information-theoretic approach to reconstruction of biological networks from time-course data. We apply this algorithm to identify significant interactions of pathways and to construct a network from microarray data during one complete cycle of yeast cell-cycle.

4.4 Case Study

In this section, we demonstrate the ability of our algorithm to construct a network for Saccharomyces cerevisiae (yeast) cell cycle [123]. Yeast cell cycle is a series of events that takes place in a yeast cell leading to its replication. The microarray data consist of 7728 probes representing 5961 genes. There are fourteen time points, and samples are taken every seven

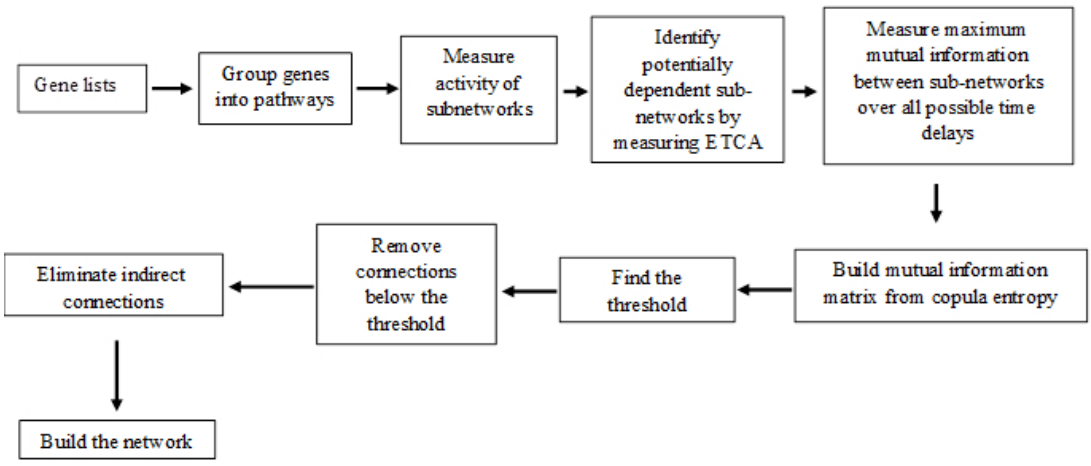

Figure 4.1: Flow chart of our algorithm to reconstruct pathway interaction networks of dynamic systems from time-course microarray data.

minutes as cells go through one complete cell cycle. After normalizing and removing genes with low expression, we group them into 94 pathways from information available in KEGG database. KEGG pathway maps for yeast cell cycles consist of four major pathways: metabolism, genetic information processing, environmental processing and cellular processes. The activity of each of 94 pathways is computed with Eq. (4.3), and the Earliest Time of Change in Activity (ETCA) in Eq. (4.4) is evaluated for each pathway. The pairs of pathways that do not satisfy Eq. (4.5) are considered non-dependent and removed from calculations. This avoids unnecessary computations for about 15% of pathway pairs. Phenylalanine metabolism (sce00360) does not satisfy Eq. (4.5) in terms of having potentially regulatory impact on other pathways. However, it satisfies Eq. (4.5) as a potential target of all other pathways. Table I lists pathways with lowest potential impact on other pathways as regulators.

Mutual information of potentially dependent pathways is obtained by computing the copula entropy in Eq. (6.4). Figure 4.2 shows the histogram of the mutual information distribution of potentially dependent pathways. The x-axis represents mutual information of all pathways

identified in step 2 and the y-axis refers to their frequencies. The solid line fits a kernel density function to the histogram. The dashed line indicates the measured threshold for $N = 1000$ and p-value $= 0.0001$. This threshold is used to detect significant interactions and discard indirect connections by applying DPI. The nodes that are not discarded are used to reconstruct the final network.

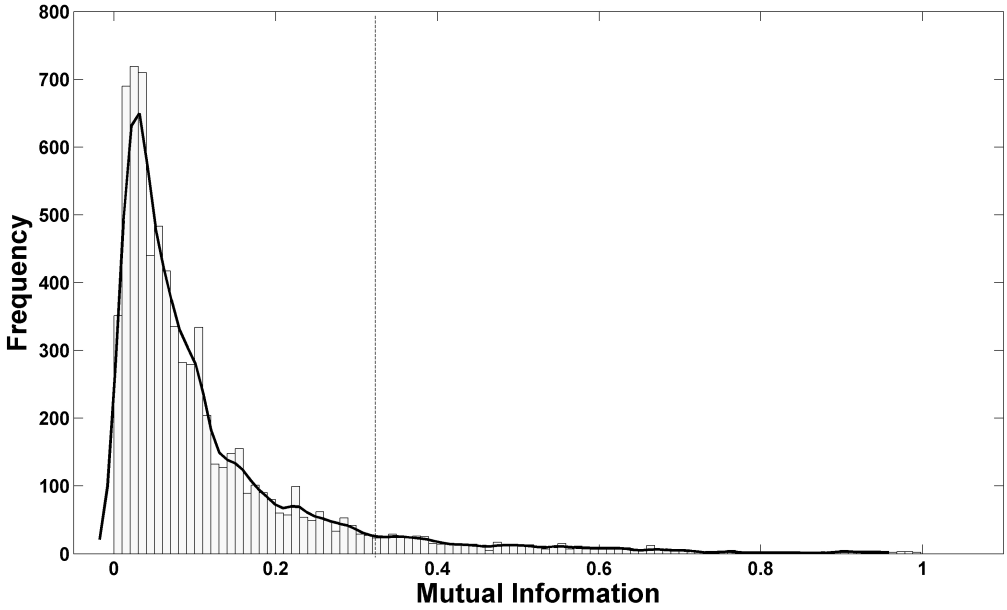

Figure 4.2: Histogram of maximum mutual information values. The dashed line indicated the selected threshold for p-value $= 0.0001$.

Figure 4.3 represents the reconstructed genetic interaction network obtained by our methodology. The nodes in this network (shown by rectangles) represent KEGG pathways (KEGG annotations have been used), and the edges (solid lines) indicate significant interactions among pathways. Arrows in this picture represent the direction of pathway interactions (since $I(X,Y) \neq I(Y,X)$ in this methodology; I is the maximum mutual information over all possible time delays).

N-Glycan biosynthesis (belonging to Glycan biosynthesis and metabolism) shows the

Table 4.1: KEGG pathways with lowest potential impact on other pathways as regulators

KEGG Annotation	name	TYPE
sce00360	Phenylalanine metabolism	Metabolism
sce00380	Tryptophan metabolism	Metabolism
sce00040	Pentose and glucuronate interconversions	Metabolism
sce0330	Arginine and proline metabolism	Metabolism

Data is from : [39].

highest activity among metabolic pathways; Ribosome (belonging to translation) has the highest activity among genetic information processing pathways; MAP Kinase signaling pathway (signal transduction) is the most active environmental information processing pathway and cell cycle represents the highest activity among the pathways related to cellular processes. Table II lists the most active pathway in each of the four major KEGG pathway maps in terms of the number of significant interactions with other pathways.

Table 4.2: KEGG pathways with largest interactions with other pathways

KEGG Annotation	name	TYPE
sce00510	N-Glycan biosynthesis	Metabolism
sce03010	Ribosome	Genetic Info.s Processing
sce04011	MAPK signaling pathway	Environmental Info. Processing
sce04011	Cell cycle-yeast	Cellular Processes

Data is from : [39].

4.5 Future Improvement

The proposed algorithm can be improved in several ways. As part of future work, we are investigating the use of conditional mutual information in addition to the DPI to detect indirect connections. The use of the DPI in mutual information-based network reconstruction (such as the ARACNE algorithm) may lead to identification of false-positive interactions for co-regulations. This is because the mutual information between two co-regulated components

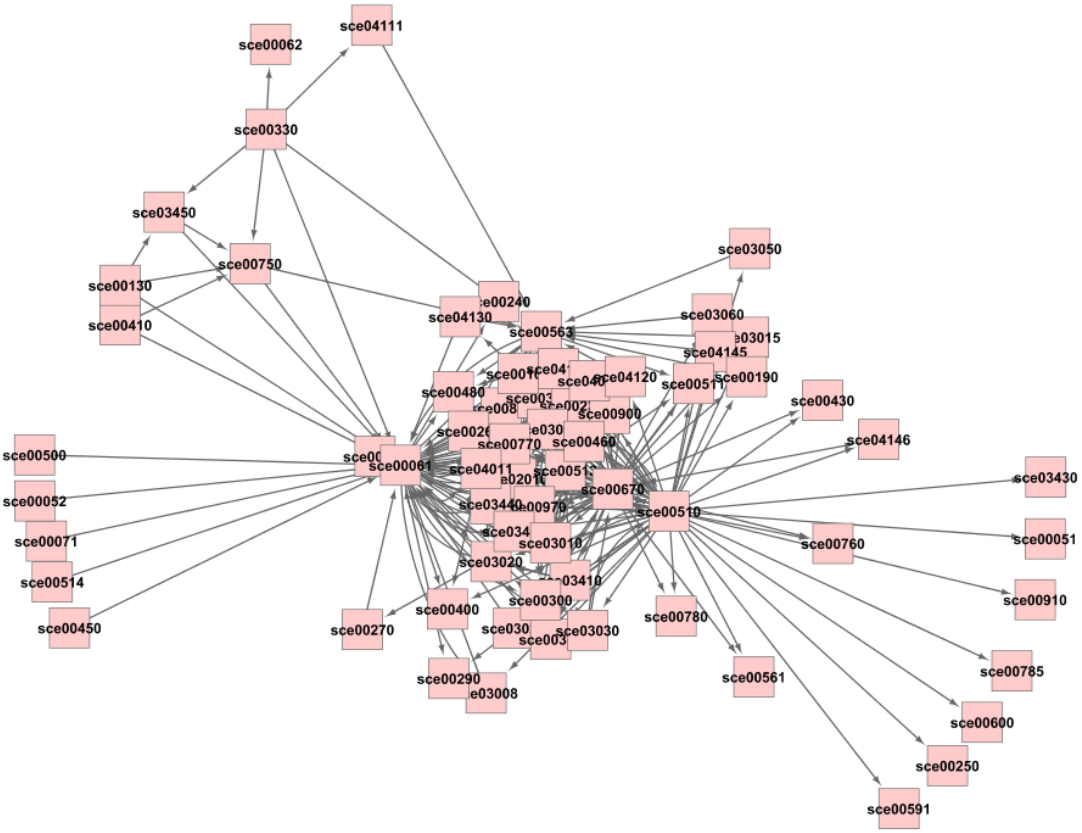

Figure 4.3: The reconstructed network for yeast cell cycle. Each rectangle represents a pathway (sub-network), and lines indicate significant connections between the associated pathways. The KEGG pathway annotations have been used as pathways identifiers.

may exceed the mutual information between regulating components, causing the DPI to identify false positive links between co-regulated components. We expect that using conditional mutual information in addition to mutual information will enable our algorithm to detect co-regulations. In addition, mutual information by itself is not a good metric in this case of interactive connections since the mutual information between components regulated by an XOR interaction will turn out to be negligible. This problem can also be overcome by computing conditional mutual information. Another limitation of the current mutual information-based methods is the lack of systematic methods to accurately ascertain a predictive model from the reconstructed network, which is necessary to model dynamic behavior of biological systems. Future work will involve development of such predictive models.

Delineation of pathway interactions is essential both to understand the underlying structure of interaction networks and to identify the pathways associated with genes not reported in KEGG database. We are investigating the functionality of poorly characterized genes, whose associated pathways are not yet identified by KEGG database.

4.6 Conclusions

We proposed an ARACNE-based algorithm for reconstruction of biological networks from time-course microarray data. This algorithm speeds up the computation by avoiding unnecessary calculations by grouping genes into their associated pathways and identifying potentially related pathways by measuring their activities. Measuring mutual information from copula entropy simplifies the process of building mutual information matrices of potentially related pathways considering all time delays. The applicability of this method has been demonstrated by developing a pathway interaction network using the yeast cell-cycle data. Our reconstructed

network shows a good agreement with the information available in the literature. In the future, we will improve our algorithm by developing probabilistic approaches to demonstrate predictive models of reconstructed dynamic networks and by supplementing DPI with conditional mutual information measurements.

4.7 Acknowledgments

This research was supported by the National Science Foundation (NSF) grant STC-0939370. The illuminating discussions with various participants in the NSF CSOI (Center for Science Of Information) program are gratefully acknowledged.

This Chapter is a reprint of F. Farhangmehr, D.M. Tartakovsky, P. Sadatmousavi, M.R. Maurya, and S. Subramaniam. An information-theoretic algorithm to data-driven genetic pathway interaction network reconstruction of dynamic systems. In Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference, pages 214-217, Dec 2013.

The dissertation author was the primary investigator and author on this paper.

Chapter 5

Statistical Approach to Reverse Engineering of Dynamic Networks from Time-Course Microarray Data

Important Symbols used

f	Probability Density Function
I	Mutual Information
h	Bandwidth
f_h	Kernel Density Estimator of f with a Bandwidth h
$MISE$	Mean Integrated Squared Error
I_0	Threshold
p	p-Value
$ETCA$	Earliest Time of Change in Activity
T_{up}	Up Threshold
T_{down}	Down Threshold
H_C	Copula Entropy
C	Copula Density
R	Rank of Observation
u	Pseudo Copula Sample

5.1 Introduction

Understanding a (typically nonlinear) relationship between inputs and outputs in dynamic systems is at the forefront of modern data science. This task is often formulated in terms of reconstruction and prediction of complex networks from time-series measurements. The structure and dynamics of networks can then be used to predict the behavior of the system and to detect new functions for its components [55].

Statistical approaches analyze input-output dependencies by using correlation measurements as a metric, without resorting to the linearity assumption. Bayesian networks (BNs) are a typical example of statistical methods. In essence, they are graphical models for describing causal interactions between variables [95, 43]. Nodes of a BN represent random variables, while its edges represent their conditional dependencies [50]. If used for network reconstruction, one chooses either a parametric or nonparametric form [91] of the conditional probability densities, and then decomposes the joint probability density into conditional probability densities among relevant nodes [105, 22].

Information-theoretic approaches provide an alternative to BNs for network reconstructions, which does not require one to specify functional and parametric forms of the random variables involved. Instead, they identify network models of systems by using mutual information (or uncertainty reduction) of interactions as a metric to establish statistical dependencies between interactions [20, 70, 52]. We propose a probabilistic algorithm that combines the strengths of Bayesian and mutual information networks. We use this algorithm to build a network model of gene interactions from microarray time-course data. We then extend this method to provide predictive models of the constructed network.

To demonstrate the applicability of the proposed methodology, we apply it to reverse-engineer gene interactions in *E. coli* from a time-course data set obtained from the GEO database. The final network identifies significant gene interactions among all possible interactions in *E. coli*, obtained over multiple time points. To decrease the computational cost, our algorithm detects pairs of genes that may potentially have regulatory effects on each others and performs the proposed computations only on the candidate pairs of genes. The algorithm computes a property, called maximum mutual information, over all possible time intervals for the genes satisfying the regulatory tests in the previous step. The use of empirical copula entropy to measure mutual information allows us to overcome challenges in measuring joint probabilities of time-dependent data and to further reduce the computational cost. The final network is constructed after selecting a proper threshold and removing statistically weak connections.

This strategy significantly decreases the computational complexity in constructing very large-scale networks from time-course data. It also facilitates the development of predictive models of the constructed network. The main idea of this predictive step is to capture the posterior probability densities of the nodes given measurements using the initially constructed net-

work.

The remainder of this section is devoted to a brief overview of the basic concepts of Bayesian networks and information-theoretic approaches to network reconstruction. In 5.2, we describe our methodology for reverse engineering of dynamic systems from time-course data. Section 5.3 contains an application of this method to *E. coli*.

5.1.1 Mutual-Information Networks

Mutual-information (MI) networks are networks reconstructed by using mutual information as a metric of statical significance of inter-node interactions. The process of developing MI networks consists of two majors steps: measurement of MI to compute a MI matrix, and selection of a proper threshold.

Measurement of Mutual Information. Consider two random variables X and Y, whose unknown (marginal and joint) probability density functions (PDFs) are to be inferred from their measurements $\{x_1,\ldots,x_n\}$ and $\{y_1,\ldots,y_n\}$. We use nonparametric kernel density estimators (KDEs) to accomplish this task. Specifically, a KDE represents $f_X(x)$, the PDF of random variable X, as [109]

$$f_X(x) = \frac{1}{nh^2\sqrt{2\pi}} \sum_{i=1}^{n} \exp\left[-\frac{(x-x_i)^2}{2h^2}\right]. \tag{5.1}$$

where the parameter h is called a kernel bandwidth. A KDE of $f_{XY}(x,y)$, the joint PDF of random variables X and Y, is

$$f_{XY}(x,y) = \frac{1}{2nh^2\pi} \sum_{i=1}^{n} \exp\left[-\frac{(x-x_i)^2+(y-y_i)^2}{2h^2}\right]. \tag{5.2}$$

Finally, MI between X and Y is defined as [87]

$$I(X,Y) = \sum_{j=1}^{n} \sum_{i=1}^{n} f_{XY}(x_i, y_j) \ln \left[\frac{f_{XY}(x_i, y_j)}{f_X(x_i) f_Y(y_j)} \right]. \tag{5.3}$$

Mutual information plays an important role in identifying indirect connections between multiple random variables. If a random variable X_i interacts with a random variable X_j through a random variable X_k then, according to a Data Processing Inequality (DPI) [83],

$$I(X_i, X_j) < \min\{I(X_i, X_k), I(X_j, X_k)\}. \tag{5.4}$$

For each triplet of random variables, whose mutual information exceeds the threshold, the DPI is applied to identify and eliminate indirect connections.

Computation of MI from time-series data can be prohibitively expensive, especially when data sets include high-dimensional vectors with different sample sizes. To speed up this computation, we will use the copula entropy. The latter is calculated by using Eqs (6.1)–(6.4) from Chapter 4.

Threshold Selection. The large deviation theory [20] states that the probability that an empirical value of mutual information I exceeds a given threshold I_0, provided that its true value is $\bar{I} = 0$, is an exponential function of the threshold. In other words, $\mathbb{P}[I \leq I_0] \equiv p \sim \exp(I_0)$. Taking the natural logarithm of both sides yields

$$\ln p = a + bI_0, \tag{5.5}$$

where p represents the p-value, and a and b are fitting parameters obtained by fitting this straight line to data. Equation (5.5) determines the threshold I_0 for any desired accuracy p (p-value). Any MI value that falls below this threshold represents a non-significant connection and is removed. This large-deviation-theory-based approximation is used in several biological network-reconstruction methods, e.g., ARACNE (Algorithm for the Reconstruction of Accurate Cellular NEtworks) [83].

To the best of our knowledge, a combination of large deviation theory for threshold selection and DPI for removing indirect connections was first introduced by ARACNE [83]. It is now widely used by various methodologies in computational systems biology.

5.1.2 Bayesian Networks

There are several equivalent definitions of a Bayesian network. According to the factorization theorem [103, 95], a network is considered a Bayesian network if the joint PDF of all of its nodes can be written as the product of the individual (marginal) PDFs of individual nodes, conditional on their parent variables. Suppose a network consists of n nodes $X = (X_1, \ldots, X_n)$, which are treated as random variables. It is called a Bayesian network if, for each X_v with a parent X_j,

$$f_X(x_1, \ldots, x_n) = \prod_{v=1}^{n} f_{X_v|X_j}(x_v \mid x_j) = \prod_{v=1}^{n} f_{X_v|X_j}(x_v \mid \text{parents}(x_v)) \tag{5.6}$$

The analysis presented below assumes conditional independence of the variables from any of their non-descendants, given the values of their parent variables.

5.2 Methodology

We propose a Reverse Engineering of GEnetic NeTworks (REGENT) algorithm for network reconstruction and modeling from time-course data. It consists of the following steps.

1. Find the minimum amount of time necessary for a significant change in the activity of a node (e.g., a gene) occurs. We refer to this time as Earliest Time of Change in Activity (ETCA). It is determined by using two up and down thresholds, T_{up} and T_{down}, respectively. The values of these thresholds may vary; for example, the thresholds used by TimeDelay-ARACNE [143] are $T_{\text{up}} = 0.83$ and $T_{\text{down}} = 1.2$. We set their values to $T_{\text{up}} = 0.85$ and $T_{\text{down}} = 1.18$. For the ith node (gene), the ETCA is calculated as

$$\text{ETCA}(G_i) = \text{argmin}_t \left\{ \frac{G_i^0}{G_i^t} \geq T_{\text{up}} \quad \text{or} \quad \frac{G_i^t}{G_i^0} \leq T_{\text{down}} \right\} \tag{5.7}$$

where G_i denotes the expression of ith gene, and After computing the ETCAs for all genes, we assume that the expression of the ith gene (G_i) can regulate the expression of the jth gene (G_j) only if

$$\text{ETCA}(G_i) \leq \text{ETCA}(G_j). \tag{5.8}$$

Interactions of gene pairs, which do not satisfy Eq. (6.7), are ignored. This assumption significantly reduces the unnecessary calculations by accounting only for the interactions of genes that may potentially regulate each others and ignoring the rest.

2. For each pair of nodes (genes) detected by Eq. (6.7), obtain MI over all possible time intervals and determine the Maximum Mutual Information (MMI). A challenge in measuring MI for time-course data is to find joint PDFs of variables with different sample sizes. To

decrease the complexity and increase the efficiency of computations, we determine the mutual information I from the empirical copula entropy H_C in Eq. (6.1) as [79]

$$I(X,Y) = -H_C(X,Y). \tag{5.9}$$

This procedure significantly reduces unnecessary computations needed to obtain joint PDFs and MI. This is in contrast to other methods (e.g., [143]), which apply Gaussian (or other types of) copula distribution to calculate joint PDFs and then use Eq. (5.3) to find MI. To the best of our knowledge, the use of empirical copula to measure MI in Eq. (5.9) is new in the context in data mining and systems biology.

3. Build a MMI matrix (MMIM) using maximum values of MI for potentially dependent genes over all possible time intervals. Eq. (5.5) is then applied to find a proper threshold for this matrix. Interactions, whose MI falls below this threshold, are removed and the DPI in Eq. (5.4) is applied to eliminate indirect connections. Finally, the network is constructed using the remaining interactions.

4. Predict posterior densities of target nodes given the measurements. Suppose that the network reconstruction procedure consisting of Steps 1-3 has identified a set X_B of m nodes $\{X_{B_1},\ldots,X_{B_m}\}$ that affect a node Y_j (Fig. 5.1). (In other words, $\{X_{B_1},\ldots,X_{B_m}\}$ are expressions of the regulator genes for a gene Y_j.). Using Bayes' theorem and with the assumption that X_B (s) are conditionally independent given Y_j, the posterior density of Y_j, given X_{test} (or measurements) is obtained by:

   Predict posterior densities of target nodes given the measurements. As Fig. 5.1 indicates, assume a probabilistic network reconstruction procedure (for example, the one described

in the previous step) has identified a network in which a node (random variable) Y_j is connected to (regulated by) a set of n nodes (random variables) $\mathbf{X} = \{X_i\}_{i=1}^n$. If this procedure has yielded a PDF $f_{Y_j}(y)$ of Y_j, this step seeks to probabilistically forecast the behavior of Y_j from a new set of k measurements $\{x_{i1}, \ldots, x_{ik}\}$ of each of the random variables X_i. We emply a Bayesian data assimilation technique [139] and treat $f_{Y_j}(y)$ as a *prior* PDF of Y_j. we treat a set of measurements $\mathbf{x}_{\text{test}}$ (an $n \times k$ matrix) as independent random variables $\mathbf{X}_{\text{test}}$ (an $n \times k$ matrix) that follow a Gaussian (conditioned on y) data model [139]

$$f_{X_{i,\text{test}}|Y_j}(\mathbf{x}_{i,\text{test}}|y) = \prod_{v=1}^{k} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(x_{iv}-y)^2}{2\sigma_i^2}\right], \qquad i = 1, \ldots, n \tag{5.10}$$

where σ_i^2 is the variance of X_i. This is a *likelihood function* to be used in Bayes' rule,

$$f_{Y_j|\mathbf{X}_{\text{test}}}(y|\mathbf{x}_{\text{test}}) = \frac{f_{\mathbf{X}_{\text{test}}|Y_j}(\mathbf{x}_{\text{test}}|y) f_{Y_j}(y)}{f_{\mathbf{X}_{\text{test}}}(\mathbf{x}_{\text{test}})}, \qquad f_{\mathbf{X}_{\text{test}}}(\mathbf{x}_{\text{test}}) = \int f_{\mathbf{X}_{\text{test}}|Y_j}(\mathbf{x}_{\text{test}}|y) f_{Y_j}(y) \mathrm{d}y \tag{5.11}$$

The *posterior* PDF $f_{Y_j|\mathbf{X}_{\text{test}}}(y|\mathbf{x}_{\text{test}})$ represents an updated distribution of Y_j obtained from new data $\mathbf{x}_{\text{test}}$. With the independence assumption of $\{X_i\}_{i=1}^n$, this yields

$$f_{Y_j|\mathbf{X}_{\text{test}}}(y|\mathbf{x}_{\text{test}}) = \frac{f_{Y_j}(y) \prod_{i=1}^n f_{X_{i,\text{test}}|Y_j}(\mathbf{x}_{i,\text{test}}|y)}{\int f_{Y_j}(y) \prod_{i=1}^n f_{X_{i,\text{test}}|Y_j}(\mathbf{x}_{i,\text{test}}|y)\, \mathrm{d}y}. \tag{5.12}$$

This procedure is repeated for all nodes of the network, Y_j ($j = 1, \ldots, m$). This step provides a probabilistic prediction of the system's response to any input $\mathbf{x}_{\text{test}}$. The predicted PDFs $f_{Y_j|\mathbf{X}_{\text{test}}}(y|\mathbf{x}_{\text{test}})$ ($j = 1, \ldots, m$) are then used in next step to construct a predictive

network.

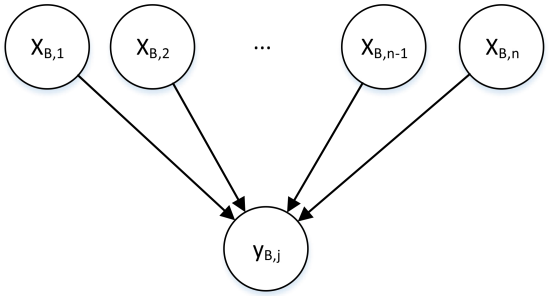

Figure 5.1: A schematic representation of a directed sub-network in which a set of nodes $\{X_{B_1}, \ldots, X_{B_n}\}$ are connected to a node Y_j, i.e., "genes $X_{B_1}, \ldots, X_{B_n}$ regulate gene Y_j".

5. Repeat steps 1 to 3 to reconstruct the predicted regulatory network by replacing the given network with the predicted network.

Figure 6.1 exhibits a flowchart of the algorithm that comprises Steps 1-5. This flowchart indicates that our algorithm can be extended to detect interactions among pathways in addition to genes. To capture pathways interactions, genes are grouped into their associated pathways first and then the mean of the genes expressions in each pathway is calculated for all possible time points. Finally, Steps 1-5 are applied to determine the pathway interactions networks. A gene might belong to more than one pathway and the mean of the gene expressions for one pathway might change at different time points. Developing genetic pathway networks would provide valuable information about the functional behavior of genes during time-course.

Some of the key features of the REGENT algorithm are listed below.

- Steps 1-3 reconstruct an initial network, while Steps 4 and 5 predict the network's performance and build a predictive network model.
  - Step 1 reduces the computational cost by identifying the pairs of nodes, which might potentially regulate each others.

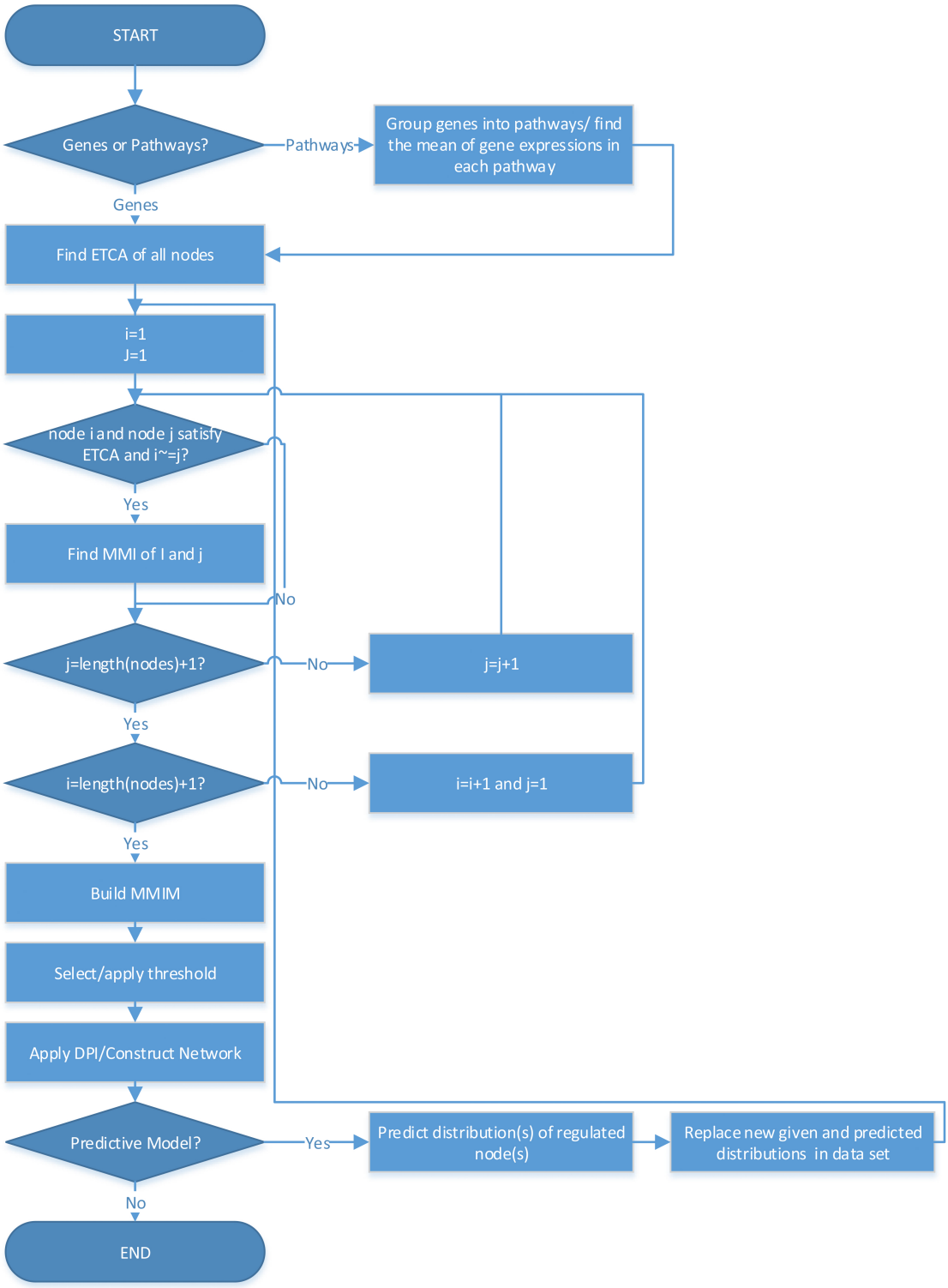

Figure 5.2: Flowchart of the proposed algorithm to data-driven network reconstruction and predictive modeling of time-course data.

  - Step 2 reduces the complexity of computing the MI for time-series data by using copula entropy.
  - Step 3 constructs a final network model by relying on the maximum MI matrix (MMIM) and selecting a proper threshold to detect significant interactions.
  - Step 4 uses this network to predict posterior densities of target nodes given new measurements.
  - Step 5 reconstructs a predictive network model by repeating Steps 1-3 with the new (predicted) data set.

- The predictive model (Steps 4 and 5) allows one to predict the behavior of nodes under new conditions, i.e., to predict posterior density of target nodes for any given data set containing information about its associated regulators.

- These predictive model can also be used to validate the initial network (built in Steps 1-3) if the same data set is used to predict posterior density of all nodes given measurements.

This algorithm reduces the complexity associated with data analysis for dynamic networks. It is worthwhile emphasizing that the constructed and predicted networks obtained in Steps 3 and 5 are built without making any assumptions about the linearity or functional forms of nodes (random variables). Instead, it assumes these variables (nodes) to be conditionally independent of variables from any of their non-descendants, given the values of their parent variables.

5.3 Case Study: E. coli Treated with Ampicillin

As an example, we analyze genes interactions of MG1655 cells in *Escherichia coli* (*E. coli*) at various time points up to 75 minutes following their treatment with 100 μg/ml of ampicillin. The ampicillin was added to cells grown in M9 media supplemented with glucose. The data used in our analysis are reported in NCBI's Gene Expression Omnibus (Edgar et al., 2002) [7] and are accessible through GEO Series accession number GSE4357. The data are part of a study tracking transcriptional responses of *E. coli* to over 30 chemical and physiological perturbations [116].

After normalizing and filtering the microarray data, we follow steps 1 to 3 to construct the network of gene interactions in *E. coli*. First, we identify and select interactions only if they satisfy Eq. (6.7). This significantly decreases the computational time and decreases unnecessary calculations by about 60%. Next, we use Eq. (5.9) to compute the maximum MI (MMI) of selected interactions during all possible time intervals. The result is a square matrix of maximum mutual information (MMIM). Figure 5.3 shows the histogram of the MMIs. The dashed line indicates the selected threshold. Only the interactions that exceed this threshold (right of the dashed line) are used for network construction.

The network is reconstructed in Step 3 by applying the DPI in Eq. (5.4) and selecting the proper threshold in Eq. (5.5) for p-value $p = 0.0001$. Figure 5.4 exhibits the resulting network of gene interactions of MG1655 cells in *E. coli*, which underwent treatment with 100 μg/ml of ampicillin. The nodes of this network represent the genes that are actively involved in the regulatory activity, and the lines indicate significant interactions.

Table 5.1 lists the most active genes in this network with respect to the number of interactions with other genes. Table 5.2 lists genes with highest activities as regulators. The third

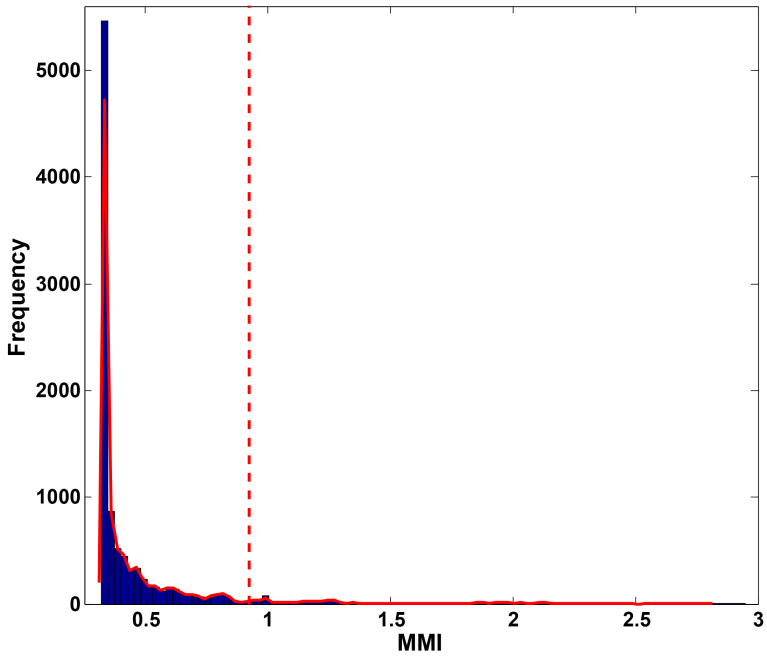

Figure 5.3: Histogram of Maximum Mutual Information (MMI) of interactions. The x-axis and y-axis in this figure indicate MMI and frequency. The red line shows the selected threshold.

columns of both tables provide definitions of the associated genes based on information available in KEGG database [61].

Table 5.1: Genes with highest activities with respect to the number of significant interactions with other genes

Gene Name	Interactions	Definition
flhE	97	proton seal during flagellar secretion [61]
yedY	97	membrane-anchored, periplasmic TMAO, DMSO reductase [61]
deaD	96	ATP-dependent RNA helicase[61]
mazG	90	nucleoside triphosphate pyrophosphohydrolase[61]
metL	89	Bifunctional aspartokinase/homoserine dehydrogenase 2 [61]
rhsE	87	pseudogene[61]
nlpA	85	cytoplasmic membrane lipoprotein-28[61]
fliY	51	cystine transporter subunit [61]
rnb	40	ribonuclease II[61]

After constructing the network in Step 3 (Figure 5.4), we use Steps 4 and 5 to develop a predictive model. This model is then used to validate the reconstructed network in Fig. 5.4 by using (5.12) and predicting the posterior densities of outputs given new measurements node by

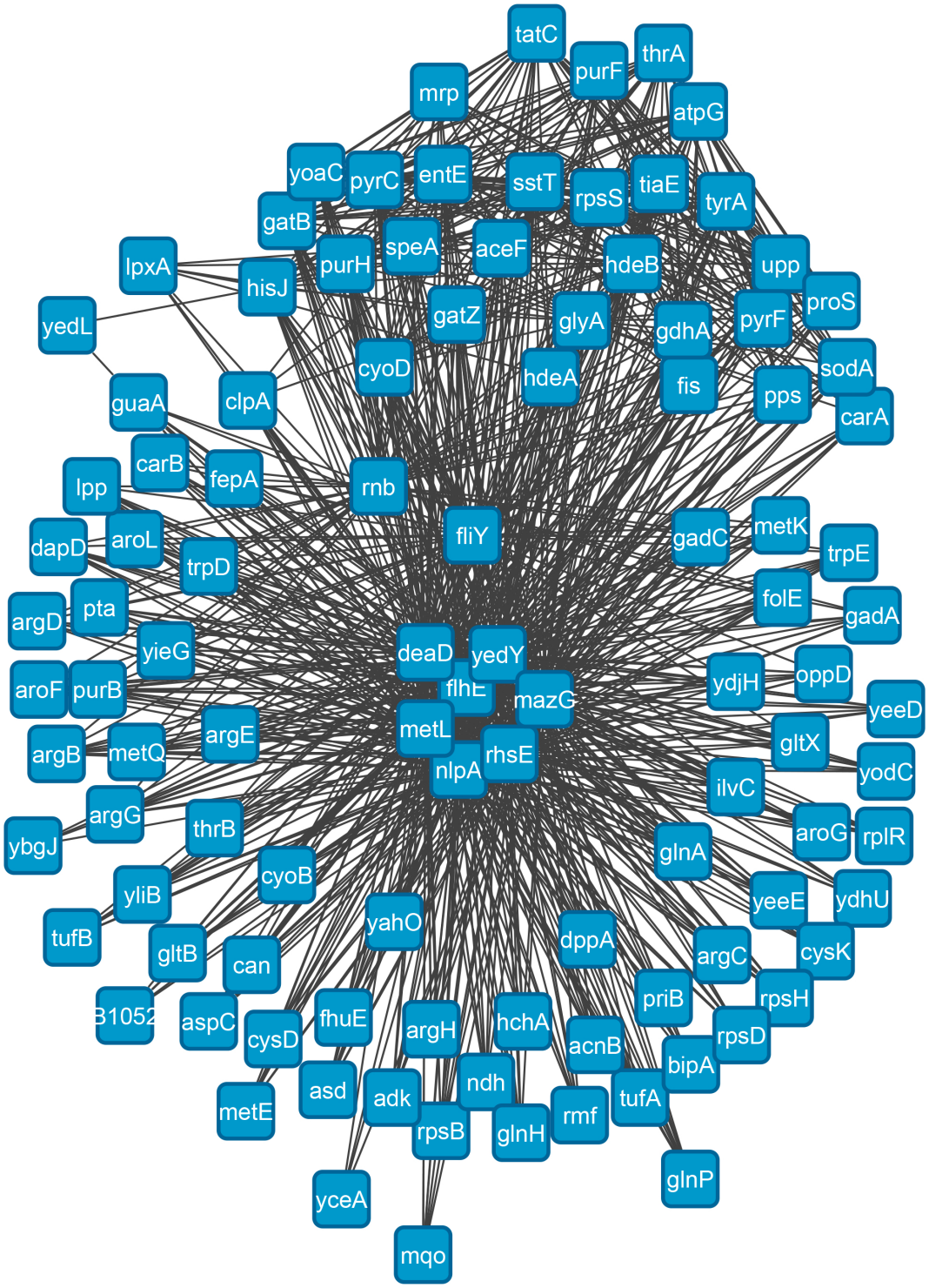

Figure 5.4: Network of gene interactions in *E. coli* subjected to treatment with 100 μg/ml ampicillin.

Table 5.2: Genes with highest activities with respect to the number of significant interactions with other genes as regulators

Gene Name	Interactions	Definition
hisJ	23	histidine/lysine/arginine/ornithine transporter subunit[61]
purH	22	fused IMP cyclohydrolase [61]
speA	22	biosynthetic arginine decarboxylase, PLP-binding [61]
upp	21	uracil phosphoribosyltransferase [61]
gatB	21	galactitol-specific enzyme IIB component of PTS [61]
yoaC	21	DUF1889 family protein [61]
gdhA	21	glutamate dehydrogenase, NADP-specific [61]
pyrF	21	orotidine-5′-phosphate decarboxylase[61]
pyrC	21	dihydro-orotase[61]
proS	21	prolyl-tRNA synthetase [61]

node. Finally, we build the predictive model by following Steps 1-3.

Figure 5.5 shows the predicted network model of *E. coli* following the above-mentioned steps. This network includes 16 false-positive and 21 false-negative interactions (see Table 5.3). The low numbers of false positives and false negatives and the high values of f-measure and accuracy demonstrate the accuracy of the initially network built in Step 3. To avoid the bias associated with using the same methodology to construct the final network, we repeat the prediction step using Eq. (5.3) to measure the maximum MI of candidate interactions. Although the predicted network model built using this method does not show any difference from Figure 5.5 obtained with REGENT, it takes about twice as long to compute as applying the empirical copula entropy.

Table 5.3: False Positive (FP), False Negative (FN), Precision, Recall, and f-measure of the predicted network

FP	16
FN	21
Precision	0.984
Recall	0.983
f-measure	0.983

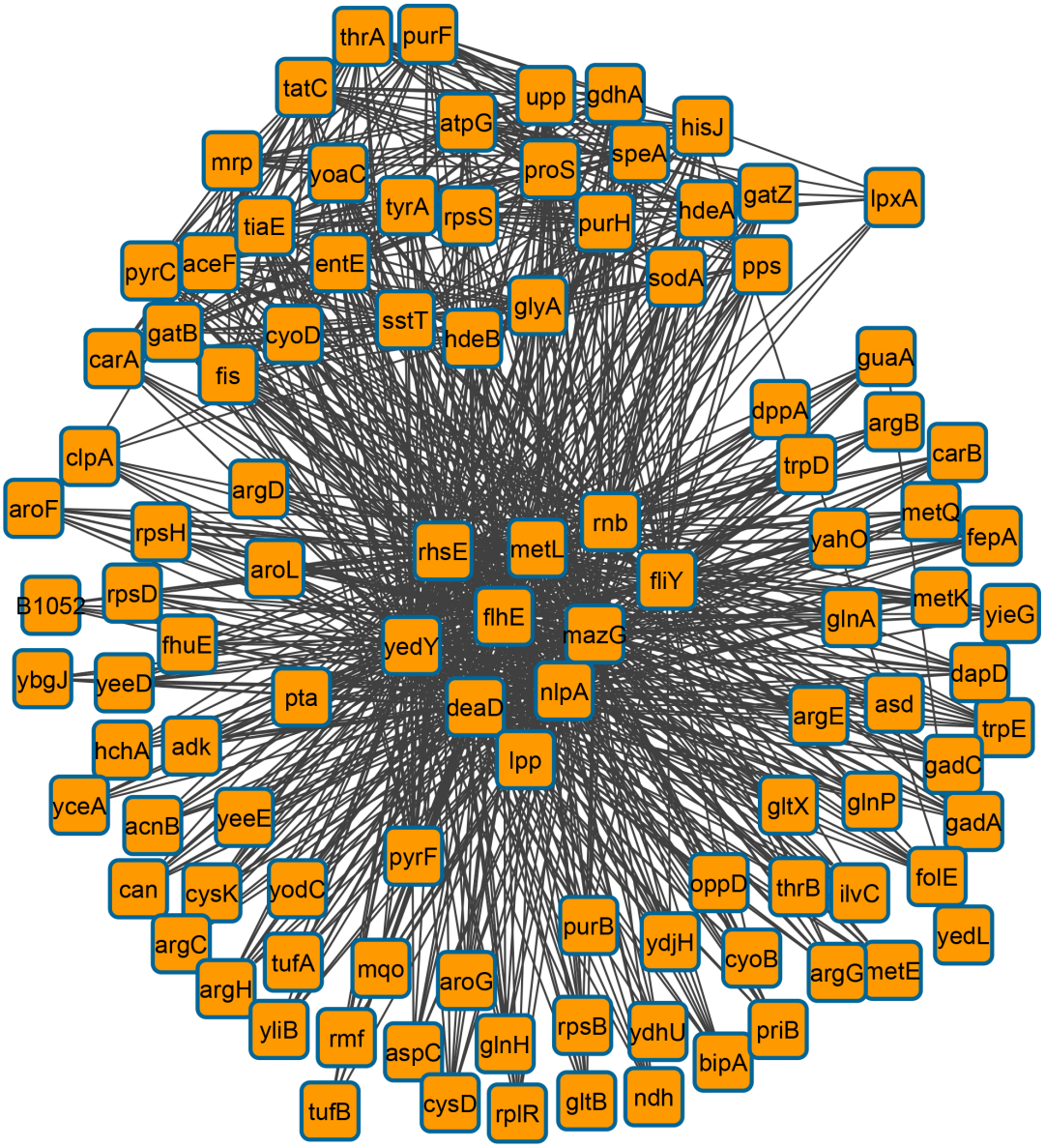

Figure 5.5: Predicted network model of *E. coli* following treatment with 100 μg/ml ampicillin.

5.4 Conclusions

We developed an algorithm, called REGENT, to accurately capture and predict the network of interactions in large-scale dynamic systems from time-course data sets. It is developed to address specific challenges in data analysis of dynamic networks. This methodology decreases the computational complexity by using a metric called ETCA and relying on empirical copula entropy to estimate mutual information. To the best of our knowledge, the latter procedure is new in the context of systems biology and data mining. The data-driven network reconstruction and predictive modeling framework make no assumptions about the linearity and functional or parametric forms of the variables. This algorithm also allows one to reverse-engineer networks of genetic pathways from time-course microarray datasets. Figure 6.1 presents the proposed methodology as a flowchart.

To demonstrate the applicability of REGENT algorithm to systems biology, we applied it to analyze data collected in *E. coli* subjected to an ampicillin treatment. Figure 5.4 shows the network of its genetic interactions and Figure 5.5 provides its predictive network model. Using the same densities as those estimated from the initial dataset for regulator nodes, predicting the posterior densities of regulated nodes given measurements and reconstructing the predictive model (Fig. 5.5), enabled us to validate our results (Fig. 5.4). Table 5.3 includes more details about this network model.

The results of this study enable one to identify hidden patterns behind dynamic biological systems, turning them into actionable information that can be used to design and develop mechanisms to engineer the system's performance. We expect REGENT algorithm to have significant applications in pharmaceutical industry, since it helps scientists and decision-makers to understand the underlying patterns and then predict the system's performance under designed

conditions.

5.5 Acknowledgments

This chapter is currently being prepared for submission for peer review and publication: Farhangmehr, F. and Tartakovsky, D. M. Statistical Approach to Reverse Engineering of Dynamic Networks from Time-Course Microarray Data.

The dissertation author is the primary investigator and author on this paper.

Chapter 6

Reverse Engineering of Gene Expression Data from Multiple Sclerosis Patients Undergoing Interferon-β Therapy

6.1 Introduction

Multiple sclerosis (MS) is an inflammatory disease, in which the insulating covers of nerve cells (myelin sheath) in the brain and spinal cord are damaged. Demyelination compromised the nervous system's ability to communicate and causes a wide range of signs and symptoms ranging from physical disabilities to cognitive dysfunctions [18]. The underlying mechanism of MS is not known but thought to be either destruction by the immune system or failure of the myelin-producing cells [93]. MS may have several forms, which are typically categorized as either relapsing-remitting forms (occurring in single attacks) or progressive forms (building up over time) [77]. Approximately 85% of MS patients are initially diagnosed with relapsing-remitting MS (RRMS) followed by improvement after relapses, compared to 10-15% with progressive forms or gradual worsening over time without periods of recovery [86]. While no current (as of 2014) treatment can change the course of progressive MS, nine disease-modifying treatments have been approved for RRMS [49, 82]. In 1993, Interferon-β (IFNβ) became the first FDA-approved drug for the treatment of RRMS [81]. IFNβ-1a is a cytokine in the interferon family, which is produced by mammalian cells; IFNβ-1b is produced in modified *E. coli* [92]. It is claimed that IFNβ decreases the rate of MS relapses about 18-38%. The mechanism by which IFNβ produces these therapeutic effects is not known; it is assumed to be associated with its immunomodulatory properties [81].

The main focus of this study is to identify novel gene patterns that may explain MS mechanisms and be used to develop new therapeutic targets by analyzing data from microarray experiments. Analyzing gene expression changes in microarray experiments may also reveal novel cellular functions associated with this disease. We use the REGENT (Reverse Engineering of GEnetic Networks from Time-series) [38] methodology presented in the previous Chapter

to construct models of genetic networks in MS patients undergoing IFNβ (both 1a and 1b) therapies. Then, we link these representative and predictive models of gene expression changes to phenotypes observed in these patients. We use the dataset that was obtained from NCBI's Gene Expression Omnibus [7] microarray experiment; it is accessible through GEO Series accession number GSE26104. In this experiment, gene expressions of RRMS patients treated with IFNβ-1b (Betaferon) or IFNβ-1a (Rebif) have been measured during a time period of 10 years [81].

The results of this study are important in discovering novel patterns that may elucidate MS mechanisms and explain how changes in gene expressions in MS patients under IFNβ therapy influence their health over time. Capturing these mechanisms not only provides new therapeutic solutions for MS patients but also allows medical professionals to make more-informed decisions.

6.2 Background: Reverse Engineering of GE Networks from Time-series (REGENT)

We use REGENT (Reverse Engineering of GEnetic Networks from Time-series) [38] approach to construct genetic networks. REGENT is a statistical approach, which relies of information-theoretic and Bayesian approaches. REGENT reconstructs a network in three main steps. First, it calculates mutual information between potentially related nodes by measuring a property called ETCA (Earliest Time in Change of Activity). Second, it builds MMIM (Maximum Mutual Information Matrix) whose elements are maximum values of mutual information between two nodes over all possible time intervals. Finally, a proper threshold is selected using large deviation theory and for a desired p-value the final network is constructed after removing

interactions whose mutual information falls bellow this threshold. This algorithm can also be used to develop interactions of genetic pathways. To construct pathway networks, genes are first grouped into pathways. Activity of a pathways is defined as the mean of expressions of genes belonging to the pathway.

6.2.1 Calculation of Mutual Information

To find mutual information, REGENT uses a quantity called copula entropy. A copula entropy, $H_C(X,Y)$, of two random variables X and Y is defined as [79]

$$H_C(X,Y) = -\int c(\mathbf{u}) \ln[c(\mathbf{u})]\, \mathrm{d}\mathbf{u} \tag{6.1}$$

where $\mathbf{u} = [F_X, F_Y]$; F_X and F_Y are marginal PDFs of variables X and Y given their samples $\{x_1,\ldots,x_n\}$ and $\{y_1,\ldots,y_m\}$;

$$c(\mathbf{u}) = \frac{\partial^2 C(\mathbf{u})}{\partial u_1 \partial u_2} \tag{6.2}$$

and $C(\mathbf{u})$ is the copula density of $\mathbf{u}$. A d-dimensional empirical copula for random variables $\{x_1^i, ..., x_d^i\}$ is obtained by [96]

$$C^d(\mathbf{u}) = \frac{1}{n} \sum_{j=1}^{n} \mathscr{I}(\widetilde{u}_1^i \leq u_1, \ldots, \widetilde{u}_d^i \leq u_d) \tag{6.3}$$

where $\mathscr{I}$ is the indicator function, $\widetilde{u}_k^i = R_k^i/n$ is called a pseudo copula sample, and R_k^i is the rank of observation x_k^i.

The copula entropy $H_C(X,Y)$ in (6.1) represents mutual information between X and Y, $I(X,Y)$, by [79]

$$I(X,Y) = -H_C(X,Y). \tag{6.4}$$

6.2.2 Threshold Selection

Based on Large Deviation theory, a proper threshold for any desired p-value p is obtained as [83]

$$\ln p = a + bI_0, \tag{6.5}$$

where a and b are fitted to the dataset and I_0 represents the selected threshold.

6.2.3 Earliest Time of Change in Activity (ETCA)

For a gene, G_i, ETCA (Earliest Time of Change in Activity) is defined as

$$\mathrm{ETCA}(G_i) = \mathrm{argmin}_t \left\{ \frac{G_i^0}{G_i^t} \geq T_{\mathrm{up}} \quad \text{or} \quad \frac{G_i^t}{G_i^0} \leq T_{\mathrm{down}} \right\} \tag{6.6}$$

where T_{up} and T_{down} are up and down thresholds selected by the user, and G_i denotes the gene expression of the ith gene in the microarray experiment. We assume that G_i can regulate the expression of the jth gene (G_j) only if

$$\mathrm{ETCA}(G_i) \leq \mathrm{ETCA}(G_j). \tag{6.7}$$

In this research, we use $T_{\mathrm{down}} = 1.18$ and $T_{\mathrm{up}} = 0.85$. Figure 6.1 represents the flowchart of REGENT algorithm.

6.3 Acknowledgments

This chapter is currently being prepared for submission for peer review and publication: Farhangmehr, F., Rangamani, P. and Tartakovsky, D. M. Reverse Engineering of Gene Interac-

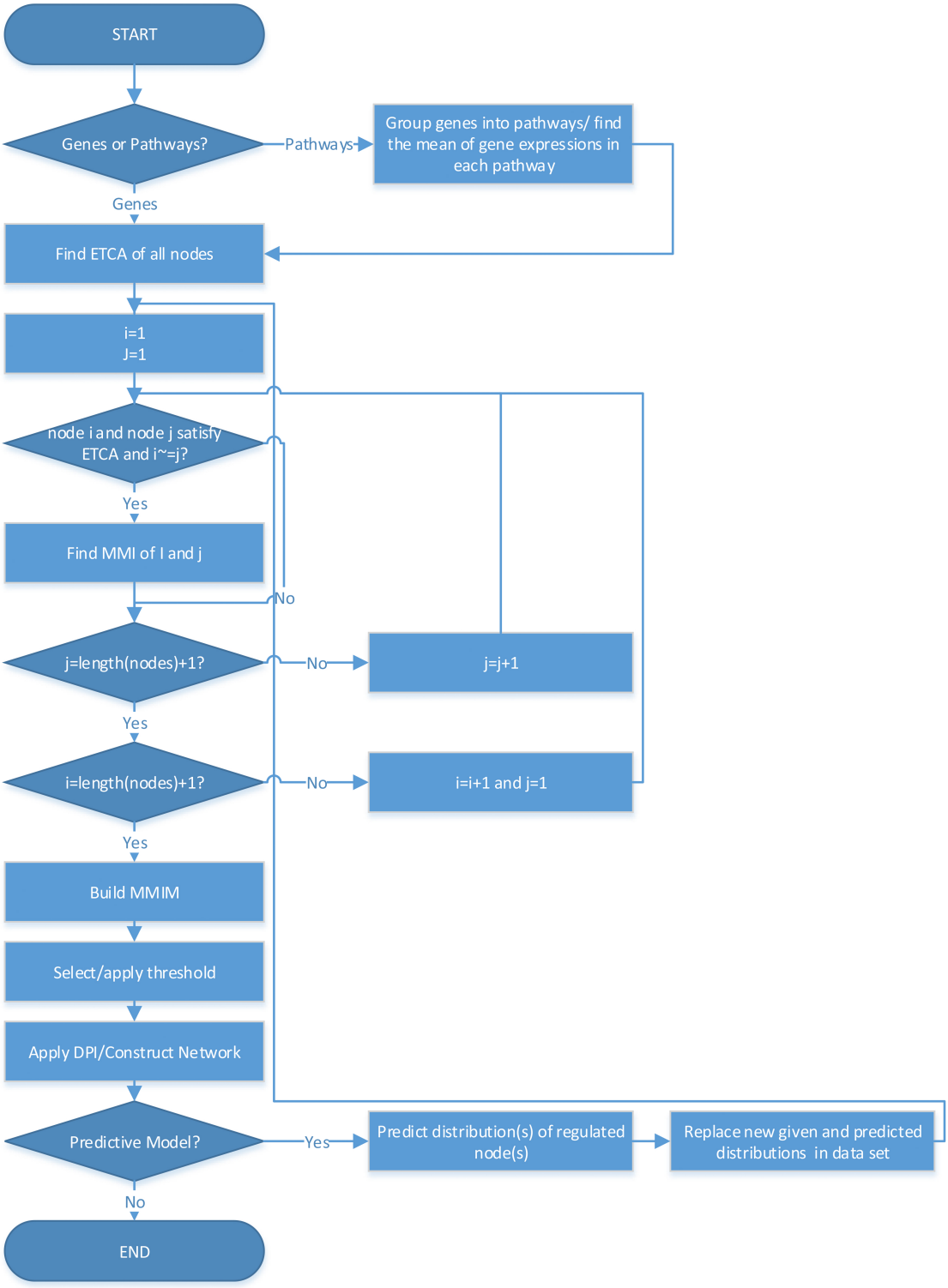

Figure 6.1: Flow .

tions for Multiple Sclerosis Patients Undergoing Interferon-β Therapy.

The dissertation author is the primary investigator and author on this paper.

Chapter 7

Summary and Conclusions

The conclusions reached in in each chapter of this dissertation are as follows.

7.1 Information-Theoretic Approach to Reconstruction of Complex Biological Networks

We used an information-theoretic approach to build an input-output model of regulatory phosphoprotein-cytokine signaling network in RAW 264.7 macrophages. The network constructed with this model was based on mutual information of interactions. The latter was computed using a nonparametric Kernel Estimation Estimator. An appropriate threshold was selected by applying the large deviation theory. According to this theory, the probability that an empirical value of mutual information exceeds a given threshold, provided that its true value is zero, is related exponentially to the threshold value. Interactions below this threshold were considered statistically insignificant and removed. The final network model was constructed using interactions above the threshold.

We developed a linear predictive model applying the Least Square Method to a test dataset. The low coefficient of determination values for this model indicate the nonlinear nature of the system. However, our methodology, which makes no assumptions about the linearity or a functional and parametric form of variables, successfully identified most of the known interactions. The network model captured by this method was validated against the published results. It also detected several new connections, such as Ribosomal S6 kinase for Tumor Necrosis Factor, which have not been previously detected by other methods. These novel regulatory components serve as testable hypotheses. Identifying the regulatory components for cytokines is critical for understanding the mechanisms that control their production and release in immune cells.

7.2 A Bayesian and Information-Theoretic Approach for Data-Driven Predictive Modeling and Reconstruction of Complex Networks

Combining Bayesian-net and information-theoretic approaches, we developed a framework for predictive modeling of complex systems. Our approach first predicts joint PDFs of output for a given input test dataset. Then, using these values, it constructs predictive network models. To demonstrate the applicability of this approach, we used it to construct predictive models of phosphoprotein-cytokine signaling networks in RAW 264.7 macrophage cells for two different networks initially constructed by using two different p-values of 0.005 and 0.001. The values for the accuracy and F-measure of these models were obtained and compared with values of a predictive model obtained by applying another methodology (the least square method used in our previous research).

High values of the accuracy and F-measure of the network models obtained with this

methodology indicate its potential uses for development of predictive models. The method is free of assumptions about the functional and parametric forms and the linearity of the systems.

7.3 An Information-Theoretic Algorithm for Data-Driven Reconstruction of a Genetic Pathway Interaction Network

We proposed an algorithm for construction of large-scale biological networks from time-course microarray data. This algorithm alleviates computational challenges in systems biology, which are related to data analysis for dynamic systems. It speeds up the computations by identifying potentially related nodes to avoid unnecessary calculations, and by using copula entropy as an estimator of maximum mutual information of interactions over all possible time intervals.

The applicability of this method was demonstrated by developing a pathway interaction network from a yeast cell-cycle microarray dataset. To construct this network, our algorithm first groups genes into their associated KEGG pathways and then captures the network model of interactions among pathways during one complete yeast cell-cycle. The proposed methodology not only decreased the complexity and computational cost of the calculations, but also showed a good agreement with the information available in the literature. The results of this study provide an important framework for capturing the functionality of genes by constructing time-course genetic pathway networks.

7.4 Statistical Approach to Reverse Engineering of Dynamic Networks from Time-Course Microarray Data

We developed an algorithm for data-driven network reconstruction and predictive modeling of large dynamic networks. This algorithm, called REGENT, first constructs a network of statistically significant interactions in large-scale systems and then, using a predictive methodology, develops a predictive network model of the constructed network for any given dataset. This approach captures the patterns of underlying systems and provides a new tool to engineer complex systems and make better-informed decisions.

We demonstrated the applicability of this algorithm in computational systems biology by applying it to *E. coli* that underwent treatment with Ampicillin. First, a network model of gene interactions in *E. coli* was constructed. Then, a predictive network was obtained using the initially constructed network. Using the same dataset as initial data for source nodes. and then predicting the probability distributions of target nodes, enabled us to computationally validate the initial network.

This algorithm detected significant gene interactions among millions of possible interactions in *E. coli* over all possible time intervals. It significantly decreased the computational time by detecting pairs of genes that may potentially have regulatory effects on each other. The main idea of the prediction step was to capture the joint PDFs of nodes using the initially built network and then, using a Bayesian-net assumption, decompose the joint PDFs into conditional PDFs among relevant nodes. This algorithm enabled us to capture and predict the performance of large-scale complex systems under various situations.

7.5 Reverse Engineering of Gene Expression Data from Multiple Sclerosis Patients Undergoing Interferon-β Therapy

This research focuses on the application of our REGENT algorithm for reverse engineering of time-course networks in a real-world problem in computational systems biology. We demonstrated how genes influence changing health over time. We applied our algorithm to a GEO data set in which the gene expressions of Multiple Sclerosis (MS) patients under Interferon-β have been measured over ten years. Capturing gene interactions and associated phenotypes provides significant information about health-related issues associated with MS patients undergoing this therapy.

Extracting the network of gene interactions for MS patients helps one to understand the impact of Interferon-β therapy on humans. It also provides potential diagnostic and therapeutic solutions for problems caused by changes in expressions of genes for MS patients. The result of this research will have significant applications in pharmaceutical industry for medical design and development. It will help medical designers and decision makers to make better-informed decisions. This research is under preparation.

7.6 Future Work

The various components of this dissertation contribute to increasing our understanding of hidden patterns and mechanisms underlying large-scale datasets. Our proposed methodologies provide solutions for reverse engineering of complex biological systems to capture these patterns and develop novel mechanisms to overcome multidisciplinary challenges associated with these complex and large-scale systems.

The future direction of our work will focus on applications of these methodologies to real-world problems. We will apply our algorithms to develop a comprehensive framework for providing potential diagnostic and therapeutic solutions for health-related problems caused by changes in genes expressions over time. We will link captured patterns of changes in expression of genes to their associated phenotypes to extract actionable information.

At the scale of technical improvement, we will use conditional mutual information in addition to mutual information. This will enable our algorithms to detect potential co-regulations. In addition, these algorithms will be more accurate (not necessarily more time efficient though), without having to rely on naive Bayesian-net logic of conditional independence assumptions. Finally, developing a user-friendly software for automating the network construction and predictive modeling processes would be a great addition to this research.

Appendix A

Development of a Linear Predictive Model

A.1 Least Square Method

To develop a predictive model using the reconstructed network, we build the following linear model between the significant inputs (X) and a chosen output (Y):

$$Y = \widehat{b}X + \varepsilon, \tag{A.1}$$

where ε represents white noise. Generally, one deals with one output at a time because the set of significant inputs differs for different outputs. Here X is mean-centered and normalized by the standard deviation and Y is mean-centered. The coefficient matrix b is estimated by least square

method [11] using training dataset:

$$\widehat{b} = (X^T X)^{-1}(X^T Y). \tag{A.2}$$

Once $\widehat{b}$ is estimated, the model can be tested on a test dataset. The test dataset generally has the same probability distribution as training dataset. Thus, given the input test data X_{test} (normalized by using the mean and standard deviation parameters obtained for the training set), the output test data Y_{test} (offset by the mean of Y) is predicted as

$$Y_{\text{test}}^{\text{pred}} = \widehat{b} X_{\text{test}}. \tag{A.3}$$

Two metrics used to measure the accuracy of the prediction are Root Mean Square Error (RSME) and coefficient of determination (R^2). They are calculated as [25]

$$\text{RMSE}_{\text{test}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_{\text{test},i} - Y_{\text{test},i}^{\text{pred}})^2} \tag{A.4}$$

and

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_{\text{test},i} - Y_{\text{test},i}^{\text{pred}})^2}{\sum_{i=1}^{n}(Y_{\text{test},i} - \bar{Y}_{\text{test}})^2} \tag{A.5}$$

where n is the number of data points, and $\bar{Y}_{\text{test}}$ is the mean value of the n data points for the chosen output. R^2 is a good quantitative metric indicating the quality of prediction by the linear model.

The scatter-plot in Figure A.1 illustrates the predictive power of the linear models made from the reconstructed network (Figure 2.4) for training (dots) and test (open circles) datasets on cytokine releases. Most of the training and test data points are inside within two root-mean-

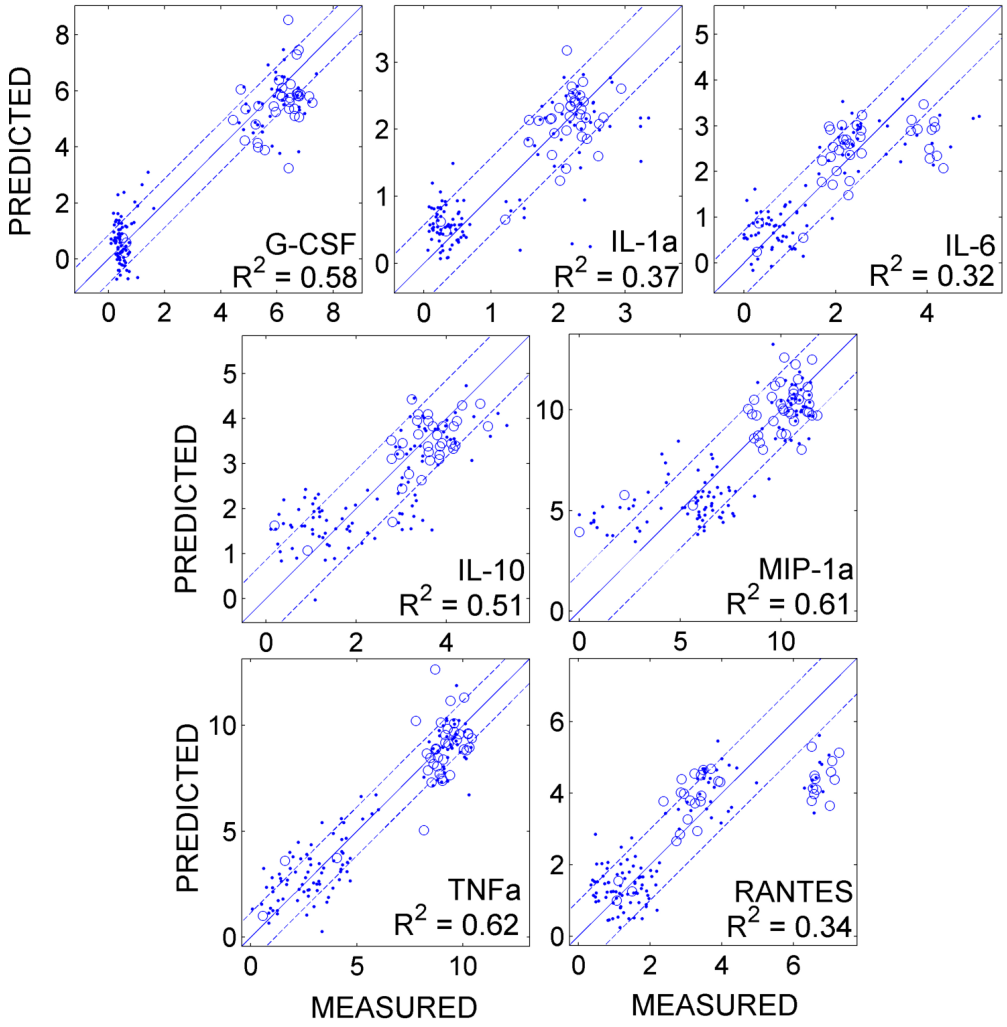

Figure A.1: Predicted (y-axis) vs. measured (x-axis) values of training (dots) and test (open circles) data for the seven cytokines.

squared errors of the training data. To provide a measure of the predictive quality of these linear models, we also computed the coefficient of determination R^2 for each cytokine as described in 2.4. The R^2 values range from 0.32 to 0.62. TNFα and MIP-1α yield the best fit ($R^2 >$ 0.6) and IL-6 and RANTES yields the lowest coefficients of determination. Although the linear model derived based on the significant components identified through the information theoretic approach is in a good agreement with the predictive models obtained with other methods, such as PCR [107] and PLS [140], the low coefficient of determination in these models demonstrate the highly non-linear nature of the phosphoprotein-cytokine signaling networks.

Bibliography

[1] The alliance for cellular signalling. http://signaling-gateway.org, 2002.

[2] MATLAB and statistics toolbox release 2012b. http://www.mathworks.com/help/stats/ksdensity.html, 2012.

[3] S. T. Ahmed, A. Mayer, J. D. Ji, and L. B. Ivashkiv. Inhibition of il-6 signaling by a p38-dependent pathway occurs in the absence of new protein synthesis. *J Leukoc Biol*, 72(1):154–62, 2002.

[4] R. Albert. Network inference, analysis, and modeling in systems biology. *Plant Cell*, 19(11):3327–3338, 2007.

[5] C. A. Amella, B. Sherry, D. H. Shepp, and H. Schmidtmayerova. Macrophage inflammatory protein 1alpha inhibits postentry steps of human immunodeficiency virus type 1 infection via suppression of intracellular cyclic AMP. *J. Virol.*, 79(9):5625–31, 2005.

[6] S. Bailly, M. Fay, N. Israel, and M. A. Gougerot-Pocidalo. The transcription factor ap-1 binds to the human interleukin 1 alpha promoter. *Eur Cytokine Netw*, 7(2):125–8, 1996.

[7] Tanya Barrett and Ron Edgar. [19] gene expression omnibus: Microarray data storage, submission, retrieval, and analysis. In Alan Kimmel and Brian Oliver, editors, *DNA Microarrays, Part B: Databases and Statistics*, volume 411 of *Methods in Enzymology*, pages 352 – 369. Academic Press, 2006.

[8] B. Beutler and A. Cerami. The biology of cachectin/tnf–a primary mediator of the host response. *Annu Rev Immunol*, 7:625–55, 1989.

[9] T. A. Bird, H. D. Schule, P. Delaney, P. de Roos, P. Sleath, S. K. Dower, and G. D. Virca. The interleukin-1-stimulated protein kinase that phosphorylates heat shock protein hsp27 is activated by map kinase. *FEBS Lett*, 338(1):31–6, 1994.

[10] J. Bondeson, K. A. Browne, F. M. Brennan, B. M. Foxwell, and M. Feldmann. Selective regulation of cytokine induction by adenoviral gene transfer of ikappabalpha into human macrophages: lipopolysaccharide-induced, but not zymosan-induced, proinflammatory cytokines are inhibited, but il-10 is nuclear factor-kappab independent. *J Immunol*, 162(5):2939–45, 1999.

[11] Otto Bretscher. *Linear algebra with applications*. Pearson Education, Boston, 5th edition, 2013.

[12] M. Brueckmann, U. Hoffmann, E. Dvortsak, S. Lang, J. J. Kaden, M. Borggrefe, and K. K. Haase. Drotrecogin alfa (activated) inhibits NF-kappa B activation and MIP-1-alpha release from isolated mononuclear cells of patients with severe sepsis. *Inflamm. Res.*, 53(10):528–33, 2004.

[13] A. J. Butte and I. S. Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput*, pages 418–29, 2000.

[14] A. J. Butte, P. Tamayo, D. Slonim, T. R. Golub, and I. S. Kohane. Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. *Proc Natl Acad Sci U S A*, 97(22):12182–6, 2000.

[15] K. H. Cho, S. M. Choo, S. H. Jung, J. R. Kim, H. S. Choi, and J. Kim. Reverse engineering of gene regulatory networks. *IET Syst Biol*, 1(3):149–63, 2007.

[16] Y. H. Cho, C. H. Lee, and S. G. Kim. Potentiation of lipopolysaccharide-inducible cyclooxygenase 2 expression by C2-ceramide via c-Jun N-terminal kinase-mediated activation of CCAAT/enhancer binding protein beta in macrophages. *Mol. Pharmacol.*, 63(3):512–23, 2003.

[17] H. Y. Chuang, E. Lee, Y. T. Liu, D. Lee, and T. Ideker. Network-based classification of breast cancer metastasis. *Mol Syst Biol*, 3:140, 2007.

[18] A. Compston and A. Coles. Multiple sclerosis. *Lancet*, 372(9648):1502–17, 2008.

[19] C. Cosentino, W. Curatola, F. Montefusco, M. Bansal, D. di Bernardo, and F. Amato. Linear matrix inequalities approach to reconstruction of biological networks. *IET Syst Biol*, 1(3):164–73, 2007.

[20] T. M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience, Hoboken, N.J., 2nd edition, 2006.

[21] X. Dai, K. Sayama, M. Tohyama, Y. Shirakata, L. Yang, S. Hirakawa, S. Tokumaru, and K. Hashimoto. The nf-kappab, p38 mapk and stat1 pathways differentially regulate the dsrna-mediated innate immune responses of epidermal keratinocytes. *Int Immunol*, 20(7):901–9, 2008.

[22] F. d'AlchÃl' Buc and V. SchÃd'chter. Modeling of biological networks, 17-20 May 2005.

[23] G. Davatelis, P. Tekamp-Olson, S. D. Wolpe, K. Hermsen, C. Luedke, C. Gallegos, D. Coit, J. Merryweather, and A. Cerami. Cloning and characterization of a cDNA for murine macrophage inflammatory protein (MIP), a novel monokine with inflammatory and chemokinetic properties. *J. Exp. Med.*, 167(6):1939–44, 1988.

[24] J. L. E. Dean, S. J. Sarsfield, E. Tsounakou, and J. Saklatvala. p38 mitogen-activated protein kinase stabilizes mRNAs that contain cyclooxygenase-2 and tumor necrosis factor AU-rich elements by inhibiting deadenylation. *J. Biol. Chem.*, 278(41):39470–39476, 2003.

[25] Morris H. DeGroot and Mark J. Schervish. *Probability and statistics*. Addison-Wesley, Boston, 4th edition, 2012.

[26] U. Dendorfer, P. Oettgen, and T. A. Libermann. Multiple regulatory elements in the interleukin-6 gene mediate induction by prostaglandins, cyclic amp, and lipopolysaccharide. *Mol Cell Biol*, 14(7):4443–54, 1994.

[27] B. Diaz and G. Lopez-Berestein. A distinct element involved in lipopolysaccharide activation of the tumor necrosis factor-alpha promoter in monocytes. *J Interferon Cytokine Res*, 20(8):741–8, 2000.

[28] C. A. Dinarello. The biological properties of interleukin-1. *Eur Cytokine Netw*, 5(6):517–31, 1994.

[29] Z. G. Dobreva, L. D. Miteva, and S. A. Stanilova. The inhibition of jnk and p38 mapks downregulates il-10 and differentially affects c-jun gene expression in human monocytes. *Immunopharmacol Immunotoxicol*, 31(2):195–201, 2009.

[30] Laurent El Ghaoui and Silviu-Iulian Niculescu. *Advances in linear matrix inequality methods in control*. Advances in design and control. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1999.

[31] F. Enault, K. Suhre, O. Poirot, C. Abergel, and J. M. Claverie. Phydbac2: improved inference of gene function using interactive phylogenomic profiling and chromosomal location analysis. *Nucleic Acids Res*, 32(Web Server issue):W336–9, 2004.

[32] Vincenzo Esposito Vinzi and SpringerLink. *Handbook of partial least squares concepts, methods and applications*. Springer, Berlin, 2010.

[33] M. Eto, A. Kouroedov, F. Cosentino, and T. F. Luscher. Glycogen synthase kinase-3 mediates endothelial cell activation by tumor necrosis factor-alpha. *Circulation*, 112(9):1316–22, 2005.

[34] L. Faggioli, C. Costanzo, M. Donadelli, and M. Palmieri. Activation of the interleukin-6 promoter by a dominant negative mutant of c-jun. *Biochimica Et Biophysica Acta-Molecular Cell Research*, 1692(1):17–24, 2004.

[35] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol*, 5(1):e8, 2007.

[36] Xiequan Fan, Ion Grama, and Quansheng Liu. Cramér large deviation expansions for martingales under Bernstein's condition. *Stochastic Processes and their Applications*, 123:3919–3942, Jun 2013.

[37] F. Farhangmehr, M. R. Maurya, D. M. Tartakovsky, and S. Subramaniam. Information theoretic approach to complex biological network reconstruction: application to cytokine release in raw 264.7 macrophages. *BMC Syst Biol*, 8:77, 2014.

[38] F. Farhangmehr and Tartakovsky. Probabilistic algorithm to data analytic of dynamic networks from time-course microarray data. *In Preparation*, 2014.

[39] F. Farhangmehr, D.M. Tartakovsky, P. Sadatmousavi, M.R. Maurya, and S. Subramaniam. An information-theoretic algorithm to data-driven genetic pathway interaction network reconstruction of dynamic systems. In *Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on*, pages 214–217, Dec 2013.

[40] C. A. Feghali and T. M. Wright. Cytokines in acute and chronic inflammation. *Front Biosci*, 2:d12–26, 1997.

[41] K. Fortney, M. Kotlyar, and I. Jurisica. Inferring the functions of longevity genes with modular subnetwork biomarkers of caenorhabditis elegans aging. *Genome Biol*, 11(2):R13, 2010.

[42] M. Frodin and S. Gammeltoft. Role and regulation of 90 kda ribosomal s6 kinase (rsk) in signal transduction. *Mol Cell Endocrinol*, 151(1-2):65–77, 1999.

[43] D. Geiger, T. Verma, and J. Pearl. Identifying independence in bayesian networks. *Networks*, 20(5):507–534, 1990.

[44] A. G. Gilman, M. I. Simon, H. R. Bourne, B. A. Harris, R. Long, E. M. Ross, J. T. Stull, R. Taussig, H. R. Bourne, A. P. Arkin, M. H. Cobb, J. G. Cyster, P. N. Devreotes, J. E. Ferrell, D. Fruman, M. Gold, A. Weiss, J. T. Stull, M. J. Berridge, L. C. Cantley, W. A. Catterall, S. R. Coughlin, E. N. Olson, T. F. Smith, J. S. Brugge, D. Botstein, J. E. Dixon, T. Hunter, R. J. Lefkowitz, A. J. Pawson, P. W. Sternberg, H. Varmus, S. Subramaniam, R. S. Sinkovits, J. Li, D. Mock, Y. Ning, B. Saunders, P. C. Sternweis, D. Hilgemann, R. H. Scheuermann, D. DeCamp, R. Hsueh, K. M. Lin, Y. Ni, W. E. Seaman, P. C. Simpson, T. D. O'Connell, T. Roach, M. I. Simon, S. Choi, P. Eversole-Cire, I. Fraser, M. C. Mumby, Y. Zhao, D. Brekken, H. Shu, T. Meyer, G. Chandy, W. D. Heo, J. Liou, N. O'Rourke, M. Verghese, S. M. Mumby, H. Han, H. A. Brown, J. S. Forrester, P. Ivanova, S. B. Milne, P. J. Casey, T. K. Harden, A. P. Arkin, J. Doyle, M. L. Gray, T. Meyer, S. Michnick, M. A. Schmidt, M. Toner, R. Y. Tsien, M. Natarajan, R. Ranganathan, and G. R. Sambrano. Overview of the alliance for cellular signaling. *Nature*, 420(6916):703–6, 2002.

[45] L. Guo, G. Wei, J. Zhu, W. Liao, W. J. Leonard, K. Zhao, and W. Paul. IL-1 family members and STAT activators induce cytokine production by Th2, Th17, and Th1 cells. *Proc. Natl. Acad. Sci. U.S.A.*, 106(32):13463–13468, 2009.

[46] C. Guzzo, N. F. C. Mat, and K. Gee. Interleukin-27 induces a stat1/3- and nf-kappa b-dependent proinflammatory cytokine profile in human monocytes. (vol 285, pg 24404, 2010). *Journal of Biological Chemistry*, 287(11):8661–8661, 2012.

[47] R. H. Hartley. Transmission of information. *Bell System Technical Journal*, 7:535–563, 1928.

[48] S. Hashimoto, K. Matsumoto, Y. Gon, S. Maruoka, K. Kujime, S. Hayashi, I. Takeshita, and T. Horie. p38 map kinase regulates tnf alpha-, il-1 alpha- and paf-induced rantes and

gm-csf production by human bronchial epithelial cells. *Clin Exp Allergy*, 30(1):48–55, 2000.

[49] D. He, Z. Xu, S. Dong, H. Zhang, H. Zhou, L. Wang, and S. Zhang. Teriflunomide for multiple sclerosis. *Cochrane Database Syst. Rev.*, 12:CD009882, 2012.

[50] D. Heckerman, D. Geiger, and D. M. Chickering. Learning bayesian networks - the combination of knowledge and statistical-data. *Machine Learning*, 20(3):197–243, 1995.

[51] T. S. Hiura, S. J. Kempiak, and A. E. Nel. Activation of the human rantes gene promoter in a macrophage cell line by lipopolysaccharide is dependent on stress-activated protein kinases and the ikappab kinase cascade: implications for exacerbation of allergic inflammation by environmental pollutants. *Clin Immunol*, 90(3):287–301, 1999.

[52] V. Hnizdo, E. Darian, A. Fedorowicz, E. Demchuk, S. Li, and H. Singh. Nearest-neighbor nonparametric method for estimating the configurational entropy of complex molecules. *J Comput Chem*, 28(3):655–68, 2007.

[53] R. M. Hobbs and F. M. Watt. Regulation of interleukin-1alpha expression by integrins and epidermal growth factor receptor in keratinocytes from a mouse model of inflammatory skin disease. *J Biol Chem*, 278(22):19798–807, 2003.

[54] D. R. Hyduke and B. O. Palsson. Towards genome-scale signalling network reconstructions. *Nat Rev Genet*, 11(4):297–307, 2010.

[55] T. Ideker, T. Galitski, and L. Hood. A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet*, 2:343–72, 2001.

[56] K. Ip, C. Colijn, and D. S. Lun. Analysis of complex metabolic behavior through pathway decomposition. *BMC Syst Biol*, 5:91, 2011.

[57] F. J. Johannes, J. Horn, G. Link, E. Haas, K. Siemienski, H. Wajant, and K. Pfizenmaier. Protein kinase Cmu downregulation of tumor-necrosis-factor-induced apoptosis correlates with enhanced expression of nuclear-factor-kappab-dependent protective genes. *Eur. J. Biochem.*, 257(1):47–54, 1998.

[58] I. T. Jolliffe. *Principal component analysis*. Springer series in statistics. Springer-Verlag, New York, 2nd edition, 2002.

[59] N. J. Jordan, M. L. Watson, T. Yoshimura, and J. Westwick. Differential effects of protein kinase c inhibitors on chemokine production in human synovial fibroblasts. *Br J Pharmacol*, 117(6):1245–53, 1996.

[60] D. E. Joseph, C. C. Paul, M. A. Baumann, and J. Gomez-Cambronero. S6 kinase p90rsk in granulocyte-macrophage colony-stimulating factor-stimulated proliferative and mature hematopoietic cells. *J Biol Chem*, 271(22):13088–93, 1996.

[61] M. Kanehisa and S. Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30, 2000.

[62] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res*, 40(Database issue):D109–14, 2012.

[63] J. Kelly, R. Spolski, K. Imada, J. Bollenbacher, S. Lee, and W. J. Leonard. A role for stat5 in cd8+ t cell homeostasis. *J Immunol*, 170(1):210–7, 2003.

[64] I. Kojadinovic. On the use of mutual information in data analysis: an overview. In *International symposium applied stochastic models data analysis*.

[65] Rikard König. Predictive techniques and methods for decision support in situations with poor data quality. 2009.

[66] D. Kontoyiannis, M. Pasparakis, T. T. Pizarro, F. Cominelli, and G. Kollias. Impaired on/off regulation of tnf biosynthesis in mice lacking tnf au-rich elements: implications for joint and gut-associated immunopathologies. *Immunity*, 10(3):387–98, 1999.

[67] S. S. Kothari, M. S. Abrahamsen, T. Cole, and W. P. Hammond. Expression of granulocyte colony stimulating factor (g-csf) and granulocyte/macrophage colony stimulating factor (gm-csf) mrna upon stimulation with phorbol ester. *Blood Cells Mol Dis*, 21(3):192–200, 1995.

[68] J. C. Kovacic, R. Gupta, A. C. Lee, M. Ma, F. Fang, C. N. Tolbert, A. D. Walts, L. E. Beltran, H. San, G. Chen, C. St Hilaire, and M. Boehm. Stat3-dependent acute rantes production in vascular smooth muscle cells modulates inflammation following arterial injury in mice. *J Clin Invest*, 120(1):303–14, 2010.

[69] Richard Kramer. *Chemometric techniques for quantitative analysis*. Marcel Dekker, New York, 1998.

[70] A. Kraskov, H. Stogbauer, and P. Grassberger. Estimating mutual information. *Phys Rev E Stat Nonlin Soft Matter Phys*, 69(6 Pt 2):066138, 2004.

[71] D. V. Kuprash, I. A. Udalova, R. L. Turetskaya, D. Kwiatkowski, N. R. Rice, and S. A. Nedospasov. Similarities and differences between human and murine tnf promoters in their response to lipopolysaccharide. *J Immunol*, 162(7):4045–52, 1999.

[72] S. Lentzsch, M. Gries, M. Janz, R. Bargou, B. Dorken, and M. Y. Mapara. Macrophage inflammatory protein 1-alpha (MIP-1 alpha) triggers migration and signaling cascades mediating survival and proliferation in multiple myeloma (MM) cells. *Blood*, 101(9):3568–73, 2003.

[73] D. Leyva-Illades, R. P. Cherla, M. S. Lee, and V. L. Tesh. Regulation of cytokine and chemokine expression by the ribotoxic stress response elicited by shiga toxin type 1 in human macrophage-like THP-1 cells. *Infect. Immun.*, 80(6):2109–2120, 2012.

[74] W. Li, Y. Liu, H. C. Huang, Y. Peng, Y. Lin, W. K. Ng, and K. L. Ong. Dynamical systems for discovering protein complexes and functional modules from biological networks. *IEEE/ACM Trans Comput Biol Bioinform*, 4(2):233–50, 2007.

[75] X. Li, S. Rao, W. Jiang, C. Li, Y. Xiao, Z. Guo, Q. Zhang, L. Wang, L. Du, J. Li, L. Li, T. Zhang, and Q. K. Wang. Discovery of time-delayed gene regulatory networks based on temporal gene expression profiling. *BMC Bioinformatics*, 7:26, 2006.

[76] D.E. Losada and J.M. Fernández-Luna. *Advances in Information Retrieval: 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21-23, 2005, Proceedings*. Lecture Notes in Computer Science / Information Systems and Applications, incl. Internet/Web, and HCI. Springer, 2005.

[77] F. D. Lublin and S. C. Reingold. Defining the clinical course of multiple sclerosis: results of an international survey. national multiple sclerosis society (usa) advisory committee on clinical trials of new agents in multiple sclerosis. *Neurology*, 46(4):907–11, 1996.

[78] J. Ma, T. Chen, J. Mandelin, A. Ceponis, N. E. Miller, M. Hukkanen, G. F. Ma, and Y. T. Konttinen. Regulation of macrophage activation. *Cell. Mol. Life Sci.*, 60(11):2334–46, 2003.

[79] Jian Ma and Zengqi Sun. Mutual information is copula entropy. *CoRR*, abs/0808.0845, 2008.

[80] W. Ma, W. Lim, K. Gee, S. Aucoin, D. Nandan, M. Kozlowski, F. Diaz-Mitoma, and A. Kumar. The p38 mitogen-activated kinase pathway regulates the human interleukin-10 promoter via the activation of sp1 transcription factor in lipopolysaccharide-stimulated human macrophages. *J Biol Chem*, 276(17):13664–74, 2001.

[81] S. Malhotra, M. F. Bustamante, F. Perez-Miralles, J. Rio, M. C. Ruiz de Villa, E. Vegas, L. Nonell, F. Deisenhammer, N. Fissolo, R. N. Nurtdinov, X. Montalban, and M. Comabella. Search for specific biomarkers of IFNbeta bioactivity in patients with multiple sclerosis. *PLoS One*, 6(8):e23634, 2011.

[82] A. Manouchehrinia and C. S. Constantinescu. Cost-effectiveness of disease-modifying therapies in multiple sclerosis. *Curr. Neurol. Neurosci. Rep.*, 12(5):592–600, 2012.

[83] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7 Suppl 1:S7, 2006.

[84] T. K. Means, R. P. Pavlovich, D. Roca, M. W. Vermeulen, and M. J. Fenton. Activation of tnf-alpha transcription utilizes distinct map kinase pathways in different macrophage populations. *J Leukoc Biol*, 67(6):885–93, 2000.

[85] T.W. Miller. *Modeling Techniques in Predictive Analytics: Business Problems and Solutions with R*. FT Press Analytics Series. Pearson Education, 2013.

[86] R. Milo and E. Kahana. Multiple sclerosis: geoepidemiology, genetics and the environment. *Autoimmun Rev*, 9(5):A387–94, 2010.

[87] Y. I. Moon, B. Rajagopalan, and U. Lall. Estimation of mutual information using kernel density estimators. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics*, 52(3):2318–2321, 1995.

[88] N. Mori and D. Prager. Transactivation of the interleukin-1 alpha promoter by human t-cell leukemia virus. *Leuk Lymphoma*, 26(5-6):421–33, 1997.

[89] D. M. Mosser and J. P. Edwards. Exploring the full spectrum of macrophage activation. *Nat Rev Immunol*, 8(12):958–69, 2008.

[90] et al. Mugdadi, A.R. A bandwidth selection for kernel density estimation of functions of random variables. *Computation Statistics and Data Analysis*, 47(1):49–62, 2004.

[91] P. Müller and F. A. Quintana. Nonparametric Bayesian data analysis. *Stat. Sci.*, 19(1):95–110, 02 2004.

[92] D. Murdoch and K. A. Lyseng-Williamson. Spotlight on subcutaneous recombinant interferon-beta-1a (rebif) in relapsing-remitting multiple sclerosis. *BioDrugs*, 19(5):323–325, 2005.

[93] J. Nakahara, M. Maeda, S. Aiso, and N. Suzuki. Current concepts in multiple sclerosis: autoimmunity versus oligodendrogliopathy. *Clin Rev Allergy Immunol*, 42(1):26–34, 2012.

[94] S. Naqvi, A. Macdonald, C. E. McCoy, J. Darragh, A. D. Reith, and J. S. Arthur. Characterization of the cellular action of the MSK inhibitor SB-747651A. *Biochem. J.*, 441(1):347–57, 2012.

[95] Richard E. Neapolitan. *Learning Bayesian networks*. Prentice Hall, Upper Saddle River, NJ, 2004.

[96] Roger B. Nelsen. *An introduction to copulas*. Springer series in statistics. Springer, New York, 2nd ed. edition, 2006.

[97] J. Numata, O. Ebenhoh, and E. W. Knapp. Measuring correlations in metabolomic networks with mutual information. *Genome Inform.*, 20:112–122, 2008.

[98] T. Ogawa, M. Kusumoto, S. Kuroki, S. Nagata, N. Yamanaka, R. Kawano, J. Yoshida, M. Shinohara, and K. Matsuo. Adjuvant gm-csf cytokine gene therapy for breast cancer. *Gan To Kagaku Ryoho*, 28(11):1512–4, 2001.

[99] Y. Ohmori, R. D. Schreiber, and T. A. Hamilton. Synergy between interferon-gamma and tumor necrosis factor-alpha in transcriptional activation is mediated by cooperation between signal transducer and activator of transcription 1 and nuclear factor kappab. *J Biol Chem*, 272(23):14899–907, 1997.

[100] V. Ollivier, G. C. Parry, R. R. Cobb, D. de Prost, and N. Mackman. Elevated cyclic amp inhibits nf-kappab-mediated transcription in human monocytic cells and endothelial cells. *J Biol Chem*, 271(34):20828–35, 1996.

[101] J. J. Oppenheim, W. J. Murphy, O. Chertox, V. Schirrmacher, and J. M. Wang. Prospects for cytokine and chemokine biotherapy. *Clin Cancer Res*, 3(12 Pt 2):2682–6, 1997.

[102] O. N. Ozes, H. Akca, L. D. Mayo, J. A. Gustin, T. Maehama, J. E. Dixon, and D. B. Donner. A phosphatidylinositol 3-kinase/akt/mtor pathway mediates and pten antagonizes tumor necrosis factor inhibition of insulin signaling through insulin receptor substrate-1. *Proc Natl Acad Sci U S A*, 98(8):4640–5, 2001.

[103] Judea Pearl. Bayesian networks. *Department of Statistics, UCLA*, 2011.

[104] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell*, 27(8):1226–38, 2005.

[105] B. E. Perrin, L. Ralaivola, A. Mazurie, S. Bottani, J. Mallet, and F. d'Alche Buc. Gene networks inference using dynamic bayesian networks. *Bioinformatics*, 19 Suppl 2:ii138–48, 2003.

[106] D. M. W. Powers. Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.

[107] S. Pradervand, M. R. Maurya, and S. Subramaniam. Identification of signaling components required for the prediction of cytokine release in raw 264.7 macrophages. *Genome Biol*, 7(2):R11, 2006.

[108] C. Qian, X. Jiang, H. An, Y. Yu, Z. Guo, S. Liu, H. Xu, and X. Cao. Tlr agonists promote erk-mediated preferential il-10 production of regulatory dendritic cells (diffdcs), leading to nk-cell activation. *Blood*, 108(7):2307–15, 2006.

[109] et al. Raykar, V. Fast optimal bandwidth selection for kernel density estimation. In *sixth SIAM International Conference on Data Mining.*

[110] M. B. Reeves and T. Compton. Inhibition of inflammatory interleukin-6 activity via extracellular signal-regulated kinase-mitogen-activated protein kinase signaling antagonizes human cytomegalovirus reactivation from dendritic cells. *J Virol*, 85(23):12750–12758, 2011.

[111] P. T. Reiss and R. T. Ogden. Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association*, 102(479):984–996, 2007.

[112] A. A. Rensink, H. Gellekink, I. Otte-Holler, H. J. ten Donkelaar, R. M. de Waal, M. M. Verbeek, and B. Kremer. Expression of the cytokine leukemia inhibitory factor and pro-apoptotic insulin-like growth factor binding protein-3 in alzheimer's disease. *Acta Neuropathol*, 104(5):525–33, 2002.

[113] D. Rossi and A. Zlotnik. The biology of chemokines and their receptors. *Annu. Rev. Immunol.*, 18:217–242, 2000.

[114] S. Saito. Cytokine cross-talk between mother and the embryo/placenta. *J Reprod Immunol*, 52(1-2):15–33, 2001.

[115] J. L. Sanders and P. H. Stern. Protein kinase C involvement in interleukin-6 production by parathyroid hormone and tumor necrosis factor-alpha in UMR-106 osteoblastic cells. *J. Bone Miner. Res.*, 15(5):885–893, 2000.

[116] Dipen P Sangurdekar, Friedrich Srienc, and Arkady B Khodursky. A classification based framework for quantitative description of large-scale microarray data. *Genome biology*, 7(4):R32, 2006.

[117] J. Scheller, A. Chalaris, D. Schmidt-Arras, and S. Rose-John. The pro- and anti-inflammatory properties of the cytokine interleukin-6. *Biochim. Biophys. Acta*, 1813(5):878–888, 2011.

[118] C. H. Schilling, S. Schuster, B. O. Palsson, and R. Heinrich. Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. *Biotechnol. Prog.*, 15(3):296–303, 1999.

[119] C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.

[120] G. Shaw and R. Kamen. A conserved au sequence from the 3' untranslated region of gm-csf mrna mediates selective mrna degradation. *Cell*, 46(5):659–67, 1986.

[121] B. W. Silverman. *Density estimation for statistics and data analysis*. Monographs on statistics and applied probability. Chapman and Hall, London ; New York, 1986.

[122] B. W. Silverman. *Density estimation for statistics and data analysis*. Monographs on statistics and applied probability. Chapman and Hall/CRC, Boca Raton, 1998.

[123] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*, 9(12):3273–97, 1998.

[124] V. Spirin and L. A. Mirny. Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci U S A*, 100(21):12123–8, 2003.

[125] A. C. Stanley and P. Lacy. Pathways for cytokine secretion. *Physiology (Bethesda)*, 25(4):218–29, 2010.

[126] M. Steinfath, D. Groth, J. Lisec, and J. Selbig. Metabolite profile analysis: from raw data to regression and classification. *Physiol Plant*, 132(2):150–61, 2008.

[127] J. Stelling, S. Klamt, K. Bettenbrock, S. Schuster, and E. D. Gilles. Metabolic network structure determines key aspects of functionality and regulation. *Nature*, 420(6912):190–3, 2002.

[128] K. Suzuki, M. Hino, F. Hato, N. Tatsumi, and S. Kitagawa. Cytokine-specific activation of distinct mitogen-activated protein kinase subtype cascades in human neutrophils stimulated by granulocyte colony-stimulating factor, granulocyte-macrophage colony-stimulating factor, and tumor necrosis factor-alpha. *Blood*, 93(1):341–9, 1999.

[129] P. J. Toscas, F. D. Shaw, and S. L. Beilken. Partial least squares (pls) regression for the analysis of instrument measurements and sensory meat quality data. *Meat Sci*, 52(2):173–8, 1999.

[130] P. Tremblay, M. Houde, N. Arbour, D. Rochefort, S. Masure, R. Mandeville, G. Opdenakker, and D. Oth. Differential effects of pkc inhibitors on gelatinase b and interleukin 6 production in the mouse macrophage. *Cytokine*, 7(2):130–6, 1995.

[131] E. Y. Tsai, J. V. Falvo, A. V. Tsytsykova, A. K. Barczak, A. M. Reimold, L. H. Glimcher, M. J. Fenton, D. C. Gordon, I. F. Dunn, and A. E. Goldfeld. A lipopolysaccharide-specific enhancer complex involving ets, elk-1, sp1, and creb binding protein and p300 is recruited to the tumor necrosis factor alpha promoter in vivo. *Mol Cell Biol*, 20(16):6084–94, 2000.

[132] B. A. Turlach. Bandwidth selection in kernel density estimation: A review. *In CORE and Institut de Statistique*, pages 23–493, 1993.

[133] B. A. Turlach. Bandwidth selection in kernel density estimation: A review. Technical report, Univ. Catholique de Louvain, 1993.

[134] S. R. S. Varadhan. Asymptotic probabilities and differential equations. *Communications on Pure and Applied Mathematics*, 19(3):261–286, 1966.

[135] S.R.S. Varadhan. *Large Deviations and Applications*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 1984.

[136] Y. Wang, T. R. Wu, S. Cai, T. Welte, and Y. E. Chin. Stat1 as a component of tumor necrosis factor alpha receptor 1-TRADD signaling complex to inhibit NF-kappab activation. *Mol. Cell Biol.*, 20(13):4505–4512, 2000.

[137] A. Y. Wen, K. M. Sakamoto, and L. S. Miller. The role of the transcription factor CREB in immune function. *J. Immunol.*, 185(11):6413–6419, 2010.

[138] Peter D. Wentzell and Lorenzo Vega Montoto. Comparison of principal components regression and partial least squares regression through generic simulations of complex mixtures. *Chemometrics and Intelligent Laboratory Systems*, 65:257âĂŞ 279, 2003.

[139] C. K. Wikle and L. M. Berliner. A Bayesian tutorial for data assimilation. *Physica D*, 230(1-2):1–16, 2007.

[140] Y. Wu, G. L. Johnson, and S. M. Gomez. Data-driven modeling of cellular stimulation, signaling and output response in raw 264.7 cells. *J Mol Signal*, 3:11, 2008.

[141] T. Yin and Y. C. Yang. Mitogen-activated protein kinases and ribosomal s6 protein kinases are involved in signaling pathways shared by interleukin-11, interleukin-6, leukemia inhibitory factor, and oncostatin m in mouse 3t3-l1 cells. *J Biol Chem*, 269(5):3731–8, 1994.

[142] Y. Zhang, Q. Zhai, Y. Luo, and M. E. Dorf. Rantes-mediated chemokine transcription in astrocytes involves activation and translocation of p90 ribosomal s6 protein kinase (rsk). *J Biol Chem*, 277(21):19042–8, 2002.

[143] P. Zoppoli, S. Morganella, and M. Ceccarelli. Timedelay-aracne: Reverse engineering of gene networks from time-course data by an information theoretic approach. *BMC Bioinformatics*, 11:154, 2010.