SINEC: LARGE SCALE SIGNALING NETWORK TOPOLOGY RECONSTRUCTION USING PROTEIN-PROTEIN INTERACTIONS AND RNAI DATA

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

SEYEDSASAN HASHEMIKHABIR

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

SEPTEMBER 2012

Approval of the thesis:

SINEC: LARGE SCALE SIGNALING NETWORK TOPOLOGY RECONSTRUCTION USING PROTEIN-PROTEIN INTERACTIONS AND RNAI DATA

submitted by **SEYEDSASAN HASHEMIKHABIR** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen
Dean, Graduate School of **Natural and Applied Sciences** ________________

Prof. Dr. Adnan Yazıcı
Head of Department, **Computer Engineering** ________________

Assoc. Prof. Dr. Tolga Can
Supervisor, **Computer Engineering Department** ________________

Examining Committee Members:

Prof. Dr. Göktürk Üçoluk
Computer Engineering, Middle East Technical University ________________

Assoc. Prof. Dr. Tolga Can
Computer Engineering, Middle East Technical University ________________

Prof. Dr. Faruk Polat
Computer Engineering, Middle East Technical University ________________

Assoc. Prof. Dr. Hasan Oğul
Computer Engineering, Baskent University ________________

Assist. Prof. Dr. Aybar Can Acar
Informatics Institute, Middle East Technical University ________________

Date: ________________

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: SEYEDSASAN HASHEMIKHABIR

Signature :

ABSTRACT

SINEC: LARGE SCALE SIGNALING NETWORK TOPOLOGY RECONSTRUCTION USING PROTEIN-PROTEIN INTERACTIONS AND RNAI DATA

Hashemikhabir, Seyedsasan
M.S., Department of Computer Engineering
Supervisor : Assoc. Prof. Dr. Tolga Can

September 2012, 40 pages

Reconstructing the topology of a signaling network by means of RNA interference (RNAi) technology is an underdetermined problem especially when a single gene in the network is knocked down or observed. In addition, the exponential search space limits the existing methods to small signaling networks of size 10-15 genes. In this thesis, we propose integrating RNAi data with a reference physical interaction network. We formulate the problem of signaling network reconstruction as finding the minimum number of edit operations on a given reference network. The edit operations transform the reference network to a network that satisfy the RNAi observations. We show that using a reference network does not simplify the computational complexity of the problem. Therefore, we propose an approach that provides near optimal results and can scale well for reconstructing networks up to hundreds of components. We validate the proposed method on synthetic and real datasets. Comparison with the state of the art on real signaling networks shows that the proposed methodology can scale better and generates biologically significant results.

Keywords: signal transduction, RNAi, protein-protein interactions, network construction

ÖZ

SINEC: PROTEİN-PROTEİN ETKİLEŞİM VE RNAI VERİLERİ KULLANARAK YÜKSEK ÖLÇEKLİ SİNYALLEME YOLAK TOPOLOJİLERİNİN OLUŞTURULMASI

Hashemikhabir, Seyedsasan
Yüksek Lisans, Bilgisayar Müğendisliği Bölümü
Tez Yöneticisi : Doç. Dr. Tolga Can

Eylül 2012, 40 sayfa

Biyolojik sinyalleme yolaklarının topolojilerinin RNA engelleme (RNAi) teknolojisi kullanılarak tespit edilmesi problemi, özellikle yolaktaki tek bir gen susturulduğunda ya da gözlemlendiğinde eksik belirtilmiş bir problemdir. Buna ek olarak, RNAi verisi ile tutarlı yolak topolojilerinin uzayının gen sayısı ile üssel olarak artması da varolan yöntemleri ancak 10-15 genle sınırlamaktadır. Bu tezde, verilen bir RNAi verisini referans bir fiziksel etkileşim ağı ile birleştirmeyi önermekteyiz. Sinyalleme yolağı oluşturma problemini verilen bir referans ağ üzerinde en az sayıda ekleme/silme operasyonu yaparak RNAi verisi ile tutarlı hale getirme problemi olarak formalize etmekteyiz. Fakat bu formulasyonun da problemi kolay bir hale getirmediğini ve oluşan problemin NP-complete sınıfında bir problem olduğunu ispatlıyoruz. Bu nedenle bu tezde bu yeni formulasyon için optimum sonuca yakın sonuçlar üreten ve yüzlerce gen içeren yolaklar için çalışabilen bir yöntem önermekteyiz. Yöntemimizi sentetik ve gerçek veriler üzerinde doğrulamaktayız. Varolan yöntemler ile kıyaslandığında önerdiğimiz yöntemin daha iyi ölçeklendiğini ve biyolojik olarak daha doğru sonuçlar ürettiğini gösteriyoruz.

Anahtar Kelimeler: sinyalleme yolakları, RNAi, protein-protein etkileşimleri, ağ oluşturma

I dedicate this thesis to my parents and brother.

ACKNOWLEDGMENTS

> “The true sign of intelligence is not knowledge but imagination.”
> *—Albert Einstein*

My ultimate dream is to contribute to human-kind health. It might seem to be unreachable goal for a computer science student but I’m here to prove that it is possible. I had heard about Bioinformatics but never truly understood the details of this area of science. I was so lucky that I got accepted to Middle East Technical University and I was much more fortunate to be given chance to work on my thesis under the advising of Dr. Tolga Can. He introduced me to the area of Bioinformatics in a way that I could integrate most of my algorithmic skills into my research without being afraid of getting lost in biological concepts. He believed in me and my work and always corrected and supported my ideas. I would like to express my deepest gratitude to Dr. Tolga Can for his enthusiasm, his encouragement, his patience and never-ending support. He is a great mentor in both scientific and social life.

This thesis is partially supported by the European Union in the 7th framework program through SysPatho, grant 260429.

I also would like to thank Dr. Kahveci and his team, Serdar Ayaz and Yusuf Kavurucu from University of Florida as our joint research colleagues for contributing in this project.

Moreover, I would like to thank my lovely parents for their moral and financial support during, not only, my master thesis, but also, my whole life. It was impossible for me to accomplish my studies without their support.

Last but not least, I would like to thank anyone who taught me the points and values both in scientific and personal life and helped me to further develop my personality.

TABLE OF CONTENTS

CHAPTERS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

Introduction

Organisms respond to changes in their surrounding environment by producing chemical or physical reactions. Humans are not an exception. Fear, anxiety, happiness are the well-known sample reasons for a reaction in the human body. Faster heartbeat, production of adrenaline hormone and changing in the color of skin are the reactions for the mentioned agents respectively. In molecular biology, this cause-and-effect relationship is studied in cellular level as well. A cell detects a signal and responds accordingly. This process occurs by participation of many components within the cell such as genes and the whole process is referred as a signal transduction pathway. There are thousands of genes in the cell and each genes performs a specific task by interacting with some of the other genes. The genes and the interactions among them can be modeled as a network such that each node represents a gene in the network whereas an interaction between a pair of genes is shown as an edge. Usually, the proteins, i.e., the gene products are the main functional elements in these networks; hence. these networks are named as Protein-Protein Interaction (PPI) networks. A signal is transferred through a cascade of genes known as signal transduction pathways in the PPI network that results in the cell reaction to an external process. There are two major problems in reconstructing the signaling networks from the network datasets. First one is to identify the set of genes involved in the signaling network. Experimental techniques often tackle this component using genetic screens. One such widely used genetic screen is RNA interference (RNAi). RNAi is a gene silencing technology that removes a specific gene and its interactions from the network. Each gene in the network is silenced one by one and then it is checked whether the effect of the signal can be observed in a destination gene known as a reporter gene. After identification of the genes involved in the signaling network, the second task is to establish the topology of the interactions that connect these genes. With the advances in gene perturbation technology, the number of screened potential genes participating in signaling mechanisms grows rapidly and this increases the complexity level of studying and identifying the signaling network's structure. Solving this problem experimentally requires advanced and costly wet-lab trials. Hence, even for many known signaling networks, a complete and accurate picture of the signaling mechanism may not be available. In this thesis, we have suggested a number of heuristics to infer the underlying signaling network topology by the use of RNAi screening data and PPI networks together. The first input for solving this problem is the PPI dataset. It is a reference graph that the edges and vertices represent proteins and the interaction among them. The PPI datasets are inaccurate and there are missing or extra edges in this graph. The second input is the RNAi data. It is a binary array that each value in the array shows that whether silencing a specific gene in the PPI network disrupts the signal transfer or not. This binary array will be used to correct the reference PPI network by adding or removing edges. These heuristics scale well into the large RNAi screening experiments with hundreds of genes and constructs the signaling network topology in a reasonable amount of time.

1.1 Background Information

In this section, we review the key biological concepts frequently used in this thesis.

1.1.1 Protein-Protein Interaction Networks

Proteins are the indispensable part of a living organism and mainly responsible for the cellular processes. Although, they rarely operate independently, binding of two or more proteins execute a specific biological function. The binding between a pair of proteins is usually interpreted as an interaction between the two proteins. Every organism has thousands of proteins with tens of thousands of pairwise interactions which constitute a large scale network of proteins. PPI networks have the properties of *Scale-Free* and *Small-World* networks. In this type of networks, nodes are usually weakly connected and there are only few nodes that are highly connected to the other ones known as *hubs*. Moreover, every node in the network can be reached through a short chain of *hub* nodes. So, PPI networks are sparse and the *Sparse Graph* data structure provides the most reasonable way to represent

Figure 1.1: Yeast PPI Network

them and it also facilitates the studying of their structural properties.

Each node in a PPI graph represents a protein and the pairwise interactions between the proteins are defined as the edges connecting the nodes. Simple undirected graphs are the basic type for representing the PPI networks; however, with the availability of extra knowledge (i.e. interaction directions, interactions weight values), the graph type can changed into directed or weighted types respectively. A sample yeast (S. cerevisiae) PPI network is shown in Figure 1.1.

1.1.2 Signal Transduction Pathways

Signaling pathways provide the main mechanism for cells to react to extracellular signals such as hormones(Berg et al., 2002). *Signal Transduction* is a process in which a cell responds to external signals by processing them and converting them to another form of the signals. The process consists of an ordered sequence of biochemical reactions that starts from a sensory gene (e.g. cell-surface protein), advances through intermediates genes and finally ends in a downstream target gene (e.g. transcription factors). The whole process can be viewed as a biological circuit. The key components of these biological circuits are proteins

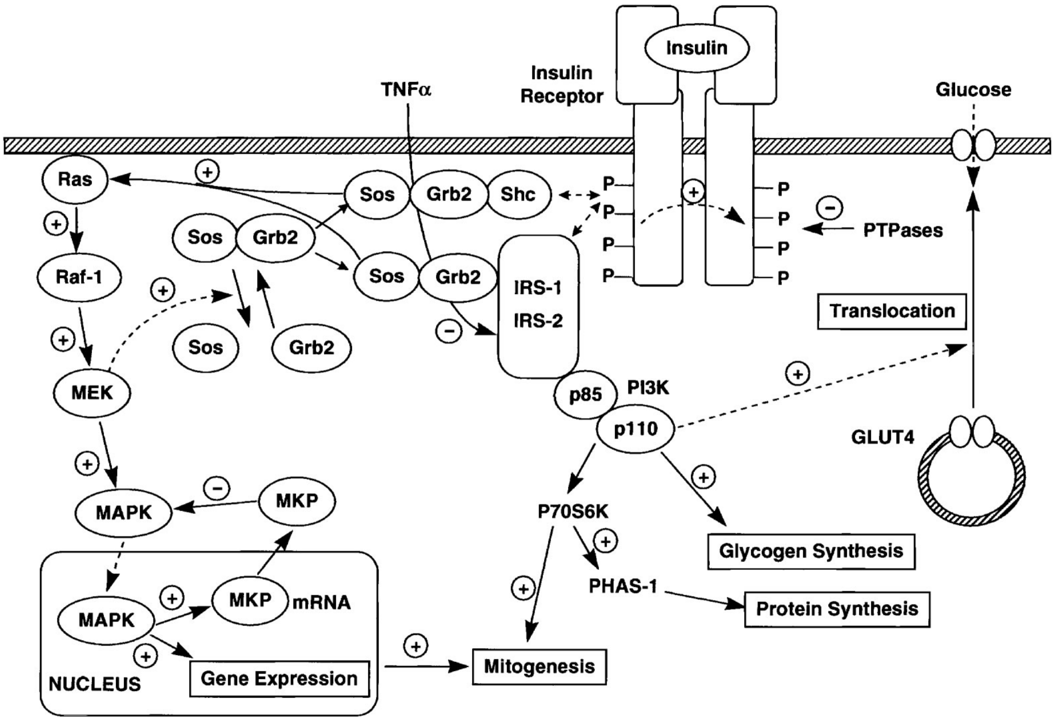


Figure 1.2: MAPK Signaling network

and a sequence of proteins that conduct a specific reaction is called, a *Signal Transduction Pathway*. Deficiencies in signal transduction often lead to disorders such as cancer, diabetes, and other genetic diseases. Signaling networks are often modeled as directed graphs where each node represents a protein or a gene involved in the network and each edge represents an effect of a protein on another one, such as phosphorylation. The effects are usually modeled as binary operations such as activation or repression. To simplify the language, we will refer to a node as "gene" rather than "protein" in the rest of the thesis unless it is necessary to distinguish between the two terms. As an example, the MAPK signaling network is depicted in Figure 1.2.

1.1.3 RNA Interference Screening

RNA Interference(RNAi) is a mechanism in living cells that regulate the gene translation process, i.e., the process to synthesize proteins from genes (Zhang, 2011). This mechanism in the cell plays an important role against the viruses and transposons. There are two types of small RNA molecules in the RNAi Pathway. The first type is known as small-interfering RNA or silencing RNA (siRNA) and the other one is micro RNA or miRNA. Both types can

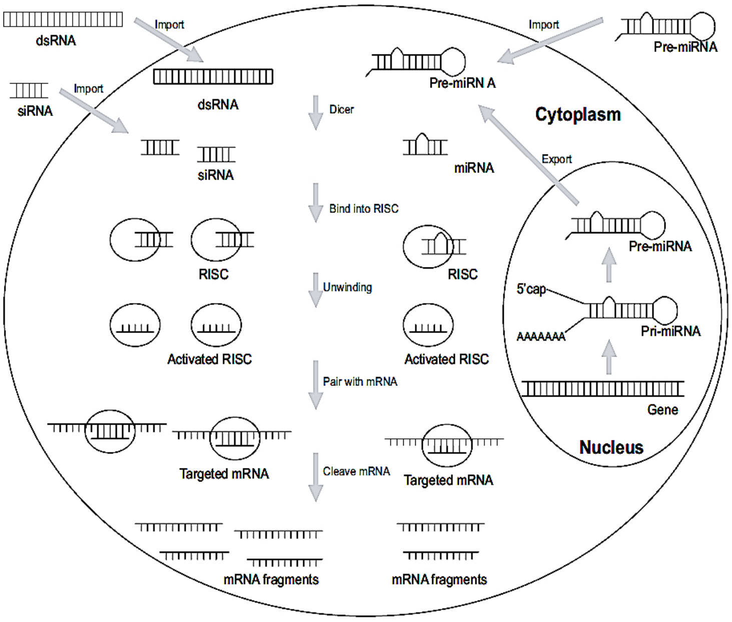


Figure 1.3: RNA Interference pathway

be attached to the specific RNAs that results in increasing or decreasing of their activities. Silencing or knocking down a gene is conducted by decreasing the expression of a gene using siRNA or miRNA that results in disruption of the gene's functionality in the pathway. RNA Interference (RNAi) Screening is a gene perturbation technology for observing the effects of the silenced gene(s) in a cell. Single gene knockdown is the most common way of observing the effect of the targeted gene on a reporter gene(Kaderali et al., 2009); however, due to the possible errors that can occur during the experiments (e.g. off-target effects, measuring faults), double or triple knocking down of genes are conducted simultaneously with multiple reporter genes to boost the precision of the resulting measured values(Sahin et al., 2009b). The structure of RNAi pathway is depicted in Figure 1.3; however, the biological details of this pathway is not within the scope of this thesis.

1.1.4 The KEGG Database

KEGG(Kanehisa and Goto, 2000) is an integrated resource for bioinformatics, consisting of three main data sources including the Gene Universe, the Chemical Universe, and the Protein Networks. The first two data source are conceptual and they present the structural similarity between genes and chemical association between the compounds respectively(Helms, 2008). Protein networks resource is the popular resource which includes a PATHWAY database. KEGG uses graph objects for representing networks where each vertex is a protein or other gene product and every edge represents a known interaction or relation between two vertices. The PATHWAY database contains several types of biological networks including Metabolic and Signaling networks for several species(Helms, 2008).

1.2 Problem Definition

Given a set G of genes known to participate in a signaling pathway and RNAi experiments performed on these genes, the problem is to construct a graph where each gene in G is a node in the graph, and the edge set E represents the signaling pathway topology. The unknown in this problem is the edge set E and our goal is to find E as biologically most accurate as possible. One of the genes in set G is designated as the *source* gene which initiates signaling and there is a separate *reporter* or *target* gene which is the final destination of the signal. These genes are known *a priori* and the connections between the intermediate genes are sought. In an RNAi experiment, each intermediate gene is turned off one by one and the signal is observed at the reporter gene. The existence/absence of the signal at the reporter genes gives us clues regarding the topology of the signaling network. As the number of possible network topologies is exponential in the number of genes in G and the number of RNAi experiments is linear in the number of genes, this is an ill defined problem. We tackle this problem by using additional data in the form of a reference PPI network. We formulate the problem as performing minimum number of edit operations on a reference PPI network so that the resulting network is consistent with the RNAi experiments.

1.3 Related Work

Protein-protein interactions (PPI) play an important role in signaling networks. High-throughput interaction assays like yeast two-hybrid provide vast amount of interaction data that can be tapped in for identification of novel networks.

Scott et al. (2006) proposed several biologically motivated extensions for Color Coding technique to extract signaling pathways from the PPI networks. In their work, three main constraints on protein network data have been applied to extract signaling pathways. First they define a set of relevant proteins to be included in the pathways, then they classify the proteins based on their functionality to order the occurrence of each protein type in the pathway, and finally they extract the pathways based on a rooted tree structure. In the last step, they evaluate the extracted pathways based on the reliability of their protein interactions.

Vinayagam et al. (2011) applied Bayesian learning framework to infer directions of edges in a PPI network and provide insights into the dynamics of known networks. However, the noisy nature of PPI networks necessitates integration with additional biological information.

Path Finder(Bebek and Yang, 2007) is another approach for determining the biologically significant pathways in PPI networks. They train a set of association rules based on the properties of already known signaling pathways and then they recover signaling pathways from PPI data using the trained rules.

In another study, Steffen et al. (2002) proposed NetSearch algorithm to find the signaling pathways by enumerating PPI networks and ranking the extracted results by clustering the genes using k-means algorithm based on their expression data.

Wang et al. (2011) presented CASCADE_SCAN method based on customized steepest descent method to predict signal transduction networks from high-throughput PPI data. In their method, they extract the signaling network with no specific starting and ending proteins. They identify *seed* nodes in the PPI data and try to expand through their neighbors in order to shape the whole signaling network.

ResponseNet(Lan et al., 2011) integrates molecular interaction data with genetic screens and transcriptomic profiling assays to identify a high-probability signaling subnetwork. This method is implemented as a Network Flow optimization problem and aims to maximize the flow between the given source and sink nodes while minimizing the cost of the extracted paths.

Ourfali et al. (2007) attacked the problem of annotating edges in a signaling network. Given a directed PPI network, they assigned signs to each edge indicating whether it is an activation or suppression to obtain paths consistent with the knockout experiments. In this approach they have formulated the problem as Integer programming optimization problem to maximize the expected number of cause-and-effects pair in the extracted networks.

The study by hsiang Yeang et al. (2004) was one of the first attempts to integrate PPI data with RNAi data. Tu et al. (2009) used PPI networks with RNAi screens to identify a more accurate set of genes involved in an expanded insulin signaling network. They used PPI networks to reduce the false positives in RNAi screens, since a large number of genes were targeted. They used functional enrichment to compare the identified genes to the genes in the KEGG database (Kanehisa and Goto, 2000) and reported that they were able to find a more enriched set of core genes.

Kaderali et al. (2009) proposed a probabilistic boolean threshold network approach to construct signaling networks by using data from single knock-down single reporter RNAi screens only. They converted RNAi scores to boolean RNAi observations by statistically postprocessing the data. These boolean RNAi observations are interpreted as constraints on the network topology. We refer to these constraints as RNAi constraints throughout the thesis. Kaderali *et al.* showed that without using additional biological knowledge, the problem is under-determined and the number of network topologies that are consistent with the RNAi constraints grows exponentially with the network size. Their method identifies the most likely networks; however, is limited to networks of small size (less than 10 genes) due to exponential increase of the search space.

Ruths et al. (2007) integrated gene knockout experiments with PPI data. They provided a solution to add new edges for repairing incomplete networks that are not consistent with the gene knockout experiments. In this method, they run the k-shortest path algorithm over the PPI data for every inconsistent knockout pair and evaluated the resulting paths based on their edge values. Finally, they included the paths with the score higher than a specified threshold to the current signaling network. However, they do not deal with false positive interactions in PPI networks.

InfluenceFlow combined PPI data with RNAi data to construct signaling networks as trees (Singh, 2011). InfluenceFlow uses RNAi scores to impose an order on the genes by assuming the influence flows from the genes with low RNAi scores to those with high RNAi scores. However, the tree topology allows for only the integration of a signal flow at the target gene and is not capable of modeling the distribution of the signal to multiple genes.

1.4 Contributions

In this thesis, we address the signaling network reconstruction problem and deal with the problem of inferring the network structure. We assume that the set of genes in the network is known *a priori*. We assume that a reference network (a PPI network or a signaling network from another organism) and RNAi experiment results specific to the studied signaling network are provided. We use the reference network as the starting topology and the RNAi constraints as the guide to the target network that needs to be constructed. We perform edit operations on the reference network, where each edit operation is an edge addition or deletion, to make it consistent with RNAi constraints. We construct a network in the simplest way, to conform RNAi data with minimum number of edge insertions/deletions on the reference network. The main intuition behind this parsimonious approach is that the reference network may not reflect the actual signaling network, but it provides a good skeleton structure to build a network upon. The inconsistency between RNAi constraints and the corresponding PPI network may be due to errors in the RNAi experiments, errors in high-throughput PPI screening, or due to a disease that alters the signaling network. The solution we develop, provides the simplest explanation for such inconsistencies.

It is important to note that, although signaling networks are directed, our formulation is able to use undirected PPI networks as the reference network. We do this by using two directed edges to represent an undirected edge. Deletion of one of these edges during edit operations, in effect, assigns a direction to this interaction.

We show that the signaling network reconstruction problem is NP-complete under our parsimonious formulation. We develop a robust method that guarantees to construct a network. It however has an exponential worst case time complexity. In our experiments, we observe that the proposed method produce results very close to the optimal one. Also, despite the theoretical exponential worst case complexity, in practice, our method can scale up to networks with hundreds of genes easily. Our results also show that the method produce biologically significant networks.

CHAPTER 2

Methods

In this chapter, we first give a formal problem definition and show that the problem we tackle is NP-complete. We then propose a heuristic, named SiNeC, for solving the defined problem. We analyze the running time and optimality of SiNeC. Finally, we propose a couple of extensions for increasing the running time performance of SiNeC.

2.1 Problem Formulation and NP-Completeness Proof

We start by defining the characteristics of RNAi data we consider in this thesis. We focus on single source, single reporter, and single knock-down RNAi screens. Although there are techniques to observe multiple reporter genes, for large scale RNAi screens, typically a single reporter is used. Also, due to combinatorial complexity, only single gene knock-downs are feasible in a large scale experiment. Single source, single reporter, and single knock-down RNAi screens are characterized by three features. The first one is the *receptor gene* which receives the extracellular signal and propagates this signal to others genes in the cell. Sometimes signaling proteins like epidermal growth factor (EGF) appear as the first gene of a network before the receptor gene. The second feature is the *reporter gene* whose expression is measured. This is typically the gene at the downstream of the receptor gene. We denote these two genes with v_s and v_t respectively. The third feature is defined by each of the remaining genes. After knocking down each gene, significant changes at the expression level of v_t implies that the knock-down affects v_t greatly. We call such genes as *critical* for receptor and reporter gene pair of that network. The following definitions formalize the concept of critical genes and establish its connection with the network topology.

Definition 2.1.1. (Simple path) *Given a directed graph $G = (V, E)$ (V and E refer to vertex and edge set), a* simple path *from u to v ($u, v \in V$) is an ordering $< v_1, v_2, \ldots, v_k >$, of a subset of the vertices of G such that $v_1 = u$, $v_k = v$, $(v_i, v_{i+1}) \in E$ and no vertex v_i is repeated $\forall i$, $1 \leq i < k$ in this ordering.*

Definition 2.1.2. (Critical and non-critical genes) *Given a network denoted by the directed graph $G = (V, E)$, a gene $v \in V$ is a* critical gene *in G if there is no simple path from v_s to v_t in G that does not contain v. Otherwise, it is a* non-critical gene.

Notice that knocking down a critical gene disrupts all possible paths from v_s to v_t and thus can affect the expression level of v_t. That of a non-critical gene, however has less effect on v_t as there will be alternative routes to propagate the signal from v_s to v_t. Although the effects of a gene knock-down on v_t is measured in a continuous real domain, following the simplification by other studies (Kaderali et al., 2009), we model these effects as binary values 1 and 0, i.e., affecting and non-affecting, if v_t's relative expression change exceeds or does not exceed a threshold respectively after the knock-down.

Definition 2.1.3. (Consistency of a network) *Assume that we are given a network denoted by the directed graph $G = (V, E)$ with n genes. Also, let us denote the constraints on the genes in V with an instantiation of the vector of binary variables $X = [b_0^1, b_0^2, \ldots, b_0^n]$ (i.e., $\forall i, b_0^i \in \{0, 1\}$. G is* consistent *with X if v_i is a critical gene when $b_0^i = 1$ and non-critical otherwise.*

Definition 2.1.4. (Distance between two networks) *Assume that we are given two networks built on the same set of genes, denoted by the directed graphs $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$. Let us denote the set difference with the "$-$" operator. We define the distance between G_1 and G_2 as*

$$dist(G_1, G_2) = |E_1 - E_2| + |E_2 - E_1|.$$

Figure 2.1 shows an example signaling network. In this example, v_s is type I IFN (gene name: IFNA2) and v_t is luciferase. Luciferase is not a native member of the signaling network; but used as a reporter whose production increases with the disruption of the signaling network. Some nodes in the network are molecular complexes and the interactions forming these complexes are modeled implicitly within the nodes. In this example, IFNAR, STAT1, JAK1 and TYK2 are critical genes whereas STAT2 and IRF9 are not.

Now, we formally define the signaling network reconstruction problem. Assume that we

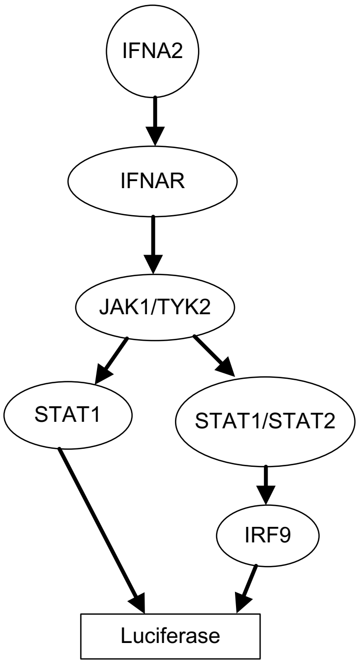


Figure 2.1: Type I IFN stimulated JAK/STAT network (Platanias, 2005).

are given a directed reference interaction network denoted by the directed graph, $G_R = (V_R, E_R)$, with n genes and designated v_s and v_t as receptor and reporter genes. Also, we are provided with a vector of constraints $X = [b_0^1, b_0^2, \ldots, b_0^n]$ that define the critical genes from RNAi experiments. Here, b_0^i is the indicator variable with $b_0^i = 1$ if the gene $v_i \in V_R$ is a critical gene in the network to be constructed. The problem is to find a network $G = (V_R, E)$, which is consistent with the constraints X and the difference dist(G_R, G) is minimum. It is worth noting that if the reference network G_R is consistent with X, then $G = G_R$ is a trivial solution to this problem. As we discuss later, the problem quickly gets challenging when G_R fails to satisfy a subset of the constraints. To establish the complexity of our problem we first define the decision version of the signaling network reconstruction problem and give the definition of an existing NP-complete problem that is used in the NP-completeness proof. The proof follows the definitions.

Problem Definition 2.1.1. (Reference Network Editing For RNAi Compliance (RNERC)) *Given a reference interaction network denoted by $G_R = (V_R, E_R)$ with designated v_s and v_t. A vector of RNAi constraints for genes are provided as $X = [b_0^1, b_0^2, \ldots, b_0^n]$.*
QUESTION: Given a non-negative integer m, is there a $G = (V_R, E)$ with dist(G_R, G) $\leq m$ that is consistent with X?

Theorem 2.1.1. *RNERC is NP-Complete.*

In the RNERC problem, the aim is to construct a graph G' that satisfies a set of constraints X, if for $\forall b_0^i \in X, 1 \leq i \leq n$, v_i is a critical gene in G' for $b_0^i = 1$ and it is a non-critical gene in G' if $b_0^i = 0$.

NP-Completeness Proof. To show the NP-Completeness of RNERC problem, we first prove that RNERC is NP. To prove RNERC is NP, it is sufficient to show that a nondeterministic algorithm needs only guess E' and in polynomial time we can check whether $G' = (V_R, E')$ satisfies X and $|E_R - E'| + |E' - E_R| \leq m$. The second part of checking whether $|E_R - E'| + |E' - E_R| \leq m$ is a simple set operation and can be done in $O(|E'| + |E_R|)$. In order to check whether $G' = (V_R, E')$ satisfies X, we run a Depth First Search (DFS) traversal of G_i' for every constraint $b_0^i \in X$, starting from v_s until we find a simple path to v_t or all the edges $e \in E' - E_{R_{v_i}}$ are visited. Since each edge $e \in E' - E_{v_i}$ is visited only once during DFS, the complexity of this check is $O(|E' - E_{R_{v_i}}|)$ for each $b_0^i \in X$ and $O(|V||E'|)$ in total, which is a polynomial time complexity. Hence RNERC is NP.

Second, we need to show that there is a polynomial time transformation from an NP-complete problem to RNERC. For this, we pick the NP-complete problem HPTP (Garey and Johnson, 1979) and show HPTP $\propto$ RNERC.

Problem Definition 2.1.2. *Hamiltonian Path Between Two Points (HPTP)*
Instance: A directed graph $G = (V, E)$ and vertices $v_s, v_t \in V$.
Question: Does G contain a Hamiltonian path beginning with v_s and ending with v_t? That is, does G contain an ordering $< v_{\pi(1)}, v_{\pi(2)}, ..., v_{\pi(k)} >$, $k = |V|$, of the vertices of G such that $v_{\pi(1)} = v_s$, $v_{\pi(k)} = v_t$, and $(v_{\pi(i)}, v_{\pi(i+1)}) \in E$ for all i, $1 \leq i < k$.

We first define a polynomial time transformation function $f : HPTP \longrightarrow RNERC$, for all instances of HPTP, then prove that this is indeed a transformation by showing that for all instances of HPTP there is a Hamiltonian path from v_s to v_t if and only if there is a G' which satisfies X and $|E_R - E'| + |E' - E_R| \leq m$.

Transformation function f is defined as follows. Given a graph $G = (V, E)$ and vertices $v_s, v_t \in V$ is an arbitrary instance of HPTP. We construct an instance of RNERC as follows.
Step 1: Remove edges $(v_i, v_s), \forall v_i \neq v_s \in V$ from E

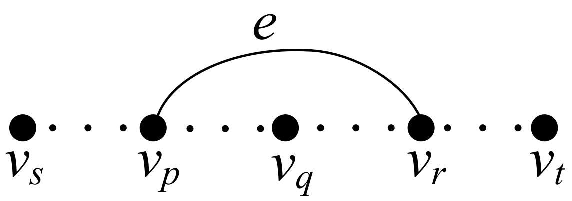


Figure 2.2: Edge e cannot exist in G.

Step 2: Remove edges $(v_t, v_i), \forall v_i \neq v_t \in V$ from E

Let the reduced edge set be E^r. The RNERC instance is $G_R = (V_R, E_R), V_R = V$ and $E_R = E^r$, source gene is v_s and the target gene is v_t and $X = \{1, 1, ..., 1\}$, $|X| = |V| - 2$ and $m = |E^r| - |V| + 1$. In other words, we set all the RNAi constraints to 1 meaning that all regular genes are critical genes and limit the number of differences between E_R and E' to $|E_R| - |V_R| + 1$. This transformation has time complexity O($|V|$) due to Steps 1 and 2, hence it is polynomial time.

To prove that this is a transformation, we need to show that there is an E' with $|E_R - E'| + |E' - E_R| \leq |E_R| - |V_R| + 1$ that satisfies X if and only if G has a Hamiltonian path between v_s and v_t.

It is easy to see that a graph $G' = (V_R, E')$ satisfies X if and only if the graph is solely a linear path, i.e., an ordering, of vertices in V_R where each consecutive pair of vertices is connected by an edge. Any other additional edge e will cause G' not to satisfy X, because removal of a node, such as v_q in Figure S2.2, will not affect the existence of a path from v_s to v_t. A connection will be possible through the edge e between v_p and v_r. Therefore e cannot exist and E' defines a Hamiltonian path from v_s to v_t with no additional edges and $|E'| = |V_R| - 1$. If $E' \subseteq E_R$, in other words if E_R contains a Hamiltonian path between v_s and v_t we only need to remove $m = |E_R| - |V_R| + 1$ edges from E_R to get E'. If there is no such path in E_R, then in addition to removing $|E_R| - |V| + 1$ edges from E_R, we need to add some edges to complete the linear path in E'. Hence the difference will be greater than m and the answer to the RNERC problem will be “No”. It is also easy to see that the graph $G_R = (V_R, E_R)$ contains a Hamiltonian path if and only if the graph $G = (V, E)$ contains a Hamiltonian path, since the only difference between the two graphs are the additional incoming edges to v_s in E and the additional outgoing edges from v_t in E which will not appear in a Hamiltonian path from v_s to v_t in G. Therefore, the RNERC problem instance has a “Yes” answer if and only if G has a Hamiltonian path from v_s to v_t. □

Next, we describe the method we have developed to tackle the signaling network reconstruction problem.

2.2 Signaling Network Constructor (SiNeC)

In a signaling network, a signal is transferred from a receptor gene v_s to a reporter gene v_t through a combination of critical and non-critical genes in at least one simple path. From the definition of critical genes, we know that all the critical genes are on all of these paths. Moreover, the critical genes in all of these paths are visited in the same order. This can be proven by contradiction. If two alternative orderings of two critical genes exist, then there exists a simple path from v_s to v_t that does not visit at least one of them. This is because the other gene appears before that one in one ordering and after in the other ordering. By combining appropriate parts of these two orderings, we can skip one of the two genes and arrive at v_t.

Our proposed method, named *Signaling Network Constructor (SiNeC)*, exploits the above observations. It works in three steps: (i) We first estimate the approximate ordering of the critical genes in the reference network. (ii) We then delete those edges that are in conflict with that order from the reference network. (iii) Finally, we insert the missing edges that are necessary to ensure the flow between consecutive critical genes and the consistency of the remaining (i.e., non-critical) genes in the reference network. The resulting network is guaranteed to be consistent with all the RNAi constraints. We elaborate on each step below.

2.2.1 Step 1: Ordering of the Critical Genes

At this step, SiNeC estimates the order in which a signal that is received at v_s is propagated among the critical genes. SiNeC uses this ordering to create the backbone of the signaling network that it constructs. This problem, in principle, is similar to the topological sorting of the nodes of a graph, which is a well known problem in graph theory. However, the topological sorting problem is defined on directed acyclic graphs; hence, it is not directly applicable for a cyclic undirected PPI network. Signaling networks are often sparse networks. Therefore, techniques that work on sparse graphs are promising to order the critical genes. Numerous methods for sparse graph traversal exist in the literature (see George and Liu (1990)). Sloan (1986) proposed a greedy algorithm in which all the nodes between predefined start and end nodes are assigned a priority value. This algorithm initially determines the priority of each gene using its degree and its distance to the end node. Priority of each gene is a weighted average of its degree and its distance to the end node. The algorithm then removes the node with highest priority in the list and updates the rest of the node priorities by recomputing the priority of each gene in the reduced graph. It continues this process iteratively to find the next node with the highest priority until there is no node left. The running time of the Sloan algorithm is linear in the size of the graph. This makes it an appropriate solution particularly for graphs with large number of nodes. We use this method on the reference network G_R to create a putative ordering of critical genes. The resulting ordering imposes that every path between v_s and v_t should traverse through the critical genes in that order.

Notice that the ordering found at this step may be violated in G_R. In the following steps, we modify G_R to ensure that this ordering holds for all the critical genes.

2.2.2 Step 2: Edge Deletions

Our goal in this step is to determine the minimum number of edges whose deletion makes the reference network consistent with the ordering of the critical genes found at the first step. In other words, we would like to ensure that, for all critical gene pairs (u, v), the signal reaches u before v, if u appears before v in the ordering imposed in Step 1.

Let u and v be two non-consecutive critical genes according to the ordering we impose (with u ordered before v). A network with c critical genes contains $\sum_{k=1}^{c-2} k$ such critical gene pairs. For each such pair (u,v), we first generate all the possible simple paths from u to v through non-critical genes. The number of such paths is exponential in the number of edges. These paths are undesirable as even the presence of one of them is in contrast with the RNAi constraints and the ordering found in Step 1. So, we need to eliminate all of these paths.

Notice that deleting a single edge on a simple path suffices to eliminate that path. Also, notice that an edge can appear on multiple simple paths. Following from these observations, we insert the edges on these paths into a priority queue according to their number of appearances in all the simple paths (i.e., number of simple paths that are eliminated by the removal of that node). Once all the paths are considered, we greedily delete the edge with the highest priority from the queue and the reference network. Notice that, if multiple edges have the same priority, we delete the edge from the queue that does not make any non-critical gene a critical one. If all the candidate edges have the same property, we break the ties randomly. We remove the paths that contain this edge from the path pool and update the priority queue for the edges that are affected by the path removal. We delete all the edges with a priority of zero from the queue. We repeat this selection and update procedure until no edges are left in the priority queue.

At the end of this step, the resulting network has no inconsistency with the ordering of the critical genes.

2.2.3 Step 3: Edge Insertions.

The final step inserts the missing edges in the reference network to guarantee that a signal can propagate from v_s to v_t by visiting the critical genes in the same order as defined in our first step. It also ensures that the resulting network is consistent with all the constraints related to non-critical genes.

Let us denote two consecutive critical genes with u and v, with u appearing before v in the generated ordering. For every such u and v, there must be at least one path from u to v. We insert an edge from u to v if at least one of the following two conditions hold. (i) No path exists between from u to v, or (ii) There is at least one non-critical gene which appears on all the paths between from u to v. The first condition above ensures that the signal can travel from one critical gene to the next. The second one guarantees that there are no critical genes between two consecutive critical genes in the resulting network. Thus, the resulting network is guaranteed to be consistent with all the RNAi constraints.

Throughout its process, SiNeC deletes the smallest number of edges to ensure ordering of critical genes and inserts the smallest number of edges to ensure that the resulting network is consistent with the constraints. Thus, it ensures that the resulting network has smallest possible distance to the reference network with the restriction that critical genes are ordered as in Step 1. The worst case time complexity of SiNeC is exponential in the number of nodes particularly for dense reference networks. This is because Step 2 generates all simple paths between critical gene pairs. However, in practice we do not observe this most of the time as biological networks are sparse (see Section 3.4).

2.3 Analysis of the SiNeC Method

Here, we answer two critical questions. Recall that our goal is to find the network whose topology is closest to the reference network. The first question we answer is how well does SiNeC achieve this goal. Every time we insert/delete an edge to/from the reference network we increase the distance between the reference network and the resulting network by one. Thus, we ideally need to apply the smallest number of edge insertions/deletions. Lemma 2.3.1 states the optimality of SiNeC when the order of the critical genes is fixed.

Second question we discuss is the performance of the SiNeC method. Finding all the paths between pairs of non-consecutive critical genes (Step 2) dominates the overall running time and space complexity. This is because, in the worst scenario, the number of paths between two genes can grow exponentially with the number of edges in the network. The growth rate increases with the density of the edges in the reference network. That said, it is worth mentioning that signaling networks are often sparse and thus the practical performance we observe on real datasets is much better.

2.3.1 Optimality of the SiNeC Method

Lemma 2.3.1. *Given a reference network, RNAi constraints, and a putative critical gene ordering,*

1. *SiNeC deletes the smallest number of edges to ensure ordering of critical genes.*
2. *SiNeC inserts the smallest number of edges to ensure that the resulting network is consistent with the constraints.*

SiNeC Optimality Proof. In order to satisfy the RNAi constraints, two stages of deletion and insertion operations are applied. In the deletion stage, we eliminate the edges that are on the paths between two non-consecutive critical genes. According to the heuristic, an edge with the highest number of occurrences in the inconsistent paths is removed from the queue and the occurrences of the other edges are updated. This process iterates until there is no edge left in the queue. We claim that this greedy approach always results in the minimum number of edge deletions. We convert this problem into the "Activity Selection Problem". Each activity in this problem, has a starting and a finishing time and the activities which do not overlap, are said to be non-conflicting. The greedy method tries to find a maximal set of non-conflicting activities. It selects an activity with the earliest finishing time that does not overlap with the previously chosen activities and it iterates until no non-conflicting activity is left. In the conversion process, we consider every edge as an activity and its number of occurrences as the finishing time of the activity. Since we have no starting time for the activities, we assume that all the starting times are set in a way that does not conflict with the other ones. In every iteration of the method, in order to find the minimal solution set, we choose an activity with latest finishing time and then, updating the finishing time of the other activities and resetting their starting time to keep them non-conflicting with the already chosen ones. The greedy approach to the "Activity Selection Problem" is proved to be optimal and so the deletion stage of the heuristic is also optimal. In the Insertion stage of the heuristic, for every two consecutive pair of critical genes with k paths between them, a non-critical gene should appear in, at most $k - 1$ of the paths. If not, a *single* edge is inserted between the critical genes pair. The same strategy applies when $k = 0$. A single addition/deletion is a minimum operation that can change a given network.

Every deleted edge in the first step, has the highest number of occurrences on the undesired paths between every two non-consecutive critical genes pairs. The same edge can also be on a path between two consecutive critical genes that ensure the function of non-critical genes between the regarding consecutive critical genes. So the deleted edge will enforce one edge insert operation for retaining the non-critical genes properties. In order to overcome this extra addition, we might choose another edge with less priority that results in non-optimal number of delete operations which at least is one more operation than the optimal case. This proves that, although, the deletion and insertion operations are applied separately, the overall distance, between the reference network and the constructed network is the smallest. □

2.4 Analysis of Time Complexity and Extensions

SiNeC algorithm has a worst-case exponential running time in the number of nodes and with the increase in the size of the network, the performance of the algorithm decreases significantly. This problem occurs in the edge deletion step of our algorithm when we find all the available paths between two non-consecutive critical genes. In each unit of time, a single process explores the paths between a given pair sequentially while other pairs wait in the operating system queue to be processed in succession of the previous one. In this section, we propose extensions to improve the performance of the original approach.

2.4.1 P-SiNeC

The waiting time for completion of edge removals increases the running time of the whole algorithm while we do consume all the available resources. Recent advance in multi-core processor technology brings the possibility for executing several tasks concurrently. This results in a shorter running time and unleashing the full power of multi-core CPUs. We have developed a new concurrent variant of the SiNeC algorithm named *P-SiNeC* (Parallel SiNeC). In this method, every pair of non-consecutive critical genes is assigned a separate process that explores the paths simultaneously. Each process updates the number of visited edges in the discovered paths concurrently in the priority queue. This extension to the original algorithm achieved faster running times on the semi-synthetic datasets (See Chapter 3 for details of the datasets). The experiments are conducted using Microsoft .NET Parallel library in a 64-bit Quad Core Intel CPU, clocked at 2.8 GHz with four gigabytes of RAM. Degree of concurrency is set to four. Figure 2.4 shows that P-SiNeC improved the running time for the semi-synthetic networks significantly (i.e. nearly three times faster). However, due to having a similar asymptotic running time with the original SiNeC, it cannot find results for networks with few hundreds of nodes.

2.4.2 L-SiNeC

P-SiNeC improves the running time of the basic SiNeC algorithm drastically, but it still cannot solve the large networks with over 300 genes in a reasonable amount of time due to its exponential running time. Note that SiNeC cannot construct a solution for any instance

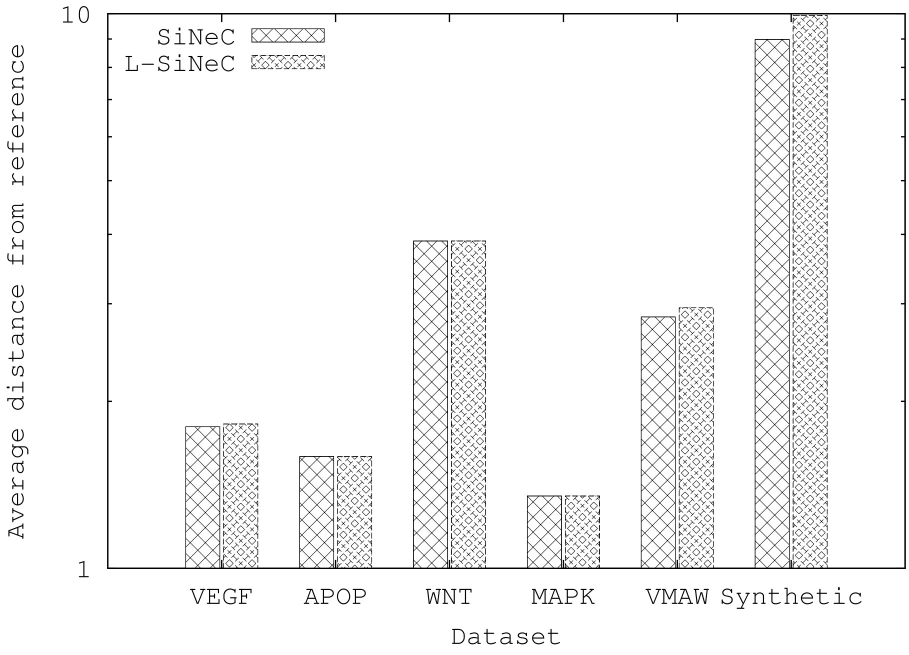


Figure 2.3: Distance from the reference network for SiNeC and L-SiNeC

of the HSA2 network (See Section 3.4). In this customized version of the SiNeC algorithm named *L-SiNeC* (Linear SiNeC), we replaced the exhaustive search in edge deletion step by a polynomial running time heuristic. In the new method, first we find the shortest path between a pair of non-consecutive critical genes then we add the visited edges of the explored path to a HashSet which holds the number of time that an edge is visited. We greedily select an edge with the highest number of occurrences from the HashSet and in case of equality, a random edge is selected. We disrupt the undesired shortest path by removing the selected edge from the HashSet. We continue this procedure by searching for the next shortest path and omitting it using the same method until there is no path left between the given pairs. The worst-case running time of L-SiNeC occurs when the all paths between a pair of nodes are disjoint, in this case, the running time equals the basic SiNeC method. Although the basic SiNeC approach is proved to provide an optimal number of additions and deletions for a given order of the critical genes, the new heuristic shows promising results over the semi-synthetic data without sacrificing the minimality in number of operations noticeably (See Figure 2.3). The average distance from the reference networks in VEGF, VMAW and Synthetic datasets are slightly higher than the SiNeC method. It also improved the running time of HSA drastically and found the results for HSA2 networks in

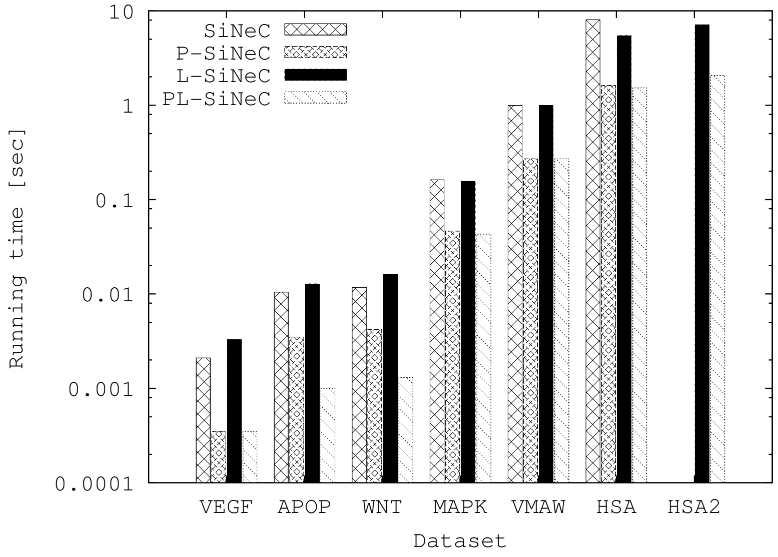


Figure 2.4: Performance comparison among the SiNeC variants

a reasonable amount of time (See Figure 2.4).

2.4.3 PL-SiNeC

Pl-SiNeC(Parallel Linear SiNeC) is a composition of both P-SiNeC and L-SiNeC methods. We added a concurrency layer to the L-SiNeC method to further improve the running time for a given network. In the deletion step of the algorithm, similar to the P-SiNeC method, every two pairs of the non-consecutive critical genes are assigned independent processes but in this extension, every process gets its own copy of shared HashSet for the visited edge counts and after the process finishes execution it updates the shared HashSet for later access by the upcoming processes. Moreover, each process puts a lock on the edge proceeding to be deleted. If another process is waiting for the same specific edge, since the edge is deleted, it will again run the shortest path algorithm to extract the possible next shortest path.

CHAPTER 3

Results

In this chapter, we evaluate the performance and reliability of our method and compare it with the state of the art.

3.1 Datasets

In this section, we describe the datasets we use to evaluate the signaling network reconstruction methods.

3.1.1 Synthetic Datasets

We randomly generated 1,000 reference networks each with nine genes. Each edge in a network is an outcome of a Bernoulli trial with probability 0.5. We also randomly generated RNAi constraints for each of the seven regular genes in the network with $p(b_i^0 = 1) = 0.5$.

3.1.2 Semi-synthetic Datasets

We have generated seven datasets (described in Table 3.1) using the signaling networks of human (*Homo sapiens*) in the KEGG database (Kanehisa and Goto, 2000). Each dataset contains 200 reference networks, each obtained from the actual signaling network using degree preserving edge shuffling method (Milo et al., 2003) with a given mutation rate. We used mutation rates of 0.05, 0.1, 0.2, 0.4 and generated equal number networks for each mutation rate. A mutation rate of r means that $r \times |E|$ edges are toggled to generate a random network.

Table 3.1: Semi-synthetic datasets generated from KEGG human signaling networks. $|V|$ and $|E|$ denote the number of nodes (genes) and edges (interactions). ([1]Oocyte meiosis, Cell cycle, P53, TGF-beta, Calcium) ([2]MTor, Phosphatidylinositol). Each dataset contains 200 reference networks with varying mutation rates.

DataSet	Description	$\|V\|$	$\|E\|$
VEGF	VEGF signaling network	28	32
APOP	Apoptosis signaling network	54	56
WNT	WNT signaling network	60	70
MAPK	MAPK signaling network	127	171
VMAW	Union of VEGF, APOP, WNT and MAPK	212	307
HSA	Union of VMAW and 5 networks[1]	357	505
HSA2	Union of HSA and 2 networks[2]	388	615

3.1.3 Real Dataset

We use the two signaling networks: the type I IFN stimulated JAK/STAT network (Platanias, 2005) and ERBB receptor-regulated G1/S transition network (Sahin et al., 2009a).

3.2 Ability to Minimize Edge Insertion/Deletions

Recall that our aim is to transform the reference network into a new network that is consistent with the RNAi constraints and that the distance to the reference network is minimum. In this experiment, we evaluated how far our proposed method is from the optimal solutions in that regard. To find the true minimum distance, we used an Integer Linear Programming (ILP) formulation that exhaustively searches the search space (Eren and Can, 2012). ILP formulation can provide solutions for small networks with up to nine genes on a standard desktop. It fails to run on larger networks due to exponential growth of the linear constraints.

Figure 3.1 reports the average number of edge insertions/deletions (i.e., distance to the reference network) for the solution reported by the ILP formulation and our method with varying number of critical genes. Several important observations follow from these results. The distance value for SiNeC is close to that for the optimal method, particularly when the number of critical genes is small. This indicates, SiNeC tends to perform better for smaller number of critical genes. This is intuitive as the success of SiNeC depends greatly on the ordering of critical genes. With increasing number of critical genes, the chances of making a mistake

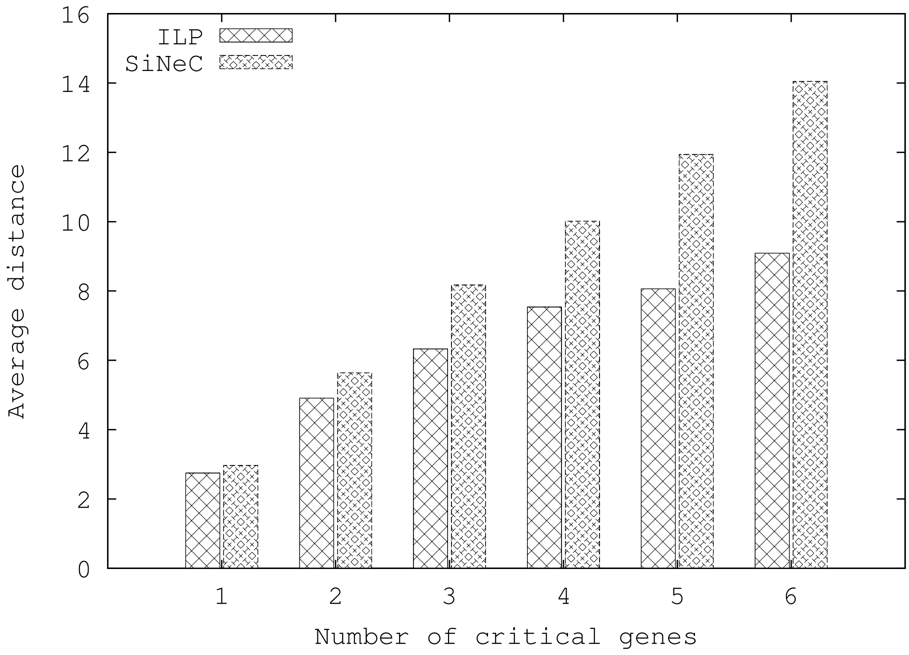


Figure 3.1: Average distance of the constructed network from the reference network increasing different number of critical genes in the reference networks in the synthetic dataset. ILP solver finds the optimal solution.

in their ordering will possibly increase. Finally, the number of edge insertion/deletions are large for such a small dataset. This is because the references in this dataset are very distant from a solution as they are created randomly.

Next, we take a closer look into how our method perform on semi-synthetic datasets. The main goal of this analysis is to answer the following two questions: (i) Most signaling networks are larger than the ones in the synthetic dataset. How does SiNeC perform on larger networks? (ii) Real networks have unique topological characteristics such as degree distributions. The semi-synthetic datasets preserve these characteristics. What is the effect of network topology on the results?

We ran SiNeC on the semi-synthetic datasets, described in Table 3.1. Thus, the networks in this experiment have the same size (same set of genes and same number of interactions) and degree distributions as those provided in KEGG human signaling networks. Furthermore, we know the gap between each reference network and the actual network that it is derived from. This allows us to evaluate the performance of our methods precisely as a function of this gap. We measured the distance between the reference and the constructed network. Figure 3.2 reports the average distance obtained by SiNeC for reference networks created at

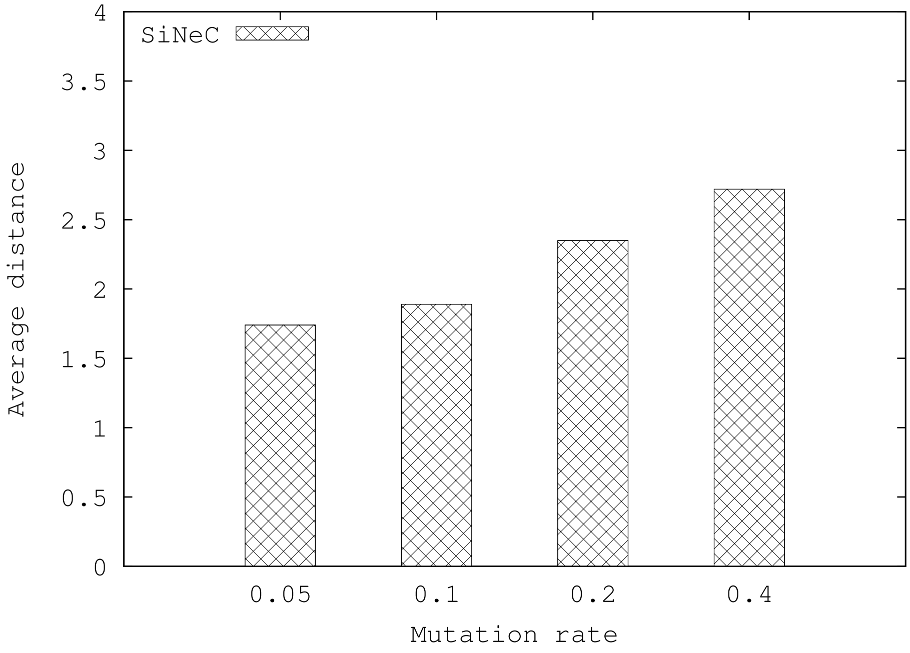


Figure 3.2: Average distance of the constructed network from the reference network for different mutation rates on the semi-synthetic dataset.

different mutation rates.

Our results demonstrate that our method can construct real sized signaling networks from several tens to several hundreds of genes using only a few edge insertions/deletions for all mutation rates (even when mutation rate is 40%). This is very promising as it demonstrates that our methods can construct signaling networks even when small amount of information on the network topology is available.

3.3 Ability to Reconstruct the Network Correctly

The purpose of network construction is to produce the true signaling network topology. Towards this goal, our methods minimize the number of network manipulations on the reference networks. This suggests that the success of our methods depend on the similarity of the topologies of the reference and actual networks. In this section, we evaluate the extent of this dependency.

We ran SiNeC on each of the semi-synthetic datasets and measured their accuracies using precision and recall metrics. Let us denote the actual and constructed networks using G and

G_c.

- (Precision) This measure reports the ratio of the number of interactions common to G and G_c to that of G_c.
- (Recall) This measure reports the ratio of the number of interactions common to G and G_c to that of G.

Clearly, high precision and recall values are preferable. For a given precision and recall value pair, there can be multiple ways to construct the network with that precision and accuracy. One way to distinguish such results, which can also help in identifying the biological relevance of the result, is to take a closer look at where the two networks G and G_c overlap. To do that, we classify each interaction into one of the following three categories.

- *Hot*: An edge (i.e., interaction) is *hot* if both of the genes defining that edge are on at least one path from v_s to v_t.
- *Cold:* An edge is *cold* if (i) at least one of the genes defining that edge cannot be reached from v_s and (ii) v_t cannot be reached from that gene.
- *Warm:* We define the remaining edges as *warm.*

In summary, hot interactions are the most important ones for the receptor and reporter genes used as they define the relation between them. Cold interactions are the least critical ones. We computed precision and recall for the interactions for each of these categories. Results are reported for hot and cold edges in Figure 3.3 and for hot and warm edges in Figure 3.4 respectively. Each point in Figure 3.3 corresponds to one constructed network. There are totally 585 networks constructed for this figure.

Figure 3.3 shows that most of the constructed networks have high precision and recall values for both hot and cold edges (75% of the overall networks have precision and recall values greater than 0.7). This suggests that SiNeC attains high accuracy most of the time even when the reference network deviates as high as 40% from the actual network. Also, we observe that for the same recall value, SiNeC has higher precision values for hot edges (the most important interactions) compared to cold edges. Moreover, the relative precision values for hot edges are greater than that of cold edges (88% of the overall networks have

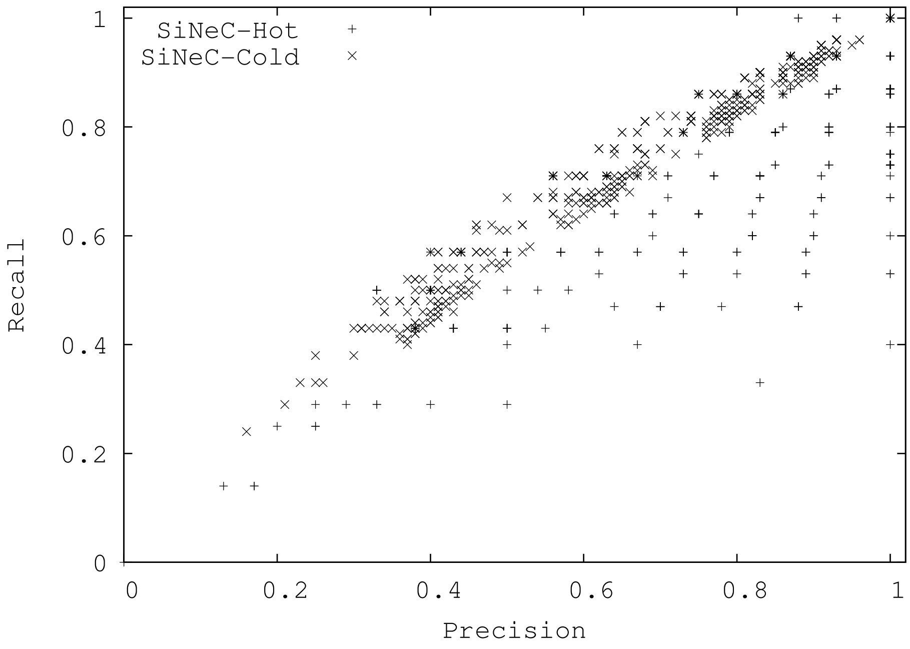


Figure 3.3: Precision and recall values for SiNeC on hot and cold edges

high precision values for hot edges than for cold edges). We observe similar results for the recall values.

3.4 Running Time Performance

A network construction method is practical if it can scale to large networks. In this section, we evaluate the running time performance of the proposed method on synthetic datasets of various sizes. The running times of SiNeC is given in Figure 3.5.

Figure 3.5 shows that SiNeC is very fast for networks with up to around 200 genes. This is a significant contribution as the ILP solution to the network construction problem does not scale beyond nine genes (see Section 3.2). As the network size and density increases, the running time of both methods increase. Recall that SiNeC has exponential running time complexity in the worst case. Our results suggest that we do not observe this for sparse networks even when the network contains few hundreds of genes. On the largest dataset (i.e., HSA2) however SiNeC fails to find a solution in one hour for a single network. This is because this dataset has dense subnetworks which quickly increase the number of simple paths that are considered by this method. We conclude that SiNeC scales well to the networks with large sizes.

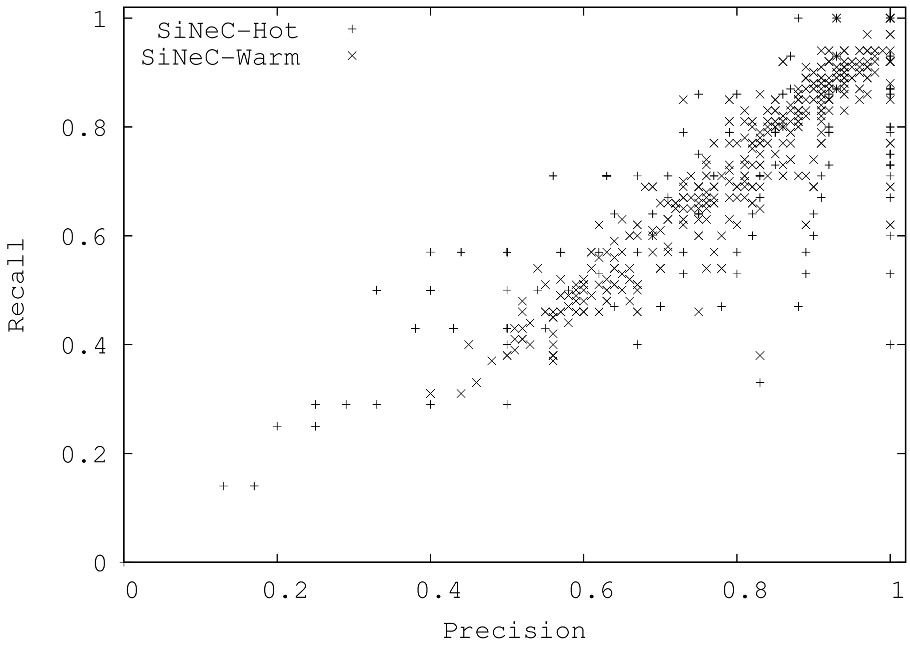


Figure 3.4: Precision versus recall for SiNeC on hot and warm edges

3.5 Significance of Constructed Networks

In this section, we evaluate the biological significance of the constructed networks. One criteria commonly used for understanding the biological relevance of a collection of proteins is the uniformity of the functions of the proteins. Uniformity, here, indicates that the resulting subnetwork collectively performs the same biological function. One measure commonly used to evaluate this is the *functional enrichment* of the resulting network (Shlomi et al., 2006). Computing this quantity relies on the terms annotated to each protein in the network by the Gene Ontology (GO) database (Consortium, 2008).

Briefly, the enrichment reports the minus log likelihood of observing common terms in a set of genes. Thus larger values of this measure mean better enrichment (See (Shlomi et al., 2006) for details on computation of the functional enrichment).

Recall from Section 3.3 that the hot edges are the most important interactions that define the signaling process between the receptor and the reporter genes. We focus on the genes that participate in these interactions and computed their functional enrichment. We limited ourselves to the functions of the reporter gene. In other words, we tested how much this

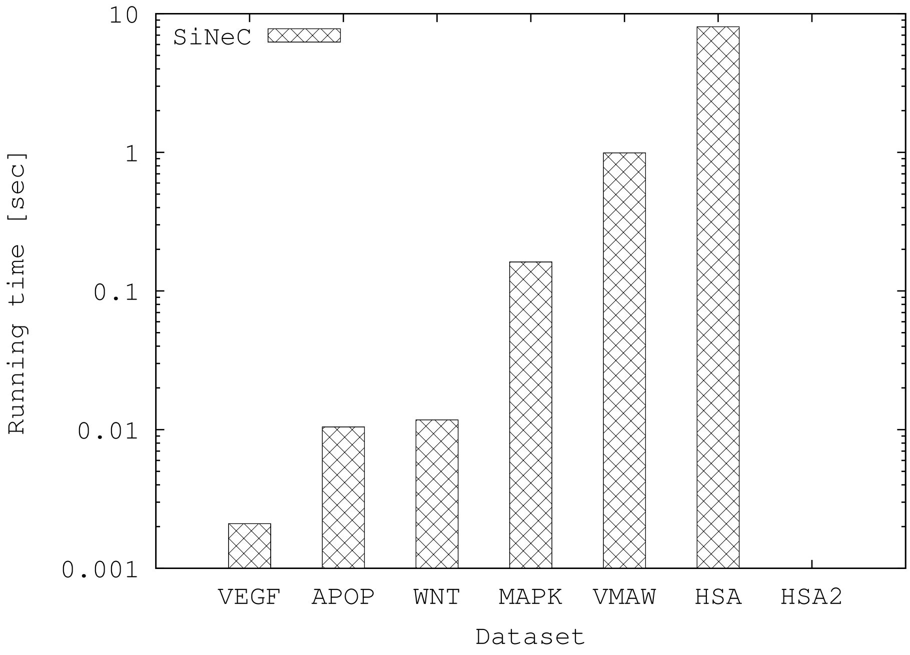


Figure 3.5: Running time for SiNeC. The time measurements are in seconds. SiNeC cannot find a solution in one hour for a single network in the largest dataset (HSA2).

path is enriched in the functions of the reporter gene. We tested this on the semi-synthetic dataset.

The functional enrichment values for the original network in semi-synthetic datasets and the networks constructed by SiNeC (according to various mutation rates and datasets) are given in Table 3.2. We observe that the constructed networks have similar functional enrichment values as those of the original network. In other words, if the same set of genes in the original network is functionally enriched, those in the constructed networks are also functionally enriched, and vice versa. Additionally, for the enriched networks (WNT and VMWA) 100% and in the overall 80% of the constructed networks have enrichment values within one standard deviation of the enrichment of the original network. This implies that the enrichment results of SiNeC are stable and does not change greatly as the reference network changes.

Table 3.2: Functional enrichment of SiNeC on KEGG datasets. The numbers when mutation rate = 0 refer to the functional enrichment value of the original network.

	Mutation rate				
Dataset	0	0.05	0.1	0.2	0.4
VEGF	1.02	1.45	1.39	1.81	1.84
APOP	0.94	0.79	0.74	0.69	0.68
WNT	2.59	2.51	3.09	3.18	3.47
MAPK	1.42	1.71	2.71	3.42	3.50
VMAW	8.98	8.52	8.79	9.46	8.71

Table 3.3: Precision and recall values of the four signaling networks constructed using SiNeC and Ruths et al. (2007).

Dataset	SiNeC		Ruths et al. (2007)	
	Precision	Recall	Precision	Recall
VEGF	0.82	0.93	0.75	1
APOP	0.96	1	1	1
WNT	0.96	1	0.90	1
MAPK	0.99	1	0.97	1

3.6 Comparison with the State of the Art

In this section, we compare SiNeC with various state of the art network construction algorithms.

In our first experiment, we compare SiNeC with Ruths et al. (2007) as this is the closest method to ours in spirit. Recall that this method alters a given network to make it consistent with the RNAi constraints. However, unlike our method, it only allows insertion of new interactions to alter the network topology. We ran both methods to construct four signaling networks, namely VEGF, APOP, WNT and MAPK. We used the corresponding signaling networks in KEGG as the gold standard for the network to be constructed. We obtained the reference networks from the PPI networks of the String database (Szklarczyk et al., 2011). We constructed each signaling network from its reference network using both methods independently. We then computed their precision and recall values. Table 3.3 summarizes the results. We observe that both methods have high precision and recall values. In three out of four datasets SiNeC has higher precision. In only one dataset Ruths *et al.* has higher recall. Even in that case SiNeC has near perfect recall. The high recall values of Ruths *et al.* is a natural outcome of the fact that their method never removed an interaction. This feature however has the drawback of reduced precision values which we observe in our

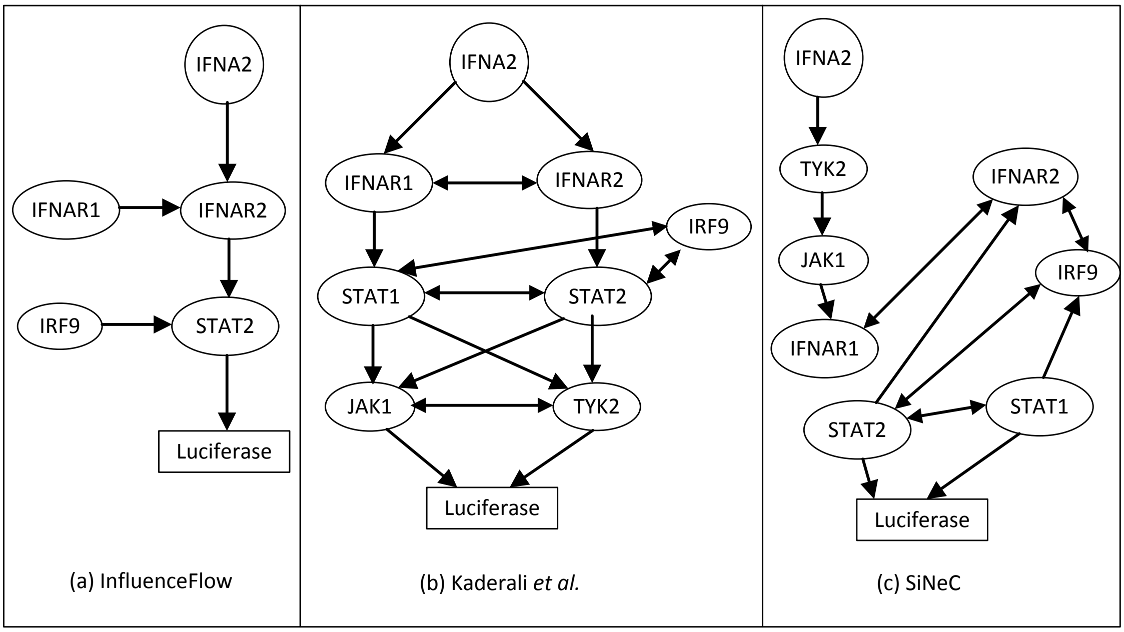


Figure 3.6: The results of the compared methods on the construction of the type I IFN stimulated JAK/STAT network

experiments. Thus, the results suggest that SiNeC performs near perfect in terms of both precision and recall.

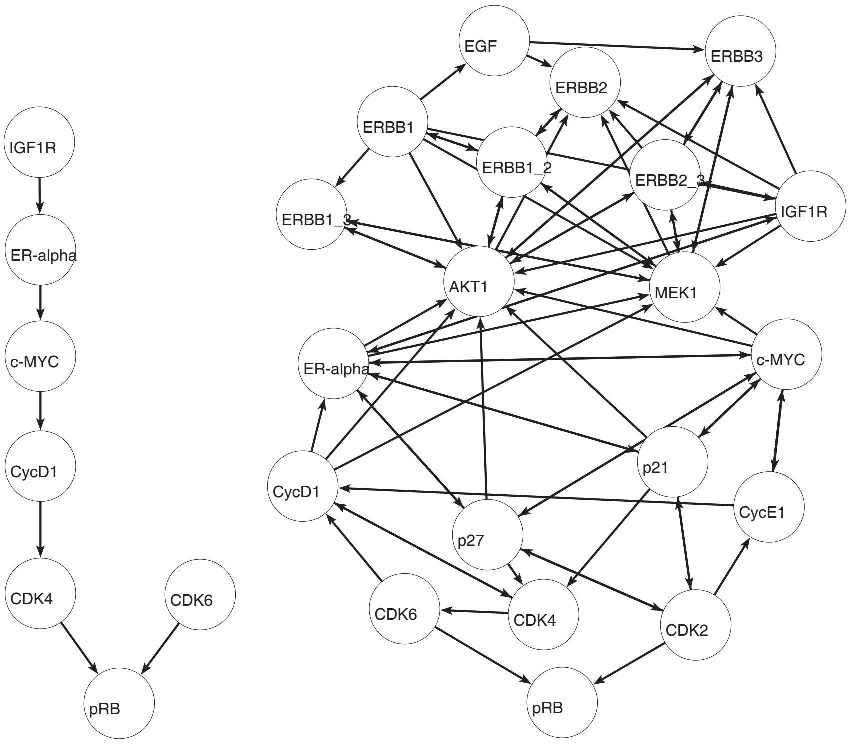
Next, we compare SiNeC with InfluenceFlow (Singh, 2011) and Kaderali *et al.*'s method (Kaderali et al., 2009) on two real datasets (JAK/STAT and the ERBB networks) described at the beginning of Section 3.1.

The JAK/STAT network is a nine node network including the signaling protein Interferon alpha-2 (encoded by the IFNA2 gene) and the reporter protein luciferase. Luciferase is not a native member of the signaling network; but used as a reporter gene whose production increases with the disruption of the signaling network. According to the RNAi screens performed by Kaderali et al. (2009), six of the regular genes are critical genes (IFNAR1, IFNAR2, JAK1, TYK2, STAT2, and IRF9) and the gene STAT1 is a non-critical gene. Figure 1 in their paper shows the true signaling network as given by Platanias (2005). We retrieved protein-protein interactions involving these genes from the EBI IntAct database (Kerrien et al., 2012). For running InfluenceFlow, we retrieved the raw RNAi scores from Lars Kaderali and we included the STAT1, STAT2, and IRF9 genes in the core cascade of the JAK/STAT network. For Kaderali *et al.*'s method we used the highest probability network reported in their paper.

Figure 3.6 shows the results of each method. Since STAT1 and JAK2 are not RNAi hits, InfluenceFlow does not include these genes in the resulting network. In addition, JAK1 and TYK2 are not included in the resulting network because there are no interactions in IntAct that connect these genes to the other genes in the network. InfluenceFlow cannot account for missing edges; hence, excludes such genes from the resulting network. Edges of the original network are pruned further, in order to give a spanning tree as the output. Kaderali *et al.* combines some of the nodes to single nodes as complexes in order to reduce the complexity of the problem. For comparison purposes, we report these complexes as separate nodes in Figure 3.6. Kaderali *et al.* use a likelihood function to search the space of all possible topologies. The most probable topology provided by Kaderali *et al.* misplaces JAK1 and TYK2 as the immediate genes before the reporter, although these are the genes that interact with the type I IFN receptors. SiNeC could discover some of the key parts of the topology correctly. For instance, it identifies JAK1, TYK, IFNAR1 and IFNAR2 as critical genes although their relative positions are switched with that given by Platanias (2005). Our results show that SiNeC and Kaderali *et al.*'s method could both reconstruct some of the key parts of the topology correctly, while InfluenceFlow failed to include some critical genes such as JAK1 and TYK2 in the final results due to incompleteness of the reference network.

Finally, we compare the results of the methods on the ERBB receptor-regulated G1/S transition network (Sahin et al., 2009a). ERBB network contains 17 genes and the three EGF receptors ERBB1, ERBB2, and ERBB3 form three heterodimer complexes which can be inserted into the network as additional nodes (ERBB1_2, ERBB1_3, and ERBB2_3). We ran InfluenceFlow and SiNeC on the PPI network collected from the literature by Sahin et al. (2009a). We ignored the directions of the edges. Kaderali *et al.*'s method did not produce a result for this network in 24 hours. We used pRB, CDK4, CDK6, and CDK2 as the core cascade for InfluenceFlow. The critical genes as reported by the RNAi screens are ERBB1, ERBB1_2, IGF1R, ER-alpha, c-MYC, CyclinD1, CyclinE1, CDK4, and CDK6. For running InfluenceFlow, we used the RNAi scores provided by Sahin *et al.*. Both SiNeC and InfluenceFlow reconstructed the ERRB signaling network within a second. The results are depicted in Figure 3.7.

Figure 3.7 shows that only 7 out of 20 genes are included in the InfluenceFlow result. InfluenceFlow excluded the genes that are not RNAi hits and the genes with no paths to the target gene. Furthermore, the highly interconnected and parallel nature of the network can-



(a) InfluenceFlow (b) SiNeC

Figure 3.7: The results of InfluenceFlow and SiNeC on the ERBB receptor-regulated G1/S transition network

not be effectively modeled by a tree. While SiNeC was able to provide directions for some of the PPI network edges, single knock-down RNAi constraints are not sufficient to uniquely determine the whole network. However, we can conclude that SiNeC is able to provide a better coverage of the true ERBB signaling network compared to InfluenceFlow.

CHAPTER 4

Discussion and Conclusions

In this thesis, we formulated the problem of signaling network inference as a reference network editing problem by integrating molecular interaction data with RNAi data. We show that this formulation is NP-complete and proposed SiNeC for construction of signaling networks involving hundreds of genes with minimum sacrifice in optimality. In this approach, first we estimated the order of the critical genes in the pathway by investigating the topological structure of the reference network using Sloan's algorithm. This ordering is crucial in our problem formulation since there should be no path between non-consecutive critical genes. In the next step, we determined and removed the edges that are in conflict with the given order of the critical genes and we proved that the count of the deleted genes is minimum for a given order. In the final step of this approach, we inserted the missing edges between consecutive critical genes in order to ensure the existing a path between them. This method is examined on semi-synthetic networks with a couple of hundred nodes and it generated the modified networks that are compatible with RNAi hits in a reasonable amount of time. We also proposed three new variants of the SiNeC algorithm for boosting performance of the method on larger networks. P-SiNeC parallelizes the exhaustive deletion stage of the algorithm by exploring the paths concurrently. L-SiNeC introduces a linear average case running time heuristic in determining the candidate edges for deletion and finally PL-SiNeC applies a concurrency layer to the L-SiNeC method. Although the precision and the accuracy of the resulting networks are significantly high, it would be beneficial to discuss the restrictions of our problem definition and the future improvements of our method. Signaling networks usually have multiple receptor and readout genes. Although, considering the single receptor and readout genes is a valid biological assumption, the future research should be focused on integration of all the input and readout genes into the problem definition. Moreover, binary identification of the critical genes imposes the presence of them in all the paths from the receptor to the readout gene. However, due to the inaccuracy of the RNAi knockdown experiments, the hit list of critical genes can be erroneous. Therefore, a more flexible definition of critical genes and their role in signaling networks can be a future research direction.

The PPI networks in our problem definition are un-weighted graphs and every edge has an equal chance to be added to or removed from the reference network. However, some PPI datasets contain confidence weighted edges indicating the likelihood of the occurrence of an interaction between the given pair of the genes. Integrating these edge scores would improve the edge deletion phase of this method and would results in more biologically accurate networks.

Finally, we use the Sloan's algorithm, a graph topological approach, for ordering the critical genes in the network. This method determines this order by simply looking at the structure of the physical network without considering their biological functionality. However, the order of the critical genes affects the quality of resulting networks significantly and different orderings can produce completely different networks. As the future work, biological knowledge such as the known functions of the genes can be integrated into determination of this ordering to make the resulting network more biologically meaningful.

Bibliography

Gurkan Bebek and Jiong Yang. Pathfinder: mining signal transduction pathway segments from protein-protein interaction networks. *BMC Bioinformatics*, 8(1):335, 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-335. URL `http://www.biomedcentral.com/1471-2105/8/335`.

Jeremy M. Berg, John L. Tymoczko, and Lubert Stryer. *Biochemistry, Fifth Edition: International Version (hardcover)*. W. H. Freeman, fifth edition edition, February 2002. ISBN 0716746840. URL `http://www.worldcat.org/isbn/0716746840`.

The Gene Ontology Consortium. The gene ontology project in 2008. *Nucleic Acids Research*, 36(suppl 1):D440–D444, 2008. doi: 10.1093/nar/gkm883. URL `http://nar.oxfordjournals.org/content/36/suppl_1/D440.abstract`.

Ouyky Eren and Tolga Can. Signaling pathway reconstruction from rnai data and ppi networks using linear programming. Manuscript in preparation, 2012.

Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. Freeman, 1979. ISBN 0716710447.

Alan George and Joseph W. H. Liu. *Computer Solutions of Large Sparse Positive Definite Systems*. Prentice-Hall Series in Computational Mathematics. Prentice-Hall, 1990.

V. Helms. *Principles of Computational Cell Biology: From Protein Complexes to Cellular Networks*. Wiley-VCH, 2008. ISBN 9783527315550. URL `http://books.google.com.tr/books?id=-Tavvybv5UwC`.

Chen hsiang Yeang, Trey Ideker, and Tommi Jaakkola. Physical network models. *Journal of Computational Biology*, 11:243–262, 2004.

Lars Kaderali, Eva Dazert, Ulf Zeuge, Michael Frese, and Ralf Bartenschlager. Reconstructing signaling pathways from RNAi data using probabilistic Boolean threshold networks. *Bioinformatics*, 25(17):2229–2235, 2009. doi: 10.1093/bioinformatics/btp375. URL `http://bioinformatics.oxfordjournals.org/content/25/17/2229.abstract`.

Minoru Kanehisa and Susumu Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000. doi: 10.1093/nar/28.1.27. URL http://nar.oxfordjournals.org/content/28/1/27.abstract.

Samuel Kerrien, Bruno Aranda, and et al. The intact molecular interaction database in 2012. *Nucleic Acids Research*, 40(D1):D841–D846, 2012. doi: 10.1093/nar/gkr1088. URL http://nar.oxfordjournals.org/content/40/D1/D841.abstract.

Alex Lan, Ilan Y. Smoly, Guy Rapaport, Susan Lindquist, Ernest Fraenkel, and Esti Yeger-Lotem. Responsenet: revealing signaling and regulatory networks linking genetic and transcriptomic screening data. *Nucleic Acids Research*, 2011. doi: 10.1093/nar/gkr359. URL http://nar.oxfordjournals.org/content/early/2011/05/14/nar.gkr359.abstract.

R. Milo, N. Kashtan, S. Itzkovitz, M. E. J. Newman, and U. Alon. On the uniform generation of random graphs with prescribed degree sequences, 2003. URL http://www.citebase.org/abstract?id=oai:arXiv.org:cond-mat/0312028.

Oved Ourfali, Tomer Shlomi, Trey Ideker, Eytan Ruppin, and Roded Sharan. Spine: a framework for signaling-regulatory pathway inference from cause-effect experiments. *Bioinformatics*, 23(13):i359–i366, 2007. doi: 10.1093/bioinformatics/btm170. URL http://bioinformatics.oxfordjournals.org/content/23/13/i359.abstract.

Leonidas C. Platanias. Mechanisms of type-i- and type-ii-interferon-mediated signalling. *Nature Reviews Immunology*, 5:375–386, 2005.

Derek Ruths, Jen-Te Tseng, Luay Nakhleh, and Prahlad T. Ram. De novo signaling pathway predictions based on protein-protein interaction, targeted therapy and protein microarray analysis. In *Proceedings of the joint 2006 satellite conference on Systems biology and computational proteomics*, RECOMB'06, pages 108–118, Berlin, Heidelberg, 2007. Springer-Verlag. ISBN 978-3-540-73059-0. URL http://dl.acm.org/citation.cfm?id=1768782.1768790.

Ozgur Sahin, Holger Frohlich, and et al. Modeling ERBB receptor-regulated G1/S transition to find novel targets for de novo trastuzumab resistance. *BMC Systems Biology*, 3(1):1, 2009a. ISSN 1752-0509. doi: 10.1186/1752-0509-3-1. URL http://www.biomedcentral.com/1752-0509/3/1.

Ozgur Sahin, Holger Frohlich, Christian Lobke, Ulrike Korf, Sara Burmester, Meher Majety, Jens Mattern, Ingo Schupp, Claudine Chaouiya, Denis Thieffry, Annemarie Poustka, Stefan Wiemann, Tim Beissbarth, and Dorit Arlt. Modeling ERBB receptor-regulated G1/S transition to find novel targets for de novo trastuzumab resistance. *BMC Systems Biology*, 3(1):1, 2009b. ISSN 1752-0509. doi: 10.1186/1752-0509-3-1. URL http://www.biomedcentral.com/1752-0509/3/1.

Jacob Scott, Trey Ideker, Richard M. Karp, and Roded Sharan. Efficient algorithms for detecting signaling pathways in protein interaction networks. *Journal of computational biology : a journal of computational molecular cell biology*, 13 (2):133–144, March 2006. ISSN 1066-5277. doi: 10.1089/cmb.2006.13.133. URL http://dx.doi.org/10.1089/cmb.2006.13.133.

Tomer Shlomi, Daniel Segal, Eytan Ruppin, and Roded Sharan. Qpath: a method for querying pathways in a protein-protein interaction network. *BMC Bioinformatics*, 7(1):199, 2006. ISSN 1471-2105. doi: 10.1186/1471-2105-7-199. URL http://www.biomedcentral.com/1471-2105/7/199.

Rohit Singh. *Algorithms for the Analysis of Protein Interaction Networks*. PhD thesis, MIT, 2011. Chapter 6: Influence Flow: Integration of PPI and RNAi Data.

S. W Sloan. An algorithm for profile and wavefront reduction of sparse matrices. *International Journal for Numerical Methods in Engineering*, 23:239–251, 1986.

Martin Steffen, Allegra Petti, John Aach, Patrik D'haeseleer, and George Church. Automated modelling of signal transduction networks. *BMC Bioinformatics*, 3(1):34, 2002. ISSN 1471-2105. doi: 10.1186/1471-2105-3-34. URL http://www.biomedcentral.com/1471-2105/3/34.

Damian Szklarczyk, Andrea Franceschini, Michael Kuhn, Milan Simonovic, Alexander Roth, Pablo Minguez, Tobias Doerks, Manuel Stark, Jean Muller, Peer Bork, Lars J. Jensen, and Christian von Mering. The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research*, 39(suppl 1):D561–D568, 2011. doi: 10.1093/nar/gkq973. URL http://nar.oxfordjournals.org/content/39/suppl_1/D561.abstract.

Zhidong Tu, Carmen Argmann, Kenny K. Wong, Lyndon J. Mitnaul, Stephen Edwards, Iliana C. Sach, Jun Zhu, and Eric E. Schadt. Integrating sirna and protein-protein interaction data to identify an expanded insulin signaling network. *Genome Research*, 19(6):1057–1067, 2009. doi: 10.1101/gr.087890.108. URL `http://genome.cshlp.org/content/19/6/1057.abstract`.

Arunachalam Vinayagam, Ulrich Stelzl, Raphaele Foulle, Stephanie Plassmann, Martina Zenkner, Jan Timm, Heike E. Assmus, Miguel A. Andrade-Navarro, and Erich E. Wanker. A directed protein interaction network for investigating intracellular signal transduction. *Sci. Signal.*, 4(189):rs8, 2011. doi: 10.1126/scisignal.2001699. URL `http://stke.sciencemag.org/cgi/content/abstract/sigtrans;4/189/rs8`.

Kai Wang, Fuyan Hu, Kejia Xu, Hua Cheng, Meng Jiang, Ruili Feng, Jing Li, and Tieqiao Wen. Cascade_scan: mining signal transduction network from high-throughput data based on steepest descent method. *BMC Bioinformatics*, 12(1):164, 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-164. URL `http://www.biomedcentral.com/1471-2105/12/164`.

X.D. Zhang. *Optimal High-Throughput Screening: Practical Experimental Design and Data Analysis for Genome-Scale RNAi Research*. Cambridge University Press, 2011. ISBN 9780521734448. URL `http://books.google.com.tr/books?id=CSxUAZo6eZEC`.