Network Discovery in Equilibrium-state and Dynamic Data: Applications to Phosphoproteomics and Kinetics

by
Yan Zhang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2012

Doctoral Committee:

Professor Philip C. Andrews, Chair
Associate Professor Anuj Kumar
Professor George Michailidis
Associate Professor Alexey Nesvizhskii
Professor Gilbert S. Omenn
Associate Professor Santiago D. Schnell

DEDICATION

To my family

ACKNOWLEDGEMENTS

First, I want to thank my advisor, Dr. Phil Andrews, for his long-term guidance, strong support and encouragement throughout my doctoral studies. He is a rich source of ideas, experience and wise remarks, and is always open-minded. He has led me into deep investigation of the interesting Proteomics, helped me to create collaborations, and supported me to explore the scientific problems that I want to give a try. He is a strong backup for me to overcome problems in my life of a scientist-to-be.

I want to thank Dr. Santiago Schnell, who has given me great help and advice over the past years. Without his help, the time-course project would not be possible. He is a really wise and patient advisor, who is always ready to help. I really appreciate him for giving me the opportunity to work on the project, to collaborate with the lab members, and to use the servers. He also gave me precious advice on career pursuit.

I want to thank my collaborators, Dr. Anuj Kumar, Christian Shively, and Dr. Hye Kyong Kweon. Without them, the phosphoproteomics bioinformatics project would not have gone so well. Dr. Anuj Kumar serves on my committee too. He brought the interesting vision of yeast filamentous growth to me. Dr. Hye Kyong Kweon performed extensive mass spectrometry experiments. Christian cultivated yeast babies, and did experimental validation for me, so that my paper is much stronger than before. I anticipate our papers to come out soon. I would also like to thank Dr. Márcio Mourão, who has answered all my questions about MIKANA, and given me useful suggestions. I have learnt a lot about mathematical modeling from him. Dr. Alla Karnovsky is an expert in metabolomics. She is a great teacher to work with. I have learnt a lot from her. Thanks to her too.

I want to thank my committee members, Dr. Gil Omenn, Dr. Alexey Nesvizhskii and Dr. George Michailidis, for all the advice and help they have given to me. I also appreciate Dr. Gil Omenn's caring about my research and community activities, and support for my career plan.

I would like to thank all the current and former lab members in Andrews Lab and Schnell Lab: Dr. Hye Kyong Kweon, Chunchao Zhang, Billy Clifford-Nunn, Dr. Angela Walker, Donna Veine, Dr. Panagiotis Papoulias, Dr. Eric Simon, Dr. Márcio A. Mourão, Erin Shellman, Michelle Wynn, Daniel DeWoskin, Dr. Conner Sandefur, Firas Midani and Caroline Adams. I thank them for the discussions, suggestions, and encouragement during my thesis studies.

Julia Eussen, Dr. Margit Burmeister and Dr. Dan Burns deserve my sincere thanks also. They have helped me go through various applications and paperwork and fortunately I got the competitive Rackham International Student Fellowship funded. They take charge of all the miscellaneous things and thus have made my life much easier. I would also thank Dr. Rajasree Menon, Dr. Jeff de Wet, Dr. David States, Dr. Yongqun Oliver He and Dr. Brian Athey at CCMB, Jing Gao at the Broad Institute, and Dr. Steve Qin at Emory University. I got to know them when I first came to the department. They are my teachers and friends. Dr. Rajasree Menon has also given me very important advice about how to write a paper well. Thanks to Dr. David States, I started to think about what my role model is. And now I have a preliminary answer.

Finally, special thanks to all my friends and family. They have supported me to accomplish my final mission as a student. My parents have been my friends and long-distance mentors, who have helped me to overcome various difficulties in my life and shared my joys as well. My husband, a new doctor of philosophy, has been quite understanding and very supportive, welcoming and encouraging me, a doctor-to-be.

I am really fortunate to have met so many wonderful people during the past years. Without their support and help, the accomplishment of this thesis is impossible. Scientific life is risky, and exciting. I have experienced a wonderful life having their company.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

Chapter I

Introduction

1.1 Proteomics and phosphoproteomics, and bioinformatics challenges

Proteomics is defined as the large-scale study of all the proteins in an organism, primarily focusing on their sequences (including splice variants), structures, localizations, abundances, post-translational modifications, and biomolecular interactions [1–6], with the goal of understanding protein functions and providing basic datasets for Systems Biology. Phosphoproteomics is one branch of proteomics, in which researchers focus on the proteins which contain the post-translational modification called phosphorylation. Phosphoproteomics is a very active area of research, since phosphorylation plays an ubiquitous and important role in controlling biological processes [4,7–13] and as such, directly impacts the functional state of proteins. Dysregulation of these processes as a result of lesions in signal transduction pathways is a major factor in the emergence of many diseases, including cancers, diabetes, and cardiovascular diseases.

In the past decade, proteomics quantification strategies have evolved from classical gel-based methods to mass spectrometry (MS)-based methods, which have enabled high-throughput global studies of proteomes, including phosphoproteomes [7,11,14–17]. The avalanche of proteomics data has brought about great challenges in bioinformatics analysis, including (1) statistical experimental design, in order to avoid experimental bias and optimize quantification efficiency [4,18]; (2) protein/peptide identification and quantification, including the topics of data reduction, abundance estimation and normalization, false discovery rate control, and so on [15,19–23]; (3)

post-identification analysis, in order to identify the characteristics of the peptides and understand protein functions; (4) dealing with the missing value issue which is haunting proteomics.

Efficient and accurate protein/peptide identification and quantification is crucial for proteomics and has been an ongoing focus in computational and statistical researches in proteomics. Different algorithms and software packages, such as MaxQuant [24], ProteinProphet [25], IDPicker [26], and many commercial products have been developed to fulfill this task for various experimental platforms.

In post-identification analysis, some common machine learning methods have been used, such as hierarchical clustering or *k*-means clustering to categorize homogeneous protein/peptide profiles into groups [4,16], which is followed by functional analysis based on Gene Ontology [27], KEGG pathways [28] and other functional terms. Identified proteins can also be mapped to known pathway networks, such as KEGG pathways, to be visualized in a functional context [16]. Due to the instrumental sampling limitations, incomplete measurement and small overlap between runs is a common issue haunting proteome data [1,14] and these issues are exacerbated by the high cost of data acquisition. I need to at least partially compensate for these issues through effective data analysis which could provide a more objective analysis of data than the informed but subjective approaches currently common for interpretation of incomplete datasets.

Due to the sampling limitation and resolution of mass spectrometers, tandem mass spectra of complex mixtures results in poor overlap of protein/peptides identified among samples [14,29]. Possible technical solution is to carry out a relatively large number of replicate experiments, however, it is time consuming and often economically impractical for large-scale projects. For this reason, a significant number of missing values exist in these datasets, which can obscure bioinformatics analysis. In this thesis (specifically Chapter 2), we develop methods to partially compensate for the missing data issue. Our approach tolerate missing values and take use of all the available values.

It has been proven evident to pick out significantly enriched functions, and identify a number of reliable candidate proteins for further validation.

1.2 Aims of the dissertation

1.2.1 Aims and hypothesis

Proteomics and phosphoproteomics are the major subjects of this thesis research. Their datasets have some similar features but also some significant differences. The general objective of my dissertation has been to develop and apply statistical, mathematical and computational methods to analyze and interpret proteome and phosphoproteome data focusing on post-identification analysis. While conventional cluster analysis and functional enrichment analysis are useful tools, a more integrative and extensive analysis is necessary to fully utilize the data for knowledge discovery. Our long term goal is to not only mapping identified proteins to known pathways, but to also infer new knowledge of those biological networks that can form the basis of further experiments. The specific outcomes of my thesis work are to (1) reveal the characteristics and structures of biological networks, such as signaling networks and kinetic networks, (2) discover predictive biomarkers, such as differentially phosphorylated proteins, and (3) find the relationships of changes in protein expression and protein modification to their cellular functions. Our hypothesis is that biological network models can be learned from the data, and a physiologically accurate network model can reflect *in silico* the regulation mechanism of the real biological network. We expect these models to suggest directions for further experimental studies.

1.2.2 Outline

This thesis contains five chapters.

Chapter 1, the Introduction, provides background introduction on proteomics and phosphoproteomics, and states the aims and focus of this thesis. Since biological network discovery is a major component in this thesis, Chapter 1 also gives a review of the general methods for network discovery. Not all of the methods described have been

successfully applied in proteomics. Their applications, advantages and disadvantages are summarized.

Chapter 2 presents a new comprehensive quantitative analysis pipeline for systematic network discovery from equilibrium-state interventional phosphorylation data. The purpose is to identify key proteins in specific pathways, discovering protein-protein relationships, and infer components of the signaling network that can be further investigated in experiments. We have also made an effort to partially compensate for the missing value issue which is a major problem in the field. We were able to successfully apply our pipeline to a series of interventional experiments identifying phosphorylation events underlying the transition from budding growth to filamentous growth in *Saccharomyces cerevisiae* strains.

Apart from investigating equilibrium-state proteomics data, I have also turned my attention to time-course biochemical data and we expect the methods developed and applied to be applicable to time-course phosphoproteome analyses in future studies. Time-course data is a richer source of information compared to equilibrium-state data and a broader range of tools can be applied to these datasets. My time-course studies are covered in Chapters 3 and 4.

Chapter 3 presents a computational method which integrates mathematical modeling of biochemical networks (*e.g.* using differential equations) with statistical methods (*e.g.* penalized regression) to infer kinetic network structure and fit reaction parameters from time-course data. It maintains the three-component framework of the previous MIKANA (Method to Infer Kinetics And Network Architecture) [30], which consists of optimization, parameter fitting and design matrix generation and made improvement or extension to each component. Non-negative LASSO and non-linear parameter fitting are applied instead of pure linear methods, and the design matrix generation algorithm is improved as well to enhance the structure prediction and parameter fitting, and allow the algorithm to now tackle networks with oscillatory behaviors.

Chapter 4 presents a Bayesian method integrated with time-delay detection to infer signaling networks from time-course data, such as the phosphorylation dynamics data in response to specific stimuli. The method was designed to fully utilize the "delay effect" between upstream and downstream proteins in time-course data, which facilitates discovery of causal influences. Examples are given to demonstrate the application and demonstrate its validity.

Chapter 5 summarizes the projects, and gives perspectives of future research directions and broader applications.

1.3 Proteomics experiments and data types

1.3.1 Mass spectrometry-based quantification approaches: labeling versus label-free

Mass spectrometry has been an indispensable tool for protein/peptide identification and quantification due to its high-throughput nature, sensitivity and increasing accuracy [1,4,31]. Multiple alternative strategies have been extensively reviewed [1,14,19], and the "bottom-up" shotgun approach of liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) of tryptic digests is the most commonly used approach (as illustrated in Figure 1.1). The input for the experiment is a mixture of proteins which is usually derived from a cell lysate. First, the proteins are digested into peptides, for example by trypsin. The peptides are then separated by liquid chromatography and analyzed by tandem mass spectrometry. Proteins and peptides are identified and quantified from the MS/MS spectra, and the information can be used as input for downstream analysis (*i.e.* post-identification analysis).

Using this experimental schema, there are two approaches to quantification: stable isotope labeling approaches and label-free approaches. Each has strengths and weaknesses. In stable isotope labeling approaches, the isotopes are introduced into proteins or peptides (1) metabolically, *e.g.* SILAC (Stable Isotope Labeling with Amino acids in Cell culture) [32,33], (2) chemically, *e.g.* iTRAQ (Isobaric Tags for Relative and Absolute Quantitation) [34,35], (3) enzymatically [36], or (4) by spiked synthetic peptide standards [14,37]. Labeling approaches enable within-run comparison of multiple

samples (differentially labeled), and improve quantification precision [1]. SILAC is generally considered the method with highest accuracy [16]. Compared to labeling approaches, label-free approaches are less accurate and thus less sensitive to changes in phosphorylation. However, they are still widely used in global proteomics, because they have no cost of labeling reagents, no time-consuming steps of labeling, and no limit for the number of samples to be directly compared. They can also provide more analytical depth (*i.e.* detect more proteins/peptides in an experiment) and higher dynamic range of quantification [14,38]. Recently, a targeted workflow has emerged, which relies on the prior information to select specific peptides and methods for measurement [39]. It is also referred to as selected reaction monitoring (SRM) or multiple reaction monitoring (MRM). It is designed to overcome the limits of dynamic range, and is highly specific and sensitive. SRM can be conducted with either labeling or label-free approaches [1].

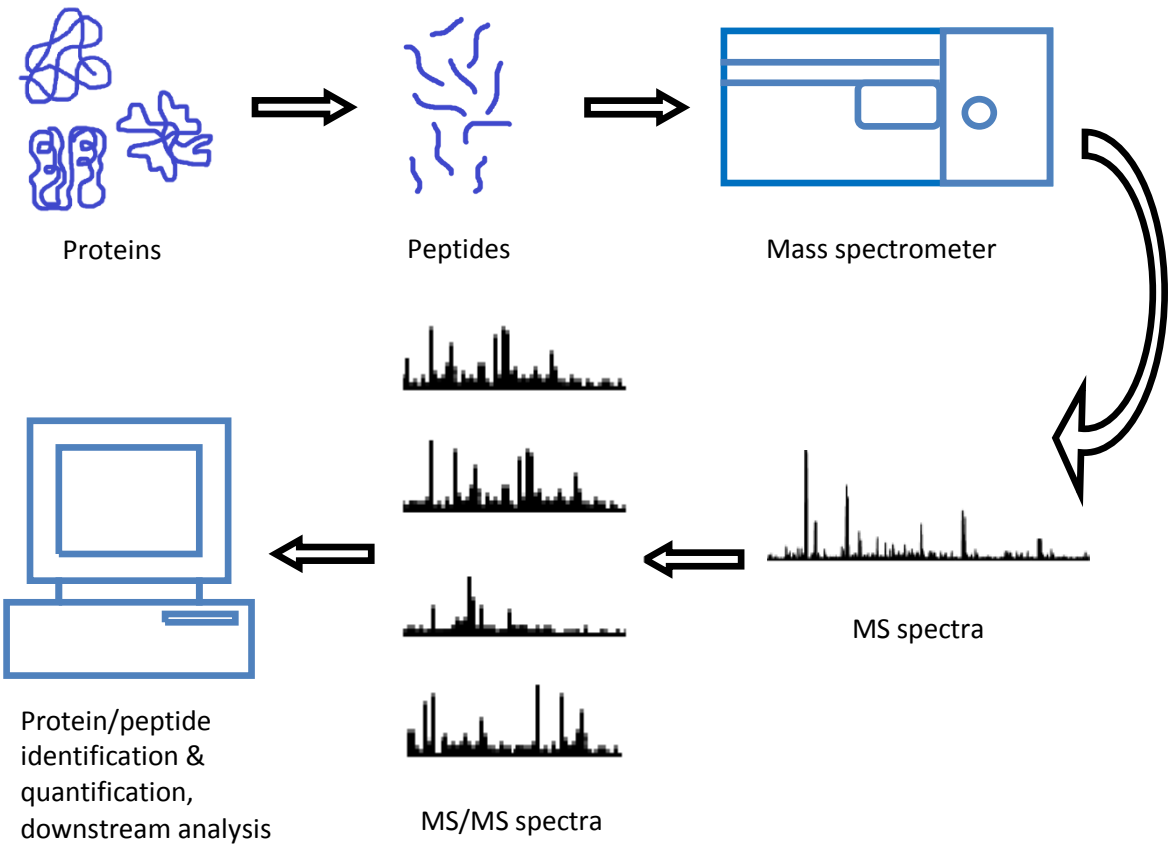


Figure 1.1 Schema of shotgun LC-MS/MS experiment.

1.3.2 Equilibrium-state data versus time-course data

Because of the high expense of running mass spectrometry experiments, most large-scale proteomics studies measure equilibrium-state data [11,12], *e.g.* abundance measurement of proteins extracted from cell culture in equilibrium state. However, proteome dynamics data, *e.g.* measurement of protein abundance changing over time, are expected to directly reveal the kinetics, and will be more informative than equilibrium-state data to uncover the molecular interactions. Several groups have studied proteome dynamics with 4 – 5 time points [17,35,40–42]. More appropriately sampled data points are expected to improve the inference from dynamic data.

1.4 Signaling networks and phosphoproteomes

Signal transduction in biology refers to all the processes in which cells transfer information to create a response, such as physiological changes in response to environmental stimuli, or morphological changes in cell cycle. The most widely studied signaling pathways involve phosphorylation cascades, which dynamically target key components of gene expression, metabolism, and other key functional proteins mediated by kinases and phosphatases. Phosphorylation/dephosphorylation directly affects protein functions (*e.g.*, activation or deactivation, localization, and binding properties). It allows rapid responses by cells to environmental changes (*e.g.* by modification of metabolic enzymes) and also plays a role in defining new cell states. For many proteins, the phosphorylation state is closer to their cellular functions than their expression level. Thus, monitoring the phosphoproteome and its changes under different conditions, and even over time, complements protein expression levels and leads to a better understanding of cell physiological states.

1.5 Categories of general methods used for network discovery

Network discovery is an important focus of my thesis studies. Here I present a review of general methods used in biological network discovery, laying particular emphasis on existing or potential applications in proteomics data analysis. Various methods can be applied to ***comparative samples***, such as multiple experimental

conditions in equilibrium-state studies (as in Chapter 2), and multiple time points in time-course studies (as in Chapter 3 and 4). Some methods are designed especially for time-course data. Besides, methods based on sequence information have also been developed. Graph theory methods can be further applied to analyze the molecular interaction networks derived from the above methods.

1.5.1 Methods for comparative samples

1.5.1.1 *Clustering methods*

Conventional clustering methods, such as hierarchical clustering methods, *k*-means and *k*-medoids algorithms [43], are among the most widely used unsupervised machine learning methods applied in both genomics and proteomics data analysis. Suppose we have comparative samples, clustering can be performed on two different dimensions: (1) along genes or proteins/peptides: genes or proteins/peptides are clustered into sub-groups sharing homogeneous quantitative profiles, which imply co-expression or co-regulation; (2) along samples: samples can be clustered into sub-groups, such as sub-classes of a disease, different stages of a disease, or sub-classes of environmental factors. The clusters along genes or proteins can be input for functional enrichment analysis. Furthermore, they can provide naive information about the network, containing core components comprised of tightly clustered genes or proteins. Each core component corresponds to some common biological functions, and might be closely connected, direct or indirect, in the network. Principal component analysis [44,45] might be an alternative for presenting relationships between genes/proteins/peptides and samples, subject to the usual concerns regarding interpretation, such as information loss.

In the study of nucleolar proteome dynamics [41], hierarchical clustering was adopted to categorize the proteins using fold-change data over time. It successfully identified the two groups of proteins strongly recruited to or depleted from an organelle. In the global study of phosphorylation dynamics of HeLa cells after epidermal growth factor stimulation [17], Olsen and co-workers clustered the time-course data of

regulated phosphopeptides using the fuzzy c-means algorithm [46]. It is a soft clustering method suitable for the case where clusters frequently overlap.

In Chapter 2 (manuscript in revision *Towards Systematic Discovery of Signaling Networks in Budding Yeast Filamentous Growth Stress Response Using Interventional Phosphorylation Data*), we focus on equilibrium-state data of comparative samples from different kinase-mutants. Peptides were clustered based on their phosphorylation changes over multiple samples. Due to lots of scattered peptides which were not stably clustered using conventional methods, we used a tight clustering method [47] to directly identify the most informative and stable clusters. Cellular functions were found to be enriched in top tight clusters.

1.5.1.2 *Correlation network*

Imamura and co-workers [48] picked Olsen's time-course data [17], and performed follow-up analysis to discover signal transduction networks. In the step of constructing the phosphorylation dynamic-based network, they calculated the correlation between each pair of phosphopeptides, and connected the strongly correlated pairs to build a correlation network. This is the general way of building the skeleton of correlation network. In Chapter 2, we generated a correlation network of phosphopeptides using the phosphorylation fold-change data over multiple kinase-mutants. A protein correlation network was then traced back from the peptides. Proteins with positive correlation have similar phosphorylation change patterns over comparative samples; while proteins with negative correlation have opposing phosphorylation change patterns.

A correlation network actually has relevance with the clustering results, if the correlation coefficient is used as the similarity metric for clustering. Tightly clustered proteins usually have high correlations between each other. Note that correlation networks have no directionality.

Instead of building only one correlation network in a study, Hendrickx and co-workers [49] built multiple correlation networks of metabolites under different experimental conditions (one correlation network per condition), based on the time-

course profiles of metabolite concentrations. They monitored the reversal of correlations between conditions, and compared them with known pathways to generate a list of possible regulation mechanisms.

1.5.1.3 *Bayesian network learning*

The Bayesian network framework has been widely applied in inferring gene networks from expression profiles. A Bayesian network is a probabilistic directed acyclic graphical (DAG) model, which represents the conditional dependencies between variables, *i.e.* the molecules of interest. It has the advantage of learning from noisy and incomplete data, as well as combining prior knowledge with data and inferring causal relationships. The theory has been thoroughly reviewed [50,51]. Based on the basic framework, extensions have been developed to apply Bayesian networks to realistic biological data. Cooper and Yoo [52] initially developed the method of learning with interventions which has been used to handle gene mutations [53]. Pe'er and co-workers [53] extended this method to further identify significant sub-networks of genes. Instead of using non-informative priors, Djebbari and Quackenbush [54] made an effort to include preliminary networks obtained from literature and/or protein-protein interaction data into the prior, to train a Bayesian network from microarray data. Shah and co-workers [55] introduced mechanistic Bayesian networks to integrate observed gene expression data with known pathway topologies in order to identify downstream targets.

Besides the application in genomics, Sachs and co-workers [56] applied it to modeling signaling pathways from multiparameter single-cell data. In Chapter 2, we obtained an equilibrium-state SILAC dataset with 8 kinase mutants as interventions. We applied a special Bayesian network, *i.e.* causal Bayesian network, for causal relationship discovery: an edge in the network only indicates a causal influence and activation or inactivation must be read out from phosphorylation profiles. Because of the common missing value issue in proteomics, the inferences from the network have to be made cautiously. To my knowledge, not many other Bayesian network applications have been reported in mass spectrometry (MS)-based proteomics studies. In equilibrium-state

studies, the reason might be the lack of replicate samples or comparative interventional conditions. In time-course studies, the reason might be insufficiency of time points.

1.5.2 Network discovery for time-course data

1.5.2.1 *Differential equations*

Differential equations have long been used for mathematical modeling of kinetic networks from time-course data [57,58]. Examples include modeling of cell cycle [59], signal transduction [60], and even details of specific regulation by phosphorylation [61]. For many cases, differential equation models are constructed based on sufficient prior information, including potential molecules involved in the network, and how they might interact with each other. The reaction parameters can be fitted from data. Differential equation models play a role in mimicking real biological networks, and help explain observed phenomena [62]. The recently reported MIKANA (Method to Infer Kinetics and Network Architecture) algorithm [30,63] is a pioneer method which predicts the network structure in terms of differential equations, and estimates corresponding kinetic parameters as well. Chapter 3 will present MIKANA Ver. 2, including new extensions and improvements to the MIKANA algorithm.

1.5.2.2 *Boolean network and fuzzy logic network modeling*

A Boolean network is a discrete-state network, in which all the variables (nodes) have binary values determined by other variables in the same network. When the network structure is complex comprising a lot of nodes, Boolean networks appear to be neat and tractable, compared to the differential equation model having large numbers of kinetic constants and parameters. It can also integrate qualitative and noisy data [64].

A fuzzy logic network is similar to a Boolean network. The only difference is that the membership function mapping a variable to 0 - 1 no longer has a crisp boundary, but rather follows a sigmoidal curve (or other curves), and the variables can be mapped to non-integer values between 0 and 1 [65]. We can consider a Boolean network as a special case of fuzzy logic networks, when the membership functions' boundaries are crisp.

In Boolean and fuzzy logic network modeling, the initial network, including nodes and linkages between nodes, is constructed based on sufficient prior information. For Boolean network, the membership functions are fixed. For general fuzzy logic networks, the shape of membership functions can be updated automatically from data [65]. Boolean and fuzzy logic networks can incorporate different molecules (proteins, DNAs, metabolites, *etc.*), and it is more feasible than differential equation models when the ranges of kinetic parameters are unknown. While reflecting realistic biomedical networks, this approach can also suggest model alterations corresponding to better explanations of mechanisms, which can guide further experimental testing and validation.

For applications of Boolean network models, see the reviews [66,67]. Aldridge and co-workers [64] used a fuzzy logic network to systematically model the kinase pathway crosstalks in TNF/EGF/insulin-induced signaling. Their model perfectly recapitulated the features in the data, and produced predictions of regulation operations in the pathway.

1.5.2.3 *Time series analysis*

Time series analysis, having overlap with signal processing, has been widely used in financial, social-economic, meteorological studies and engineering applications. Time series analysis quantifies the features and variations of time-course data. It helps summarize the past and predict the future [68]. Some techniques in time series analysis, such as Granger causality, have been applied to analyze biological time-course data to infer interactions. Examples include inferring gene regulation from microarray data, inferring signal delivery on distributed sites in a hemisphere, and so on [69,70]. The Granger causality approach requires relatively large number of time points. It might be applicable to proteome data when we can obtain a sufficient number of time points.

1.5.3 Sequence motif-based methods

Besides the relative and absolute abundance information, peptide sequence information can also be utilized to extract enriched sequence motifs. In eukaryotic proteins, phosphorylation only occurs on serine, threonine, tyrosine and histidine

residues, although phosphohistidine is unstable to isolation. The sequence patterns (***sequence motifs***) common among multiple peptides can be summarized. In phosphorylation studies, sequence motifs can help identify the kinase-substrate binding domain and provide insights into how the upstream kinases regulates the proteome [16]. But in isolation, they provide little or no predictive value on the status of a single phosphorylation site in a particular cell/physiological state without experimental support. A recent algorithm NetworKIN [71] has been developed to predict *in vivo* kinase-substrate relationships based on motif information of kinases and phosphoproteins. Sequence information is powerful to match kinases and their substrates; however, it might also identify false positive protein pairs having matching sequences but no functional interactions. Combining sequence information with interventional phosphorylation responses or phosphorylation dynamics might help lower the false discovery rate.

1.5.4 Graph theory methods for summarizing and comparing network attributes

All the above methods are used for inferring network characteristics and structure. For whatever network we generated, directed or non-directed, we can summarize the network attributes, such as network motifs [72], node degrees [73], and shortest path lengths [74]. The distributions of network attributes represent the nature of the network, and provide measures to compare different networks.

We can extract small, repeated sub-graphs from the entire network. A sub-graph is a significant pattern, called a ***network motif***, if it has significantly higher frequency in the network compared to a random network [72]. Imamura and co-workers [48] generated a correlation network of peptides based on time-course data, then applied network motif analysis to the non-directed network. Distribution of other graph theory-based measures, such as node degrees and shortest path lengths, were inspected as well to describe the network and compare it with known network topology.

In Chapter 2, we analyze the equilibrium-state data of phosphorylation level changes induced by kinase mutations. Part of our analyses is the correlation network analysis. Our correlation network of phosphoproteins is generated based on

phosphorylation responses over multiple mutants. The proteins with high degrees of connectivity are predicted as hub proteins involved in the network. Samples of the hub proteins are then used for causal Bayesian network structure learning to identify the causal influences between them.

Similar analysis can be applied to various networks, including genome, proteome and transcriptome networks, and might provide insight into the relationships between different molecular levels.

1.5.5 Summary of review

Various methods can be applied to both equilibrium-state and/or time-course data for network discovery. Clustering methods and correlation network analysis are among the simplest and most widely used methods for revealing some nature of networks. In order to infer connections between nodes and even directed causal influences, more sophisticated network modeling methods can be applied, including Bayesian network structure learning, Boolean network modeling and so on. Clustering methods, correlation network analysis and Bayesian network structure learning require multiple comparative samples. Bayesian network learning has been widely used in microarray analysis; however, its application in proteomics has not been as wide. One of the reasons might be the much fewer comparative samples available in proteome studies due to high experimental expenses. Differential equations have long been used in kinetic modeling. The method benefits from a large number of time points. Boolean network and fuzzy network modeling are mainly used in the metabolomics community. They require a fair amount of prior information of the network structure. Time-series analysis brought from engineering can also be applied in analyzing biological time-course data. The methods making use of sequence information do not rely on comparative samples. They identify the sequence motifs among multiple peptides and identify the relationship between motifs.

Besides the main categories of methods mentioned above, other methods have also been proposed to analyze biological data. Albert *et al.* [75] proposed a method to infer signaling network from indirect experimental evidence. Specifically, observed

causal relationships are represented as paths in a network, and then combinational optimization is used to find the sparsest network structure fitting the experimental observations. Yeung *et al.* [76] used singular value decomposition and robust regression to reconstruct connectivity topology of large sparse networks.

When a molecular interaction network is generated using the above methods, graph theory methods can be further applied to the network attributes, which present the nature of network and facilitate network comparison.

Chapter II

Towards Systematic Discovery of Signaling Networks in Budding Yeast Filamentous Growth Stress Response Using Interventional Phosphorylation Data

2.1 Abstract

Reversible phosphorylation is one of the major mechanisms of signal transduction. Phosphorylation signaling networks are critical regulators of cell growth and development, but to date few such networks have been delineated extensively. Towards this end, quantitative phosphoproteomics is emerging as a useful tool enabling the large-scale determination of relative phosphorylation levels. However, phosphoproteomics differs from classical proteomics by a more extensive sampling limitation due to the limited number of detectable sites per protein. Here, we propose a comprehensive quantitative analysis pipeline customized for phosphoproteome data from interventional experiments for identifying key proteins in specific pathways, discovering the protein-protein interactions and inferring the signaling network. We made an effort to partially compensate the missing value issue haunting proteomics studies as well. For developing our pipeline, we used mass spectrometry-based SILAC (Stable Isotope Labeling with Amino acids in Cell culture) data with interventional experiments in the form of kinase-inactivating mutations. The major building blocks of the pipeline include phosphopeptide meta-analysis, correlation network analysis and Bayesian method-based causal relationship discovery. We were able to successfully apply our pipeline to a series of interventional experiments identifying phosphorylation

events underlying the transition to a filamentous growth form in *Saccharomyces cerevisiae*. We identified 5 high-confidence proteins from meta-analysis, and 19 hub proteins from correlations analysis (Pbi2p and Hsp42p are identified by both analyses). Nine of them have direct or indirect evidence of involvement in filamentous growth. All of these proteins are involved in stress responses. In addition, five of our predicted novel proteins, Nth1p, Pbi2p, Pdr12p, Rcn2p and Pbp1p, have been tested by interventional phenotypic experiments; and all of them present differential invasive growth, providing validation of our approach in this system. Our pipeline was able to infer the phosphoprotein interaction networks, which suggested potential proteins that can be intervened in future studies. The new comprehensive pipeline presents an effective systematic way for discovering signaling networks using interventional phosphoproteome data, and we anticipate the methodology to be applicable as well to other interventional studies via different experiment platforms.

2.2 Introduction

Cells exchange and receive information from the environment through signaling pathways, which are crucial for cells to maintain normal functions and properly respond to stress and stimuli. Dysregulation of these processes is a major factor in the emergence of many diseases, including cancers, diabetes, and cardiovascular disease. Reversible phosphorylation is one of the major forms of signal transduction and can affect protein function and gene expression [7–13]. Investigations into phosphorylation provide insight into signaling pathways by providing the target sites of phosphorylation and the quantitative changes in phosphorylation level in response to genetic or environmental perturbations. Effective, sensitive identification of candidate proteins for further studies remains a challenge in the face of experimental limitations of current technologies which have a high cost component, provide incomplete coverage of the phosphoproteome, and have sampling limitations which affect replicate runs.

Large-scale phosphoproteomics studies on a number of organisms have been carried out using mass spectrometry (MS)-based approaches (reviewed in [77–79]). These include two recent global phosphoproteomic studies of the budding yeast

(*Saccharomyces cerevisiae*) [11,12]. In the study carried out by Bodenmiller *at al.* [11], protein kinases and phosphatases were systematically perturbed through gene deletions. The system-wide responses to the perturbations were measured by label-free MS-based quantification, and the results evaluated to determine their contributions to understanding the relationships between these signal transduction proteins and cell pathways. Another global interaction study focused on kinase and phosphatase interactions [12] by capturing protein-protein interactions by affinity capture-immunoblot and identifying the isolated protein complexes by mass spectrometry. These two global studies both adopted label-free, cost-effective quantitative approaches. However, label-free methods typically increase variance relative to isotope enrichment methods [14]. For the purpose of this study, we have used isotope labeled SILAC (Stable Isotope Labeling with Amino acids in Cell culture) method [32,33] to increase sensitivity to change.

The general scope of this manuscript is the description of a comprehensive pipeline, incorporating statistical and mathematical methods, for investigating and evaluating the quantitative phosphoproteome data, identification of candidate proteins and processes to be pursued in subsequent molecular biology and genetic studies. The phosphoproteome data utilized in this analysis was obtained from interventional experiments of a subset of yeast kinases. Eight yeast kinases, Ksp1p, Kss1p, Sks1p, Ste20p, Snf1p, Tpk2p, Elm1p and Fus3p, that were identified as components of budding yeast filamentous growth response [28,80–82] were mutated (inactivated alleles) individually as genetic interventions. Each of these kinases exhibits a filamentous growth deletion phenotype, with the deletion of *KSP1*, *KSS1*, *SKS1*, *STE20*, *SNF1*, and *TPK2* yielding a loss of filamentous growth and the deletion of *ELM1* and *FUS3* yielding enhanced filamentation. Classic studies have identified key kinase-based signaling networks that regulate the filamentous growth transition. In particular, yeast filamentous growth is regulated by mitogen-activated protein kinase (MAPK) and protein kinase A (PKA) pathways [80,83,84] as well as being impacted by other signaling pathways. MAPK pathways are evolutionarily conserved across phyla and consist of

three-kinase cascades serving central roles in signal transduction in eukaryotic cells [85]. A graphical illustration of currently recognized budding yeast filamentous growth pathways, integrating information from authoritative pathway databases and reviews, is shown in Figure 2.1. The phosphoproteome data for this study was obtained by SILAC approach, and we used the Mascot search engine [86] followed by MaxQuant software [24] to identify and quantify peptides and proteins. We obtained phosphorylation level changes from the MaxQuant analysis for mutants versus wild type control for the comprehensive quantitative analyses.

The broad focus of the filamentous growth kinase networks in particular has made it difficult to tease out important kinase targets (direct or indirect). Bioinformatics methods provide a promising avenue with which local kinase signaling relationships can be identified. While traditional cluster analysess associated with functional enrichment analysis are useful tools, their performance might be affected by the missing value issue. We need to deal with it in order to obtain reliable clusters and enriched functions. Furthermore, a more integrative and extensive analysis is necessary to find new components of the pathways, uncover relationships between the pathway components, and to elaborate the signaling network structure. Thus we propose this comprehensive quantitative analysis pipeline customized for SILAC data, and partially compensate the missing value issue. The major building blocks include phosphopeptide meta-analysis, correlation network analysis, causal relationship discovery, and validation by literature mining. We have successfully applied the pipeline to analyze our current yeast data. Candidate proteins predicted to contribute to the filamentous growth response were selected by phosphopeptide meta-analysis and correlation network analysis. Causal relationship discovery was performed on candidate proteins identified from our analysis and validated proteins from the literature. The inferred causal relationships, along with the interactions inferred from phosphorylation changes in response to individual mutants, have suggested potential proteins that can be further intervened and studied in the future.

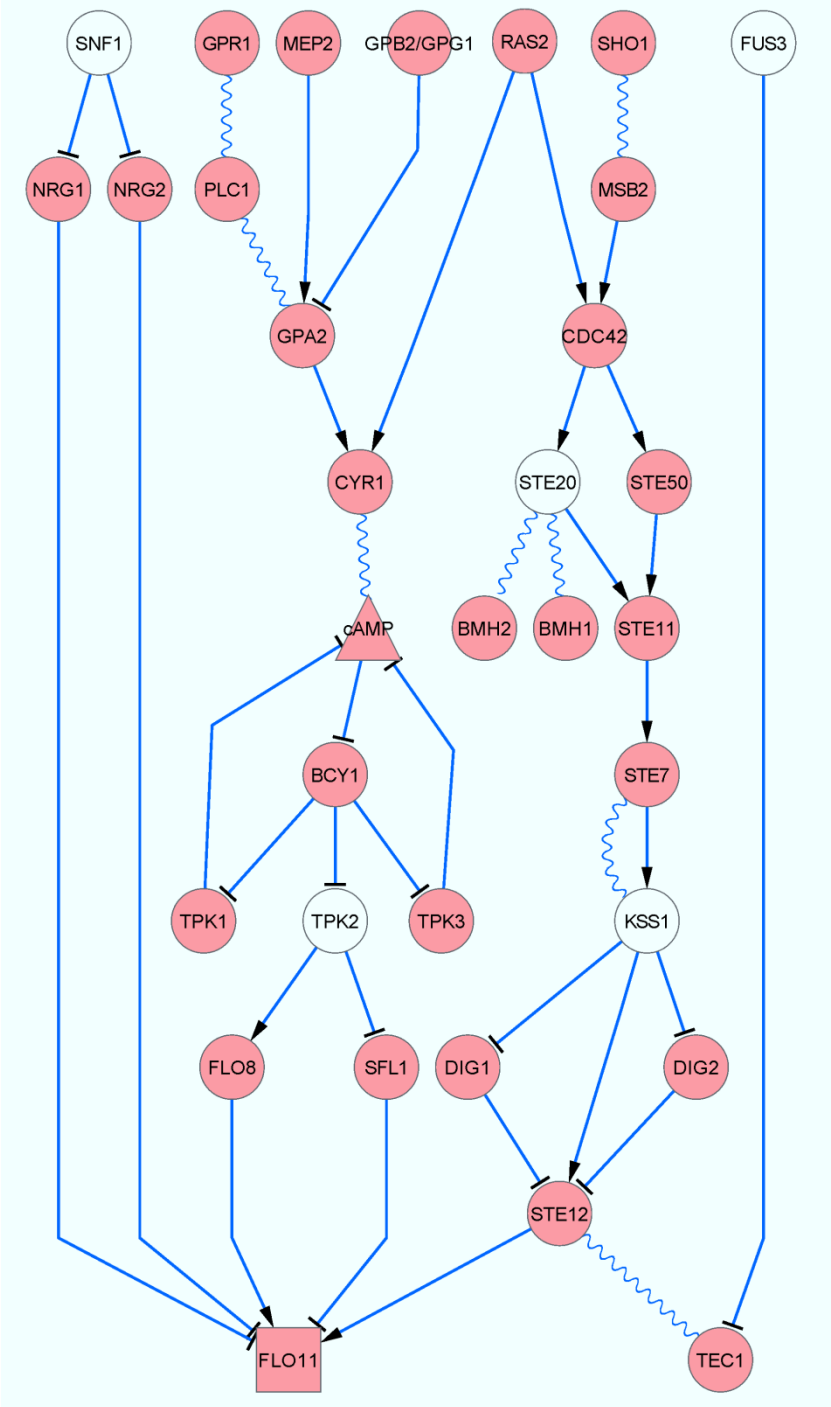


Figure 2.1 Graphical illustration of the filamentous growth pathway in budding yeast from previous studies.

The ellipses ◯ are proteins; the rectangles □ are genes; and the triangles △ are metabolites. The linkage between shapes: →indicates stimulation, ⊣indicates inhibition, and ∿indicates association. The information were extracted from *Science Signaling* Database of Cell Signaling [80] and KEGG database [28]. The white ellipses are five of the eight kinases selected to be mutated in our experiments.

2.3 Materials and Methods

2.3.1 Mass spectrometry data

Tandem mass spectrometry data were generated from a series of triplex SILAC [32,87,88] experiments of kinase-dead mutant (KD) strains versus the wild type (WT) haploid filamentous yeast Y825 strain. Different Lys and Arg isotope forms were used to label the three samples in a triplex SILAC experiment: light (Lys0/Arg0) for WT control sample, medium (Lys4/Arg6) and heavy (Lys8/Arg10) for two different mutant samples. We then obtained peptide phosphorylation fold changes in medium versus control and heavy versus control samples.

Eight yeast kinases, KSP1, KSS1, SKS1, STE20, SNF1, TPK2, ELM1 and FUS3, all known to be involved in filamentous growth [28,80,81], were chosen to be mutated (inactivated alleles) individually to cultivate mutant cell cultures. All strains were auxotrophic for Lys and Arg, and were grown on defined medium supplemented with the appropriate isotopic forms of Lys and Arg. The cultures were grown to log phase, and treated with 1% (vol/vol) butanol to induce filamentous growth [81]. The treated samples were incubated for another 16 hours to obtain enough proteins for mass spectrometry analysis. The final O.D. at 600nm reached a high value usually between 1.0 and 1.5. Cells were harvested by centrifugation and lysed in the presence of protease and phosphatase inhibitors. Protein levels were determined by the Bradford protein assay and the proteins from the triplex labeling were then pooled, and were digested by trypsin. The digest was separated into fractions using strong cation-exchange (SCX) fractionation, followed by selective enrichment of phosphorylated peptides using titanium dioxide [89,90] and then analyzed by LC-MS/MS using a Thermo Fisher Orbitrap XL mass spectrometer. Peptides were identified using MaxQuant software [24] following the Mascot search engine [86], and filtered requiring peptide identification FDR < 1%. The method for calculating peptide identification FDR based on concatenated databases was described by Cox J and Mann M [24]. A total of 3,312 phosphopeptides representing 1,063 proteins were identified. Among those, 73 unique phosphopeptides representing 66 common proteins were commonly identified in all the 8 kinase-dead

mutants; while, 882 phosphopeptides representing 486 proteins were common to at least half of the kinase-dead mutants. These summary numbers are listed in Table 2.1.

Table 2.1 Summary of the dataset.

Summary	Number of phosphopeptides	Number of proteins
Identifications in the whole dataset	3,312	1,063
Identifications common among all 8 kinase-dead mutants (KDs)	73	66
Identifications common among 4 - 8 KDs	882	486
Identifications that are significant in at least 1 KD	863	452
Globally significant differential phosphorylation sites	28 (5 from complete measurements – high-confidence)	26 (5 from complete measurements – high-confidence; 17 have inner connections supported by STRING[91,92])
High-confidence hub proteins identified from the stringent correlation network	-	19
Proteins known to be involved in filamentous growth from literature mining, and detected in our dataset	-	20 (15 of them are significant in at least 1 KD)

2.3.2 Post-identification analyses

2.3.2.1 *Phosphopeptide meta-analysis*

In the meta-analysis, we contrast and combine the results from different KD-versus-WT experiments, so that to find the correlations between kinase-dead mutants, categorize peptide phosphorylation patterns over experiments, and identify differentially phosphorylated peptides.

Overview of the influences inferred from kinase-dead mutations

The relative phosphorylation level obtained for each phosphopeptide is represented as a ratio for each of the 8 kinase-dead mutants (KD) versus wild type (WT) under filamentous growth conditions. Two examples of phosphopeptides identified in all 8 kinase-dead mutants are shown in Table 2.2. The ratio lists of all the identified phosphopeptides are aligned to constitute a ratio matrix. The quantity measuring

statistical significance of each ratio, *i.e.*, the significance B value, was calculated with MaxQuant [24]. The ratios shown in Table 2.2 were extracted before filtering by statistical significance.

Table 2.2 Ratio lists for two representative phosphopeptides from the ratio matrix.

Phosphorylation fold-change / Phosphopeptide	Sks1-KD vs. WT	Ste20-KD vs. WT	Snf1-KD vs. WT	Tpk2-KD vs. WT	Elm1-KD vs. WT	Fus3-KD vs. WT	Kss1-KD vs. WT	Ksp1-KD vs. WT
ADDEEDLS(ph)DENIQPELR	0.72	0.71	0.70	0.52	1.0	0.88	0.83	0.86
ADGTGEAQVDNS(ph)PTTESNSR	2.3	3.7	2.1	2.2	0.33	0.58	0.75	0.69

Phosphorylation level of each phosphopeptide is represented in a list of ratios. We used the peptide ratios provided by the MaxQuant output, which have been normalized for each LC-MS/MS run [24]. The significance B values provided by MaxQuant are not shown here. For the cluster analysis, if a phosphopeptide is detected multiple times under the same KD-versus-WT condition, the median of all its ratios are taken. S(ph) or T(ph) indicates that the specific amino acid, serine or threonine, is phosphorylated, respectively.

For the purpose of evaluating similar or reciprocal effects on phosphorylation changes in response to different kinase mutations, we generated a correlation heatmap of the kinase-dead mutants (see Figure 2.2), which is presented as Spearman correlations between pairs of mutants. In order to avoid the strong correlation dominated by the majority of peptides whose phosphorylation do not change significantly, only the peptides having at least 2-fold changes in both mutants were used for calculation. Positive or negative correlations can be interpreted as similar or reciprocal effects on phosphorylation induced by different kinase mutations.

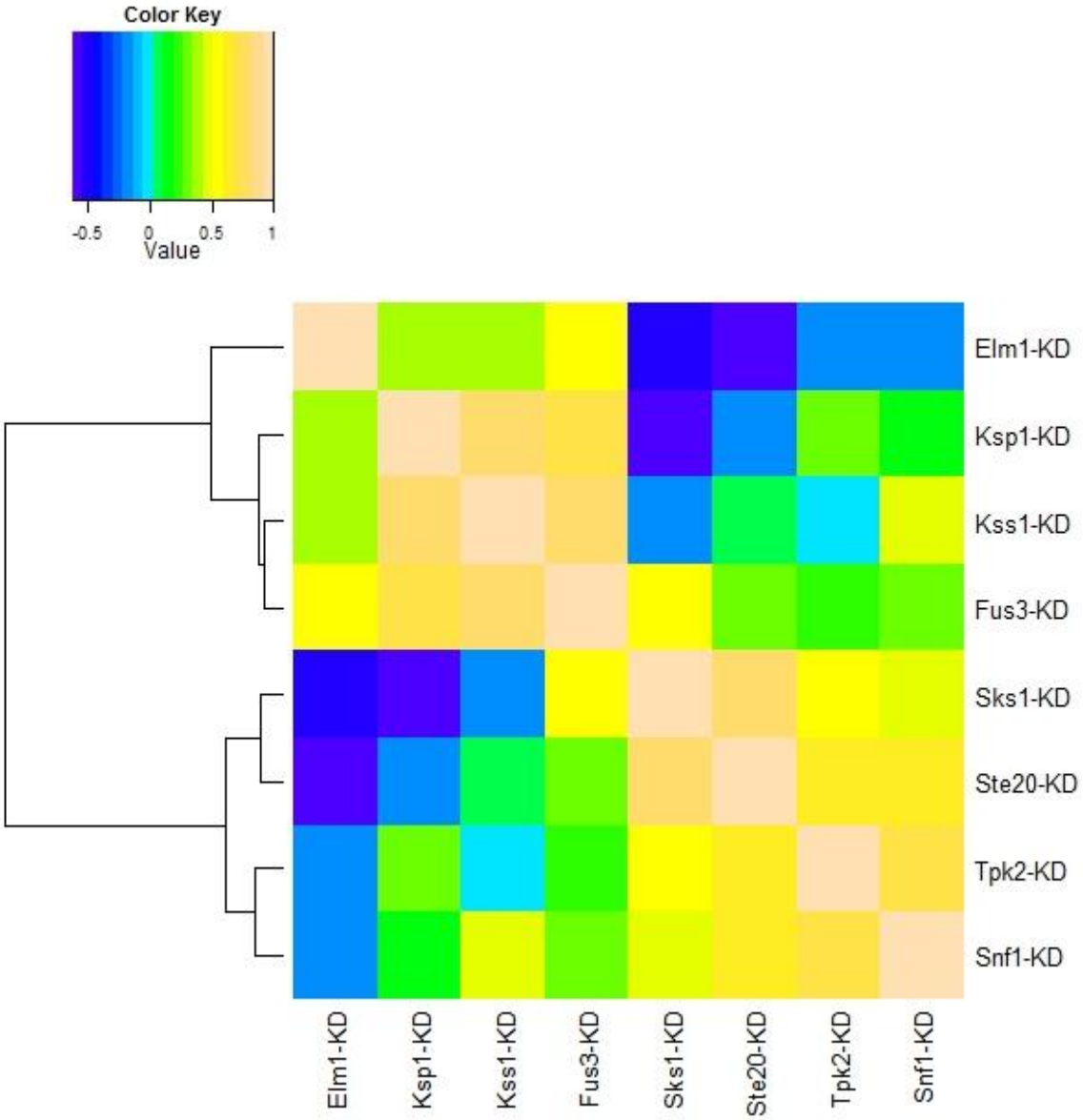


Figure 2.2 Correlation heat map of the kinase-dead mutants (log2 ratios adopted).

The hierarchical clustering tree using Spearman correlation as the similarity metric is drawn along the left side of the heatmap.

Clustering phosphopeptides

Our goal of this cluster analysis is to find the groups of phosphopeptides sharing similar phosphorylation change patterns, which are likely to be involved in similar functional pathways. The phosphopeptides commonly identified in 4 - 8 KD-versus-WT conditions were selected, and the missing values were imputed (on log2 scale) using 5-

nearest neighbor averaging [93,94]. The imputed dataset was analyzed using the tight clustering method [47], which sequentially identified the most informative, tight and stable clusters from the data, without enforcing all peptides to be clustered.

We also attempted several traditional clustering methods, such as hierarchical clustering methods [95] and PAM (Partitioning Around Medoids) method [43]. Traditional methods enforce all peptides to be clustered. However, due to lots of scattered peptides which only loosely match to the patterns, it is problematic to estimate the number of clusters and the clustering results were not satisfactory. While the tight clustering method is more suitable for data with scatter peptides. It advantageously identifies the tight clusters in the decreasing order of stability, and the selection of the number of clusters is less crucial [47].

Note that the cluster analysis was performed at peptide level rather than protein level, because many proteins contain multiple phosphorylation domains whose responses may correlate or not, depending on the function of phosphorylation at those sites and the physiological conditions examined. Proteins can be traced back from the peptides.

Functional annotation within each tight cluster

The functional terms were annotated for the proteins in top tight clusters to survey functional enrichment. The Functional Annotation Tool on DAVID v6.7 [96,97] was used to facilitate annotation.

Identification of differential phosphorylation in each mutant

The phosphopeptides that change phosphorylation level significantly in each individual KD-versus-WT experiment were selected by the significance B value < 0.05.

Identification of globally significant differential phosphorylation

The kinases selected to be dead mutated are all known to be involved in filamentous growth. The proteins which have globally significant responses in the

mutants versus WT controls are potential components involved in filamentous growth or expression products of the gene targets. Detecting globally differentially phosphorylated peptides combining the results from all the KD-versus-WT experiments is a multiple testing problem [98]. Due to the missing data issue common in proteome data, it is too stringent and impractical to require a candidate to be completely significant in all the experiments. Thus, we relax the requirement, and use less stringent methods which can still identify the candidates having global significance. We extended the Fisher's combined probability test [99] to allow missing values, and it was applied to solve the multiple testing problem.

In the framework of Fisher's method, the two-tailed p-value p_i for an individual significance test in a KD-versus-WT experiment is calculated as twice the significance B value. In our dataset, $i = 1, 2, \ldots, 8$, corresponding to the 8 KD-versus-WT conditions, and the total number of individual tests, $N = 8$. The test statistic $X^2 = -2\sum_{i=1}^{N} \ln(p_i)$ follows a chi-square distribution with $2N$ degrees of freedom. Thus, the p-value for the test statistic X^2 can be determined, which is the combined p-value for all 8 individual tests. Each identified phosphopeptide has a combined p-value as a measure of global significance. The extension of Fisher's method: for each phosphopeptide, its combined p-value was calculated from all of its available significance B values. All the non-missing values were retained for calculating the combined p-value, rather than excluding the incomplete data from the dataset. The FDR of multiple testing is controlled using the *Benjamini-Hochberg* procedure [100].

We also adapted an adaptively weighted statistic-based method (missing values not allowed) [101], which was initially developed for detecting differential gene expression, for detecting differential phosphorylation from our common peptides appearing in all KDs. The globally significant phosphorylation sites detected by these two methods were generally consistent.

2.3.2.2 *Correlation network analysis*

A correlation network of all the 73 common phosphopeptides with complete measurements was generated based on their phosphorylation changes under all 8 KD-

versus-WT conditions. The Pearson correlation coefficient between each pair of distinct phosphopeptides was calculated. Strong correlations meet the following criterion: p-value of the Pearson's correlation test < 0.05, and a stringent requirement of |Pearson correlation coefficient| ≥ 0.9. The protein identifications can be traced back from the phosphopeptides.

The correlation network among proteins is an undirected network. Degrees of connectivity for each protein in the network can provide an assessment of importance of the protein. The higher the degree, the more frequently the protein is involved in interactivities with other proteins in the network. From this measurement, we predict core-components in the correlation network.

2.3.2.3 *Literature mining*

In addition to the candidate proteins predicted by global differential phosphorylation and the core-components identified from the correlation network, we also retrieved a list of proteins reported as known or potential components involved in filamentous growth from literature as well as authoritative databases, such as SGD [102,103], BIOGRID [104] and *Science Signaling* Database of Cell Signaling [80]. Note that people have usually used different terms to refer to filamentous growth in haploid cells; "filamentous growth", "filamentation" might all refer to the same biological process. In SGD database, we search both key words for Descriptions and GO Biological Process terms associated with the proteins.

2.3.2.4 *Causal Bayesian network modeling*

The correlation network is intuitive; however, it is not directed, and direction information for networks is quite useful for interpretation. For this reason we went beyond correlation analysis to causal Bayesian network modeling. Because different phosphopeptides from the same protein do not definitely change phosphorylation level in the same direction, the network modeling must be performed on peptide level, and then traced back to their parent proteins.

Data preprocessing

If a phosphopeptide was detected more than once in a specific mutant, the median of the fold-changes was taken as a representative of the response in this mutant. The phosphorylation fold-changes of peptides were discretized into three states based on the 2-fold change criterion [50]: if the ratio is smaller than 0.5, the state is categorized into *under-phosphorylation*; if the ratio is greater than 2, the state is categorized into *over-phosphorylation*; otherwise, the state is categorized into *baseline*. This discretization criterion worked effectively for this dataset.

Causal relationship discovery

A causal Bayesian network is a Bayesian network in which a directed edge is interpreted as a causal influence from the parent node to the child node [105,106]. In our study, each protein (represented by unique phosphopeptides) is considered as one node in the network, and a directed edge starting from the node of protein X pointing to the node of protein Y represents a causal influence of protein X on Y. Disregarding confounding influences, there are three simple model structures between two proteins X and Y: (1) X has causal influence on Y; (2) the opposite; (3) no causal relationship between X and Y. Note that the directed edge only indicates the direction of causal influence, but do not tell whether the influence is activation or inhibition.

Non-informative prior distribution of the model structures is used. For given data, D, and prior knowledge, K, we want to find the model structure, S, that has the highest posterior probability, $P(S|D,K)$. According to Bayes' theorem, $P(S|D,K) \propto P(D|S,K)$. While all the nodes have been discretized in ***Data preprocessing***, assuming the causal mechanisms are local and independent, and the prior distribution of the parameters associated with each node is Dirichlet, the marginal likelihood $P(D|S,K)$ can be obtained by the Bayesian Dirichlet equivalent (BDe) metric [52,105,107,108]. For the mixture of observational and interventional data, only the passively observed cases are counted in the BDe metric calculation [52,105]. The structure with the highest posterior probability is assigned to the corresponding pair of proteins.

The analyses were implemented in R v2.15.1 and MATLAB R2012a. The causal Bayesian network structure learning was performed in MATLAB using BNT (Bayes Net Toolbox for MATLAB) v1.0.7 [109]. Cytoscape v2.8.3 [110,111] was used for network visualization.

2.4 Results

2.4.1 Workflow

An overview of the analytical workflow is shown in Figure 2.3. Following peptide identification and quantification, the comprehensive post-identification analyses performed consisted of phosphopeptide meta-analysis, correlation network analysis, and literature mining, followed by causal relationship discovery to infer signaling network characteristics. The inferred protein-protein relationships involving hub proteins were backed up by literature, and suggested potential proteins to be intervened in the future studies of yeast filamentous growth pathways. Details of the methodologies are described in **Materials and Methods**. Table 2.1 lists several important summary numbers of this dataset.

2.4.2 Similar or reciprocal effects induced by kinase-dead mutations

The relationships of the eight kinase mutants and their effects on global phosphorylation patterns were subjected to correlation analysis (see ***Overview of the influences inferred from kinase-dead mutations*** in **Materials and Methods**). The results were visualized in a correlation heatmap (Figure 2.2). The negative correlation between kinase mutants of *SKS1* and *ELM1* are apparent from Figure 2.2 as are the similarities between some of the mutants (*e.g.*, *SNF1* and *TPK2*). *SKS1* mutants inhibit filamentous growth and *ELM1* promotes it, while *SNF1* and *TPK2* have similar phenotypes. The general correlations between kinases are consistent with their filamentous growth phenotypes and reinforce the identification of core target proteins.

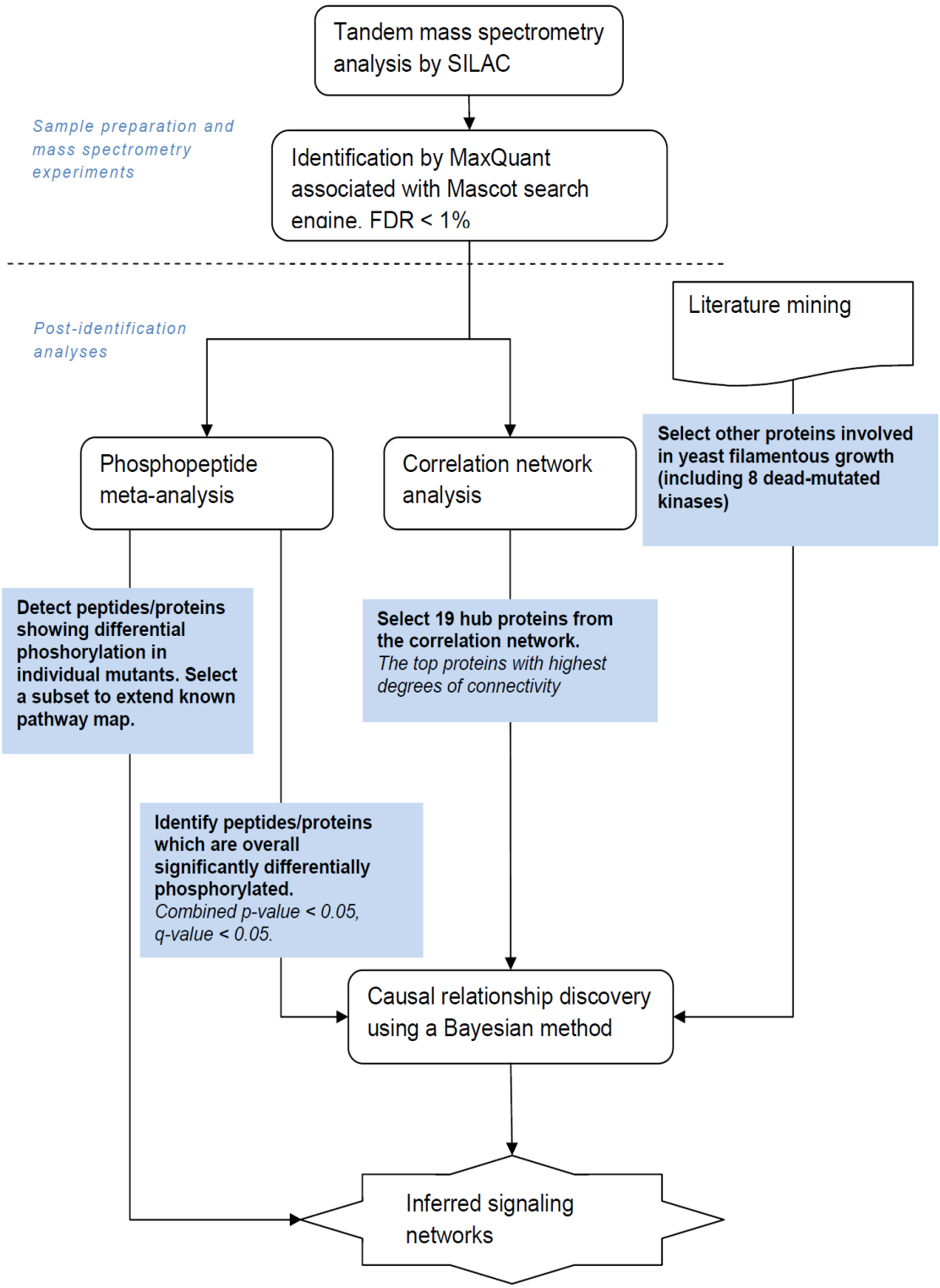


Figure 2.3 Summary flow chart of the analytical workflow.

We must be cautious when interpreting the correlations for evaluating partially multiplexed data, such as in triplex SILAC experiments where two kinase mutant samples and a control sample were analyzed in the same MS run. The identification and quantification of phosphopeptides in all the samples in a triplex experiment tended to be tied, because a peptide quantified for one sample should also be quantifiable for the other two samples. The overlap within a triplex run should be 100% but the overlap between runs will be lower due to instrument sampling limitations. A high number of replicates may contribute to minimize missing data, and compensate the possible bias brought by tied identification and quantification.

2.4.3 Phosphopeptide clusters based on phosphorylation changes

A total of 882 phosphopeptides representing 486 proteins were commonly identified in 4 – 8 kinase-dead mutants (KD). After the missing values being imputed, the tight clustering method [47] was used to assign those phosphopeptides into groups, and identify the most informative, tight and stable clusters (see ***Clustering phosphopeptides*** in **Materials and Methods**). The results are illustrated in Figure 2.4. The assignment of proteins and peptides in the top 8 tight clusters is provided in Table 2.3. We also surveyed enriched functions in the tight clusters (Table 2.3), in terms of functional categories, Gene Ontology, pathways and proteins Domains [27,28,96,97]. In summary, similar phosphorylation change patterns over multiple mutants (compared to wild type) tends to suggest involvement in similar biological functions. Enriched functional terms include nucleotide phosphate-binding domains, ribosome biogenesis, fructose and mannose metabolism, and glycolysis. Differential carbohydrate metabolism is consistent with the invasive nutrition forage observed under environmental stresses leading to filamentous growth.

Table 2.3 Top 8 tight clusters and functional enrichment.

Cluster	Proteins (traced back from phosphopeptides)	Enriched terms
1	YRO2, BUG1, VPS74, HXK1, PIL1, FBP26, PTK2, NPA3, BIR1, MYO3, UTP14, ARE2, DBP5, RUD3	Nucleotide phosphate-binding region:ATP (P-value=6.54E-04, Benjamini=3.4E-2) ** Nucleotide-binding (P-value=1.8E-3, Benjamini=4.2E-2) ** ATP-binding (P-value=6.0E-3, Benjamini=9.3E-2) *
2	VMA2, SEC31, GLY1, PEA2, VTC2, KEM1, UFD1, TIF4631, BCY1, SPA2, MFT1, NEW1, KRE6	-
3	NUP60, SLA1, STU1, YCL020W, VBA4, HOM2, YDR365W-B, VPS74, PSP1, CHD1, NUP145, SPT6, HSE1, ABF1, MEH1, CKI1, YLR413W, SPT5, HRB1, LCB4, CAF20, MRL1	Endosome (P-value=1.6E-3, Benjamini=6.6E-2) * RNA polymerase II transcription elongation factor activity (P-value=1.4E-3, Benjamini=9.6E-2) * Transcription elongation regulator activity (P-value=2.8E-3, Benjamini=9.9E-2) *
4	FAP7, ITR1, LSB3, LEU1, FLC3, SPT6, YGR125W, CRP1, KEL1, LCB3, YBT1, BDF1, YMR031C, DDR48, YMR295C, GPD2, ZEO1, CAF20, SNF2	-
5	PIN4, CYC8, BUD3, LYS20, CDC34, MAK21, BFR2, SUM1, GLY1, NUP145, PRP43, SPT6, ENP2, YOR1, SSZ1, NUP2, YLR345W, SUB1, ESC1, BDP1, DCP2, RPC31, SLA2, NOP8, ALE1, MSB1, SNU66	Nucleus (P-value=1.0E-4, Benjamini=3.4E-3) *** Nuclear lumen (P-value=3.4E-4, Benjamini=2.7E-2) **
6	SIF2, PPH22, VAC8, HSP12, RTF1, RSC30, TRA1, LCB3, NAP1, SIC1, RPN13, YMR196W, MRE11, MCK1, LEM3, FPK1, LSP1	-
7	IST2, AIM3, RPC53, YDR186C, ECM32, MIG1, HXK2, VHS2, RNR2, UTR1, FBA1, EAP1, YLR257W, PFK2, PFK2, ACC1, YOR052C	Fructose and mannose metabolism (P-value=3.0E-3, Benjamini=3.9E-2) ** Glycolysis (P-value=1.6E-3, Benjamini=4.3E-2) ** Glycolysis / gluconeogenesis (P-value=9.8E-3, Benjamini=6.2E-2) *
8	AKL1, IST2, MAK5, FEN1, LHP1, RPC53, SAS10, SHS1, MAK21, DOP1, GCD6, GUK1, CHO1, PDA1, LEU1, NOP7, SPT6, TFG1, HXT1, AIM21, URA2, CDC11, MAK11, VPS13, CBF5, VTA1, CRN1, YMR031C, EFR3, ADE4, NOP12, MAM3, CAF20, PEX25, TIF5	Ribosome biogenesis (P-value=1.0E-4, Benjamini=5.0E-3) ***

Functional enrichment P-value and Benjamini-Hochberg corrected p-value (Benjamini) were calculated with DAVID Functional Annotation Tool [96,97]. They are given in the brackets following corresponding terms.

* Benjamini < 0.1, ** Benjamini < 0.05, *** Benjamini < 0.01.

All the clusters are highly enriched in the term “phosphoprotein” (not listed above).

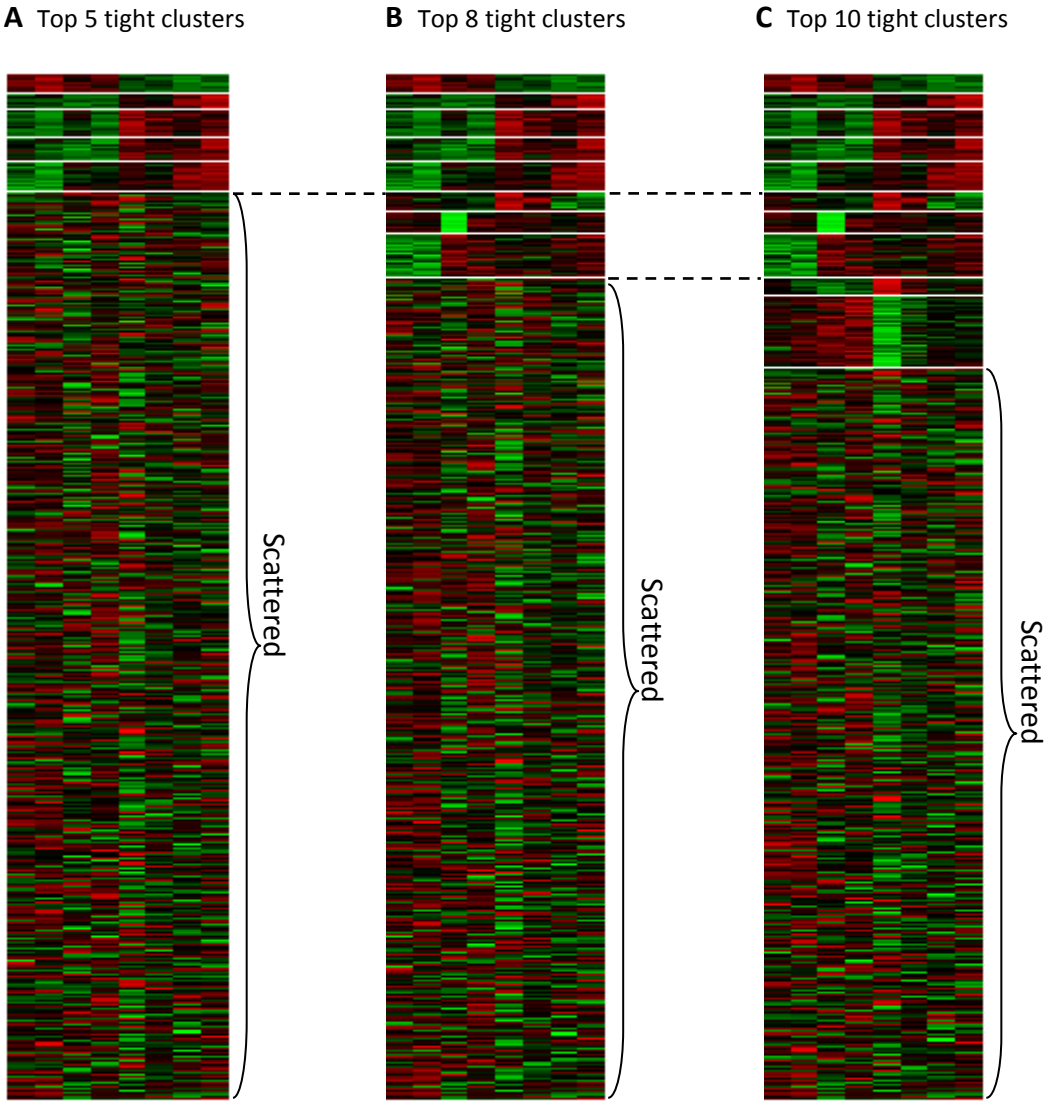


Figure 2.4 Top clusters selected by tight clustering.

The phosphopeptides commonly identified in 4 – 8 KD-versus-WT conditions were used. After missing value being imputed, the tight cluster method [47] was used to pick out the top tightest and stablest clusters. R package tightClust was used, adopting the suggested parameters. R code: tight.clust(data.impute, target, k.min=15, random.seed=12345), the value of target is 5, 8 and 10, respectively, for **A**, **B** and **C**. The columns of a map correspond to SKS1-KD, STE20-KD, SNF1-KD, TPK2-KD, ELM1-KD, FUS3-KD, KSS1-KD and KSP1-KD. The rows correspond to phosphopeptides.

A - Top 5 tight clusters were sequentially selected, and plotted on the top. The order of stability decreases from the top down. Scattered peptides were not clustered. **B** - Top 8 tight clusters sequentially selected. **C** - Top 10 tight clusters sequentially selected.

We found several different sites on the same protein share similar phosphorylation change patterns, thus end up appearing in the same cluster. For example, "_KGS(ph)FTTELSCR_" (position: 520) and "_RSS(ph)YISDTLINHQMPDAR_" (position: 238/239) on Psp1p in Cluster 3. It is possible that those phosphorylation sites are co-regulated by the same biological process. They might be closely located in protein tertiary structure and phosphorylated by the same enzyme. Another example is two adjacent phosphorylation sites, "_DQDQSSPKVEVTS(ph)EDEK_" (position: 495) and "_VEVT(ph)SEDEKELESAAYDHAEPVQPEDAPQDIANDELK_" (position: 494) on Leu1p in Cluster 4. Both of these two sites were identified in a WT/SNF1/TPK2 experiment, where the serine (S) at position 495 in the former has phosphorylation probability 0.999 (reported by MaxQuant), while the threonine (T) at position 494 in the latter has phosphorylation probability 0.962. These two sites might be competing and alternative. The dominancy of either site might be affected by protein cellular localization.

On the other hand, we also found the same protein, such as Spt6p, to be clustered in multiple functional groups. Those different sites do not necessarily change phosphorylation in a similar pattern, since they might be regulated by different biological processes. All the above observations are worth further investigation.

2.4.4 Identification of differential phosphorylation in each mutant

A total of 863 unique phosphopeptides representing 452 proteins were identified to have significant phosphorylation changes in at least one kinase-dead mutant. We can then infer the downstream proteins regulated by the kinases. The inferred regulation might be direct or indirect. A total of 1588 significant kinase-phosphopeptide regulation pairs were identified (Dataset S2). Sixty-one pairs of them directly contain the components in the known pathway map (Figure 2.1). We incorporated these proteins and generated an extended pathway map (Figure 2.5).

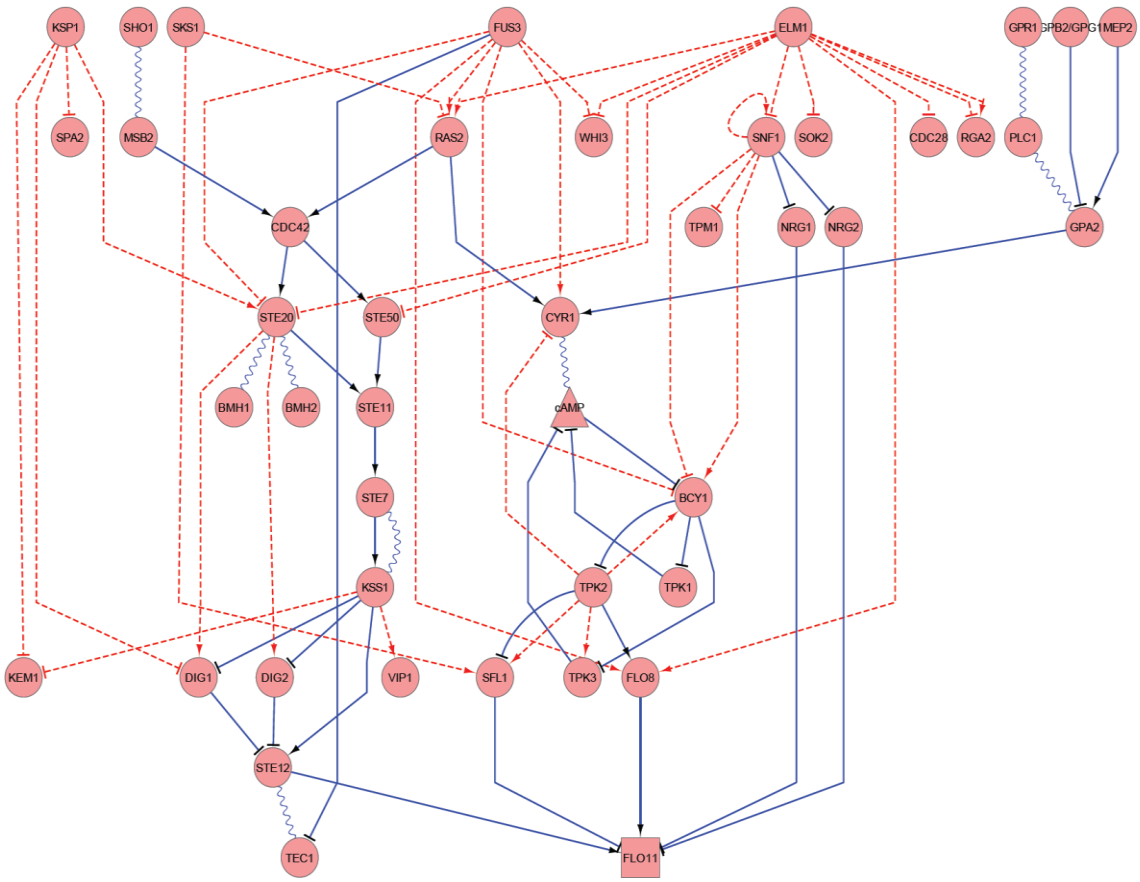


Figure 2.5 Extended filamentous pathway map.

The extended filamentous growth pathway map integrating the known knowledge (Figure 2.1) and the regulation inferred from significant differential phosphorylation in individual KDs. The inferred regulation might be direct or indirect.

The ellipses are proteins; the rectangles are DNAs; and the triangles are metabolites. The linkage between shapes: → indicates stimulation, —| indicates inhibition, and ∿∿ indicates association. Solid lines indicate physical interactions, while dashed lines indicate changes in phosphorylation.

2.4.5 Phosphopeptides with globally significant phosphorylation changes

A total of 28 phosphopeptides representing 26 proteins from the entire dataset were found to have globally significant phosphorylation changes. These candidates were picked out without using prior knowledge. The Fisher's probability test [99] was extended to allow missing values (see **Materials and Methods**), and it was used for detecting global significance. Each selected phosphopeptide satisfies the criterion: the combined p-value < 0.05, q-value < 0.05 for controlling false discovery rate (FDR) [100], and the significance B value < 0.05 in at least 4 out of 8 kinase-dead mutant (KD) versus wild type (WT) conditions. The combined p-value is a measure of global significance,

while the significance B value [24] is a measure of significance in an individual experiment. Five of the globally significant phosphopeptides, Nth1p, Hsp42p, Pbi2p, Rcn2p and Pdr12p, came with complete measurements (Table 2.4). We consider them have high-confidence. Another adaptively weighted statistic [101] was applied to all complete measurements for validation. Adopting the same selection criterion as above, Nth1p, Pbi2p, Rcn2p and Pdr12 were again identified as globally significant. Retrospective and prospective evidence has been found to support some of our predictions.

Table 2.4 Globally significant phosphopeptides selected from the complete measurements (high-confidence).

ENSEMBL ID [112]	Standard name	Name description[a]	Modified sequence	Stress response
YDR001C	NTH1	Neutral trehalase;Alpha,alpha-trehalase;Alpha,alpha-trehalose glucohydrolase	_RGS(ph)EDDTYSSSQGNR_	Nth1p is a multiple stress responsive protein [113,114].
YNL015W	PBI2	Protease B inhibitors 2 and 1;Proteinase inhibitor I(B)2	_HNDVIENVEEDKEVHT(ph)N_	Pbi2 gene deletion leads to decreased resistance to hyperosmotic stress . [115].
YOR220W	RCN2	Regulator of calcineurin 2;Weak suppressor of PAT1 ts protein 1	_NKPLLSINT(ph)DPGVTGVDSSSLNK_	Rcn2p is Induced in response to DNA-damaging agent methyl methanesulphonate [116].
YPL058C	PDR12	ATP-dependent permease PDR12	_HLSNILS(ph)NEEGIER_	Pdr12 is strongly induced by weak acid stress [117] and is a target of the transcription factor War1p [118] which elicits weak organic acid stress adaptation through active efflux [119,120].
YDR171W	HSP42	Heat shock protein 42	_KS(ph)S(ph)SFAHLQAPSPIPDPLQVSKPETR_	Protein expression is induced by stresses such as heat shock, salt shock and starvation [121].

[a] Annotated with MaxQuant

Nth1p is a key enzyme in the trehalose pathway which plays a crucial role in glucose homeostasis and stress responses [113,122] and is a substrate phosphorylated for both Tpk1p and Tpk2p [123]. The *NTH1* gene also has been reported to have genetic interactions with the *TPK1* and *TPK2* genes [124]. It has been reported to physically

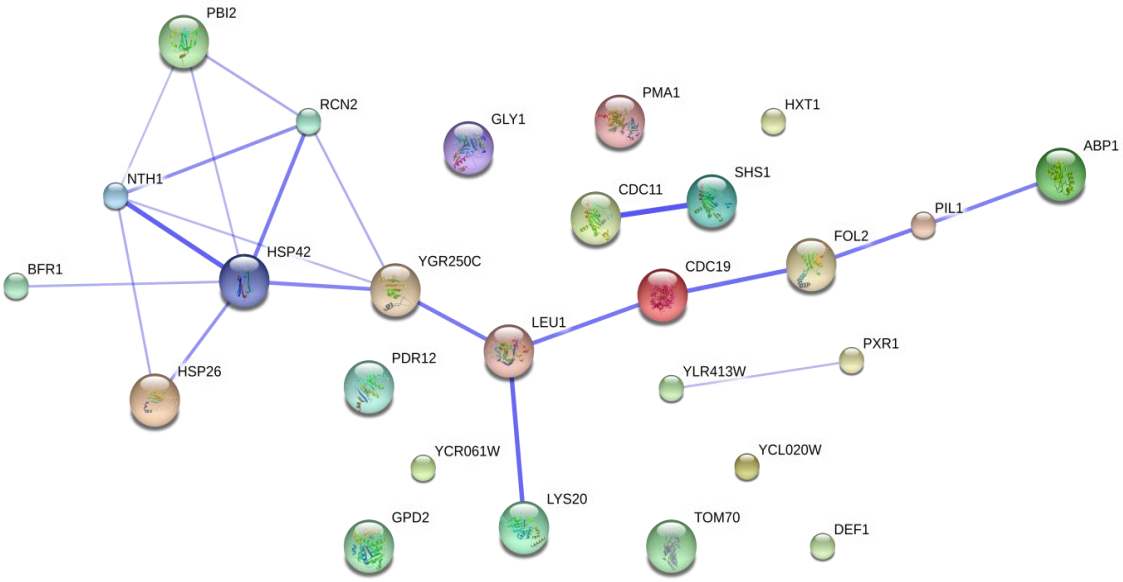
interact with the kinase Sks1p [7] and with Bmh1p [125]. The above direct interactors of Nth1p, *i.e.*, Tpk1p, Tpk2p, Sks1p and Bmh1p, are all known to play roles in filamentous growth [81,126–130]. The Rcn2p protein was also reported to physically interact with Bmh1p [125], which associates with the Ste20p protein involved in filamentous growth [130,131]. Bmh1p may also interact with Tpk1p [132–134]. Thus, Nth1p and Rcn2p have been closely associated with a number of proteins known to be involved in filamentous growth. Hsp42p has physical association with Fus3p [125], and its expression is induced under starvation [121]. The remaining two proteins in Table 2.4 have not yet been closely linked to filamentous growth but play roles in other stress responses and represent new leads.

We also searched the STRING database [91,92] to investigate the inner connections between the 26 globally significant proteins (shown in Figure 2.6). STRING assigns the confidence of protein-protein interactions integrating high-throughput experiments, genetic context, co-expression and other previous knowledge. In Figure 2.6, 17 proteins, including Nth1p, Hsp42p, Rcn2p, Pbi2p, Hsp26p, Bfr1p, YGR250C protein, Leu1p, Lys20p, Cdc19p, Fol2p, Pil1p, Abp1p, Cdc11p, Shs1p, YLR413W protein and Pxr1p, have direct or indirect connections. It presents a closely inter-connected sub-network embodying Nth1p, Pbi2p, Rcn2p, Hsp42, YGR250C protein and Hsp26p.

2.4.6 Correlation network analysis

All possible pairs among the 73 common phosphopeptides with complete measurement were tested using the Pearson correlation. A total of 45 strongly correlated phosphopeptide pairs were identified, each satisfying the following criteria: the correlation test p-value < 0.05, and the stringent requirement of $|\text{Pearson correlation coefficient}| \geq 0.9$. Detailed information on the 45 pairs of phosphopeptides is provided in Dataset S4. Twenty-seven of the pairs have positive correlations, while 18 pairs have negative correlations. A stringent protein correlation network containing 35 proteins (Figure 2.7) was generated by connecting the strongly correlated peptide pairs and then tracing the peptides back to their parent proteins.

A Confidence view



B Evidence view

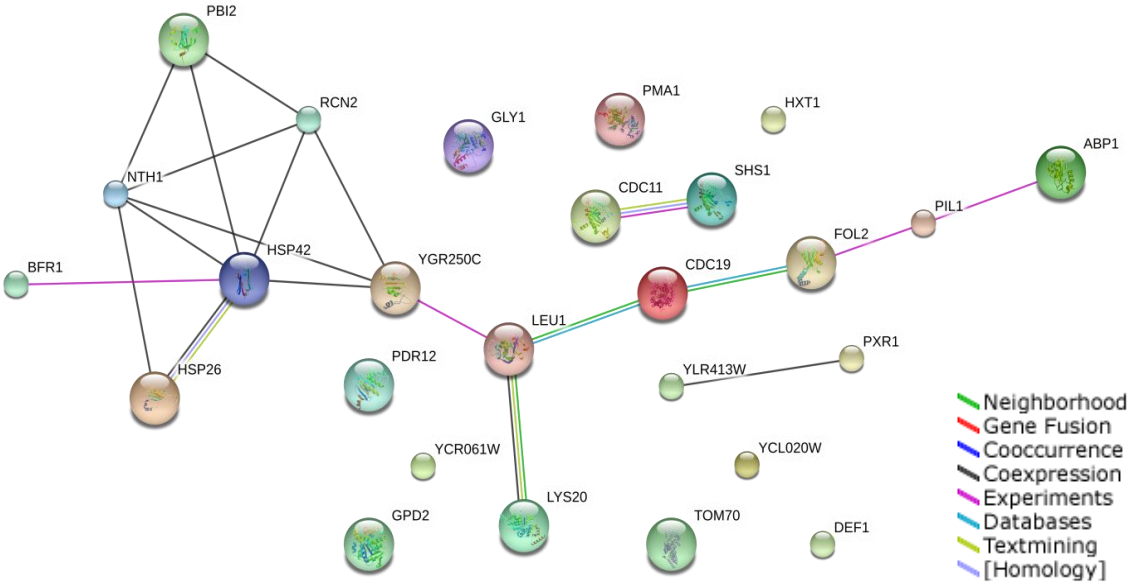


Figure 2.6 STRING reported inner connections of the globally significant proteins.

The network was generated using STRING v9.0 [91,92] using the default parameters with median confidence. **A** – Confidence view. The thicker the edge it, the more confidence the interaction is. **B** – Evidence view. Different colors of the edges indicate different evidence types.

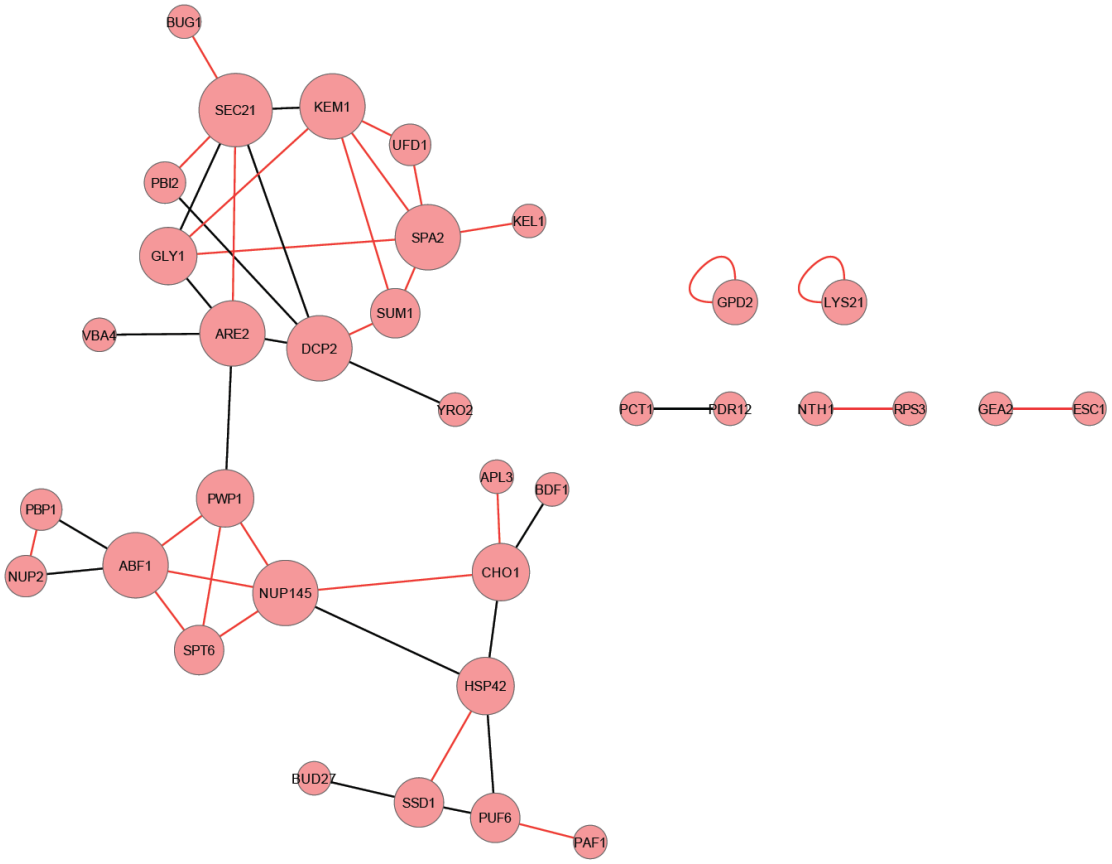


Figure 2.7 Stringent correlation network of phosphoprotein pairs.

Red lines indicate positive correlations, while black lines indicate negative correlations. The larger the node size, the greater the degree of connectivity.

Identifying core-components in the correlation network

In the protein correlation network, proteins with the highest degrees of connectivity are considered core components in the network. The 19 proteins having degrees greater than 1 (protein self-connection ignored) in the stringent protein correlation network were predicted to be core components of the network. Detailed descriptions and evidence of the proteins are summarized in Table 2.5. Kem1p, Spa2p and Spt6p have been reported to be directly involved in filamentous growth in previous literature. Six other proteins, Are2p, Dcp2p, Hsp42p, Ssd1p, Sum1p and Ufd1p, have reported evidence in terms of genetic and/or physical interactions with known components of filamentous growth. The remaining proteins have been implicated in various stress responses, including the unfolded protein response (*e.g.*, sensitivity to

tunicamycin), osmotic shock, and thermal shock, but not previously linked to filamentous growth. Our predictions, Pbi2p and Pbp1p, have never been reported as involved in filamentous growth, however, they are validated by our experiments (see ***Experimental validation*** in **Results**). The invasive growth assay of haploid strains (under 1% butanol treatment, which would induce filamentous growth in wild type strains) shows that the deletion of *PBI2* results in decreased invasive growth. In diploid strains, *PBP1* deletion strain does not form filamentous growth under nitrogen. The results indicate that Pbi2p and Pbp1p are involved in filamentous growth, as well as stress response.

Gpd2p and Lys21p are two self-connected proteins. The self-connection was built up by two distinct phosphorylation sites on the protein. Gpd2p have not been related to filamentous growth in *Saccharomyces cerevisiae*. Its homolog Gpd2p in *Candida albicans*, is involved in core stress response, and is induced upon pseudohyphal growth [42–49].

Table 2.5 Phosphoproteins having degrees of connectivity greater than 1 in the stringent correlation network.

Index	Standard name	Degree of connectivity	Evidence of involvement in filamentous growth	Remarks	Stress response
1	SEC21	6	-	In the stringent correlation network, SEC1 correlated with KEM1 (known involved in filamentous growth), and ARE2 (indirect evidence of involvement).	*SEC21* overexpression leads to decreased rapamycin resistance. [135]. Sec21p mutants have reduced resistance to tunicamycin decreased [136].
2	ABF1	5	-	In the stringent correlation network, ABF1 correlated with SPT6 (known involved in filamentous growth), and PBP1 (validated by our experiment).	Implicated in oxidative stress [137]. Abf1p mutants exhibit decreased sensitivity to tunicamycin [137].
3	ARE2	5	Are2p has positive physical interactions with Fus3p, Tpk1p and Ste20p [7].	Several other proteins involved in sterol biosynthesis or response are differentially phosphorylated in the kinase-dead mutants	Are2p is linked to cell wall construction and plays a key role in sterol biosynthesis whose regulation

				(YML008C, YHR073W, YKL140W)	is important in specific stress responses [138,139].
4	DCP2	5	(1) Dcp2p has physical interactions with Spt6p [140] and Kem1p [140–142], both are involved in filamentous growth. (2) Dcp2p is a phosphorylation substrate of Ste20p [143].	In the stringent correlation network, DCP2 is highly correlated with SUM1 (positive) and ARE2 (negative).	Dcp2p is involved in stress granule assembly [143].
5	KEM1 (Alias XRN1)	5	Kem1p plays a direct role in yeast filamentous growth, affecting *FLO11* transcription [144].		*KEM1* deletion has increased sensitivity to hyperosmotic stress [145].
6	NUP145	5	-	In the stringent correlation network, NUP145 is highly correlated with SPT6 (known involved in filamentous growth) and HSP42(indirect evidence).	*NUP145* deletion has decreased resistance to sodium arsenite [146]. Reduced functioning of Nup145p causes the strain to also have decreased resistance to tunicamycin [136].
7	SPA2	5	When *SPA2* is disrupted, filamentous growth decreases [147]. Refer to SGD.		Required for recovery from osmotic stress [148].
8	CHO1	4	-		
9	GLY1	4	-	In the stringent correlation network, GLY1 has positive correlation with KEM1 and SPA2, both known to be involved in filamentous growth.	The *GLY1* deletion strain has decreased resistance to hyperosmotic stress [115].
10	HSP42	4	Hsp42p has physical association with Fus3p [125].		Protein expression is induced by stresses such as heat shock, salt shock and starvation [121], which might also contribute to the phosphorylation level change that we detected.
11	PWP1	4	-	In the stringent correlation network, PWP1 has positive correlation with SPT6.	Pwp1p mutants have decreased resistance to tunicamycin [136].
12	PUF6	3	-		The *PUF6* deletion strain has decreased resistance to

					multiple chemical stresses [103,149,150] and is more sensitive to both cold [151] and heat [152].
13	SPT6	3	(1) Filamentous growth decreases when *SPT6* is disrupted [147]. Refer to SGD. (2) *SPT6* has a genetic interaction with *RAS2* [153] which regulates filamentous growth [154]. (3) Spt6p has physical interaction with Kss1p [7].		Spt6p is involved in regulation of transcription from RNA polymerase II promoter in response to stress [155].
14	SSD1	3	(1) Ssd1p has physical interactions with Ste20p [7] and Kem1p [156]. (2) *SSD1* has positive genetic interaction with *KEM1* [157], and negative genetic interaction with *STE50* [158], known to be involved in filamentous growth [159]. (3) *SSD1* also has negative genetic interaction with *ILM1* [160]. Ilm1p is required for slowed DNA synthesis-induced filamentous growth [161,162]. (4) Overexpression of the *SSD1* homologue, ropy, in N. crassa has been shown to suppress mutations of POD6 and COT1 which play essential roles in hyphal tip extension [163], a process required for filamentous growth.		Required for thermotolerance [164] and migrates to stress granules [165].
15	SUM1	3	*SUM1* has a negative genetic interaction with *ELM1* and *TPK3* [166].	In the stringent correlation network, SUM1 has positive correlation with SPA2 and KEM1, both known to be involved in filamentous growth.	Involved in osmotic stress [167]. Predicted to be involved in stress response [168].
16	NUP2	2	-	In the stringent correlation network, NUP2 are positively correlated PBP1(predicted and validated by our experiment).	Nup2p is involved in mRNA export from nucleus in response to heat stress [169]. When Nup2 is deleted, the strain has decreased resistance to bortezomib [170] and arsenite(3-)

					[171].
17	PBI2	2	-		*PBI2* deletion has decreased resistance to hperosmotic stress [115] and multiple chemicals [103,149,172].
18	PBP1	2	Pbp1p has not been reported related to filamentous growth. However, our experiment has validated that *PBP1* mutant strain does not have filamentous growth under 1% butanol treatment, which will induce filamentous growth in WT strain.		Pbp1p is involved in stress granule assembly [173]. *PBP1* mutant has decreased resistance to stress.
19	UFD1	2	*UFD1* has a negative genetic interaction with *KEM1* [174].	In the stringent correlation network, UPD1 has positive correlation with KEM1 and SPA2, both known to be involved in filamentous growth.	Involved in ER and heat stress responses [175,176].

2.4.7 Literature mining

In addition to the candidate proteins predicted from our dataset, we retrieved from the literature and authoritative databases [28,80–82,102,103] a list of proteins involved in filamentous growth. A total of 69 unique proteins, not all being phosphoproteins, were extracted, and 20 of them have been detected in our phosphoproteome dataset. Among those, 15 proteins, including Bcy1p, Cdc28p, Cyr1p, Dig1p, Dig2p, Flo8p, Kem1p, Ras2p, Sfl1p, Snf1p, Spa2p, Ste20p, Ste50p, Tpk3 and Tpm1p, showed significant phosphorylation changes in at least one kinase-dead mutant, and are displayed in our extended pathway map (Figure 2.5).

2.4.8 Causal Bayesian network

The interactions retrieved from the differentially phosphorylated proteins in individual kinase-dead mutants (the dashed edges in Figure 2.5) did not make use of phosphorylation change pattern over different kinase-dead mutants, and the protein pairs must contain a mutated kinase. In contrast, the correlation network is a network of the common peptides, taking into account the protein responses in all the kinase-dead

mutants, and the correlated protein pairs do not necessarily contain the mutated kinases. Note that this network is not directed and more information may be gleaned from a causal analysis. We implemented causal relationship discovery to detect the direction of influences between proteins with the understanding that the relationships may be direct or indirect. A total of 46 unique proteins were selected to construct the network. All of them are listed in Table 2.6.

Table 2.6 Focus proteins used for causal relationship discovery.

Mutated kinases	Globally significant (high-confidence)	Hub proteins (high-confidence)	From literature mining and detected in our dataset
KSP1	NTH1	SEC21	BCY1
KSS1	PBI2	ABF1	BMH1
SKS1	RCN2	ARE2	BUD2
STE20	PDR12	DCP2	CDC28
SNF1	HSP42	KEM1	CYR1
TPK2		NUP145	DIG1
ELM1		SPA2	DIG2
FUS3		CHO1	FLO8
		GLY1	GPR1
		HSP42	KEM1
		PWP1	NRG1
		PUF6	PEA2
		SPT6	RAS2
		SSD1	SFL1
		SUM1	SNF1
		NUP2	SPA2
		PBI2	STE20
		PBP1	STE50
		UFD1	TPK3
			TPM1

Bayesian network modeling identified causal influences for 22 protein pairs (44 phosphopeptide pairs), satisfying the posterior probability of the relationship greater than 0.5. The network comprising all the causal relationships is presented in Figure 2.8. Among those, only 6 protein pairs have the posterior probability higher than 0.7. The other protein pairs do not have high probability since the samples available for training

the model is limited due to the missing data issue caused by instrument limitation. The arrows in Figure 2.8 only indicate the existence of causal influence, but do not specify whether the influence is activation or inhibition. The causal relationship discovered might be between proteins that are not immediately adjacent in pathways so the relationship could be quite indirect.

Through another inspection of the phosphorylation change patterns of the peptide pairs detected with relatively strong causal influences (posterior probability higher than 0.7), we observed that: Ste20p has opposing phosphorylation changes compared to Are2p, Pdr12p and Sec21p; two phosphopeptides (the same amino acid sequence but different phosphorylation sites) on Hsp42p present opposing phosphorylation changes compared to Ste20p; and Pbp1p presents consistent phosphorylation change compared to Ste20p. With caution we predict that the opposing pattern implicates an inhibitive influence of Are2p, Pdr12p and Sec21p to Ste20p; and similarly, inhibition of Hsp42p to Ste20p; while Pbp1p shed activating influence to Ste20p. Again, we emphasize that the influence might be quite indirect and even be influenced by multiple pathways.

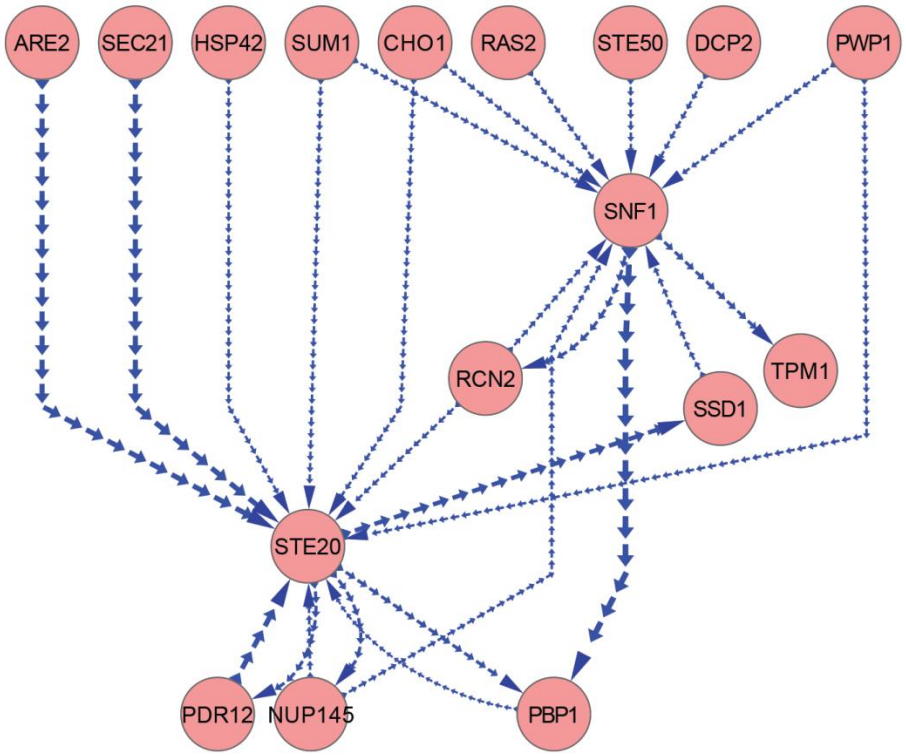


Figure 2.8 Causal Bayesian network.

Each edge →→→ indicates a potential causal influence between proteins, which might be a direct or indirect influence. It does not distinguish activation and inhibition. The thicker the edge, the higher the posterior probability.

2.4.9 Experimental validation

Five predicted novel high-confidence proteins, Nth1p, Pbi2p, Pdr12p, Rcn2p and Pbp1, have been tested in invasive growth assays.

Mutant strains compared to WT haploid

Gene deletion strains were compared with Y825 wild type strains (WT haploid) under 1% butanol treatment, which will induce filamentous growth in Y825. Invasive growth assays were done by plate-washing. The plate after wash is shown in Figure 2.9A. From our results, the deletion of PBI2 results in decreased invasive growth, while deletion of NTH1, PDR12, and RCN2 results in increased invasive growth. In haploid yeast, filamentous growth refers to invasive growth, which allows cells to grow penetrating the medium [177].

Mutant strain compared to WT diploid

In addition, in experiments of diploid strains, PBP1 deletion strain does not show filamentous growth under nitrogen stress, while wild type diploid strain (WT diploid) forms filamentous growth (Figure 2.9B).

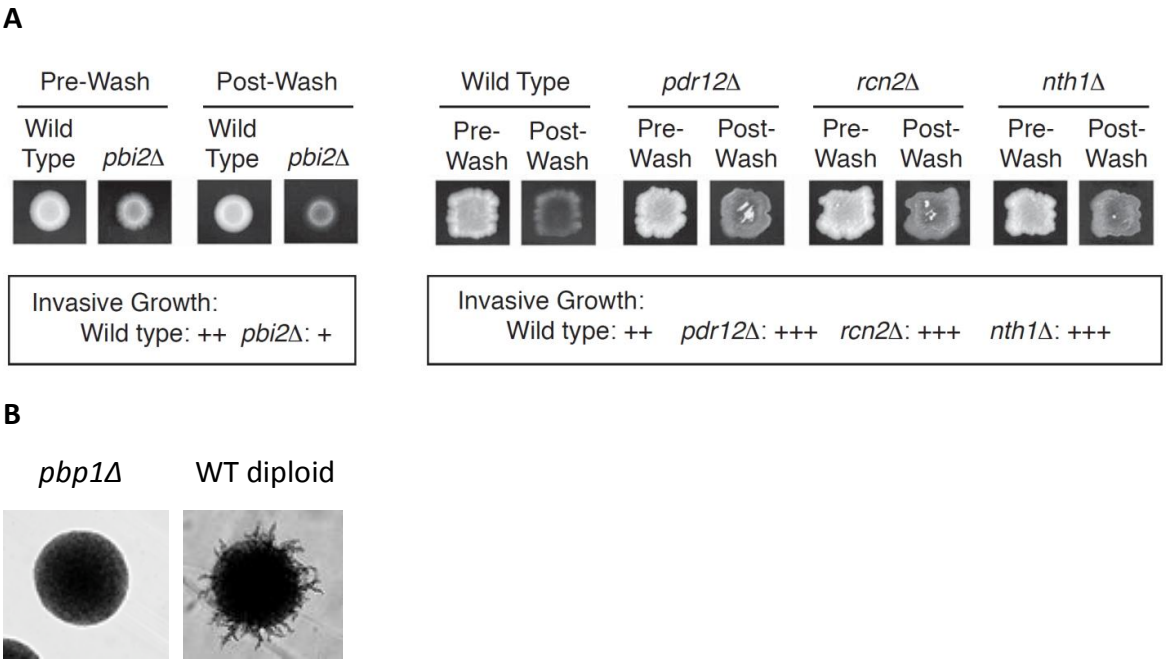


Figure 2.9 Phenotypic result of five deletion strains and wild type strains.

A - Mutant strains compared to WT haploid, **B** - Mutant strain compared to WT diploid.

In summary, all of the deletions result in differential filamentous growth compared to wild type controls, which means they are involved in filamentous growth. The results have validated our approach to identification of candidate proteins in this biological system from phosphoproteomics data alone.

2.5 Discussion

In this study, we demonstrate that interventional phosphoproteome studies can provide new insight into signaling pathways involved in biological processes, such as yeast filamentous growth. In order to increase sensitivity to smaller changes in phosphorylation relative to previous yeast global phosphoproteome studies [11,12], we used SILAC, an isotope labeling approach. Isotope labeling approaches are generally more accurate relative to label-free approaches [14], but require greater resources to

implement, resulting in trade-offs between precision and missing data due to sampling limitations. We proposed and developed a comprehensive computational and statistical analysis pipeline for the post-identification studies of phosphoproteome data. The analyses are aimed at discovering candidate components of significant pathways involved in filamentous growth as well as the potential targets of the pathways, and to provide more information on the signaling network structure by monitoring changes in phosphorylation in response to mutational interventions. We applied the pipeline to analyze our interim yeast phosphoproteome datasets and a total of 882 unique phosphopeptides representing 486 proteins were identified as significantly influenced by at least one out of 8 kinase-dead mutants. Twenty-eight unique phosphopeptides having globally significant phosphorylation were identified from the whole dataset among which 5 peptides representing 5 proteins, Nth1p, Pbi2p, Rcn2p, Pdr12p and Hsp42p, were identified as high-confidence candidates. Nineteen candidate proteins with relatively high degrees of connectivity were selected in the stringent correlation network. Among the high-confidence candidate proteins, 3 proteins have been reported to be directly involved in filamentous growth and another 6 proteins were also supported, in terms of genetic and physical interactions with known components involved in filamentous growth. The remaining proteins have been implicated in various other stress responses and may play roles in filamentous growth or may be secondary stress responders. Pbp1p is one of our novel predictions, and it has been validated to be involved in filamentous growth by our mutational phenotypic experiments. Causal relationship discovery was further performed on the candidates and validated proteins. The inferred causal relationships, along with the interactions inferred from phosphorylation changes in response to individual mutants, form phosphoprotein interaction networks, which suggested potential proteins to be intervened in future studies.

Each of the kinases mutated in this study had previously been implicated in filamentous growth. Many of these kinases are known to also affect pathways that are not involved directly in filamentous growth. However, the proteins which change

phosphorylation level in response to multiple mutants are reasonable candidates involved in filamentous growth. The sensitivity of such detection is constrained by the degree of overlap between pathways, the coverage of pathways by the mutants, and the extent of missing data issue. Upstream components of isolated pathways may be missed, while downstream core components are more likely to be identified.

A remaining challenge for quantitative phosphoproteome analysis arises from the sampling limitations and resolution of mass spectrometers [14]. This feature of tandem mass spectra of complex mixtures results in poor overlap of peptides identified among samples unless a relatively large number of replicate experiments are carried out (which is time consuming and often economically impractical for large-scale projects). For this reason, a significant number of missing values exist in these datasets which can obscure potential candidates for further validation studies. We developed methods to partially compensate for missing data issue. In the phosphopeptide meta-analysis, an extension of Fisher's combined probability test was made to relax the restrictions of complete measurements. The causal network modeling component was also developed to allow missing values without excluding the incomplete measurements. We also performed cluster analysis of phosphopeptides. Instead of adopting traditional clustering methods, we directly identified the most stable clusters using missing value-imputed data. Our approach was able to pick out significantly enriched functions, and identify a number of reliable candidate proteins for further validation.

This analysis pipeline has been developed to study the yeast filamentous growth pathways; however, the methodology is not limited to yeast or this biological process. It can be applied to other complex organisms to facilitate investigation into various biological processes. We anticipate the methodology to be applicable as well to other interventional studies via different experiment platforms.

Chapter III

Inferring Kinetic Networks and Parameters from Time-course Data

3.1 Introduction

Apart from investigating equilibrium-state proteomics data, I have also made an effort to study the methods for analyzing time-course biochemical data. Time-course data is a richer source of information compared to equilibrium-state data and a broader range of tools can be applied to these datasets.

Time-course data measuring concentrations of chemical species can be used to investigate the kinetics of reactions whether those species are small molecule metabolites or phosphoproteins. Time-course data is more information-rich than equilibrium-state data. For instance, transient behaviors and oscillations can be detected by time-course data, but might not be detected by equilibrium-state data. It is also possible to take advantage of the time dimension and the implicit relationships between time points. Studies utilizing time-course data are carried out on different experimental platforms, such as nuclear magnetic resonance (NMR) [178], microarrays [179], and mass spectrometry (MS)-based proteomics experiments [17]. However, the task of reconstructing biochemical networks from time-course data is challenging, because of the large amount of potential combinatorial interactions between chemical species, and potentially unknown parameters related to those interactions. Several approaches have been proposed to solve this challenge. Yeung *et al.* [76] used singular value decomposition and robust regression to reconstruct connectivity topology of large sparse gene networks. They only consider the gene network dynamics near steady state.

Ross and co-workers [180,181] proposed a method for determining causal connectivities of species, based on the systematic response due to a pulse change of one species. Fromentin *et al.* [182] presented a hybrid model mixing delay parameters and an ordered pattern of concentration peaks to infer networks. Feng and co-workers [69,70] applied the Granger causality approach to infer interactions between species in the frequency domain. Schnell's team [30,63,183] developed MIKANA (Method to Infer Kinetics And Network Architecture) Ver. 1, a computational tool with a user interface, which not only infers network topology of reactions, but also estimates the kinetic parameters. Its major components include pseudo-linear model generation, model selection, and parameter fitting, which altogether serve as a tool for the reconstruction of reaction mechanisms from time-course data. In this chapter I present an improved version of this tool, MIKANA Ver. 2. It maintains the three-component framework of MIKANA, and makes improvements and extensions to each component.

3.1.1 The MIKANA framework

The MIKANA framework is shown in Figure 3.1. MIKANA considers only elementary reactions which, when combined, can create complex reaction mechanisms. An elementary reaction is a single step reaction, in which one or more chemical species react directly to form products without intermediates; and it has a single transition state [184]. Given time-course data as input, the elementary reaction generator identifies the species and generates a set of all the possible realistic elementary reactions between them (up to predefined and customized restrictions). MIKANA supports exclusion of specific reactions from the set if specified. The elementary reactions will be expressed using differential equations, whose terms are filled into a model design matrix. The derivative vector generator generates the concatenated velocity vector for all species over time. Then a pseudo-linear model can be constructed using the model design matrix and the concatenated velocity vector. The initial full model contains all the possible elementary reactions. The subsequent step is the iterative procedure to select the optimal subset of elementary reactions and estimate their kinetic parameters, *i.e.* rate constants. When an optimal subset is determined, the ODE (ordinary differential

equation) generator generates corresponding differential equations. Fitted data can be simulated from the ODEs. Results can be outputted and plotted.

The elementary reaction generator of MIKANA Ver. 1 can support most realistic zero-order to second-order elementary reactions. It uses the general-to-specific method [30,63] to select the optimal reduced model, that is, it starts from the full model, and shrinks the model size by one iteratively, until the cost function is optimized. Srividhya *et al.* [63] verified its superior performance relative to the specific-to-general approach, which starts from single-reaction models and increases the model size by one iteratively until the cost function is optimized. MIKANA Ver. 1 performs model parameter fitting using linear least square method.

3.1.2 Motivation of improvements

3.1.2.1 *Autocatalytic reactions and third-order reactions*

MIKANA Ver. 1 has performed well on simple non-oscillatory models presented in [30] fitting zero-order to second-order reactions, which contains 0 to 2 reactant molecules (stoichiometry). However, for the models containing oscillations or other behavior contributed by autocatalytic reactions, it does not fit well. The following is an example with two species.

Example 1 – A simple oscillatory model

This model system contains two species y_1 and y_2. They are involved in a second-order autocatalytic reaction, and have mass exchange with external environment. The reactions are schemed as:

$$\begin{aligned} &\xrightarrow{k_1} y_2, \\ y_1 &\xrightarrow{k_2}, \\ y_1 + y_2 &\xrightarrow{k_3} 2y_1. \end{aligned} \qquad (Model\ 3.1)$$

When the rate constants $k_1 = 0.30$, $k_2 = 1.30$, $k_3 = 0.64$, the system shows damped oscillation. I simulated the data as input with initial values [1, 1] in time range [0, 50]. MIKANA Ver. 1 can detect the reactions; however, the fitted curve deviates from the input (Figure 3.2), with predictions $\hat{k}_1 = 0.2120$, $\hat{k}_2 = 1.7252$ and $\hat{k}_3 = 0.4382$.

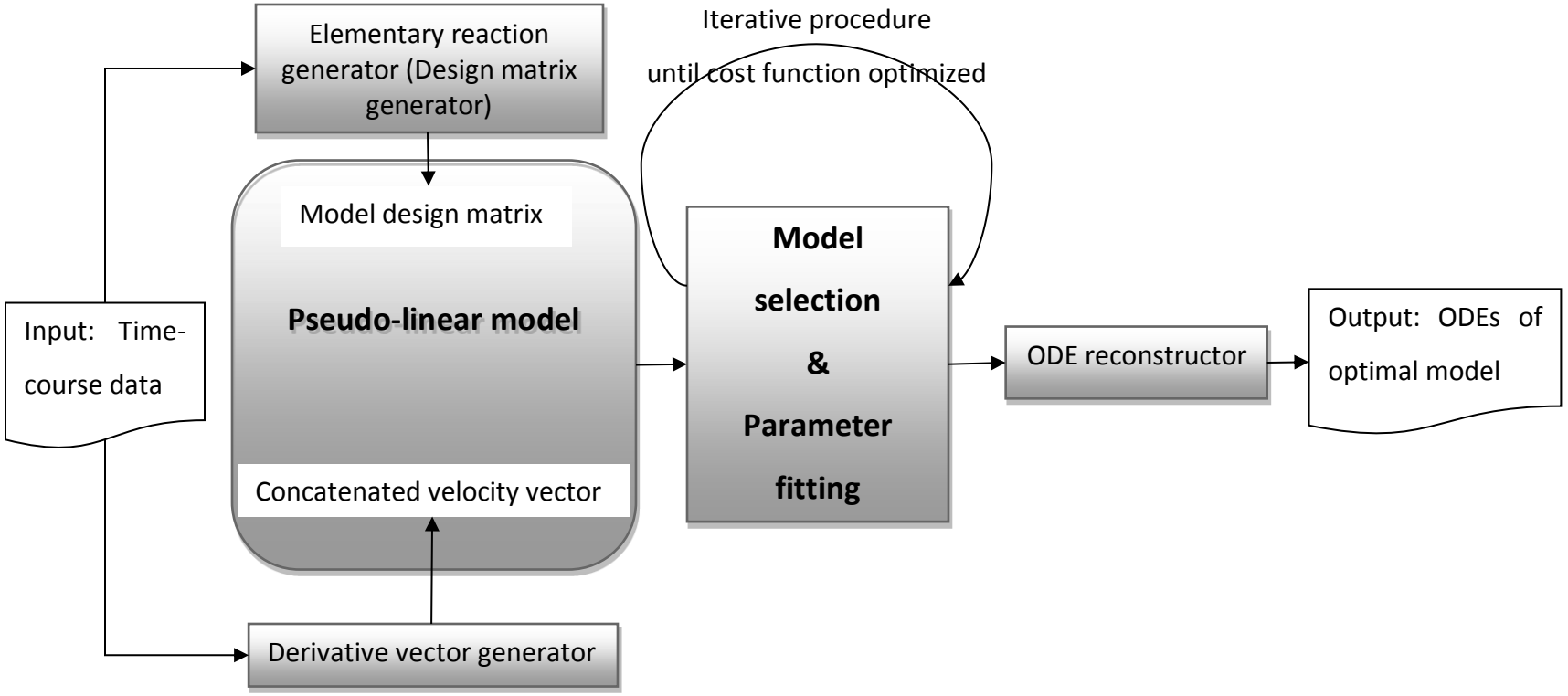


Figure 3.1 Algorithm framework.

Prediction of reactions containing more species will be more challenging. In order to obtain better prediction, I improve the design matrix in MIKANA Ver. 2 to better support autocatalytic reactions (see ***Pseudo-linear model generation*** in **Materials and Methods**). Besides, known elementary reactions contain no more than three reactant molecules [185–187]. Several forms of third-order elementary reactions are also realistic. Thus in MIKANA Ver. 2, I extend the design matrix to support them as well.

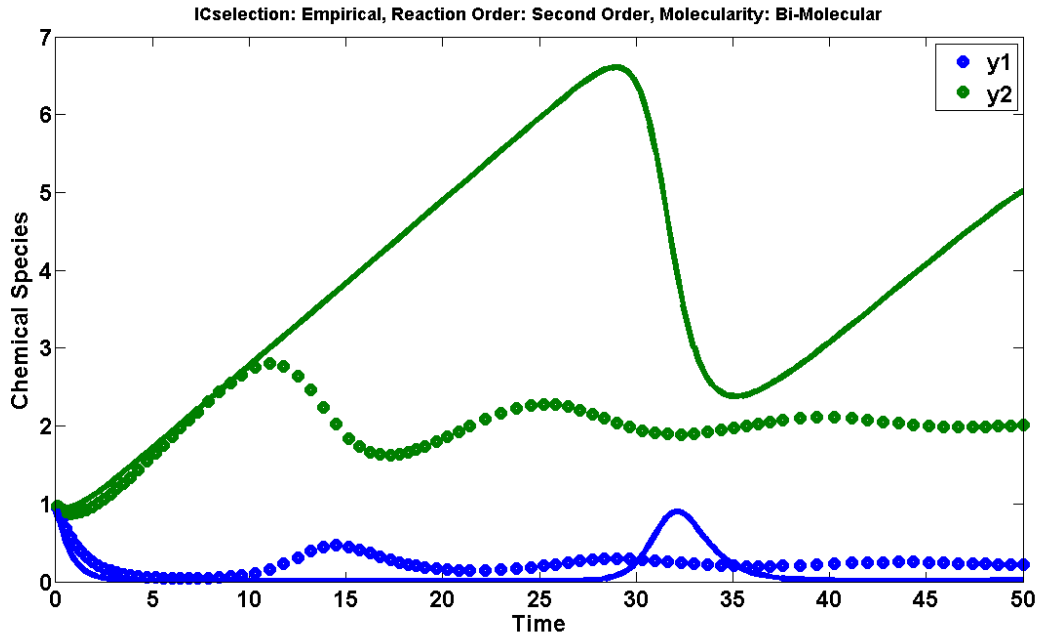


Figure 3.2 Input/Output plot for the simple oscillatory model, fitted by MIKANA Ver. 1.

Dots indicate input data, solid lines indicate fitted curves.
The Initial reaction set contains all the zero-order to second-order reactions supported by MIKANA Ver. 1.
Empirical information criterion is used for model selection.

3.1.2.2 *Model selection and parameter fitting*

As illustrated in Figure 3.1, the model selection and parameter fitting are coupled in an iterative procedure. MIKANA Ver. 1 adopts the general-to-specific approach to perform model selection [30,63]. All the possible elementary reactions are considered in the initial model, and then the model is reduced by discarding reactions iteratively, until the defined cost function is minimized. This approach outperforms the

specific-to-general approach, which starts from a single elementary reaction then increases model size until stop criterion is met [63]. MIKANA Ver. 1 uses linear non-negative least square algorithm [188] to estimate the rate constants. Since each iterative step is performed based on the result of its previous step (called stepwise subset selection), the algorithm might be trapped in local optima. The computing time will grow exponentially with the number of species [183].

In MIKANA Ver. 2, I seek an alternative optimization solution, which does not trap in local optima, and does not turn computationally intractable when the number of species grows large. My new solution was inspired by the Lasso [189], which uses L_1-norm penalized regression, and directly assigns less contributing coefficients to 0 hence gives interpretable parsimonious models (see ***Model selection and parameter fitting*** in **Materials and methods**). I add non-negative coefficient constraint to the Lasso. The non-negative Lasso is also a convex optimization problem thus any local optimum is also a global optimum. The computing time of non-negative Lasso is determined by the number of values the penalty tuning parameter takes, and does not grow exponentially with the number of species. Non-negative Lasso is used to select the optimal subset of reactions. Non-linear non-negative least square regression is used to fit the rate constants of the reduced models. It can be applied to both linear and nonlinear fitting, and has better performance than linear least square method, especially for fitting oscillatory behaviors.

Besides, in order to reduce the influence of large noise, MIKANA Ver. 2 provides an option to smooth the input data and rule out outliers using robust spline smoothing [190]. The feature updates I have carried out are summarized in Table 3.1.

Table 3.1 Feature updates of MIKANA.

Algorithm component	MIKANA Ver. 1	MIKANA Ver. 2
Design matrix generation	Support zero-order, first-order and second-order elementary reactions, excluding autocatalytic reactions.	Further support third-order elementary reactions, including autocatalytic reactions.
Model selection	General-to-specific method [30,63]	Non-negative Lasso
Parameter fitting	Non-negative linear least square	Non-negative non-linear least square (LMF method [191,192] implemented)
Noise tolerance (optional)	None	Spline smoothing

3.2 Materials and Methods

The MIKANA framework is illustrated in Figure 3.1. The most important components in the algorithm are: (1) pseudo-linear model generation, (2) model selection through optimization method, and (3) parameter fitting.

The kinetic network topology can be constructed by elementary reactions between species. The problem of identifying the reactions fitting the data can be transformed into a regression or optimization problem. In simple words, from the time-course data a pseudo-linear model is posed,

$$X \cdot k = Y,$$

in which X is the model design matrix generated from differential equations, Y is the concatenated velocity vector, and k is the vector of rate constants. X and Y are directly known from the data, and k is the vector of rate constant that we want to estimate. The goal is to select an interpretable reduced model which minimizes the cost function and

fit rate constants to a reduced model. The methodology applied in MIKANA Ver. 2 is described in detail below.

3.2.1 Pseudo-linear model generation

Non-autocatalytic reactions

Let's first consider the following elementary reaction:

$$n^A A + n^B B \xrightarrow{k_i} n^C C + n^D D, \qquad (Scheme\ \mathrm{I})$$

in which k_i is the non-negative rate constant for reaction i; n^A, n^B, n^C and n^D are the number of molecules of the reactant species A, B, and product species C, D.. By the number of molecules, we mean the molecularity of the species. In the case of law of mass action, it can be identical to the stoichiometry. Then the reaction rate can be calculated:

$$rate = k_i[A]^{n^A}[B]^{n^B} = k_i\phi_i([A],[B]),$$

where $[A]$ and $[B]$ are the concentrations of A and B. Since $[A]$ and $[B]$ depend on time, $\phi_i([A],[B])$ also depends on time. For time point t_j, it can be written as $\phi_i(t_j)$. Then the velocities of A, B, C and D in reaction i at time point t_j are

$$V_A = -n^A k_i \phi_i(t_j) = \theta_{ij}^A k_i,$$
$$V_B = -n^B k_i \phi_i(t_j) = \theta_{ij}^B k_i,$$
$$V_C = n^C k_i \phi_i(t_j) = \theta_{ij}^C k_i,$$
$$V_D = n^D k_i \phi_i(t_j) = \theta_{ij}^D k_i.$$

In MIKANA Ver. 1, for a species s, an element of the model design matrix θ_{ij}^s is defined as $\sigma_i^s n^s \phi_i(t_j)$. In this model, σ_i^s has unit magnitude with either positive or negative sign: $\sigma_i^s = +1$, if species s is a product; and $\sigma_i^s = -1$, if species s is a reactant. This model works perfect for non-autocatalytic reactions. However, it ignores the case in which at least one of the product species is also a reactant, which happens in autocatalytic reactions. In MIKANA Ver. 2, I define a modified version of of θ_{ij}^s, so that it can support both non-autocatalytic reactions and autocatalytic reactions (see immediately below).

General elementary reactions

In MIKANA Ver. 2, for a species s, I define an element of the model design matrix as $\theta_{ij}^{s} = \delta_i^s \phi_i(t_j)$, in which δ_i^s equals the number of molecules of product s minus the number of molecules of reactant s in reaction i. For non-autocatalytic reactions, the element defined in MIKANA Ver. 2 is equal to the one defined in MIKANA Ver. 1.

Let's consider the following general elementary reaction in $Scheme$ II that can represent any reactions supported by MIKANA,

$$n_{react}^{A}A + n_{react}^{B}B + n_{react}^{C}C + n_{react}^{D}D \xrightarrow{k_i} n_{prod}^{A}A + n_{prod}^{B}B + n_{prod}^{C}C + n_{prod}^{D}D. \qquad (Scheme\ \text{II})$$

I have $\delta_i^A = n_{prod}^A - n_{react}^A$, $\delta_i^B = n_{prod}^B - n_{react}^B$, $\delta_i^C = n_{prod}^C - n_{react}^C$, and $\delta_i^D = n_{prod}^D - n_{react}^D$.

Assume reaction i ($i = 1, \dots, I$) is any elementary reaction in $Scheme$ II. Each species is measured at J time points. For a species s (indexed $1, \dots, S$), the velocity at time point t_j ($j = 1, \dots, J$) contributed by reaction i is

$$\delta_i^s \cdot rate = \delta_i^s \cdot k_i \phi_i(t_j) = \theta_{ij}^s \cdot k_i. \qquad (eqn.\ 3.1)$$

Therefore, the total velocity of species s at time point t_j for all reactions is

$$\frac{d[s]}{dt}|_{t=t_j} = \sum_{i=1}^{I} \delta_i^s \phi_i(t_j) \cdot k_i = \begin{bmatrix} \delta_1^S \phi_1(t_j) & \dots & \delta_I^S \phi_I(t_j) \end{bmatrix} \begin{bmatrix} k_1 \\ \vdots \\ k_I \end{bmatrix}.$$

The block of model design matrix of species s for all I reactions and all J time points is

$$X_s = \begin{bmatrix} \delta_1^S \phi_1(t_1) & \cdots & \delta_I^S \phi_I(t_1) \\ \vdots & \ddots & \vdots \\ \delta_I^S \phi_1(t_J) & \cdots & \delta_I^S \phi_I(t_j) \end{bmatrix}_{J \times I};$$

and the corresponding velocity vector is

$$Y_s = \begin{bmatrix} \frac{d[s]}{dt}|_{t=t_1} \\ \vdots \\ \frac{d[s]}{dt}|_{t=t_J} \end{bmatrix}_{J \times 1}.$$

The **full pseudo-linear model** for all species, all reactions and all time points can be written as follows:

$$\begin{bmatrix} & X_1 & \\ & \vdots & \\ & X_S & \end{bmatrix}_{JS\times I} \begin{bmatrix} k_1 \\ \vdots \\ k_I \end{bmatrix}_{I\times 1} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_S \end{bmatrix}_{JS\times 1}, \qquad (eqn.\ 3.1)$$

where $\begin{bmatrix} & X_1 & \\ & \vdots & \\ & X_S & \end{bmatrix}_{JS\times I}$ is the **full model design matrix** $\boldsymbol{X}$, $\begin{bmatrix} Y_1 \\ \vdots \\ Y_S \end{bmatrix}_{JS\times 1}$ is the **concatenated velocity vector** $\boldsymbol{Y}$, and $\begin{bmatrix} k_1 \\ \vdots \\ k_I \end{bmatrix}_{I\times 1}$ is the rate constant vector k I want estimate. With random noise $\mathcal{E}$, the model in $eqn.\ 3.1$ is written as

$$Y = Xk + \mathcal{E}. \qquad (eqn.\ 3.2)$$

Supported elementary reactions

MIKANA Ver. 1 supports most of zero-order, first-order and second-order reactions, excluding autocatalytic reactions. MIKANA Ver. 2 has been extended to support third-order reactions, and autocatalytic reactions. The forms of reactions supported by MIKANA Ver. 2 are listed in Table 3.2.

Autocatalytic reactions are an essential part of several nonlinear dynamics models , such as the Brusselator [193,194] and Oregonator [195]. Several examples of the newly supported reactions, including third-order reactions, are as follows:

$1A + 1B \rightarrow 1C + 1D$: $BrO_3^- + Br^- \rightarrow HBrO_2 + HOBr$ [196];

$2A + 1B \rightarrow 2C$: $2NO + O_2 \rightarrow 2NO_2$ [197];

$2A + 1B \rightarrow 3A$ is a reaction in the Brusselator autocatalytic model [195,198];

$2A + 1B \rightarrow B + C$: $2H + M \rightarrow M + H_2$, in which M is inert gas such as N_2 or Ar [185].

Table 3.2 Forms of elementary reactions supported by MIKANA Ver. 2.

Reaction order	Elementary reactions	
	1 reactant species (A)	2 reactant species (A and B)
Zero order	$\rightarrow 1A$	
First order	$1A \rightarrow$ $1A \rightarrow 1B$ $1A \rightarrow 2B$ $1A \rightarrow 1B + 1C$ $1A \rightarrow 2B + 1C$ $1A \rightarrow 2B + 2C$	
Second order	$2A \rightarrow 1B$ $2A \rightarrow 2B$ $\mathbf{2A \rightarrow 1A + 1B}$ $\mathbf{2A \rightarrow 1A + 2B}$ $2A \rightarrow 1B + 1C$ $2A \rightarrow 2B + 1C$ $2A \rightarrow 2B + 2C$	$\mathbf{1A + 1B \rightarrow 2A}$ $\mathbf{1A + 1B \rightarrow 2B}$ $1A + 1B \rightarrow 1C$ $1A + 1B \rightarrow 2C$ $\mathbf{1A + 1B \rightarrow 2A + 1C}$ $1A + 1B \rightarrow 1C + 2D$ $\mathbf{1A + 1B \rightarrow 1C + 1D}$
Third order		$\mathbf{2A + 1B \rightarrow 2C}$ $\mathbf{2A + 1B \rightarrow 3A}$ $\mathbf{2A + 1B \rightarrow 1B + 1C}$

A, B, C and D each represents a species. An equation in the table only represents a scheme of reactions. For instance, $1Y_1 \rightarrow 2Y_2 + 1Y_3$ and $1Y_1 \rightarrow 1Y_2 + 2Y_3$ correspond to the same reaction scheme "$1A \rightarrow 2B + 1C$", but they have different stoichiometry.

When we have more than two species in the input data, all the possible combinations between the species are considered to create the initial full set of elementary reactions.

$\rightarrow 1A$ is a reaction from external source. $1A \rightarrow$ is a reaction to external sink.

The reaction schemes highlighted in red are newly supported or improved by MIKANA Ver. 2.

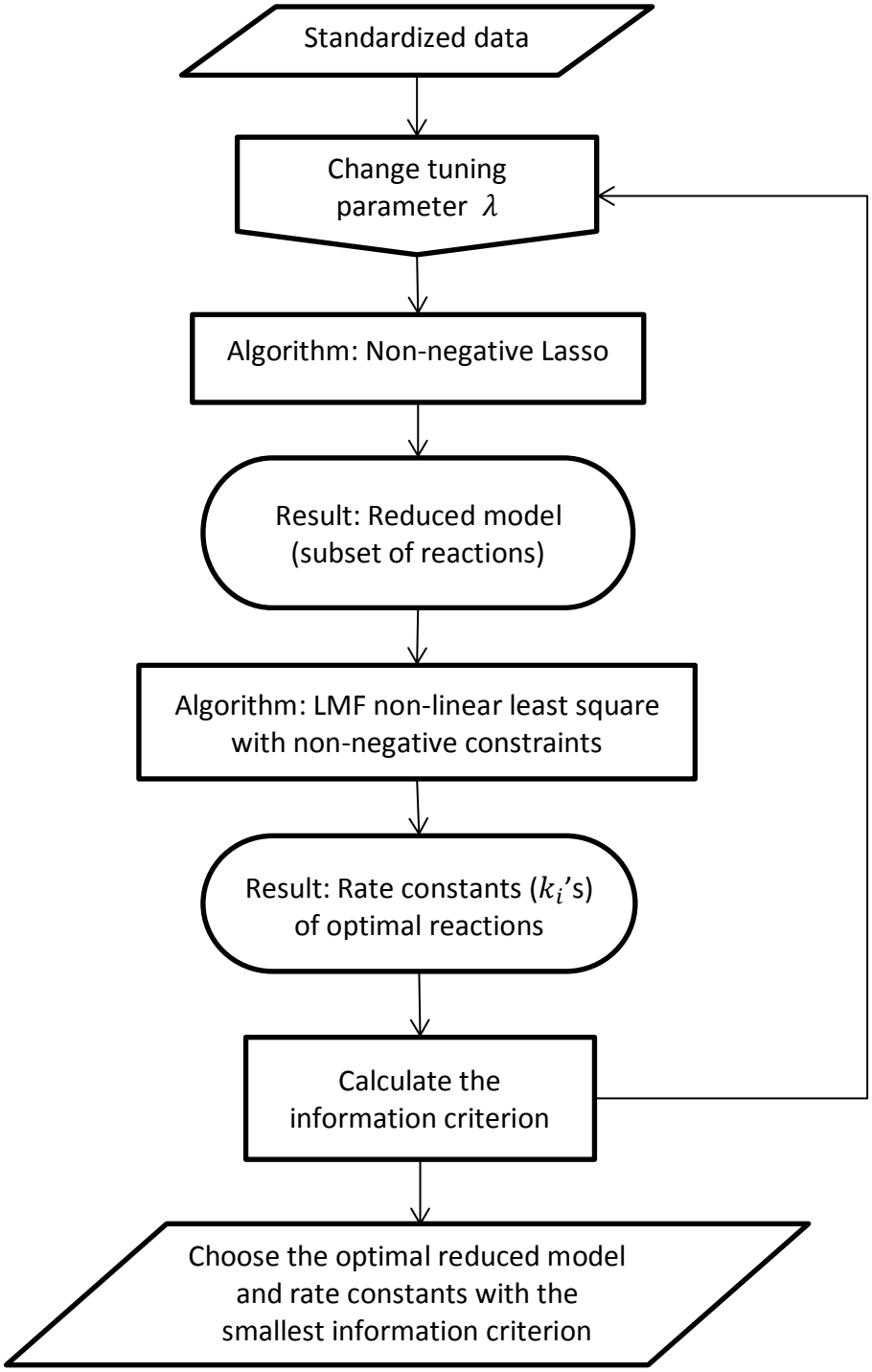


Figure 3.3 Scheme of MIKANA Ver. 2 for model selection and parameter fitting.

λ is the tuning parameter defined in $eqn.\ 3.4$. It is tuned to have increasing values from 0.01 to 0.1 with step 0.01 (unless otherwise stated). For each λ value, one reduced model is selected by the non-negative Lasso. Then LMF method is performed to estimate the model's rate constants. For the reduced models, information criterion (empirical, Bayesian, or Akaike) is calculated balancing the 5-fold cross-validation mean square error and the excessive number of reactions. It is used to choose the final optimal reduced model.

3.2.2 Model selection and parameter fitting

3.2.2.1 *The scheme*

MIKANA Ver. 2 applies the non-negative Lasso for model reduction. It uses L_1-norm penalized regression, and directly assigns less contributing coefficients to 0 hence gives interpretable reduced models. The non-negative Lasso solves a convex optimization problem thus any local optimum is also a global optimum. I apply the Levenberg-Marquardt-Fletcher (LMF) algorithm [191,192], a nonlinear least square method to fit the rate constants of the reduced models. The scheme of model selection and parameter fitting in MIKANA Ver. 2 is shown in Figure 3.3.

3.2.2.2 *The non-negative Lasso for model reduction*

Suppose after data standardization, y_r ($r = 1, \dots, JS$) are the elements of Y, $x_{r,i}$ is the element at the r-th row and i-th column in X. Letting $\hat{k} = \begin{bmatrix} \hat{k}_1 \\ \vdots \\ \hat{k}_I \end{bmatrix}_{I\times 1}$, the non-negative Lasso estimate $\hat{k}$ is defined by

$$\hat{k} = \arg\min \left\{ \sum_{r=1}^{JS} \left(y_r - \sum_i k_i x_{r,i} \right)^2 \right\} \text{ subject to } \sum_i k_i \le t, \text{ and } k_i \ge 0 \text{ for all } i, \quad (eqn.\ 3.3)$$

where t is the tuning parameter. Consider the case when $I = 2$, the estimation illustration of non-negative Lasso is shown in Figure 3.4. The constraint region is the triangle. The elliptical contours center at the ordinary least square estimates. The non-negative solution is the first place that the contours touch the triangle. The solutions at the acute angles of the triangle will give a zero k_i, $i = 1\ or\ 2$. Reactions with zero rate constants will be discarded in the reduced models.

For the convenience of programming, the above expression $eqn.\ 3.3$ can be re-written in the form:

$$\hat{k} = \arg\min \left\{ \sum_{r=1}^{JS} \left(y_r - \sum_i k_i x_{r,i} \right)^2 + \lambda \sum_i k_i + Pen \cdot \sum_i Ind(k_i < 0) \right\}, \quad (eqn.\ 3.4)$$

where λ is the tuning parameter. Ind is the indicator function: if $k_i < 0$, $Ind(k_i < 0) = 1$; otherwise, $Ind(k_i < 0) = 0$. Pen is a predefined large constant, such as 10^5, penalizing negative k_i values. I adopt this form in MIKANA Ver. 2. The non-negative Lasso problem is a regularized minimization problem. I solve it using the Dual Augmented Lagrangian (DAL) algorithm [199] for computing efficiency.

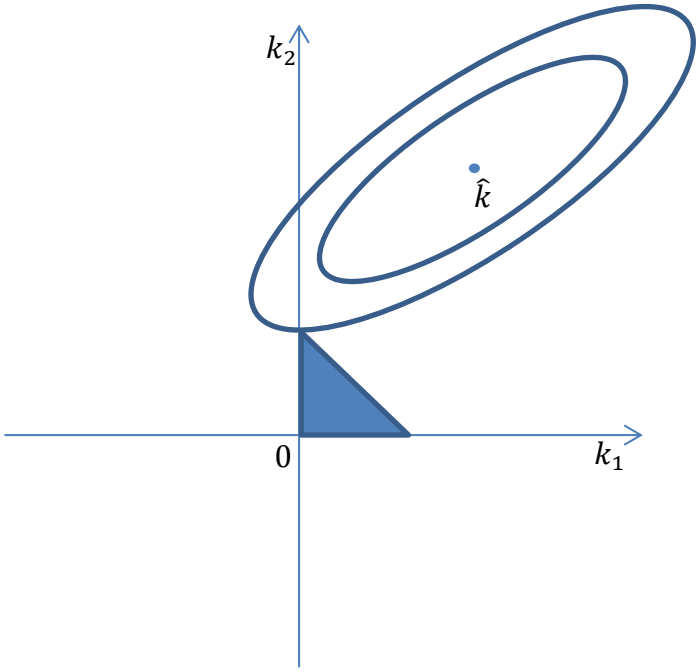


Figure 3.4 Estimation illustration of non-negative Lasso when $I = 2$.

3.2.2.3 *Parameter fitting*

The full pseudo-linear model $(eqn. 3.2)$ is always transformed from nonlinear systems, which contain differential equations higher than first order. Both linear and non-linear least square methods can be used to fit the parameters. The Levenberg-Marquardt-Fletcher (LMF) algorithm [191,192] is among the standard nonlinear least square routines in practice, which can fit both linear and non-linear systems. It has improved fitting than linear methods on oscillatory systems.

3.2.3 Noise tolerance

When the input data contains noise, the derivative of the time-course curve might become discontinuous; even deviate much from underlying true model. In order to avoid the influence of large noise, I provide an option to smooth the input data and rule out outliers using robust spline smoothing [190] before model selection and parameter fitting.

3.3 Results

Here I present five examples to evaluate algorithm performance, both the improvements and limitations. Example 1 is the simple oscillatory model I mentioned in the introduction. Example 2 - 3 are the same examples used for assessing MIKANA Ver. 1 [30]. Example 5 is a second-order model containing autocatalytic degradation. Example 6 is a third-order model.

MIKANA has two parameters for customizing the supported elementary reactions included in the initial set of reactions:

(1) Reaction mode:

Option 1 – Backbone mode: It only supports reactions in the form of $n^A A \rightarrow n^B B$, in which *A* and *B* are species, and n^A and n^B are the number of molecules.

Option 2 – General mode: No restriction.

(2) Molecularity per species:

Option 1 – Unimolecular: The maximum number of molecules per species is 1; and it supports up to second-order reactions.

Option 2 – Bimolecular: The maximum number of molecules per species is 2; and it only supports up to second-order reactions.

Option 3 – Termolecular: The maximum number of molecules per species is 3. It supports all the reactions in the above two options, and further supports third-order reactions.

In order to assess the performance of network structure prediction, we compare the predicted structure with the underlying true structure using *sensitivity* and *specificity* of edge predictions, when the network prediction is stably achieved (such as Example 1 - 3). These measures are defined as (# means "number of")

$$sensitivity = \frac{\#\ true\ edges\ in\ the\ predicted\ structure}{\#\ edges\ in\ the\ true\ structure},$$

$$specificity =$$

$$(\#total\ possible\ edges - \#edges\ in\ the\ true\ structure - \#edges\ in\ the\ predicted\ structure + \#true\ edges\ in\ the\ predicted\ structure)\ /\ (\#\ total\ possible\ edges\ - \#\ edges\ in\ the\ true\ structure),$$

They supplement the comparison of elementary reactions and differential equations.

3.3.1 Example 1 – The simple oscillatory model

The model system is schemed as $Model\ 3.1$ (see Example 1 in **Introduction)**. Figure 3.2 shows the result generated by MIKANA Ver. 1 with General mode and Bimolecular per species. Empirical information criterion is used for model selection. Adopting the same settings, I performed MIKANA Ver. 2 on the same data, and the Input/Output plot is shown in Figure 3.5. It predicts exactly correct reactions, thus both the sensitivity and specificity of edge predictions are 100%. Additionally, the estimates of rate constants are $\hat{k}_1 = 0.3011$, $\hat{k}_2 = 1.3024$ and $\hat{k}_3 = 0.6418$, which are very close to the true parameter values.

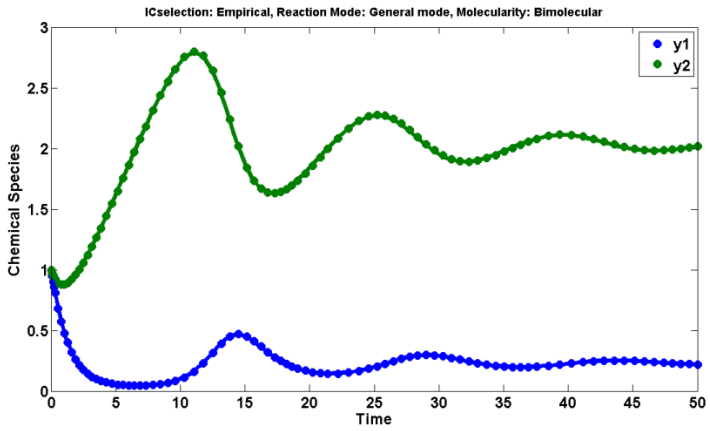


Figure 3.5 Input/Output plot for the simple oscillatory model, fitted by MIKANA Ver. 2.

Dots indicate input data, solid lines indicate fitted curves. Settings: General mode, Biomolecular per species. Empirical information criterion is used for model selection.

3.3.2 Example 2 – Michaelis-Menten mechanism

A typical Michaelis-Menten mechanism of enzyme action [200] is represented schematically as

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} C \overset{k_2}{\rightarrow} E + P,$$

in which E is the enzyme, S is the substrate, C is the enzyme-substrate complex, and P is the product. k_1, k_{-1} and k_2 are the rate constants. The example input data was provided in [30], with k_1=0.068 $mM^{-1}min^{-1}$, k_{-1}=0.0136 $mM^{-1}min^{-1}$ and k_2=0.0068 $mM^{-1}min^{-1}$ [201], and 10% noise. I performed network prediction using MIKANA Ver. 1 and Ver. 2 respectively, adopting settings suggested by [30]: General mode and Unimolecularity per species.

Figure 3.6 includes the results of MIKANA Ver. 1. The best prediction without prior information is obtained using Akaike information criterion. If we have prior information that the reactions in the schemes $\rightarrow X$ and $X \rightarrow$ do not exist in the mechanism, excluding them from the initial elementary set enables MIKANA Ver. 1 to correctly predict all the reactions and get good fitting, shown in Figure 3.6B.

Figure 3.7 contains the results of MIKANA Ver. 2. Whether incorporating the prior information or not does not affect the predictions here. Using either Akaike or Bayesian information criterion, MIKANA Ver. 2 predicts correct reactions and obtains good fitting. Both the sensitivity and specificity of edge predictions are 100%.

A

Without prior information;

Akaike information criterion

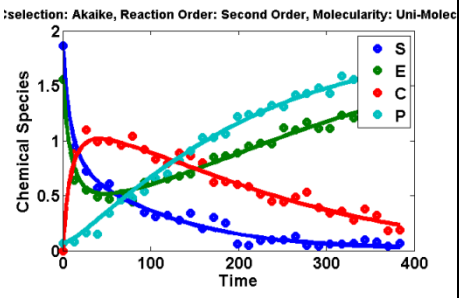
→ E

E →

C → S

<u>C → E + P</u>

<u>S + E → C</u>



k_1=0.0478

k_2=0.0066

B

With prior information:

excluding → X and X →;

Akaike/Bayesian information criterion

<u>C → S + E</u>

<u>C → E + P</u>

<u>S + E → C</u>

k_1=0.0478

k_{-1}=0.0082

k_2=0.0066

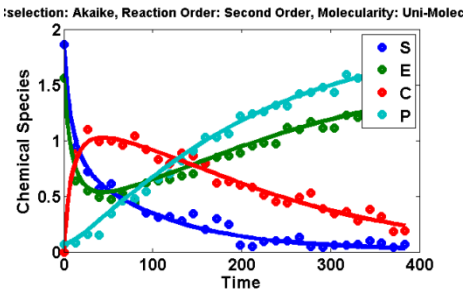


Figure 3.6 Results of Example 2 - Michaelis-Menten mechanism, obtained using MIKANA Ver. 1.

A – Best result without prior information. **B** – Best result with prior information (excluding → X and X →). Settings: General mode and Unimolecular per species. Predicted reactions and Input (dots)/Output (lines) plots are provided. Predicted reactions are listed by the side of plots, and the correctly predicted ones are underlined. For the correctly predicted reactions, the rate constants are provided, unit omitted.

<u>C → S + E</u>
<u>C → E + P</u>
<u>S + E → C</u>

k_1=0.0478
k_{-1}=0.0082
k_2=0.0066

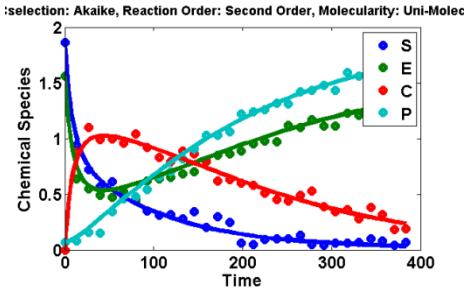


Figure 3.7 Results of Example 2 - Michaelis-Menten mechanism, obtained using MIKANA Ver. 2.

Settings: General mode and Unimolecular per species. Using either Akaike or Bayesian information criterion, with or without prior information, the reactions can all be correctly predicted and the results are the same. Predicted reactions, rate constants and Input (dots)/Output (lines) plots are provided.

3.3.3 Example 3 – The *Lactococcus lactis* glycolytic pathway

The *Lactococcus lactis* glycolytic pathway has been studied experimentally [178,202]. We used the data from nuclear magnetic resonance experiments [178], containing 7 species measured at 25 time points over 15.75 minutes. The 7 species are glucose (coded as X_1), glucose-6-phosphate (G6P) (X_2), fructose-1,6-biphosphate (FBP) (X_3), 3-phosphoglyceric acid (3-PGA) (X_4), phosphoenolpyruvate (PEP) (X_5), pyruvate (X_6) and lactate (X_7). A simplified topology of the glycolytic pathways is shown in Figure 3.8. Glucose (X_1) is converted into several intermediate products sequentially, and finally turns into lactate (X_7). PEP (X_5) also contributes to the conversion of glucose into G6P [63]. Pyruvate kinase (PK) catalyzes the transformation from PEP to pyruvate. Lactate dehydrogenase (LDH) performs the terminal conversion from pyruvate to lactate [203]. High concentration of FBP (X_3) activates PK and LDH, directing the pathway towards producing lactate [204]. Additionally, the free metabolite inorganic phosphate (Pi) is an inhibitor of PK [178,202]. In the network containing 7 nodes, consider directed edges and edges from/to environment, there are $2C_7^2 + 7 = 49$ possible edges. The underlying true structure shown in Figure 3.8 contains 12 edges.

The same dataset is analyzed using MIKANA Ver. 1 and Ver. 2, considering Backbone mode and Bimolecular per species to catch the backbone of the interactions and regulations; empirical information criterion is used. The results are shown in Figure 3.9 and Figure 3.10. Overall the predicted topologies are similar. MIKANA Ver. 1 predicted 15 interactions (arrows in the topology figure), and 6 of them are consistent with known topology shown in Figure 3.8.

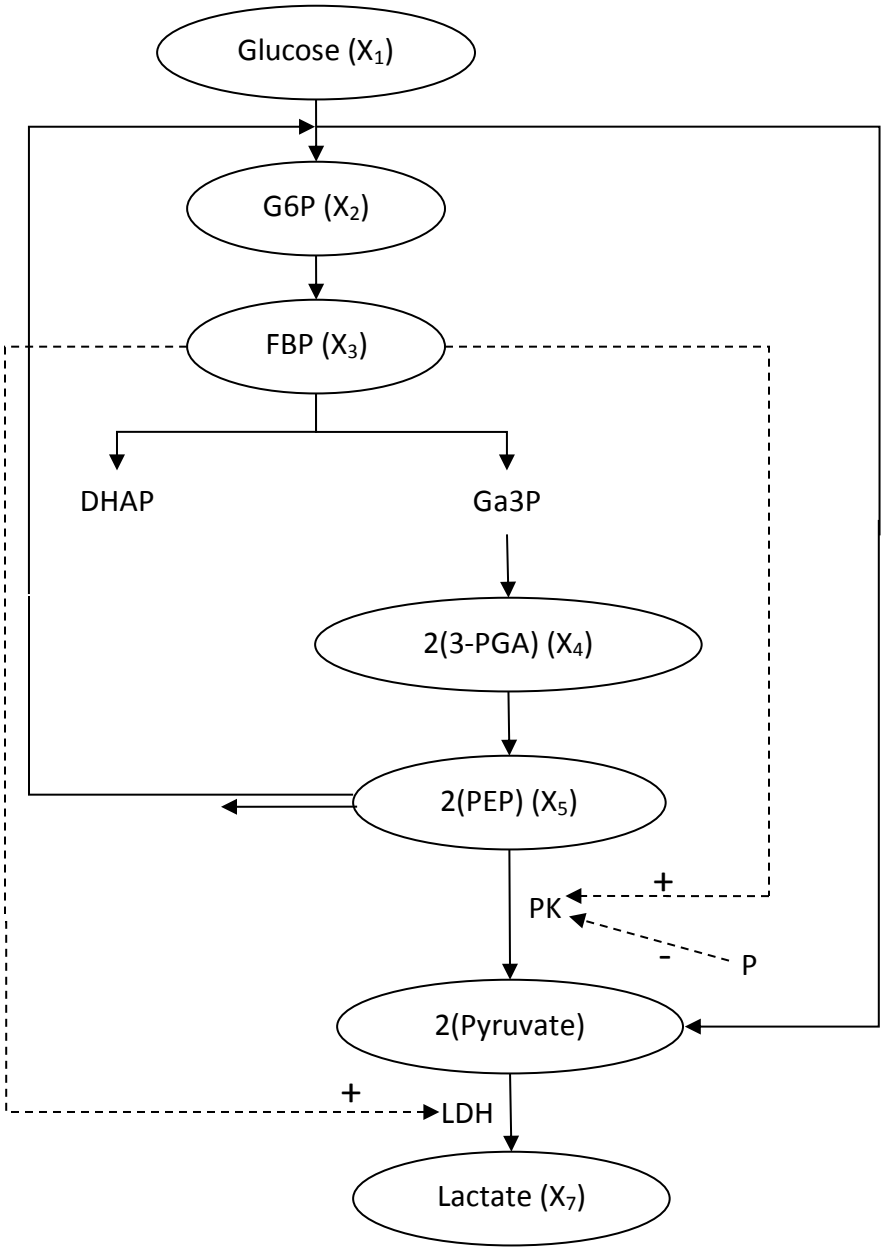


Figure 3.8 A simplified topology of the *Lactococcus lactis* glycolytic pathway.

The information comes from *Srividhya et al.* [63], Neves *at al.* [178], Hoefnagel *et al.* [202] and Ramos et al. [204] All the species coded from $X_1 – X_7$ in the ellipses have time-course measurements from Neves *at al.* [178] Other species, including dihydroxy acetone phosphoate (DHAP), glyceraldehyde-3-phosphate (Ga3P), pyruvate kinase (PK), inorganic phosphate (Pi) and lactate dehydrogenase (LDH) were not measured in the experiment. Solid lines indicate reactions, while dotted lines indicate activation (+) or inhibition (-).

For the result of MIKANA Ver. 1,

$$sensitivity = \frac{6}{12} = 50\%,$$

$$specificity = \frac{49-12-15+6}{49-12} \approx 76\%.$$

MIKANA Ver. 2 predicted 17 interactions, and 8 of them are consistent with known topology. For the result of MIKANA Ver. 2,

$$sensitivity = \frac{8}{12} \approx 67\%,$$

$$specificity = \frac{49-12-17+8}{49-12} \approx 76\%.$$

Therefore, for this example, MIKANA Ver. 2 improves the sensitivity and maintains specificity at the same value.

MIKANA Ver. 1 correctly captures the conversions of G6P to FBP (*i.e.* $X_2 \rightarrow X_3$) and 3-PGA to PEP ($X_3 \rightarrow X_4$), as well as the depletion of PEP (X_5). It predicted that G6P is produced by glucose ($X_1 \rightarrow X_2$), but missed the contribution of PEP (X_5). Pyruvate (X_6) is another product from this step, and it was predicted as $X_1 \rightarrow X_6$. The regulation of FBP to LDH is captured as $X_3 \rightarrow X_7$. MIKANA Ver. 2 correctly predicted the above information as well. It further identifies the contribution of PEP to G6P productions ($X_5 \rightarrow X_2$); and the conversion of PEP to pyruvate ($X_5 \rightarrow X_6$).

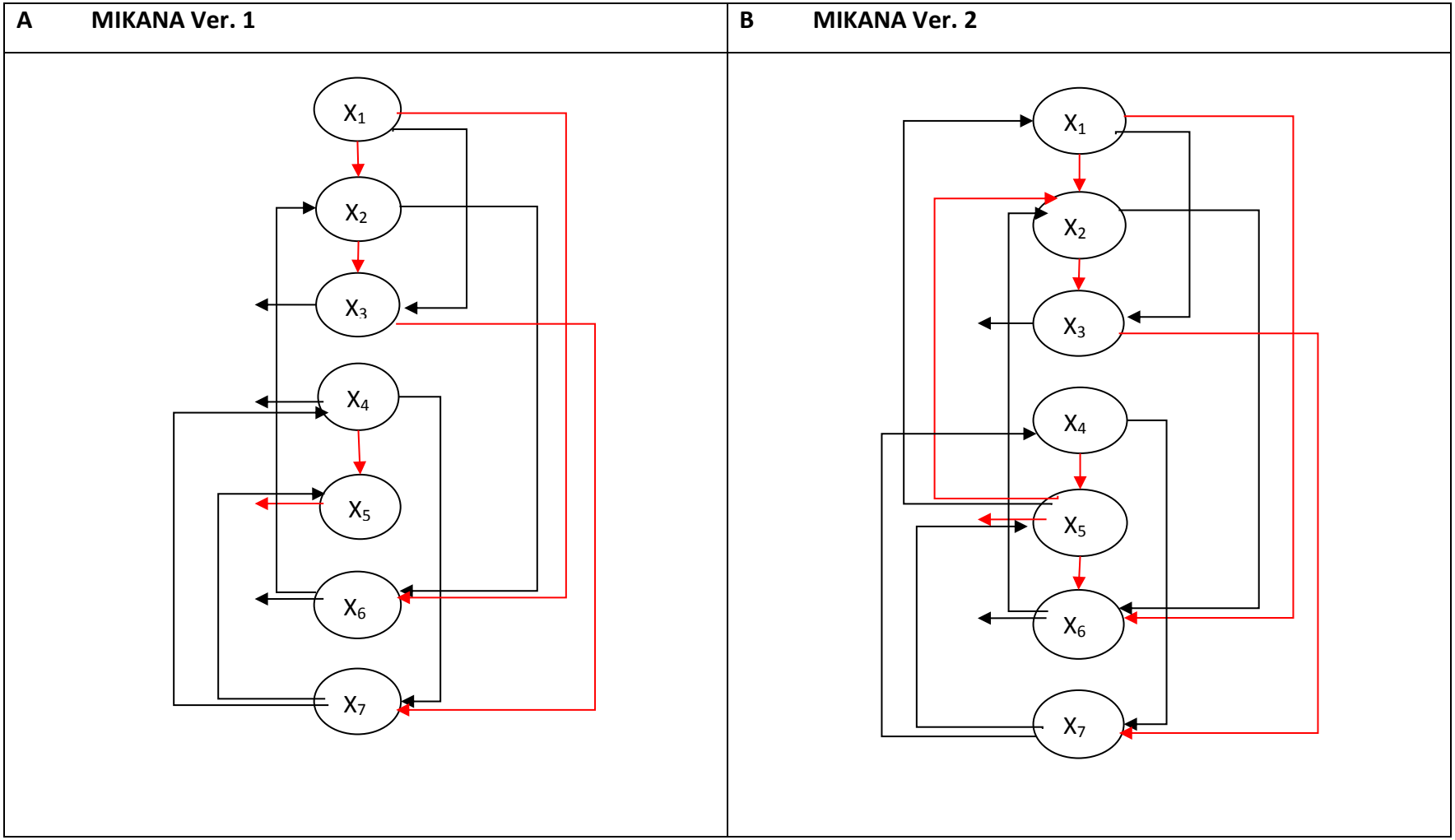


Figure 3.9 Predicted network topology for the *Lactococcus lactis glycolytic* pathway.

Settings: Backbone mode, Bimolecular per species; Information criterion: empirical. The number of molecules are disregarded. **A** – Topology predicted by MIKANA Ver. 1. **B** – Topology predicted by the MIKANA Ver. 2. Red lines are consistent with known topology shown in Figure 3.8.

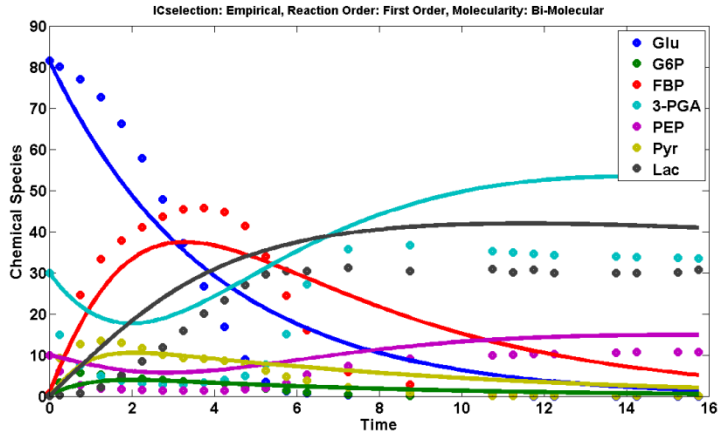
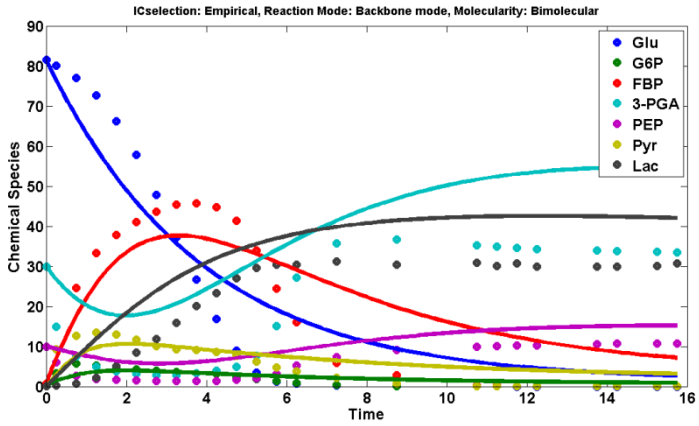
A MIKANA Ver. 1	B MIKANA Ver. 2
	

Figure 3.10 Input/Output plots for the *Lactococcus lactis* glycolytic pathway.

Settings: Backbone mode, Bimolecularity per species; Information criterion: empirical. Dots indicate input data, solid lines indicate fitted curves.
A – Output predicted by MIKANA Ver. 1. **B** – Output predicted by MIKANA Ver. 2.

3.3.4 Example 4 – Autocatalytic degradation

This second-order system containing autocatalytic degradation is schemed as below:

$$2y_1 \overset{k_1}{\rightarrow} y_1 + y_2,$$

$$2y_2 \overset{k_2}{\rightarrow} y_2 + y_3,$$

$$y_3 \overset{k_3}{\rightarrow} y_4,$$

$$y_1 + y_4 \underset{k_5}{\overset{k_4}{\rightleftarrows}} y_5,$$

$$y_5 \overset{k_6}{\rightarrow} y1,$$

$$y_2 \overset{k_7}{\rightarrow},$$

$$y_3 \overset{k_8}{\rightarrow},$$

$$y_4 \overset{k_9}{\rightarrow}.$$

I generated the input data with rate constants $k_1 = 2$, $k_1 = 1$, $k_3 = 0.1$, $k_4 = 100$, $k_5 = 1$, $k_6 = 0.1$, $k_7 = 0.17$, $k_8 = 0.17$ and $k_9 = 0.07$; initial values all 1. Besides, normally distributed random noise is added, whose standard deviation is 10% of the range of each species. The theoretical time-course curves (without noise) and noisy curves are shown in Figure 3.11A and B. I also apply the option of smoothing the noisy data (see ***Noise tolerance*** in **Materials and methods**). Smoothed curves are shown in Figure 3.11C.

The results of running MIKANA Ver. 1 and Ver. 2 are shown in Figure 3.12. Neither version can fully recover the reactions in the true model. They only detect no more than three reactions, but accompanied with more redundant reactions not in the model. The fitting of MIKANA Ver. 1 deviates. There tends to be over-fitting, although the curve seems fitted well in MIKANA Ver. 2. The noise worsens MIKANA's fitting performance.

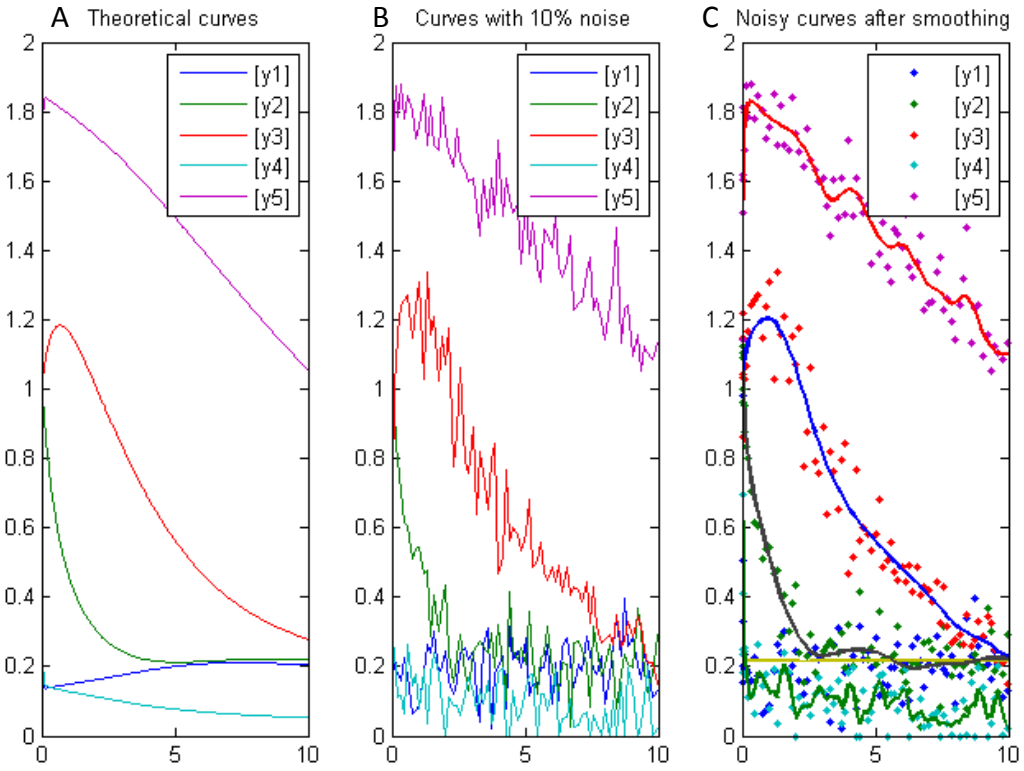


Figure 3.11 Time-course curves for Example 6.

A – Theoretical curves without noise. **B** – Curves with 10% noise. **C** –Noisy curves after smoothing. The dots are the noisy data with 10% noise. The lines are the smoothed curves.

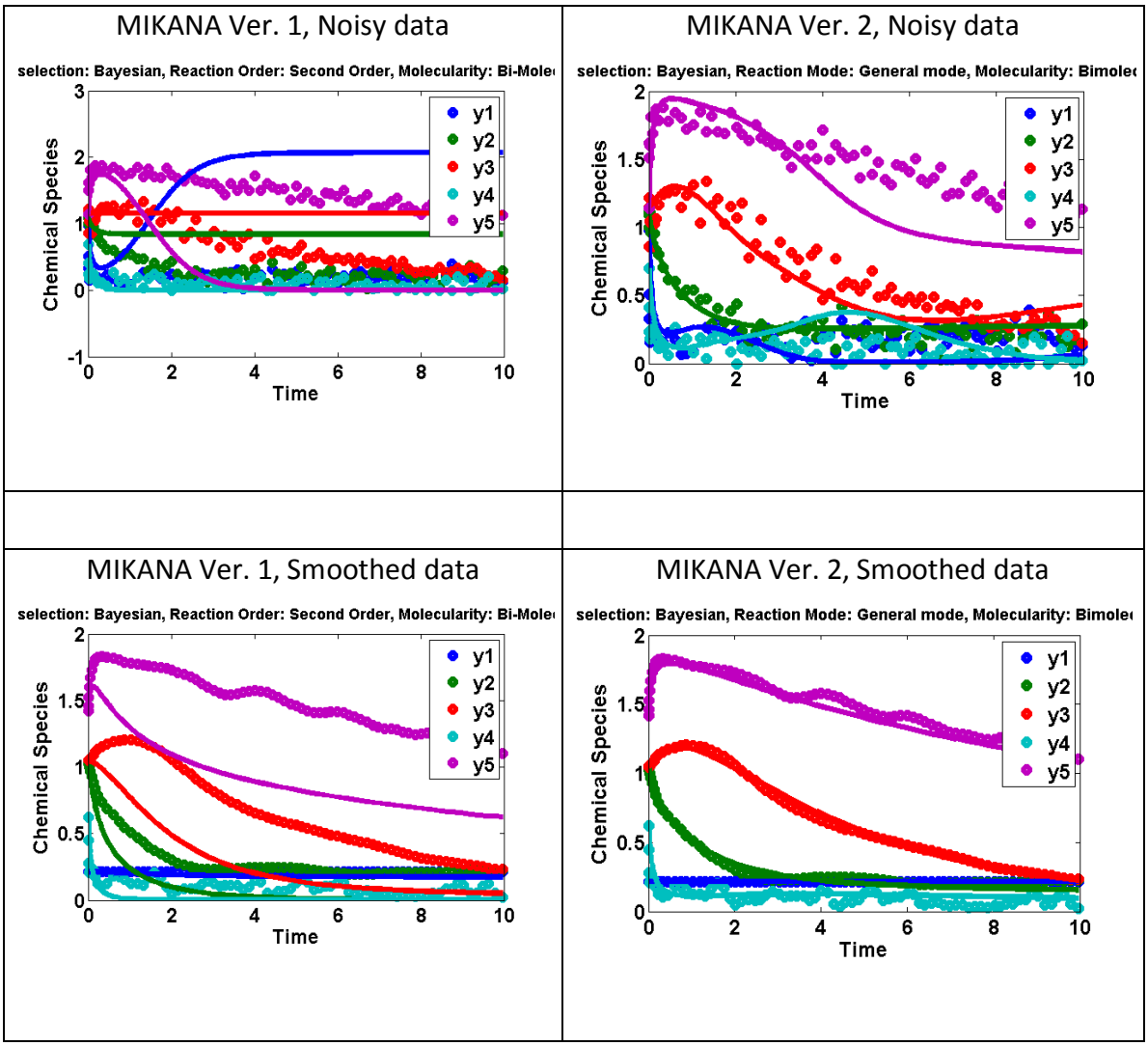


Figure 3.12 Influence of noise to MIKANA.

Settings: General mode and Bimolecular per species; Information criterion: Bayesian.

3.3.5 Example 5 – The standard mitogen model

Tyson and Light [205] presented a modification of the Brusselator model [193,194]:

$$A \xrightarrow{k_1} X,$$
$$B + X \xrightarrow{k_2} Y + D,$$
$$X + 2Y \xrightarrow{k_3} 3Y, \qquad (Model\ 3.2)$$
$$Y \xrightarrow{k_4} E.$$

It can display limit cycle oscillations depending on the destabilizing self-limiting autocatalytic third-order reaction, $X + 2Y \rightarrow 3Y$. When $k_1 = k_2 = k_3 = k_4 = 1$, $[A]$ is fixed and $[B] \rightarrow 0$, $[X]$ and $[Y]$ present limit cycles, which give interpretation of the standard mitogen model, helping understanding the cell division control [205,206]. I want a model without concentration constraints $[A]$ and $[B]$. Thus I re-write the model as $Model\ 3.3$:

$$\begin{aligned} &\overset{k_1}{\rightarrow} X, \\ X &\overset{k_2}{\rightarrow} Y + D, \\ X + 2Y &\overset{k_3}{\rightarrow} 3Y, \\ Y &\overset{k_4}{\rightarrow} E. \end{aligned} \qquad (Model\ 3.3)$$

Let $k_1 \equiv [A] = 0.5$, $k_2 \equiv [B] = 0.01$, and $k_3 = k_4 = 1$, then species X, Y, B and D in $Model\ 3.3$ will have the same time-course behavior as in $Model\ 3.2$, as shown in Figure 3.13., as. The input data of four species was simulated in time range [0, 60] with zero initial concentrations.

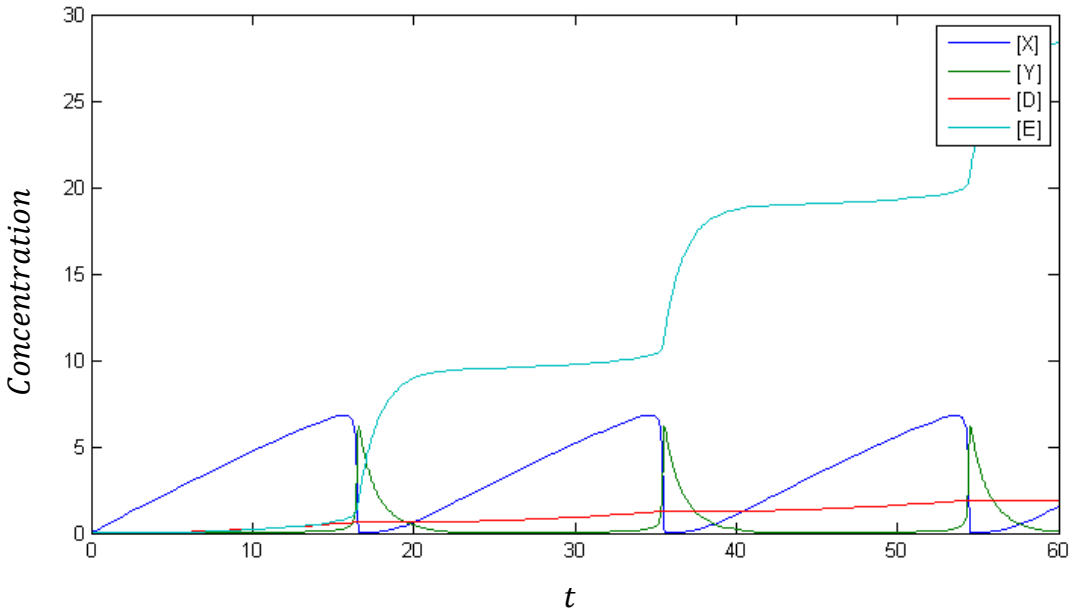


Figure 3.13 The theoretical time-course curves of $[X]$, $[Y]$, $[D]$ and $[E]$ in the standard mitogen model.

This model contains third-order reactions, which is among the extensions that can be generated in the initial set of reactions by MIKANA Ver. 2. I perform MIKANA Ver. 2 on the dataset simulated from $Model\ 3.3$. Settings: General mode and Termolecular per species, and empirical information criterion is used for model selection. I increase the parameter λ from 0.01 by step 0.01 until an empty model is reached. Then a really sparse reduced model is selected as the optimal. The predicted reactions are

$$\rightarrow 1\ X,$$
$$1\ Y \rightarrow 1\ E,$$
$$1\ Y \rightarrow 2\ E,$$
$$2\ Y \rightarrow 2\ E,$$
$$2\ Y \rightarrow 1\ Y\ +\ 1\ E,$$
$$1\ X\ +\ 1\ Y \rightarrow 2\ Y\ ,$$
$$2\ Y\ +\ 1\ X \rightarrow 3\ Y.$$

Among those, three reactions (highlighted in red) are in the true model schemed as $Model\ 3.3$. Predicted differential equations:

X' = + 0.5233 - 0.2980*X^1*Y^1 - 0.3672*Y^2*X^1

Y' = - 0.8121*Y^1 + 0.2980*X^1*Y^1 - 0.0529*Y^2 + 0.3672*Y^2*X^1

D' = 0

E' = + 0.7836*Y^1 + 0.0529*Y^2

The Input/Output plot and fitting of first and second derivatives are shown in Figure 3.14. Both the first and second derivatives are fitted well, which means the pseudo-linear model has been fitted well. MIKANA reports the errors of first derivatives are X: $7.888642e-02$, Y: $8.227607e-02$, D: $1.544635e-03$, and E: $1.635387e-02$. However, the time-course curves are still not fitted well. Take the curve of X for example, its concentration is frequently oscillating. Except those sharp turning points on the curve, the fitted curve has the same derivatives as the input.

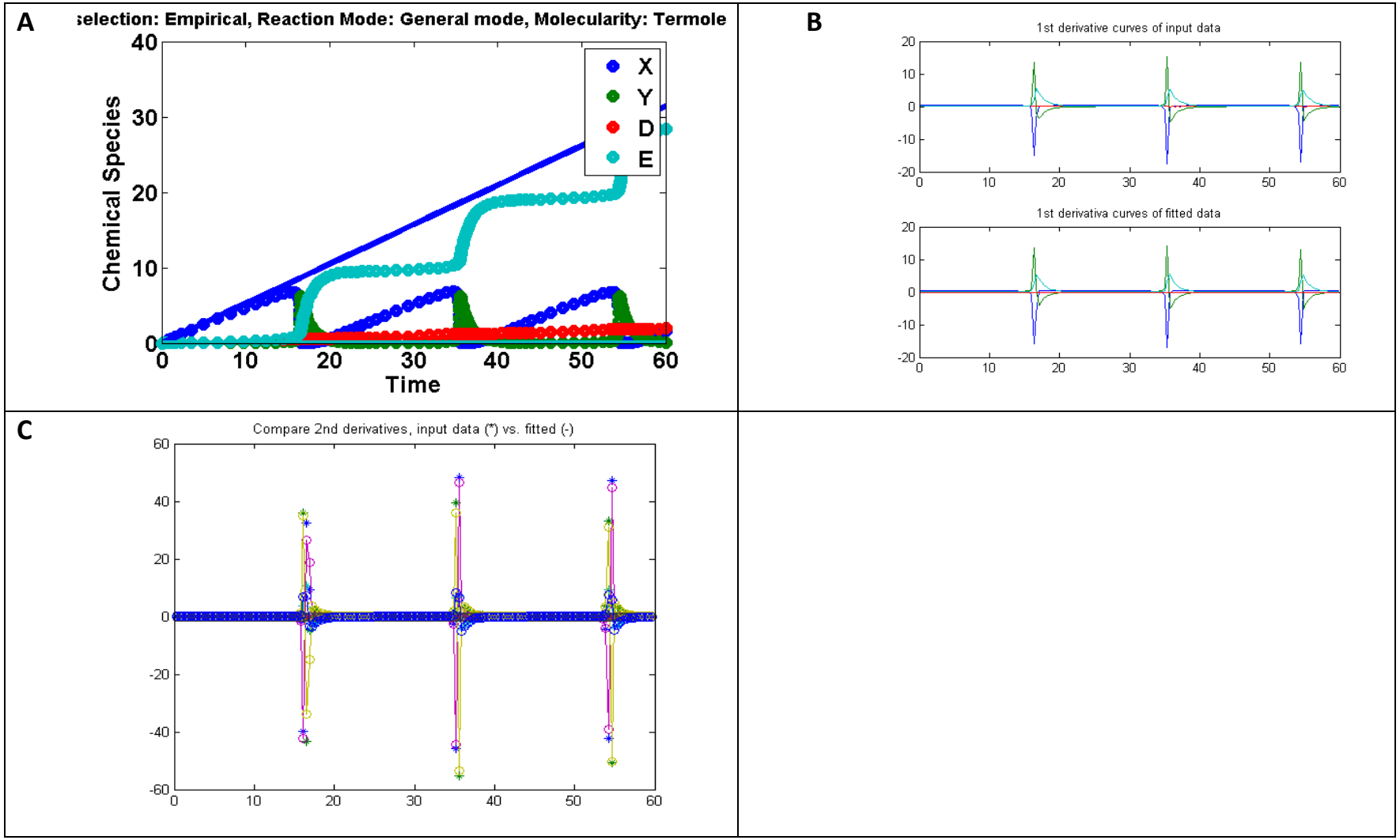


Figure 3.14 Result of Example 5 – the standard mitogen model, obtained by MIKANA Ver. 2.

A – Input(dots)/Output(lines) plot. **B** – First derivatives of input data (up) and fitted data (down). **C** – Second derivatives of input data (*) and fitted data (◯–).

3.4 Discussion

In this chapter, I present MIKANA Ver. 2, an improved version of MIKANA. It is a general method for inferring kinetic reactions and fitting parameters. The design matrix generator is improved to support autocatalytic reactions and third-order elementary reactions. It applies non-negative Lasso for model reduction and non-linear LMF algorithm for parameter fitting. Non-negative Lasso is optimized towards global optima, and its computing time does not grow exponentially with the number of species. Non-linear LMF algorithm outperforms linear least square regression, especially for oscillatory systems. LMF is more suitable for our overdetermined systems, where there are more equations than unknown parameters. Besides, the smoothing method is provided as an option to preprocess input data so that to avoid the influence of large noise. MIKANA Ver. 2 has improved performance on simple models, but the reconstruction of complex non-linear models is still not successful. MIKANA Ver. 2 can often generate a model which fits the data, but the over-fitting issue is haunting. Limitations of MIKANA are discussed below. Possible solutions for these issues are also posed for future studies.

Challenges of fitting higher-order oscillatory systems

Predicting the topology of higher-order oscillatory systems is challenging. Example 5 is an example. Even though both first derivatives and second derivatives have been fitted well, the time-course curves do not fit well. The pseudo-linear model is fitted to obtain the precise prediction of first derivatives. When the oscillatory curve and a straight line have the same derivatives except several sparse turning points, they are almost indistinguishable to the fitting method. More sampling points around the turning points might increase their influence on parameter fitting, thus help distinguishing the two curves. Alternative methods other than pseudo-linear model fitting, such as genetic algorithm, might be used to grow the possible set of reactions from single reactions.

Challenges of fitting complex systems

When the number of species and the order of reactions raises, the number of possible elementary reactions increases dramatically. Example 5 contains four species, and all possible elementary reactions up to third order are considered. A total of 314 reactions are included in the initial set of Example 5. MIKANA tends to predict a model containing more reactions than the true model, and the result might be over-fitted. Besides, different sets of reactions with specific range of parameters might degenerate to display very similar behavior in data collection time period. They are almost indistinguishable.

It is hard to predict a model exactly the same as the unique underlying true model for complex systems. Instead of providing only a single final predicted model, we might provide a set of highly possible models. With prior information, some models of the set will be selected as optimal.

Conflicting reactions have not been avoided in MIKANA predicted models. We might categorize reactions into conflicting groups. All the members in a group cannot appear in the same reaction system. At most one member from a group can be picked to build a model. Generating models without conflicting reactions will improve the model prediction.

Collinearity in design matrix

The pseudo-linear model $(eqn. 3.2)$ is prone to collinearity issues, where the explanatory variables (the columns in the design matrix) are not independent, and might have linear relationships. This collinearality might influence prediction precision and stability. Possible solutions include principal component regression, partial least square regression and so on [207], which worth investigation.

Prediction leverage of coefficients (rate constants)

We say a coefficient has higher prediction leverage if its change has more impact on the change of dependent variable(s). In model selection, our goal is to keep the coefficients that have large prediction leverage, and discard those having small prediction leverage. It is possible that some small rate constants tend to be

discarded from the model, but it does not means that they definitely have low prediction leverage. We standardize the data before model selection to reduce the influence of coefficient magnitude.

Chapter IV

Systematic Discovery of Signaling Networks Using Phosphorylation Dynamics Data

4.1 Introduction

This chapter is another attempt to discover network structures from time-course data, with applications in signal transductions. The novelty of the method presented in this chapter is the combination of Bayesian structure learning with time-delay detection.

Signal transduction is the process in which cells pass information between each other, and transduce extracellular signals to trigger intracellular responses. Defects and dysfunctions of signaling cascades can affect cell survival and cause diseases. Protein phosphorylation events are the most extensively studied of the signaling processes and in human cells, more than 500 different kinases are thought to be regulating signal transduction [208]. There are at least 107 genes in human genome code tyrosine phosphatases [209]. Phosphorylation and dephosphorylation of specific amino acid residues in target proteins plays an important role in these processes [210]. Proteins may be multiply phosphorylated within a small physical domain and may contain more than one phosphorylation domain, each having different or related functions. A single domain may be the target of several different kinases, and multiple phosphorylation events may be required to affect protein function.

The role of phosphorylation in cell physiology is discussed in more detail in Chapters 1 and 2 of my thesis, and here is a brief summary. Protein phosphorylation has been extensively studied throughout the past couple of decades. Initially, studies

were low-throughput, mainly focused on specific molecules in specific pathways. More recently, *in vitro* assays and protein chip arrays were developed which enables high-throughput system-wide studies; however, the kinase specificity *in vitro* and *in vivo* are often not the same [17]. The emergence of mass spectrometry (MS)-based approaches has enabled high-throughput large-scale studies of *in vivo* phosphoproteomics. Various MS-based methods, whether measuring dynamics or equilibrium-state, labeled or label-free, have been widely used for experimental research [7,11,17].

In this study, I focus on phosphorylation dynamics data, which are usually generated by labeled or label-free MS-based technologies, presenting abundance (or spectral count) changes of phosphopeptides over time due to certain stimuli. Towards the discovery of signaling networks, existing methods for MS time-course data include the Pearson correlation and clustering-based methods [48], the maximum entropy principle-based approach [211], NetworKIN which augments motif-based predictions with context for kinases and phosphoproteins [71] and so on. Here, I propose a Bayesian approach integrated with time delay detection to fully utilize the time delay between proteins in response to a stimulus to discover the causal relationship between proteins. The time-delay pattern might be easily missed by correlation coefficient-based cluster analysis. Phosphorylation dynamics also contain other information not included in this framework, such as peptide sequences, kinase/phosphatase motifs, localization, and even physicochemical characteristics. This method is not a perfect one taking all these features into account, but I focus on taking full advantage of the temporal information, and the approach can be a complementary analysis strategy in combination with other methods. Integrating multiple types of information is a challenge, and is worth investigation. Subsequent graph-theory-based analysis can be applied to the signaling network reconstructed with all the detected relationships to characterize network attributes.

In the **Results**, I evaluate the performance and stability of the approach on three simulated time-course datasets, including (1) a demonstration model containing two proteins with time-delay pattern, (2) the three-tiered cascade in MAPK (mitogen-activated protein kinase) pathway, and (3) EFGR (epidermal growth factor receptor)-mediated signal transduction. The latter two datasets are simulated

from ordinary differential equation (ODE) models which have been approved by previous experiments. In the demonstration model, I evaluate the results of my approach compared with correlation analysis. I sought to identify the critical number of time points, at and above which the approach can obtain higher confidence network structures. Although my approach is developed for phosphorylation dynamics data, it can be readily applied to other biological networks, which display time-delay patterns.

4.2 Materials and Methods

I propose an approach for detecting relationships between species, such as phosphorylated proteins, utilizing time delay patterns between the time-course profiles of different chemical species. The workflow is shown in Figure 4.1. I define the term *causal relationship* between species A and B, meaning there is an influence between the two species: A and B might be co-regulated, or A regulates B (A→B), or *vice versa* (B→A). The influence might be direct or indirect. I first use pairwise causal relationship discovery (based on causal Bayesian network structure learning) to identify the pairs of species having causal relationships. If the direction of a relationship is determined, either A→B or B→A, I further identify whether the relationship is up-regulation or down-regulation. If a causal relationship is detected, but the direction is not determined, I perform time delay detection to decide the direction. Subsequently, I determine whether the relationship is an up- or down-regulation. All the relationships can be connected to reconstruct the network structure. Since my method uses a Bayesian method (for causal relationship discovery) integrated with time delay detection (to help determine the direction of relationships), I call the method *BTD* (Bayesian + Time Delay detection).

4.2.1 Pairwise causal relationship discovery

The relative phosphorylation abundance is first discretized into three states – low, medium and high. In **Simulation 1 -** the demonatration model, I use the 2-fold change criterion [50] for simplification. In the other two simulations of signal transductions, I calculate the mean and standard deviation of relative phosphorylation abundance for each species. The log (logarithm) fold-change lower than the mean minus 1 standard deviation is set as low; the log fold-change higher

than the mean plus 1 standard deviation is set as high; otherwise it is set as medium. The causal relationship discovery is performed on the discretized data.

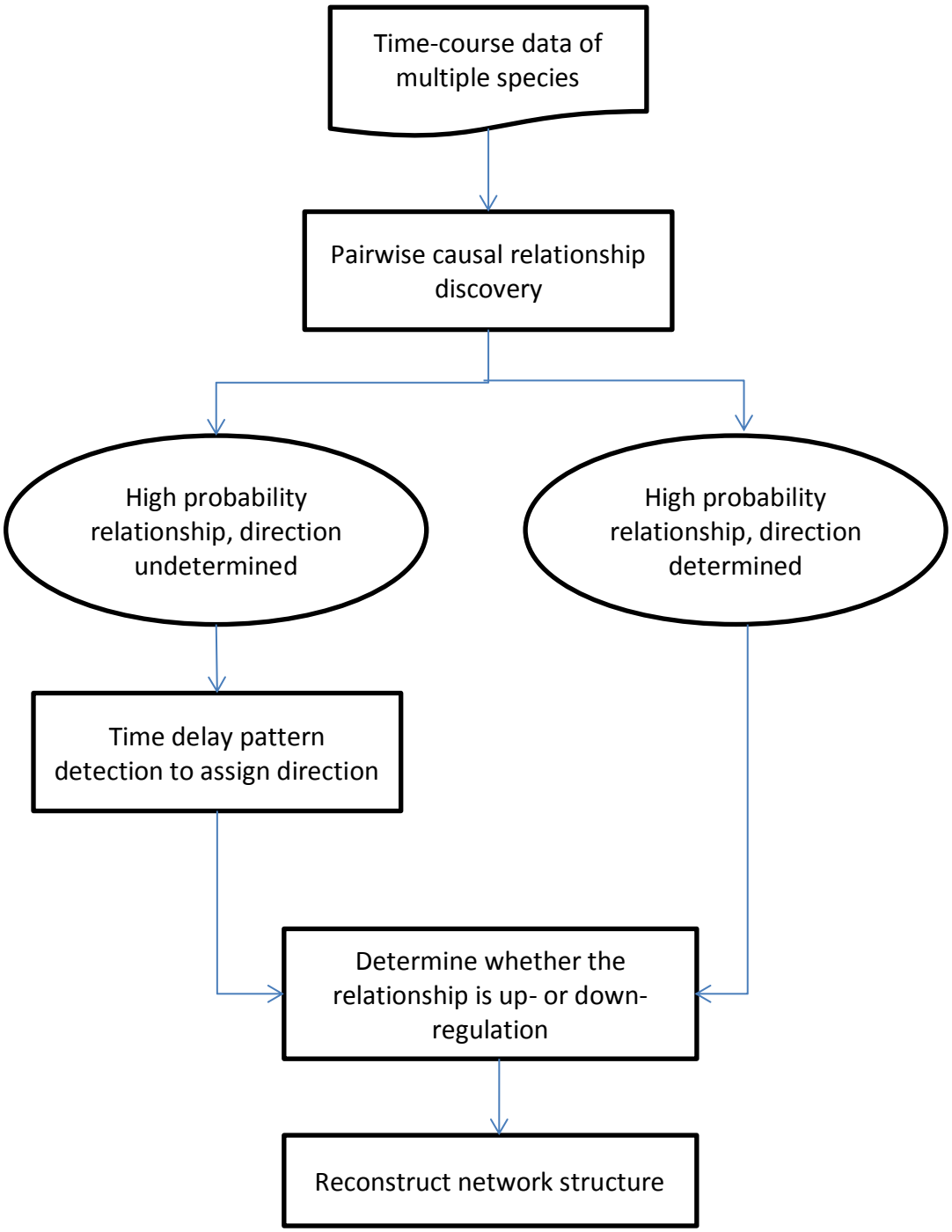


Figure 4.1 Workflow of BTD for reconstructing network structure from dynamics data.

The method of causal relationship discovery has been described in detail in Chapter 2. Instead of having different mutant conditions as in Chapter 2, here I have different time points. Phosphorylation abundance measurements at different time points for the same pair of species are used for causal Bayesian network structure learning. I only consider three possible structures of causal relationships between two species A and B: A→B, B→A or no causal relationship between A and B. The posterior probability of a structure is proportional to the Bayesian Dirichlet equivalent [105,107]. The structure with largest posterior probability is assigned as the relationship between species. In the case that A→B and B→A have the same probability and the sum of their probabilities is close to 1, I apply the time delay detection to help choose one from A→B and B→A.

4.2.2 Time delay detection

I detect the first extreme value that each time-course profile reaches in response to stimuli. The time-delay between extreme values is used to represent the time-delay between species. The species that reaches the extreme value earlier is inferred to be more upstream in a pathway.

Based on the second derivative at the extreme point, I can determine whether it is a maximum or minimum. If in a relationship A→B, both species reach a maximum (or both minimum), the relationship is inferred as up-regulation. If one reaches a maximum and the other reaches a minimum, the relationship is inferred as down-regulation. The regulation might be direct or indirect.

4.3 Results

To evaluate the performance of my approach, I simulate three datasets. The first dataset (**Simulation 1**) is generated from an artificial two-protein model, which is a toy model for demonstrating the key points of why I propose my approach. However, the case used is likely too simple to represent real biological dynamics. Thus, I move on to other representative models intended to more closely resemble real biological systems. I first analyze the three-tiered cascade in MAPK (mitogen-activated protein kinase) pathway (**Simulation 2**). Then I extend the analysis to a more complex model, epidermal growth factor receptor (EGFR)-mediated signal transduction, containing more protein complexes and intermediates (**Simulation 3**).

4.3.1 Simulation 1: A demonstration model of two proteins

Up-regulation case

Suppose I have two interacting proteins, A and B; and A up-regulates B. I assume protein A responses instantly after stimulation, which occurs at time $t = 0$. The time-course data of both proteins are collected simultaneously. The shape of their time courses $f_A(t)$ and $f_B(t)$ are simulated in proportion to the Gamma density functions, so that they roughly resemble some experimental curves of phosphorylation dynamics [7]. f_A follows a $Gamma(2,5)$ density function timed by 30, while f_B is just a shifted and scaled f_A:

$$\begin{cases} f_A(t) = 30 \cdot Gamma(2,5) = \frac{30}{25^2\Gamma(2)} te^{-\frac{t}{5}} \\ f_B(t) = M \cdot f_A(t - t_{delay}), \qquad M > 0, t_{delay} \geq 0 \end{cases} \qquad (eqn.\ 4.1)$$

in which, t is the time (min); f_A and f_B are functions of t; t_{delay} and M are pre-assigned constants which can be tuned. t_{delay} is the time delay of protein B after protein A. M is the scale of response magnitude of protein B compared with protein A. Figure 4.2A shows an example when $t_{delay} = 5$ and $M = 1$. For simplification, I only vary M in a mild range, so that the maxima of f_A and f_B are in (2, 5), when the 2-fold change criterion [50] is valid for identifying significant fold-changes.

Set $eqn.\ 4.1$ as the underlying true model, I compare BTD (*Bayesian + Time Delay detection*) and the Pearson correlation on the performance of identifying the truth. Let $M = 1$, I change t_{delay} from integers 0 to 19. In each of the 20 time delay conditions, I collect data from $t = 0$ to 40 by step size 0.1, thus the number of sample points is 401 per condition. The comparison result is shown in Figure 4.2B. The BTD is always able to detect the true relationship, by assigning it a probability equal to 1. By contrast, Pearson correlation can detect the true relationship only when the time delay is small. When $t_{delay} > 3$, Pearson correlation coefficient falls below 0.76. It even assigns a strong negative correlation when the time delay approaches 19, opposing the true relationship that protein A up-regulates protein B.

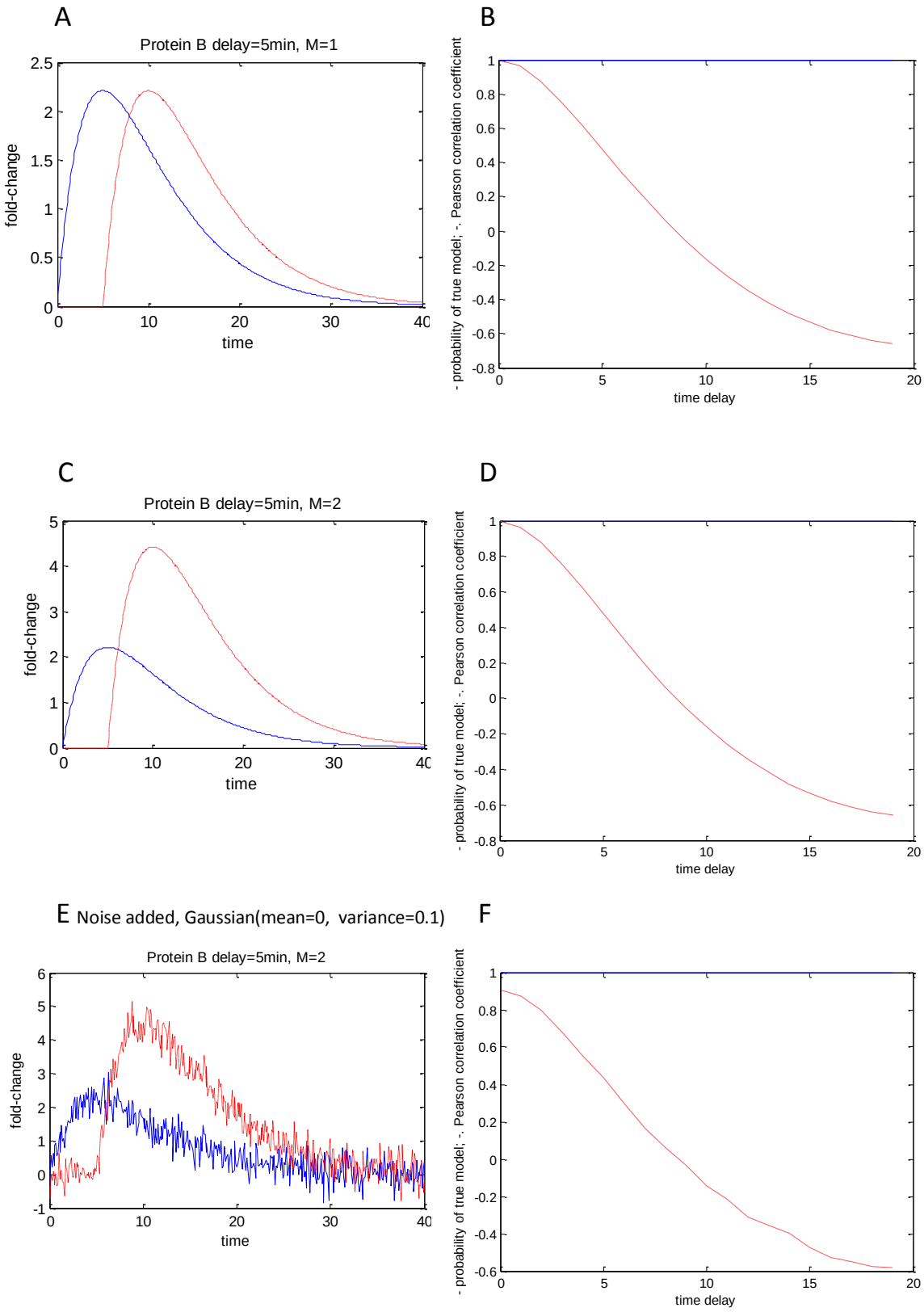


Figure 4.2 Comparison of BTD (Bayesian + Time Delay detection) with Pearson correlation, detecting up-regulation with time delay.

A - An example of the assumed underlying true models, with $\boldsymbol{M = 1}$, $\boldsymbol{t_{delay} = 5}$ (min). The blue line is the curve of protein A, while the red dash line is the curve of protein B. **B** - Set $\boldsymbol{M = 1}$, and $\boldsymbol{t_{delay}}$ increases from integers 0 to 19 (min). The number of sample points is 401, from 0 to 40 (min) by step size 0.1 (min). The blue line is the probability assigned by BTD to the true model. It remains 1, invariant with $\boldsymbol{t_{delay}}$. The red dash line is the Pearson correlation coefficient of the two 401-element vectors of proteins A and B, in the time range [0, 40]. The Pearson correlation coefficient decreases while $\boldsymbol{t_{delay}}$ increases. **C** - An example of the assumed underlying true models, with $\boldsymbol{M = 2}$, $\boldsymbol{t_{delay} = 5}$ (min). The legends are the same as in plot A. **D** - Set $\boldsymbol{M = 2}$, and all the other settings are the same as in plot B. **E**, **F** - Add Gaussian white noise with mean 0 and variance 0.1. The comparison is carried out again.

I further perform comparisons while varying M in a range ensuring the maxima of f_A and f_B between 2 and 5, and keeping the other settings unchanged. The comparisons always favor BTD over applying Pearson correlation alone on detecting up-regulation relationships having time delay. The result of $M = 2$ is shown in Figure 4.2C and D. After adding independent Gaussian white noise with mean 0 and variance 0.1, the comparison is carried out again. The conclusion does not change (Figure 4.2E and F). (Negative fold-changes might be obtained after introducing random noise. For BTD, the negatives are considered as zeros. For Pearson correlation, f_A or f_B are shifted on the vertical axis to make them non-negative. It does not change the Pearson correlation coefficient between f_A and f_B.)

Down-regulation case

Suppose protein A down-regulates protein B, and their time- course curves follow the equations below:

$$\begin{cases} f_A(t) = 30 \cdot Gamma(2,5) = \frac{30}{25^2\Gamma(2)} te^{-\frac{t}{5}} \\ f_B(t) = A - M \cdot f_A\left(t - t_{delay}\right), \qquad A = 1.5, M > 0, t_{delay} \geq 0 \end{cases} \qquad (eqn.\ 4.2)$$

I also carry out multiple comparisons of BTD and Pearson correlation, by changing M and adding white noise into the model. A is a constant assigned to keep f_B in a reasonable magnitude range. One example is shown in Figure 4.3A and B, in which $M = 1.5$, and independent Gaussian white noise with mean 0 and variance 0.1 is added. Again, 401 time points are sampled from time 0 to 40 (min) by step size 0.1 (min). I obtain the same conclusion as in the up-regulation case that BTD performs

perfectly for detecting the true down-regulation relationships; it assigns a constant probability 1 to the truth. Pearson correlation only works when the time delay is small, where it assigns a strong negative correlation between the two proteins. When the time delay increases, it starts to assign a positive correlation coefficient which is against the underlying truth.

Too much noise will hide the information implicit in the data. When the standard deviation of the Gaussian white noise increases to 1 (comparing Figure 4.3C and A), both BTD and Pearson correlation do not work well. As shown in Figure 4.3D, among the 20 time-delay conditions ($t_{delay} = 0, 1, \dots, 19$), BTD detects the down-regulation relationship with probability higher than 0.8 in 7 conditions; while, Pearson correlation barely detects it in only 1 condition, *i.e.* $t_{delay} = 0$, where the Pearson correlation coefficient is smaller than -0.5.

Random case

To evaluate the false-discovery rate, I simulate two random series of data for proteins A and B. They are generated as random numbers from a $Gaussian(0, 1)$ distribution and the values are shifted along the vertical axis by 2, ensuring a reasonable range (Figure 4.4A). There is no assumed relationship between the two random series. The simulation is repeated 1000 times and in each repeat, 401 sample points are taken from 0 to 40 (min) by step size 0.1 (min).

BTD can assign a probability to the existence of a causal relationship, as indicated by the blue line in Figure 4.4B. Throughout all the 1000 repeats, the probability almost remains at a low value close to 0 except one outlier. It has a 95% quantile-based confidence interval (0, 0.0104), and specificity = 99.9%. Pearson correlation coefficient is also calculated, as indicated by the red dash line in Figure 4.4B, with a 95% quantile-based confidence interval (-0.0939, 0.0965), and specificity = 100%. The confidence intervals are constructed using Monte Carlo methods. The results obtained from both BTD and Pearson correlation are consistent. Based on the probabilities assigned by BTD, there is no significant causal relationship between proteins A and B. No strong correlation are detected between A and B based on Pearson correlation coefficient.

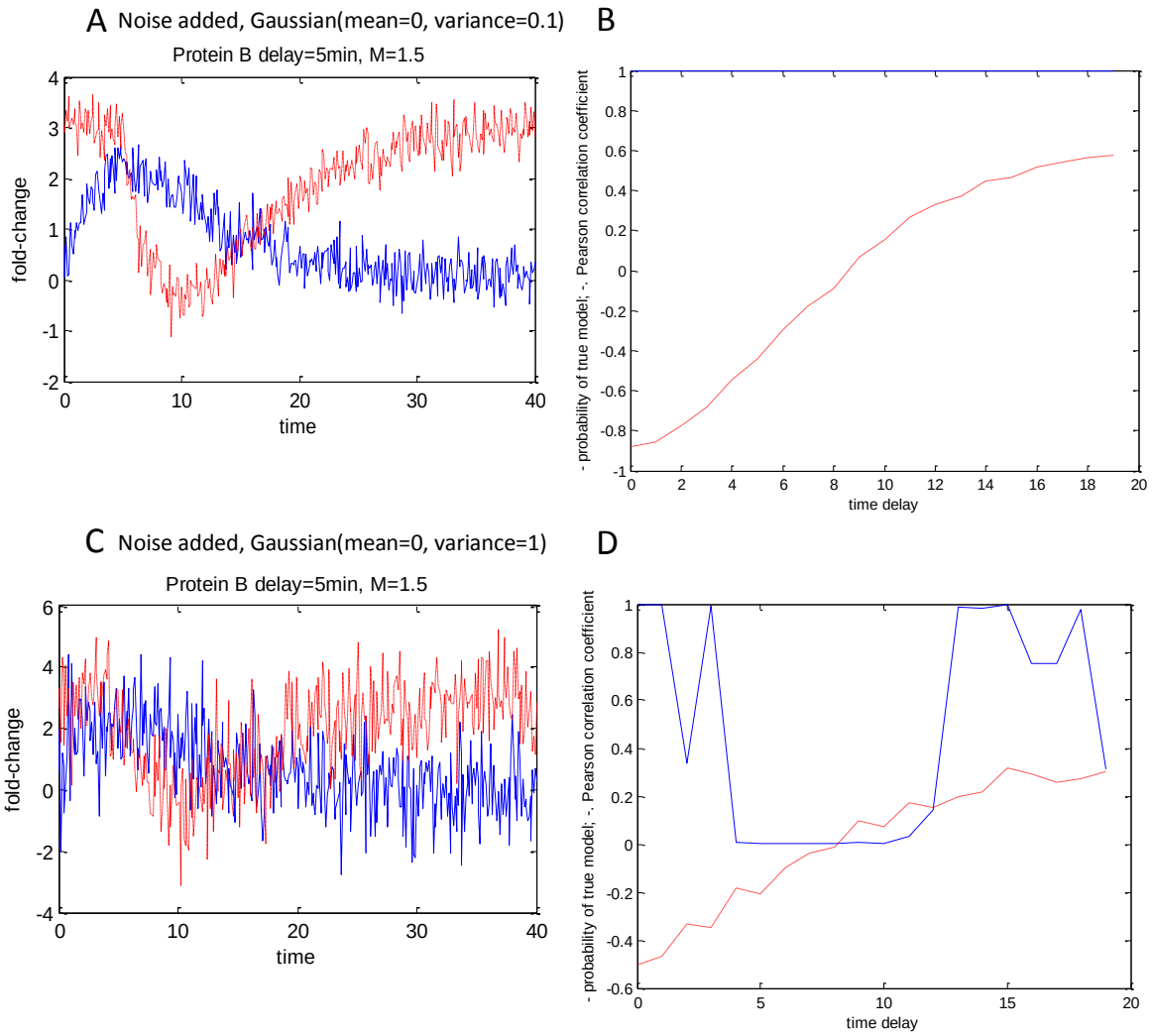


Figure 4.3 Comparison of BTD (Bayesian + Time Delay detection) with Pearson correlation, detecting down-regulation with time delay.

A - An example of the assumed underlying true models, with $\boldsymbol{M = 1.5}$, $\boldsymbol{t_{delay} = 5}$ (min). White Gaussian noise with mean 0 and variance 0.1 is added. The blue line is the curve of protein A; while the red dash line is the curve of srotein B. **B** - Set $\boldsymbol{M = 1.5}$, and $\boldsymbol{t_{delay}}$ increases from integers 0 to 19 (min). The number of sample points is 401, from 0 to 40 (min) by step size 0.1 (min). The blue line is the probability assigned by BTD to the true model. It remains 1, invariant with $\boldsymbol{t_{delay}}$. The red dash line is the Pearson correlation coefficient of the two 401-element vectors of proteins A and B, in the time range [0, 40]. The Pearson correlation coefficient increases while $\boldsymbol{t_{delay}}$ increases. **C**, **D** - Increase the variance of Gaussian white noise to 1. The comparison is carried out again.

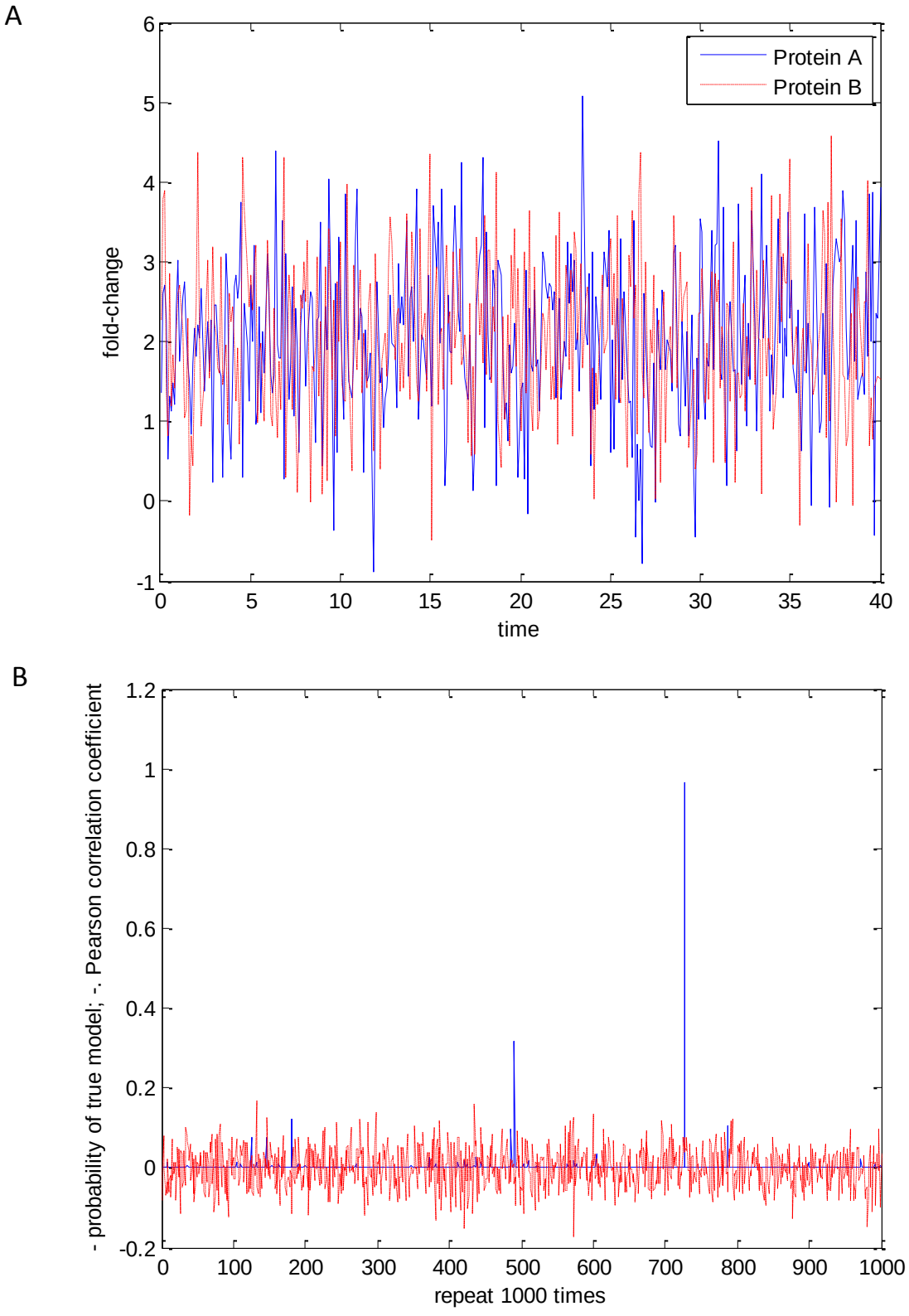


Figure 4.4 Comparison of BTD (Bayesian + Time Delay detection) with Pearson correlation, on random data, repeat 1000 times.

A - Two series of random data for proteins A and B are generated from $\boldsymbol{Gaussian(0,1)+2}$. The simulation is performed 1000 times. One repeat is shown. The blue line is the curve of protein A; while the red dash line is the curve of protein B. **B** - The number of sample points is 401, from 0 to 40 (min) by step size 0.1 (min). The blue line is the probability assigned by BTD to the existance of a causal relationship between proteins A and B. The red dash line is the Pearson correlation coefficient of the two 401-element vectors of proteins A and B. The horizontal axis is the index of repeats, from 1 to 1000.

Sensitivity and specificity

I evaluate the performance of both BTD and Pearson correlation depending on the time range and the number of data points. For calculating sensitivity and specificity, suppose the *positive* means protein A up-regulates protein B, and the data are generated from $eqn.4.1$, with $M=0$ and $t_{delay}=0,1,\dots,19$. The *negative* is defined as protein A and protein B having no relationship, and the data are generated as totally random numbers from $Gaussian(0,1)+2$. In real mass spectrometry experiments, the choice of time range for collecting data depends on whether the transient response or the long-term response is of interest. I take two time ranges, 0-20 (min) and 0-40 (min), and tune the sampling interval, *i.e*. the step size.

The results for detecting positives are shown in Table 4.1 and Figure 4.5A. The time range 0-40 (min) with step size 0.1 (min) can ensure the BTD will detect all the up-regulation relationships without misidentification. When the step size increases, the sensitivity of BTD decreases as expected. In the smaller time range 0-20 (min), BTD can obtain 30% ~ 95% sensitivity. Although Pearson correlation achieves a 100% sensitivity in 0~40(min) range when there are 3 time points, the 3-time-point setting is improper (see Figure 4.6). There is a high probability that I can obtain two random series with strong Pearson correlation. (In reality, consensus time-course curves averaged from replicate experiments might compensate for this issue.) Using the significance cut-off values adopted in Table 4.2, the probability is higher than 0.5. In summary, compared with Pearson correlation, BTD is still superior on sensitivity. More sampling points on longer time ranges do not ensure Pearson correlation to have a higher sensitivity; contrarily, the result might be even worse

with more misidentifications, such as assigning a strong negative correlation to an up-regulation model.

The results for detecting negatives are shown in Table 4.2 and Figure 4.5B. The data are generated totally randomly from $Gaussian(0,1) + 2$. For each setting, the simulation is repeated 1000 times. The specificity only relies on the number of time points. While the number of time points decreases, the specificity of both approaches decreases. As shown in Figure 4.5B, the Pearson correlation has a little higher specificity than BTD in the gross, except for the improper 3 time-point setting (Table 4.2). With 14 time points, BTD reaches specificity greater than 90%. Twenty time points is enough for both approaches to reach high specificity greater than 95%.

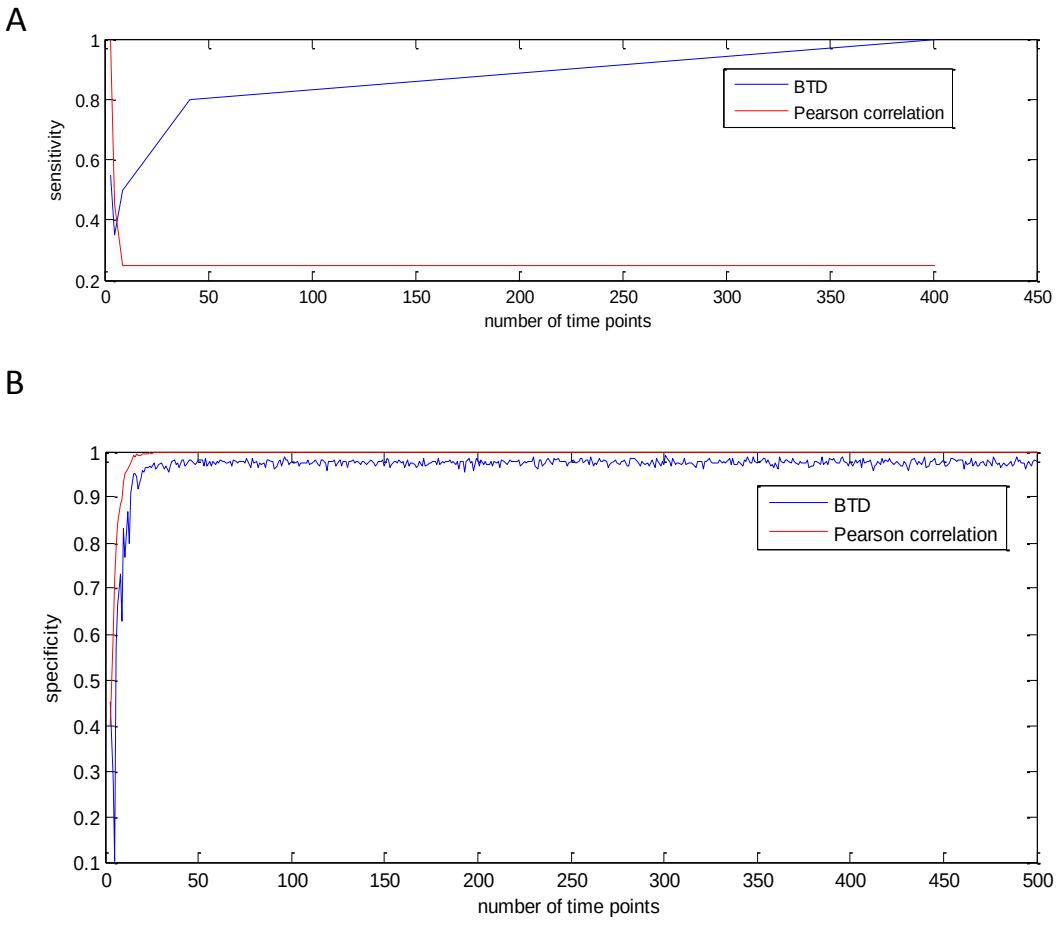


Figure 4.5 Sensitivity and specificity, relying on the number of time points.

A - Sensitivity. Data for plotting comes from Table 4.1. The time range is set as 0~40 (min). **B** - Specificity. The time-course data of the two proteins are all generated from $\boldsymbol{Gaussian(0,1)+2}$. The simulation is repeated 1000 times for each number of time points increasing from 3 to 500.

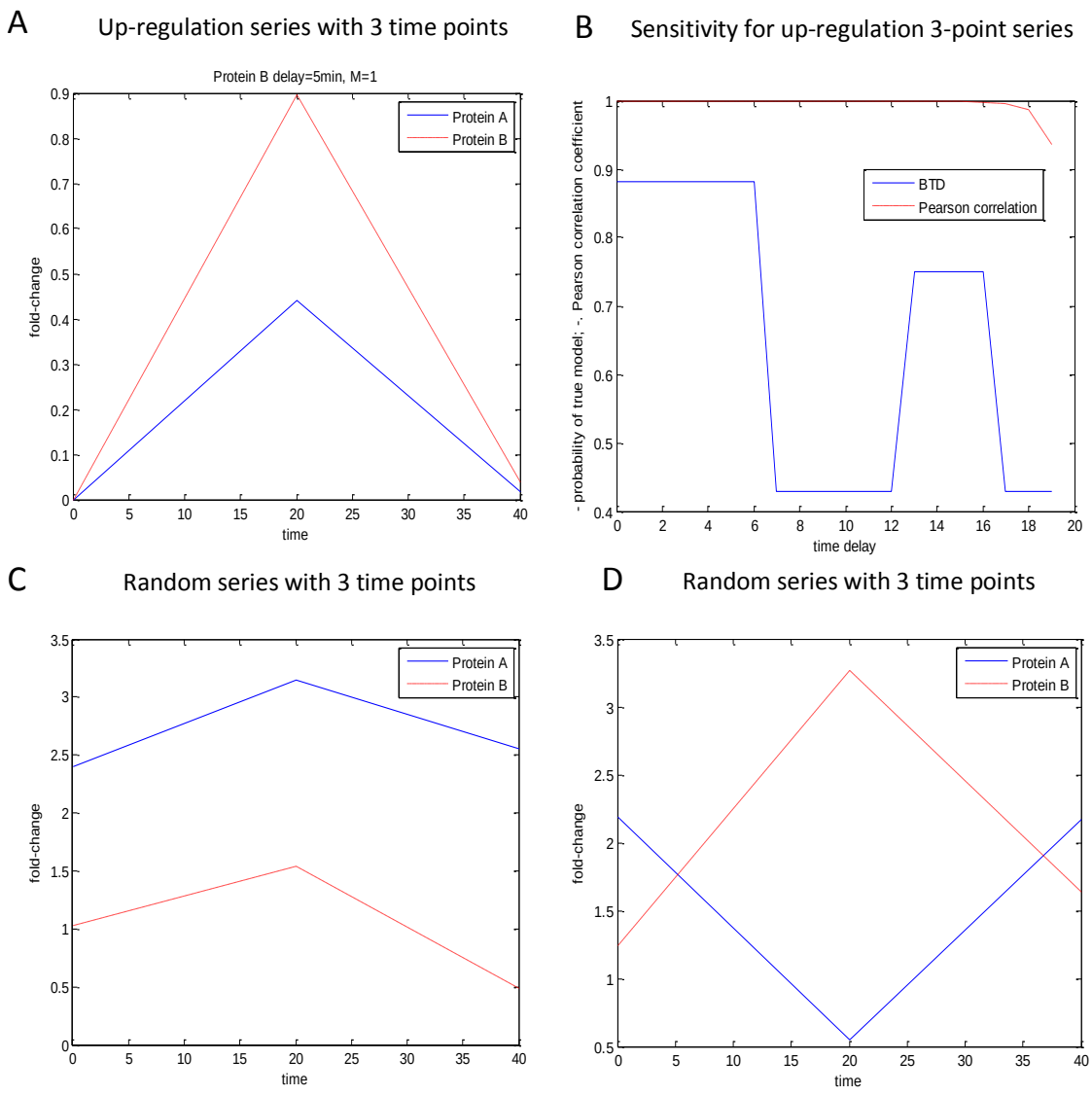


Figure 4.6 Three-time-point settings in the up-regulation case and the random case.

A - protein A up-regulates protein B. $\boldsymbol{M = 1, t_{delay} = 5}$, 3 time points. **B** - Sensitivity of BTD and Pearson correlation on detecting up-regulation as true positive. The data are sample at 3 time points: 0, 20 and 40 (min). **C**, **D** - Two random 3-time-point series are generated from $\boldsymbol{Gaussian(0,1) + 2}$.

Table 4.1 Test the performance of BTD and Pearson correlation on detecting the assumed up-regulation relationship[a].

Time range (min)	Step size (min)	Number of data points	Assigned by BTD (out of 20 time-delay conditions)			Assigned by Pearson correlation (out of 20 time-delay conditions)		
			Up-regulation[b] **(sensitivity)**	Down-regulation[c]	No significant relationship[d]	Up-regulation[e] **(sensitivity)**	Down-regulation[f]	Not correlated
0~40	0.1	401	20 (100%) [g]	0	0	5 (25%)	3 (15%)	12 (60%)
0~40	1	41	16 (80%)	0	4 (20%)	5 (25%)	2 (10%)	13 (65%)
0~40	5	9	10 (50%)	0	10 (50%)	5 (25%)	0	15 (75%)
0~40	10	5	7 (35%)	0	13 (65%)	9 (45%)	0	11 (55%)
0~40	20	3[h]	11(55%)	0	9 (45%)	20 (100%)	0	0
0~20	0.1	201	19 (95%)	0	1 (5%)	3 (15%)	8 (40%)	9 (45%)
0~20	1	21	6 (30%)	0	14 (70%)	3 (15%)	6 (30%)	11 (55%)
0~20	5	5	10 (50%)	0	10 (50%)	5 (25%)	0	15 (75%)
0~20	10	3	10 (50%)	0	10 (50%)	9 (45%)	0	11 (55%)

[a] The assumed truth is determined by equations (1), with $M = 1$. $\mathrm{t}_{\mathrm{delay}} = 0, 1, \ldots, 19$ corresponds to 20 time-delay conditions.

[b,c] The total probability of up- and down-regulation is greater than 0.6. The direction of either up- or down-regulation is determined by whether the first extrema of the pair are both local maxima (or minimA - or not.

[d] The total probability of up- and down-regulation is smaller than 0.6.

[e] It satisfies Pearson correlation coefficient > 0.6.

[f] It satisfies Pearson correlation coefficient < -0.6.

[g] The percentiles are provided in the brackets.

[h] Improper setting.

Table 4.2 Test the performance of BTD and Pearson correlation on random data[a].

Number of data points	Assigned by BTD				Assigned by Pearson correlation				
	95% CI[b] of (probability of causal relationship)		P_1=Percentage of (probability of causal relationship > 0.5)	**Specificity** *i.e.* 1-P_1	95% CI[c] of Pearson correlation coefficient		P_2=Percentage of (Pearson correlation coefficient > 0.6)	P_3=Percentage of (Pearson correlation coefficient < -0.6)	**Specificity** *i.e.* 1-P_2-P_3
3	0.42857	0.88235	0.548	0.452	-0.9976	0.99617	0.303	0.299	0.398
4	0.15789	0.94937	0.717	0.283	-0.95476	0.95587	0.202	0.201	0.597
5	0.29183	0.78097	0.897	0.103	-0.88313	0.88785	0.145	0.137	0.718
6	0.10297	0.95842	0.429	0.571	-0.83414	0.79731	0.103	0.109	0.788
7	0.22697	0.89155	0.339	0.661	-0.77946	0.75473	0.075	0.083	0.842
8	0.091668	0.91072	0.269	0.731	-0.70062	0.66896	0.056	0.058	0.886
9	0.07884	0.77034	0.371	0.629	-0.69614	0.67197	0.049	0.055	0.896
10	0.076611	0.88538	0.167	0.833	-0.61841	0.6282	0.036	0.033	0.931
11	0.067683	0.9048	0.231	0.769	-0.59369	0.59664	0.024	0.024	0.952
12	0.050019	0.88936	0.13	0.87	-0.59125	0.58081	0.018	0.021	0.961
13	0.061804	0.87924	0.203	0.797	-0.56221	0.57432	0.018	0.012	0.97
14	0.059366	0.87111	0.093	0.907	-0.54818	0.51903	0.011	0.015	0.974
15	0.055939	0.82548	0.049	0.951	-0.47838	0.47955	0.005	0.004	0.991
16	0.036966	0.67571	0.047	0.953	-0.49545	0.48621	0.006	0.007	0.987
17	0.025204	0.78212	0.054	0.946	-0.44601	0.4592	0.002	0.003	0.995
18	0.049685	0.88935	0.082	0.918	-0.45796	0.49352	0.002	0.005	0.993
19	0.020282	0.64439	0.057	0.943	-0.44223	0.4589	0.003	0.004	0.993
20	0.017038	0.6097	0.041	0.959	-0.46005	0.42189	0.003	0.002	0.995
21	0.017264	0.7115	0.044	0.956	-0.41698	0.43623	0.004	0.002	0.994
22	0.011694	0.65702	0.034	0.966	-0.43103	0.43122	0.001	0.004	0.995
23	0.010686	0.67288	0.035	0.965	-0.41797	0.41085	0.003	0.001	0.996
24	0.011273	0.58448	0.032	0.968	-0.39153	0.39027	0.001	0.001	0.998
25	0.011716	0.60351	0.035	0.965	-0.38695	0.37794	0	0.004	0.996

[a] The time-course data of the two proteins are all generated from $Gaussian(0,1)+2$. The simulation is repeated 1000 times for each setting.

[b,c] 95% quantile-based confidence interval. They are constructed using Monte Carlo methods. The number in the left column is the lower bound, while the number in the right column is the upper bound.

In conclusion, BTD beats Pearson correlation on the aspect of sensitivity. On the aspect of specificity, BTD does not perform as well as Pearson correlation; however, when the number of time points is relatively large, say over 20, their performance can be close.

Choice of time range and sampling interval

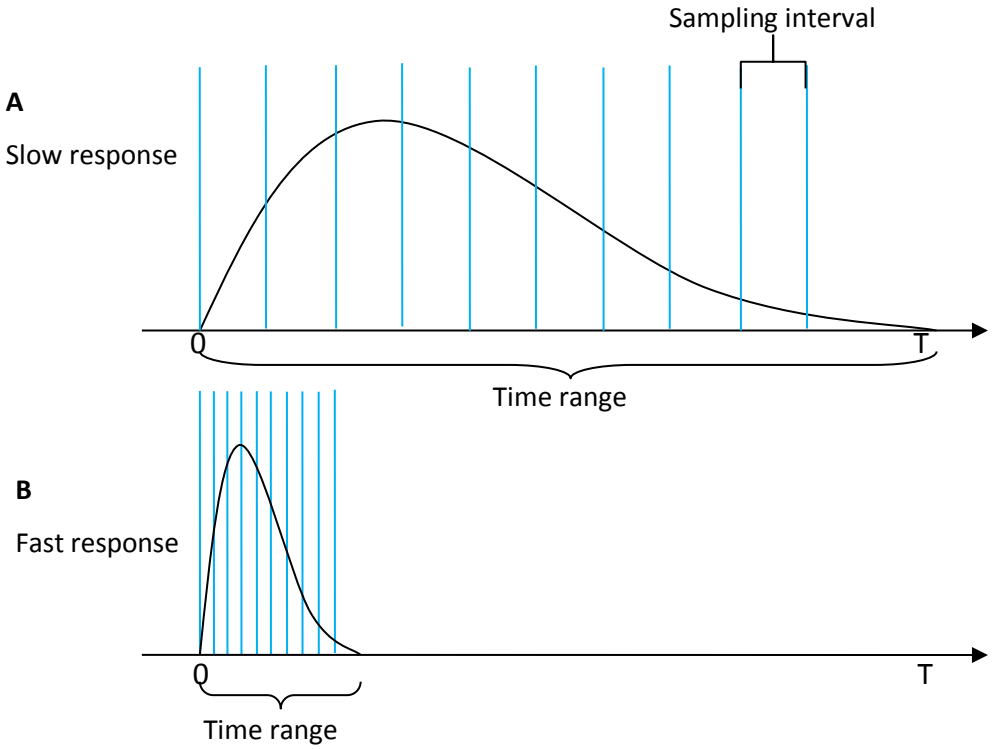


Figure 4.7 Choice of time range and sampling interval.

A – Slow response, **B** – Fast response. Both are on the same time scale from 0 to T. Blue lines indicate sampling time points. The proper time range for collecting data and sampling intervals are not the same for slow and fast reponses.

For signal transduction, early events may be quite fast (1-5 minutes), while later steps in the response may be slower (many minutes to hours). In biological experiments, the time ranges are always chosen to match the research focus (illustrated in Figure 4.7). There is no fixed setting for the choice of time range and sampling interval. For slow responses, the time range is chosen to be longer, and the sampling interval wider; while

for fast responses, the time range is chosen to be shorter, and the sampling interval narrower. The time range for collecting data should cover the response curve for the entire biological process under study and the sampling interval should support enough resolution to present the characteristics of the curves.

4.3.2 Simulation 2: The three-tiered cascade in MAPK pathway

The MAPK (mitogen-activated protein kinase) pathway is a commonly used paradigm for signal transduction, and it has been extensively studied. It is highly conserved from yeast to human and plays a central role in many crucial cellular processes, such as proliferation and differentiation [212]. A typical structure of the three-tiered cascade is described by Hornberg *et al.* [213]. It involves three kinases in succession (RAF-MEK-MAPK) forming interconnected cycles; in each cycle, a kinase phosphorylates a substrate, and a phosphatase dephosphorylates the substrate [213]. The cascade can be activated by various extracellular stimuli, including epidermal growth factor (EGF) via the epidermal growth factor receptor (EGFR). I use the model on JWS Online Cellular Systems Modeling Database [214] to simulate the concentrations of species in the cascade, including of RAF, MEK and MAPK and their phosphorylated forms, as well as the active and inactive EGF receptors. The simulated curves are shown in Figure 4.8A. Note that the phosphorylation dynamics obtained following the common experimental procedure [7] is not the concentration of species, but the concentration fold-change of the phosphorylated species after stimulation ($t > 0$) versus control ($t = 0$). Thus I calculate the fold-change, and the curves in 0-40 min having 11 time points and 41 time points respectively are shown in Figure 4.8B and C.

Before performing BTD, the fold-changes need to be discretized into three states – low, medium and high. Instead of using the simple 2-fold change criterion adopted in **Simulation 1**, the discretization is performed based on the sample mean and standard deviation of the log (logarithm) fold-changes (see **Materials and Methods**).

The 41-point case works perfectly to detect the underlying truth: RAF up-regulates MEK, and meanwhile MEK up-regulates MAPK. Both the sensitivity and specificity (as defined in Chapter 3) are 100%.

The 11-time-point case can tell that MAPK is the last protein in the cascade, and RAF and MEK have correlation; however, it cannot differentiate the role of RAF and MAPK, because this information is lost in the sparsely sampled data (see Figure 4.8B). Considering all the direct causal influences between proteins, the sensitivity is 50% and the specificity is 75%.

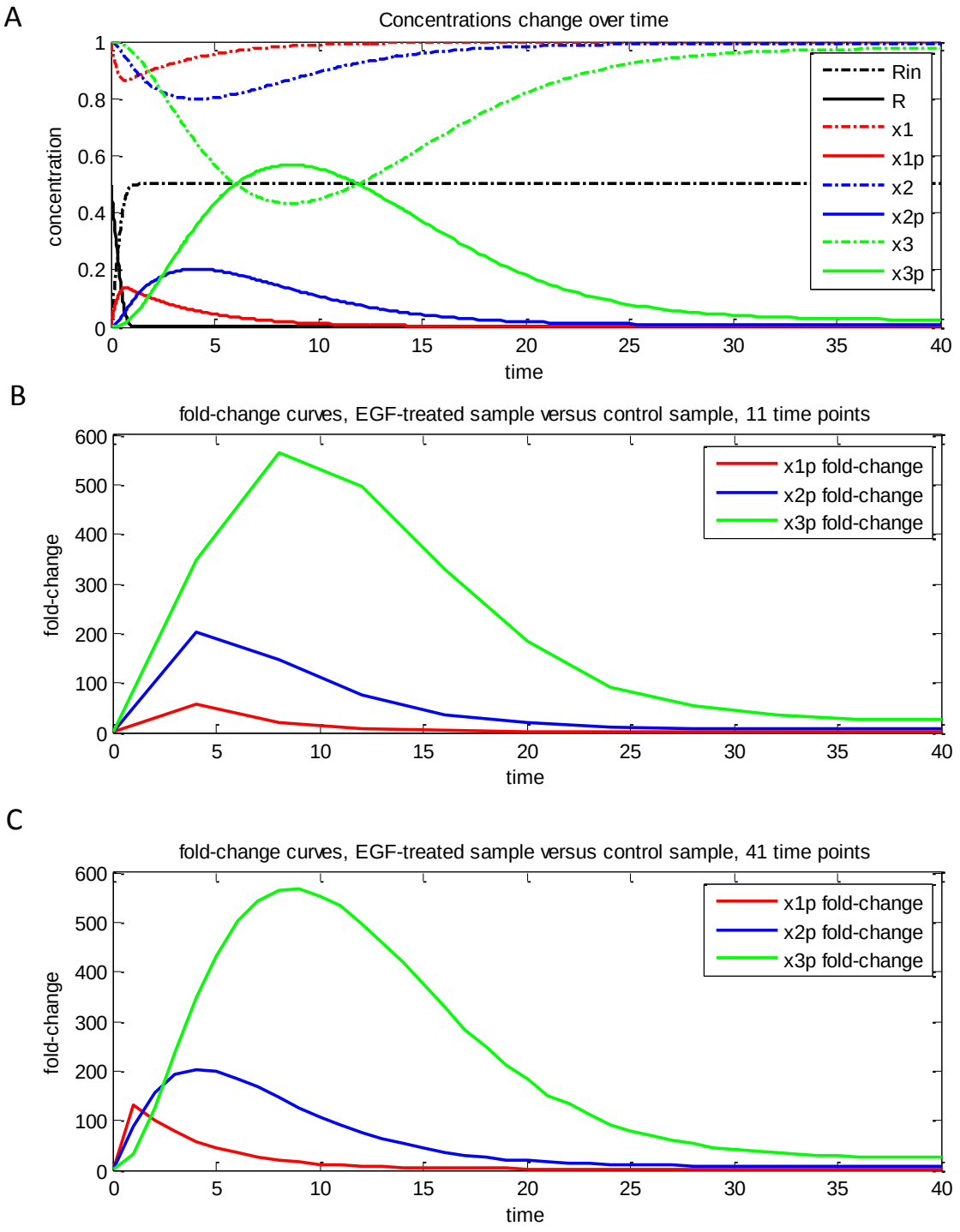


Figure 4.8 Time-course curves generated from the simplified MAPK pathway model.

RAF, MEK and MAPK are coded as x1, x2 and x3, respectively; while their corresponding activated forms are coded as x1p, x2p and x3p, respectively. The active and inactive EGF receptors are coded as R and Rin, respectively. **A** - Concentrations change over time after the EGF stimulation at time = 0. At time = 0, the sample is considered as untreated control. The raw data are generated on the platform of JWS Online

Cellular Systems Modelling Database [214], which contains the Hornberg mathematical model [213]. Simulation settings: Start time = 0, End time = 40; Steady-state analysis: steady state; Metabolic control analysis: Control coefficients. Time = 0 is when the EGF is added into the system, and R's concentration increases to 0.5 instantly. **B** - Fold-change curves, EGF-treated versus control. 11 time points taken at closest points to 0, 4, 8, 12, …,40 (units). **C** - EGF-treated-versus-control fold-change curves, 41 time points taken at closest points to 0, 1, 2, 3, …, 40 (units). The control is set when time = 0. In plot B and C, the fold-changes are calculated as EGF-treated-versus-control concentration ratios of x1p, x2p, and x3p at each selected time point.

4.3.3 Simulation 3: Epidermal growth factor receptor (EGFR)-mediated signal transduction

EGFR-mediated signal transduction plays a crucial role in controlling fundamental cellular functions. The pathway is often hyperactivated in cancer cells and is considered a promising target for cancer therapy [215–219]. Mathematical models have been developed to describe the mechanism in great detail [219–222]. I use the extensive Wolf model with feed-forward inhibition via transient activated RAS-GAP proteins [219] to generate *in silico* time course data of protein concentrations. This model consists of 61 reactions, which describe the complex processes including receptor activation, receptor-adaptor complex formation, the phospholipase C-γ (PLCγ) cycle, the RAS cycle, and the MAPK cascade. The concentrations of all the species in the model, including the complexes and intermediates, are calculated using JWS Online [214]. Parameters and initial values are the same as mentioned in [219], except $[R2P]_0$ = $[R\text{-}PLP]_0$ = $[R\text{-}ShP]_0$ = $[Rafa]_0$ = $[MEKP]_0$ = $[MEKPP]_0$ = $[ERKP]_0$ = $[ERKPP]_0$ = 0.01 to ensure non-zero concentrations of phosphorylated proteins as control (see the equations for calculating total phosphorylated proteins immediately below).

Only six proteins in the model have single and/or double phosphorylated forms, and their concentrations can be calculated as below. The species in the square brackets are the protein complexes and intermediates defined in Wolf model [219].

- Total phosphorylated EGFR = 2×([R2P] + [R-PL] + [R-PLP] + [R-GAP] + [R-G] + [R-G-S] + [R-Sh] + [R-ShP] + [R-ShP-G] + [R-ShP-G-S] + [R-Sh-G-S-Ras-GDP] + [R-Sh-G-S-Ras] + [R-Sh-G-S-Ras-GTP] + [R-G-S-Ras-GDP] + [R-G-S-Ras]

+ [R-G-S-Ras-GTP] + [R-GAP-Ras-GTP]);

- Total phosphorylated PLCγ = [R-PLP] + [PLCγP]+ [PLCgPI];
- Total phosphorylated SHC = [R-ShP] + [R-ShP-G] + [R-ShP-G-S] + [ShP] + [ShP-G]+ [ShP-G-S] + [R-Sh-G-S-Ras-GDP] + [R-Sh-G-S-Ras] + [R-Sh-G-S-Ras-GTP];
- Total activated RAF = [Rafa] + [Rafa-Pase1] + [MEK-Rafa] + [MEKP-Rafa];
- Total single phosphorylated MEK = [MEKP] + [MEKP-Rafa] + [MEKP-Pase2];
- Total double phosphorylated MEK = [MEKPP] + [ERK-MEKPP] + [ERKP-MEKPP] + [MEKPP-Pase2];
- Total single phosphorylated ERK = [ERKP] + [ERKP-MEKPP] + [ERKP-Pase3];
- Total double phosphorylated ERK = [ERKPP] + [ERKPP-Pase3].

Following the principle of the common experimental procedure [7] (also adopted in **Simulation 2**), I calculate the dynamics of the above phosphorylated forms. The simulated phosphorylation dynamics is shown in Figure 4.9.

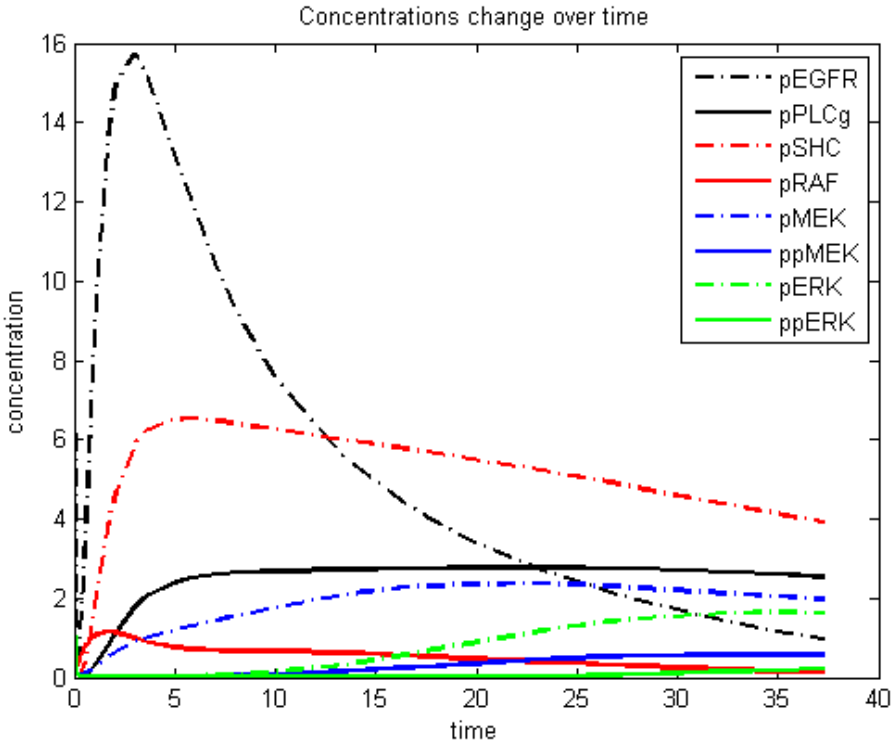


Figure 4.9 Phosphorylation dynamics simulated for EGFR-mediated signal transduction.

Twelve pairs of the phosphorylated species have been assigned the probability of causal influence > 0.999. No directional causal influence is determined by causal relationship discovery. Thus time delay detection is applied to these pairs to determine the possible direction of the causal influence and up-/down-regulation. The result is shown in Figure 4.10. Each species in the figure is the total concentration of a phosphorylated form, including those involved in complexes with other proteins. It is a very much simplified pathway without all those intermediate species. In the complex full pathway model [219], EGFR is activated by phosphorylation. Phosphorylated EGFR and SHC form a complex, which promotes phosphorylation of SHC. Phosphorylated EGFR also forms a complex with PLCγ, then PLCγ will be phosphorylated. pMEK is downstream of pSHC. The influence between them is not direct. Double phosphorylated MEK promotes phosphorylation of ERK. The parameters and initial values of the ODE model used for generating input data might affect the direction of influences I observe. We did not calculate the sensitivity or specificity for this complex example, because the nodes in the predicted network are total concentrations, and they are not the same entities as in the Jana model [219].

4.4 Discussion

I evaluated BTD on three simulated datasets, generated respectively from an artificial two-protein model, and two ordinary differential equation (ODE) models of signal transductions. In **Simulation 1**, I demonstrated the advantage of BTD over traditionally used correlation analysis in terms of detecting regulations when time delay exists. I sought to identify the critical number of time points, at and above which the approach can obtain high confidence network structure. In **Simulation 2**, BTD performs well in reconstructing the simple three-tiered cascade. More sample points improves the reconstruction. In **Simulation 3**, the phosphorylated proteins are found to be closely related. Directions of causal influences are determined by time delay detection. However, I have to be cautious while explaining up- and down- regulation. The dynamics of the total concentration of a phosphorylated protein might not be the same as the

dynamics of its addends. The current time delay detection can be applied to oscillatory dynamics, still requiring the stimulus occurring at $t = 0$. More sophisticated methods other than detecting first extremes might be applied to improve delay pattern detection [223,224].

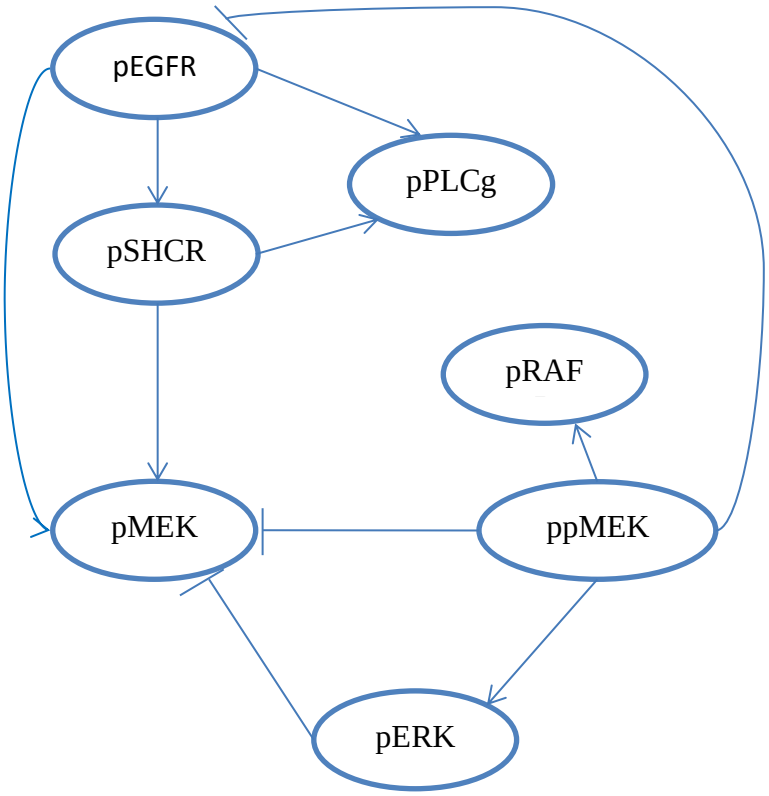


Figure 4.10 Predicted network of EGFR-mediated signal transduction.
pEGFR: total phosphorylated EGFR, pSHC: total phosphorylated SHC, pRLCg: total phosphorylated RLCγ, pRAF: total phosphorylated single phosphorylated RAF, pMEL: total single phosphorylated MEK, ppMEK: total double phosphorylated MEK, pERK: total single phosphorylated ERK.

The BTD method is a new approach for reconstructing networks, including signaling networks, from dynamics data. In BTD, I integrate causal relationship discovery with time-delay pattern detection to identify relationships between species, such as phosphoproteins. I first identify the protein pairs having causal relationships and determine the direction of influence; then assign up-regulation or down-regulation to the protein pairs. Inferred relationships can be connected to reconstruct networks. In real biological experiments, the data on intermediate species in a pathway might not be

collected. The predicted relationships might be direct or indirect, thus further interventional experiments will usually be needed to distinguish them. Reconstructing signaling networks from phosphorylation dynamics remains challenging. Despite the high experimental cost, more time points will improve the performance, especially for fast changing dynamics.

Chapter V

Conclusion and Future Work

Reconstructing biochemical networks from data is challenging, because of the large amount of potential combinational interactions between species, unknown parameters related to those interactions, and multiple potential states. I have made an effort to address a part of this challenge for both equilibrium-state data and dynamic data. My work contains methodology development, methodology evaluation by simulation, and their applications to large-scale datasets. Discovery of networks, such as signaling networks, protein-protein interaction networks and kinetic networks, is the overall goal of my studies. In this thesis, Chapter 2 describes a quantitative analysis pipeline for equilibrium-state phosphoproteomics data with mutational interventions; Chapter 3 presents a general method for analyzing time-course kinetics data, using a differential equation-based method; and Chapter 4 concentrates on dynamic phosphoproteomics and applies a Bayesian method integrated with time delay detection to analyze the time-course data.

In Chapter 2, I presented a novel comprehensive analysis pipeline, incorporating statistical and mathematical methods, for investigating and evaluating quantitative phosphoproteome data, identifying key proteins in specific pathways, discovering the protein-protein interactions and inferring the signaling network. Recent statistical research in proteomics concentrates on protein and peptide identification, while I focused on post-identification analysis of the data. I demonstrated that going far beyond conventional methods such as clustering and functional enrichment which are applied in most phosphoproteomics studies yields real benefits in identification of new

leads for network components and enables further experimental studies. Furthermore, I developed the pipeline to partially compensate the missing value issue haunting proteomics studies, without excluding the incomplete measurements which also contain information.

The major building blocks of the pipeline include phosphopeptide meta-analysis, correlation network analysis and Bayesian method-based causal relationship discovery. I have successfully applied the pipeline to a series of interventional experiments identifying phosphorylation events underlying the transition to filamentous growth in *Saccharomyces cerevisiae* strains. Five of the predicted proteins have been prospectively tested by interventional phenotypic experiments and all of them exhibit differential invasive growth, validating my approach.

The pipeline provides an objective analysis of incomplete phosphoproteomics data and has practical applications for discovering signaling networks using interventional phosphoproteome data. I anticipate the methodology to be applicable as well to other interventional studies via different experiment platforms.

In Chapter 3, I present a computational method MIKANA Ver. 2, an improved version of MIKANA Ver. 1, for network reconstruction from time-course data. It integrates mathematical modeling of biochemical networks with statistical methods to infer kinetic reactions and estimate reaction parameters from data. MIKANA Ver. 1 is a pioneer work that predicts the optimal set of reactions from a large number of possible reactions, and predicts the reaction parameters as well. The MIKANA framework has three-components, pseudo-linear model generation, optimization and parameter fitting, which altogether serve as a tool for network reconstruction without previous knowledge about the reaction mechanism. I have made improvements and extensions to each MIKANA component.

The pseudo-linear model generator has been improved to support autocatalytic reactions and third-order reactions. Non-negative Lasso is used for model selection targeting global optima, instead of the previously used stepwise model selection

method, which might be trapped in local optima. It is coupled with LMF [191,192] non-linear parameter fitting that can be applied to both oscillatory and linear systems, and is more suitable for overdetermined systems, where there are more equations than unknown parameters. Spline smoothing [190] is provided as an option for filtering out noise from input data, which also improves curve fitting. The prediction precision and stability has been improved on simple models, including second-order oscillatory models. Although MIKANA Ver. 2 improves some of the deficiencies of the currently available algorithm, it still has limitations. I present a thorough assessment and discussions of the advantages and limitations of the improved method, and propose possible solutions to overcome the latter.

One of the limitations is that, degeneration of complex systems causes different differential equation systems to have similar behavior in some range of parameters. These systems are indistinguishable to our algorithm without prior information. The reverse engineering task of reconstructing networks (represented by differential equations) from data might give another system as the solution, which degenerates to have almost the same behavior. In the future, instead of presenting only one reconstructed system (supported by MIKANA Ver. 1 and Ver. 2), a set of systems all having perfect fitting might be presented. Sufficient prior information might help to choose the most appropriate one in specific conditions.

Predicting the topology of higher-order oscillatory systems is challenging. The pseudo-linear model is designed to fit the coefficients to precisely predict first derivatives of the concentration. When the oscillatory curve of concentration has the same derivatives with a straight line except several sparse turning points, they are almost indistinguishable to the fitting method. Alternative methods other than pseudo-linear model fitting, such as genetic algorithm, might be used to grow the possible set of reactions from single reactions.

The pseudo-linear model is prone to collinearity issues, where the explanatory variables (the columns in the design matrix) are not independent, and might have linear

relationships. Possible solutions include principal component regression, partial least square regression and so on [207], which worth investigation.

In Chapter 4, I presented a method called BTD (Bayesian + Time Delay detection), a Bayesian method integrated with time delay detection to infer signaling networks from time-course data, such as phosphorylation dynamics data, in response to specific stimuli. The method was designed to fully utilize the "time delay" between upstream and downstream proteins in time-course data, which facilitates discovery of causal influences.

In the BTD workflow, I first identify the protein pairs having causal relationships and determine the direction of influence; then assign up-regulation or down-regulation to the protein pairs. Inferred relationships can be connected to reconstruct networks. Although my approach is developed aiming at phosphorylation dynamics data, it can be readily applied to other biological networks which display time-delay patterns.

I evaluated the performance and stability of my approach on simulated datasets. The approach is able to detect regulations with time delays which are ignored or wrongly identified using only correlation coefficient. I also sought to define the critical number of time points at and above which the approach can obtain higher confidence network structures. The method works well on my simple models (Simulation 1 & 2). For complex models with intermediates that are unable to measure (Simulation 3), the total concentrations of phosphorylated forms are calculated by summing up all intermediates. These total concentrations are used to reconstruct a simplified approximate structure. The predicted network structure of total phosphorylated forms in Simulation 3 can be partially explained by biological knowledge.

In real biological experiments, the intermediates in a pathway might be missed in the datasets. The predicted relationships might be direct or indirect, thus further interventional experiments are needed to distinguish them. Despite the high experimental cost, more time points will improve the performance, especially for fast changing dynamics.

The current time delay detection can be applied to oscillatory dynamics, still requiring the stimulus occurring at $t = 0$. More sophisticated methods might also be developed and applied to improve delay pattern detection. Time series analysis and signal processing techniques provide a resource of tools that I can investigate in future studies.

In summary, I have developed and applied methods for analyzing large-scale phosphoproteomics data, and also performed mathematical modeling and reaction reconstruction for smaller-scale kinetics. The pipeline presented in Chapter 2 for analyzing large-scale interventional equilibrium-state phosphorylation data has been validated by experiments, future applications are anticipated for the newly developed pipeline. Chapter 3 presented my efforts to reconstruct kinetic networks from time-course data. Both improvements and limitations are thoroughly evaluated. Possible solutions have been proposed to reduce limitations and improve the performance of the algorithm. Chapter 4 presented a new method integrating a Bayesian method with time delay detection to identify upstream-downstream relationships between proteins, with a special focus on phosphorylation dynamics. More sophisticated methods for time delay detection might improve the performance of network prediction. All of these studies have benefited me by providing a better understanding of biological and chemical networks, which helps to elucidate the regulation of biological processes, such as signaling pathways.

Bibliography

1. Käll L, Vitek O (2011) Computational mass spectrometry-based proteomics. PLoS computational biology 7: e1002277.

2. Anderson NL, Anderson NG (1998) Proteome and proteomics: new technologies, new concepts, and new words. Electrophoresis 19: 1853–1861.

3. Blackstock WP, Weir MP (1999) Proteomics: quantitative and physical mapping of cellular proteins. Trends in biotechnology 17: 121–127.

4. Vitek O (2009) Getting started in computational mass spectrometry-based proteomics. PLoS computational biology 5: e1000366. doi:10.1371/journal.pcbi.1000366.

5. Menon R, Omenn GS (2011) Identification of alternatively spliced transcripts using a proteomic informatics approach. Methods in molecular biology (Clifton, NJ) 696: 319–326. doi:10.1007/978-1-60761-987-1_20.

6. Menon R, Zhang Q, Zhang Y, Fermin D, Bardeesy N, et al. (2009) Identification of novel alternative splice isoforms of circulating proteins in a mouse model of human pancreatic cancer. Cancer research 69: 300–309. doi:10.1158/0008-5472.CAN-08-2145.

7. Ptacek J, Devgan G, Michaud G, Zhu H, Zhu X, et al. (2005) Global analysis of protein phosphorylation in yeast. Nature 438: 679–684. doi:10.1038/nature04187.

8. Manning G, Plowman GD, Hunter T, Sudarsanam S (2002) Evolution of protein kinase signaling from yeast to man. Trends in biochemical sciences 27: 514–520.

9. Ficarro SB, McCleland ML, Stukenberg PT, Burke DJ, Ross MM, et al. (2002) Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. Nature biotechnology 20: 301–305. doi:10.1038/nbt0302-301.

10. Cohen P (2000) The regulation of protein function by multisite phosphorylation--a 25 year update. Trends in biochemical sciences 25: 596–601.

11. Bodenmiller B, Wanka S, Kraft C, Urban J, Campbell D, et al. (2010) Phosphoproteomic analysis reveals interconnected system-wide responses to perturbations of kinases and phosphatases in yeast. Science signaling 3: rs4. doi:10.1126/scisignal.2001182.

12. Breitkreutz A, Choi H, Sharom JR, Boucher L, Neduva V, et al. (2010) A global protein kinase and phosphatase interaction network in yeast. Science 328: 1043–1046. doi:10.1126/science.1176495.

13. Yachie N, Saito R, Sugiyama N, Tomita M, Ishihama Y (2011) Integrative features of the yeast phosphoproteome and protein-protein interaction map. PLoS computational biology 7: e1001064. doi:10.1371/journal.pcbi.1001064.

14. Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B (2007) Quantitative mass spectrometry in proteomics: a critical review. Analytical and bioanalytical chemistry 389: 1017–1031. doi:10.1007/s00216-007-1486-6.

15. Cappadona S, Baker PR, Cutillas PR, Heck AJR, van Breukelen B (2012) Current challenges in software solutions for mass spectrometry-based quantitative proteomics. Amino acids 43: 1087–1108. doi:10.1007/s00726-012-1289-8.

16. Kumar C, Mann M (2009) Bioinformatics analysis of mass spectrometry-based proteomics data sets. FEBS letters 583: 1703–1712. doi:10.1016/j.febslet.2009.03.035.

17. Olsen JV, Blagoev B, Gnad F, Macek B, Kumar C, et al. (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell 127: 635–648.

18. Oberg AL, Vitek O (2009) Statistical design of quantitative mass spectrometry-based proteomic experiments. Journal of proteome research 8: 2144–2156. doi:10.1021/pr8010099.

19. Noble WS, MacCoss MJ (2012) Computational and statistical analysis of protein mass spectrometry data. PLoS computational biology 8: e1002296.

20. Martens L (2011) Bioinformatics challenges in mass spectrometry-driven proteomics. Methods in molecular biology (Clifton, NJ) 753: 359–371. doi:10.1007/978-1-61779-148-2_24.

21. Jones AR (2012) Bioinformatics Challenges and Solutions in Proteomics as Quantitative Methods Mature. Omics : a journal of integrative biology. doi:10.1089/omi.2012.0051.

22. Nesvizhskii AI (2010) A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. Journal of proteomics 73: 2092–2123. doi:10.1016/j.jprot.2010.08.009.

23. Nesvizhskii AI (2012) Computational and informatics strategies for identification of specific protein interaction partners in affinity purification mass spectrometry experiments. Proteomics 12: 1639–1655.

24. Cox J, Mann M (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nature biotechnology 26: 1367–1372. doi:10.1038/nbt.1511.

25. Nesvizhskii AI, Keller A, Kolker E, Aebersold R (2003) A statistical model for identifying proteins by tandem mass spectrometry. Analytical chemistry 75: 4646–4658.

26. Ma Z-Q, Dasari S, Chambers MC, Litton MD, Sobecki SM, et al. (2009) IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. Journal of proteome research 8: 3872–3881. doi:10.1021/pr900360j.

27. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics 25: 25–29. doi:10.1038/75556.

28. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research 28: 27–30.

29. Glatter T, Wepf A, Aebersold R, Gstaiger M (2009) An integrated workflow for charting the human interaction proteome: insights into the PP2A system. Molecular systems biology 5: 237. doi:10.1038/msb.2008.75.

30. Mourão MA, Srividhya J, McSharry PE, Crampin EJ, Schnell S (2011) A graphical user interface for a method to infer kinetics and network architecture (MIKANA). PLoS ONE 6: e27534. doi:10.1371/journal.pone.0027534.

31. Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. Nature 422: 198–207. doi:10.1038/nature01511.

32. Ong S-E, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, et al. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Molecular & cellular proteomics : MCP 1: 376–386.

33. Mann M (2006) Functional and quantitative proteomics using SILAC. Nature reviews Molecular cell biology 7: 952–958. doi:10.1038/nrm2067.

34. Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, et al. (2004) Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Molecular & cellular proteomics : MCP 3: 1154–1169. doi:10.1074/mcp.M400129-MCP200.

35. Keshamouni VG, Jagtap P, Michailidis G, Strahler JR, Kuick R, et al. (2009) Temporal quantitative proteomics by iTRAQ 2D-LC-MS/MS and corresponding mRNA expression analysis identify post-transcriptional modulation of actin-cytoskeleton regulators during TGF-beta-Induced epithelial-mesenchymal transition. Journal of proteome research 8: 35–47. doi:10.1021/pr8006478.

36. Yao X, Afonso C, Fenselau C (n.d.) Dissection of proteolytic 18O labeling: endoprotease-catalyzed 16O-to-18O exchange of truncated peptide substrates. Journal of proteome research 2: 147–152.

37. Ong S-E, Mann M (2005) Mass spectrometry-based proteomics turns quantitative. Nature chemical biology 1: 252–262. doi:10.1038/nchembio736.

38. Wolf-Yadlin A, Hautaniemi S, Lauffenburger DA, White FM (2007) Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks. Proceedings of the National Academy of Sciences of the United States of America 104: 5860–5865. doi:10.1073/pnas.0608638104.

39. Reker D, Malmström L (2012) Bioinformatic challenges in targeted proteomics. Journal of proteome research. doi:10.1021/pr300276f.

40. Blagoev B, Ong S-E, Kratchmarova I, Mann M (2004) Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics. Nature biotechnology 22: 1139–1145. doi:10.1038/nbt1005.

41. Andersen JS, Lam YW, Leung AKL, Ong S-E, Lyon CE, et al. (2005) Nucleolar proteome dynamics. Nature 433: 77–83. doi:10.1038/nature03207.

42. Zhang Y, Wolf-Yadlin A, Ross PL, Pappin DJ, Rush J, et al. (2005) Time-resolved mass spectrometry of tyrosine phosphorylation sites in the epidermal growth factor receptor signaling network reveals dynamic modules. Molecular & cellular proteomics : MCP 4: 1240–1250. doi:10.1074/mcp.M500089-MCP200.

43. Theodoridis S, Koutroumbas K (2006) Pattern Recognition, Third Edition: 635.

44. Abdi H, Williams LJ (2010) Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics 2: 433–459. doi:10.1002/wics.101.

45. Pearson K (1901) On lines and planes of closest fit to systems of points in space. Philosophical Magazine 2: 559–572.

46. Futschik ME, Carlisle B (2005) Noise-robust soft clustering of gene expression time-course data. Journal of bioinformatics and computational biology 3: 965–988.

47. Tseng GC, Wong WH (2005) Tight clustering: a resampling-based approach for identifying stable and tight patterns in data. Biometrics 61: 10–16.

48. Imamura H, Yachie N, Saito R, Ishihama Y, Tomita M (2010) Towards the systematic discovery of signal transduction networks using phosphorylation dynamics data. BMC bioinformatics 11: 232. doi:10.1186/1471-2105-11-232.

49. Hendrickx DM, Hoefsloot HCJ, Hendriks MMWB, Vis DJ, Canelas AB, et al. (2012) Inferring differences in the distribution of reaction rates across conditions. Molecular BioSystems 8: 2415–2423. doi:10.1039/c2mb25015b.

50. Friedman N, Linial M, Nachman I, Pe'er D (2000) Using Bayesian networks to analyze expression data. Journal of computational biology : a journal of computational molecular cell biology 7: 601–620. doi:10.1089/106652700750050961.

51. Needham CJ, Bradford JR, Bulpitt AJ, Westhead DR (2007) A primer on learning in Bayesian networks for computational biology. PLoS computational biology 3: e129. doi:10.1371/journal.pcbi.0030129.

52. Cooper G, Yoo C (1999) Causal Discovery from a Mixture of Experimental and Observational Data. Proceedings of Uncertainty in Artificial Intelligence. pp. 116–125.

53. Pe'er D, Regev A, Elidan G, Friedman N (2001) Inferring subnetworks from perturbed expression profiles. Bioinformatics (Oxford, England) 17 Suppl 1: S215–24.

54. Djebbari A, Quackenbush J (2008) Seeded Bayesian Networks: constructing genetic networks from microarray data. BMC systems biology 2: 57. doi:10.1186/1752-0509-2-57.

55. Shah A, Tenzen T, McMahon AP, Woolf PJ (2009) Using mechanistic Bayesian networks to identify downstream targets of the sonic hedgehog pathway. BMC bioinformatics 10: 433. doi:10.1186/1471-2105-10-433.

56. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP (2005) Causal protein-signaling networks derived from multiparameter single-cell data. Science (New York, NY) 308: 523–529. doi:10.1126/science.1105809.

57. Crampin EJ, Schnell S, McSharry PE (2004) Mathematical and computational techniques to deduce complex biochemical reaction mechanisms. Progress in biophysics and molecular biology 86: 77–112. doi:10.1016/j.pbiomolbio.2004.04.002.

58. Tyson JJ, Chen K, Novak B (2001) Network dynamics and cell physiology. Nature reviews Molecular cell biology 2: 908–916. doi:10.1038/35103078.

59. Novák B, Tyson JJ (2004) A model for restriction point control of the mammalian cell cycle. Journal of theoretical biology 230: 563–579. doi:10.1016/j.jtbi.2004.04.039.

60. Brightman FA, Fell DA (2000) Differential feedback regulation of the MAPK cascade underlies the quantitative differences in EGF and NGF signalling in PC12 cells. FEBS letters 482: 169–174.

61. Queralt E, Lehane C, Novak B, Uhlmann F (2006) Downregulation of PP2A(Cdc55) phosphatase by separase initiates mitotic exit in budding yeast. Cell 125: 719–732. doi:10.1016/j.cell.2006.03.038.

62. Wynn ML, Ventura AC, Sepulchre JA, García HJ, Merajver SD (2011) Kinase inhibitors can produce off-target effects and activate linked pathways by retroactivity. BMC systems biology 5: 156. doi:10.1186/1752-0509-5-156.

63. Srividhya J, Crampin EJ, McSharry PE, Schnell S (2007) Reconstructing biochemical pathways from time course data. Proteomics 7: 828–838. doi:10.1002/pmic.200600428.

64. Aldridge BB, Saez-Rodriguez J, Muhlich JL, Sorger PK, Lauffenburger DA (2009) Fuzzy logic analysis of kinase pathway crosstalk in TNF/EGF/insulin-induced signaling. PLoS computational biology 5: e1000340. doi:10.1371/journal.pcbi.1000340.

65. Fuzzy Logic Toolbox: User's Guide (r2012a). (2012). The MathWorks, Inc. Available:http://www.mathworks.com/help/toolbox/fuzzy/fp754.html. Accessed 24 August 1BC.

66. Bornholdt S (2008) Boolean network models of cellular regulation: prospects and limitations. Journal of the Royal Society, Interface / the Royal Society 5 Suppl 1: S85–94. doi:10.1098/rsif.2008.0132.focus.

67. Wang R-S, Saadatpour A, Albert R (2012) Boolean modeling in systems biology: an overview of methodology and applications. Physical Biology 9: 055001. doi:10.1088/1478-3975/9/5/055001.

68. Cowpertwait PSP, Metcalfe AV (2009) Introductory Time Series with R. Gentleman R, Hornik K, Parmigiani G, editors Springer. p. doi:10.1007/978-0-387-88698-5.

69. Guo S, Wu J, Ding M, Feng J (2008) Uncovering interactions in the frequency domain. PLoS computational biology 4: e1000087. doi:10.1371/journal.pcbi.1000087.

70. Wu J, Liu X, Feng J (2008) Detecting causality between different frequencies. Journal of neuroscience methods 167: 367–375. doi:10.1016/j.jneumeth.2007.08.022.

71. Linding R, Jensen LJ, Ostheimer GJ, van Vugt MATM, Jørgensen C, et al. (2007) Systematic discovery of in vivo phosphorylation networks. Cell 129: 1415–1426. doi:10.1016/j.cell.2007.05.052.

72. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, et al. (2002) Network motifs: simple building blocks of complex networks. Science (New York, NY) 298: 824–827. doi:10.1126/science.298.5594.824.

73. Diestel R (2005) Graph theory, volume 173 of Graduate Texts in Mathematics. 4th, editor Springer-Verlag, Heidelberg. p.

74. Watts DJ, Strogatz SH (1998) Collective dynamics of “small-world” networks. Nature 393: 440–442. doi:10.1038/30918.

75. Albert R, DasGupta B, Dondi R, Kachalo S, Sontag E, et al. (2007) A novel method for signal transduction network inference from indirect experimental evidence. Journal of Computational Biology 14: 927–949. doi:10.1089/cmb.2007.0015.

76. Yeung MKS, Tegnér J, Collins JJ (2002) Reverse engineering gene networks using singular value decomposition and robust regression. Proceedings of the National Academy of Sciences of the United States of America 99: 6163–6168. doi:10.1073/pnas.092576199.

77. Schmelzle K, White FM (2006) Phosphoproteomic approaches to elucidate cellular signaling networks. Current opinion in biotechnology 17: 406–414. doi:10.1016/j.copbio.2006.06.004.

78. Macek B, Mann M, Olsen JV (2009) Global and site-specific quantitative phosphoproteomics: principles and applications. Annual review of pharmacology and toxicology 49: 199–221. doi:10.1146/annurev.pharmtox.011008.145606.

79. Kosako H, Nagano K (2011) Quantitative phosphoproteomics strategies for understanding protein kinase-mediated signal transduction pathways. Expert review of proteomics 8: 81–94. doi:10.1586/epr.10.104.

80. Thorner J, Truckses D, Garrenton L (2012) Filamentous Growth Pathway in Yeast. Science Signaling Database of Cell Signaling. Available:http://stke.sciencemag.org/cgi/cm/stkecm;CMP_14554. Accessed 3 February 1BC.

81. Bharucha N, Ma J, Dobry CJ, Lawson SK, Yang Z, et al. (2008) Analysis of the yeast kinome reveals a network of regulated protein localization during filamentous growth. Molecular biology of the cell 19: 2708–2717. doi:10.1091/mbc.E07-11-1199.

82. Edgington NP, Blacketer MJ, Bierwagen TA, Myers AM (1999) Control of Saccharomyces cerevisiae filamentous growth by cyclin-dependent kinase Cdc28. Molecular and cellular biology 19: 1369–1380.

83. Palecek SP, Parikh AS, Kron SJ (2002) Sensing, signalling and integrating physical processes during Saccharomyces cerevisiae invasive and filamentous growth. Microbiology (Reading, England) 148: 893–907.

84. Ceccato-Antonini SR, Sudbery PE (2004) Filamentous growth in Saccharomyces cerevisiae. Brazilian Journal of Microbiology 35: 173–181. doi:10.1590/S1517-83822004000200001.

85. Saito H (2010) Regulation of cross-talk in yeast MAPK signaling pathways. Current opinion in microbiology 13: 677–683. doi:10.1016/j.mib.2010.09.001.

86. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20: 3551–3567. doi:10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2.

87. Kratchmarova I, Blagoev B, Haack-Sorensen M, Kassem M, Mann M (2005) Mechanism of divergent growth factor effects in mesenchymal stem cell

differentiation. Science (New York, NY) 308: 1472–1477. doi:10.1126/science.1107627.

88. Boersema PJ, Aye TT, van Veen TAB, Heck AJR, Mohammed S (2008) Triplex protein quantification based on stable isotope labeling by peptide dimethylation applied to cell and tissue lysates. Proteomics 8: 4624–4632. doi:10.1002/pmic.200800297.

89. Larsen MR, Thingholm TE, Jensen ON, Roepstorff P, Jørgensen TJD (2005) Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns. Molecular & cellular proteomics : MCP 4: 873–886. doi:10.1074/mcp.T500007-MCP200.

90. Thingholm TE, Jørgensen TJD, Jensen ON, Larsen MR (2006) Highly selective enrichment of phosphorylated peptides using titanium dioxide. Nature protocols 1: 1929–1935. doi:10.1038/nprot.2006.185.

91. Von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, et al. (2003) STRING: a database of predicted functional associations between proteins. Nucleic Acids Research 31: 258–261.

92. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, et al. (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Research 39: D561–8. doi:10.1093/nar/gkq973.

93. Hastie T, Tibshirani R, Sherlock G, Eisen M, Brown P, et al. (1999) Imputing missing data for gene expression arrays. Stanford University Statistics Department Technical report.

94. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, et al. (2001) Missing value estimation methods for DNA microarrays. Bioinformatics (Oxford, England) 17: 520–525.

95. Ward JHJ (1963) Hierarchical Grouping to Optimize an Objective Function. 58: 236–244.

96. Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols 4: 44–57. doi:10.1038/nprot.2008.211.

97. Huang DW, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research 37: 1–13. doi:10.1093/nar/gkn923.

98. Miller RG (1981) Simultaneous Statistical Inference. 2nd ed. New York: Springer-Verlag. p.

99. Fisher RA (1925) Statistical methods for research workers. Boyd OA, editor Oliver and Boyd. p.

100. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B 57: 289–300.

101. Li J, Tseng GC (2011) An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. The Annals of Applied Statistics 5: 994–1019. doi:10.1214/10-AOAS393.

102. SGD project (2012) Saccharomyces Genome Database. Available:http://www.yeastgenome.org/. Accessed 16 August 2012.

103. Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, et al. (2011) Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic acids research. doi:10.1093/nar/gkr1029.

104. Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, et al. (2006) BioGRID: a general repository for interaction datasets. Nucleic acids research 34: D535–9. doi:10.1093/nar/gkj109.

105. Yoo C, Thorsson V, Cooper G (2002) Discovery of Causal Relationships in a Gene-regulation Pathway from a Mixture of Experimental and Observational DNA Microarray Data. Pacific Symposium on Biocomputing 7. pp. 498–509.

106. Mani S, Cooper GF (2004) Causal discovery using a Bayesian local causal discovery algorithm. Studies in health technology and informatics 107: 731–735.

107. Cooper GF, Herskovits E (1992) A Bayesian Method for the Induction of Probabilistic Networks from Data. Machine Learning 9: 309–347. doi:10.1023/A:1022649401552.

108. David Heckerman DMC (1995) Learning Bayesian networks: The combination of knowledge and statistical data. Machine Learning 20: 197–243.

109. Murphy K (2011) Bayes Net Toolbox for Matlab. Available:http://code.google.com/p/bnt/. Accessed 12 May 1BC.

110. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research 13: 2498–2504. doi:10.1101/gr.1239303.

111. Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics (Oxford, England) 27: 431–432. doi:10.1093/bioinformatics/btq675.

112. Hubbard TJP, Aken BL, Beal K, Ballester B, Caccamo M, et al. (2007) Ensembl 2007. Nucleic acids research 35: D610–7.

113. Zähringer H, Burgert M, Holzer H, Nwaka S (1997) Neutral trehalase Nth1p of Saccharomyces cerevisiae encoded by the NTH1 gene is a multiple stress responsive protein. FEBS letters 412: 615–620.

114. Nwaka S, Holzer H (1998) Molecular biology of trehalose and the trehalases in the yeast Saccharomyces cerevisiae. Progress in nucleic acid research and molecular biology 58: 197–237.

115. Yoshikawa K, Tanaka T, Furusawa C, Nagahisa K, Hirasawa T, et al. (2009) Comprehensive phenotypic analysis for identification of genes affecting growth under ethanol stress in Saccharomyces cerevisiae. FEMS yeast research 9: 32–44. doi:10.1111/j.1567-1364.2008.00456.x.

116. Lee M-W, Kim B-J, Choi H-K, Ryu M-J, Kim S-B, et al. (2007) Global protein expression profiling of budding yeast in response to DNA damage. Yeast (Chichester, England) 24: 145–154. doi:10.1002/yea.1446.

117. Hatzixanthis K, Mollapour M, Seymour I, Bauer BE, Krapf G, et al. (2003) Moderately lipophilic carboxylate compounds are the selective inducers of the Saccharomyces cerevisiae Pdr12p ATP-binding cassette transporter. Yeast (Chichester, England) 20: 575–585. doi:10.1002/yea.981.

118. Schüller C, Mamnun YM, Mollapour M, Krapf G, Schuster M, et al. (2004) Global phenotypic analysis and transcriptional profiling defines the weak acid stress response regulon in Saccharomyces cerevisiae. Molecular biology of the cell 15: 706–720. doi:10.1091/mbc.E03-05-0322.

119. Kren A, Mamnun YM, Bauer BE, Schüller C, Wolfger H, et al. (2003) War1p, a novel transcription factor controlling weak acid stress response in yeast. Molecular and cellular biology 23: 1775–1785.

120. Holyoak CD, Thompson S, Ortiz Calderon C, Hatzixanthis K, Bauer B, et al. (2000) Loss of Cmk1 Ca(2+)-calmodulin-dependent protein kinase in yeast results in

constitutive weak organic acid resistance, associated with a post-transcriptional activation of the Pdr12 ATP-binding cassette transporter. Molecular microbiology 37: 595–605.

121. Wotton D, Freeman K, Shore D (1996) Multimerization of Hsp42p, a novel heat shock protein of Saccharomyces cerevisiae, is dependent on a conserved carboxyl-terminal sequence. The Journal of biological chemistry 271: 2717–2723.

122. François J, Parrou JL (2001) Reserve carbohydrates metabolism in the yeast Saccharomyces cerevisiae. FEMS Microbiology Reviews 25: 125–145. doi:10.1016/S0168-6445(00)00059-0.

123. Galello F, Portela P, Moreno S, Rossi S (2010) Characterization of substrates that have a differential effect on Saccharomyces cerevisiae protein kinase A holoenzyme activation. The Journal of biological chemistry 285: 29770–29779. doi:10.1074/jbc.M110.120378.

124. Panni S, Landgraf C, Volkmer-Engert R, Cesareni G, Castagnoli L (2008) Role of 14-3-3 proteins in the regulation of neutral trehalase in the yeast Saccharomyces cerevisiae. FEMS yeast research 8: 53–63. doi:10.1111/j.1567-1364.2007.00312.x.

125. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, et al. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440: 637–643. doi:10.1038/nature04670.

126. Santangelo GM (2006) Glucose signaling in Saccharomyces cerevisiae. Microbiology and molecular biology reviews : MMBR 70: 253–282. doi:10.1128/MMBR.70.1.253-282.2006.

127. Estruch F (2000) Stress-controlled transcription factors, stress-induced genes and stress tolerance in budding yeast. FEMS microbiology reviews 24: 469–486.

128. Norbeck J, Blomberg A (2000) The level of cAMP-dependent protein kinase A activity strongly affects osmotolerance and osmo-instigated gene expression changes in Saccharomyces cerevisiae. Yeast (Chichester, England) 16: 121–137. doi:10.1002/(SICI)1097-0061(20000130)16:2<121::AID-YEA511>3.0.CO;2-A.

129. Barbieri M, Bonafè M, Franceschi C, Paolisso G (2003) Insulin/IGF-I-signaling pathway: an evolutionarily conserved mechanism of longevity from yeast to humans. American journal of physiology Endocrinology and metabolism 285: E1064–71. doi:10.1152/ajpendo.00296.2003.

130. Roberts RL, Mösch HU, Fink GR (1997) 14-3-3 proteins are essential for RAS/MAPK cascade signaling during pseudohyphal development in S. cerevisiae. Cell 89: 1055–1065.

131. Gancedo JM (2001) Control of pseudohyphae formation in Saccharomyces cerevisiae. FEMS microbiology reviews 25: 107–123.

132. Bertram PG, Zeng C, Thorson J, Shaw AS, Zheng XF (1998) The 14-3-3 proteins positively regulate rapamycin-sensitive signaling. Current biology : CB 8: 1259–1267.

133. Irie K, Gotoh Y, Yashar BM, Errede B, Nishida E, et al. (1994) Stimulatory effects of yeast and mammalian 14-3-3 proteins on the Raf protein kinase. Science (New York, NY) 265: 1716–1719.

134. Gelperin D, Weigle J, Nelson K, Roseboom P, Irie K, et al. (1995) 14-3-3 proteins: potential roles in vesicular transport and Ras signaling in Saccharomyces cerevisiae. Proceedings of the National Academy of Sciences of the United States of America 92: 11539–11543.

135. Butcher RA, Bhullar BS, Perlstein EO, Marsischky G, LaBaer J, et al. (2006) Microarray-based method for monitoring yeast overexpression strains reveals small-molecule targets in TOR pathway. Nature chemical biology 2: 103–109. doi:10.1038/nchembio762.

136. Breslow DK, Cameron DM, Collins SR, Schuldiner M, Stewart-Ornstein J, et al. (2008) A comprehensive strategy enabling high-resolution functional analysis of the yeast genome. Nature methods 5: 711–718. doi:10.1038/nmeth.1234.

137. Woo DK, Phang TL, Trawick JD, Poyton RO (2009) Multiple pathways of mitochondrial-nuclear communication in yeast: intergenomic signaling involves ABF1 and affects a different set of genes than retrograde regulation. Biochimica et Biophysica Acta 1789: 135–145.

138. Borah S, Shivarathri R, Kaur R (2011) The Rho1 GTPase-activating protein CgBem2 is required for survival of azole stress in Candida glabrata. The Journal of biological chemistry 286: 34311–34324. doi:10.1074/jbc.M111.264671.

139. Montañés FM, Pascual-Ahuir A, Proft M (2011) Repression of ergosterol biosynthesis is essential for stress resistance and is mediated by the Hog1 MAP kinase and the Mot3 and Rox1 transcription factors. Molecular Microbiology 79: 1008–1023. doi:10.1111/j.1365-2958.2010.07502.x.

140. Fromont-Racine M, Mayes AE, Brunet-Simon A, Rain JC, Colley A, et al. (2000) Genome-wide protein interaction screens reveal functional networks involving Sm-like proteins. Yeast (Chichester, England) 17: 95–110. doi:10.1002/1097-0061(20000630)17:2<95::AID-YEA16>3.0.CO;2-H.

141. Gavin A-C, Aloy P, Grandi P, Krause R, Boesche M, et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature 440: 631–636. doi:10.1038/nature04532.

142. Collins SR, Kemmeren P, Zhao X-C, Greenblatt JF, Spencer F, et al. (2007) Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. Molecular & cellular proteomics : MCP 6: 439–450. doi:10.1074/mcp.M600381-MCP200.

143. Yoon J-H, Choi E-J, Parker R (2010) Dcp2 phosphorylation by Ste20 modulates stress granule assembly and mRNA decay in Saccharomyces cerevisiae. The Journal of cell biology 189: 813–827. doi:10.1083/jcb.200912019.

144. Kim J, Kim J (2002) KEM1 is involved in filamentous growth of Saccharomyces cerevisiae. FEMS microbiology letters 216: 33–38.

145. Cai H, Kauffman S, Naider F, Becker JM (2006) Genomewide screen reveals a wide regulatory network for di/tripeptide utilization in Saccharomyces cerevisiae. Genetics 172: 1459–1476. doi:10.1534/genetics.105.053041.

146. Pan X, Reissman S, Douglas NR, Huang Z, Yuan DS, et al. (2010) Trivalent arsenic inhibits the functions of chaperonin complex. Genetics 186: 725–734. doi:10.1534/genetics.110.117655.

147. Jin R, Dobry CJ, McCown PJ, Kumar A (2008) Large-scale analysis of yeast filamentous growth by systematic gene disruption and overexpression. Molecular biology of the cell 19: 284–296. doi:10.1091/mbc.E07-05-0519.

148. Yuzyuk T, Amberg DC (2003) Actin Recovery and Bud Emergence in Osmotically Stressed Cells Requires the Conserved Actin Interacting Mitogen-activated Protein Kinase Kinase Kinase Ssk2p/MTK1 and the Scaffold Protein Spa2p. Molecular Biology of the Cell 14: 3013–3026.

149. Parsons AB, Brost RL, Ding H, Li Z, Zhang C, et al. (2004) Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. Nature biotechnology 22: 62–69. doi:10.1038/nbt919.

150. Kapitzky L, Beltrao P, Berens TJ, Gassner N, Zhou C, et al. (2010) Cross-species chemogenomic profiling reveals evolutionarily conserved drug mode of action. Molecular systems biology 6: 451.

151. Li Z, Lee I, Moradi E, Hung N-J, Johnson AW, et al. (2009) Rational extension of the ribosome biogenesis pathway using network-guided genetics. PLoS biology 7: e1000213.

152. Sinha H, David L, Pascon RC, Clauder-Münster S, Krishnakumar S, et al. (2008) Sequential elimination of major-effect contributors identifies additional quantitative trait loci conditioning high-temperature growth in yeast. Genetics 180: 1661–1670.

153. Howard SC, Hester A, Herman PK (2003) The Ras/PKA signaling pathway may control RNA polymerase II elongation via the Spt4p/Spt5p complex in Saccharomyces cerevisiae. Genetics 165: 1059–1070.

154. Gimeno CJ, Ljungdahl PO, Styles CA, Fink GR (1992) Unipolar cell divisions in the yeast S. cerevisiae lead to filamentous growth: regulation by starvation and RAS. Cell 68: 1077–1090.

155. Klopf E, Paskova L, Solé C, Mas G, Petryshyn A, et al. (2009) Cooperation between the INO80 complex and histone chaperones determines adaptation of stress gene transcription in the yeast Saccharomyces cerevisiae. Molecular and cellular biology 29: 4994–5007. doi:10.1128/MCB.01858-08.

156. Kurischko C, Kim HK, Kuravi VK, Pratzka J, Luca FC (2011) The yeast Cbk1 kinase regulates mRNA localization via the mRNA-binding protein Ssd1. The Journal of cell biology 192: 583–598. doi:10.1083/jcb.201011061.

157. Koh JLY, Ding H, Costanzo M, Baryshnikova A, Toufighi K, et al. (2010) DRYGIN: a database of quantitative genetic interaction networks in yeast. Nucleic acids research 38: D502–7. doi:10.1093/nar/gkp820.

158. Fiedler D, Braberg H, Mehta M, Chechik G, Cagney G, et al. (2009) Functional organization of the S. cerevisiae phosphorylation network. Cell 136: 952–963. doi:10.1016/j.cell.2008.12.039.

159. Ramezani Rad M, Jansen G, Bühring F, Hollenberg CP (1998) Ste50p is involved in regulating filamentous growth in the yeast Saccharomyces cerevisiae and associates with Ste11p. Molecular & general genetics : MGG 259: 29–38.

160. Hoppins S, Collins SR, Cassidy-Stone A, Hummel E, DeVay RM, et al. (2011) A mitochondrial-focused genetic interaction map reveals a scaffold-like complex

required for inner membrane organization in mitochondria. The Journal of Cell Biology 195: 323–340. doi:10.1083/jcb.201107053.

161. Entian KD, Schuster T, Hegemann JH, Becher D, Feldmann H, et al. (1999) Functional analysis of 150 deletion mutants in Saccharomyces cerevisiae by a systematic approach. Molecular & general genetics : MGG 262: 683–702.

162. Kang CM, Jiang YW (2005) Genome-wide survey of non-essential genes required for slowed DNA synthesis-induced filamentous growth in yeast. Yeast (Chichester, England) 22: 79–90. doi:10.1002/yea.1195.

163. Seiler S, Vogt N, Ziv C, Gorovits R, Yarden O (2006) The STE20/germinal center kinase POD6 interacts with the NDR kinase COT1 and is involved in polar tip extension in Neurospora crassa. Molecular Biology of the Cell 17: 4080–4092.

164. Mir SS, Fiedler D, Cashikar AG (2009) Ssd1 Is Required for Thermotolerance and Hsp104-Mediated Protein Disaggregation in Saccharomyces cerevisiae. Molecular and Cellular Biology 29: 187–200.

165. Kurischko C, Kuravi VK, Herbert CJ, Luca FC (2011) Nucleocytoplasmic shuttling of Ssd1 defines the destiny of its bound mRNAs. Molecular Microbiology 81: 831–849. doi:10.1111/j.1365-2958.2011.07731.x.

166. Bandyopadhyay S, Mehta M, Kuo D, Sung M-K, Chuang R, et al. (2010) Rewiring of genetic networks in response to DNA damage. Science (New York, NY) 330: 1385–1389. doi:10.1126/science.1195618.

167. Humphrey T, Enoch T (1998) Sum1, a highly conserved WD-repeat protein, suppresses S-M checkpoint mutants and inhibits the osmotic stress cell cycle response in fission yeast. Genetics 148: 1731–1742.

168. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, et al. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nature genetics 34: 166–176. doi:10.1038/ng1165.

169. Thomsen R, Saguez C, Nasser T, Jensen TH (2008) General, rapid, and transcription-dependent fragmentation of nucleolar antigens in S. cerevisiae mRNA export mutants. RNA (New York, NY) 14: 706–716. doi:10.1261/rna.718708.

170. Fleming JA, Lightcap ES, Sadis S, Thoroddsen V, Bulawa CE, et al. (2002) Complementary whole-genome technologies reveal the cellular response to proteasome inhibition by PS-341. Proceedings of the National Academy of Sciences of the United States of America 99: 1461–1466. doi:10.1073/pnas.032516399.

171. Zhou X, Arita A, Ellen TP, Liu X, Bai J, et al. (2009) A genome-wide screen in Saccharomyces cerevisiae reveals pathways affected by arsenic toxicity. Genomics 94: 294–307. doi:10.1016/j.ygeno.2009.07.003.

172. Shin C-S, Kim SY, Huh W-K (2009) TORC1 controls degradation of the transcription factor Stp1, a key effector of the SPS amino-acid-sensing pathway in Saccharomyces cerevisiae. Journal of cell science 122: 2089–2099.

173. Buchan JR, Muhlrad D, Parker R (2008) P bodies promote stress granule assembly in Saccharomyces cerevisiae. The Journal of cell biology 183: 441–455.

174. Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, et al. (2010) The genetic landscape of a cell. Science (New York, NY) 327: 425–431. doi:10.1126/science.1180823.

175. Chen M, Gutierrez GJ, Ronai ZA (2011) Ubiquitin-recognition protein Ufd1 couples the endoplasmic reticulum (ER) stress response to cell cycle control. Proceedings of the National Academy of Sciences of the United States of America 108: 9119–9124. doi:10.1073/pnas.1100028108.

176. Hsieh M-T, Chen R-H (2011) Cdc48 and Cofactors Npl4-Ufd1 Are Important for G1 Progression during Heat Stress by Maintaining Cell Wall Integrity in Saccharomyces cerevisiae. PLoS ONE 6: 13.

177. Cullen PJ, Sprague GF (2000) Glucose depletion causes haploid invasive growth in yeast. Proceedings of the National Academy of Sciences of the United States of America 97: 13619–13624. doi:10.1073/pnas.240345197.

178. Neves AR, Ventura R, Mansour N, Shearman C, Gasson MJ, et al. (2002) Is the glycolytic flux in Lactococcus lactis primarily controlled by the redox charge? Kinetics of NAD(+) and NADH pools determined in vivo by 13C NMR. The Journal of biological chemistry 277: 28088–28098. doi:10.1074/jbc.M202573200.

179. Greenall A, Lei G, Swan DC, James K, Wang L, et al. (2008) A genome wide analysis of the response to uncapped telomeres in budding yeast reveals a novel role for the NAD+ biosynthetic gene BNA2 in chromosome end protection. Genome biology 9: R146. doi:10.1186/gb-2008-9-10-r146.

180. Vance W, Arkin A, Ross J (2002) Determination of causal connectivities of species in reaction networks. Proceedings of the National Academy of Sciences of the United States of America 99: 5816–5821. doi:10.1073/pnas.022049699.

181. Torralba AS, Yu K, Shen P, Oefner PJ, Ross J (2003) Experimental test of a method for determining causal connectivities of species in reactions. Proceedings of the

National Academy of Sciences of the United States of America 100: 1494–1498. doi:10.1073/pnas.262790699.

182. Fromentin J, Eveillard D, Roux O (2010) Hybrid modeling of biological networks: mixing temporal and qualitative biological properties. BMC systems biology 4: 79. doi:10.1186/1752-0509-4-79.

183. Srividhya J, Mourão MA, Crampin EJ, Schnell S (2010) Enzyme catalyzed reactions: from experiment to computational mechanism reconstruction. Computational biology and chemistry 34: 11–18. doi:10.1016/j.compbiolchem.2009.10.007.

184. Compendium of chemical terminology. Version 2.3.2 (2012). International Union of Pure and Applied Chemistry. p.

185. Chang R (2005) Physical chemistry for the biosciences. Sansalito, CA: University Science. p.

186. UC Davis ChemWiki, the Dynamic Textbook Project. (2012). Available:http://chemwiki.ucdavis.edu. Accessed 2 September 2012.

187. Kerr JA (1987) CRC Handbook of bimolecular and termolelucar gas reactions. Boca Raton, Florida: CRC Press. p.

188. Lawson CL, Hanson RJ (1974) Solving Least Squares Problems. Prentice-Hall. p. 161.

189. Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological) 58: 267–288.

190. Garcia D (2010) Robust smoothing of gridded data in one and higher dimensions with missing values. Computational Statistics & Data Analysis 54: 1167–1178. doi:10.1016/j.csda.2009.09.020.

191. Fletcher R (1971) A modified Marquardt subroutine for non-linear Least Squares. Report AERE-R 6799.

192. Moré JJ (1978) The Levenberg-Marquardt algorithm: implementation and theory. NUMERICAL AND STATISTICS Lecture notes in mathematics 630: 105–116.

193. Prigogine I (1967) On Symmetry-Breaking Instabilities in Dissipative Systems. The Journal of Chemical Physics 46: 3542. doi:10.1063/1.1841255.

194. Prigogine I (1968) Symmetry Breaking Instabilities in Dissipative Systems. II. The Journal of Chemical Physics 48: 1695. doi:10.1063/1.1668896.

195. Field RJ, Noyes RM (1974) Oscillations in chemical systems. IV. Limit cycle behavior in a model of a real chemical reaction. The Journal of Chemical Physics 60: 1877–1884. doi:10.1063/1.1681288.

196. Dyson R, Maeder M, Puxty G, Neuhold Y-M (2002) Simulation of complex chemical kinetics. Inorganic Reaction Mechanisms 5: 39–46.

197. Olbregts J (1985) Termolecular reaction of nitrogen monoxide and oxygen: A still unsolved problem. International Journal of Chemical Kinetics 17.

198. Glandsdorff P, Prigogine I (1971) Thermodynamic theory of structure, stability and fluctuations. New York: Wiley. p.

199. Tomioka R, Suzuki T, Sugiyama M (2011) Super-linear convergence of dual augmented-lagrangian algorithm for sparsity regularized estimation. Journal of Machine Learning Research 12: 1537–1586.

200. Segel IH (1993) Enzyme kinetics: behavior and analysis of rapid equilibrium and steady state... John Wiley & Sons, New York: 1123.

201. Bicknell R, Waley SG (1985) Single-turnover and steady-state kinetics of hydrolysis of cephalosporins by beta-lactamase I from Bacillus cereus. The Biochemical journal 231: 83–88.

202. Hoefnagel MHN, van der Burgt A, Martens DE, Hugenholtz J, Snoep JL (2002) Time dependent responses of glycolytic intermediates in a detailed glycolytic model of Lactococcus lactis during glucose run-out experiments. Molecular biology reports 29: 157–161.

203. Hols P, Kleerebezem M, Schanck AN, Ferain T, Hugenholtz J, et al. (1999) Conversion of Lactococcus lactis from homolactic to homoalanine fermentation through metabolic engineering. Nature biotechnology 17: 588–592. doi:10.1038/9902.

204. Ramos A, Neves AR, Ventura R, Maycock C, López P, et al. (2004) Effect of pyruvate kinase overproduction on glucose metabolism of Lactococcus lactis. Microbiology (Reading, England) 150: 1103–1111.

205. Tyson JJ, Light JC (1973) Properties of two-component bimolecular and trimolecular chemical reaction systems. Journal of Chemical Physics 59: 4164–4173.

206. Mitchison JM (1971) The biology of cell cycle. . Cambridge: Cambridge University Press. p.

207. Naes T, Mevik B-H (2001) Understanding the collinearity problem in regression and discriminant analysis. Journal of Chemometrics 15: 413–426. doi:10.1002/cem.676.

208. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S (2002) The protein kinase complement of the human genome. Science (New York, NY) 298: 1912–1934. doi:10.1126/science.1075762.

209. Alonso A, Sasin J, Bottini N, Friedberg I, Friedberg I, et al. (2004) Protein tyrosine phosphatases in the human genome. Cell 117: 699–711. doi:10.1016/j.cell.2004.05.018.

210. Pawson T, Scott JD (2005) Protein phosphorylation in signaling--50 years and counting. Trends in biochemical sciences 30: 286–290. doi:10.1016/j.tibs.2005.04.013.

211. Locasale JW, Wolf-Yadlin A (2009) Maximum entropy reconstructions of dynamic signaling networks from quantitative proteomics data. PloS one 4: e6522. doi:10.1371/journal.pone.0006522.

212. Pearson G, Robinson F, Beers Gibson T, Xu BE, Karandikar M, et al. (2001) Mitogen-activated protein (MAP) kinase pathways: regulation and physiological functions. Endocrine reviews 22: 153–183.

213. Hornberg JJ, Bruggeman FJ, Binder B, Geest CR, De Vaate AJMB, et al. (2005) Principles behind the multifarious control of signal transduction. ERK phosphorylation and kinase/phosphatase control. The FEBS journal 272: 244–258.

214. Olivier BG, Snoep JL (2004) Web-based kinetic modelling using JWS Online. Bioinformatics (Oxford, England) 20: 2143–2144. doi:10.1093/bioinformatics/bth200.

215. Arteaga CL (2003) ErbB-targeted therapeutic approaches in human cancer. Experimental cell research 284: 122–130.

216. Downward J (2003) Targeting RAS signalling pathways in cancer therapy. Nature Reviews Cancer 3: 11–22. doi:10.1038/nrc969.

217. Gschwind A, Fischer OM, Ullrich A (2004) The discovery of receptor tyrosine kinases: targets for cancer therapy. Nature Reviews Cancer 4: 361–370.

218. Kolch W (2002) Ras/Raf signalling and emerging pharmacotherapeutic targets. Expert Opinion On Pharmacotherapy 3: 709–718.

219. Wolf J, Dronov S, Tobin F, Goryanin I (2007) The impact of the regulatory design on the response of epidermal growth factor receptor-mediated signal transduction towards oncogenic mutations. The FEBS journal 274: 5505–5517.

220. Kholodenko BN, Demin OV, Moehren G, Hoek JB (1999) Quantification of short term signaling by the epidermal growth factor receptor. The Journal of Biological Chemistry 274: 30169–30181.

221. Markevich NI, Moehren G, Demin OV, Kiyatkin A, Hoek JB, et al. (2004) Signal processing at the Ras circuit: what shapes Ras activation patterns? Systems Biology 1: 104–113.

222. Schoeberl B, Eichler-Jonsson C, Gilles ED, Müller G (2002) Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. Nature Biotechnology 20: 370–375.

223. Quian Quiroga R, Kreuz T, Grassberger P (2002) Event synchronization: a simple and fast method to measure synchronicity and time delay patterns. Physical Review E - Statistical, Nonlinear and Soft Matter Physics 66: 041904.

224. Wessel N, Suhrbier A, Riedl M, Marwan N, Malberg H, et al. (2009) Detection of time-delayed interactions in biosignals using symbolic coupling traces. Europhysics Letters 87: 10004. doi:10.1209/0295-5075/87/10004.