IDENTIFICATION OF TRANSITION BASED MODELS FOR THE INFERENCE OF SUB-CELLULAR NETWORKS

DEEPIKA VATSA

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI
SEPTEMBER 2020

IDENTIFICATION OF TRANSITION BASED MODELS FOR THE INFERENCE OF SUB-CELLULAR NETWORKS

by

DEEPIKA VATSA

DEPARTMENT OF ELECTRICAL ENGINEERING

Submitted

in fulfillment of the requirements of the degree of Doctor of Philosophy

to the

INDIAN INSTITUTE OF TECHNOLOGY DELHI

SEPTEMBER 2020

To my loving parents

Certificate

This is to certify that the thesis titled **Identification of transition based models for the inference of sub-cellular networks** being submitted by **Ms. Deepika Vatsa** to the Department of Electrical Engineering, Indian Institute of Technology Delhi, for the award of **Doctor of Philosophy** is a record of the bona-fide research work carried out by her under my guidance and supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree. The work presented in this thesis has not been submitted elsewhere, either in part or full, for the award of any other degree or diploma.

Dr. Sumeet Agarwal
Assistant Professor
Department of Electrical Engineering
Indian Institute of Technology Delhi
New Delhi - 110 016, India.

Date:

Place: New Delhi

Acknowledgements

I would have to thank many people for their support, guidance, and motivation who get me through my Ph.D. journey. First and foremost, I would like to express my sincere gratitude towards my advisor Dr. Sumeet Agarwal who is an inspiration for me. I am thankful for his constant guidance and support throughout these years. His intelligence, ideas, and constructive discussions in our weekly meetings has motivated me throughout and has helped me to establish the overall direction of this thesis. He had always given me time when needed and understood me in times of difficulties. This thesis is an outcome of his vision. I would also like to thank Prof. Ashwin Srinivasan from BITS Pilani, Goa for his ideas, and guidance in my research work. I am also grateful to the members of my SRC committee, Dr. Sumantra Dutta Roy, Prof. James Gomes, and Dr. Shaunak Sen for the insightful suggestions and constructive feedback for my work at end semester's progress meetings. I also wish to thank my examiners for their feedback and suggestions which has greatly improved my thesis.

I thank technical staff members of the Multimedia Lab for allowing me to own this lab and to let me work in day and night as per my convenience. I had been very lucky to share the lab with labmates like Manoj, Ronak, Richa, Anupama, Shruti, and Snigdha who had been there for me and provided a friendly environment in the lab. I especially thank Anupama for being a source of motivation for me. I thank my friends Richa, Lucky and Jagriti for all the friendliness and support through good and bad.

Finally, I owe my deepest gratitude to my parents, sisters, and brother for all the love and care. Thank you for being the pillar of support throughout. I thank my father for always believing in me and encouraging me to achieve the best. I dedicate this thesis to him and my mom. I am also deeply thankful to my extended family for understanding me and supporting me in the crucial years of my Ph.D. I especially thank my husband for being there for me always and supporting me throughout. No words can thank you enough!

Deepika Vatsa

Abstract

This thesis is devoted to the study of biological networks at the sub-cellular level. Network inference is one of the branches of systems biology that aims to contribute in this direction. It focuses on the inference of interactions between the molecular species in biological networks. Computational approaches are employed to retrieve these interactions from the molecular interaction data. The objective of this thesis is to develop mathematical learning approaches for network inference and contribute to the understanding of biological subsystems.

We can classify the work done in this thesis along three dimensions. The first is biological networks studied. We particularly looked at signalling networks and gene regulatory networks. The second is the organisms studied. We looked at signalling pathways and expression data sets of organisms such as *E. coli*, *S. cerevisiae* and *A. thaliana*. The third is computational models used for the network inference task. All work done in this thesis revolves around transition-based models (specifically Petri nets).

Though a large amount of biological data is available, mostly, this data is sparse, noisy, and limited, which creates a challenge in the network inference procedure. Probabilistic extensions of the transition-based approaches are developed to tackle these challenges. The probabilistic formulation of the approaches tries to model the inherent stochastic behaviour of the biological subsystem and includes a provision for the incorporation of domain knowledge that helps in dealing with the limited data problem. To model the nature of interactions in the two subsystems, i.e., signalling networks and gene regulatory networks, we proposed two network reconstruction approaches, Probabilistic Logic-Guarded Transition System (PLGTS) and Probabilistic Extended Petri Net for Gene Regulatory Networks (PEPN-GRN), respectively for each subsystem. We also tried to fit the probabilistic formulation of Petri nets in the family of graphical models and proved their equivalence with the Dynamic Bayesian Networks for a particular class of transition systems that exhibit the regulation mechanism.

The ultimate goal of systems biology is to provide a holistic view of a biological system. The unification of proposed modelling approaches PLGTS and PEPN-GRN in a single framework is a step towards achieving this goal, where along with the inference of individual subsystems, the

interplay between the different subsystems is also illustrated.

सार

यह थीसिस उप-सेलुलर स्तर पर जैविक नेटवर्क के अध्ययन के लिए समर्पित है। नेटवर्क इंट्रेंस सिस्टम जीव विज्ञान की शाखाओं में से एक है जिसका उद्देश्य इस दिशा में योगदान करना है। यह जैविक नेटवर्क में आणविक प्रजातियों के बीच बातचीत की शुरूआत पर केंद्रित है। आणविक इंटरैक्शन डेटा से इन इंटरैक्शन को पुनर्प्राप्त करने के लिए कम्प्यूटेशनल दृष्टिकोण कार्यरत हैं। इस थीसिस का उद्देश्य नेटवर्क के आविष्कार के लिए गणितीय मॉडल विकसित करना और जैविक उप-प्रणालियों की समझ में योगदान करना है।

हम इस शोध में किए गए कार्यों को तीन आयामों के साथ वर्गीकृत कर सकते हैं। पहले जैविक नेटवर्क का अध्ययन किया जाता है। हमने विशेष रूप से सिग्नलिंग नेटवर्क और जीन विनियामक नेटवर्क को देखा। दूसरा अध्ययन किया गया जीव है। हमने ई कोलाई, एस सेरेविसिया और ए थलियाना जैसे जीवों के मार्ग और अभिव्यक्ति डेटा सेट को देखा। तीसरा कम्प्यूटेशनल मॉडल है, जिसका उपयोग नेटवर्क इंट्रेंस टास्क के लिए किया जाता है। इस थीसिस में किए गए सभी कार्य संक्रमण-आधारित मॉडल (विशेष रूप से पेट्री नेट) के चारों ओर घूमते हैं।

यद्यपि जैविक डेटा की एक बड़ी मात्रा उपलब्ध है, ज्यादातर, यह डेटा विरल, शोर, और सीमित है, जो नेटवर्क की प्रक्रिया में एक चुनौती पैदा करता है। इन चुनौतियों से निपटने के लिए संक्रमण-आधारित मॉडल के संभावित विस्तार विकसित किए गए हैं। मॉडल का संभाव्य सूत्रीकरण जैविक सबसिस्टम के निहित स्टोकेस्टिक व्यवहार को मॉडल करने की कोशिश करता है, और इसमें डोमेन ज्ञान के समावेश के लिए एक प्रावधान शामिल है जो सीमित डेटा समस्या से निपटने में मदद करता है। दो उपतंत्रों, अर्थात सिग्नलिंग नेटवर्क और जीन विनियामक नेटवर्क में परस्पर क्रियाओं की प्रकृति को मॉडल करने के लिए, हमने दो नेटवर्क पुनर्निर्माण दृष्टिकोण, प्रोबेबिलिस्टिक लॉजिक-गार्डेड ट्रांजिशन सिस्टम (पीएलजीटीएस) और जिनेबुल नेटवर्क्स नेटवर्क (पीईपीएन-जीआरएन) के लिए संभावित विस्तारित पेट्री नेट का प्रस्ताव दिया। क्रमश: प्रत्येक सबसिस्टम के लिए। हमने ग्राफिकल मॉडल के परिवार में पेट्री नेट के संभाव्य सूत्रीकरण को भी फिट करने की कोशिश की और विनियमन प्रणाली को प्रदर्शित करने वाले संक्रमण प्रणालियों के एक विशेष वर्ग के लिए डायनेमिक बेयसियन नेटवर्क के साथ अपनी समानता साबित की।

सिस्टम बायोलॉजी का अंतिम लक्ष्य एक जैविक प्रणाली का एक समग्र दृष्टिकोण प्रदान करना है। प्रस्तावित मॉडलिंग के एकीकरण से पीएलजीटीएस और एकल ढांचे में पीईपीएन-जीआरएन इस लक्ष्य को प्राप्त करने की दिशा में एक कदम है, जहां व्यक्तिगत उपप्रणालियों के हस्तक्षेप के साथ-साथ विभिन्न उपप्रणालियों के बीच परस्पर क्रिया को भी चित्रित किया गया है।

Contents

List of Figures

List of Tables

List of Acronyms

DNA Deoxyribo Nucleic Acid

cDNA complmentary Deoxyribo Nucleic Acid

mRNA Messenger Ribonucleic Acid

PPI Protein Protein Interaction

TF Transcription Factor

SMD Stanford Microarray Database

Y2H Yeast two-Hybrid

AP Affinity Purification

MS Mass Spectroscopy

SELDOM enSEmbLe of Dynamic lOgic-based Model

PCC Pearson Correlation Coefficient

PN Petri Net

P-invariant Place invariant

T-invariant Transition invariant

SPN Stochastic Petri Net

GSPN Generalized Stochastic Petri Net

BCN Boolean Control Network

PBN Probabilistic Boolean Network

ODE Ordinary Differential Equation

SDE Stochastic Differential Equation

BN Bayesian Network

DAG Directed Acyclic Graph

CTMC Continuous Time Markov Chain

CPD Conditional Probability Distribution

CPT Conditional Probability Table

DBN Dynamic Bayesian Network

HMM Hidden Markov Model

GRN Gene Regulatory Network

LGTS Logic Guarded Transition System

PLGTS Probabilistic Logic Guarded Transition System

PEPN Probabilistic Extended Petri Net

PEPN-GRN Probabilistic Extended Petri Net for Gene Regulatory Network

MAPK Mitogen-Activated Protein Kinase

PRISM PRogramming In Statistical Modelling

NFA Non-deterministic Finite Automata

PFA Probabilistic Finite Automata

CES Cause-Effect Structure

CTBN Continuous Time Bayesian Network

GCTBN Generalized Continuous Time Bayesian Network

LTL Linear Temporal Logic

PNFL Petri Nets with Fuzzy Logic

DREAM Dialogue for Reverse Engineering Assessments and Methods

EFD Equal Frequency Discretization

EWD Equal Width Discretization

ARACNE Algorithm for the Reconstruction of Accurate Cellular NEtworks

CLR Context Likelihood of Relatedness

MRNET Minimum-Redundancy NETwork

E. coli Escherichia coli

S. cerevisiae Saccharomyces cerevisiae

A. thaliana Arabidopsis thaliana

TP True Positive

FP False Positive

TN True Negative

FN False Negative

TPR True Positive Rate

FPR False Positive Rate

AUC Area Under the Curve

ROC Receiver-Operating Characteristic

PR Precision-Recall

AUROC Area Under Receiver-Operating Characteristic

AUPR Area Under Precision-Recall

PPV Positive Predictive Value

LR Logistic Regression

SMOTE Synthetic Minority Oversampling TEchnique