UNIVERSITY OF CALIFORNIA, SAN DIEGO

# Modeling Nonlinear Behavior of Dynamic Biological Systems

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

in

Electrical Engineering (Signal and Image Processing)

by

Maryam Masnadi-Shirazi

Committee in charge:

Professor Shankar Subramaniam, Chair
Professor Pamela Cosman, Co-chair
Professor Todd P. Coleman
Professor Sadik Esener
Professor Kenneth Kreutz-Delgado

2018

The Dissertation of Maryam Masnadi-Shirazi is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

Co-chair

_____

Chair

University of California, San Diego

2018

# DEDICATION

In the Name of God, the Most Gracious, the Most Merciful.

I dedicate this dissertation to my mother, Mehri Daneshvar, my father, Dr. Mohammad Ali Masnadi-Shirazi, and my dearest husband, Mahyar Madjlessi-Kupai, for their endless support and love.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

# LIST OF SUPPLEMENTAL FILES

Masnadishirazi_G1_Phase_Interactions.xlsx

Masnadishirazi_S_Phase_Interactions.xlsx

Masnadishirazi_G2M_Phase_Interactions.xlsx

ACKNOWLEDGEMENTS

It is said that one thanks God by showing appreciation towards His subjects. And so, I would like to acknowledge many people who have helped me throughout my education and doctoral work.

First of all, I would like to express my most sincere gratitude to my supervisor, Professor Shankar Subramaniam. Besides providing exceptional scientific guidance and insight, which is undoubtedly reflected in this dissertation, I should also thank him for giving me the opportunity to explore new ideas and new areas while being patient towards my trial-and-errors and further showing to me what it means to be a true scientific professional. I want to thank my ECE PhD advisor, Professor Pamela Cosman, for facilitating and supporting my goal of pursuing interdisciplinary research between the ECE and Bioengineering departments. I am also thankful for having an outstanding doctoral committee, and wish to thank Professor Kenneth Kreutz-Delgado, Professor Sadik Esener, and Professor Todd P. Coleman for their valuable feedback, accessibility and informative courses that have formed the foundation of my research.

I would like to thank my colleague at the Subramaniam lab, Dr. Mano R. Maurya for his collaboration and help. Also, I would like to thank the Subramaniam lab members, Dr. Merril Gersten, Dr. Andrew Caldwell, Dr. Shakti Gupta, Dr. Kavitha Mukund, Dr. Pam Bhattacharya, Shamim Mollah, Julian Nitka, and Carol Kling for their support and assistance throughout the years.

I would like to thank my friends who have helped me in so many ways both in the good times and the bad: Dr. Mahmoud Tarokh and Mrs. Ency Tarokh, Dr. Charles Kunkel and Mrs. Seddigheh Kunkel, Dr. Ebraheem Fontaine and Danielle Fontaine, Dr. Khosrow Behbehani and

his family, Dr. Koohyar Minoo and his family, Dr. Mohammadreza Keshtkaran, Dr. Alireza Dehghani, Dr. Zeinab Taghavi, Justin and Fatemeh Mashouf, Zeinab Mofrad, Samaneh Keshavarz and her family, Bahareh Marzban and her family.

I would like to thank my piano teacher Ms. Nadia Dabbagh who mentored me since I was just a child; I will forever carry her teachings with me. I would like to thank my extended family, grandparents, my aunts, uncles and cousins for supporting me and being there for me through difficult times.

I want to thank my mother Mehri Daneshvar, and my father Dr. Mohammad Ali Masnadi-Shirazi for simply everything. I know I can never be grateful enough for the things my parents have done for me, I just hope that they are pleased with me and know that without them, I wouldn't be where I am today. I want to thank my brothers, Drs. Hamed and Alireza Masnadi-Shirazi whom I've always looked up to, for encouraging me and making me believe in myself. I would like to thank my sisters in law, Hanieh Hadaegh, Mina Hosseini, and Mahtab Madjlessi-Kupai for completing my family and being the sisters I never had. I reserve a special thanks for my mother and father in law, Farkhondeh and Mahmood Madjlessi-Kupai, for their endless kindness and grace. And last but not least, I want to thank my best friend, soulmate, and husband Mahyar Madjlessi-Kupai from the bottom of my heart, for his unconditional support, love and patience, and for bringing so much peace and happiness into my life.

Chapter II, in full, is a reprint of the material as it appears in Time-Varying Causal Inference from Phosphoproteomic Measurements in Macrophage Cells, 2014. Masnadi-Shirazi, Maryam; Maurya, Mano R.; Subramaniam, Shankar., IEEE Transactions on Biomedical Circuits and Systems, Volume 8, 2014. The dissertation author was a primary investigator and author of this paper.

Chapter III, in full, has been submitted for publication of the material as it may appear in Time Varying Causal Network Reconstruction of a Mouse Cell Cycle, 2018, Masnadi-Shirazi, Maryam; Maurya, Mano R.; Pao, Gerald; Ke, Eugene; Verma, Inder; Subramaniam, Shankar., PLOS Computational Biology, 2018. The dissertation author was a primary investigator and author of this paper.

Chapter IV, is currently being prepared for submission for publication of the material. Masnadi-Shirazi, Maryam; Subramaniam, Shankar. The dissertation author was the primary investigator and author of this material.

# VITA

| | |
|---|---|
| 2006-2010 | Bachelor of Science, Electrical Engineering, Shiraz University |
| 2010-2012 | Master of Science, Electrical Engineering (Signal and Image Processing), University of California, San Diego |
| 2012-2018 | Research Assistant, Electrical Engineering (Signal and Image Processing), University of California, San Diego |
| 2018 | Doctor of Philosophy, Electrical Engineering (Signal and Image Processing), University of California, San Diego |

# PUBLICATIONS

"Time-Varying Causal Inference from Phosphoproteomic Measurements in Macrophage Cells" IEEE transactions on Biomedical Circuits and Systems, vol. **8**, pp 74-86, February 2014.

"Time Varying Causal Network Reconstruction of a Mouse Cell Cycle", submitted to PLOS Computational Biology.

"Multiview Radial Basis Function Network: A New Approach on Nonparametric Forecasting of Chaotic Dynamic Systems", in preparation for submission.

ABSTRACT OF THE DISSERTAION

Modeling Nonlinear Behavior of Dynamic Biological Systems

by

Maryam Masnadi-Shirazi

Doctor of Philosophy in Electrical Engineering
(Signal and Image Processing)

University of California, San Diego, 2018

Professor Shankar Subramaniam, Chair
Professor Pamela Cosman, Co-Chair

With the availability of large-scale data acquired through high-throughput technologies, computational systems biology has made substantial progress towards partially modeling biological systems. In this dissertation we intend to focus on deciphering the dynamics of such systems through data-driven analysis of multivariate time-course data. We develop integrative frameworks to study the following problems: 1) time-varying causal inference when the number of samples exceeds the number of variables (overdetermined case), 2) dynamic causal network reconstruction when the number of variables exceeds the data samples (underdetermined case), 3) forecasting the dynamic behavior of complex chaotic systems from short and noisy time-series

data. In the first problem we utilize the notion of Granger causality identified by a first-order vector autoregressive (VAR) model on phosphoproteomic measurements to unravel the crosstalk between various phosphoproteins in three distinct time intervals. In problem 2 we use a non-parametric change point detection (CPD) algorithm on transcriptional time series data from a mouse cell cycle to estimate temporal patterns that can be associated with different phases of the cell cycle. The second problem becomes more complex as the number of variables exceeds the number of time-series data and we use a higher order VAR model to estimate causal interactions among cell cycle genes. To solve this ill-posed problem we use Least Absolute Shrinkage and Selection Operator (LASSO) and select the regularization parameters through Estimation Stability with Cross Validation (ES-CV) leading to more biologically meaningful results. LASSO + ES-CV is applied to temporal intervals associated with the G1, S and G2/M phases of the cell cycle to estimate phase-specific intracellular interactions. In problem 3, we develop a nonparametric forecasting algorithm for chaotic dynamic systems, Multiview Radial Basis Function Network (MV-RBFN) that outperforms a model-free approach, Multiview Embedding (MVE). In this algorithm, the forecast skill of all possible manifolds (views) reconstructed from a combination of variables and their time lags is assessed and ranked from best to worst. MV-RBFN uses the top $k$ views as the inputs of a neural network to approximate a nonlinear function $f(.)$ that maps the past events of a dynamic system as the input, to future values as the output.

# Chapter I

# Introduction

## I.A Modeling the Dynamics of Biological and Ecological Systems from Time Series Data

In recent years, high-throughput technologies such as nextgen sequencing, DNA microarray expression profiling, and high-content imaging have made it possible to make concurrent quantitative measurement of multiple cellular components including mRNA levels, protein phosphorylation and metabolites. The application of mathematical and statistical approaches to such data has been used widely to understand the relationship between different components in the cell to partially reconstruct data-driven networks. Conventional methods of network reconstruction such as correlation based methods (*1, 2*), principal component regression (PCR) (*3*), and partial least squares (PLS) (*4*) offer a static characterization of network topologies, devoid of any temporality which is an ingrained feature of biological systems. Boolean network (BN) and dynamic Bayesian learning provide a temporally evolving picture of the network but either require discretization of data values and thus oversimplification of the network topology (*5*), or perform poorly on high dimensional data (*6*).

In the case of forecasting the behavior of complex natural systems, many studies assume linearity and use generalized linear models while such systems exhibit nonlinear dynamics with time lags, reciprocal feedback loops and unpredictable surprises (*7*). Equation based methods such as differential equations may also be used to model the dynamics of chaotic systems but require prior knowledge about the actual interaction of system components (*8*). Even if the underlying structure is known, dimensionality poses a challenge on accurate estimation of model parameters. An alternative equation-free method suitable for chaotic behavior is state space reconstruction (SSR) which provides substantial flexibility in the nonlinearity of the system (*9*).

In this dissertation, we look into the following problems:

1. Estimating time-varying causal network from phosphoproteomic measurements in macrophage cells for a set of 17 phosphoproteins.

2. Dynamic network reconstruction from RNA-sequencing data in mouse embryonic fibroblast primary cells for a set of 63 cell cycle genes.

3. Prediction of future behavior of chaotic dynamic ecological systems (using simulated and real data).

In both problems 1 and 2, we assume that the corresponding biological systems (macrophage cells and embryonic fibroblast cells) have a stochastic behavior and assume linearity of the model structure and use the notion of Granger causality identified by vector autoregression (VAR) for causal inference (*10*). The first scenario is a simpler problem where number of data samples exceed the number of variables (overdetermined problem) and we use first-order vector autoregression (VAR) to model the system's dynamics. The second problem becomes more complex as we consider higher-order VAR to model the cell cycle, resulting in an underdetermined problem that cannot be solved uniquely. In problems 1 and 2, we develop integrative frameworks to investigate the temporal behavior of the data-driven reconstructed networks. In the third scenario, we no longer assume linearity of the model structure where we develop a nonparametric forecasting algorithm that takes advantage of the dimensionality of complex chaotic systems in nature.

## I.B Contribution of the Dissertation

The contribution of this dissertation has three main components summarized in the following subsections.

### I.B.1 Time-varying Causal Inference from Phosphoproteomic Measurements in Macrophage Cells

Protein phosphorylation is a key reversible modification that acts as a switch to turn "on" or "off" a protein activity or cellular pathway (*11*). Activation of proteins through phosphorylation serves as the flux in the signaling pathways; errors in transferring cellular information can alter normal function and may lead to diseases such as chronic anti-inflammatory diseases, autoimmunity and cancer. Since biological systems evolve through time, it is important to study the dynamic behavior of the topology of the signaling pathways and networks. Thus, we allow the network topology (the set of connections or edges presented in the network) to evolve with time. Our objective in this study is to derive a time-varying model for the phosphoproteomic network to understand the dynamics of signaling pathways using the notion of Granger causality. We have applied the notion of Granger causality and statistical hypothesis testing to estimate causal relationships between different phosphoproteins using time-series data.

According to Granger's definition of causality, it is said that signal $X(t)$ causes signal $Y(t)$, if the future values of $Y(t)$ can be better predicted using the past values of $X(t)$ and $Y(t)$ than only using the past of itself (*10*). Utilizing the notion of Granger causality, we apply first-order vector autoregression (VAR) to infer causal relationships among phosphoproteins by analyzing the time-varying fold changes of phosphoprotein n ns in response to single and double ligand stimuli. The quantitative levels of phosphoproteins are measured through western blot experiments in RAW 264.7 macrophage cells.

The availability of multiple single and double experiments in the phosphoprotein time-series provides more data samples and makes the VAR model an overdetermined problem that can be solved uniquely via Least Squares (LS) estimation. We further test the significance of predicted LS coefficients by performing a two-tailed t-test to infer statistically significant causal connections. Moreover, by partitioning the time-series data into three segments, we investigate the evolution of the underlying topology of the estimated phosphoprotein network in three distinct time intervals.

### I.B.2 Time-varying Causal Network Reconstruction of a Mouse Cell Cycle

The progression of a eukaryotic cell cycle is governed by a complex, dynamical network of molecular interactions that regulate a series of directional and irreversible events such as cell growth, DNA replication, mitosis, and cell division. The biochemical pathways controlling the order and timing of cell cycle phases, called cell cycle checkpoints, play an essential role in maintaining genomic stability of the cell. Dysregulation of these checkpoints can alter the ability of the cell to undergo cell-cycle arrest in response to DNA damage and may lead to cancer. Significant progress has been made in identifying molecular players and pathways involved in cell cycle mechanisms through extensive investigations on model systems like yeast. Protein assays, transcriptional studies, fluorescent imaging, and protein interaction mapping have all contributed to our current understanding of the cell cycle. From these studies and other phenotypic assays, molecular players engaged in distinct phases of the cell cycle, namely, G1, S, G2, and M phases, have been identified, resulting in a static pathway map of the cell cycle (*12*). These maps lack dynamical information, owing to the absence of systematic time series measurements. Fine-grained time series measurements of a mammalian cell cycle, can enrich the understanding of dynamical networks through which the temporal relationships between molecular players and

modules can be inferred, and further provide insights into mechanistic causality. In this study, we present a systematic fine-grained RNA sequencing study of the transcriptional profiles during a mammalian cell cycle. Although these measurements are at the transcript level, we anticipate that given the strong transcriptional mechanisms that are concomitant with the cell cycle, these data have the potential to provide detailed dynamical mechanisms of the cell cycle. While there has been several attempts at identifying different regimes in long time-series, mainly in the signal processing community (*13-15*), they have not been used to further develop evolving dynamical models and networks for biological systems.

We have developed a framework to investigate the temporal changes in the cell cycle network using RNA-seq time series data from Mouse Embryonic Fibroblast (MEF) primary cells. We use a non-parametric change point detection (CPD) algorithm (*16*) based on Singular Spectrum Analysis (SSA) (*17*) to infer the mechanistic changes in the time-course data for a set of 63 cell cycle genes to estimate cell cycle phases. We also use the notion of Granger causality implemented through vector autoregressive (VAR) model (*18*) to predict the future expression levels of each gene as a function of the past expression levels of other genes yielding directionality of gene regulation among the 63 cell cycle genes. Furthermore, we utilize the concept of Minimum Description Length (MDL) (*19*) to use past expression levels of genes, up to 9 time lags (equivalent to 4.5 hours), to determine the minimum data information from past events required for a robust prediction of values at the current time.

Considering the fact that we use a higher order VAR model to predict causality, the linear inverse problem becomes an undetermined problem that cannot be solved uniquely. However, if the solution is sufficiently sparse, it is actually possible to recover the solution by solving an $\ell_1$-norm regularization problem which is strictly related to the Least Absolute Shrinkage and

Selection Operator (LASSO) problem (*20*). The regularization parameter in LASSO sets a trade-off between the fit error and the sparsity of the solution. The conventional methods in selecting the regularization parameters include Akaike's information criterion (AIC) (*21*) and Bayesian information criterion (BIC) (*22*). These criteria can be easily computed but depend on model assumptions and even if the model assumptions are met, they may not be valid in the finite sample cases. The regularization parameter is often selected through the model-free Cross-validation (CV) approach (*23, 24*). When sample size is large, CV leads to estimators with good predictive performance. However, when sample size is small, CV does not yield an interpretable model since LASS+CV is unstable and not reliable for scientific interpretations (*25*). In this study we observed that using the Estimation Stability with Cross Validation (ES-CV) criterion (*26*) leads to more meaningful results that make biological sense. Estimation stability (ES) is based on the idea that the solution is not meaningful if it varies considerably from sample to sample.

This computational scheme enables us to (i) estimate the timing of cell cycle phases, (ii) infer the duration of the G1, S and G2/M phases of the MEF cell cycle to be 14.5, 10 and 4 hours, respectively, (iii) reconstruct three successive directed graphs representing the key regulatory mechanisms among the 63 cell cycle genes in the G1, S and G2/M phases of the cell cycle, (iv) infer the temporal impact that biological processes have on one another, as well as the dynamic changes in temporal dependencies as the cell evolves through successive phases, and (v) reflect the chronological order of regulatory events that are crucial to cell cycle control. The main power of our work is its ability to capture key pathways and important causal interactions over time, providing a broad picture of the dynamics of a cell cycle regulatory network. We validate the reliability of our time-varying network for cell cycle progression by comparing the interactions detected in our results to the well-known regulatory pathways in the literature as well as estimating

temporal interdependences (time-delay) between important biological processes as the cell evolves through successive phases of the cell cycle.

### I.B.3 Multiview Radial Basis Function Network: A New Approach on Nonparametric Forecasting of Chaotic Dynamic Systems

In recent years, the availability of large time-course datasets in multiple disciplines, including biology, ecology and finance has brought forth the problem of handling such data for scientific analysis (*27-29*). In many studies, generalized linear models and vector autoregressive models are used for structural estimation and inference, where such systems exhibit nonlinear dynamics with time lags, reciprocal feedback loops and unpredictable surprises (*7, 30*). On the other hand, equation-based models such as difference and differential equations may be used to analyze the evolution of a dynamic system, but often require some degree of prior knowledge about the nature of interactions among various system components (*8*), or even if the model structure is known, dimensionality poses a challenge on accurate parameter estimation of variables (*31*). Furthermore, prior work has established that ecological and biological models are often ineffective in predicting the future due to the highly nonlinear nature of component interactions (*32, 33*).

An alternative equation-free approach suitable for non-equilibrium dynamics (including chaos) and nonlinearity is state space reconstruction (SSR) which is a model-free approach in the sense that there is no analytic formula assumption thus allowing substantial flexibility in the nonlinearity of the system (*9, 34*). SSR uses lagged coordinate embeddings to reconstruct attractors that map the time-series evolution from time domain into state space trajectories. In a notable theorem, Takens proved that the overall behavior of a chaotic dynamic system can be reconstructed from lags of a single variable (*35*). Later Takens' theorem was generalized and it was demonstrated

that the information from a combination of multiple time-series (and their lags) can be used in an attractor reconstruction to provide a more mechanistic model (*36, 37*). Nonetheless, since attractor reconstruction relies only on experimental data, the limitations of short or noisy time-series restricts the ability to infer system dynamics as a whole. Namely, SSR from short time series provide a scarce view of a system's mechanism, diminishing reliability of inferences. In addition, when time-series data is corrupted with observational noise, data may become meaningless and irrelevant in providing useful information for predictability. Ye *et al.* (2016) introduced an analytical approach, multiview embedding (MVE), which harnesses the complexity of short and noisy ecological time series as a way to improve forecasting (*38*). MVE is a method based on nearest neighbors that looks into the predictability of all possible manifold reconstructions using the method of simplex projection (*34*). In this work, we treat prediction of the dynamical system as an inverse problem that involves interpolation and approximating an unknown function from time series data. Instead of relying on single nearest neighbors of the top attractor reconstructions as carried out in MVE, here we introduce multiview radial basis function network (MV-RBFN) autoregressive model that calculates a distance-weighted average of all points in the top manifold reconstructions through a nonlinear kernel estimation method. Similar to MVE, attractors from combinations of variables and their lags are reconstructed. Each manifold (view) comprises information that is particular to that embedding. By ranking the reconstructed manifolds according to their forecast skill (prediction errors), and merging the top views and the information contained in them, MV-RBFN is capable of recovering the dynamics of the system in a manner that outperforms MVE and nonlinear univariate and multivariate autoregressive models.

We show that our approach, multiview radial basis function network (MV-RBFN) provides a better forecast performance than that obtained using a model-free approach, multiview

embedding (MVE), owing to the universal approximation property of radial basis function networks. We demonstrate this for simulated ecosystems and a long term mesocosm experiment on a multi-species plankton community obtained from the Baltic Sea. By taking advantage of dimensionality, we show that MV-RBFN overcomes the shortcomings of noisy and short time-series.

## I.C Organization of the dissertation

The rest of the dissertation is organized as follows. In Chapter II, we look into estimating time-varying causal network from phosphoproteiomic measurements in macrophage cells for a set of 17 phosphoproteins. In Chapter III, we develop an integrative framework that deals with an undetermined inverse problem for dynamic network reconstruction from RNA-seq data in mouse embryonic fibroblast primary cells for a set of 63 cell cycle genes. In Chapter IV, we develop a nonparametric forecasting algorithm that predicts future behavior of chaotic dynamic ecological systems (using simulated and real data). Finally in Chapter V, the conclusions of the dissertation are presented.

Chapter II

# Time-varying Causal Inference from Phosphoproteomic Measurements in Macrophage Cells

## II.A Introduction

The understanding of cellular function at the molecular level involves the study of intracellular signaling, metabolic pathways and gene regulatory networks, through "omics" measurements on biological systems. Protein phosphorylation is one of the main steps in intracellular signaling from the activated proteins located at the plasma membrane to the cytosolic space and nucleus. Phosphorylation is one of the most studied post-translational modification of proteins since is it vital for many protein interactions that regulate cellular processes such as cell growth, cell differentiation and development to cell cycle control and metabolism (*39*). Phosphorylation is a key reversible modification with the combined involvement of protein kinases and phosphatases to activate and deactivate proteins (*11*). Phosphorylation mainly occurs on serine, threonine and tyrosine residues that can regulate enzymatic activity, subcellular localization, complex formation and degradation of proteins. Activation of proteins through phosphorylation serves as the flux in the signaling pathways. Several signaling pathways such as the nuclear factor kappa B (NF-κB), mitogen-activated protein kinases (MAPK), and signal transducer and activator of transcription (STAT) play essential roles in transmitting signals that trigger the release of cytokines, which are central to the processes of inflammation and modulation of immune function (*40*). The signaling pathways act as modules to regulate the transcription and release of various cytokines, some of which are involved in the pathogenesis of many diseases, e.g., chronic inflammatory diseases, autoimmunity and cancer. Thus, reconstructing protein networks from "omics" measurements can help us not only understand and model cellular signaling pathways but also assist in uncovering the mechanisms of disease progression. Since knowledge of protein-protein interaction is sparse, it is difficult to simultaneously analyze the dynamics of various proteins *in vitro* or *in vivo*. High-throughput technologies, such as nextgen

sequencing, DNA microarray expression profiling, phosphoproteomics, metabolomics and high-content imaging, have made it possible to make concurrent quantitative measurements of various components of the cell, including mRNA levels, protein phosphorylation and metabolites, enabling the reconstruction of large-scale cellular networks. Complexities such as feedback and feed-forward loops and the cross-talk between different signaling pathways have hindered the problem of developing reliable mathematical approaches within an integrative framework, taking into account the dynamics of signaling networks (*40*).

During the last decade, the application of mathematical and statistical approaches to high-throughput biological data has been used extensively to decipher the relationship between different components in the cell to partially reconstruct intracellular networks. With the availability of large-scale omics data, computational systems biology has made substantial progress towards modeling and reconstruction of data-driven networks using (1) input/output-based models such as Partial Least Squares (PLS) (*4*) and Principal Component Regression (PCR) (*41*), (2) probabilistic graphical models such as Bayesian network-based models (*42-44*), probabilistic Boolean network models (*45, 46*), and (3) information theory-based methods such as integrated correlation and transfer entropy based approach (*47*) and C3NET (*48, 49*). Other approaches using differential equations (*50*), structural equation methods (*51*) and state-space models (*52*) have also been proposed during the past few years.

Biological systems evolve through time and it is important to study the dynamic behavior of the topology of the signaling pathways/networks themselves (*52*). Thus, we allow the network topology (the set of connections/edges present in the network) to evolve with time. Our objective in this study is to derive a time-varying model for the phosphoprotein network to understand the dynamics of signaling pathways using the notion of Granger causality. Causality can be

determined by prior biological information. However, in many cases, no "*a priori"* knowledge is available to provide causal relationships in network reconstruction. Furthermore, it is appealing to discover new causal relationships, rather than already known ones. In the present work, we have applied the notion of Granger causality and statistical hypothesis testing to estimate causal relationships between different phosphoproteins using time-series data. According to Granger's definition of causality, it is said that signal X(t) causes signal Y(t), if future values of Y(t) can be better predicted using the past values of X(t) and Y(t) than only using the past of itself (*10*).

Due to the fact that intracellular networks are not static, we use time series data in order to determine these dynamic changes in the network topology. In the present work, we use a vector autoregressive (VAR) model to infer relationships of Granger causality among phosphoproteins by analyzing the time-varying fold changes of phosphoproteins in response to single and double ligand stimuli. The quantitative levels of phosphoproteins were measured through western blot experiments by the Alliance for cellular Signaling (AfCS) (*53*) in RAW 264.7 macrophage cells. We infer the topology of the phosphoprotein networks in three distinct time intervals.

## II.B Approach

### II.B.1 Granger Causality

Granger causality was first introduced by the Noble prize- winning economist, Clive Granger, and has proven useful for analyzing the relationships and influences among macroeconomic time series (e.g. income, exchange rate, etc.) (*10*). We note that Granger causality is not meant to be equivalent to the true causality, but is intended to provide useful information regarding causation and the direction of information flow. Formally, a time series $x_t$ is said to Granger-cause a time series $y_t$, if the future value of $y_t$ can be predicted given the past values of $y_t$ and $x_t$, $(y_{t-1}, y_{t-2}, \dots, x_{t-1}, x_{t-2}, \dots)$, better than predicting the future of $y_t$ given only the past

values of $y_t$ ($y_{t-1}, y_{t-2}, ...$). Commonly, Granger causality is identified by VAR models (*54*). A VAR model of p-order and k-dimensional time series is given by:

$$y_t = v + A_1 y_{t-1} + A_2 y_{t-2} + A_3 y_{t-3} + \ . \ . \ . \ + A_p y_{t-p} + \ \varepsilon_t \quad \text{(II.1)}$$

where $y_t = (y_{1t}, y_{2t}, \dots, y_{kt})'$ is a $(k \times 1)$ random vector, $y_{it}$ is the measurement at time t of the $i^{th}$ random variable, $A_l$ is a $(k \times k)$ autoregressive coefficient matrix, $\boldsymbol{v}$ is a $(k \times 1)$ vector of intercepts and $\varepsilon_t = (\varepsilon_{1t}, \varepsilon_{2t}, \dots, \varepsilon_{kt})'$ is a $k$-dimensional error vector of random variables with zero mean and covariance matrix $\sum$.

The optimal order of the VAR model can be found through approaches such as Minimum Description Length (*19*) which requires many samples in time. In the present work, since there are only three original samples in time, we consider the following first order VAR model:

$$y_t = v + A_1 y_{t-1} + \ \varepsilon_t \quad \text{(II.2)}$$

VAR allows identification of Granger causality for linear relationships. In order to find causal relationships, we analyze the elements of matrix $A_1$. An important outcome of this approach is that the series $y_{jt}$ causes of $y_{it}$ if and only if the $ij^{th}$ entry of matrix $A_1$ is statistically significant. Therefore, it is sufficient to estimate the autoregressive coefficient matrix of the VAR model in order to identify the direction of Granger causality.

This approach can be applied to the analysis of phosphoprotein time-course data to interpret functional connectivity between phosphoproteins to reconstruct their underlying network by testing the statistical significance of the estimated components of $A_1$. Considering the time series $(y_1, \dots, y_T)$ for each of the $k$ variables, the first-order VAR model in (2) can be written in the following matrix form (*55*):

$$Y = XB + \varepsilon \quad \text{(II.3)}$$

where $Y = (y_1, \dots, y_T)'$ is a $(T \times k)$ matrix whose columns are time series for each of the $k$ random variables with sample size $T$, $B = (v, A_1)'$ is a $((k+1) \times k)$ matrix, $X = (X_0, \dots, X_{T-1})'$ is a $(T \times (k+1))$ matrix with $X_t = (1;\ y_t)$, and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_T)'$ is a $(T \times k)$ matrix. For each of the $k$ columns of matrices $Y, B$, and $\varepsilon$, we can write the following linear regression model

$$Y_i = XB_i + \varepsilon_i;\ i = 1, \dots, k \tag{II.4}$$

where vector $Y_i$ represents the $i^{th}$ column of matrix $Y$, vector $B_i$ is the $i^{th}$ column of matrix $B$ and vector $\mathcal{E}_i$ is the $i^{th}$ column of matrix $\mathcal{E}$. In this linear model, we seek to estimate the unknown coefficients in matrix $B$. We can use least squares (LS) estimation method in order to compute the unknown parameters/coefficients. Therefore, each column of matrix $B$ is estimated through the LS estimation shown below:

$$\widehat{B_i} = (X'X)^{-1}X'Y_i;\ i = 1, \dots, k \tag{II.5}$$

After estimating the coefficient vectors for each of the outputs, they can be concatenated to construct the estimated matrix $\hat{B}$, and therefore, the autoregressive coefficient matrix $A_1$ can be computed. The proposed VAR model analyzes causality between different variables in terms of how the future of a variable can be predicted using the past values of itself and other variables. According to this model, as stated earlier, variable $j$ is said to Granger-cause variable $i$, if the $ij^{th}$ entry of matrix $A_1$ is nonzero. However, the least squares criteria favors solutions with many nonzero entries, which is contrary to the goal of finding purely zero entries to identify whether or not causations between pairs of variables exist. Hence, we need to apply statistical significance test to examine the significance of the estimated parameters. We know that LS estimation minimizes the root mean squared error (RMSE), and by computing the RMSE, we can perform a two-tailed t-test on the coefficients. The RMSE is computed as follows:

$$RMSE_{LS} = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(Y_i - \widehat{Y}_i)^2} = std\left(Y - \hat{Y}\right) \times \sqrt{\frac{T-1}{T}} \quad \text{(II.6)}$$

where $\widehat{Y}_i$ is the estimation of $Y_i$:

$$\widehat{Y}_i = X\hat{B}_i \quad \text{(II.7)}$$

**Significant Connections:** The standard-deviation of the model parameters are estimated as:

$$\sigma_{b,LS} \approx diag((X^T X)^{-1})^{\frac{1}{2}} \times RMSE_{LS} \times \left(\frac{T}{v}\right)^{\frac{1}{2}} ; v = T - k - 1 \quad \text{(II.8)}$$

where $T$ is the length of the time series, $k$ is the number of variables, and $v$ is defined as the degrees of freedom. Then the ratio $r_{ji} = \hat{B}_{ji}/\sigma_{b,LS}$ is computed for the $j^{th}$ entry of the $i^{th}$ column of the estimated matrix $\hat{B}$ and $|r_{ji}|$ is compared against $R = tinv(1 - \alpha/2, v)$, where $tinv(.)$ denotes the inverse of the cumulative t-distribution and $\alpha = 0.01$ (two-tailed) for a confidence interval of 99%. The estimated coefficients are considered statistically significant if their corresponding ratios are greater than $R$ and insignificant otherwise (t-test on the model coefficients). We also computed the p-value, false-discovery rate (FDR) using the Benjamini–Hochberg (BH) method (*56*) for the connections retained. As presented in the Results section, the Benjamini-Hochberg FDR for the connections retained is less than 0.026.

**Performance Metrics:** Type I error, Type II error, and accuracy of the network is computed (*57*) as follows using the False Positives (FP), False Negatives (FN), True Positives (TP) and True Negatives (TN) in the network identified:

$$Type\ I\ Error = \frac{FP}{FP+TN} \quad \text{(II.9)}$$

$$Type\ II\ Error = \frac{FN}{FN+TP} \quad \text{(II.10)}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{(II.11)}$$

### II.B.2 Application of the VAR model to Phosphoproteomic Data

We applied this method to time-course data on the level of phosphorylation of proteins in RAW 264.7 macrophages in response to stimuli, provided by the Alliance for Cellular Signaling (AfCS) (*53*) . This data set consists of fold changes of 21 phosphoproteins at 4 time points; i.e., data at 1,3,10 and 30 minutes, in response to treatments with 22 single ligands and their double ligand combinations measured using the western blot method. The fold changes of the phosphoproteins are determined by dividing the volume of each phosphoprotein band for the ligand-treated samples by the average volume of the corresponding bands for the untreated samples (volume is the sum of the image pixel values within the area of the band). The replicates for the experiments with unique combination of ligand(s) for each phosphoprotein were averaged. Out of 327 unique ligand combinations, the number of combinations with 1, 2, 3, 4 and more than 4 replicates was 68, 68, 123, 37 and 31, respectively. Thus, most ligand combinations have three replicates, hence resulting in only a small bias due to the difference in the number of replicates.

Due to the fact that the time intervals are not equal, we interpolated the data using linear interpolation with steps of one minute. Other interpolation methods (e.g., cubic) may result in large deviations at the intermediate time points, and this may not be close to the real variation of the fold change of the phosphoproteins in the biological system. We excluded the last sample in the original data, since it was taken 20 minutes after the previous one, which is considered to be too large an interval for accurate interpolation. In these experiments, we had missing data for 4 of the 21 phosphoproteins, signal transducer and activator of transcription (STAT) 3, STAT5, c-Jun N-terminal kinases (JNK) long (JNKL) and JNK short (JNKS). Therefore, we excluded these variables from further analysis. We assumed that at a given time, the underlying phosphoprotein network that represents the structure or the topology of the biological system is the same across all

experiments, i.e., the topology of the phosphoprotein network representing the behavior of the biological system remains unchanged regardless of which ligand(s) is stimulating the system. Thus, to deal with the problem of rank deficiency of matrix $X$ in (4), we stacked the data from multiple experiments for both the output data in matrix $Y$ (data related to present) and the input data in matrix $X$ (data related to the past). This ensures that matrix $X$ will have full column rank and there will be a unique solution to the least squares problem. Figure II.1 shows a schematic of how the input and output data from multiple experiments were stacked. Before implementing the VAR model, the data in matrix $X$ was normalized and matrix $Y$ was mean-centered for each variable.

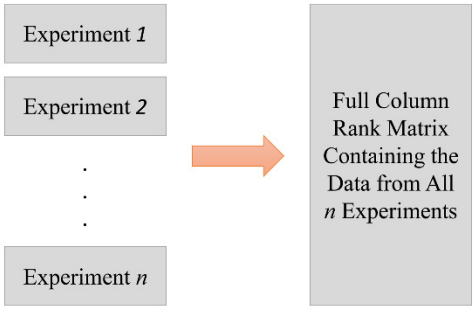



**Figure II.1 Schematic to show the stacking of the data matrices.** Each column corresponds to the time series data of each of the $k$ variables.

In addition to implementing the VAR model, the correlation between the past and present values for each pair of variables was studied and the correlation matrix between the input and output variables was computed. Figure II.2 visualizes the correlation matrix as a heat-map, where the rows and columns of the heat-map are the input (at time $t$-1) and the output variables (at time $t$) for the whole time-series data, respectively.

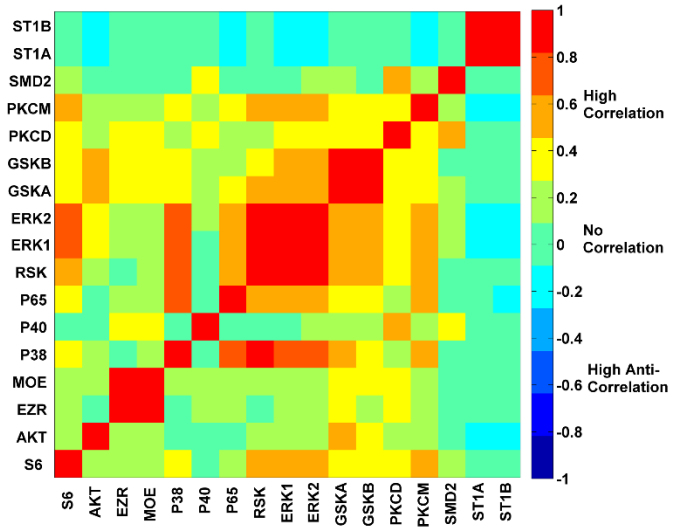



**Figure II.2 Heat-map of the correlation matrix between the input and output variables.** This matrix contains the pairwise correlation coefficient between columns of matrix X and Y for the whole time series [1-10] minutes.

In order to investigate how the underlying topology of the network is changing, we partition the time series for all the variables into three segments and then apply the VAR model for each segment separately. Since we are considering the time-course data for 1 to 10 minutes, and the granularity of the measurements is not fine, three overlapping segments, [1-4], [3-7] and [6-10] minutes were considered using interpolated data. Next, in order to investigate how the causal relationships are evolving with time, we estimate the causality coefficients and perform a statistical significance test (t-test) for each segment separately. We also compute the correlation matrix for each segment independently. It is expected that the results based on the interpolated data in the [3-7] minute interval are more affected by the actual experimental value at 3 minute, whereas those based on [6-10] minute interval are more affected by the actual experimental value at 10 minute. Among the statistically significant causal relationships that were estimated through the VAR model, only those with high correlation coefficients ($\geq 0.4$; p-value is quite significant since the number of rows in the matrices $X$ and $Y$, 2943, is very large) were selected to reconstruct the final network for each time interval. Therefore, the network identified contains likely causal connections

which also exhibit high correlation.

It can be noted that since we are considering three separate time intervals to study the temporal evolution of the network, we expect that the information provided in the time series data may differ from stage to stage. Therefore, a causal relationship A → B that exists at an earlier stage need not exist at the following stage, i.e., the past value of A may no longer contribute to predicting the future value of B at the following stage. Thus, according to Granger's definition of causality, there will be no causal relationship at the following stage. This implies that the weights of edges (resulting in fluxes through connections) change though time. For example, if the weight of a connection decreases and the corresponding p-value becomes more than the threshold of 0.01 (for a confidence interval of 99%), we no longer consider that connection to exist as a strong causal relationship even though we may observe the connection in the underlying network.

## II. C Results and Discussion

### II.C.1 Graphical Network Reconstruction

We have reconstructed the phosphoprotein signaling network that represents the underlying network corresponding to the full time series data shown in Figure II.3. In this network, out of 17 × 17 possible connections, only 35 were significant, many of which have negative coefficients in matrix $A_1$. Connections with negative coefficients are considered as inhibitory relationships shown in Figure II.3. Important inhibitory edges include AKT → GSKα/β (*58-60*), ERK1/2 → RSK (*61, 62*). Different edge-widths are used to indicate edges with low, medium or high correlation.

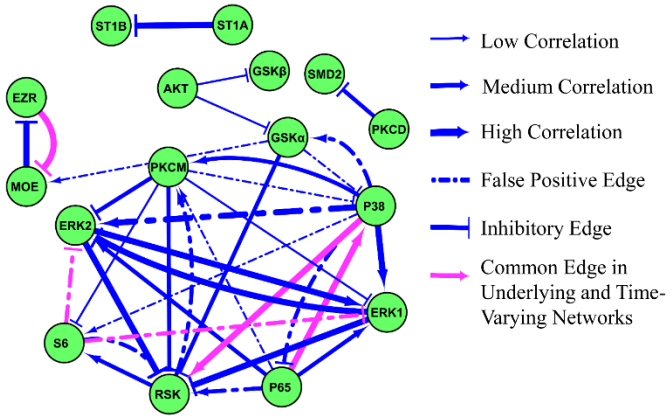



**Figure II.3 The reconstructed network for the underlying signaling network in RAW 264.7 macrophages.** This network represents the cross-talk between phosphoproteins considering the whole time-series for [1-10] minute period. The pink connections are common edges in the underlying network and the timevarying network (Figure II.5). Different edge-widths are used to represent low ($0.4 \le r < 0.5$), medium ($0.5 \le r < 0.75$) and high ($r \ge 0.75$) correlation coefficients corresponding to the edges. Inhibitory connections are shown with a blunt end instead of an arrow.

To test the robustness of our model to the choice of $\alpha$ and correlation threshold, we used different correlation thresholds and confidence intervals (for the two tailed t-test) to reconstruct the underlying network. To evaluate the performance of each trial, we compared the significant connections identified for the underlying network to the true connections from the literature.

Table II.1 implies that by increasing α from 0.01 to 0.02 and 0.05, i.e., reducing the confidence interval from 99% to 98% and to 95%, the number of False Positives increase and thus, Type I error increases. We also tested the results for different correlation thresholds that result in further trimming of the parameters. The optimal correlation threshold for which Type I and Type II errors are both minimized, is C = 0.4.

**TABLE II.1**
**Robustness of Results of the Underlying Network to the Choice of Different Thresholds**

| Correlation Threshold | α | Type I Error | Type II Error | Accuracy |
|---|---|---|---|---|
| **C=0.4** | 0.01 | 0.07 | 0.56 | 0.86 |
| **C=0.5** | 0.05 | 0.05 | 0.66 | 0.86 |
| **C=0.4** | 0.02 | 0.07 | 0.53 | 0.86 |
| **C=0.4** | 0.05 | 0.10 | 0.48 | 0.84 |
| **C=0.6** | 0.01 | 0.02 | 0.79 | 0.87 |
| **C=0.7** | 0.01 | 0.02 | 0.84 | 0.86 |

We also studied the effect of more fine time-intervals. If we interpolate with steps of half a minute instead of one minute, the accuracy of the model does not change significantly. With a sample time of one minute, accuracy is 0.86, and with that of half a minute, accuracy is 0.87. We found that by using the cubic interpolation rather than linear interpolation, Type II error increases, justifying the use of linear interpolation.

Many of the connections found using our approach (underlying network, Figure II.3) were also identified using a PLS-based approach (*4*). There are some differences between our network and the network obtained using the PLS approach. The connections p38 ↔ p65, p65 → ERK1/2 and GSKα → RSK are found in our network (Figure II.3), but not in the PLS-based network. However, the connections PKCD → EZR, MOE/EZR → RSK and p38 → AKT are found using the PLS approach, but are absent in our network.

The correlation coefficients with their corresponding p-values, along with the Benjamini-Hochberg FDR and p-values based on the t-test on the model coefficients for the connections retained in the underlying network (Figure II.3) are listed in Table II.2. It can be noted, that the Benjamini-Hochberg FDR for all these connections/edges are less than 0.026. The distribution of the p-values (t-test on the model coefficients) from all 17×17 possible connections for the underlying network is shown in Figure II.4 (implicitly used to calculate FDR).

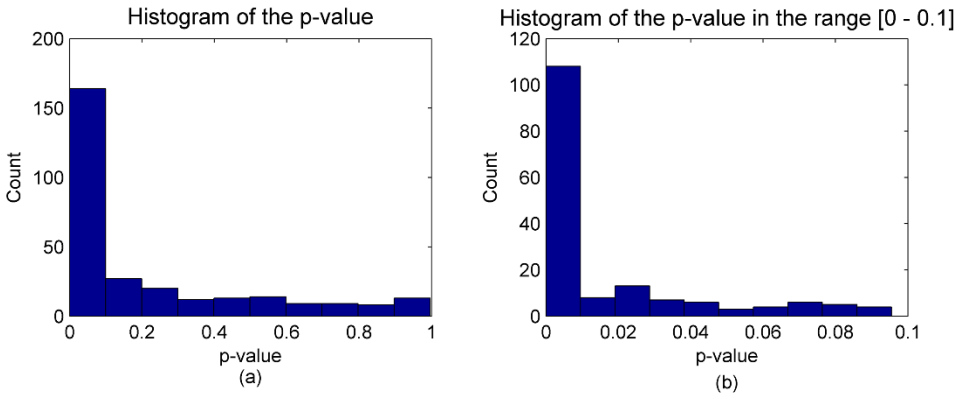



**Figure II.4 Histogram of the p-values (t-test on the model coefficients) for the underlying network generated from17×17 p-value numbers.**

**TABLE II.2**
**Correlation Coefficients and Statistical Significance of Edges Retained in the Underlying Network.**
Abbreviation: Benjamini-Hochberg (BH), false-discovery rate (FDR).

| Edges | | Correlation Based | | Model Coefficient Based | |
|---|---|---|---|---|---|
| Source Node | Target Node | Correlation Coefficient | p-value | FDR (BH) | p-value |
| ERK2 | ERK1 | 0.96 | 0 | 3.75E-10 | 3.89E-11 |
| ST1A | ST1B | 0.96 | 0 | 1.11E-02 | 3.74E-03 |
| ERK1 | ERK2 | 0.94 | 0 | 4.56E-09 | 5.52E-10 |
| MOE | EZR | 0.9 | 0 | 2.33E-08 | 2.98E-09 |
| EZR | MOE | 0.9 | 0 | 9.42E-03 | 3.00E-03 |
| ERK2 | RSK | 0.85 | 0 | 1.29E-06 | 1.92E-07 |
| P38 | RSK | 0.82 | 0 | 5.13E-03 | 1.56E-03 |
| ERK1 | RSK | 0.82 | 0 | 1.57E-06 | 2.44E-07 |
| P65 | P38 | 0.81 | 0 | 0 | 0 |
| P38 | ERK2 | 0.8 | 0 | 1.04E-03 | 2.83E-04 |
| P38 | ERK1 | 0.77 | 0 | 1.89E-05 | 3.60E-06 |
| P38 | P65 | 0.72 | 0 | 1.50E-06 | 2.28E-07 |
| P65 | ERK2 | 0.62 | 0 | 0 | 0 |
| RSK | S6 | 0.62 | 0 | 2.19E-04 | 4.99E-05 |
| S6 | ERK1 | 0.59 | 1.19E-279 | 0 | 0 |
| S6 | ERK2 | 0.59 | 1.21E-277 | 0 | 0 |
| P65 | RSK | 0.59 | 2.50E-270 | 0 | 0 |
| RSK | PKCM | 0.58 | 6.52E-269 | 2.60E-02 | 9.80E-03 |
| P65 | ERK1 | 0.58 | 2.02E-262 | 0 | 0 |
| P38 | PKCM | 0.54 | 6.74E-227 | 1.42E-11 | 1.43E-12 |
| P38 | GSKA | 0.54 | 1.75E-218 | 4.38E-08 | 6.21E-09 |
| S6 | RSK | 0.53 | 6.00E-214 | 0 | 0 |
| PKCD | SMD2 | 0.51 | 1.29E-197 | 0 | 0 |
| PKCM | RSK | 0.51 | 2.73E-196 | 5.74E-06 | 9.92E-07 |
| PKCM | ERK2 | 0.51 | 7.58E-193 | 7.67E-06 | 1.38E-06 |
| GSKA | RSK | 0.5 | 9.80E-188 | 2.12E-02 | 7.78E-03 |
| PKCM | ERK1 | 0.5 | 5.13E-182 | 1.09E-03 | 3.09E-04 |
| AKT | GSKA | 0.48 | 4.45E-166 | 1.91E-04 | 4.16E-05 |
| PKCM | S6 | 0.46 | 3.36E-156 | 2.04E-02 | 7.33E-03 |
| P65 | PKCM | 0.45 | 7.45E-150 | 8.90E-04 | 2.34E-04 |
| AKT | GSKB | 0.45 | 5.72E-147 | 2.64E-06 | 4.39E-07 |
| P38 | S6 | 0.45 | 5.40E-144 | 1.03E-03 | 2.78E-04 |
| GSKA | MOE | 0.45 | 1.49E-143 | 3.00E-04 | 7.05E-05 |
| PKCM | P38 | 0.44 | 9.25E-139 | 9.85E-03 | 3.20E-03 |
| GSKA | P38 | 0.44 | 3.39E-136 | 2.66E-03 | 8.00E-04 |

We also present the dynamic evolution of the network in three temporal stages shown in Figure II.5. The topology of the phosphoprotein network changes through time. Figure II.5.a corresponds to the reconstructed network in the first stage of the network development. Figure II.5.b and Figure II.5.c correspond to the reconstructed phosphoprotein networks for the second and third stages of the network evolution, respectively. The inhibitory edges such as AKT → GSKα/β are shown in Figure II.5.

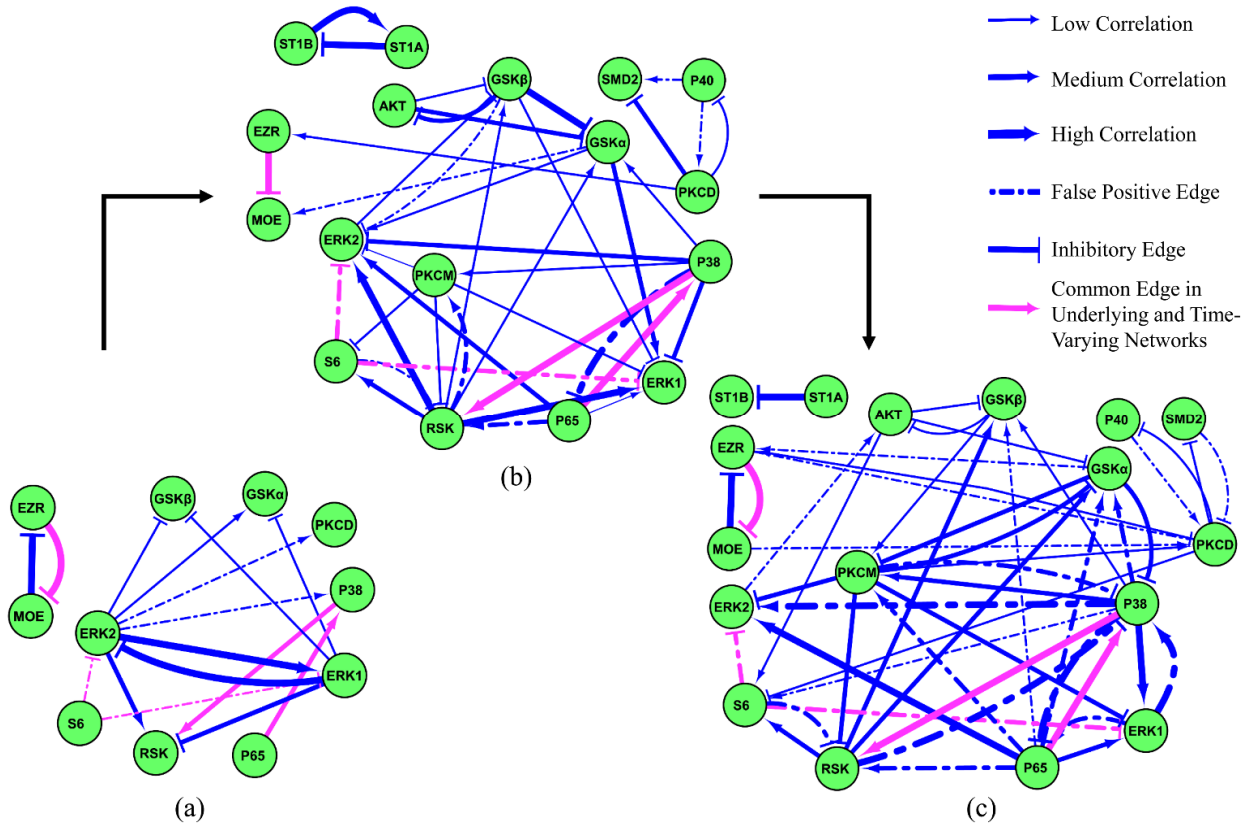



**Figure II.5 Time-dependent cascade of the phosphoprotein signaling network in RAW 264.7 macrophages in three stages.** (a) Reconstructed network in stage 1 related to [1-4] minute interval. (b) Reconstructed network in stage 2 related to [3-7] minute interval. (c) Reconstructed network in stage 3 related to [6-10] minute interval. The pink connections are common to all the three networks as well as the underlying network (Figure II.3). Different edge-widths are used to represent low ($0.4 \leq r < 0.5$), medium ($0.5 \leq r < 0.75$) and high ($r \geq 0.75$) correlation coefficients corresponding to the edges. Inhibitory connections are shown with a blunt end instead of an arrow.

**Effect of single-ligand data vs. double-ligand data:** To evaluate the consistency of the data across experiments involving different ligand combinations, we applied the VAR model to single ligand experiments (22 experiments). According to our results, the reconstructed network based on only single ligand experiments has higher Type I and Type II error. We also used only the double ligand experiments to model the network, and as we anticipated, the performance does not change significantly. It can be noted that the double ligand combinations result in activation of the signaling pathway in ways that are functionally distinct from single ligand experiments. Furthermore, as an estimate of the differences in the variability for different phosphoproteins

across time and treatment, we computed the ratio of the standard deviation of the standard deviation (*std*) to the mean of the *std* of every phosphoprotein (*std* is computed at every time for every treatment, using the replicate data), and found that this measure is of the same order (about 1) for all phosphoproteins across experiments.

### II.C.2 Temporal Evolution of the Phosphoprotein Network

In this subsection, we discuss the dynamic nature of the phosphoprotein network evolving in three successive temporal stages. For the sake of simplicity in our discussions, we treat each phosphoprotein as a node and each regulatory interaction as an edge in the network analysis.

Stage 1 (Figure II.5.a) shows the initiation of interactions among phosphoproteins. Since this network captures the early phase of the response of the system to the ligands, there are very few interactions taking place in the network. Extracellular signal-regulated kinase (ERK) plays a crucial role in the regulation and phosphorylation of most of the proteins that are present in the first stage of the network including p38 MAP Kinase (p38), p90 ribosomal S6 kinase (RSK), glycogen synthase kinase-3 (GSK), and protein kinase C (PKC) D. Ribosomal protein S6 (S6) affects ERK1 and ERK2. There is also a regulatory interaction between Nuclear Factor Kappa B (NF-κB p65) and p38. In addition, it is evident that Moesin (MOE) and Ezrin/Radixin (EZR) are part of the same pathway since a bidirectional link exists between them. As the network progresses to stage 2, several other interactions emerge. Figure II.5.b shows that protein kinase B (AKT) arises in stage 2 and regulates the phosphorylation of GSKα/β. The signal transducer and activator of transcription 1 A and B (STAT1A/B, also ST1A/B for short) pairs are variants of the same protein and are expected to be activating one another. Indeed, they show a bidirectional relationship. PKCD that was regulated by ERK2 in stage 1, now promotes the phosphorylation of EZR and mother against decapentaplegic homolog 2 (SMD2), as well as mutually regulating

neutrophil cytosolic factor 4 (p40). In stage 2, PKCM also appears and plays role in the regulation of RSK, S6 and ERK1/2, while being activated by p38. Role of S6 almost stays unchanged; i.e., it continues to regulate ERK1/2, except that as a result of the network progression from stage 1 to stage 2, we also see its interaction with RSK. This progression also brings about the phosphorylation of GSKα/β by RSK. In stage 1, p38 was activated by p65 and ERK2, whereas in the second stage, p38 regulates ERK1/2 along PKCM and gets involved in a mutual regulatory relationship with p65. p65 also affects ERK1/2 as well as RSK.

The evolution of the network to stage 3 provides not only most of the links that existed in stage 2, but also includes some new interactions. For instance, AKT proceeds to phosphorylate GSKα/β, while other nodes such as p65, RSK and p38 start to have causal influences on the activation of GSKα/β. Furthermore, in this phase, PKCD is regulated with the activation of PKCM, p40, SMD2 and EZR. Another interesting change is that p65 takes part in the activation of PKCM and ERK1/2. Moreover, AKT, broadly known for the activation of GSK, gets involved in the activation of S6, while being activated by ERK2.

Upon careful investigation of the time-dependent cascade of the network, we realize that there are very few stable interactions that exist in all three stages. Moreover, the well-known signaling pathways such as the MAPK, STAT1A/B, AKT/GSK and NF-κB pathways emerge only in the last two stages and not in stage 1. The few causal interactions that persist throughout the temporal progression of the network are S6→ERK1/2, EZR→MOE, p38→RSK and p65→p38. The time-varying succession of the significant interactions along with the related literature which validates some of these connections is shown in Table II.3.

**TABLE II.3**
**Comparison of our Results with the Current Literature**

| Correlated pairs | Stage 1 | Stage 2 | Stage 3 | Underlying network | Current knowledge | References |
|---|---|---|---|---|---|---|
| **(GSK, AKT)** | — | AKT → GSKα/β | AKT → GSKα/β | AKT → GSKα/β | AKT → GSK | (*58-60*) |
| **(GSK, RSK)** | — | RSK → GSKα/β | RSK → GSKα/β | GSKα → RSK | RSK →GSK | (*63, 64*) |
| **(GSK, P38)** | — | P38 → GSKα | P38 → GSKα/β | P38 ↔ GSKα | P38 → GSK | (*65, 66*) |
| **(GSK, ERK)** | ERK2/1 →GSKα/β | ERK2 ↔ GSKβ | — | — | ERK → GSK | (*67, 68*) |
| **(GSK, P65)** | — | — | P65 → GSKα/β | — | GSK → P65 | (*69, 70*) |
| **(RSK, S6)** | — | S6 ↔ RSK | S6 ↔ RSK | S6 ↔ RSK | RSK → S6 | (*62, 71*) |
| **(RSK, ERK)** | ERK2 → RSK | RSK → ERK1/2 | — | ERK1/2 → RSK | ERK → RSK | (*61, 62*) |
| **(RSK, P38)** | p38 → RSK | P38 → RSK | P38 ↔ RSK | P38 → RSK | P38 → RSK | (*72*) |
| **(PKC, S6)** | — | PKCM → S6 | PKCD → S6 | PKCM → S6 | PKC → S6 | (*73*) |
| **(PKC, P38)** | — | P38 → PKCM | P38 ↔ PKCM | P38 ↔ PKCM | P38 → PKCM | (*74, 75*) |
| **(PKC,ERK)** | — | PKCM → ERK1/2 | PKCM → ERK1/2 | PKCM → ERK1/2 | PKCM → ERK | (*76-78*) |
| **(PKC, EZR)** | — | PKCD → EZR | PKCD ↔ EZR | — | PKC → EZR | (*79, 80*) |
| **(PKC, MOE)** | — | — | MOE → PKCD | — | PKC → MOE | (*79, 80*) |
| **(PKC, P65)** | — | — | P65 → PKCM | P65 → PKCM | PKC → P65 | (*81-83*) |
| **(PKC, RSK)** | — | PKCM ↔ RSK | PKCM → RSK | PKCM ↔ RSK | PKC → ERK →RSK | (*62, 84*) |
| **(S6, ERK)** | S6 → ERK1/2 | S6 → ERK1/2 | S6 → ERK1/2 | S6 → ERK1/2 | ERK → S6 | (*62, 85*) |
| **(P65, RSK)** | — | P65 → RSK | P65 → RSK | P65 → RSK | RSK → P65 | (*86-88*) |
| **(P65, ERK)** | — | P65 → ERK1/2 | ERK1 → P65 | P65 → ERK1/2 | P65 → ERK | (*89-91*) |
| **(P65, P38)** | P65 → P38 | P65 ↔ P38 | P65 ↔ P38 | P65 ↔ P38 | P65 → P38 | (*89-91*) |
| **(P38, ERK)** | ERK2 → P38 | P38 → ERK1/2 | P38 → ERK2<br>P38 ↔ ERK1 | P38 → ERK1/2 | P38 → ERK | (*92-94*) |
| **(P38, S6)** | — | — | P38 → S6 | P38 → S6 | P38 → RSK → S6 | (*65, 95*) |
| **(AKT, ERK)** | — | — | ERK2 → AKT | — | AKT → ERK | (*96, 97*) |
| **(SMD, PKC)** | — | PKCD → SMD2 | PKCD ↔ SMD2 | PKCD → SMD2 | PKC → SMD | (*98, 99*) |
| **(SMD, P40)** | — | P40 → SMD2 | — | — | — | — |
| **(P40, PKC)** | — | PKCD ↔ P40 | PKCD ↔ P40 | — | PKC → P40 | (*100-102*) |

**TABLE II.3 (Continued)**
**Comparison of our Results with the Current Literature**

| Correlated pairs | Stage 1 | Stage 2 | Stage 3 | Underlying network | Current knowledge | References |
|---|---|---|---|---|---|---|
| **(AKT, S6)** | — | — | AKT → S6 | — | AKT → RSK → S6 | (*62, 71, 103*) |
| **(GSK,ERM)** | — | GSKα → MOE | GSK → EZR | GSK → MOE | — | — |

### II.C.3 Summary of Results

We have used a linear-model structure, least-squares regression and statistical hypothesis testing (t-test) on the coefficients of the linear model to identify significant edges in the network. Two types of networks have been identified, (1) based on the entire (interpolated) time-course data during [1-10] min, referred to as the underlying network (Figure II.3), and (2) temporally evolving network, in three-stages, based on three overlapping temporal regimes (Figure II.5). There is considerable overlap between our networks and a network obtained by a PLS-based approach published in the literature. The temporally-evolving network of Figure II.5.a shows the initiation of interactions among the phosphoproteins in stage 1 (e.g., ERK → p38/RSK,/GSK/PKCD and S6 → ERK1/2), and the addition (e.g., AKT → GSKα/β and PKCM → RSK,/S6/ ERK1/2 during stage 1 → stage 2) or deletion (ERK2 → PKCD during stage 1 → stage 2) of specific connections with progress to stages 2 and 3. Persistent connections throughout the temporal progression of the network are S6 → ERK1/2, EZR → MOE, p38 → RSK and p65 → p38. We also found that the reconstructed network based on only single ligand experiments has higher Type I and Type II error as compared to using both single- and double-ligand data.

## II.D Validation of Results and Discussion

The results shown above are acquired through data-driven reconstruction of the network

with no *a priori* information about the behavior of the underlying biological system. Here, we inspect our results and compare them with the existing information in the biology literature. In Table II.3, every causal relationship between pairs of phosphoproteins is shown by a directed arrow, and each mutual interaction is shown by a bi-directed arrow.

**Role of AKT/GSK:** GSK mediates protein phosphorylation and is involved in various intracellular pathways, metabolism and cancer. In mammalian cells GSK is encoded by two genes GSKα and GSKβ, with similar biochemical and substrate properties. GSK targets proteins that are involved in Alzheimer's disease and neurological disorders. AKT is broadly known for activation and inhibition of GSK phosphorylation in HEK293 (Human Embryonic Kidney 293) cells, zebrafish and xenopus embryo (*58-60*). We can readily see that the relationships AKT → GSKα and AKT → GSKβ, representing phosphorylation of GSKα and GSKβ by AKT, are captured in our model. Our results also indicate that the bidirectional connection AKT ↔ GSKβ exists in second and third stages. In addition to AKT, recent studies show that RSK plays a role in modulating the activity of GSK in cerebral granule neurons, xenopus development and intracellular neural signaling systems (*104-106*). There is also indication that the activation of RSK is responsible for the phosphorylation of GSKβ induced by epidermal growth factor (EGF) in human epidermoid A431 cells (*63*), and that GSKβ expressed in HeLa cells (from human cervical cancer cell line) is phosphorylated on Ser-9 by activation of p90Rsk (*64*). Our model suggests the connection RSK → GSKα/β in stages 2 and 3, and the reverse connection GSKα→RSK in the underlying network. In previous studies it has been discovered *in vitro* that GSK is differentially regulated by the stimulation of PKC in rabbit skeletal muscle cells, Sf9 cells and HEK293 cells (*107-109*).

Another phosphoprotein involved in the regulation of GSK is p38. Recent studies indicate

that p38 induces GSK phosphorylation in brain, thymocytes and human breast cancer cells (MDA-MB-231 cells) (*65, 66*) which is detected in the last two stages in our network. Furthermore, ERK activates GSK through phosphorylation in Hep-G2 cells and myocardial tissue cells in mice (*67, 68*). We detect this relationship in the first two stages. Moreover, the existing knowledge illustrates that GSK is involved in the activation of p65 in hepatocytes from mice and HeLa cells (*69, 70*) while our model captures the reverse connection p65 → GSKα/β in stage 3.

**EZR and MOE:** EZR and MOE are part of the same pathway, called Ezrin/radixin/moesin (ERM) protein pathway. The ERM proteins regulate actin cytoskeleton and are involved in signaling, transport, and structural functions of the cell (*110, 111*). As we can see in Figure II.2, the heat-map shows high correlation between these variables. In addition, the pairs ERK1/2 and STAT1A/B are variants of the same protein and are expected to be regulated similarly. Thus, as expected, high correlations and bidirectional causal relationships are observed between the members of each pair in Figure II.3 and 5. Despite the fact that the heat-map in Figure II.2 shows very high correlation between GSKα and β in all stages, we observe the connection GSKβ → GSKα only in stage 2. This is an interesting result confirming the fact that "correlation does not imply causality" in the sense that the two variables may be highly correlated but there is no information in the past of one of them that can be used to predict the future of the other. The same result was found for PKCD/M. The connection PKCM → PKCD was found only in stage 3.

**S6 and RSK:** Ribosomal protein S6, which is involved in cell growth and regulation of cellular translation, is phosphorylated at several serine residues with mitogen stimulation by activation of one or more protein kinase cascades. It is well known that in mammalian cells, phosphorylation of ribosomal protein S6 *in vitro* and *in vivo* is regulated by the activation of RSK (*62, 71*), while our results indicate the existence of a bidirectional connection S6 ↔ RSK. RSK is

involved in receptor-mediated signal transduction. Phosphorylation of RSK, which promotes cell survival and proliferation, lies at the end of the signaling cascade mediated by ERK and is regulated through the activation of ERK subfamily of MAP kinases (*61, 62*). We observed this relationship in the first and second stages. Furthermore, our network suggests that RSK can be activated by p38 through the connection p38 → RSK in stages 1 and 2 and p38 ↔ RSK in stage 3. In current literature there is some evidence confirming this interaction in HEK293 cells (*72*). Protein kinase C (PKC) is a family of fatty acid-activated protein kinase enzymes that is involved in regulating cell growth, learning and memory, transcription and mediating immune response. PKC which exists in various isoforms, is known to be involved in the activation of ERK in HEK293 cells (*84*), which then results in the activation of RSK through the MAP kinase pathway (*62*). Therefore it is anticipated that RSK and PKC have a hidden indirect relationship that was captured in our model where the connection PKCM → RSK is found in stage 3 and the underlying network and the connection PKCM ↔ RSK is found in stage 2. Our model still captured this connection by considering a faster time step (half a minute) in the model. In addition, PKC mediates the phosphorylation of S6 *in vivo* in HEK 293 cells (*73*). PKCM → S6 can be found in stage 2 and the underlying network and PKCD → S6 in stage 3.

**ERK and p38 (MAPK):** There are three distinct subfamilies of MAPK pathway: ERK1/2, JNK and p38 MAP kinases that have substantial impact on mediating various cellular signaling functions and physiological processes. These three enzymes are part of a phosphorylation system in which they regulate and phosphorylate one another (*112*). In this study we do not analyze the role of JNK in the signaling pathway, and we focus on the role of ERK1/2 and p38 in regulation and phosphorylation of one another and other phosphoproteins. The activation or inhibition of p38 potentiates the activation of ERK (*92-94*). Unlike other pathways that appear only in the last two

stages in our results, the crosstalk between ERK and p38 is found in all three stages. The activation of NF-kappa B (p65) can be triggered by the phosphorylation of ERK1/2 and recent research affirms the existence of cross-talk between ERK and p65 and between p65 and p38 (*89-91*) that can be seen in Figure II.5. p38 MAPK plays a critical role as downstream effector of PKC enzymes in LNCaP human prostate cancer cells and SK-Hep-1 hepatocellular carcinoma cells (*74, 75*). Our results indicate the connections p38 ↔ PKCM in stage 3 and the underlying network, and p38 → PKCM in stage 2. Furthermore, p38 modulates the phosphorylation of subfamilies of RSK such as 70 kDa ribosomal S6 kinase (p70S6K) and ribosomal S6 kinase 1 (S6K1) (*65, 95*) . We also know that RSK's target substrate is S6 (*62, 71*). This implies that p38 may indirectly play a role in the phosphorylation of S6. Our findings indicate that the connection p38 → S6 exists in stage 3 and the underlying network. There is no evidence in the existing literature confirming this relationship. The correlation coefficients for these edges are close to the correlation threshold. With a faster time step in the model, this connection is no longer significant. Hence, this interaction can be considered as false positive in our results. Moreover, phosphorylation of ribosomal protein S6 is known to be dependent upon the activation of ERK in HeLa cells and in mouse dentate gyrus (*62, 85*) whereas our model captured the reverse connection.

Recent evidence implies that stimulation of PKC activates ERK1 and ERK2 in myocardial cells of rabbit, glomeruli of diabetic rats and glomerular mesangial cell cultures under high glucose conditions and in human neutrophil cells (*76-78*). In our results, this relationship arises in the last two stages.

**p65**: Nuclear Factor Kappa B (NF-κB) exists in almost all animal cell types and is involved in mRNA transcription, regulation of inflammation, apoptosis and immune responses. There is some evidence that p65 NF-κB exists in the cytoplasm of unstimulated cells in an inactive form,

and that it can be activated by exposure to PKC in human YT cells (*81-83*), whereas our results captured the reverse connection p65 → PKCM. It is interesting that previous computational methods such as those in (*4*) also captured the same reverse connection. Furthermore, there is some evidence that activation of NF-κB requires RSK-dependent p65 phosphorylation in vascular smooth muscle cells (*87, 88*) but extended analysis is needed to thoroughly understand the role of p65 in the biological function of RSK (*86*). Our model estimated the opposite relationship p65 → RSK in stages 2, 3 and in the underlying network. Interestingly, in our analysis, the coefficient for RSK → p65 is just below the threshold and hence is not included in the network.

**Other Pathways**: Recent studies show evidence that activation of AKT inhibits the activation of the ERK pathway in C2C12 mouse myoblast cells (*97*) and that specific drugs unravel the crosstalk between the AKT and ERK pathways in neural stem cells (*96*). In fact, we found the connection ERK2 → AKT in stage 3. SMD2 relays extracellular signals from transforming growth factor beta (TGF-β) ligands to the nucleus (*113, 114*). There is some evidence that activation of SMAD (SMAD2, also SMD2) is modulated by protein kinase C in NIH-3T3 cells (*98, 99*), while the connection PKCD → SMD2 in stage 2 and the underlying network and PKCD ↔ SMD2 in stage 3 is captured in our networks. Some evidence provide affirmation that phosphorylation of ezr/radixin/moesin (ERM) is dependent upon catalytic function of PKC in MCF-7 breast cancer cells and in endothelial cells (*79, 80*). Our network reconstruction captures PKCD → EZR in stage 2, PKCD ↔ EZR and the reverse connection, MOE → PKCD, in stage 3.

The current knowledge confirms that p40 is phosphorylated *in vitro* by protein kinase C in HL-60 cells and human neutrophils (*100-102*). The bidirectional connection PKCD ↔ p40 was found in stage 2 and 3 of our reconstructed network. Our model also captures the connection p40 → SMD2 in stage 2.

AKT → S6 appears in stage 3 of our networks. It is known that protein kinase B (AKT) plays a role in the phosphorylation of RSK in human 293 cells (*103*) and ribosomal protein S6 (S6) is a substrate of RSK (*62, 71*). Thus, it can be anticipated that AKT is capable of having an indirect impact on the phosphorylation of S6. This connection is statistically significant even with a faster time step in the model. Another potential novel connection is the crosstalk between GSK and ezrin/radixin/moesin (ERM), GSK → ERM (*4*).

**Relationship of signaling pathways with diseases:** Some of these pathways such as p38 and NF-κB regulate the transcription of the cytokine tumor necrosis factor α (TNFα) which is a target for rheumatoid arthritis (*115*). NF-κB is involved in the regulation of pro-inflammatory chemokines and cytokines in meningitis (*116*). Furthermore, deviations in the levels of MAPKs from their normal cellular levels have been implicated in the development of cancer (*117*).

## II.E Conclusion

We have applied the notion of Granger causality through the vector autoregressive model to develop a novel framework for reconstructing dynamic networks from large-scale multi-experiment multivariate high-throughput data sets. We used an approach based on a linear-model template and statistical hypothesis testing (t-test) of the coefficients of the model to find significant or potentially causal connections. We have applied this methodology to phosphoprotein time-course data generated by the Alliance for Cellular Signaling (AfCS) in RAW 264.7 macrophage cells in single and double ligand experiments. We were able to predict connectivity, causality and dynamics of information flow in the progression of the phosphoprotein network. We also found that the reconstructed network based on only single ligand data has higher Type I and Type II error as compared to using both single- and double-ligand data.

Since the intracellular networks have a dynamic nature and their topology changes with time, in this work, our main goal was to investigate the temporal evolution of the phosphoprotein network. During the early stage, ERK plays an important role in regulating p38, RSK, PKCD and GSK, while ERK itself is regulated by S6. As the network evolves to the second and third stages, the well-known signaling pathways such as the MAPK, STAT1A/B, AKT/GSK and NF-κB pathways appear to play role in the network. These results have enhanced our knowledge about the important signaling pathways that activate macrophage cells and play an essential role in the secretion of cytokines during an inflammatory response, and may contribute to finding novel targets for inflammation-related diseases.

The method we have developed and applied here provides a strategy for reconstructing and analyzing dynamical networks in biological systems. In addition to providing networks in the temporal context, our method provides the directionality and potential causality of molecular interactions. We note that we built our methodology based on the notion of Granger causality, which is not meant to be equivalent to the true causality.

## II.F Acknowledgements

Chapter II, in full, is a reprint of the material as it appears in Time-Varying Causal Inference from Phosphoproteomic Measurements in Macrophage Cells 2014. Masnadi-Shirazi, Maryam; Maurya, Mano R.; Subramaniam, Shankar., IEEE Transactions on Biomedical Circuits and Systems, Volume 8, 2014. The dissertation author was a primary investigator and author of this paper.

Chapter III

# Dynamic Causal Network Reconstruction of a Mouse Cell Cycle

## III.A Abstract

Biochemical networks are often described through static or time-averaged measurements of the component macromolecules. Temporal variation in these components plays an important role in both describing the dynamical nature of the network as well as providing insights into causal mechanisms. In this study, we use well-constructed temporal transcriptional measurements in a mammalian cell during a cell cycle, to identify dynamical networks and mechanisms describing the cell cycle. The methods we have used and developed in part deal with Granger causality, vector autoregression and change point detection algorithms that are traditionally employed in engineering. From the temporal measurements in mouse embryonic fibroblasts, we identify precisely the timing of different phases of the cell cycle, namely, G1, S and G2/M phases, as well as the key regulators in each of the phases. We also pinpoint the temporal dependence of each of the proteins in the network on their own past and that of others that are causally linked to them. In addition, we provide a modular analysis of the temporal networks paving the way for design of precise experiments for modulating the regulation of the cell cycle.

## III.B Introduction

The progression of a eukaryotic cell cycle is governed by a complex, dynamical network of molecular interactions that regulate a series of directional and irreversible events such as cell growth, DNA replication, mitosis, and cell division. The biochemical pathways controlling the order and timing of cell cycle phases, called cell cycle checkpoints, play an essential role in maintaining genomic stability of the cell. Dysregulation of these checkpoints can alter the ability of the cell to undergo cell-cycle arrest in response to DNA damage and may lead to cancer. Significant progress has been made in identifying molecular players and pathways involved in cell cycle mechanisms through extensive investigations on model systems like yeast. Protein assays,

transcriptional studies, fluorescent imaging, and protein interaction mapping have all contributed to our current understanding of the cell cycle. From these studies and other phenotypic assays, molecular players engaged in distinct phases of the cell cycle, namely, G1, S, G2, and M phases, have been identified, resulting in a static pathway map of the cell cycle (*12*). These maps lack dynamical information, owing to the absence of systematic time series measurements. Fine-grained time series measurements of a mammalian cell cycle, can enrich the understanding of dynamical networks through which the temporal relationships between molecular players and modules can be inferred, and further provide insights into mechanistic causality. In this work, we present a systematic fine-grained RNA sequencing study of the transcriptional profiles during a mammalian cell cycle. Although these measurements are at the transcript level, we anticipate that given the strong transcriptional mechanisms that are concomitant with the cell cycle, these data have the potential to provide detailed dynamical mechanisms of the cell cycle.

Inferring causality from time-series data poses considerable challenges; conventional methods of network reconstruction offer a static characterization of the network topologies, devoid of any temporality which is an ingrained feature of biological systems. For example, correlation-based methods (*1, 2*), matrix-based methods such as least-squares, principal component regression (PCR) (*3*), and partial least squares (PLS) (*4*), L1-penalty based approaches such as least absolute shrinkage and selection operator (LASSO) and fused LASSO (*118, 119*), Gaussian graphical models (*120*), and information-theory based approaches relying on mutual information (*121, 122*) are among the methods primarily used for static network reconstruction. Boolean network (BN) is among the approaches proposed to model dynamic gene regulatory networks through parameter estimation (*5, 123-125*). Although BN captures temporal relationships, it requires discretization of gene expression levels to binary values and simplification of the network topology based on prior

knowledge to permit parameter estimation. Bucci *et al.* (2016) propose an approach called MDSINE that models and predicts the dynamics of microbial systems (*126*), but does not provide a time-varying view of the causal interactions. A dynamic Bayesian learning approach provides a temporally evolving picture of the network (*6, 127*), but is computationally expensive and tends to perform poorly on high dimensional data. Even though time series data can be used to easily construct correlation networks, developing quantitative models from these data is complicated due to the inherent nonlinearity of biological systems. However, it is possible to capture this nonlinearity using successive linear models over distinct time windows or temporal regimes. The assumption is that within a given regime, the topology of the network does not change. This is an alternative to building non-linear models which require substantially larger amounts of data due to the substantial increase in dimensionality even if only the quadratic terms are considered. While there has been several attempts at identifying different regimes in long time-series, mainly in the signal processing community (*13-15*), they have not been used to further develop evolving dynamical models and networks for biological systems.

We have developed a framework to investigate the temporal changes in the cell cycle network using RNA-seq time series data from Mouse Embryonic Fibroblast (MEF) primary cells. We use a non-parametric change point detection (CPD) algorithm (*16*) based on Singular Spectrum Analysis (SSA) (*17*) to infer the mechanistic changes in the time-course data for a set of 63 cell cycle genes to estimate cell cycle phases. We also use the notion of Granger causality implemented through a vector autoregressive (VAR) model (*18*) to predict the future expression levels of each gene as a function of the past expression levels of other genes yielding directionality of gene regulation among the 63 cell cycle genes. Furthermore, we utilize the concept of Minimum Description Length (MDL) to use past expression levels of genes, up to 9 time lags (equivalent to

4.5 hours), to determine the minimum data information from past events required for a robust prediction of values at the current time.

This computational scheme enabled us to (i) estimate the timing of cell cycle phases, (ii) infer the duration of the G1, S and G2/M phases of the MEF cell cycle to be 14.5, 10 and 4 hours, respectively, (iii) reconstruct three successive directed graphs representing the key regulatory mechanisms among the 63 cell cycle genes in the G1, S and G2/M phases of the cell cycle, (iv) infer the temporal impact that biological processes have on one another, as well as the dynamic changes in temporal dependencies as the cell evolves through successive phases, and (v) reflect the chronological order of regulatory events that are crucial to cell cycle control. The main power of our work is its ability to capture key pathways and important causal interactions over time, providing a broad picture of the dynamics of a cell cycle regulatory network. We validate the reliability of our time-varying network for cell cycle progression by comparing the interactions detected in our results to the well-known regulatory pathways in the literature as well as estimating temporal interdependences (time-delay) between important biological processes as the cell evolves through successive phases of the cell cycle.

## III.C Materials and Methods

### III.C.1 RNA-seq Data

The gene expression profiles are acquired through a RNA-seq experiment for serum response of Cf-1 MEF primary cells (E13 embryos), the purpose being to transcriptionally characterize the changes in the cell cycle genes as the cell cycle progresses. After the cells have been incubated in starvation medium (0.5% FCS) for 36 hours, serum is added to reach 20%. RNA isolation is performed under Trizol RNA extraction protocol. The RNA-seq data is aligned using

the STAR RNA-seq aligner (*128*) and the read counts are normalized using the HOMER software (*129*). Samples are taken one hour before the addition of serum, right before the addition of serum and every half hour after serum addition. This sampling routine is carried out for approximately two cell cycles. The raw time-series data is then processed to determine the fold-change in expression for each gene by dividing the expression level of each sample by the average of the expression levels at samples taken one hour before serum addition and right before serum addition.

### III.C.2 Change Point Detection Algorithm

Change Point Detection (CPD) is a non-parametric method based on sequential application of Singular Spectrum Analysis (SSA) to detect changes in time-series (*16, 130*). SSA is a powerful method for time-series analysis that is based on applying principal component analysis to the trajectory matrix acquired from the original time series. Basic SSA has four main steps:

1. Embedding

Let $x_1, x_2, \dots, x_T$ be a time series of length $T$, $M (M \leq T/2)$ be some integer called 'lag', and let $K = T - M + 1$. Define the trajectory matrix

$$X = (x_{ij})_{ij=1}^{M,K} = \begin{pmatrix} x_1 & x_2 & x_3 & \dots & x_K \\ x_2 & x_3 & x_4 & \dots & x_{K+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_M & x_{M+1} & x_{M+2} & \dots & x_T \end{pmatrix} \quad \text{(III.1)}$$

Note that the columns of the trajectory matrix $X_j$ $(j = 1, \dots, K)$ are vectors that lie in an $M$-dimensional space $\mathbb{R}^M$space.

2. Singular Value Decomposition

Let $R = XX^T$ be the lag-covariance matrix. The Singular Value decomposition (SVD) of $R$ provides us with $M$ eigenvalues, eigenvectors and principal components. $\gamma_1, \gamma_2, \ldots, \gamma_M$ denote the eigenvalues of $R$ and $U_1, U_2, \ldots, U_M$ are the corresponding orthonormal eigenvectors of $R$. If $b$ is the number of non-zero eigenvalues, and $V_i$ the eigenvector of $X^T X$, we have $V_i = X^T U_i$ for $i = 1, \ldots, b$. Then SVD of X will yield $\mathrm{X} = X_1 + X_2, + \cdots, X_b$, where $X_i = \sqrt{\gamma_i} U_i V_i; i = 1, \ldots, b$.

3. Grouping The indices $\{1, 2, \ldots, b\}$ can be split into two groups $I = \{i_1, \ldots, i_l\}$
4. and $I' = \{1, \ldots, b\} \backslash I$. Matrices $X_I = \sum_{i \in I} X_i$ and $X_{I'} = \sum_{i \notin I} X_i$ correspond to group $I$ and $I'$ and lead to the decomposition $X = X_I + X_{I'}$.
5. Diagonal Averaging (Hankelization)

This step transforms each matrix of the grouped decomposition in the previous step into new time series of length $T$ and is performed by averaging the diagonals $i + j = const$ of the Hankel matrices, $X_I$ and $X_{I'}$. Hankelization is an optimal procedure that uniquely defines the one-to-one correspondences between the Hankel matrices $X_I$ and $X_{I'}$ and their respective time-series $z_t$ and $\varepsilon_t$ of length $N$, leading to the decomposition of series $x_t$ into two series $z_t$ and $\varepsilon_t$

$$x_t = z_t + \varepsilon_t. \tag{III.2}$$

$z_t$ and the residual series $\varepsilon_t$ can be associated with signal and noise respectively.

The SSA captures the structure of the time-series by selecting the $l$ eigen-vectors, which span an $l$-dimensional subspace. Figure III.1 shows the scree plot and explained variance of eigenvalues, respectively when SVD is applied to the time-course data of Cdkn2d. This helps

choose the number of eigenvalues that capture sufficient variation in the time series (see the trend of Cdkn2d time-series displayed by the three largest eigenvalues in Figure III.2).

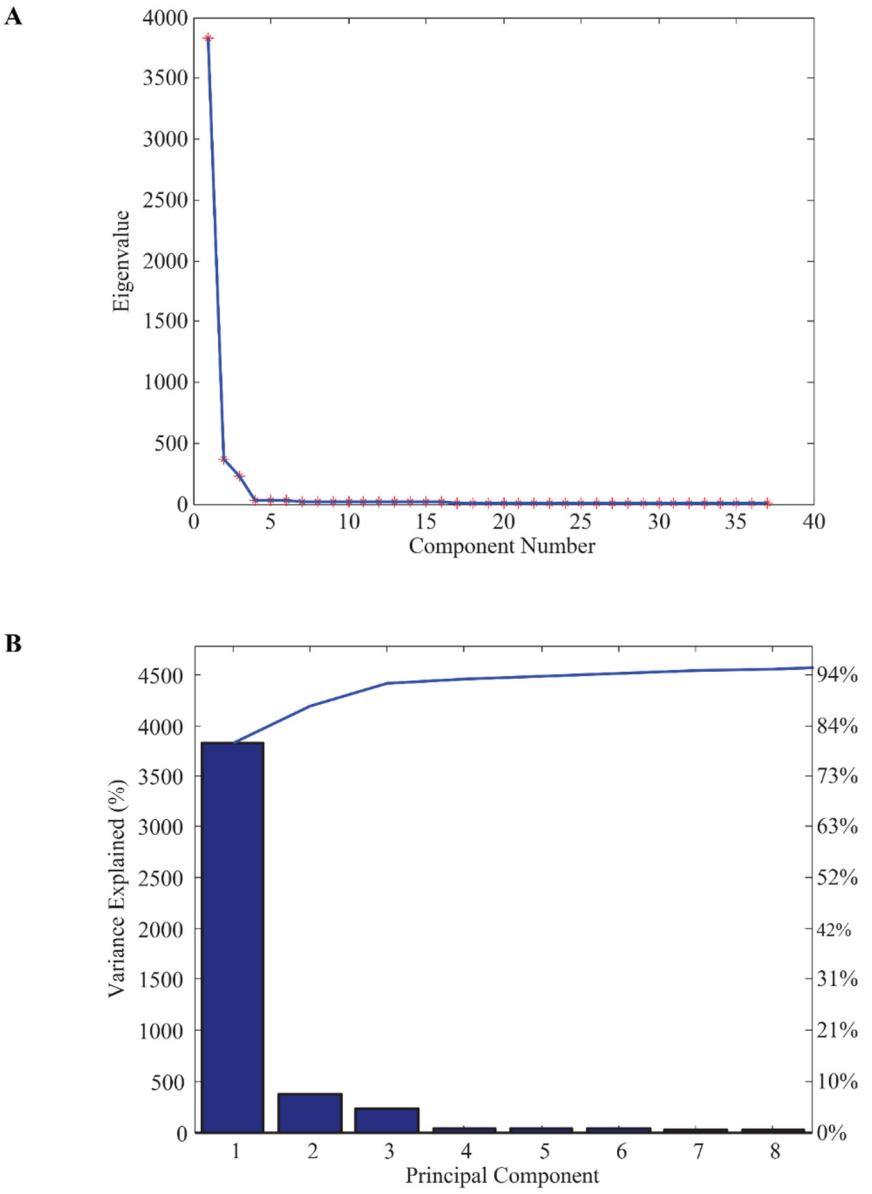



**Figure III.1 Principal Components of Cdkn2d gene expression profile. (A)** Scree plot of the Eigenvalues shows the ordered eigenvalues of the lag-covariance matrix corresponding to the gene expression profile of Cdkn2d. We can see a dramatic change in slope of the eigenvalue plot at the fourth component. Therefore, from what is observed in this plot, it is reasonable to retain the first three largest eigenvalues and group them together to select the set ***I***. **(B)** Explained variance for the first eight largest principal components that explain 95% of the cumulative variation. The

fourth and higher components explain very little variation and thus the first three largest eigenvalues can be grouped together in group $\boldsymbol{I}$.

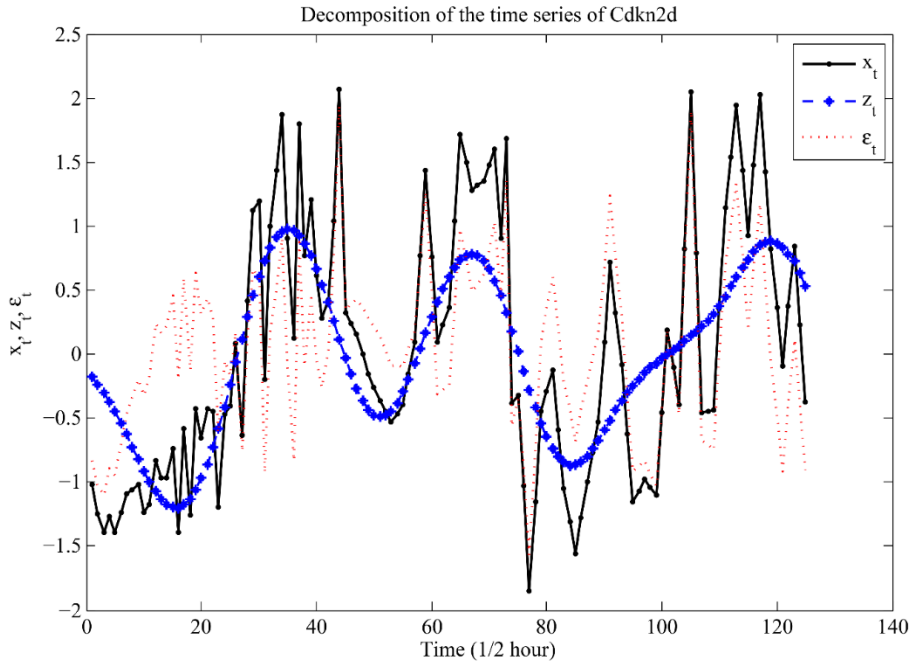



**Figure III.2 Decomposition of Cdkn2d time series into the main signal and noise.** This plot depicts the time series $\boldsymbol{x_t}$ for the gene expression profile of Cdkn2d (black curve), along with its decomposition into two time series $\boldsymbol{z_t}$ and $\boldsymbol{\varepsilon_t}$. $\boldsymbol{z_t}$ (blue curve) corresponds to the time series reconstruction from matrix $\boldsymbol{X_I}$ that is built from the three largest eigenvalues of the lag-covariance matrix, and $\boldsymbol{\varepsilon_t}$ (red dotted curve) corresponds to the time series reconstruction from matrix $\boldsymbol{X_{I'}}$ that is built from the remaining eigenvalues of the lag-covariance matrix.

The distance between the $l$-dimensional subspace selected in step three of the basic SSA and the vectors $X_j$ in equation III.1 should stay fairly small for $X_j, j > K$, if the time series $x_t, t = 1, \dots, T$ continutes for $t > T$ and there is no change in the mechanism generating $x_t$. Nonetheless, if at a certain time point $t + \tau$ the mechanism generating $x_t$ $(t > T + \tau)$ has altered, then we can expect to see an increase in the distance between the $l$-dimensional subspace and the vectors $X_j$ for $j > K + \tau$. This is equivalent to saying that a change in the structure of the time series pushed the vectors $X_j$ out of the subspace.

Change point detection can be achieved by sequentially applying the SVD to the lag-covariance matrices computed in time intervals of length $N$, $[n+1,\ n+N]$, for each $n$ to accommodate the change point detection algorithm to slow changes in the time series structure.

Let $x_1, x_2, \dots, x_T$ be a time series of length $T$. Let us choose two integers: the window width N (N≤T), and the lag parameter M (M≤N/2). Also, set K=N-M+1. The iterative change point detection algorithms has the following four steps.

Step 1. Construction of the *l*-dimensional space

1. For every suitable $n \geq 0$ we construct the trajectory matrix considering the time interval [n+1, n+N]

$$X_B^{(n)} = \begin{pmatrix} x_{n+1} & x_{n+2} & x_{n+3} & \cdots & x_{n+K} \\ x_{n+2} & x_{n+3} & x_{n+4} & \cdots & x_{n+K+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n+M} & x_{n+M+1} & x_{n+M+2} & \cdots & x_{n+T} \end{pmatrix} \tag{III.3}$$

These matrices are called *base matrices*. The columns of the base matrix $X_B^{(n)}$ are vectors $X_j^{(n)}$:

$$X_j^{(n)} = (x_{n+j}, \dots, x_{n+j+M-1})^T$$

2. For each n=0, 1,… we define the lag-covariance matrix $R_n = X_B^{(n)}(X_B^{(n)})^T$. The singular value decomposition of $R_n$ gives us a collection of M eigenvectors.

3. We select a distinct group $I = \{i_1, \dots, i_l\}$ of *l*<M of these eigenvectors; this determines an *l*-dimensional subspace $\mathcal{L}_{n,I}$ of the M-dimensional space $\mathbb{R}^M$ of the vectors $X_j^{(n)}$.

Step 2. Construction of the Test Matrix

Construct the matrix $X_T^{(n)}$ of size M×Q, whose columns are vectors $X_j^{(n)}$, $(j = p + 1, \dots, p + Q)$; that is,

$$X_T^{(n)} = \begin{pmatrix} x_{n+p+1} & x_{n+p+2} & x_{n+p+3} & \cdots & x_{n+q} \\ x_{n+p+2} & x_{n+p+3} & x_{n+p+4} & \cdots & x_{n+q+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n+p+M} & x_{n+p+M+1} & x_{n+p+M+2} & \cdots & x_{n+q+M-1} \end{pmatrix} \quad \text{(III.4)}$$

where $q = p + Q$. This matrix is called *test matrix*.

Step 3. Computation of the Detection Statistics

The detection statistics are:

- $\mathcal{D}_{n,I,p,q}$, the sum of squared distances between the vectors $X_j^{(n)}$, $(j = p + 1, \dots, q)$ and the $l$-dimensional subspace $\mathcal{L}_{n,I}$ of $\mathbb{R}^M$ is calculated as following:

$$\mathcal{D}_{n,I,p,q} = \sum_{j=p+1}^{q} (X_j^{(n)})^T X_j^{(n)} - (X_j^{(n)})^T U U^T X_j^{(n)} \quad \text{(III.5)}$$

  where $U$ is the $M \times l$ matrix whose columns $U_{i_1}, \dots, U_{i_l}$ are the orthonormal eigenvectors that span the $\mathcal{L}_{n,I}$ subspace.

- $S_n = \widetilde{\mathcal{D}}_{n,I,p,q} / \mu_{n,I}$, the normalized sum of squares of distances. Here

$$\widetilde{\mathcal{D}}_{n,I,p,q} = \frac{1}{MQ} \mathcal{D}_{n,I,p,q} \quad \text{(III.6)}$$

  and $\mu_{n,I}$ is an estimator of the normalized sum of squared distances $\widetilde{\mathcal{D}}_{j,I,p,q}$ at the time intervals $[j + 1, j + m]$ where the hypothesis of no change can be accepted. It is suggested to use $\mu_{n,I} = \widetilde{\mathcal{D}}_{m,I,0,K}$ where $m$ is the largest value of $m \leq n$ so that the hypothesis of no change has been accepted.

- Cumulative sum-type statistic

$$W_1 = S_1, \quad W_{n+1} = \max\{0, (W_n + S_{n+1} - S_n - 1/3MQ)\}, n \geq 1. \qquad \text{(III.7)}$$

Step 4: Decision Rule

The algorithm announces a structural change in the time series, if for some *n* we observe $W_n > h$ with the threshold $h = \frac{2t_\alpha}{MQ}\sqrt{\frac{1}{3}Q(3MQ - Q^2 + 1)}$, where $t_\alpha$ is the $(1-\alpha)$-quantile of the standard normal distribution.

**Choice of Parameters**: Window length N and lag M have to be chosen reasonably. The choice of N determines the smoothness or the effect of changes in the time series, i.e., if N is too large then we may miss changes in the time series. Alternatively, if N is too small we can have too many false alarms and outliers will be recognized as structural changes in the time Series. M is usually chosen to be M=N/2. The choice of $l$ is such that the largest $l$ principal components provide a good description of the signal and the lower $l - M$ components correspond to noise. It is advised to make a visual inspection of the SSA decomposition of the whole time series to choose $l$. A general recommendation for the choice of $p$ is that $p \geq K$ so that the columns of the base and test matrices do not coincide and thus, the change point detection algorithm is more sensitive to changes.

Figure III.1.A shows the plot of the ordered set of eigenvalues of the lag-covariance matrix corresponding to the gene expression profile (time-series) of Cdkn2d. We can notice that the fourth and higher components only explain 5-6% of the cumulative variation (Figure III.1.B). Therefore, the first three largest eigenvalues of the lag-covariance matrix will provide a good description of the original time series for Cdkn2d. Hence, it is appropriate to group the largest three eigenvalues

in set $I$ and the remaining eigenvalues in set $I'$to decompose the time series of Cdkn2d into the main signal $z_t$ and noise $\varepsilon_t$.

Figure III.2 displays the decomposition of Cdkn2d time series into two separate time series that are reconstructed from the decomposition of the trajectory matrix $X$ into $X_I$ and $X_{I'}$. $z_t$ is reconstructed from $X_I$ which corresponds to group $I$ of eigenvalues, and $X_{I'}$ corresponds to group $I'$ of the eigenvalues.

The cumulative sum-type statistic $W_n$ is computed based on the distance between the $l$-dimensional subspace and the vectors $X_j$ and compared against a threshold; every time the test statistic exceeds the threshold $h$, a change point is detected. In the case of Cdkn2d time series, $l = 3$ is chosen. Once $l$ is chosen the CPD algorithm is performed on the time series data. Figure III.3 depicts the detection of change points in the time-series for Cdkn2d time series. The change points detected are representative of a structural change in the mechanism generating the time-series.

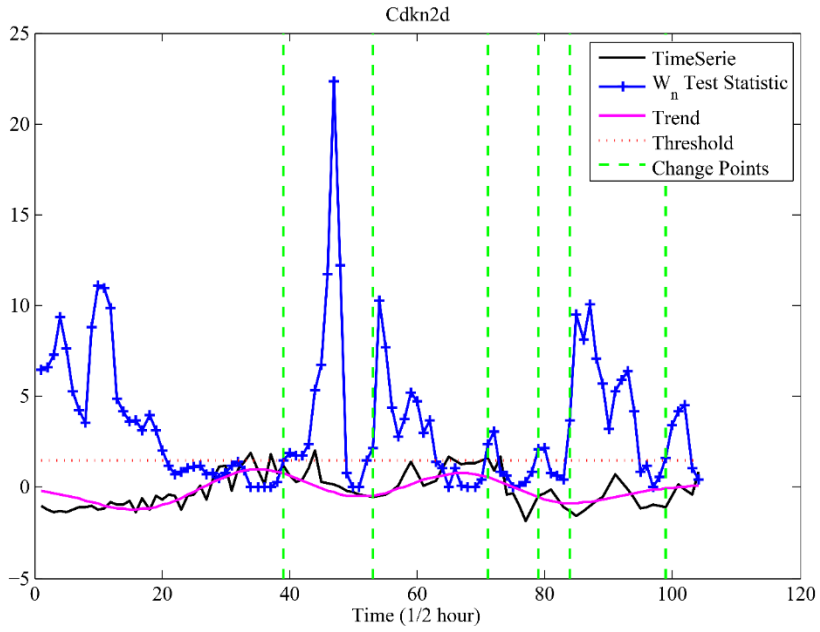



**Figure III.3 Plot of change points for Cdkn2d time series.** The black curve is the original RNA-seq time series data for Cdkn2d, $x_t$. The pink curve corresponds to $z_t$ which is the trend of the time series. The blue curve is the plot of the $W_n$ test statistic calculated through the CPD algorithm. The dotted red curve is the threshold $h$ which is used in the decision rule. Every time the $W_n$ test statistic exceeds the threshold, a change point is selected (green dotted lines).

### III.C.3 Granger Causality

Granger causality is a notion based on the ability to predict the future value of one process using the past values of another process (*131*). This notion was first introduced in macroeconomics and has proven useful in providing the direction of information flow, however it is not equivalent to true causality. Granger causality provides information about numerical information and prediction, while true causality is profoundly related to the influence of one variable onto another. Formally, a time series $x$ is said to Granger-cause a time series $y$ if the future value of $y$ can be better predicted given the past values of $x$ and $y$, $(x_{t-1}, x_{t-2}, \dots, y_{t-1}, y_{t-2}, \dots)$, than predicting the future of $y_t$ given only the past values itself, $(y_{t-1}, y_{t-2}, \dots)$. This statistical concept of causality can be well represented by the VAR model for linear relationships (*18*). A $d$-order VAR model of a $k$ dimensional time series is given by:

$$y(t) = v + A_1 y(t-1) + A_2 y(t-2) + \ \ldots \ + A_d y(t-d) + \ \varepsilon(t) \qquad \text{(III.8)}$$

where $y(t) = (y_1(t), y_2(t), \ldots, y_k(t))^T$ is a $(k \times 1)$ random vector, $y_i(t)$ is the measurement at time $t$ of the $i^{th}$ random variable, $A_l$ is a $(k \times k)$ autoregressive coefficient matrix, $\boldsymbol{v}$ is a $(k \times 1)$vector of intercepts and $\varepsilon(t) = (\varepsilon_1(t), \varepsilon_2(t), \ldots, \varepsilon_k(t))^T$ is a $k$-dimensional error vector of random variables with zero mean and covariance matrix $\sum$.

A necessary and sufficient condition for variable $y_j$ to be Granger-causal for $y_i$ is that the corresponding coefficient $a_{ijl}$ ($ij^{th}$ entry of $A_l$, $l = 1, \ldots, d$) is statistically significant (*54, 132*). Therefore, the direction of information flow can be determined by estimating the autoregressive coefficient matrices of the VAR model. The optimal order of the VAR model can be estimated via the minimum description length (MDL) principle.

### III.C.4 Estimation Stability with Cross Validation

Considering the time series $(y_1, \ldots, y_T)$ for each of the $k$ variables, the VAR model in Equation III.8 can be written compactly in the following matrix form (*133*):

$$Y = \varphi X + \varepsilon \qquad \text{(III.9)}$$

where $Y = (y_1, \ldots, y_T)^T$ is a $(T \times k)$ matrix whose columns are time series for each of the $k$ random variables with sample size $T$, $\varphi = (\varphi_0, \ldots, \varphi_{T-1})^T$ is a $(T \times (kd+1))$ matrix with $\varphi_t = (1; y(t); \ldots; \ y(t-d+1))^T$, $X = (v, A_1, A_2, \ldots, A_d)^T$ is a $((kd+1) \times k)$ coefficient matrix and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_T)^T$ is a $(T \times k)$ matrix. For each of the $k$ columns of matrices $Y, X$, and $\varepsilon$, we have the following linear regression model:

$$y_i = \varphi x_i + \varepsilon_i, \ i = 1, \ldots, k \qquad \text{(III.10)}$$

We are interested in recovering vector $x_i \in \mathbb{R}^{kd+1}$ from the observation $y_i \in \mathbb{R}^T$ and $\varphi$. Since $\varphi \in \mathbb{R}^{T \times (kd+1)}$, and $T \ll kd + 1$, we have an underdetermined system of linear equations,

and this linear inverse problem cannot be solved uniquely. However, if $x_i$ is sufficiently sparse, i.e., the support of $x_i$ has small cardinality, it is actually possible to recover $x_i$ by solving the following $\ell_0$ minimization problem (*134, 135*):

$$\hat{x}_i = \min ||x_i||_0 \quad subject\ to\ \ y_i = \varphi x_i\ ,\ i = 1, \dots,\ k \tag{III.11}$$

where $||x_i||_0$ denotes the number of nonzero coefficients of $x_i$. Since $\ell_0$ minimization is an NP-hard problem, it can be relaxed to an $\ell_1$-norm regularization that can be a heuristic for finding a unique sparse solution (*136*):

$$\hat{x}_i = \min\ ||y_i - \varphi x_i||_2 + \lambda\ ||x_i||_1\ ,\ i = 1, \dots,\ k \tag{III.12}$$

Note that $\ell_1$-norm regularization in Equation III.11 is strictly related to the Least Absolute Shrinkage and Selection Operator (LASSO) problem (*20*):

$$\hat{x}_i = \min \frac{1}{2} ||y_i - \varphi x_i||_2^2 + \lambda\ ||x_i||_1\ ,\ \ i = 1, \dots,\ k \tag{III.13}$$

The regularization parameter $\lambda$ in the LASSO sets a trade-off between the fit error $||y_i - \varphi x_i||_2^2$ and the sparsity of the signal $x_i$. In order to choose the desired $\lambda$, one can use traditional model selection criteria, such as Akaike's information criterion (*AIC*) (*21*) and Bayesian information criterion (*BIC*) (*22*). These criteria are easily computed, though are dependent on model assumptions and even if model assumptions are met, they may not be valid in the finite sample cases. The regularization parameter $\lambda$ is often selected through the model-free Cross-validation (CV) approach (*23, 24*). CV often leads to estimators with good predictive performance when sample size is large. In the cases where sample size is small, CV does not yield a good interpretable model because LASSO + CV is unstable and not reliable for scientific interpretations (*25*). In this work, we observed that selecting $\lambda$ through *Estimation Stability with Cross Validation* (ES-CV) leads to more meaningful and interpretable results (*26*). Estimation stability (ES) is based

on the idea that the solution is not meaningful if it varies considerably from sample to sample. The LASSO generates a family of solutions known as the solution path:

$$\hat{x}_i[\lambda] = minimize\ ||y_i - \varphi x_i||_2^2 + \lambda||x_i||_1 \tag{III.14}$$

We want to choose $\lambda$ in the solution path based on estimation stability. Since ES is tightly tied to the sampling scheme, we need multiple solution paths to evaluate stability. Cross-validation data perturbation is used to randomly partition the $T$ samples into $V$ groups of pseudo data sets by leaving out one group at a time. Let $\varphi^*[j],\ y_i{}^*[j]$ represent the $j^{th}$ pseudo data set (random partition) derived from $\varphi$ and $y_i$, respectively. The pseudo solutions are given by:

$$\hat{x}_i[j;\lambda] = minimize\ ||y_i{}^*[j] - \varphi^*[j]x_i||_2^2 + \lambda||x_i||_1 \tag{III.15}$$

for $j = 1, \dots, V, i = 1, \dots, k$. ES measures the stability or similarity of pseudo solutions across different groups of samples. For each $\lambda$, the stability of the following estimates

$$\hat{y}_i[j;\lambda] = \varphi\hat{x}_i[j;\lambda],\ j = 1, \dots, V,\ i = 1, \dots, k \tag{III.16}$$

are studied by looking at the sample variance of the estimates

$$\widehat{VAR}(\hat{y}_i[\lambda]) = \frac{1}{V}\sum_{j=1}^{V} ||\hat{y}_i[j;\lambda] - \bar{\hat{y}}_i[\lambda]||_2^2,\ j = 1, \dots, V,\ i = 1, \dots, k \tag{III.17}$$

where

$$\bar{\hat{y}}_i[\lambda] = \frac{1}{V}\sum_{j=1}^{V} \hat{y}_i[j;\lambda].$$

The normalized version of the sample variance is defined as the estimation stability metric:

$$ES(\lambda) = \frac{\widehat{VAR}(\hat{y}_i[\lambda])}{||\bar{\hat{y}}_i[\lambda]||_2^2} \quad \text{(III.18)}$$

*ES* is the reciprocal of the test statistic for testing the null hypothesis $H_0$: $\varphi x_i = 0$, and can be viewed as a selection of $\lambda$ as a set of hypothesis tests; for each $\lambda$ we are testing to see if the fit $\hat{y}_i[\lambda]$ is statistically different from fitting the null model ($\varphi x_i = 0$).

The most statistically significant solution along the solution path is the one whose ES metric has the largest reciprocal. Therefore, the most statistically significant solution is the one that locally minimizes the *ES* metric. In the case where noise overwhelms the signal (high noise), $y$ bears no relation to $\varphi$ and *ES* proposes inadvertent local minima. Thus, cross-validation is incorporated into finding the solution (ES-CV) (see Figure III.4.A). ES-CV further limits the choice of $\lambda$ to the local minimum of $ES(\lambda)$ that is greater than or equal to the choice of cross-validation (see Figure III.4.B) (*25, 26*).

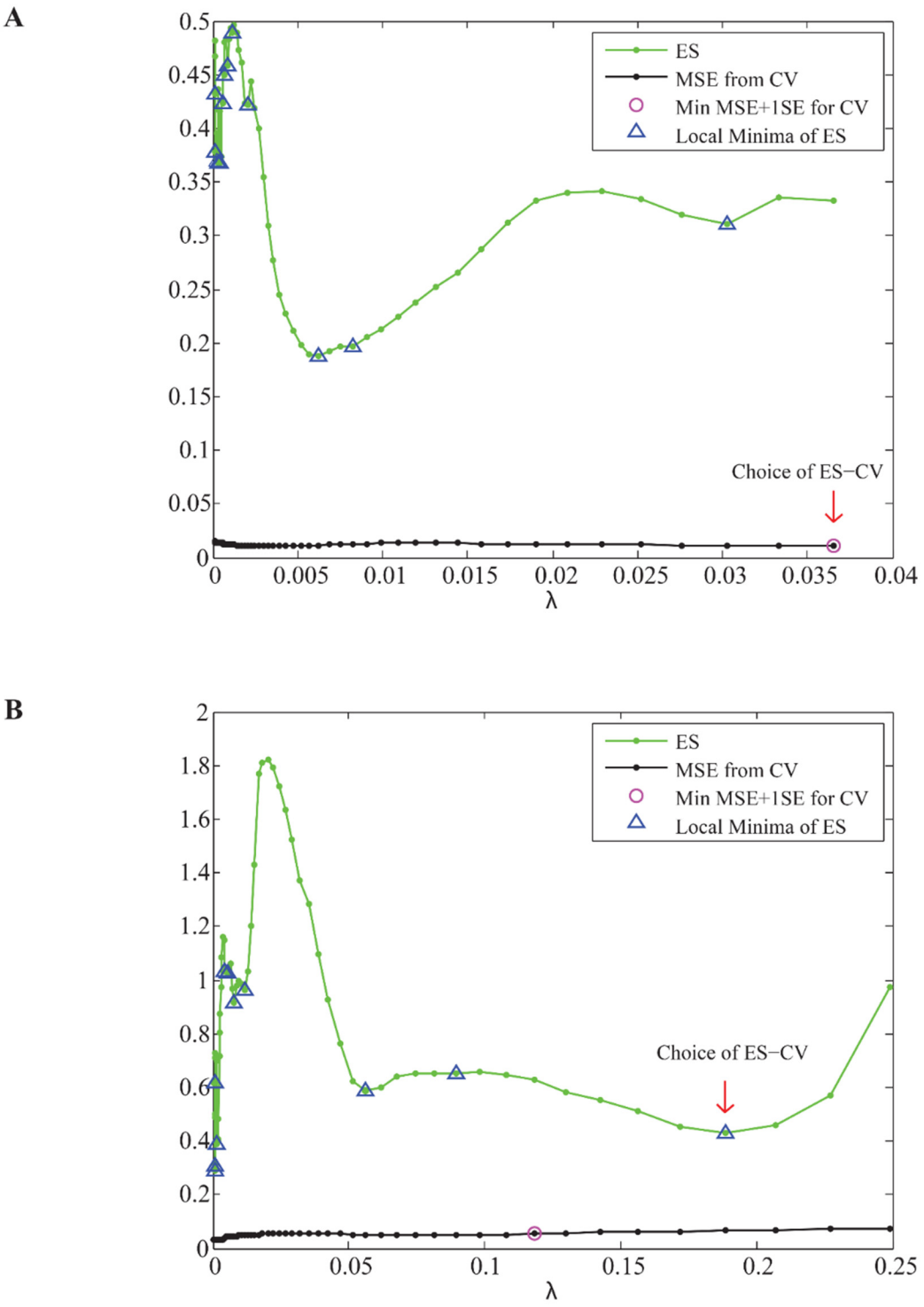



**Figure III.4 Estimation Stability with Cross Validation (ES-CV).** The green curves are the plot of the ES metric. The black curves are the plot of mean squared error (MSE) through cross validation. The blue triangles identify the $\lambda$ at which the local minima of the ES metric occur. The pink circles indicate the largest $\lambda$ such that MSE is within one standard error of the minimum MSE. **(A)** In the case where noise overwhelms the data, ES fails and CV is incorporated. We can note that between the choice of CV (pink circle) and the choice of ES (blue triangles), ES-CV picks the larger $\lambda$. **(B)** We can note that the ES-CV approach selects a larger $\lambda$ compared to the choice of cross validation. Hence, the choice of $\lambda$ selected through ES-CV leads to a sparser solution than that of CV.

### III.C.5 Minimum Description Length

The optimal order of the VAR model can be estimated through model selection approaches such as Minimum Description Length (MDL) (*19*). MDL selects a model that provides the shortest description of data. Description length for observations $y^T = \{y_1, y_2, \dots, y_T\}$ from a parametric family $\mathcal{M} = \{f(y^T|\theta): \theta \in \Theta\}$ is $-logf_\theta(y^T) + \mathcal{L}(\theta)$, where the first term is the cost function and the second term is the cost of transmitting the estimated parameter $\theta$. For a linear regression model in Equation (III.8), the observation $y$ has the following description length:

$$DL = \frac{T}{2}\log RSS + \frac{d}{2}\log T\,;\;\; d = 1, 2, \dots, d_{max} \tag{III.19}$$

where $RSS$ denotes the residual sum of squares and $d$ is the order of the VAR model in Equation 3. The optimal order is selected such that the code length in Equation III.19 is minimized:

$$d_{opt} = minimize\ \frac{T}{2}\log RSS + \frac{d}{2}\log T\,;\;\; d = 1, 2, \dots, d_{max} \tag{III.20}$$

### III.C.6 Evaluating Association between the Time-series of Two Cell Cycles

The RNA-seq experiment was done for two cell cycles (for mouse embryonic fibroblast primary cells) following serum starvation and the addition of serum. Serum starvation and refeeding for mammalian cell division does not necessarily result in synchronization of the entire cell population (*137, 138*). Thus, the two time-series data acquired through RNA-seq does not reflect the behavior of synchronized cells, and therefore they may not have begun at the same occasion of measurements. The time interval separating the start of the two cell cycles is called *delay* or *offset*. A common approach to finding the association between events in two time-series is cross-correlation in which the Pearson product moment correlation is computed for the two time-series (*139*). The offset is determined by finding the sample at which the highest cross-correlation

between the two time-series occurs. Figure III.5 shows the plot of the cross correlation of the two available time-series for Smc1a gene.

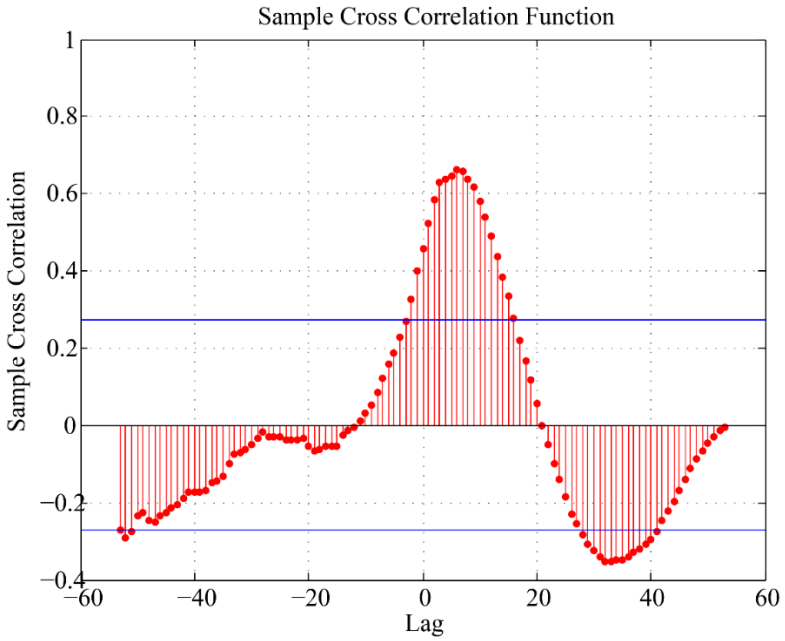



**Figure III.5 Cross correlation of two time-series of Smc1a gene.** The cross correlation plot of the two time-series shows that maximal association for the two time-series occurs with an offset of 7 samples.

### III.C.7 Precision of Results

Precision or confidence indicates the proportion of predicted positive edges that are real positives (*140*). In other words, Precision is a measure of accuracy of the predicted positives:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

## III.D Results

Gene expression in MEFs is measured at 96 different time points at intervals of 0.5 hr or 1 hr (later interpolated to every 0.5 hr), covering more than one full cycle and the G1, S and part of G2/M phases of another cycle. Of the 4248 differentially expressed genes, i.e., genes whose

expression values change more than 2-fold as compared to that at t=0 at one or more time points, 63 are cell-cycle genes included in the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways database (*12*). We first detected the different stages of the cell cycle using the CPD algorithm. Then we developed a VAR model for each stage through the estimation of optimal time-lags. Finally, we carried out an in-depth analysis of the temporally evolving networks as the cell cycle progresses.

### III.D.1 Detecting Temporal Changes and Stages in the Cell Cycle Time Series Data

In order to synchronize the cell cycle, the MEF cells were serum starved and the time-series RNA-seq measurements were initiated following the addition of serum to re-initiate the cell cycle. In order to identify different phases of the cell cycle from the time-series data, we use a model-free CPD algorithm (*16*) (discussed in the Materials and Methods section). The CPD algorithm captures the ongoing mechanistic changes as the cell cycle progresses and partitions the time series data into intervals with dominant trends, associated with cell cycle phases. It can be noted that no *a priori* assumptions on the duration of the cell cycle phases were incorporated in our analysis. In this study, we apply the CPD algorithm to 63 cell cycle genes presented in the KEGG pathway for mouse cell cycle (*12*). Table III.1 presents the list of these 63 cell cycle genes and their abbreviated gene symbols for mouse (mus musculus).

**Table III.1 List of 63 cell cycle genes presented in the KEGG pathway (Mus musculus).**

| Gene symbol | Gene full name |
|---|---|
| Abl1 | Abelson murine leukemia viral oncogene homolog 1 |
| Anapc1 | Anaphase-promoting complex subunit |
| Atm | Ataxia telangiectasia mutated |
| Bub1 | Mitotic checkpoint serine/threonine-protein kinase BUB1 |
| Bub1b | Mitotic checkpoint serine/threonine-protein kinase BUB1 beta |
| Bub3 | Mitotic checkpoint protein BUB3 |
| Ccnb2 | Cyclin B2 |
| Ccnd1 | Cyclin D1 |
| Ccne1 | Cyclin E1 |
| Ccnh | Cyclin-H |
| Cdc14b | Cell division cycle 14B |
| Cdc20 | Cell division cycle 20 |
| Cdc25a | Cell division cycle 25A |
| Cdc25b | Cell division cycle 25B |
| Cdc45 | Cell division cycle 45 |
| Cdc6 | Cell division cycle 6 |
| Cdc7 | Cell division cycle 7 |
| Cdh1 | Cadherin 1 |
| Cdk1 | Cyclin-dependent kinase 1 |
| Cdk2 | Cyclin-dependent kinase 2 |
| Cdk4 | Cyclin-dependent kinase 4 |
| Cdkn1a | Cyclin-dependent kinase inhibitor 1A |
| Cdkn1b | Cyclin-dependent kinase inhibitor 1B |
| Cdkn2a | Cyclin-dependent kinase inhibitor 2A |
| Cdkn2b | Cyclin-dependent kinase inhibitor 2B |
| Cdkn2c | Cyclin-dependent kinase inhibitor 2C |
| Cdkn2d | Cyclin-dependent kinase inhibitor 2D |
| Chek1 | Checkpoint Kinase 1 |
| Crebbp | CREB binding protein |
| Dbf4 | Dbf4 zinc finger |
| E2f1 | E2F transcription factor 1 |
| E2f4 | E2F transcription factor 4 |
| Espl1 | Extra spindle pole bodies 1, separse |
| Gadd45a | Growth arrest and DNA-damage-inducible 45 alpha |
| Gsk3b | Glycogen synthase kinase 3 beta |
| Hdac2 | Histone deacetylase 2 |
| Mad1l1 | MAD1 mitotic arrest deficient 1-like 1 |
| Mad2l1 | MAD2 mitotic arrest deficient-like 1 |
| Mcm3 | Minichromosome maintenance complex component 3 |
| Mdm2 | transformed mouse 3T3 cell double minute 2 |
| Myc | Myelocytomatosis oncogene |
| Orc1 | Origin recognition complex, subunit 1 |
| Pcna | Proliferating cell nuclear antige |
| Pkmyt1 | Protein kinase, membrane associated tyrosine/threonine 1 |
| Plk1 | Polo-like kinase 1 |
| Prkdc | Protein kinase, DNA activated, catalytic polypeptide |
| Pttg1 | Pituitary tumor-transforming gene 1 |
| Rad21 | RAD21 cohesin complex component |

**Table III.1 (Continued) List of 63 cell cycle genes presented in the KEGG pathway (Mus musculus).**

| Gene symbol | Gene full name |
|---|---|
| Rb1 | Retinoblastoma 1 |
| Rbl1 | Retinoblastoma-like 1 (p107) |
| Sfn | Stratifin |
| Skp2 | S-phase kinase-associated protein 2 |
| Smad2 | SMAD family member 2 |
| Smad4 | SMAD family member 4 |
| Smc1a | Structural maintenance of chromosomes 1A |
| Smc3 | Structural maintenance of chromosomes 3 |
| Stag1 | Stromal antigen 1 |
| Tfdp1 | Transcription factor Dp 1 |
| Tgfb1 | Transforming growth factor, beta 1 |
| Trp53 | Transformation related protein 53 |
| Ttk | Ttk protein kinase |
| Wee1 | WEE 1 homolog 1 |
| Zbztb17 | zinc finger and BTB domain containing 17 |

For every gene, the time-course data for approximately two consecutive cell cycles are available. We use cross-correlation between the two time-series data to obtain the offset between the two cycles by finding the time point at which the maximum association between the two time-series occurs (see Figure III.5). When the offset is computed for every gene, the gene expression profile is derived by properly concatenating the two time-series according to the offset and then the CPD algorithm is applied. This algorithm may detect more than one change point in the expression profile of each of the 63 cell cycle genes.

Figure III.6 is a radar chart that depicts the count of genes for which the CPD algorithm detects change points at every time point (1/2 *hour*) (data from 5 hours to 35 hours after the start of the first cell cycle is shown in Figure III.6). There are three significant peaks in the radar chart at 14.5, 24.5 and 28.5 hours at which the CPD algorithm detects change points for 29, 16 and 14 genes, respectively. We consider these peaks as break-points between the consecutive G1, S and G2/M phases of the cell cycle. According to the radar chart in Figure III.6, the duration of the G1, S and G2/M phases of the cell cycle is estimated to be 14.5, 10 and 4 hours respectively. Therefore,

we presume the intervals [1-14.5], [14.4-24.5] and [24.5-28.5] *hours* represent the expression profile of genes in the G1, S and G2/M phases of the cell cycle.

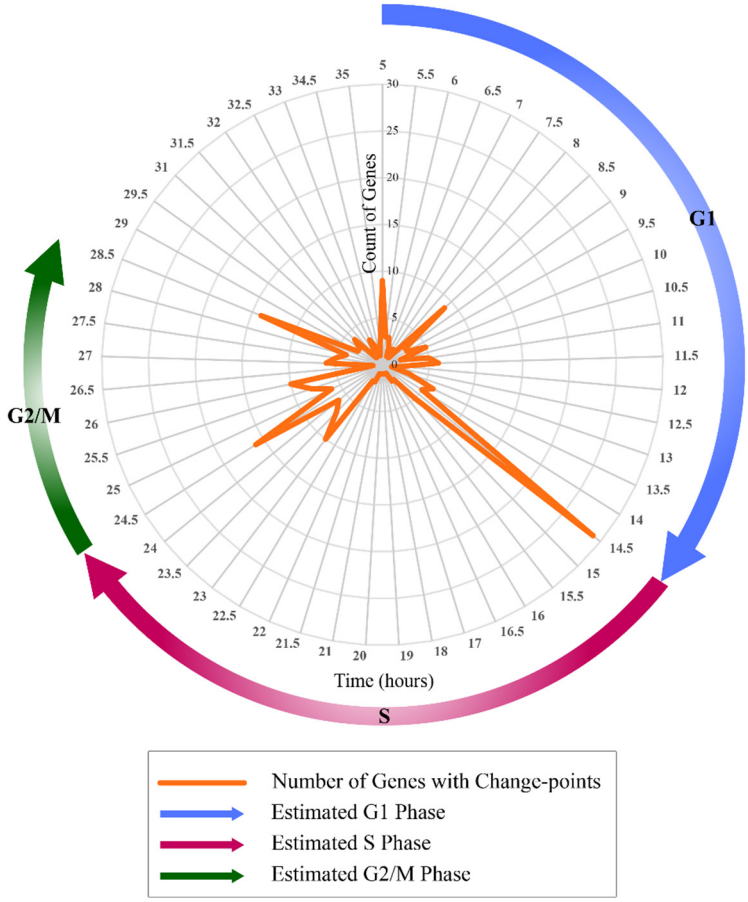



**Figure III.6 Segmentation of MEF cell cycle data with the change-point detection algorithm.** Radar chart displays the count of genes that were detected to have change points at every sample (1/2 *hour*) in the gene expression profiles of the 63 cell cycle genes.

### III.D.2 Network Reconstruction from Cell Cycle Time-series Data

After detection of the major temporal intervals associated with cell cycle phases, the successive directed graphs reflecting causal relationships of 63 cell cycle genes are reconstructed as the cell progresses through the G1, S and G2/M phases. In this work, the notion of Granger causality is used to predict directionality of links in the networks. Based on the definition of Granger causality, a series $X(t)$ is said to cause series $Y(t)$ if the future value of $Y(t)$ is better

predicted using the past values of $X(t)$ and $Y(t)$ than when the future value of $Y(t)$ is predicted using only the past values of itself (*10*). With the assumption that gene expressions may be modeled through a linear regression, one can identify Granger causality through Vector Autoregressive (VAR) models (see Materials and Methods section). A $d$-order VAR model of a $k$ dimensional time series is given by Equation III.8. Since the VAR model can be of any arbitrary order $1, 2, \ldots d$, the question of what the optimal order is arises. The optimal order of a variable $y_i(t)$ in the VAR model determines the number of time-lags that is necessary to take into account, in order to extract sufficient information from the lagged values of all variables that can provide the most accurate prediction of $y_i(t)$. This optimal order is estimated with the Minimum Description Length (MDL) principle (*19*). Here we compute the description length of the VAR model for each gene separately up to order $d_{max} = 9$. Figure III.7 shows the plot of the description length of four genes in the estimated G1 phase.

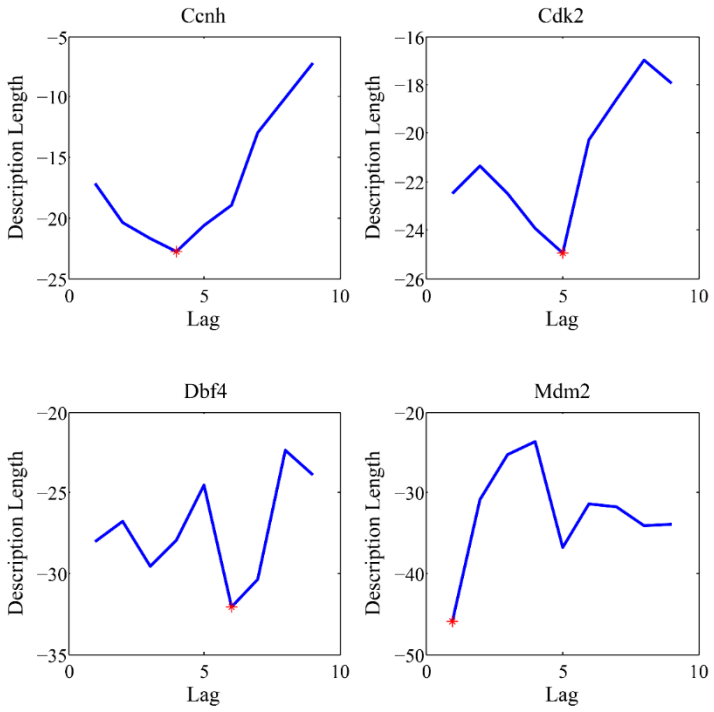



**Figure III.7 The plot of the description length for up to order $d_{max} = 9$ in the estimated G1 phase.** The optimal order, shown in a red asterisk, is the order at which the description length is minimized. As shown, the description length is minimized when the expression profiles of Ccnh, Cdk2, Dbf4 and Mdm2 are modeled through VAR models of order 4, 5, 6 and 1 respectively.

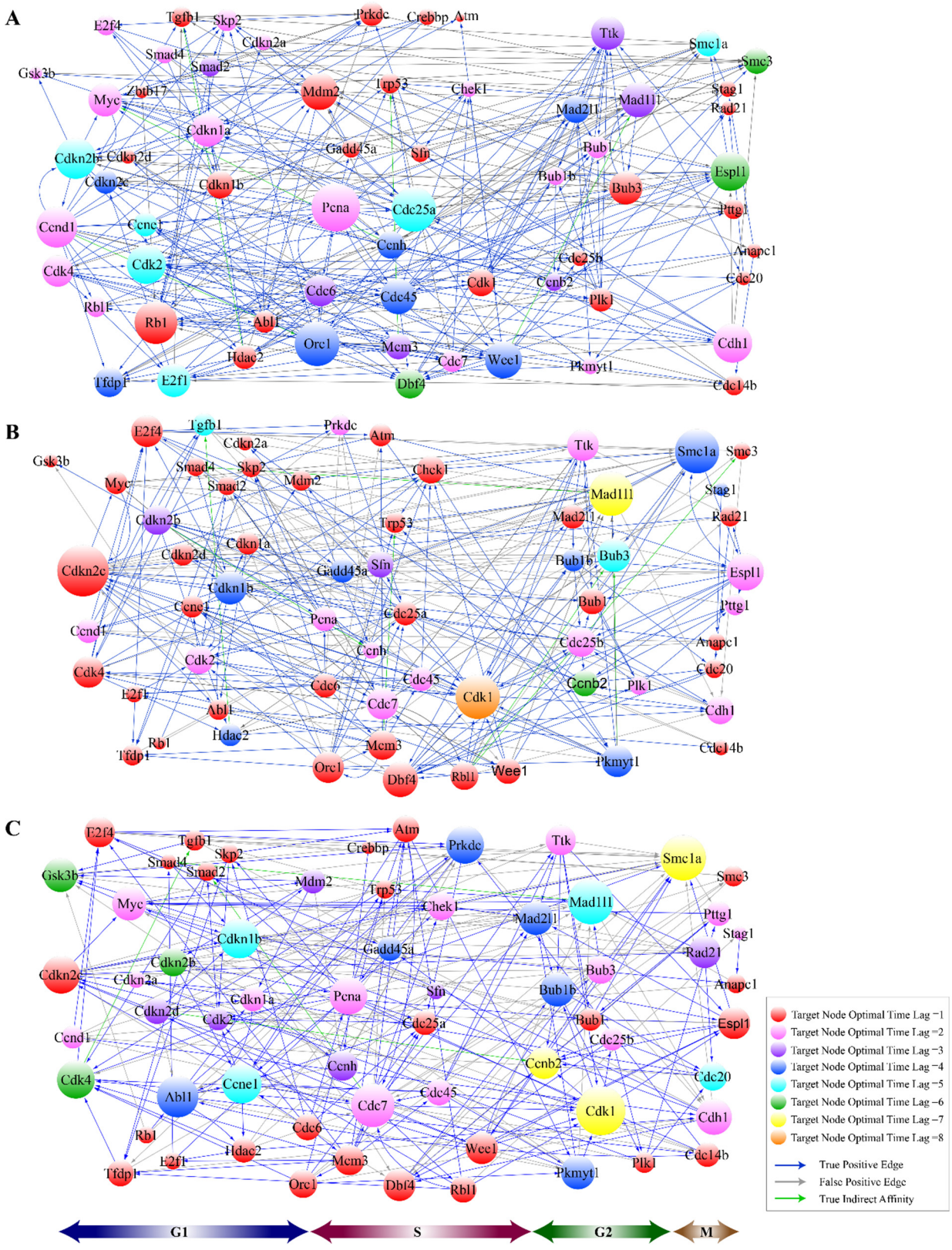
Once the optimal order for each gene is computed through MDL, we reconstruct three successive networks that reveal the evolution of the gene regulatory network of the 63 cell cycle genes through a complete cell cycle. Towards this, we use the expression profiles of genes for the three intervals [1-14.5], [1-24.5], and [1-28.5] *hours* derived through the CPD algorithm. Figure III.8.A depicts the gene regulatory network related to the [1-14.5] *hour* interval of the cell cycle associated with the G1 phase, Figure III.8.B shows the network reconstructed for the [1-24.5] *hour* interval associated with the G1 phase followed by the S phase, and Figure III.8.C illustrates the network representing the [1-28.5] *hour* interval related to the complete cell cycle (G1 and S phases followed by the G2/M phase). The resulting interactions have been validated with prior literature

and the interactions in the STRING database. Table III.2 presents the precision and false discovery rate of predictions in the reconstructed networks in Figures III.8.A, III.8.B and III.8.C.

**Table III.2 Statistics for the reconstructed network of the G1, S and G2 phases in Figure III.8.**

| Reconstructed Network | Number of true positive edges | Number of false positive edges | Precisions | False Discovery Rate |
|---|---|---|---|---|
| G1 phase | 268 | 76 | 0.78 | 0.22 |
| S phase | 198 | 78 | 0.72 | 0.28 |
| G2/M phase | 203 | 103 | 0.61 | 0.39 |

**Figure III.8. Time-varying cascade of the MEF cell cycle network for G1, S and G2/M phases. (A)** The graphic reconstruction of the network representing the causal interactions of 63 cell cycle genes obtained by using only the data samples in the interval [1-14.5] *hour* of the cell cycle associated with the G1 phase. **(B)** The network obtained by using only the data samples in the interval [1-24.5] *hour* of the cell cycle associated with the G1 phase followed by the S phase. **(C)** The network obtained by using the data samples in the interval [1-28.5] *hour* of the cell cycle associated with G1 and S phase followed by the G2/M phase. The blue edges represent the true positive (TP) connections validated though the known literature (STRING database). The green edges represent true indirect affinities between the pairs of genes they are connected to, and the gray edges are interactions captured in our model with no further evidence in the literature. The node colors denote the optimal time lag corresponding to every target gene in the VAR model.


A
B
C
Target Node Optimal Time Lag =1
Target Node Optimal Time Lag =2
Target Node Optimal Time Lag =3
Target Node Optimal Time Lag =4
Target Node Optimal Time Lag =5
Target Node Optimal Time Lag =6
Target Node Optimal Time Lag =7
Target Node Optimal Time Lag =8
True Positive Edge
False Positive Edge
True Indirect Affinity
G1
S
G2
M

### III.D.3 Temporal Dependence of Biological Processes in the Cell Cycle

In order to understand the temporal aspect of cell cycle processes, we analyze the transient length of influence of dynamic processes on one another; our primary question seeks to ask if one biological event induces the occurrence of another event in the cell, what is the duration of its influence? We sought to explore the temporal dependence of intracellular processes by considering 16 time-dependent biological processes governing the progression of the cell cycle. S3 Table shows these biological mechanisms listed in the chronological order of their occurrence during a cell cycle along with their members (genes) according to the Reactome pathway database (*141*). In the three successive networks in Figure III.8, we group cell cycle genes that belong to each of the 16 biological processes into modules and infer the temporal dependence of modules on one another. The temporal interdependences of these processes are assessed by taking into account the average of directed edge time-lags between pairs of processes. For instance, Figures III.9.A, III.9.B, and III.9.C display the links from the nodes in G1/S transition module to the nodes in the G2/M DNA replication checkpoint mechanism as the cell goes through the G1, S, and G2/M phases, respectively. The numbers labeling these links denote the optimal number of time-lags required in the VAR model when assessing Granger causality.

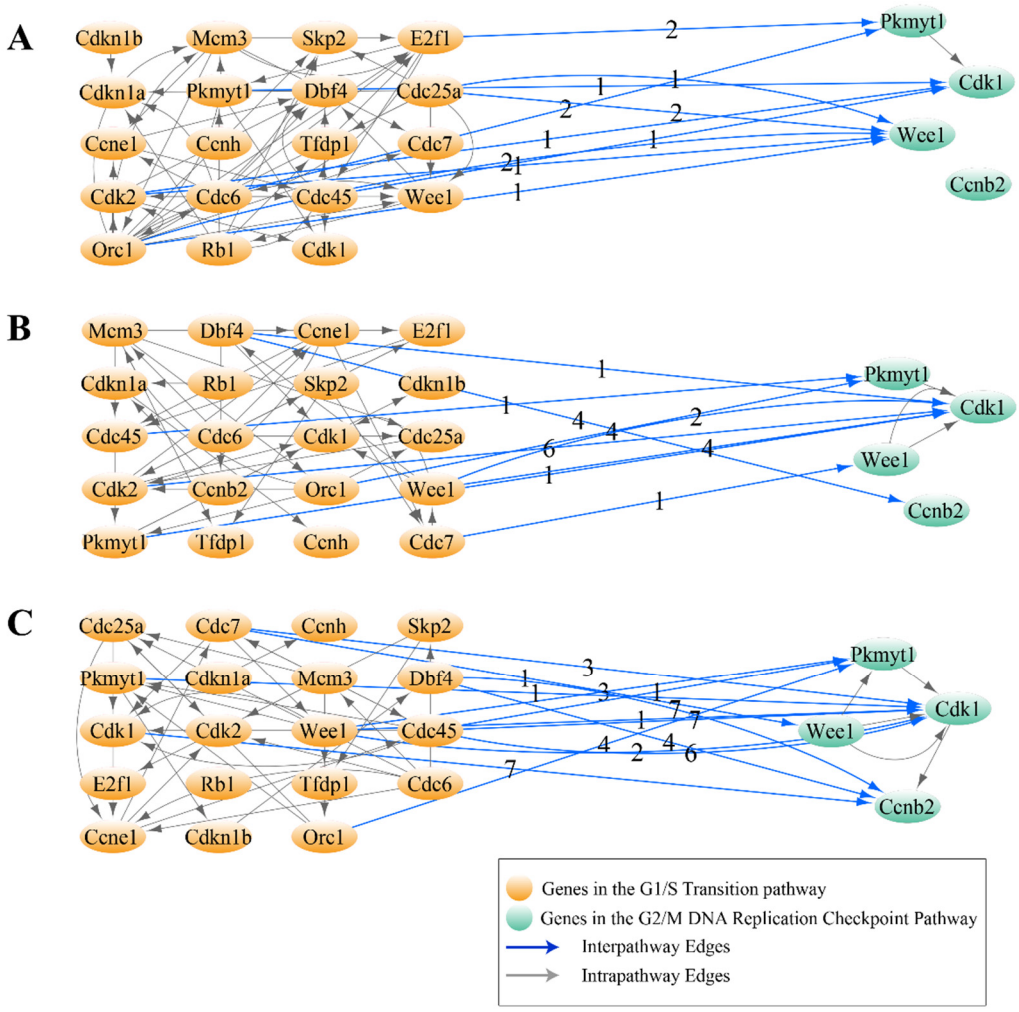



**Figure III.9 Temporal dependence of G2/M DNA replication checkpoint mechanism on the G1/S transition mechanism.** Orange nodes are genes that take part in G1/S transition mechanism of the cell cycle and the green nodes are genes that take part in G2/M DNA replication pathway. Every edge label denotes the temporal dependence of the target node on the source node. In this example, the farthest dependence is 7 time lags. **(A)** Temporal dependence of G2/M DNA replication pathway on the G1/S-transition pathway in the [1-14.5] *hour* interval. **(B)** Temporal dependence of G2/M DNA replication pathway on the G1/S-transition pathway in the [1-24.5] *hour* interval. **(C)** Temporal dependence of G2/M DNA replication pathway on the G1/S-transition pathway in the [1-28.5] *hour* interval.

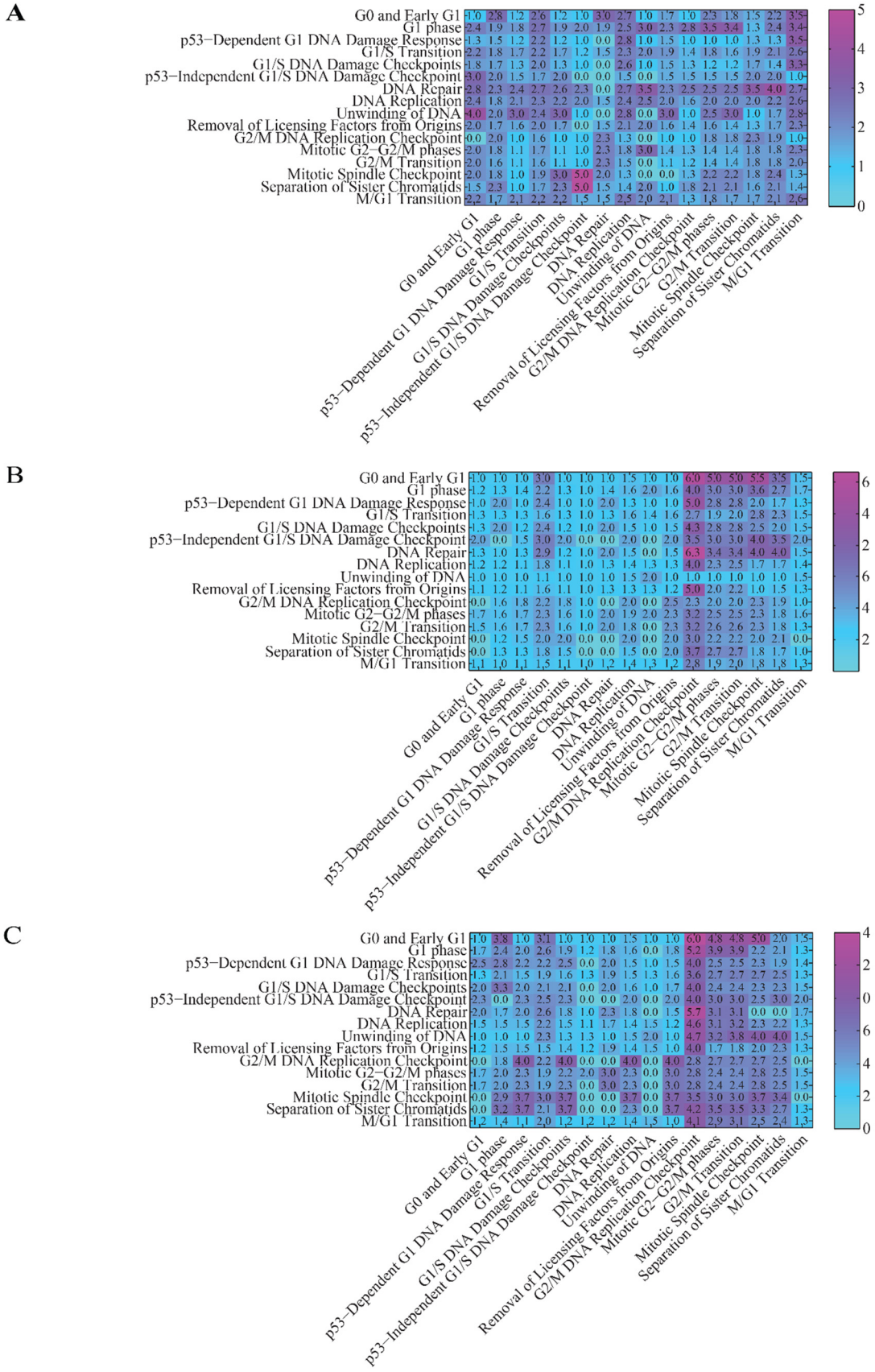
The average time-lag of edges in the three graphs in Figures III.9.A, III.9.B, and III.9.C are 1.4, 2.67, and 3.62 respectively. Here, as the cell evolves through a complete cell cycle, the average time-lag of the causal effect the G1/S transition mechanism has on the G2/M DNA replication mechanism increases. To further explore the length of intertwined temporal dependence these biological processes have on one another, we extend this analysis to all 16 intracellular processes listed in Table III.3.

Figures III.10.A, III.10.B, and III.10.C show the heat map plot displaying the average time-lag of edges between each pair of the 16 processes as the cell completes the G1, S, and G2/M phases. The heat map images identify temporal dependence of biological events on one another in different stages of the cell cycle.

**Table III.3. List of time-dependent biological processes according to the Reactome pathway database.**

| | Biological Process | Members (Genes) |
|---|---|---|
| **1** | G0 and Early G1 | Cdk2; E2f4; Rbl1; Tfdp1; Ccne1 |
| **2** | G1 Phase | Cdkn2b; Cdkn2c; E2f4; Skp2; E2f1; Cdkn2d; Ccnd1; Cdkn1b; Cdkn1a; Rb1; Cdkn2a; Cdk4; Ccnh; Tfdp1; Rbl1 |
| **3** | p53-Dependent G1 DNA Damage Response | Ccne1; Trp53; Cdkn1b; Cdkn1a; Atm; Cdk2; Mdm2 |
| **4** | G1/S Transition | Rb1; Skp2; E2f1; Cdkn1b; Cdc25a; Ccne1; Pkmyt1; Orc1; Cdk1; Cdk2; Dbf4; Tfdp1; Cdc45; Mcm3; Wee1; Ccnh; Cdkn1a; Cdc6; Cdc7 |
| **5** | G1/S DNA Damage Checkpoints | Ccne1; Trp53; Cdkn1b; Cdc25a; Atm; Cdk2; Mdm2; Cdkn1a; Chek1 |
| **6** | p53-Independent DNA Damage Response | Cdc25a; Atm; Chek1 |
| **7** | DNA Repair | Pcna; Atm; Ccnh; Prkdc |
| **8** | DNA Replication | Pcna; Cdkn1b; Cdkn1a; Rb1; Orc1; Cdk2; Dbf4; Cdc45; Mcm3; Cdc6; Cdc7 |
| **9** | Unwinding of DNA | Cdc45; Mcm3 |
| **10** | Removal of licensing factors from origins | Cdkn1b; Cdkn1a; Rb1; Orc1; Cdk2; Mcm3; Cdc6 |
| **11** | G2/M DNA replication checkpoint | Cdk1; Wee1; Pkmyt1; Ccnb2 |
| **12** | Mitotic G2-G2/M phases | Ccnb2; E2f1; Plk1; Cdc25b; Cdc25a; Pkmyt1; Cdk1; Cdk2; Wee1; Ccnh |
| **13** | G2/M Transition | Ccnb2; Plk1; Cdc25b; Cdc25a; Pkmyt1; Cdk1; Cdk2; Wee1; Ccnh |
| **14** | Mitotic Spindle Checkpoint | Mad2l1; Bub3; Bub1b; Mad1l1; Cdc20; Anapc1 |
| **15** | Separation of Sister Chromatids | Smc1a; Mad2l1; Bub1; Bub3; Bub1b; Mad1l1; Smc3; Stag1; Cdc20; Plk1; Espl1; Anapc1; Pttg1; Rad21 |
| **16** | M/G1 Transition | Orc1; Cdk2; Dbf4; Cdc45; Mcm3; Cdc6; Cdc7 |

**Figure III.10 Temporal interdependencies of biological processes as the cell goes through the G1, S and G2/M phases.** Each row and column in the heat map represents one of the 16 time-dependent biological processes. The number in every pixel represents the average time-lag of edges sourcing from its corresponding row process and targeting its column process (one lag is equivalent to ½ hour). **(A)** Heatmap of temporal dependence of processes as the cell goes through the G1 phase, **(B)** Heatmap of temporal dependence of processes as the cell goes through the G1 followed by the S phase. **(C)** Heatmap of temporal dependence of processes as the cell goes through the G1, S and G2/M phases.


A
G0 and Early G1
G1 phase
p53−Dependent G1 DNA Damage Response
G1/S Transition
G1/S DNA Damage Checkpoints
p53−Independent G1/S DNA Damage Checkpoint
DNA Repair
DNA Replication
Unwinding of DNA
Removal of Licensing Factors from Origins
G2/M DNA Replication Checkpoint
Mitotic G2−G2/M phases
G2/M Transition
Mitotic Spindle Checkpoint
Separation of Sister Chromatids
M/G1 Transition
B
G0 and Early G1
G1 phase
p53−Dependent G1 DNA Damage Response
G1/S Transition
G1/S DNA Damage Checkpoints
p53−Independent G1/S DNA Damage Checkpoint
DNA Repair
DNA Replication
Unwinding of DNA
Removal of Licensing Factors from Origins
G2/M DNA Replication Checkpoint
Mitotic G2−G2/M phases
G2/M Transition
Mitotic Spindle Checkpoint
Separation of Sister Chromatids
M/G1 Transition
C
G0 and Early G1
G1 phase
p53−Dependent G1 DNA Damage Response
G1/S Transition
G1/S DNA Damage Checkpoints
p53−Independent G1/S DNA Damage Checkpoint
DNA Repair
DNA Replication
Unwinding of DNA
Removal of Licensing Factors from Origins
G2/M DNA Replication Checkpoint
Mitotic G2−G2/M phases
G2/M Transition
Mitotic Spindle Checkpoint
Separation of Sister Chromatids
M/G1 Transition

### III.D.4 G1 Phase

The G1 phase, also known as the Gap 1 phase, is the first of the four phases that occur in one complete eukaryotic cell cycle. During the G1 phase, the cell grows in size and synthesizes mRNA and proteins required for DNA synthesis. In this section, we investigate the role of key regulatory proteins and their corresponding phase specific interactions found in the reconstructed G1 phase network (Figure III.8.A). The complete list of the edges estimated in the G1 phase network is presented in supplementary table G1_Phase_Interactions.xlsx.

**Rb1/Rbl1:** In Figure III.8.A, we note Rb1 interacts with Cdkn1a, Cdkn2a, Skp2, Cdh1, and Anapc1. It is known that Cdkn1a forms a physical complex with Rb1 and can activate Rb1 to bring about cell cycle arrest (*142, 143*). Furthermore, Rb1 activity is mainly regulated by Cdkn2a's inhibition of Ccnd1 to prevent phosphorylation of retinoblastoma (Rb) proteins, while Ccnd1 initiates the phosphorylation of Rb1 in mid-G1 phase (*144, 145*). Rb1 also physically interacts with Skp2 to inhibit Cdkn1b ubiquitination and induce G1 arrest (*146*). Further, Anapc1 and its activator Cdh1 interact with Rb1 and are required for Rb1-induced cell cycle arrest which leads to Rb1-induced accumulation of P27 (Cdkn1b) during G1 arrest (*147*). Detection of the Rb1→Abl1 edge is illustrated in Figure III.8.A. Rb1 is known to form a complex with Abl1 in the late-G1/early-S-phase as a result of its hyperphosphorylation by the cyclin-D/cdk4-6 complex (*148-150*).

The Rb1→Tfdp1 and Rbl1→E2f1 edges are captured in the reconstruction of the network representing G1 phase in Figure III.8.A. It is widely accepted that Rb1 and Rbl1 genes negatively regulate the G1/S transition of the cell cycle and enable cell growth by targeting key transcription factors, including E2Fs and transcription factor DP subunits (*151-153*). In addition, trans-

activation by the E2f1-Tfdp1 heterodimers is known to be inhibited by the retinoblastoma protein family (*154*).

**E2F1-4:** In Figure III.8.A, E2f1 is seen to interact with Mcm3, Cdc6, Orc1, and Cdc45. The E2F transcription factor upregulates the transcription of Mcm3 gene in the late G1 phase (*155, 156*). Besides the minichromosome maintenance complex (MCM) genes, Cdc6, ORC, and Cdc45 genes that are components of the pre-replication complex are well-known E2F-inducible genes during the late G1 and G1/S boundary in the cell cycle (*157-160*). The Tfdp1→ E2f1 interaction is also detected; it is widely established that Tfdp1 interacts and form heterodimers with E2f1 to regulate the cell cycle progression from G1 to S phase (*161-163*).

**Ccnd1/Cdk4:** We can note the Ccnd1-Cdkn2b and Cdk4-Cdkn1b interactions in Figure III.8.A. Cdkn2b can physically interact with and inhibit the activity of D-type cyclin dependent kinases and Cyclin D/CDK complexes while the Cip/Kip proteins, including Cdkn1a and Cdkn1b, can inhibit G1 CDKs such as Cdk4 (*144, 164-166*). We also see the Ccnd1→Rbl1 and Cdk4→Rbl1 interactions in Figure III.8.A. It is well-known that in late G1 phase, Cyclin D/Cdk4-6 complexes perform the main phosphorylation of Rbl1, a member of the retinoblastoma family, leading to dissociation of Rbl1 from Rb-E2F/DP complexes (*167-169*). Furthermore, the phosphorylation of Rbl1 by Cyclin D/Cdk4 complex inactivates Rbl1 to promote G1/S transition (*169*).

Ccnd1→E2f1 and Ccnd1→Tgfβ1 interactions are seen in Figure III.8.A. E2f1 is known to promote cell cycle progression through the induction of G1 phase cyclin, Cyclin D1 (*170, 171*). Tgfβ1 blocks the progression of cell cycle during G1 and this is associated with Tgfβ1 inhibition of Ccnd1 expression (*172*). We also note the Ccnd1→Cdh1 and Cdk4→Cdh1 interactions; Cdh1 is known to limit the accumulation of the G1 mitotic cyclin/CDK complexes to prevent pre-mature S-phase entry (*173*). Ccnd1→Ccne1 is also captured in Figure III.8.A. Analyses by Geng *et al.*

(1999) suggest that Cyclin E is a major downstream target of Cyclin D enabling the cell to progress through G1 and enter the S phase (*174*).

**Pre-Replicative Complex:** The Orc1↔Mamc3, Orc1→Cdc6 and Mcm3→Orc1 interactions are also seen in Figure III.8.A. According to multiple studies, in late mitosis and during G1 phase, Orc1 bound to replication origins recruits and serves as a platform for the assembly of Cdc6 followed by Mcm3 to form the pre-replicative complex (*175-178*). Orc1 interacts with Cdc6 throughout the G1 phase but not during other phases (*176*).

**Kip/Cip Cyclin Dependent Kinase Inhibitors (Cdkn1a, Cdkn1b, and Cdkn2a):** The Cdkn1b→Tgfβ1, Mdm2→Cdkn1a and Cdkn2a→Mdm2 regulatory links can be observed in Figure III.8.A. Tgfβ1 is reported to downregulate Cdkn1b during G1 phase (*179*) and Mdm2 has been shown to negatively regulate Cdkn1a and promote its proteasomal degradation which controls cell cycle progression during the G1 phase (*180, 181*). Several studies have shown that Cdkn2a physically interacts with Mdm2 to impede Mdm2-induced degradation of Trp53 and enhances Trp53 role in transcription and apoptosis (*182, 183*). This particular interaction stabilizes p53 and restores a p53-dependent G1 cell cycle arrest that is otherwise abrogated by MDM2 (*166, 184, 185*). See S1 Text for extended description of interactions.

**Myc:** In Figure III.8.A, we can see the connections Myc↔Cdc25a, Myc↔Cdkn2b, Myc←Cdkn1b and Crebbp→Myc. It is known that Cdc25a is capable of augmenting Myc-induced apoptosis in G1 (*186*). Myc represses cyclin dependent kinase inhibitors Cdkn2b during G1 arrest (*187, 188*) and takes part in Cdkn1b degradation (*189, 190*). Crebbp is known to regulate and stabilize Myc in G1 to prevent inappropriate S phase entry (*191, 192*). Furthermore, the Myc-Smad2 interaction is captured, while Myc is known to physically interact with Smad2 to inhibit TGF$\beta$ mediated induction of Cdkn2b in the G1 phase (*193*).

**Smad2-4:** We observe the Smad2→Cdkn1a, Smad4→Cdkn1b, Smad2→Rb1 and Skp2→Smad2 connections in Figure III.8.A (G1 phase). Studies show that Smad2 knockdown decreases Cdkn1a and releases G0/G1 arrest in mouse embryonic palate mesenchymal (MEPM) cells (*194, 195*). Also, loss of Smad4 as a tumor suppressor is associated with Cdkn1b downregulation and decreases Rb1 phosphorylation that results in G1-S transition and cell proliferation (*196*). A recent study shows that Smad2 overexpression results in an increase in Rb1, leading to cell cycle arrest at the G1 to S phase boundary (*197*). Liu *et al.* (2007) have shown that Tgf$\beta$-induced Skp2 degredation is mediated by the Smad cascade, thereby facilitating cell cycle arrest at the G1/S transition (*198*).

**Cyclin E/Cdk2:** The interaction of Ccne1 with Cdk2, Mcm3, Cdc45 and Cdc6 can be noted in Figure III.8.A. It is well-known that Ccne1 forms a complex with Cdk2, whose activity is required for the G1/S transition (*199*). Li *et al.* (2011) have indicated that Mcm3's phosphorylation by Cyclin E is involved in its loading onto the chromatin during G1 phase and before DNA replication (*200*) and that Cyclin E promotes chromatin loading of Cdc45 and phosphorylation of Cdc6 at the replication origins during the G1/S transition (*201, 202*). We can also notice the Cdk2→Trp53 interaction where it's been shown the activation of Trp53 tumor suppressor is required for Cdk2 phosphorylation and progression through G1 phase (*203, 204*).

**Pcna:** The interaction of Pcna with Gadd45a and Trp53 can be observed in Figure III.8.A. Multiple studies have shown that Gadd45a binds to and interacts with Pcna (*205-207*) and inhibits entry of cell into S phase (*208*). Furthermore, studies have shown that Trp53 mediates the activation of Pcna expression leading to arrest of cell growth at late G1 phase (*209-211*).

**Abl1 and Hdac2:** We can note the Abl1↔Mdm2, Bub3→ Hdac2 in Figure III.8.A. Research has revealed the role of Abl1 in phosphorylation of Mdm2 which neutralizes the

inhibitory effect of Mdm2 on Trp53 in response to DNA damage and stabilizes p53 in an active form (*212-214*). Yoon *et al.* (2004) have indicated that Bub3 directly interact with Hdac2 sauggesting that the Bub3–HDAC complexes are constituitively present thoughout G1 and G2 phases and may interact with Mad1l1 (*215*).

### III.D.5 S Phase

S (synthesis) phase is the second phase of the cell cycle occurring after the G1 phase and before the G2 phase in which DNA is replicated. Here we delve into the results for key S-phase proteins we obtained through our analysis (depicted in Figure III.8.B). The full list of the edges identified for S phase is presented in supplementary table S_Phase_Interactions.xlsx.

**Chek1:** We note the Chek1→Trp53 and Orc1→Chek1 edges in Figure III.8.B. It is well established that Chek1 regulates Trp53 activity during DNA damage-induced S and G2 phase arrests (*216-218*). Moreover, it has been extensively studied that cells with replicative initiation mutants defective in the Orc1 gene require the checkpoint kinase Chek1 during S phase to maintain cell viability by stabilizing DNA replication forks (*219-221*). One can note the interaction of Chek1 with Cdc45 and Cdk2 in Figure III.8.B. Cdc45 is a target of the Chek1-mediated S-phase checkpoint (*222, 223*). During the S-phase checkpoint, Chek1 activity increases which leads to Cdk2 inhibition and blockage of the S-phase transit in response to DNA damage (*224, 225*). We can further note that Chek1 interacts with Smc1a and Wee1 in Figure III.8.B. Syljuåsen *et al.* (2005) have shown that inhibition of Chek1 in S-phase cells triggers rapid phosphorylation of Smc1a, therefore suggesting a regulatory association between the two genes during S phase of the cell cycle to protect DNA breakage and promote DNA repair (*223*). Chek1 phosphorylates and positively regulates Wee1 in the DNA replication checkpoint (*226*) and in the G2 DNA damage

checkpoint (*227*). Additionally, Wee1 inhibition diminishes Chek1 phosphorylation in cells that are undergoing replicative stress (*228*).

**Atm:** We note the E2f4→Atm, Skp2→Atm and Cdc7→Atm edges in Figure III.8.B. E2F transcription factors not only regulate many genes required for entry into S phase, but also take part in DNA repair by transcriptionally regulating Atm (*229*). Wu *et al.* (2012) have examined the role of Skp2 in DNA damage response and repair by showing its recruitment and activation of Atm during DNA double-strand breaks (*230*). Cdc7, involved in initiation and progression of DNA replication during S phase, further plays role in DNA repair by activating the Atm/Atr-Chek1 checkpoint pathway (*231*).

**Trp53:** The interaction of Trp53 with Mcm3 and Orc1, both of which are key components of the pre-replicative complex, is shown in Figure III.8.B. Trp53 controls the initiation of replication and entry into S phase by regulating proliferation related genes such as Mcm3, Orc1, and Cdc6 (*232, 233*). Furthermore, the Pkmyt1→Trp53 interaction has been detected in the reconstruction of the S phase regulatory network. Price *et al.* (2002) have shown that Pkmyt1 can negatively regulate Trp53-induced apoptosis in response to DNA damage in the S phase or the G2 phase (*234*).

**Mdm2:** The Cdk1→Mdm2 and Ttk→Mdm2 interactions can be seen in Figure III.8.B. Mdm2 is known to be phosphorylated by Cyclin A-Cdk1 complexes at the onset of S phase to reduce its interaction with Trp53 (*235*). Moreover, Ttk phosphorylates Mdm2 which facilitates oxidative DNA damage repair and cell survival during the S-phase (*236*).

**Pre-replicative complex**: We can see the interaction of Mcm3 with Cdc45 in Figure III.8.B. Mcm3 and Cdc45, both interacting components of the pre-replicative complex (*237-239*),

are known to dissociate from the origin DNA and associate with non-origin DNA and move with replication forks at the beginning of S phase (*240, 241*). In addition, Cdc45 loading onto the chromatin in the S phase is required to activate the helicase activity of the MCM complex (*242, 243*). We further note the Cdc6→Cdk2 edge; Cdc6 has been shown to activate Cdk2 to initiate DNA replication and G1-S phase progression (*202, 244*). Cdc6 is also known to activate Cdk2 to prevent re-replication during S and G2 phases (*245*). Dbf4→Cdk1 can be seen in Figure III.8.B; Cdk1 is known to target the Dbf4-Cdc7 kinase at the end of S phase to prevent re-replication in G2/M (*246, 247*).

### III.D.6 G2/M Phase

G2 phase is the third phase of the cell cycle in which the cell rapidly grows, protein synthesis occurs, and the cell prepares to enter mitosis. During mitosis, the replicated chromosomes are separated into two nuclei and the cell is divided into two daughter cells. Supplementary table G2M_Phase_Interactions.xlsx, consists of the entire list of interactions estimated in reconstruction of the G2/M phases. In this section, we investigate the main G2/M signaling pathways predicted in our study (shown in Figure III.8.C).

**Ttk:** We note the Ttk→Bub1, Ttk→Mad2l1, and Ttk→Bub1b interactions in Figure III.8.C. Studies have revealed that Mph1 (Ttk homologue), which localizes to the kinetochores only at prometaphase (second phase of mitosis), is required for the recruitment of Bub1 and other spindle assembly checkpoint components (*248, 249*). Ttk promotes closed Mad2l1 production and subsequent assembly of the mitotic checkpoint complex (MCC) to activate the spindle checkpoint assembly (*250*). Huang *et al.* (2008) have reported that Ttk is one of the major kinases required for Bub1b phosphorylation which is essential for the mitotic checkpoint and also for kinetochores to establish microtubule attachments during G2/M (*251*).

**Mad2l1-Mad1l1:** The Espl1→Mad2l1, Mad2l1→Bub1b, Bub3→Mad1l1, and Rad21→Mad1l1 edges can be seen in Figure III.8.C. The Espl1-Mad2l1 interaction has been confirmed as a regulatory mechanism required for sister chromatid segregation (*252*). Further, the spindle assembly checkpoint components Mad2l1 and Bub1b are known to act cooperatively to assemble the mitotic checkpoint complex and to prevent premature chromatid separation at the mitotic checkpoint (*253-255*). Multiple studies have indicated that Mad1l1 forms a complex with Bub3 during the cell cycle and is crucial for spindle checkpoint function (*256-258*). There is also evidence that knockdown of MAD proteins is correlated with Rad21 cleavage to promote sister chromatid segregation (*259*).

**Bub1b-Bub1-Bub3:** The Bub1b→Cdc20 and Bub1b→Plk1 edges can be seen in Figure III.8.C. Studies have shown that a checkpoint function of Bub1b is to inhibit the activity of Anaphase Promoting Complex (APC/C) by blocking the binding of Cdc20 to APC/C (*260-262*). Furthermore, Bub1b binds to Cdc20 to inhibit APC activity in interphase, allowing the accumulation of Cyclin B in G2 phase prior to M-phase entry (*263*). Bub1b localizes to centrosomes and suppresses centrosome amplification via regulating Plk1 activity during interphase (*264*). In addition, Bub1b brings about the action of Plk1 at kinetochores for appropriate chromosome alignment during prometaphase (*265*).

**Cdk1:** We can see the interaction of Cdk1 with Bub1b and Rbl1 in Figure III.8.C. Phosphorylation of Bub1b by Cdk1 is required for mitotic spindle checkpoint arrest and promotes the formation of the kinetochore during G2/M (*266*). It has been widely reported that Cdk1 phosphorylates pRB (retinoblastoma protein) in mitotic cells (*150, 267, 268*), while our model captures the interaction of Cdk1 with the pRB-related protein, Rbl1.

The network of Figure III.8.C depicts the edges Cdk1→Ccnb2, Wee1→Cdk1, and Pkmyt1→Cdk1. B-type cyclins form a complex with Cdk1 and this complex accumulates through late S and G2 phases of the cell cycle (*269*) and the activation of the Cyclin B-Cdk1 kinase is needed for entry into the G2/M phase (*270, 271*). It is widely accepted that Cdk1 activity is regulated through its inhibitory phosphorylation by Wee1 and Pkmyt1, leading to activation of the G2/M arrest which prevents premature entry into mitosis (*272-275*).

**Ccnb2:** We can see the Cdc20→Ccnb2 and Cdc25b→Ccnb2 edges in Figure III.8.C. It is known that APC/C-Cdc20 interaction can mediate cyclin B degradation which consequently prevents Cdk1 activity from reaching excessively high levels (*276*) and that the spindle assembly checkpoint acts on Cdc20 to block the degradation of Cyclin B during metaphase (*277*). The Cdc25 phosphatases are known to dephosphorylate and therefore activate the Cdk1-Cyclin B complexes (*278-280*).

**Espl1:** The Espl1→Ccnb2, Cdk1→Espl1, Espl1→Smc1a, and Espl1-Bub1 interactions are shown in Figure III.8.C. Espl1 binds to Cyclin B during anaphase, a required step in anaphase to shut down Cdk1 activity, to achieve abrupt and simultaneous separation of sister chromatids (*281-283*). It is widely accepted that Espl1 triggers anaphase (fourth phase of mitosis) by initiating cleavage of cohesin multiprotein complex which includes the Smc1a subunit (*284*). Studies have determined the role of Bub1 in the timing of Espl1 activation and hence regulation of anaphase (*285, 286*).

## III.E Discussion and Conclusion

Mammalian cell cycle is a dynamic process orchestrated by the activation of distinct molecular players across time. Canonical characterization of the cell cycle as a static network fails

to provide temporal mechanistic insights on the control exerted by the proteins during different phases of the cell cycle. In this study, we use an exhaustive and fine-grained time series expression dataset capturing the cell cycle of MEF primary cells to develop a temporally evolving dynamical network for the cell cycle progression. Using a set of 63 key cell cycle genes, we show that our causality-driven approach provides a temporal map of the phases of the cell cycle.

The mechanistic changes in the RNA-seq time-course data are identified by a change point detection algorithm which enables us to infer the timing of cell cycle phases and their duration with no prior biological knowledge. Through our computational analysis, the G1, S, and G2/M phases are estimated to be 14.5 hours, 10 hours, and 4 hours long, respectively. For a typical proliferating mammalian cell with an average cycle span of 24 hours, G1 phase lasts about 11 hours, S phase about 8 hours, G2 phase about 3-4 hours, and M phase about one hour (*287*). However, cell cycle duration varies from one cell type to another; for instance, the average phase duration for the rat embryo PC12 cell line when serum starved for 24 hours and then serum treated for 37 hours, is roughly 15 hours, 13.3 hours, and 4 hours for the G1, S, and G2/M phases, respectively (*288*), whereas reports show that the average cell cycle length for MEF cell line is 25.3 hours (*289, 290*).

The three successive directed graphs depicted in Figure III.8, representing the interaction of cell cycle genes as the cell evolves through the G1, S and G2/M phases of the cell cycle, are derived by utilizing the notion of Granger causality identified by a VAR model. This enables us to detect the main regulatory pathways and checkpoints essential to cell cycle regulation and reconstruct phase-specific gene regulatory networks at each stage of the cell cycle. Moreover, this approach allows for the inference of temporal length of influences each gene has on others. The temporal dependencies are obtained by estimating the optimal order of the VAR model that reveals

the sufficient number of lags required to extract useful past information that may influence the expression of other genes. Figure III.11 shows the three successive networks with key regulatory interactions that have been detected in Figure III.8 for the networks capturing the G1, S, and G2/M phases.

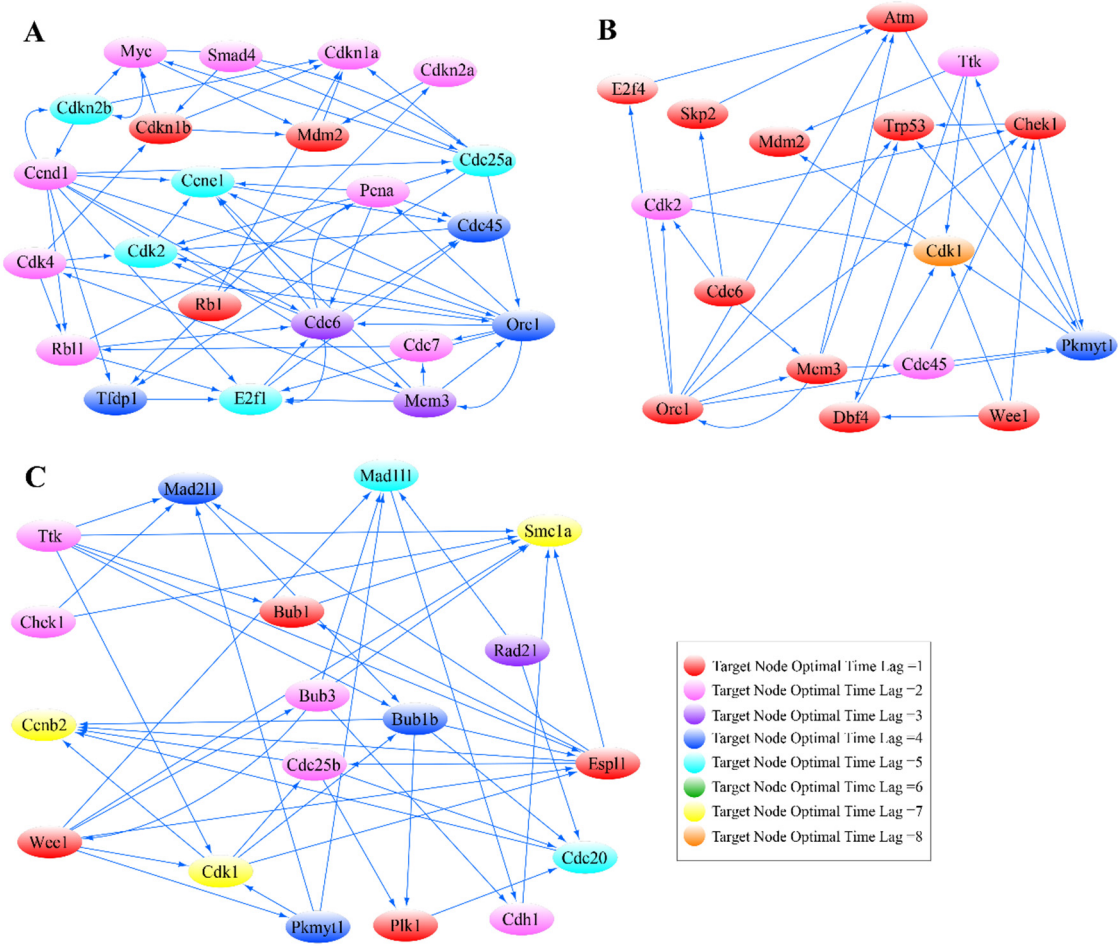



**Figure III.11 Key signaling pathways captured in the G1, S and G2/M phases of the cell cycle.** Diagram showing a subset of genes from the reconstructed networks in Figure III.8 depicting the key phase-specific regulatory interactions in the (A) G1 phase, (B) G1 phase followed by the S phase and (C) G1 and S phases followed by the G2/M phase.

Among the key G1 phase mechanisms (Figure III.11.A), we were able to detect the regulation of Rbl1 by Ccnd1 and Cdk4 as a promoting factor in the G1/S transition (*169*), the role of the retinoblastoma protein in enabling cell growth by targeting E2f and DP transcription factors

(*151-153*), as well as the function of cyclin dependent kinase inhibitors in inducing growth arrest in the G1 phase (*165, 166*). The Cdkn2a-Mdm2 interaction which stabilizes the tumor suppressor protein Trp53 (*185*), the Cyclin E-Cdk2 interaction required for G1/S transition (*199*), as well as Ccne1's role in the loading of Mcm3 and Cdc45 onto the chromatin (*200-202*) were detected in our reconstruction of the G1 phase network. We were able to detect the recruitment and assembly of Mcm3 and Cdc6 by Orc1 leading to the formation of the pre-replication complex and its assembly onto replication origins prior to S phase (*175, 177*).

The G1 phase events prepare the cell to initiate DNA replication in the S phase of the cell cycle. Regulated and monitored replication ensures the duplication of the entire genome in a timely fashion. The pre-replicative complex is assembled onto each origin prior to S phase and creates licensed origins that can initiate replication by origin firing. Once the cell transitions from G1 phase to the S phase, the licensed origin are converted into active replication forks (*291, 292*).

Major S-phase regulatory pathways are shown in Figure III.11.B. The loading of the replicative polymerases through Mcm3's recruitment of Cdc45 (*293*), along with the intra S-phase checkpoint exerted by Chek1's targeting of Cdc45 and regulation of Cdk2 (*222, 224*), are among the major S-phase pathways. The network in Figure III.11.B further describes the role of Chek1 in stabilizing the replication forks and protecting against DNA breakage through its interaction with Orc1 and Smc1a (*220, 223*).

During S-phase, Trp53 is involved in regulating initiation of replication by targeting replication-related genes Cdc6, Orc1, and Mcm3 (*232*). Furthermore, we detected Cdk1's role in preventing re-replication during S phase by regulating Dbf4 (*247*), along with the function of Atm in regulation of DNA damage and DNA repair, captured through Atm's interaction with Dbf4 and Skp2 (*229, 230*).

Figure III.11.C represents significant regulatory pathways characterized in the G2/M phases of the cell cycle such as spindle assembly checkpoint (SAC), mitotic checkpoint assembly, and chromosome segregation. Among these pathways, we detected the formation of mitotic checkpoint complex and the establishment of microtubule attachments during G2/M phases through the function of Ttk-Mad2l1 and Ttk-Bub1 interactions respectively (*249, 251*). Moreover, our proposed model for the G2/M phase identifies the Bub3-Mad1l1 interaction essential for spindle checkpoint function (*256-258*), the cooperative interaction of Mad2l1 and Bub1b that is required for prevention of premature sister chromatid segregation (*254*), as well as Mad2l1-Chek1 interaction which ensures fidelity of mitotic segregation (*294*). In addition, we detected the interactions suggesting the blockage of premature entry into mitosis through Cdk1's phosphorylation by Wee1 and Pkmyt1 (*272-275*). It is interesting that Cdk1's phosphorylation not only happens at early G2 phase, but may also occur during late S phase (*275*) as shown in Figure III.11.B and Figure III.11.C. We detected Cdk1's role in preventing re-replication during G2/M by targeting Cdc7 (*246*), along with the activation of the Cdk1-Cyclin B complex required for G2/M entry (*270*). Additionally, we spotted Plk's regulation of Cdc20 which activates the anaphase promoting complex, triggering the separation of sister chromatids (*295*), Plk1's role in mitotic exit through its interaction with Cdc25b (*296*), as well as the concurrent and abrupt segregation of sister chromatids through the Espl1-Ccnb2-Cdk1 pathway (*281, 283*) (Figure III.11.C).

It is worth noting that some interactions that are described in the literature as specific to certain phases may be found in other phases of the cell cycle as well. For instance, Mad1l1-Bub3 which is specific to G2/M, was also captured in G1 and S phase reconstruction. This is due to the

fact that such complexes/interactions are present throughout the cell cycle but exist at significantly higher levels during the phases they are generally known for (*256*).

In summary, we reconstruct causal mechanisms and networks across time during a mammalian cell cycle. While our reconstruction is based on using the transcriptome, and there could be differences between the transcriptome and the proteome abundances (*297-300*), we believe that the broad conclusions are substantiated by mechanisms reported in the literature. For example, studies have revealed that certain classes of genes, such as cell cycle genes, have higher correlation of mRNA expression with the corresponding protein expression across a large number of genes (*301, 302*), validating our use of the transcriptome across time to investigate the cell cycle. Through our integrative framework, we are able to provide insights into the temporal behavior of the MEF cell cycle describing information such as duration of cell cycle phases, identification of phase specific regulatory networks, and detection of key regulatory interactions essential to passage of the cell through cell cycle checkpoints. Moreover, the utilization of higher order VAR models lead to determining the temporal dependencies between multiple biological pathways in the three successive cell cycle regimes. The causal and temporally-dependent pathways also point to potential temporally specific perturbations and potential therapeutic targets that can help with repairing aberrant cell cycle mechanisms associated with pathologies (*300*).

### III.F Summary

Causal molecular mechanisms in cellular functions can only be inferred from temporal and longitudinal measurements. Few methods exist for analyzing time series data to identify distinct temporal regimes and the corresponding time-varying causal networks and mechanisms. In this study, we have developed an integrative framework that allows the detection of distinct temporal regimes, along with temporally evolving directed networks that provide a comprehensive picture

of the crosstalk among different molecular components (nodes) in each regime. We have applied our approach to RNA-Seq time-course data spanning nearly two cell cycles from Mouse Embryonic Fibroblast (MEF) primary cells. This strategy enabled us to, without any prior knowledge, extract information on duration and timing of cell cycle phases, phase-specific causal interaction of cell cycle genes as well as temporal interdependencies of biological mechanisms through a complete cell cycle. Our inference of dynamic interplay of multiple intracellular mechanisms can be used to predict time-varying cellular reponses and to explore the effect of drug dose and timing in therapeutic interventions.

## III.G Acknowledgements

Chapter III, in full, has been submitted for publication of the material as it may appear in Time Varying Causal Network Reconstruction of a Mouse Cell Cycle, 2018, Masnadi-Shirazi, Maryam; Maurya, Mano R.; Pao, Gerald; Ke, Eugene; Verma, Inder; Subramaniam, Shankar., PLOS Computational Biology, 2018. The dissertation author was a primary investigator and author of this paper.

Chapter IV

# Multiview Radial Basis Function Network: A New Approach on Nonparametric Forecasting of Chaotic Dynamic Systems

## IV.A Abstract

The curse of dimensionality has long been a hurdle in the analysis of complex data in areas such as computational biology, ecology and econometrics. In this work, we present a forecasting algorithm that exploits the dimensionality of data in a nonparametric autoregressive framework. The main idea is that the dynamics of a chaotic dynamical system consisting of multiple time-series can be reconstructed using a combinations of multiple variables. This nonlinear autoregressive algorithm uses attractors reconstructed from a combination of variables as the inputs of a neural network to predict the future. We show that our approach, multiview radial basis function network (MV-RBFN) provides a better forecast than that obtained using a model-free approach, multiview embedding (MVE). We demonstrate this for simulated ecosystems and a mesocosm experiment. By taking advantage of dimensionality, we show that MV-RBFN overcomes the shortcomings of noisy and short time-series.

## IV.B Introduction

In recent years, the availability of large time-course datasets in multiple disciplines, including biology, ecology and finance has brought forth the problem of handling such data for scientific analysis (*27-29*). In many studies, generalized linear models and vector autoregressive models are used for structural estimation and inference, where such systems exhibit nonlinear dynamics with time lags, reciprocal feedback loops and unpredictable surprises (*7, 30*). On the other hand, equation-based models such as difference and differential equations may be used to analyze the evolution of a dynamic system, but often require some degree of prior knowledge about the nature of interactions among various system components (*8*), or even if the model structure is known, dimensionality poses a challenge on accurate parameter estimation of variables (*31*). Furthermore, prior work has established that ecological and biological models are often

ineffective in predicting the future due to the highly nonlinear nature of component interactions (*32, 33*).

An alternative equation-free approach suitable for non-equilibrium dynamics (including chaos) and nonlinearity is state space reconstruction (SSR) which is a model-free approach in the sense that there is no analytic formula assumption thus allowing substantial flexibility in the nonlinearity of the system (*9, 34*). SSR uses lagged coordinate embeddings to reconstruct attractors that map the time-series evolution from time domain into state space trajectories. In a notable theorem, Takens proved that the overall behavior of a chaotic dynamic system can be reconstructed from lags of a single variable (*35*). Later Takens' theorem was generalized and it was demonstrated that the information from a combination of multiple time-series (and their lags) can be used in an attractor reconstruction to provide a more mechanistic model (*36, 37*). Nonetheless, since attractor reconstruction relies only on experimental data, the limitations of short or noisy time-series restricts the ability to infer system dynamics as a whole. Namely, SSR from short time series provide a scarce view of a system's mechanism, diminishing reliability of inferences. In addition, when time-series data is corrupted with observational noise, data may become meaningless and irrelevant in providing useful information for predictability. Ye *et al.* (2016) introduced an analytical approach, multiview embedding (MVE), which harnesses the complexity of short and noisy ecological time-series as a way to improve forecasting (*38*). MVE is a method based on nearest neighbors that looks into the predictability of all possible manifold reconstructions using the method of simplex projection (*34*). In this work, we treat prediction of the dynamical system as an inverse problem that involves interpolation and approximating an unknown function from time series data. Instead of relying on single nearest neighbors of the top attractor reconstructions as carried out in MVE, here we introduce a multiview radial basis function network (MV-RBFN)

autoregressive model that calculates a distance-weighted average of all points in the top manifold reconstructions through a nonlinear kernel estimation method. Similar to MVE, attractors from combinations of variables and their lags are reconstructed. Each manifold (view) comprises information that is particular to that embedding. By ranking the reconstructed manifolds according to their forecast skill (prediction errors), and merging the top views and the information contained in them, MV-RBFN is capable of recovering the dynamics of the system in a manner that outperforms MVE and nonlinear univariate and multivariate autoregressive models.

## IV.C Materials and Methods

### IV.C.1 Multiview Radial Basis Function Network (MV-RBFN)

MV-RBFN utilizes radial basis function networks (RBFN) initially proposed to perform accurate interpolation of data points in the multidimensional space (*303*). Suppose we are interested in forecasting variable $y$ in a three-species food chain with components $x$, $y$, and $z$. By constructing the attractors from combination of variables of the three-species food chain, one can look into the forecast skill of each multivariate manifold (Figure IV.1). For example, the blue manifold in Figure IV.1.A is an embedding constructed from variables $z$, $y$, and variable $x$ delayed by two time lags. Each multivariate embedding in Figure IV.1 is mapped using a Gaussian RBFN which approximates a nonlinear function that transforms the input space of past values in each manifold to the output space of future target values:

$$Y_i(t+1) = \Psi(M_i)\alpha_i + \varepsilon_i, \quad i = 1, 2, \dots m \tag{IV.1}$$

$\Psi(M_i)$ is a data matrix of nonlinear Gaussian kernel functions with the inputs being points on the $i^{th}$ manifold $M_i$, and $\alpha_i$ is a $p$ dimensional vector of output weights that can be fixed such that the prediction error is minimized in the minimum mean squared error sense. $p$ is the number

of centers in each manifold that can be chosen through a k-means clustering algorithm and $m$ is the number of all possible manifold reconstructions from a combination of variables and their time lags. Given $N$ variables and $L$ lags for each variable, the possible number of reconstructions in an $E$-dimensional space grows combinatorially:

$$m = \binom{NL}{E} - \binom{N(L-1)}{E} \tag{IV.2}$$

where, the first term is the number of manifolds formed by choosing $E$ of the $NL$ possible coordinates, and the second term is subtracted to account for the number of unacceptable manifolds where all $E$ coordinates are lagged; an acceptable manifold is one with at least one coordinate at the current time $t$.

The black manifolds in Figure IV.1 are reconstructed from the actual future observations of variable $y$ and the red dots are the predicted values. One can rank constructed embeddings based on their prediction accuracy (mean absolute error or correlation between observation and predictions) from the best (Figure IV.1.A) to the worst (Figure IV.1.C).

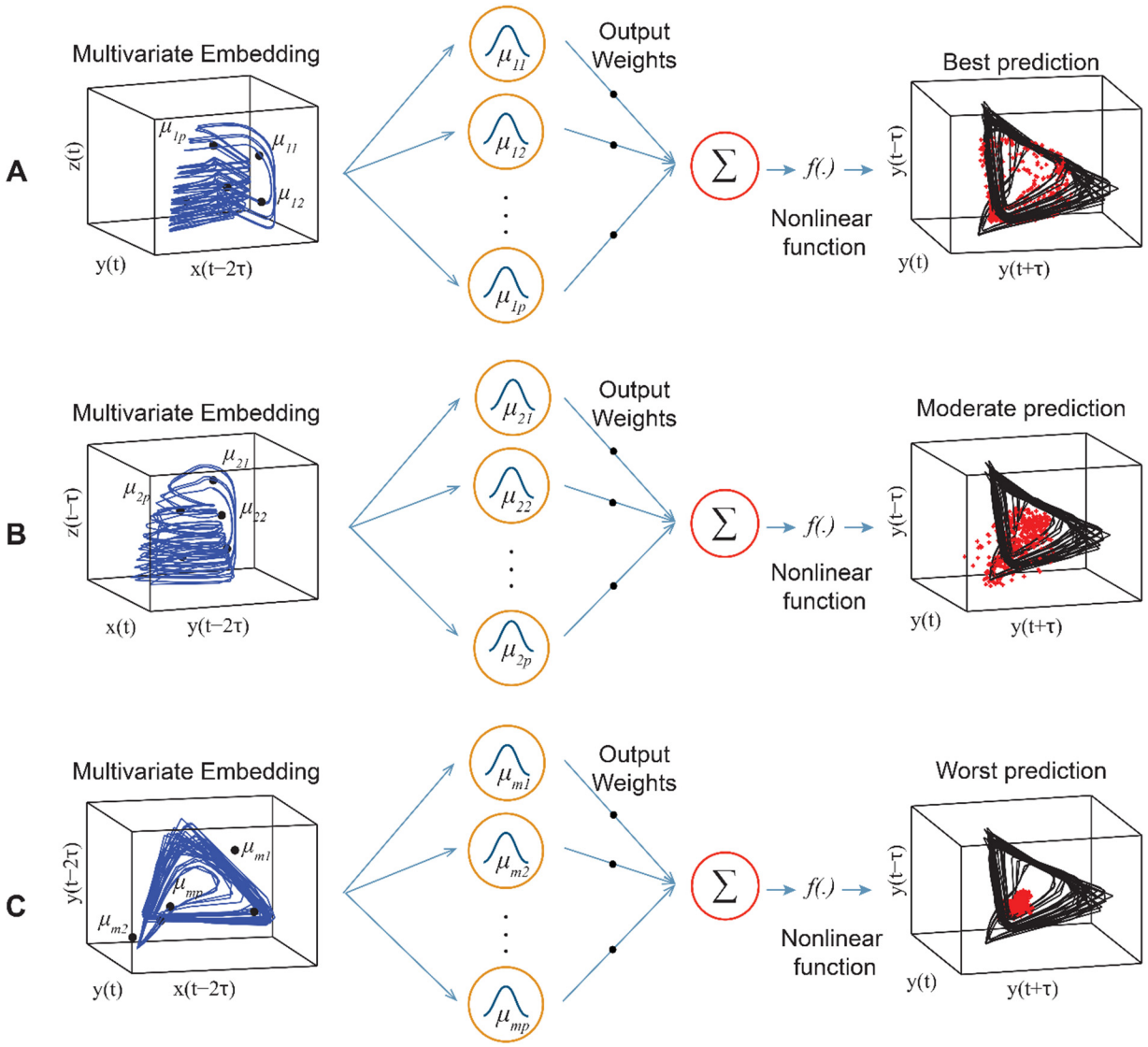



**Figure IV.1 Schematic showing forecast skill of multivariate embeddings in the three-species food chain model.** **(A)** Multivariate embedding reconstructed from $z(t)$, $y(t)$ and $x(t-2\tau)$ in 3-dimensional space provides the best forecast of variable $y$ using Gaussian radial basis functions with centers $\{\mu_{11}, \ldots, \mu_{1p}\}$.**(B)** Multivariate embedding reconstructed from $z(t-\tau)$, $x(t)$ and $y(t-2\tau)$ in 3-dimensional space provides moderate forecast of variable $y$ using Gaussian radial basis functions with centers $\{\mu_{21}, \ldots, \mu_{2p}\}$. **(C)** Multivariate embedding reconstructed from $y(t-2\tau)$, $y(t)$ and $x(t-2\tau)$ in 3-dimensional space provides the worst forecast of variable $y$ using Gaussian radial basis functions with centers $\{\mu_{m1}, \ldots, \mu_{mp}\}$.

Once all reconstructions are ordered based on their prediction skill in the in-sample portion of the data, one can identify the top $k$ manifold $M_1, \ldots M_k$ in an $E$-dimensional space that will further be used in the MV-RBFN forecast of the out-of-sample portion of the data. Figure IV.2.A

shows that the inputs of the MV-RBFN model are the top $k$ manifolds in the prediction of variable $y$ that are fed into the three-layer neural network. Each node in the hidden layer uses a Gaussian RBF with centers $\{\mu_{l\rho}\}_{\rho=1}^{p}, l = 1, \dots, k$ as nonlinear activation functions. The one-step forecast of $y$ through the multiview RBFN, and the actual one-step-ahead observations of $y$ are shown in Figure IV.2.B in the red and black curves respectively.

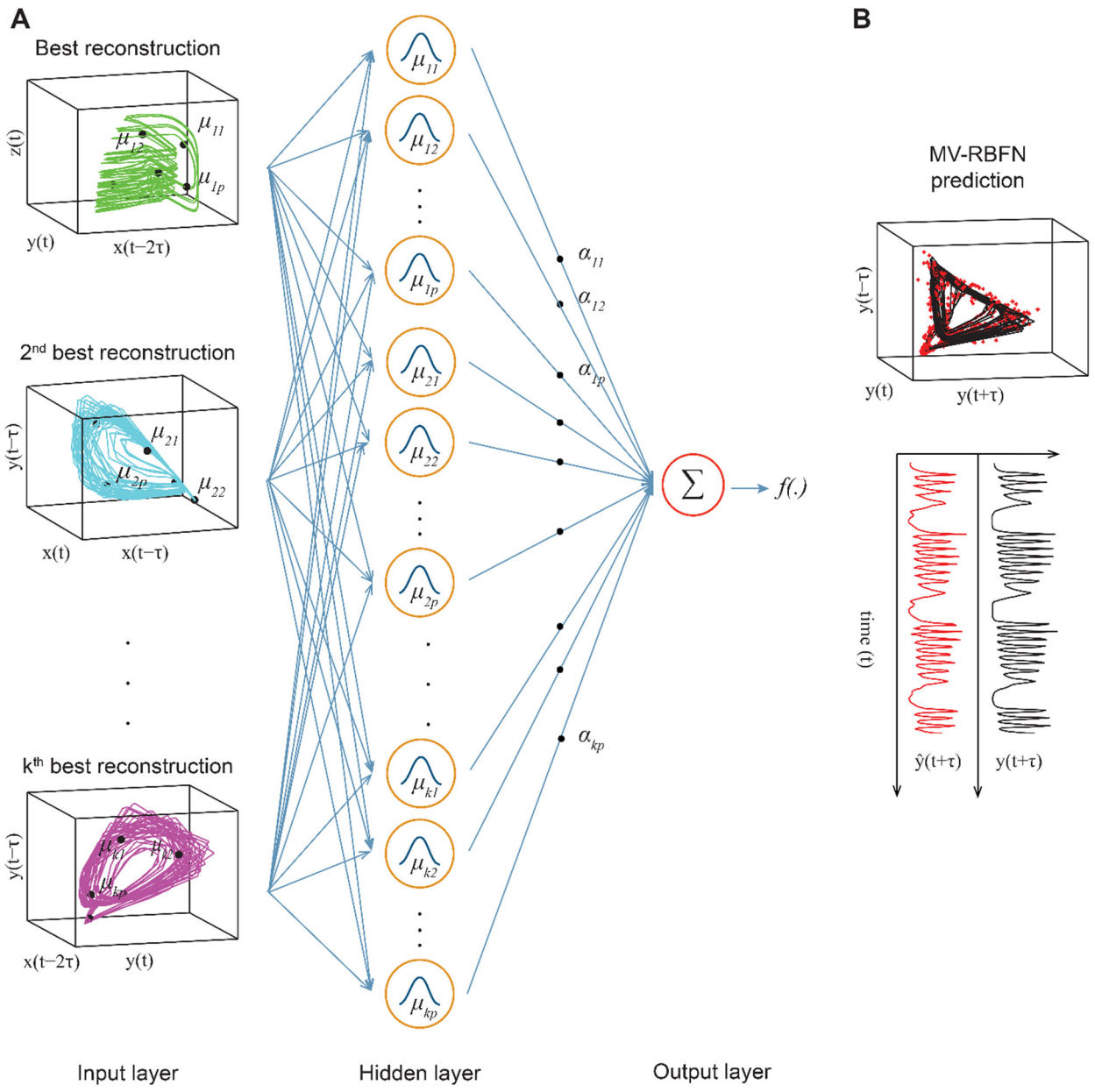



**Figure IV.2 Multiview radial basis function network. (A)** Three-layer neural network takes the best $k$ predictive embeddings as its inputs. The nonlinear function $f(.)$ is estimated by fixing the $\alpha$ weights through linear optimization. **(B)** The predicted forecast and future observation are shown by the red and black curve (manifold) in time domain (in state space) respectively.

Given multivariate times series of $N$ variables $X = \{x_1(t), x_2(t), \dots x_N(t)\}; t = 1, \dots, T$, the nonlinear multiview RBFN model maps the top $k$ manifolds such that the likelihood of the nonlinear autoregressive model is maximized:

$$X_j(t+1) = \Psi(M)\alpha_j + \varepsilon_j,\ \ j = 1, 2, \dots\ N \tag{IV.3}$$

where,

$$X_j(t+1) = [x_j(E+1)\ x_j(E+2)\ \dots\ x_j(T)]^T \tag{IV.4}$$

$$\Psi(M) = [\Psi(M_1)\ \Psi(M_2)\ \dots\ \Psi(M_k)] \tag{IV.5}$$

$$\alpha_j = [\alpha_{j1}\ \alpha_{j2}\ \dots\ \alpha_{jk}]^T \tag{IV.6}$$

$$\alpha_{jl} = \left[\alpha_{jl}(1)\ \alpha_{jl}(2)\ \dots\ \alpha_{jl}(p)\right]^T,\ \ l = 1, 2, \dots k \tag{IV.7}$$

$$M_l = [M_l^1 \quad M_l^2 \quad \cdots \quad M_l^{T-E}] = \begin{bmatrix} x_d(E) & x_d(E+1) & \cdots & x_d(T-1) \\ x_e(E-n_1\tau) & x_e(E-n_1\tau+1) & \cdots & x_e(T-n_1\tau-1) \\ \vdots & \vdots & \ddots & \vdots \\ x_b(E-n_f\tau) & x_b(E-n_f\tau) & \cdots & x_b(T-n_f\tau-1) \end{bmatrix},$$

$$d, e, b\ \in \{1, 2, \dots N\},\ \ n_1, n_f \in \{0, 1, \dots, E-1\} \tag{IV.8}$$

$$\Psi\left(M_l^g\right) = \left[\psi_1\left(M_l^g\right)\psi_2\left(M_l^g\right)\dots\ \psi_p\left(M_l^g\right)\right],\ \ g = 1, 2, \dots, T-E \tag{IV.9}$$

$$\psi_\rho\left(M_l^g\right) = \exp(-||M_l^g - \mu_{l\rho}||^2/2{\sigma_l}^2),\ \ \rho = 1, 2, \dots p \tag{IV.10}$$

N is the number of variables in the chaotic system. T is the time-series length. The value of the $j^{th}$ variable at time $t$ is denoted by $x_j(t)$. $\{\mu_{l\rho}\}_{\rho=1}^{p}$ is the set of $p$ centers in the space of the $l^{th}$ manifold $M_l$ of the top $k$ manifold reconstructions. The centers are determined by a k-means clustering procedure. $\alpha_j$ is the vector of weights between the target variable $x_j$ and $\Psi(M)$. $\sigma_l$ is the width or radii of the Gaussian RBF in the space of $M_l$ which is selected as the average of the Euclidean distances between each center $\mu_{lp}$ and its nearest neighbor $\mu_{l\rho}$, (*304*). $\alpha_{jl}(\rho)$ is the weight corresponding to the kernel function $\psi_p\left(M_l^g\right)$. Here the type of the radial basis function $\psi_1\left(M_l^g\right)$ is taken as Gaussian kernels whose inputs are $E$-dimensional vectors of a combination of

variables and time lags. $\alpha$ vectors are weights that are fixed such that the prediction error is minimized, and $\varepsilon(t)$ denotes Gaussian white noise independent of the time series. In general, one can use the least squares method to adjust the $\alpha$ weights in the minimum mean squared error sense. Once the $\alpha_j$ vector is estimated via least squares on the library data that is selected randomly from the in-sample portion of the data, they are tested on the out-of-sample test set to calculate the out-of-sample forecast.

### IV.C.2 Simulated Data

The simulated data used in this work is generated from ecosystem simulations of a three-species food chain (*305*), a three species couple logistic model (*38*), a flour beetle model (*306*) and a five species model (*307*).

*Three-species food chain model*

The following differential equations model a chaotic three-species food chain of variables $x$, $y$, and $z$ (*305*):

$$dx/dt = x(1-x) - f_1(x)y \tag{IV.11}$$

$$dy/dt = f_1(x)y - f_2(y)z - d_1 y \tag{IV.12}$$

$$dz/dt = f_2(y)z - d_2 z \tag{IV.13}$$

with

$$f_i(u) = a_i u/(1 + b_i u) \tag{IV.14}$$

The parameter values used in the simulations are as follows: $a_1 = 2.5$, $a_2 = 0.1$, $b_1 = 3.2$, $b_2 = 2$, $d_1 = 0.2$, and $d_2 = 0.015$. The initial conditions used are $x_0 = 0.8$, $y_0 = 0.2$ and $z_0 = 8$.

*Three species coupled logistic model*

The three interacting species $x$, $y$, and $z$ are model through the following coupled logistic map as mentioned in Ye et al. (2016):

$$\begin{bmatrix} x_{t+1} \\ y_{t+1} \\ z_{t+1} \end{bmatrix} = \begin{bmatrix} 3.6 \\ 3 \\ 3 \end{bmatrix} \circ \begin{bmatrix} x_t \\ y_t \\ z_t \end{bmatrix} \circ \left( \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} - \begin{bmatrix} 1 & 0.2 & 0.2 \\ 0.2 & 1 & -0.2 \\ 0.2 & -0.2 & 1 \end{bmatrix} \right) \tag{IV.15}$$

where ° is the entry wise product. The initial conditions used in the simulations are $\begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix} = \begin{bmatrix} 0.2 \\ 0.2 \\ 0.2 \end{bmatrix}$.

*Flour beetle model*

The chaotic behavior of an insect population, *Tribolium Castaneum*, is modeled through the following equations for different life stages (larvae, pupae, and adults) of flour beetle suggested by Dennis et al. (*306*):

$$L_{t+1} = bA_t \exp(-c_{e1}L_t - c_{ea}A_t) \tag{IV.16}$$

$$P_{t+1} = L_t(1 - \mu_1) \tag{IV.17}$$

$$A_{t+1} = P_t \exp\left(-c_{pa}A_t\right) + A_t(1 - \mu_a) \tag{IV.18}$$

with the following parameter values used in the simulations: $b = 10.67$, $\mu_1 = 0.1955$, $\mu_a = 0.96$, $c_{e1} = 0.01647$, $c_{ea} = 0.01313$, $c_{pa} = 0.35$ . The initial values are $L_1 = 250$, $P_1 = 5$ and $A_1 = 100$.

*Five-species model*

The following equations identify a chaotic five-species competition model for variables $Y_1$, $Y_2$, $Y_3$, $Y_4$, and $Y_5$ suggested by Sugihara et al. (*307*):

$$Y_1(t) = Y_1(t)[4 - 4\,Y_1(t) - 2\,Y_2(t) - 0.4\,Y_3(t)] \quad \text{(IV.19)}$$

$$Y_2(t) = Y_2(t)[3.1 - 0.31\,Y_1(t) - 3.1\,Y_2(t) - 0.93\,Y_3(t)] \quad \text{(IV.20)}$$

$$Y_3(t) = Y_3(t)[2.12 + 0.636\,Y_1(t) + 0.636\,Y_2(t) - 2.12\,Y_3(t)] \quad \text{(IV.21)}$$

$$Y_4(t) = Y_4(t)[3.8 - 0.111\,Y_1(t) - 0.011\,Y_2(t) + 0.131\,Y_3(t) - 3.8\,Y_4(t)] \quad \text{(IV.22)}$$

$$Y_5(t) = Y_5(t)[4.1 - 0.082\,Y_1(t) - 0.111\,Y_2(t) - 0.125\,Y_3(t) - 4.1\,Y_5(t)] \quad \text{(IV.23)}$$

with the initial conditions $Y_1(1)= Y_5(1) =0.1$, $Y_2(1)=0.02$, $Y_3(1) = Y_4(1)=0.01$.

### IV.C.3 Real Data

*Mesocosm plankton community data*

The data drawn from the mesocosm 8-year experiment on a plankton community isolated from the Baltic Sea has been shown to represent the dynamics of a chaotic system. We use the transformed data of the abundance of Rotifers, Calanoid Copepods, Picocyanobacteria and Nanoflagellates from the supplementary material of Benica et al. (*308*). The data transformation in Benica et al. is done such that the raw data is interpolated by hermite cubic interpolation to obtain data with equidistant time intervals of 3.35 days, and then rescaled by a fourth-root transformation to suppress sharp peaks. The transformed data is of length 794 samples.

### IV.C.4 Manifold Reconstruction

As described in Ye et al. (*38*), the possible $m$ number of 3-dimensional manifold reconstructions of combination of variables and their time lags of $0, \tau$ and $2\tau$ is:

$$m = \binom{NL}{E} - \binom{N(L-1)}{E} \quad \text{(IV.24)}$$

where $N$ is the number of variables in the dynamic system, $L$ is the number of possible lags for each variable, and $E$ is the embedding dimension. The first term is the number of manifolds formed by choosing $E$ of the $NL$ possible coordinates, and the second term is subtracted to eliminate the number of invalid manifolds with $E$ lagged coordinates. A valid manifold is one with at least one unlagged coordinate. For example, the possible number of valid manifold reconstructions for a 3 and 4 variable system is 64, and 164 respectively. Unlike Ye et al. (*38*) that suggests $k = \sqrt{m}$, we found out that for the multiview radial basis function network (MV-RBFN) approach the best number of top $k$ reconstructions to incorporate into MV-RBFN is $k = N$, where $N$ is the number of variables in the interconnected dynamic system. This is because for any $N$-variate system, if we let $k$ be equal to $\sqrt{m}$ ($\sqrt{m} \geq N$) we will have too many hidden units in the hidden layer of the radial basis function network. Particularly in cases where the time series is noisy, too many hidden units in the hidden layer of the neural network leads to overfitting of the training samples and poor generalization (*309*). In this work, we choose $\tau = 1$ and $E = 3$ for the ecosystem simulated data and mesocosm experiment data.

### IV.C.5 Out-of-sample Forecasting

In order to quantitatively evaluate the one-step-ahead forecast skill of the MV-RBFN, we performed an out of sample forecast scheme on the simulated ecosystem data. We generated 3000 samples for all variables in the simulated ecosystem models and discarded the first 500 samples to exclude the transient behavior of the time series. The last 500 samples [2501 to 3000] are kept as the out of sample test set. Radial basis function based autoregressive model is performed on each of the $m$ manifold reconstruction to rank them based on their forecast skill in the in-sample portion of the data. For the simulated time series data, 100 libraries are randomly chosen in the in-sample portion of the data [501 to 2000]; the starting point of each library is chosen from a uniform

distribution distributed in the [501 to 2000] interval. The top $k$ manifold reconstructions are selected to perform MV-RBFN forecasting (as shown in Figure IV.2). The forecast skill is then calculated by averaging the mean absolute values of the 100 randomly sampled libraries. The libraries are selected in various lengths of 25, 50, 75 and 100 samples.

We use the same out-of-sample forecasting scheme to calculate the performance of the MVE approach proposed by Ye et al. (*38*).

### IV.C.6 Pseudo Out-of-sample Forecasting

Due to the limited length of the mesocosm data, we used a pseudo out-of-sample forecast scheme to evaluate the forecast performance of the MV-RBFN and multi-view embedding approaches; the first 3/4 of the time-series were used as the training set, and the last 1/4 portion of the data was used as the test set. This forecast scheme is also known as the method of time-series cross-validation for one-step ahead forecasts. In the pseudo-out-of-sample strategy, the one-step-ahead forecast at time $t + 1$ is estimated using data through time $t$, then moving forward to time $t + 1$ and repeating until all test data samples are covered in the recursive estimation . In this work, we used an increasing data window in the recursive forecast of samples.

## IV.D Results

To assess the performance of the MV-RBFN approach, we compare the forecast performance (correlation) between the out-of-sample forecast estimates and the one-step-ahead observations using our proposed MV-RBFN autoregressive model with that of MVE proposed by Ye et al. (2016) (*38*). Figure IV.3 depicts the forecast skill (correlation) of the MV-RBF and the MVE approaches for simulated ecological systems with 10% added noise for a three-species food chain (*305*), a three-species coupled logistic and a three-stage flour-beetle model (*306*) (for additional details see Supplementary Materials). In almost all cases, MV-RBFN outperforms MVE resulting in better forecast skills with higher correlations. As expected, the forecast performance improves as the length of time-series increases.

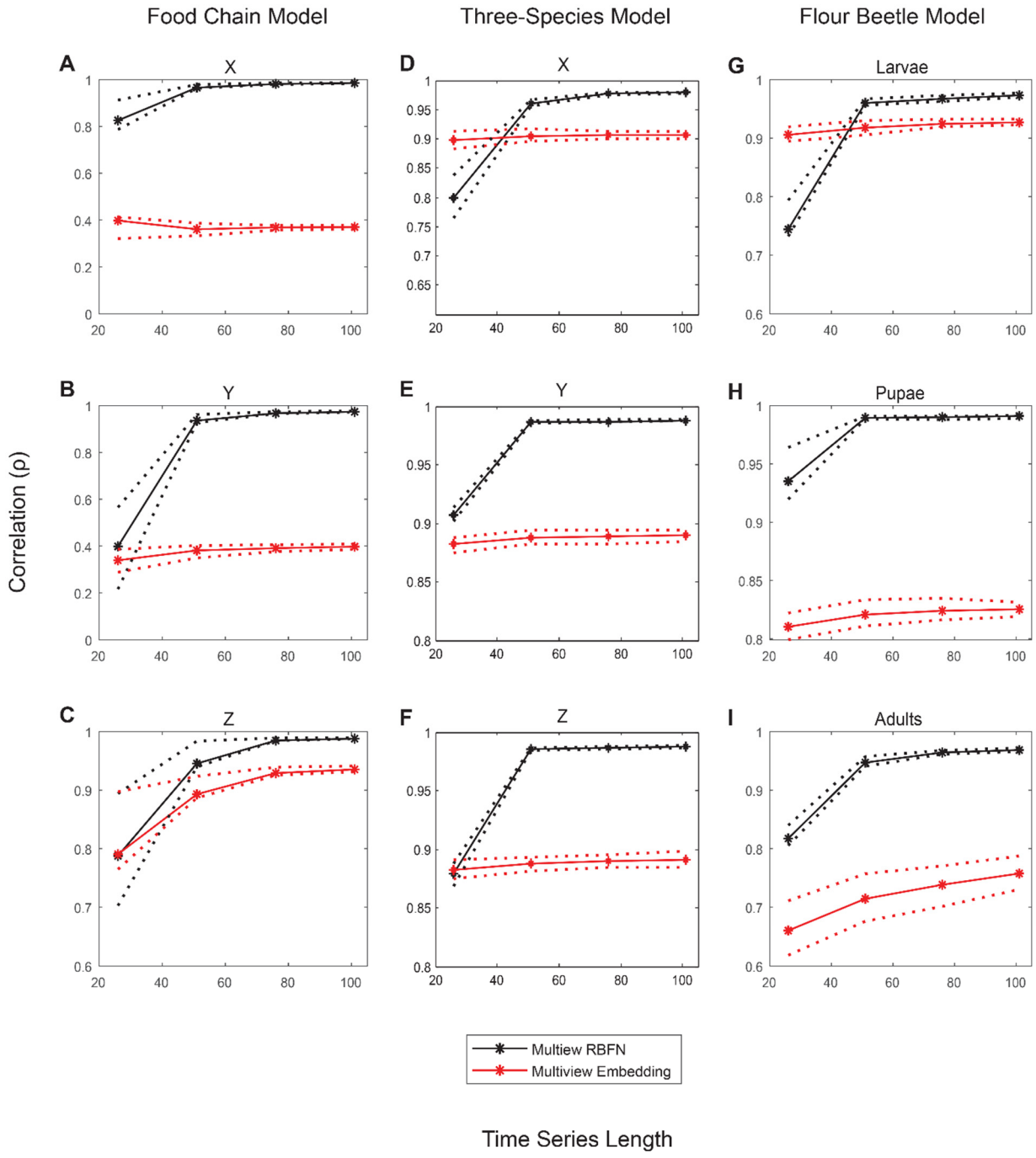



**Figure IV.3 Comparison of forecast performances for MV-RBFN and MVE in simulated ecological data with 10% added noise. (A to C)** forecast skill (correlation between estimated forecast and one-step-ahead observation) versus length of the libraries for variables X, Y, and Z in three-species food chain model. **(D to F)** same as A to but for the three-species coupled logistic model. **(G to I)** same as A to C for the flour beetle model. Solid lines show the averaged values for 100 randomly selected libraries, and the dotted lines indicate the upper and lower quartiles.

To further assess the forecast skill of MV-RBF on real world data, we extend this analysis to time-series data from along term mesocosm experiment on a four-species marine plankton community obtained from the Baltic Sea (*308*). The mesocosm data consists of the plankton population of Nanoflagellates and Picocyanobacteria that fall prey to two predators, Rotifers and Calanoid Copepods. Coupling of predator-prey oscillations where preys have a causal effect on the predators exhibit chaotic patterns. Figure IV.4 shows the comparison of the forecast performances of MV-RBFN and MVE for the long-term plankton community data; for all four species, MV-RBFN outperforms MVE in forecasting. Using the MAE metric provides similar results when comparing MVE and MV-RBFN (Figures IV.5 and IV.6).

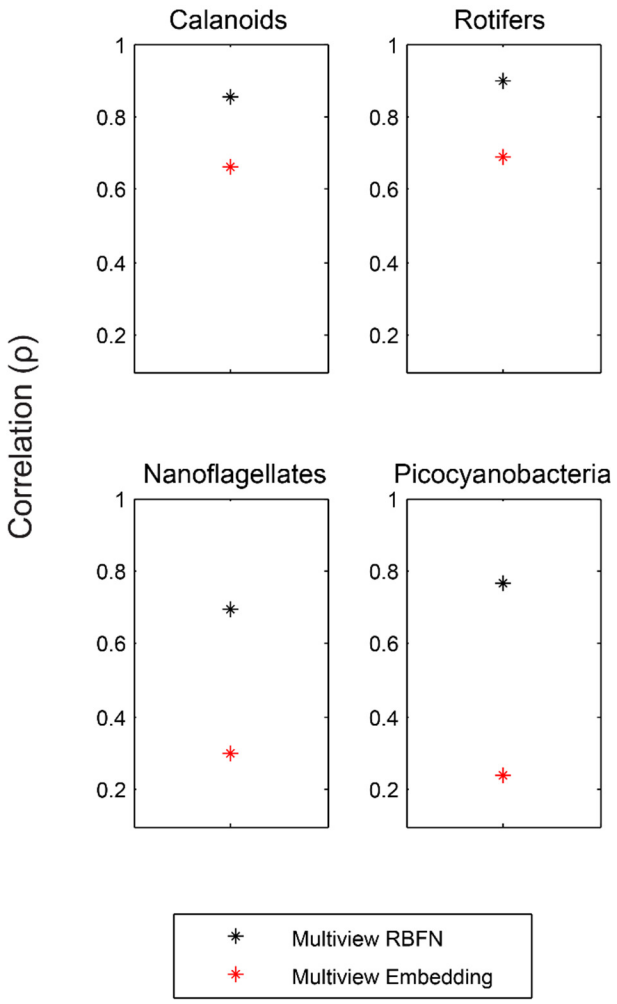



**Figure IV.4 Comparison of forecast performance of MV-RBFN and MVE for the long-term mesocosm experiment.** Correlation between the predictions and observations for plankton communities of calanoids, rotifers, nanoflagellates and picocyanobacteria.

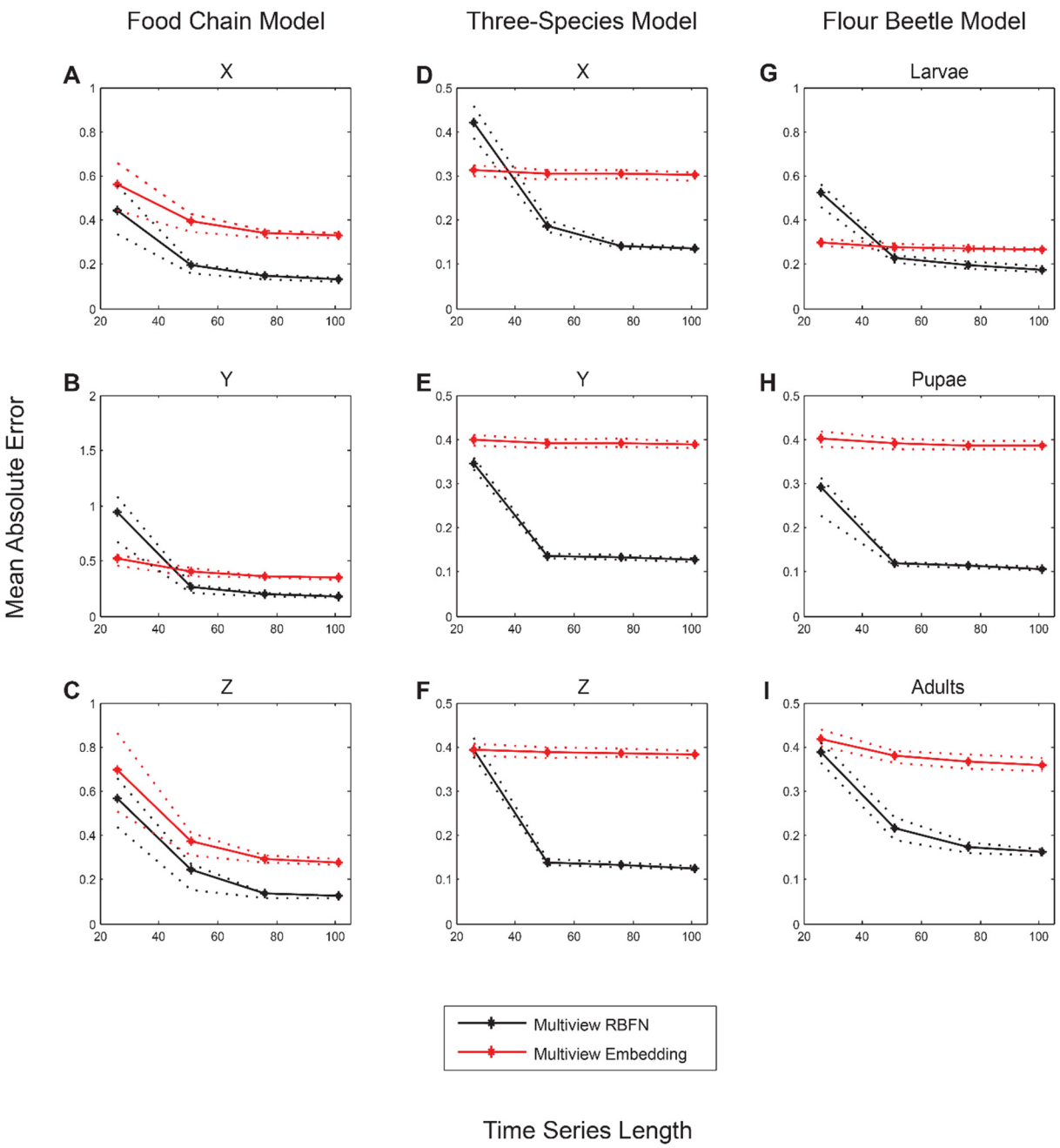



**Figure IV.5 Comparison of forecast performance (mean absolute error) for MV-RBFN and MVE in simulated ecological data with 10% added noise. (A to C)** forecast skill (mean absolute error between estimated forecast and one-step-ahead observation) versus length of the libraries for variables X, Y, and Z in three-species food chain model. **(D to F)** same as A to but for the three-species coupled logistic model. **(G to I)** same as A to C for the flour beetle model. Solid lines show the average values for 100 randomly selected libraries, and the dotted lines indicate the upper and lower quartiles.

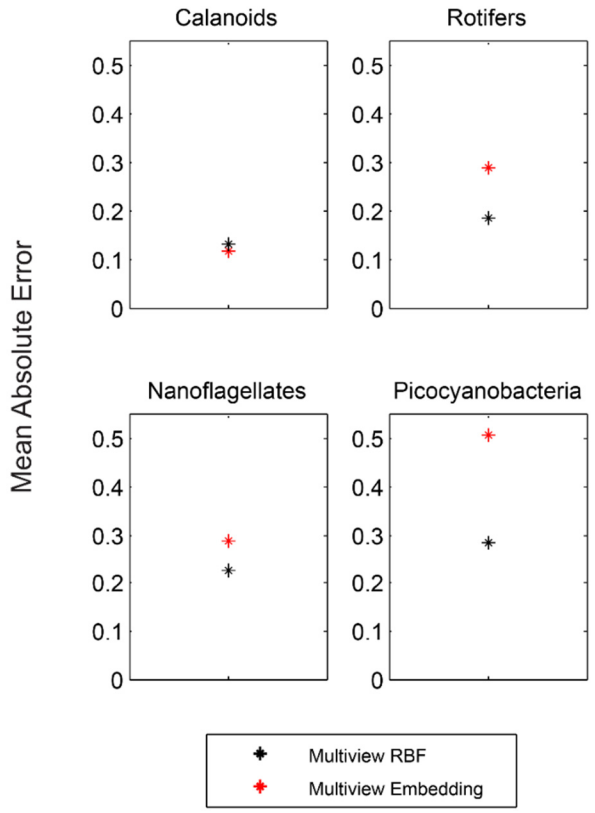



**Figure IV.6 Comparison of forecast performance (mean absolute error) of MV-RBFN and MVE for the long-term mesocosm experiment.** Mean absolute error between the predictions and observations for plankton communities of calanoids, rotifers, nanoflagellates and picocyanobacteria.

We compare the forecast skill of the MV-RBFN approach with that from a univariate radial basis function method and a multiview radial basis function approach using the best single view in terms of mean absolute error (MAE) and correlation ($\rho$). Figures IV.7 to IV.10 show that MV-RBFN yields a better forecast performance than that from the univariate RBFN approach and the best single view RBFN for the three-species models and a five-species model (*307*). Furthermore to study the modeling framework of MV-RBFN, we look into the effect of observational noise in the time-series data. Figures IV.11 to IV.16 indicate that as more noise is added to the data, the forecast error increases.

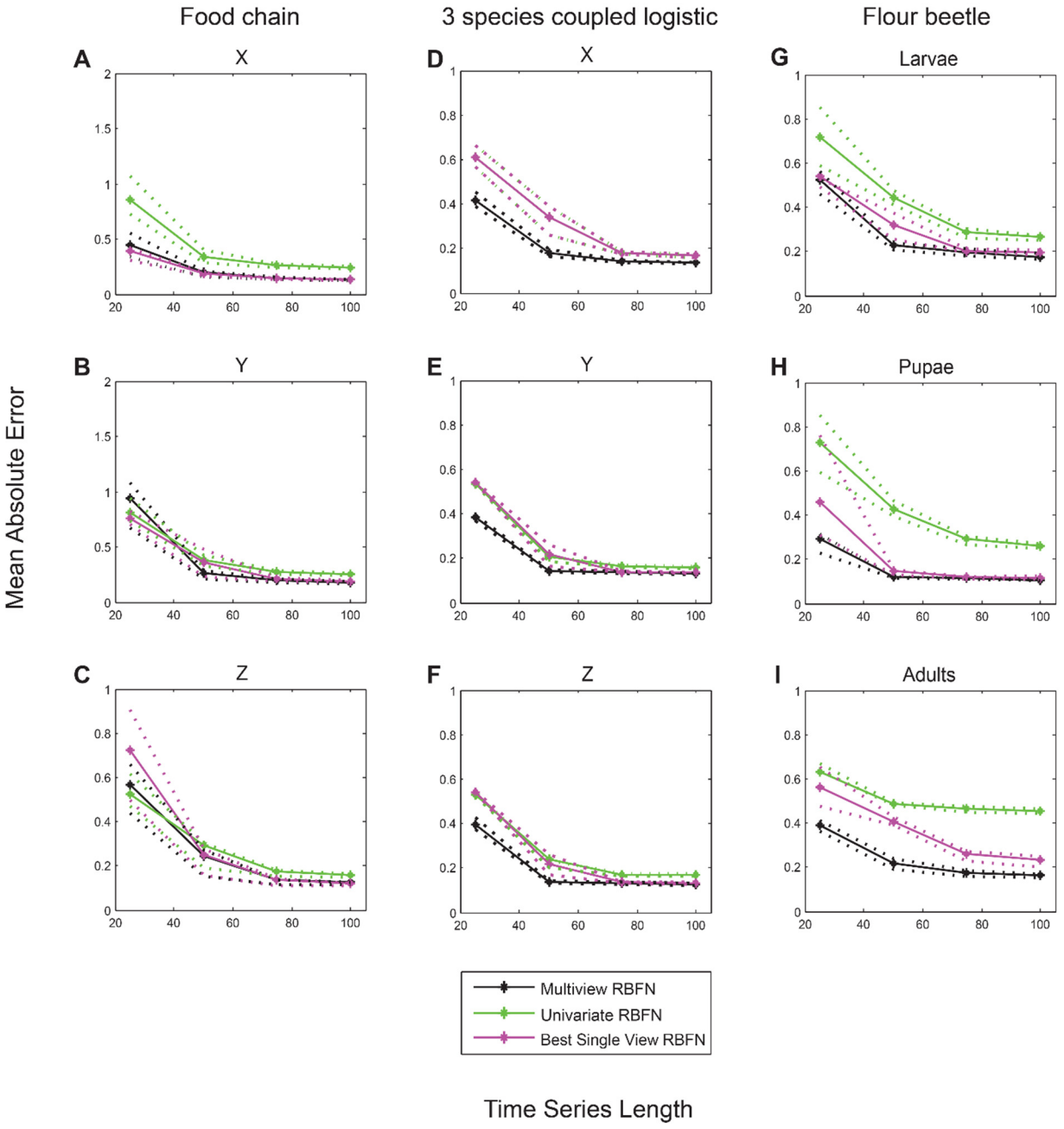



**Figure IV.7 Forecast performance (mean absolute error) vs. time series length of libraries with 10% added noise. (A to C)** average mean absolute error between predictions and observations for 100 randomly sampled libraries for variables $X$, $Y$., and $Z$ vs. length of the libraries in the food chain model. **(D to F)** same as A to C but for the 3 species coupled logistic model. **(G to I)** same as A to C but for the variables larvae, pupae and adults in the flour beetle model. The solid black curves are the average mean absolute errors for the Multiview RBFN approach for the top $k$ manifold reconstructions. The solid green curves are the average mean absolute errors for the univariate RBFN approach, and the solid pink curves are the average mean absolute error using the single best view (manifold) in the RBFN autoregressive approach. The dotted lines are the upper and lower quartiles.

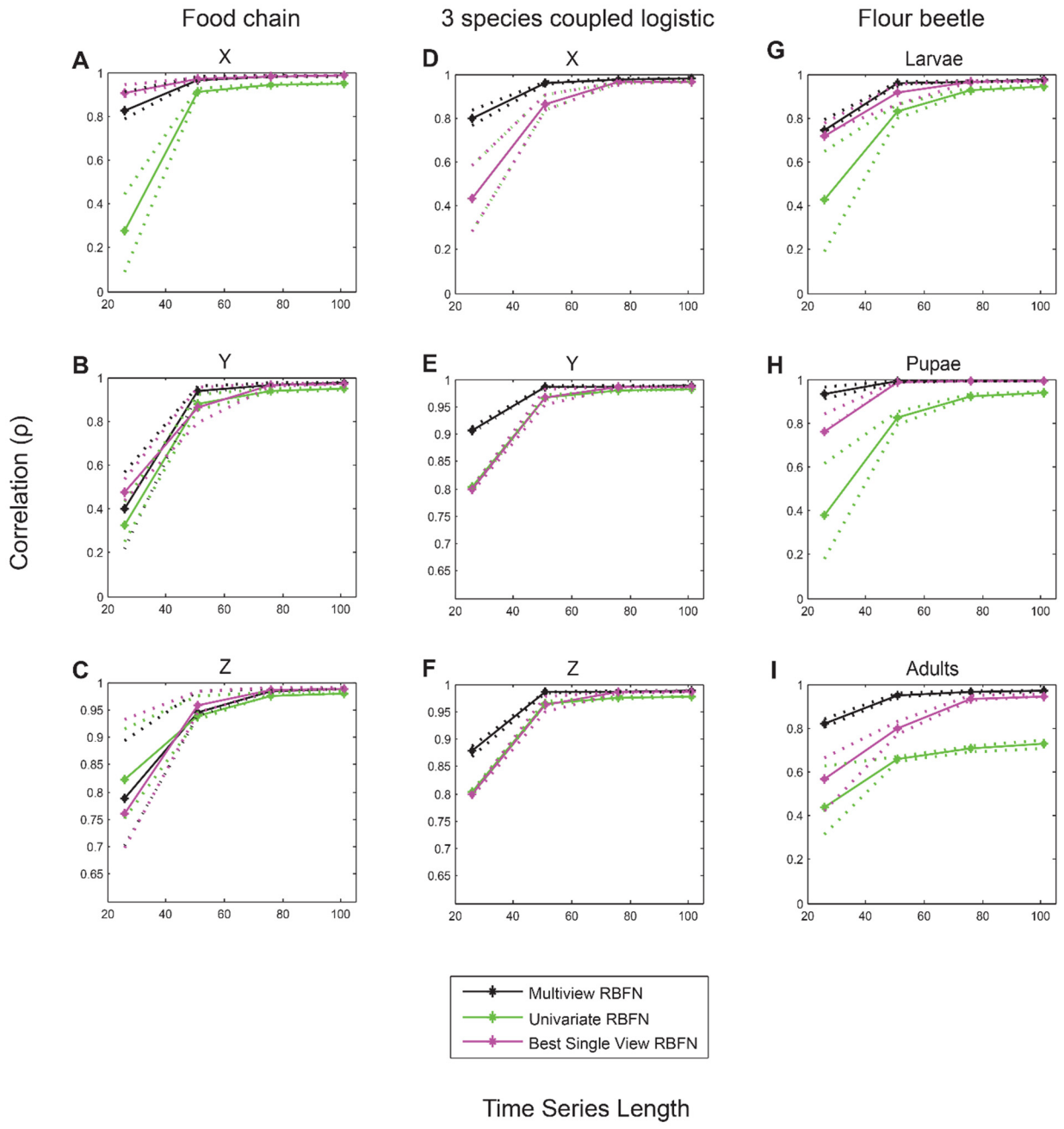



**Figure IV.8 Forecast performance (correlation) vs. time series length of libraries with 10% added noise. (A to C)** average correlation between predictions and observations for 100 randomly sampled libraries for variables $X$, $Y$., and $Z$ vs. length of the libraries in the food chain model. **(D to F)** same as A to C but for the 3 species coupled logistic model. **(G to I)** same as A to C but for the variables larvae, pupae and adults in the flour beetle model. The solid black curves are the average correlation for the Multiview RBFN approach for the top $k$ manifold reconstructions. The solid green curves are the average correlation for the univariate RBFN approach, and the solid pink curves are the average correlation using the single best view (manifold) in the RBFN autoregressive approach. The dotted lines are the upper and lower quartiles.

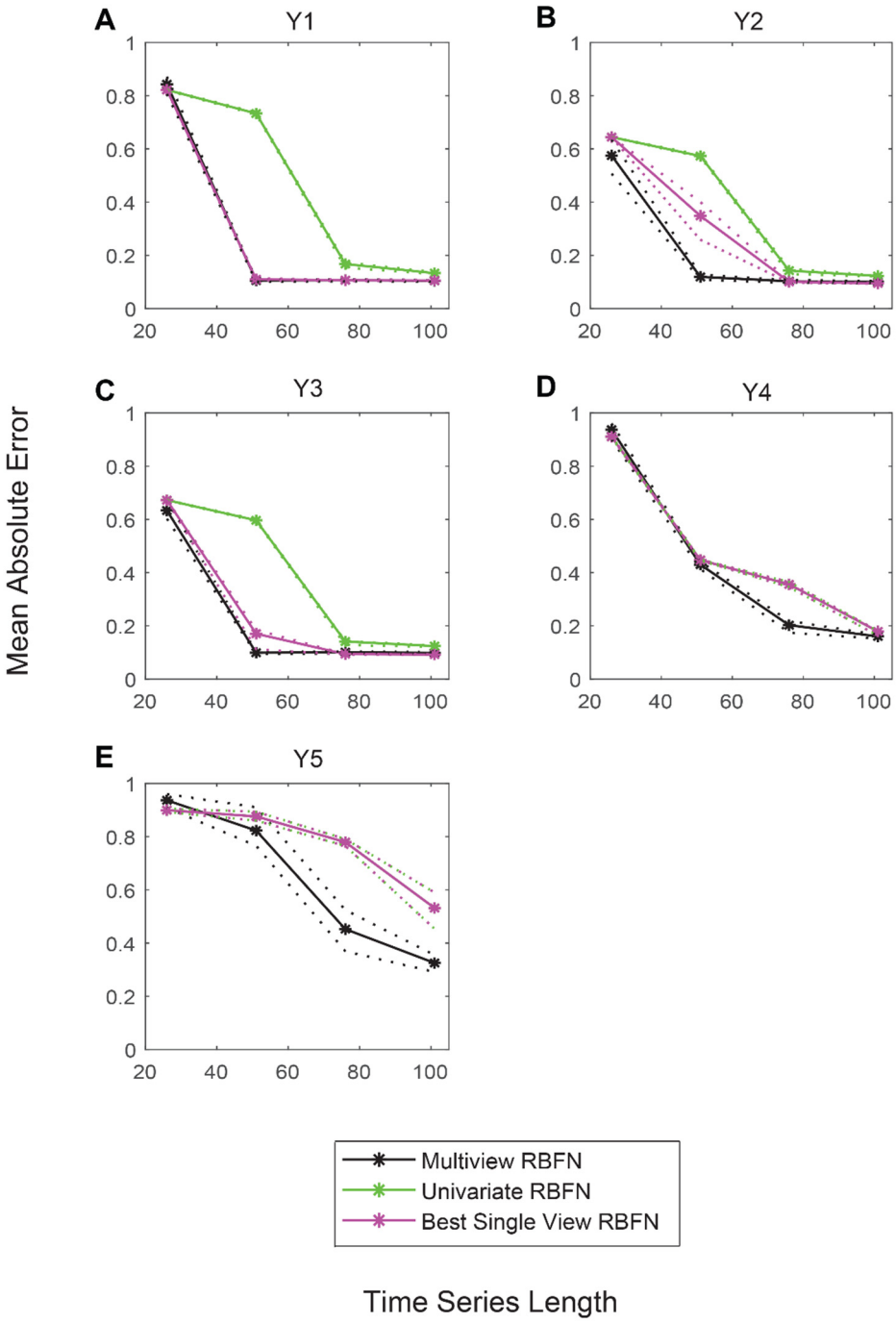



**Figure IV.9 Forecast performance (mean absolute error) vs. time series length of libraries for the five-species model with 10% added noise. (A to E)** average mean absolute error between predictions and observations for 100 randomly sampled libraries for variables $Y_1$, $Y_2$, $Y_3$, $Y_4$, $Y_5$ vs. length of the libraries. The solid black curves are the average mean absolute error for the Multiview RBFN approach for the top $k$ manifold reconstructions. The solid green curves are the average mean absolute error for the univariate RBFN approach, and the solid pink curves are the average mean absolute error using the single best view (manifold) in the RBFN autoregressive approach. The dotted lines are the upper and lower quartiles. In figure panels D and E, the manifolds of the univariate and the best single view coincide.

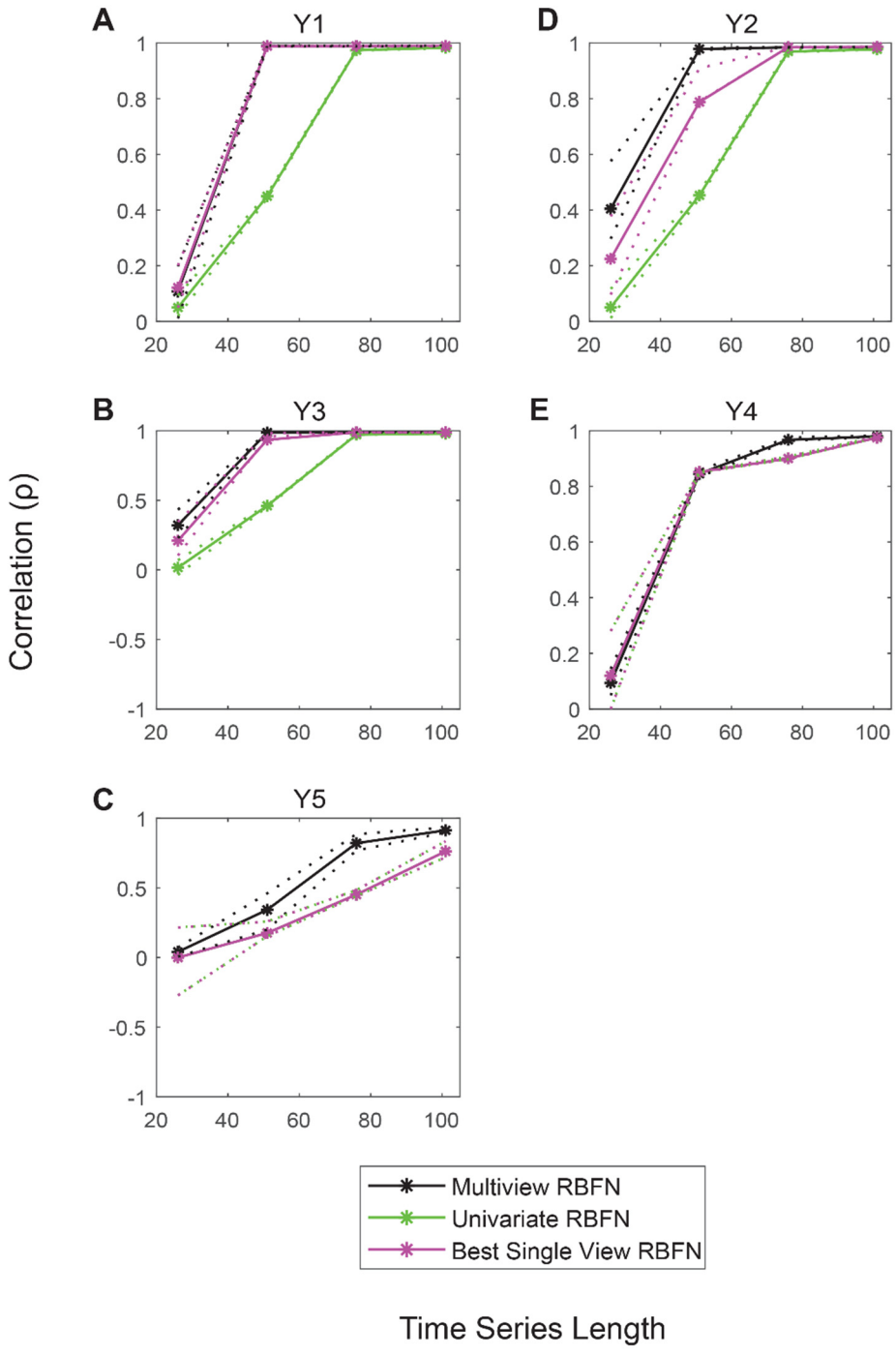



**Figure IV.10 Forecast performance (correlation) vs. time series length of libraries for the five-species model with 10% added noise. (A to E)** average mean absolute error between predictions and observations for 100 randomly sampled libraries for variables $Y_1$, $Y_2$, $Y_3$, $Y_4$, $Y_5$ vs. length of the libraries. The solid black curves are the average correlation for the Multiview RBFN approach for the top $k$ manifold reconstructions. The solid green curves are the average correlation for the univariate RBFN approach, and the solid pink curves are the average correlation using the single best view (manifold) in the RBFN autoregressive approach. The dotted lines are the upper and lower quartiles. In figure panels D and E, the manifolds of the univariate and the best single view coincide.

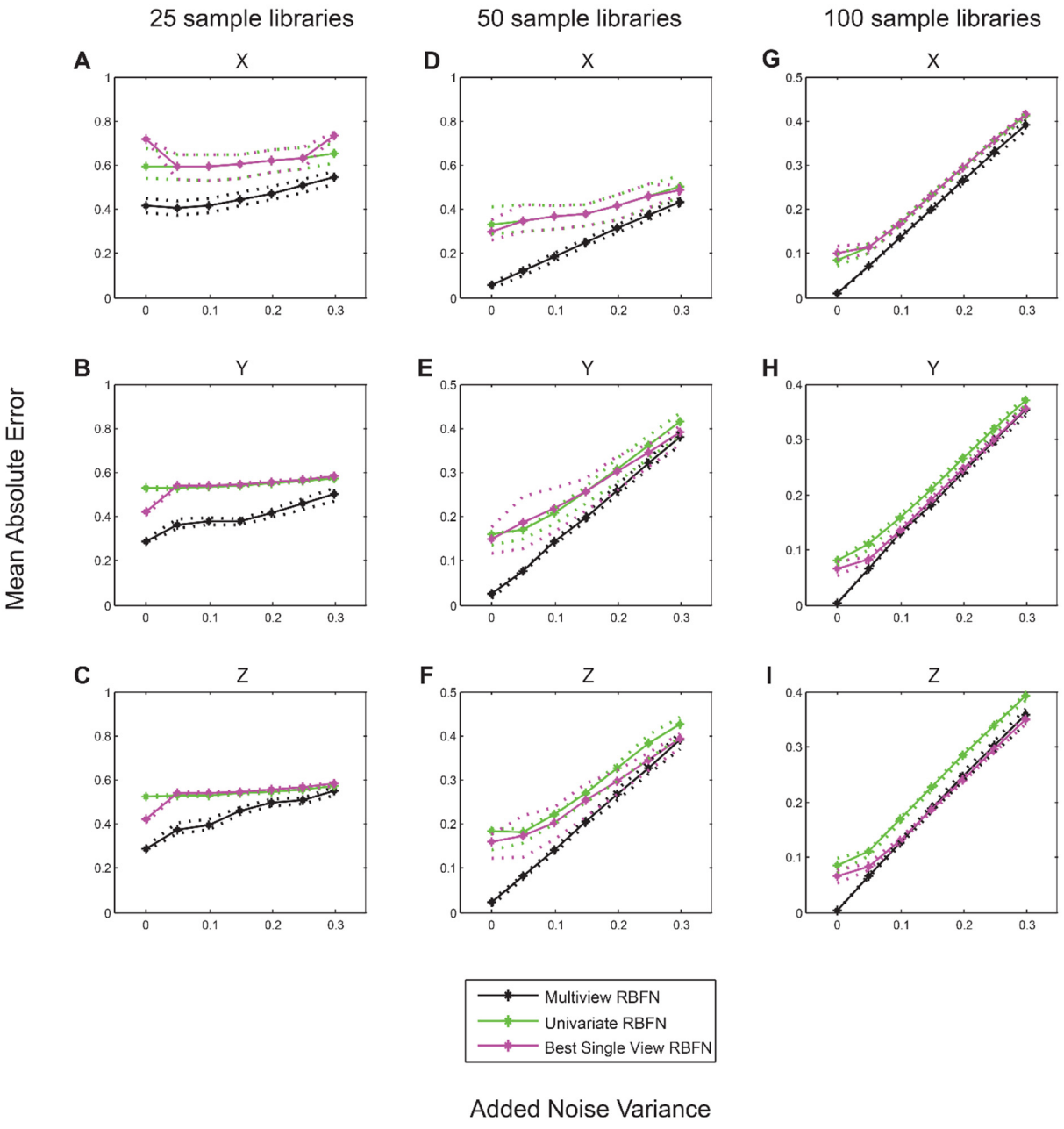



**Figure IV.11 Forecast performance (mean absolute error) vs. noise for the 3 species coupled logistic model. (A to C)** average mean absolute error between predictions and observations for 100 randomly sampled libraries of length 25 for variables $X$, $Y$, and $Z$. **(D to F)** same as A to C but for 100 randomly sampled libraries of length 50. **(G to I)** same as A to C but for 100 randomly sampled libraries of length 100. The solid black curves are the average mean absolute errors for the Multiview RBFN approach for the top $k$ manifold reconstructions. The solid green curves are the average mean absolute errors for the univariate RBFN approach, and the solid pink curves are the average mean absolute error using the single best view (manifold) in the RBFN autoregressive approach. The dotted lines are the upper and lower quartiles.

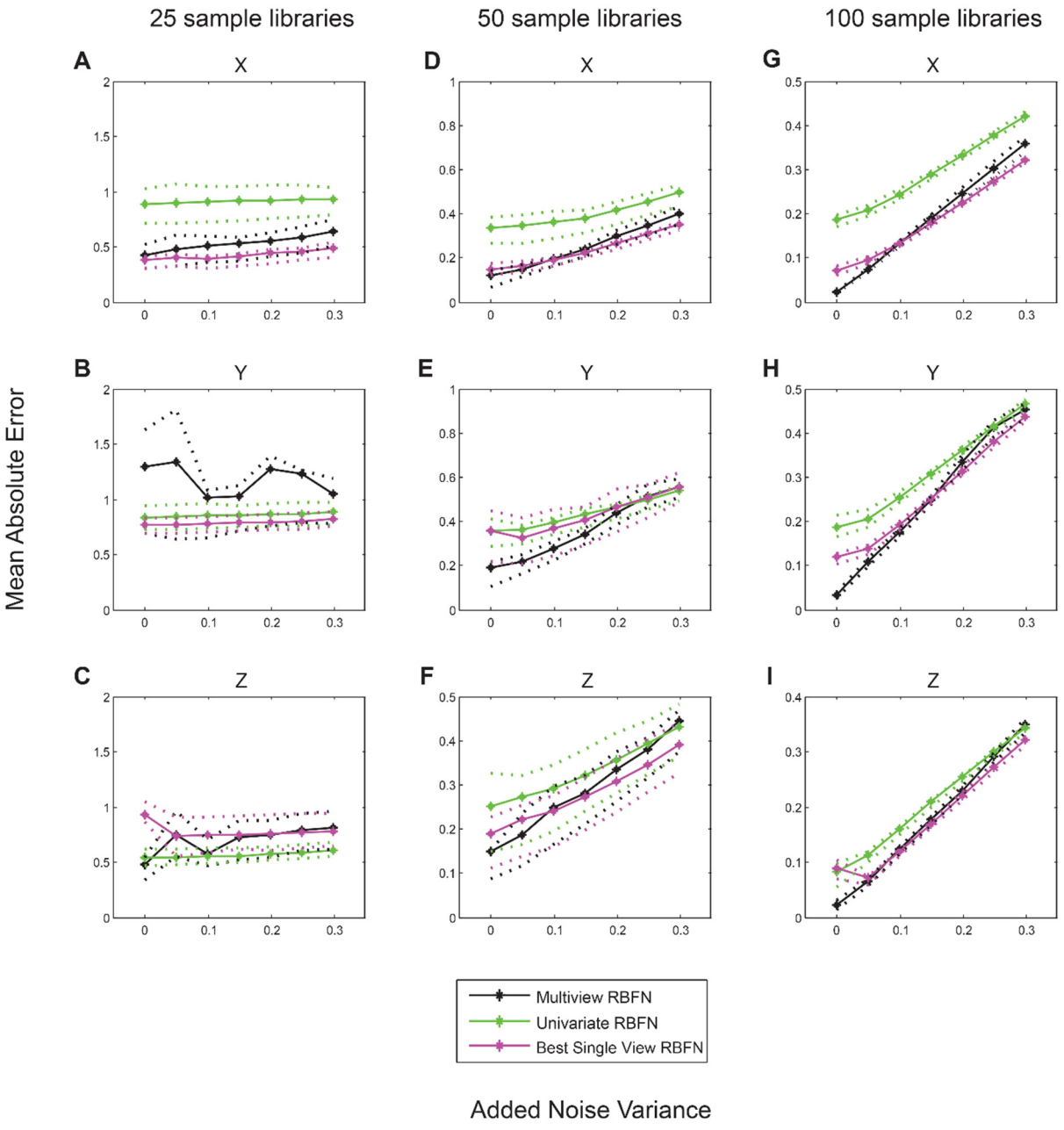



**Figure IV.12 Forecast performance (mean absolute error) vs. noise for the food chain model. (A to C)** average mean absolute error between predictions and observations for 100 randomly sampled libraries of length 25 for variables $X$, $Y$, and $Z$. **(D to F)** same as A to C but for 100 randomly sampled libraries of length 50. **(G to I)** same as A to C but for 100 randomly sampled libraries of length 100. The solid black curves are the average mean absolute errors for the Multiview RBFN approach for the top $k$ manifold reconstructions. The solid green curves are the average mean absolute errors for the univariate RBFNapproach, and the solid pink curves are the average mean absolute error using the single best view (manifold) in the RBFN autoregressive approach. The dotted lines are the upper and lower quartiles.

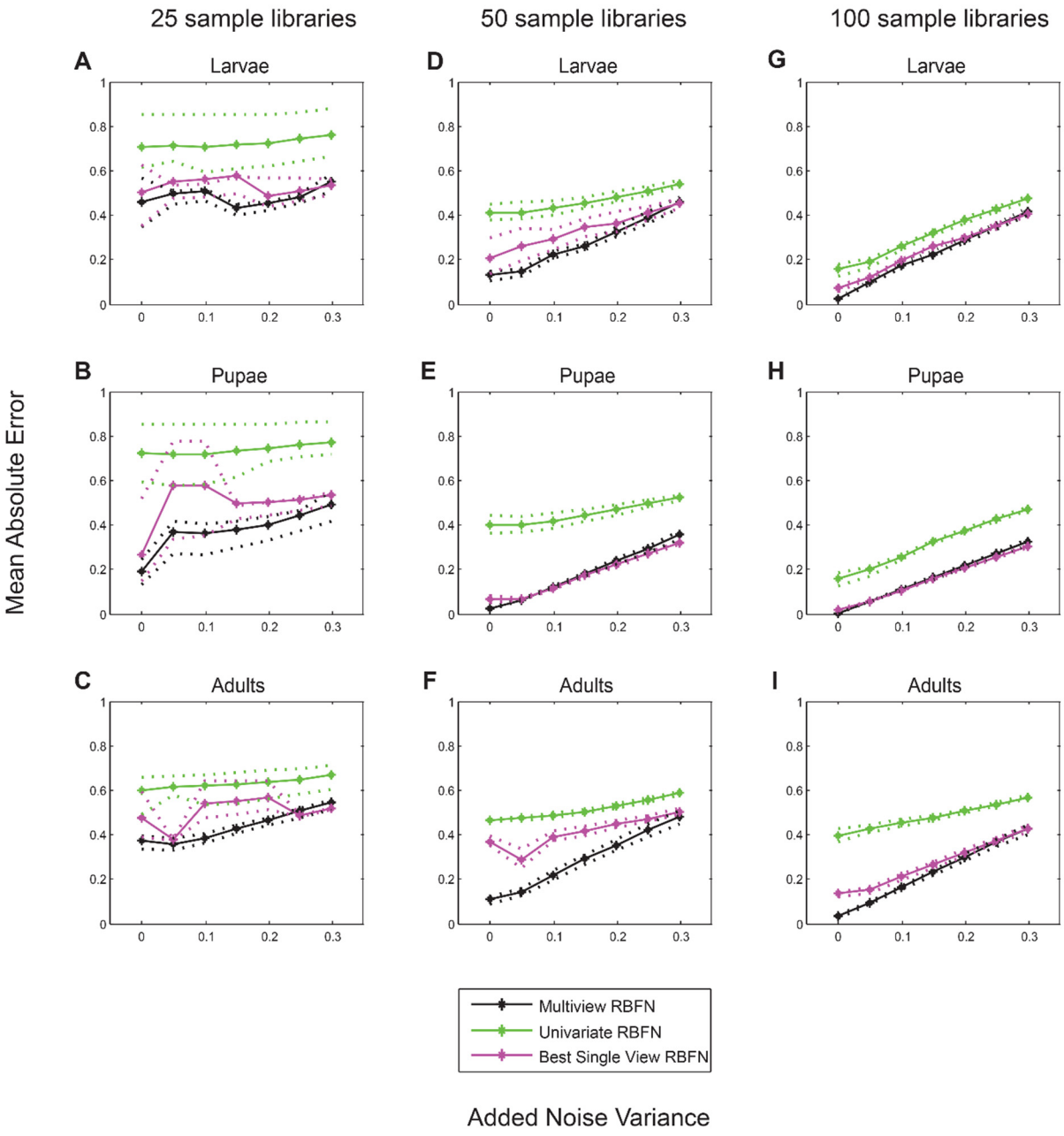



**Figure IV.13 Forecast performance (mean absolute error) vs. noise for the flour beetle model. (A to C)** average mean absolute error between predictions and observations for 100 randomly sampled libraries of length 25 for variables $X$, $Y$, and $Z$. **(D to F)** same as A to C but for 100 randomly sampled libraries of length 50. **(G to I)** same as A to C but for 100 randomly sampled libraries of length 100. The solid black curves are the average mean absolute errors for the Multiview RBFN approach for the top $k$ manifold reconstructions. The solid green curves are the average mean absolute errors for the univariate RBFN approach, and the solid pink curves are the average mean absolute error using the single best view (manifold) in the RBFN autoregressive approach. The dotted lines are the upper and lower quartiles.

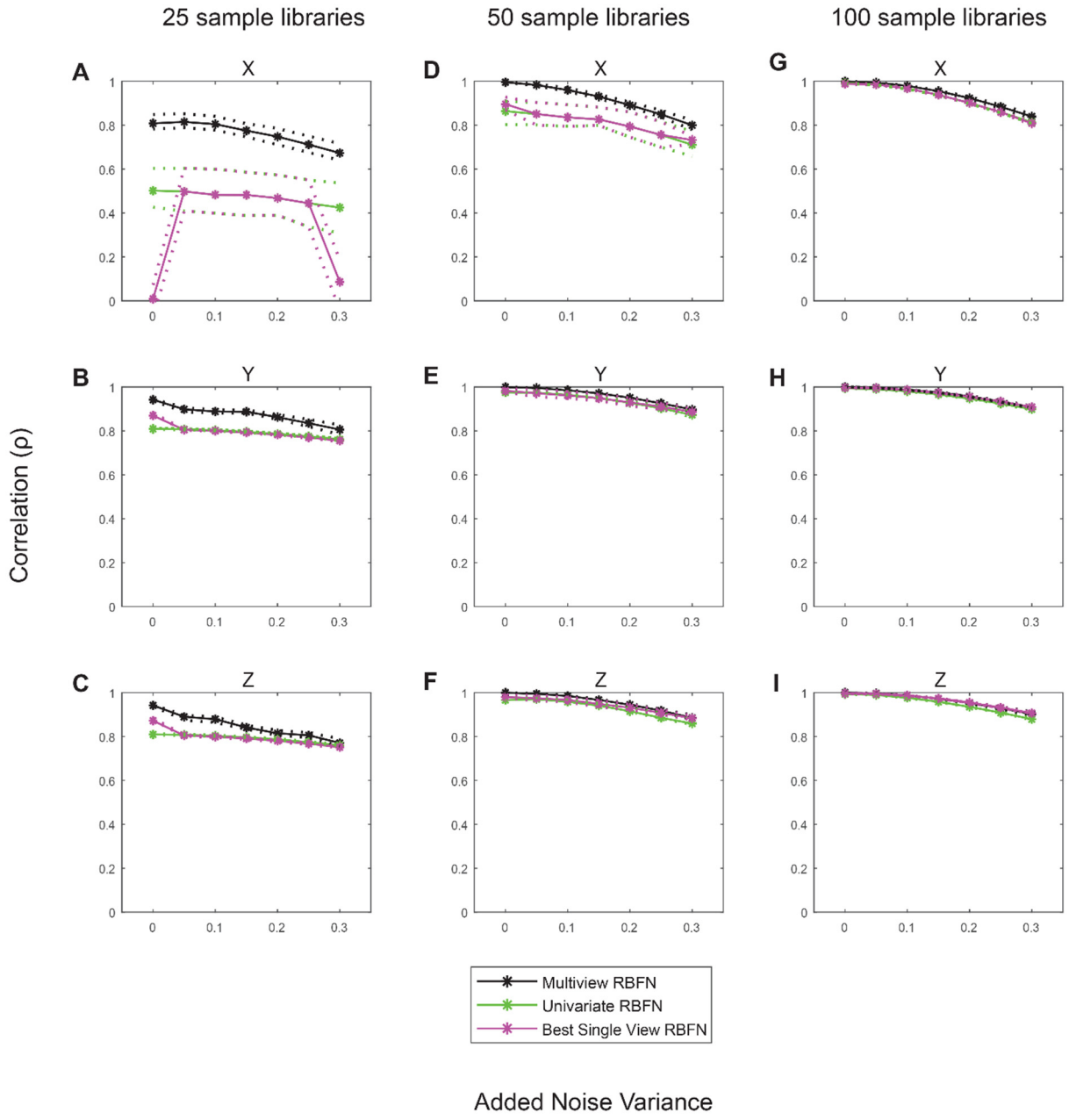



**Figure IV.14 Forecast performance (correlation) vs. noise for the 3 species coupled logistic model. (A to C)** average correlation between predictions and observations for 100 randomly sampled libraries of length 25 for variables $X$, $Y$, and $Z$. **(D to F)** same as A to C but for 100 randomly sampled libraries of length 50. **(G to I)** same as A to C but for 100 randomly sampled libraries of length 100. The solid black curves are the average correlation for the Multiview RBFN approach for the top $k$ manifold reconstructions. The solid green curves are the average correlation for the univariate RBFN approach, and the solid pink curves are the average correlation using the single best view (manifold) in the RBFN autoregressive approach. The dotted lines are the upper and lower quartiles.

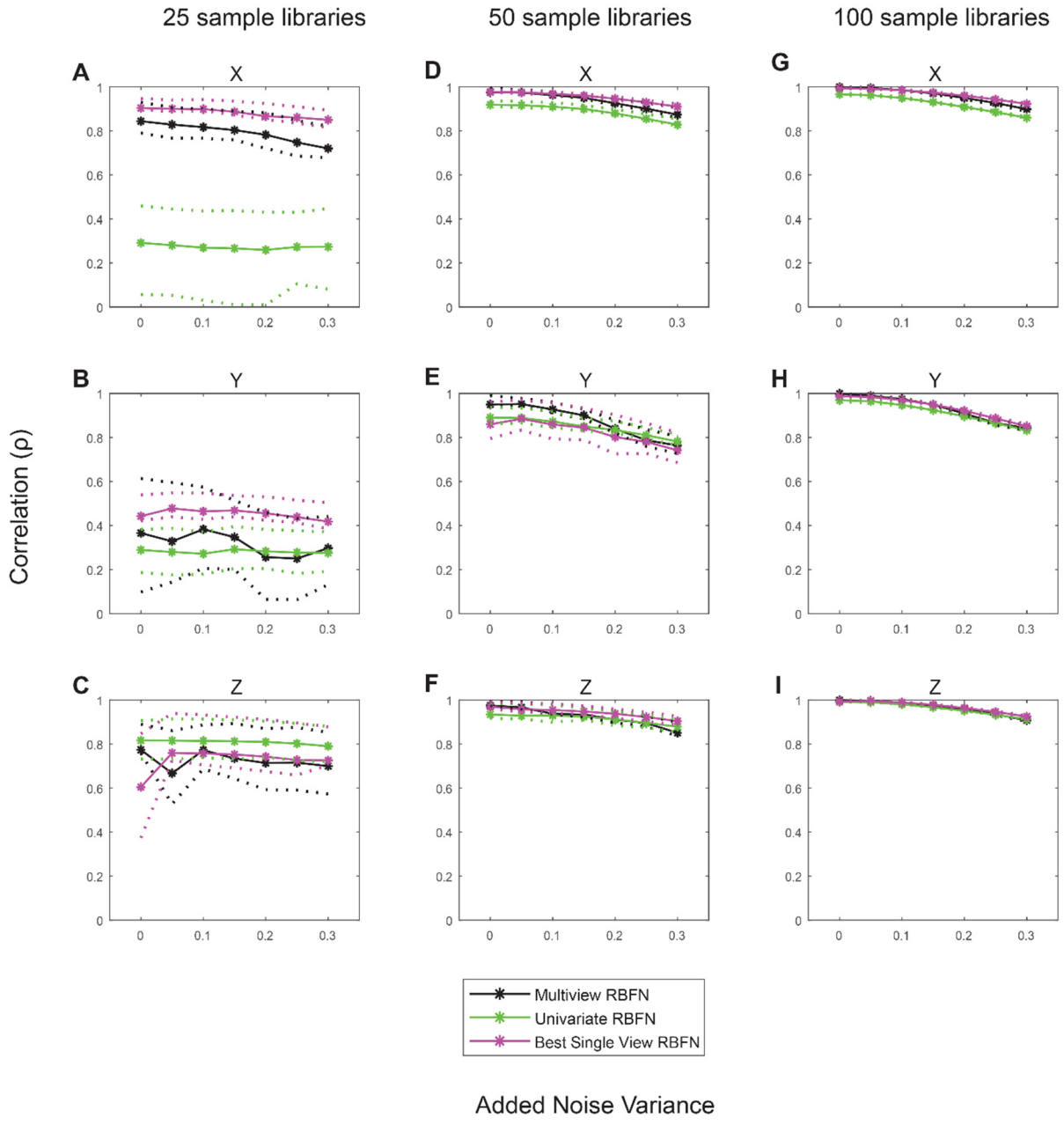



**Figure IV.15 Forecast performance (correlation) vs. noise for the food chain model. (A to C)** average correlation between predictions and observations for 100 randomly sampled libraries of length 25 for variables $X$, $Y$, and $Z$. **(D to F)** same as A to C but for 100 randomly sampled libraries of length 50. **(G to I)** same as A to C but for 100 randomly sampled libraries of length 100. The solid black curves are the average correlation for the Multiview RBFN approach for the top $k$ manifold reconstructions. The solid green curves are the average correlation for the univariate RBFN approach, and the solid pink curves are the average correlation using the single best view (manifold) in the RBFN autoregressive approach. The dotted lines are the upper and lower quartiles.

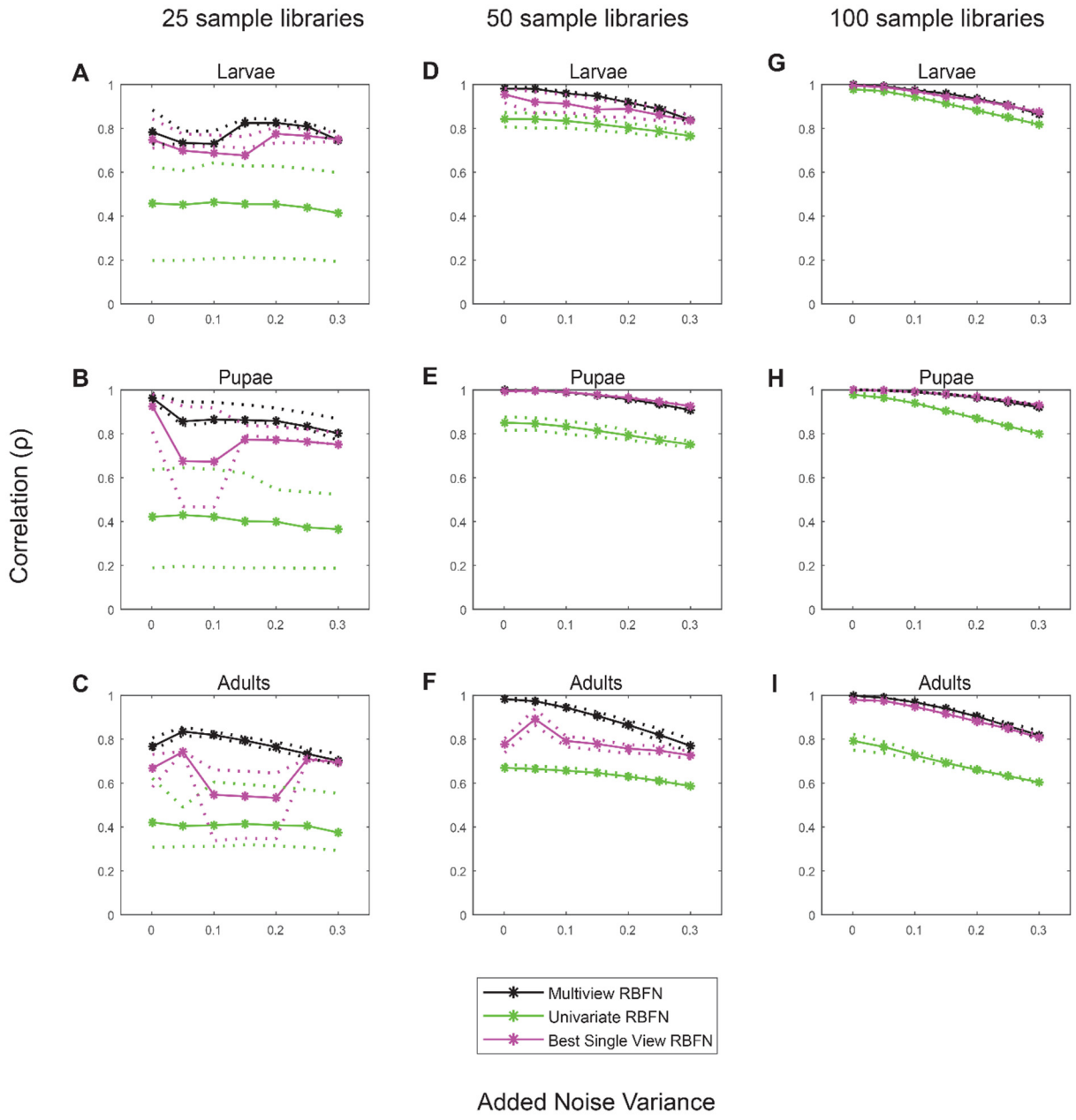



**Figure IV.16 Forecast performance (correlation) vs. noise for the flour beetle model. (A to C)** average correlation between predictions and observations for 100 randomly sampled libraries of length 25 for variables $X$, $Y$, and $Z$. **(D to F)** same as A to C but for 100 randomly sampled libraries of length 50. **(G to I)** same as A to C but for 100 randomly sampled libraries of length 100. The solid black curves are the average correlation for the Multiview RBFN approach for the top $k$ manifold reconstructions. The solid green curves are the average correlation for the univariate RBFN approach, and the solid pink curves are the average correlation using the single best view (manifold) in the RBFN autoregressive approach. The dotted lines are the upper and lower quartiles.

## IV.E Discussion and Conclusion

### IV.E.1 Computational Complexity

Ranking the manifold reconstructions in MVE algorithm involves using the Simplex Projection approach which is based on nearest neighbors. The search for the nearest neighbors in all valid manifold reconstructions in the simplex projection method leads to computational complexity of order $O(N.m.T)$, where $N$ is the number of variables, $m$ is the number of manifold reconstructions, and $T$ is the number of samples in the time series. This computational complexity is of order $O(m.p + N.m)$ for the MV-RBFN algorithm, where $m.p$ is related to time needed to find the $p$ centers (prototypes) in each of the $m$ manifold reconstructions and $N.m$ is related to the computational time required for building Gaussian radial basis functions (activation functions). Since $O(m.p + N.m) < O(N.m.T)$, MV-RBFN is of a lower computational complexity compared to MVE.

### IV.E.2 Forecast Skill

In MVE, forecasting relies on the ranking of manifold reconstruction through simplex projection's search for nearest neighbors, leading to higher computational complexity. In addition, the time-index of the true single nearest neighbors in MVE may be misplaced due to the effect of noise and therefore the MVE forecast may not accurately indicate resemblance to the target point. In contrast, MV-RBFN computes the distance-weighted average of all points in the top $k$ manifolds (Figure IV.2.A). The Gaussian radial basis function (activation function) in the hidden layer produces higher values when the distance between the data points in the input manifolds and their corresponding prototypes (centers) are small; the activation values fall off exponentially as the distance between data points and prototypes increases (*310*).

Similar to MVE, MV-RBFN exploits the pooled information contained in the top $k$ manifold reconstructions. When components of a complex dynamic system have cause-and-effect relationships with one another, relying on univariate information towards prediction of the system's dynamics does not yield good prediction skills (Figures IV.7 to IV.10). The advantage of a multiview prediction scheme is particularly evident when the time series are short and noisy, which is very common in biological and ecological data sets. The estimated nonlinear function $f(.)$ in MV-RBFN is a smooth map which produces better forecast performance than MVE due to the universal approximation property of radial basis function networks.

## IV.F Acknowledgements

Chapter IV, in full, is currently being prepared for submission for publication of the material. Masnadi-Shirazi, Maryam; Subramaniam, Shankar. The dissertation author was the primary investigator and author of this material.

# Chapter V

# Conclusions

In this dissertation we have looked into two temporal aspects of dynamic biological and ecological systems: 1) Estimating time-varying intracellular signaling pathways and regulatory interactions from data, 2) Forecasting the behavior of chaotic dynamic systems. Towards this, we worked on three different projects. In the first two projects we used the notion of Granger causality to reconstruct time-varying intracellular networks from biological data to investigate the dynamics of signaling pathways and regulatory interactions within the cell. Whereas in the last project, we developed a nonparametric approach to improve the forecasting of the dynamic behavior of complex chaotic systems by exploiting the dimensionality of the system.

In the first project, we applied the notion of Granger causality through the vector autoregressive model to develop a novel framework for reconstructing dynamic networks from large-scale multi-experiment multivariate high-throughput data sets. We used an approach based on a linear-model template and statistical hypothesis testing (t-test) of the coefficients of the model to find significant or potentially causal connections. Due to the availability of data from multiple experiments, this linear inverse problem was an overdetermined problem that could be solved via least squares estimation. We were able to predict connectivity, causality and dynamics of information flow in the progression of the phosphoprotein network.

Causal molecular mechanisms in cellular functions can only be inferred from temporal and longitudinal measurements. Few methods exist for analyzing time series data to identify distinct temporal regimes and the corresponding time-varying causal networks and mechanisms. In the second project, we developed an integrative framework that allows the detection of distinct temporal regimes using a nonparametric change point detection algorithm, along with temporally evolving directed networks that provide a comprehensive picture of the crosstalk among different molecular components (nodes) in each regime. We applied our approach to RNA-Seq time-course

data spanning nearly two cell cycles from Mouse Embryonic Fibroblast (MEF) primary cells. Due to the limited data samples, the linear autoregressive model used to infer causality was an underdetermined problem where the number of parameters exceeded the number of samples. Using LASSO and Estimation Stability with Cross Validation (ES-CV), we were able to, without any prior knowledge, extract information on duration and timing of cell cycle phases, phase-specific causal interaction of cell cycle genes as well as temporal interdependencies of biological mechanisms through a complete cell cycle. Our inference of dynamic interplay of multiple intracellular mechanisms can be used to predict time-varying cellular reponses and to explore the effect of drug dose and timing in therapeutic interventions.

In the third project, we developed a nonparametric forecasting algoirhtm, multiview radial basis function networks (MV-RBFN) that improves the forecast skill of chaotic dynamic systems in simulated ecosystem models and real data from a mesocosm experiment on plankton population. MV-RBFN exploits the dimnesionality of the complex dynamic systems by using the pooled information from attractors (manifolds) reconstructed from combination of variables and time lags as the inputs of a neural network. MV-RBFN approximates a nonlinear function $f(.)$ from the time-series data that maps the input space of past values of a dynamic system into the future values using Gaussian radial basis function netoworks. We showed that MV-RBFN outperforms univariate RBFN and multivariate RBFN approaches as well as a model-free appraoch, multiview embnedding (MVE) which is a forecasting algorithms based on empirical dynamic modeling. The strength of MV-RBFN in providing better forecast skill is particularly evident when time-series are short and noisy which is very common in ecology and biology. The estimated nonlinear function $f(.)$ in MV-RBFN is a smooth map which produces better forecast performance than MVE due to the universal approximation property of radial basis function networks.

# Bibliography


1. S. Horvath, J. Dong, Geometric interpretation of gene coexpression network analysis. *PLoS comput biol* **4**, e1000117 (2008).
2. S. Kumari, J. Nie, H.-S. Chen, H. Ma, R. Stewart, X. Li, M.-Z. Lu, W. M. Taylor, H. Wei, Evaluation of gene association methods for coexpression network construction and biological knowledge discovery. *PloS one* **7**, e50411 (2012).
3. S. Pradervand, M. R. Maurya, S. Subramaniam, Identification of signaling components required for the prediction of cytokine release in RAW 264.7 macrophages. *Genome biology* **7**, 1 (2006).
4. S. Gupta, M. R. Maurya, S. Subramaniam, Identification of crosstalk between phosphoprotein signaling pathways in RAW 264.7 macrophage cells. *PLoS Comput Biol* **6**, e1000654 (2010).
5. Y. M. Zou, Dynamics of Boolean networks. *arXiv preprint arXiv:1307.0757*, (2013).
6. M. Zou, S. D. Conzen, A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics (Oxford, England)* **21**, 71 (2005).
7. V. G. Ivancevic, T. T. Ivancevic, *Complex nonlinearity: chaos, phase transitions, topology change and path integrals*. (Springer Science & Business Media, 2008).
8. S. Daun, J. Rubin, Y. Vodovotz, G. Clermont, Equation-based models of dynamic biological systems. *Journal of critical care* **23**, 585 (2008).
9. H. Liu, M. J. Fogarty, S. M. Glaser, I. Altman, C.-h. Hsieh, L. Kaufman, A. A. Rosenberg, G. Sugihara, Nonlinear dynamic features and co-predictability of the Georges Bank fish community. *Marine Ecology Progress Series* **464**, 195 (2012).
10. C. W. Granger, Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 424 (1969).
11. T. Hunter, Protein kinases and phosphatases: the yin and yang of protein phosphorylation and signaling. *Cell* **80**, 225 (1995).
12. M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**, 27 (2000).
13. N. Omranian, B. Mueller-Roeber, Z. Nikoloski, Segmentation of biological multivariate time-series data. *Sci Rep* **5**, 8937 (2015).
14. A. Same, F. Chamroukhi, G. Govaert, P. Aknin, Model-based clustering and segmentation of time series with changes in regime. *Adv Data Anal Classi* **5**, 301 (Dec, 2011).

15. N. Dobigeon, J. Y. Tourneret, J. D. Scargle, Joint segmentation of multivariate astronomical time series: Bayesian sampling with a hierarchical model. *Ieee T Signal Proces* **55**, 414 (Feb, 2007).

16. V. Moskvina, A. Zhigljavsky, An algorithm based on singular spectrum analysis for change-point detection. *Communications in Statistics-Simulation and Computation* **32**, 319 (2003).

17. N. Golyandina, D. Stepanov, SSA-based approaches to analysis and forecast of multidimensional time series. (2005).

18. M. Masnadi-Shirazi, M. R. Maurya, S. Subramaniam, Time-varying causal inference from phosphoproteomic measurements in macrophage cells. *IEEE transactions on biomedical circuits and systems* **8**, 74 (Feb, 2014).

19. M. H. Hansen, B. Yu, Model Selection and the Principle of Minimum Description Length. *Journal of the American Statistical Association* **96**, 746 (2001/06/01, 2001).

20. R. Tibshirani, Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**, 273 (2011).

21. H. Akaike, A new look at the statistical model identification. *Automatic Control, IEEE Transactions on* **19**, 716 (1974).

22. G. Schwarz, Estimating the dimension of a model. *The annals of statistics* **6**, 461 (1978).

23. D. M. Allen, The relationship between variable selection and data agumentation and a method for prediction. *Technometrics* **16**, 125 (1974).

24. M. Stone, Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)* **36**, 111 (1974).

25. B. Yu, Stability. *Bernouli* **19**, 1484 (2013).

26. C. Lim, B. Yu, Estimation Stability with Cross Validation (ES-CV). *arXiv.org/abs/1303.3128*, (2013).

27. O. J. Reichman, M. B. Jones, M. P. Schildhauer, Challenges and opportunities of open data in ecology. *Science* **331**, 703 (2011).

28. V. Marx, Biology: The big challenges of big data. *Nature* **498**, 255 (2013).

29. D. Boyd, K. Crawford, Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society* **15**, 662 (2012).

30. E. R. Berndt, B. H. Hall, R. E. Hall, J. A. Hausman, in *Annals of Economic and Social Measurement, Volume 3, number 4*. (NBER, 1974), pp. 653-665.

31. D. L. Donoho, High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture* **1**, 32 (2000).

32. R. H. Peters, *A critique for ecology*. (Cambridge University Press, 1991).

33. O. H. Pilkey, L. Pilkey-Jarvis, *Useless arithmetic: why environmental scientists can't predict the future*. (Columbia University Press, 2007).

34. G. Sugihara, R. M. May, Nonlinear forecasting as a way of distinguishing chaos from. *Nonlinear Physics for Beginners: Fractals, Chaos, Solitons, Pattern Formation, Cellular Automata and Complex Systems*, 118 (1998).

35. F. Takens, Detecting strange attractors in turbulence. *Lecture notes in mathematics* **898**, 366 (1981).

36. E. R. Deyle, G. Sugihara, Generalized theorems for nonlinear state space reconstruction. *PLoS One* **6**, e18295 (2011).

37. T. Sauer, J. A. Yorke, M. Casdagli, Embedology. *Journal of statistical Physics* **65**, 579 (1991).

38. H. Ye, G. Sugihara, Information leverage in interconnected ecosystems: Overcoming the curse of dimensionality. *Science* **353**, 922 (2016).

39. F. Delom, E. Chevet, Phosphoprotein analysis: from proteins to proteomes. *Proteome science* **4**, 15 (2006).

40. J. A. Papin, T. Hunter, B. O. Palsson, S. Subramaniam, Reconstruction of cellular signalling networks and analysis of their properties. *Nat Rev Mol Cell Biol* **6**, 99 (Feb, 2005).

41. S. Pradervand, M. R. Maurya, S. Subramaniam, Identification of signaling components required for the prediction of cytokine release in RAW 264.7 macrophages. *Genome Biol.* **7**, R11 (2006).

42. N. Dojer, A. Gambin, A. Mizera, B. Wilczynski, J. Tiuryn, Applying dynamic Bayesian networks to perturbed gene expression data. *BMC bioinformatics* **7**, 249 (2006).

43. N. Friedman, M. Linial, I. Nachman, D. Pe'er, Using Bayesian networks to analyze expression data. *J Comput Biol* **7**, 601 (2000).

44. K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, G. P. Nolan, Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308**, 523 (Apr 22, 2005).

45. S. Haider, R. Pal, in *Genomic Signal Processing and Statistics (GENSIPS), 2011 IEEE International Workshop on*. (2011), pp. 162-165.

46. R. Pal, A. Datta, M. L. Bittner, E. R. Dougherty, Intervention in context-sensitive probabilistic Boolean networks. *Bioinformatics* **21**, 1211 (Apr 1, 2005).

47. C. Damiani, P. Lecca, in *Computer Modeling and Simulation (EMS), 2011 Fifth UKSim European Symposium on*. (2011), pp. 129-134.

48. G. Altay, Empirically determining the sample size for large-scale gene network inference algorithms. *IET Syst Biol* **6**, 35 (Apr, 2012).

49. G. Altay, F. Emmert-Streib, Inferring the conservative causal core of gene regulatory networks. *BMC systems biology* **4**, 132 (2010).

50. T. Mestl, E. Plahte, S. W. Omholt, A Mathematical Framework for Describing and Analyzing Gene Regulatory Networks. *J Theor Biol* **176**, 291 (Sep 21, 1995).

51. M. Xiong, J. Li, X. Fang, Identification of genetic networks. *Genetics* **166**, 1037 (Feb, 2004).

52. Y. H. Chang, C. Tomlin, in *Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on*. (2011), pp. 3706-3711.

53. P. Patel, B. Asbach, E. Shteyn, C. Gomez, A. Coltoff, S. Bhuyan, A. L. Tyner, R. Wagner, S. W. Blain, Brk/Protein Tyrosine Kinase 6 Phosphorylates p27KIP1, Regulating the Activity of Cyclin D–Cyclin-Dependent Kinase 4. *Molecular and cellular biology* **35**, 1506 (2015).

54. A. Fujita, J. R. Sato, H. M. Garay-Malpartida, P. A. Morettin, M. C. Sogayar, C. E. Ferreira, Time-varying modeling of gene expression regulatory networks using the wavelet dynamic vector autoregressive method. *Bioinformatics* **23**, 1623 (Jul 1, 2007).

55. H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*. (Springer-Verlag Berlin Heidelberg, 2005).

56. Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met* **57**, 289 (1995).

57. B. Asadi, M. R. Maurya, D. M. Tartakovsky, S. Subramaniam, Comparison of statistical and optimisation-based methods for data-driven network reconstruction of biochemical systems. *IET Syst Biol* **6**, 155 (Oct, 2012).

58. A. Ali, K. P. Hoeflich, J. R. Woodgett, Glycogen synthase kinase-3: properties, functions, and regulation. *Chemical reviews* **101**, 2527 (Aug, 2001).

59. D. A. Cross, D. R. Alessi, P. Cohen, M. Andjelkovich, B. A. Hemmings, Inhibition of glycogen synthase kinase-3 by insulin mediated by protein kinase B. *Nature* **378**, 785 (Dec 21-28, 1995).

60. H. C. Lee, J. N. Tsai, P. Y. Liao, W. Y. Tsai, K. Y. Lin, C. C. Chuang, C. K. Sun, W. C. Chang, H. J. Tsai, Glycogen synthase kinase 3 alpha and 3 beta have distinct functions during cardiogenesis of zebrafish embryo. *BMC developmental biology* **7**, 93 (2007).

61. S. A. Richards, J. Fu, A. Romanelli, A. Shimamura, J. Blenis, Ribosomal S6 kinase 1 (RSK1) activation requires signals dependent on and independent of the MAP kinase ERK. *Current biology : CB* **9**, 810 (Jul 29-Aug 12, 1999).

62. P. P. Roux, D. Shahbazian, H. Vu, M. K. Holz, M. S. Cohen, J. Taunton, N. Sonenberg, J. Blenis, RAS/ERK signaling promotes site-specific ribosomal protein S6 phosphorylation via RSK and stimulates cap-dependent translation. *The Journal of biological chemistry* **282**, 14056 (May 11, 2007).

63. Y. Saito, J. R. Vandenheede, P. Cohen, The mechanism by which epidermal growth factor inhibits glycogen synthase kinase 3 in A431 cells. *Biochem J* **303 ( Pt 1)**, 27 (Oct 1, 1994).

64. V. Stambolic, J. R. Woodgett, Mitogen inactivation of glycogen synthase kinase-3 beta in intact cells via serine 9 phosphorylation. *Biochem J* **303 ( Pt 3)**, 701 (Nov 1, 1994).

65. C. H. Choi, B. H. Lee, S. G. Ahn, S. H. Oh, Proteasome inhibition-induced p38 MAPK/ERK signaling regulates autophagy and apoptosis through the dual phosphorylation of glycogen synthase kinase 3beta. *Biochemical and biophysical research communications* **418**, 759 (Feb 24, 2012).

66. T. M. Thornton, G. Pedraza-Alva, B. Deng, C. D. Wood, A. Aronshtam, J. L. Clements, G. Sabio, R. J. Davis, D. E. Matthews, B. Doble, M. Rincon, Phosphorylation by p38 MAPK as an alternative pathway for GSK3beta inactivation. *Science* **320**, 667 (May 2, 2008).

67. A. Das, F. N. Salloum, L. Xi, Y. J. Rao, R. C. Kukreja, ERK phosphorylation mediates sildenafil-induced myocardial protection against ischemia-reperfusion injury in mice. *American journal of physiology. Heart and circulatory physiology* **296**, H1236 (May, 2009).

68. Q. Ding, W. Xia, J. C. Liu, J. Y. Yang, D. F. Lee, J. Xia, G. Bartholomeusz, Y. Li, Y. Pan, Z. Li, R. C. Bargou, J. Qin, C. C. Lai, F. J. Tsai, C. H. Tsai, M. C. Hung, Erk associates with and primes GSK-3beta for its inactivation resulting in upregulation of beta-catenin. *Molecular cell* **19**, 159 (Jul 22, 2005).

69. H. Buss, A. Dorrie, M. L. Schmitz, R. Frank, M. Livingstone, K. Resch, M. Kracht, Phosphorylation of serine 468 by GSK-3beta negatively regulates basal p65 NF-kappaB activity. *The Journal of biological chemistry* **279**, 49571 (Nov 26, 2004).

70. R. F. Schwabe, D. A. Brenner, Role of glycogen synthase kinase-3 in TNF-alpha-induced NF-kappaB activation and apoptosis in hepatocytes. *American journal of physiology. Gastrointestinal and liver physiology* **283**, G204 (Jul, 2002).

71. T. W. Sturgill, J. Wu, Recent progress in characterization of protein kinase cascades for phosphorylation of ribosomal protein S6. *Biochimica et biophysica acta* **1092**, 350 (May 17, 1991).

72. B. Pierrat, J. S. Correia, J. L. Mary, M. Tomas-Zuber, W. Lesslauer, RSK-B, a novel ribosomal S6 kinase family member, is a CREB kinase under dominant control of p38alpha mitogen-activated protein kinase (p38alphaMAPK). *The Journal of biological chemistry* **273**, 29661 (Nov 6, 1998).

73. T. Valovka, F. Verdier, R. Cramer, A. Zhyvoloup, T. Fenton, H. Rebholz, M. L. Wang, M. Gzhegotsky, A. Lutsyk, G. Matsuka, V. Filonenko, L. Wang, C. G. Proud, P. J. Parker, I. T. Gout, Protein kinase C phosphorylates ribosomal protein S6 kinase betaII and regulates its subcellular localization. *Molecular and cellular biology* **23**, 852 (Feb, 2003).

74. Y. Tanaka, M. V. Gavrielides, Y. Mitsuuchi, T. Fujii, M. G. Kazanietz, Protein kinase C promotes apoptosis in LNCaP prostate cancer cells through activation of p38 MAPK and inhibition of the Akt survival pathway. *The Journal of biological chemistry* **278**, 33753 (Sep 5, 2003).

75. T.-T. W. Y.-H. Hsieh, C.-Y. Huang, Y.-S. Hsieh, J.-M. Hwang, and J.-Y. Liu, p38 Mitogen-Activated Protein Kinase Pathway Is Involved in Protein Kinase Cα–Regulated Invasion in Human Hepatocellular Carcinoma Cells. *Cancer Res* (May 1, 2007).

76. G. P. Downey, J. R. Butler, J. Brumell, N. Borregaard, L. Kjeldsen, A. Q. A. K. Sue, S. Grinstein, Chemotactic peptide-induced activation of MEK-2, the predominant isoform in human neutrophils. Inhibition by wortmannin. *The Journal of biological chemistry* **271**, 21005 (Aug 30, 1996).

77. M. Haneda, S. Araki, M. Togawa, T. Sugimoto, M. Isono, R. Kikkawa, Mitogen-activated protein kinase cascade is activated in glomeruli of diabetic rats and glomerular mesangial cells cultured under high glucose conditions. *Diabetes* **46**, 847 (May, 1997).

78. P. Ping, J. Zhang, X. Cao, R. C. Li, D. Kong, X. L. Tang, Y. Qiu, S. Manchikalapudi, J. A. Auchampach, R. G. Black, R. Bolli, PKC-dependent activation of p44/p42 MAPKs during myocardial ischemia-reperfusion in conscious rabbits. *The American journal of physiology* **276**, H1468 (May, 1999).

79. M. Koss, G. R. Pfeiffer, 2nd, Y. Wang, S. T. Thomas, M. Yerukhimovich, W. A. Gaarde, C. M. Doerschuk, Q. Wang, Ezrin/radixin/moesin proteins are phosphorylated by TNF-alpha and modulate permeability increases in human pulmonary microvascular endothelial cells. *Journal of immunology* **176**, 1218 (Jan 15, 2006).

80. T. Ng, M. Parsons, W. E. Hughes, J. Monypenny, D. Zicha, A. Gautreau, M. Arpin, S. Gschmeissner, P. J. Verveer, P. I. Bastiaens, P. J. Parker, Ezrin is a downstream effector of trafficking PKC-integrin complexes involved in the control of cell motility. *The EMBO journal* **20**, 2723 (Jun 1, 2001).

81. F. Shirakawa, S. B. Mizel, In vitro activation and nuclear translocation of NF-kappa B catalyzed by cyclic AMP-dependent protein kinase and protein kinase C. *Molecular and cellular biology* **9**, 2424 (Jun, 1989).

82. D. M. Silberman, M. Zorrilla-Zubilete, G. A. Cremaschi, A. M. Genaro, Protein kinase C-dependent NF-kappaB activation is altered in T cells by chronic stress. *Cellular and molecular life sciences : CMLS* **62**, 1744 (Aug, 2005).

83. M. W. Wooten, Function for NF-kB in neuronal survival: regulation by atypical protein kinase C. *Journal of neuroscience research* **58**, 607 (Dec 1, 1999).

84. S. H. I. Brändlin, Protein Kinase C (PKC)η-mediated PKCμ activation Modulates ERK and JNK Signal Pathways. *J. Biol. Chem* **277**, 6490 (2002).

85. G. Gangarossa, E. Valjent, Regulation of the ERK pathway in the dentate gyrus by in vivo dopamine D1 receptor stimulation requires glutamatergic transmission. *Neuropharmacology* **63**, 1107 (Nov, 2012).

86. A. Carriere, H. Ray, J. Blenis, P. P. Roux, The RSK factors of activating the Ras/MAPK signaling cascade. *Frontiers in bioscience : a journal and virtual library* **13**, 4258 (2008).

87. J. C. L. Zhang, Y. Ma, W. Thomas, J. Zhang, and J. Du, Dual Pathways for Nuclear Factor κB Activation by Angiotensin II in Vascular Smooth Muscle Phosphorylation of p65 by IκB Kinase and Ribosomal Kinase. *Circ. Res.* **97**, 975 ( Nov. 2005).

88. L. Zhang, Y. Ma, J. Zhang, J. Cheng, J. Du, A new cellular signaling mechanism for angiotensin II activation of NF-kappaB: An IkappaB-independent, RSK-mediated phosphorylation of p65. *Arteriosclerosis, thrombosis, and vascular biology* **25**, 1148 (Jun, 2005).

89. A. Bhattacharyya, S. Pathak, S. Datta, S. Chattopadhyay, J. Basu, M. Kundu, Mitogen-activated protein kinases and nuclear factor-kappaB regulate Helicobacter pylori-mediated interleukin-8 release from macrophages. *Biochem J* **368**, 121 (Nov 15, 2002).

90. M. S. Hayden, S. Ghosh, Shared principles in NF-kappaB signaling. *Cell* **132**, 344 (Feb 8, 2008).

91. N. D. Perkins, Post-translational modifications regulating the activity and function of the nuclear factor kappa B pathway. *Oncogene* **25**, 6717 (Oct 30, 2006).

92. F. Chiacchiera, V. Grossi, M. Cappellari, A. Peserico, M. Simonatto, A. Germani, S. Russo, M. P. Moyer, N. Resta, S. Murzilli, C. Simone, Blocking p38/ERK crosstalk affects colorectal cancer growth by inducing apoptosis in vitro and in preclinical mouse models. *Cancer letters* **324**, 98 (Nov 1, 2012).

93. D. Fey, D. R. Croucher, W. Kolch, B. N. Kholodenko, Crosstalk and signaling switches in mitogen-activated protein kinase cascades. *Frontiers in physiology* **3**, 355 (2012).

94. Q. Liu, P. A. Hofmann, Protein phosphatase 2A-mediated cross-talk between p38 MAPK and ERK in apoptosis of cardiac myocytes. *American journal of physiology. Heart and circulatory physiology* **286**, H2204 (Jun, 2004).

95. H. Zbinden-Foncea, L. Deldicque, N. Pierre, M. Francaux, J. M. Raymackers, TLR2 and TLR4 activation induces p38 MAPK-dependent phosphorylation of S6 kinase 1 in C2C12 myotubes. *Cell biology international* **36**, 1107 (2012).

96. W. Huang, Y. Zhao, X. Zhu, Z. Cai, S. Wang, S. Yao, Z. Qi, P. Xie, Fluoxetine upregulates phosphorylated-AKT and phosphorylated-ERK1/2 proteins in neural stem cells: evidence for a crosstalk between AKT and ERK1/2 pathways. *Journal of molecular neuroscience : MN* **49**, 244 (Feb, 2013).

97. C. Rommel, B. A. Clarke, S. Zimmermann, L. Nunez, R. Rossman, K. Reid, K. Moelling, G. D. Yancopoulos, D. J. Glass, Differentiation stage-specific inhibition of the Raf-MEK-ERK pathway by Akt. *Science* **286**, 1738 (Nov 26, 1999).

98. S. Itoh, F. Itoh, M. J. Goumans, P. Ten Dijke, Signaling of transforming growth factor-beta family members through Smad proteins. *European journal of biochemistry / FEBS* **267**, 6954 (Dec, 2000).

99. I. Yakymovych, P. Ten Dijke, C. H. Heldin, S. Souchelnytskyi, Regulation of Smad signaling by protein kinase C. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **15**, 553 (Mar, 2001).

100. A. Fontayne, P. M.-C. Dang, M.-A. Gougerot-Pocidalo, J. E. Benna, Phosphorylation of p47phox Sites by PKC α, βII, δ, and ζ: Effect on Binding to p22phox and on NADPH Oxidase Activation. *Biochemistry (Mosc.)* **41**, 7743 (2002).

101. A. P. Bouin, N. Grandvaux, P. V. Vignais, A. Fuchs, p40(phox) is phosphorylated on threonine 154 and serine 315 during activation of the phagocyte NADPH oxidase. Implication of a protein kinase c-type kinase in the phosphorylation process. *The Journal of biological chemistry* **273**, 30097 (Nov 13, 1998).

102. V. H. Olavarria, J. E. Figueroa, V. Mulero, Prolactin-induced activation of phagocyte NADPH oxidase in the teleost fish gilthead seabream involves the phosphorylation of p47phox by protein kinase C. *Developmental and comparative immunology* **36**, 216 (Jan, 2012).

103. A. Dufner, M. Andjelkovic, B. M. Burgering, B. A. Hemmings, G. Thomas, Protein kinase B localization and activation differentially affect S6 kinase 1 activity and eukaryotic translation initiation factor 4E-binding protein 1 phosphorylation. *Molecular and cellular biology* **19**, 4525 (Jun, 1999).

104. C. A. Grimes, R. S. Jope, The multifaceted roles of glycogen synthase kinase 3beta in cellular signaling. *Progress in neurobiology* **65**, 391 (Nov, 2001).

105. M. Li, X. Wang, M. K. Meintzer, T. Laessig, M. J. Birnbaum, K. A. Heidenreich, Cyclic AMP promotes neuronal survival by phosphorylation of glycogen synthase kinase 3beta. *Molecular and cellular biology* **20**, 9356 (Dec, 2000).

106. M. A. Torres, H. Eldar-Finkelman, E. G. Krebs, R. T. Moon, Regulation of ribosomal S6 protein kinase-p90(rsk), glycogen synthase kinase 3, and beta-catenin in early Xenopus development. *Molecular and cellular biology* **19**, 1427 (Feb, 1999).

107. X. Fang, S. Yu, J. L. Tanyi, Y. Lu, J. R. Woodgett, G. B. Mills, Convergence of multiple signaling cascades at glycogen synthase kinase 3: Edg receptor-mediated phosphorylation and inactivation by lysophosphatidic acid through a protein kinase C-dependent intracellular pathway. *Molecular and cellular biology* **22**, 2099 (Apr, 2002).

108. N. Goode, K. Hughes, J. R. Woodgett, P. J. Parker, Differential regulation of glycogen synthase kinase-3 beta by protein kinase C isotypes. *The Journal of biological chemistry* **267**, 16878 (Aug 25, 1992).

109. K. Hughes, E. Nikolakaki, S. E. Plyte, N. F. Totty, J. R. Woodgett, Modulation of the glycogen synthase kinase-3 family by tyrosine phosphorylation. *The EMBO journal* **12**, 803 (Feb, 1993).

110. M. Sala-Valdes, A. Ursa, S. Charrin, E. Rubinstein, M. E. Hemler, F. Sanchez-Madrid, M. Yanez-Mo, EWI-2 and EWI-F link the tetraspanin web to the actin cytoskeleton through their direct association with ezrin-radixin-moesin proteins. *The Journal of biological chemistry* **281**, 19665 (Jul 14, 2006).

111. U. Tepass, FERM proteins in animal morphogenesis. *Current opinion in genetics & development* **19**, 357 (Aug, 2009).

112. G. L. Johnson, R. Lapadat, Mitogen-activated protein kinase pathways mediated by ERK, JNK, and p38 protein kinases. *Science* **298**, 1911 (Dec 6, 2002).

113. C. H. Heldin, K. Miyazono, P. ten Dijke, TGF-beta signalling from cell membrane to nucleus through SMAD proteins. *Nature* **390**, 465 (Dec 4, 1997).

114. J. Massague, TGF-beta signal transduction. *Annual review of biochemistry* **67**, 753 (1998).

115. B. Foxwell, K. Browne, J. Bondeson, C. Clarke, R. de Martin, F. Brennan, M. Feldmann, Efficient adenoviral infection with IkappaB alpha reveals that macrophage tumor necrosis factor alpha production in rheumatoid arthritis is NF-kappaB dependent. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 8211 (Jul 7, 1998).

116. K. S. Doran, G. Y. Liu, V. Nizet, Group B streptococcal beta-hemolysin/cytolysin activates neutrophil signaling pathways in brain endothelium and contributes to development of meningitis. *The Journal of clinical investigation* **112**, 736 (Sep, 2003).

117. E. K. Kim, E. J. Choi, Pathological roles of MAPK signaling pathways in human diseases. *Biochimica et biophysica acta* **1802**, 396 (Apr, 2010).

118. B. Asadi, M. Maurya, D. Tartakovsky, S. Subramaniam, Comparison of statistical and optimisation-based methods for data-driven network reconstruction of biochemical systems. *IET systems biology* **6**, 155 (2012).

119. N. Omranian, J. M. Eloundou-Mbebi, B. Mueller-Roeber, Z. Nikoloski, Gene regulatory network inference using fused LASSO on multiple data sets. *Scientific reports* **6**, 20533 (2016).

120. J. Schäfer, K. Strimmer, An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics (Oxford, England)* **21**, 754 (2005).

121. C. O. Daub, R. Steuer, J. Selbig, S. Kloska, Estimating mutual information using B-spline functions–an improved similarity measure for analysing gene expression data. *BMC bioinformatics* **5**, 1 (2004).

122. A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. D. Favera, A. Califano, ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics* **7**, S7 (2006).

123. S. A. Kauffman, Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of theoretical biology* **22**, 437 (1969).

124. T. Akutsu, S. Miyano, S. Kuhara, Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics (Oxford, England)* **16**, 727 (2000).

125. L. Raeymaekers, Dynamics of Boolean networks controlled by biologically meaningful functions. *Journal of Theoretical Biology* **218**, 331 (2002).

126. V. Bucci, B. Tzen, N. Li, M. Simmons, T. Tanoue, E. Bogart, L. Deng, V. Yeliseyev, M. L. Delaney, Q. Liu, MDSINE: Microbial Dynamical Systems INference Engine for microbiome time-series analyses. *Genome biology* **17**, 121 (2016).

127. N. Friedman, M. Linial, I. Nachman, D. Pe'er, Using Bayesian networks to analyze expression data. *Journal of computational biology* **7**, 601 (2000).

128. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* **29**, 15 (2013).

129. S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, C. K. Glass, Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell* **38**, 576 (2010).

130. N. Golyandina, V. Nekrutkin, A. A. Zhigljavsky, *Analysis of time series structure: SSA and related techniques*. (CRC press, 2001).

131. C. W. J. Granger, Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica* **37**, 424 (1969).

132. A. Shojaie, G. Michailidis, Discovering graphical Granger causality using the truncating lasso penalty. *Bioinformatics (Oxford, England)* **26**, i517 (2010).

133. H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*. (Springer Berlin Heidelberg, 2006).

134. D. L. Donoho, Compressed sensing. *Information Theory, IEEE Transactions on* **52**, 1289 (2006).

135. E. J. Candes, J. K. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics* **59**, 1207 (2006).

136. S. Boyd, L. Vandenberghe, *Convex optimization*. (Cambridge university press, 2004).

137. S. Cooper, Mammalian cells are not synchronized in G1 phase by starvation or inhibition: considerations of the fundamental concept of G1 phase synchronization. *Cell proliferation* **31**, 9 (1998).

138. S. Cooper, Rethinking synchronization of mammalian cells for cell cycle analysis. *Cellular and Molecular Life Sciences CMLS* **60**, 1099 (2003).

139. S. M. Boker, J. L. Rotondo, M. Xu, K. King, Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological Methods* **7**, 338 (2002).

140. D. M. Powers, Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. (2011).

141. D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, The Reactome pathway knowledgebase. *Nucleic acids research* **42**, D472 (2014).

142. M. Haubold, A. Weise, H. Stephan, N. Dünker, Bone morphogenetic protein 4 (BMP4) signaling in retinoblastoma cells. *International journal of biological sciences* **6**, 700 (2010).

143. M. Nakanishi, Y. Kaneko, H. Matsushime, K. Ikeda, Direct interaction of p21 cyclin-dependent kinase inhibitor with the retinoblastoma tumor suppressor protein. *Biochemical and biophysical research communications* **263**, 35 (1999).

144. C. J. Sherr, J. M. Roberts, CDK inhibitors: positive and negative regulators of G1-phase progression. *Genes & development* **13**, 1501 (1999).

145. C. J. Sherr, F. McCormick, The RB and p53 pathways in cancer. *Cancer cell* **2**, 103 (2002).

146. P. Ji, H. Jiang, K. Rekhtman, J. Bloom, M. Ichetovkin, M. Pagano, L. Zhu, An Rb-Skp2-p27 pathway mediates acute cell cycle inhibition by Rb and is retained in a partial-penetrance Rb mutant. *Molecular cell* **16**, 47 (2004).

147. U. K. Binné, M. K. Classon, F. A. Dick, W. Wei, M. Rape, W. G. Kaelin, A. M. Näär, N. J. Dyson, Retinoblastoma protein and anaphase-promoting complex physically interact and functionally cooperate during cell-cycle exit. *Nature cell biology* **9**, 225 (2007).

148. I. Neganova, M. Lako, G1 to S phase cell cycle transition in somatic and embryonic stem cells. *Journal of anatomy* **213**, 30 (2008).

149. P. J. Welch, J. Y. Wang, A C-terminal protein-binding domain in the retinoblastoma protein regulates nuclear c-Abl tyrosine kinase in the cell cycle. *Cell* **75**, 779 (1993).

150. E. S. Knudsen, J. Y. Wang, Differential regulation of retinoblastoma protein function by specific Cdk phosphorylation sites. *Journal of Biological Chemistry* **271**, 8313 (1996).

151. K. Helin, Regulation of cell proliferation by the E2F transcription factors. *Current opinion in genetics & development* **8**, 28 (1998).

152. R. A. Weinberg, The retinoblastoma protein and cell cycle control. *Cell* **81**, 323 (1995).

153. C. Attwooll, E. L. Denchi, K. Helin, The E2F family: specific functions and overlapping interests. *The EMBO journal* **23**, 4709 (2004).

154. K. Helin, C.-L. Wu, A. R. Fattaey, J. A. Lees, B. D. Dynlacht, C. Ngwu, E. Harlow, Heterodimerization of the transcription factors E2F-1 and DP-1 leads to cooperative trans-activation. *Genes & Development* **7**, 1850 (1993).

155. K. Ohtani, R. Iwanaga, M. Nakamura, M.-a. Ikeda, N. Yabuta, H. Tsuruga, H. Nojima, Cell growth-regulated expression of mammalian MCM5 and MCM6 genes mediated by the transcription factor E2F. *Oncogene* **18**, 2299 (1999).

156. Y. Arata, M. Fujita, K. Ohtani, S. Kijima, J.-y. Kato, Cdk2-dependent and-independent pathways in E2F-mediated S phase induction. *Journal of Biological Chemistry* **275**, 6337 (2000).

157. K. Ohtani, A. Tsujimoto, M.-a. Ikeda, M. Nakamura, Regulation of cell growth-dependent expression of mammalian CDC6 gene by the cell cycle transcription factor E2F. *Oncogene* **17**, 1777 (1998).

158. G. Bosco, W. Du, T. L. Orr-Weaver, DNA replication control through interaction of E2F–RB and the origin recognition complex. *Nature cell biology* **3**, 289 (2001).

159. Z. Yan, J. DeGregori, R. Shohet, G. Leone, B. Stillman, J. R. Nevins, R. S. Williams, Cdc6 is regulated by E2F and is essential for DNA replication in mammalian cells. *Proceedings of the National Academy of Sciences* **95**, 3603 (1998).

160. R. Stevens, M. Grelon, D. Vezon, J. Oh, P. Meyer, C. Perennes, S. Domenichini, C. Bergounioux, A CDC45 homolog in Arabidopsis is essential for meiosis, as shown by RNA interference–induced gene silencing. *The Plant Cell* **16**, 99 (2004).

161. M. Vidal, P. Braun, E. Chen, J. D. Boeke, E. Harlow, Genetic characterization of a mammalian protein-protein interaction domain by using a yeast reverse two-hybrid system. *Proceedings of the National Academy of Sciences* **93**, 10321 (1996).

162. K. Martin, D. Trouche, C. Hagemeier, T. Kouzarides, Regulation of transcription by E2F1/DP1. *Journal of Cell Science* **1995**, 91 (1995).

163. C.-L. Wu, L. R. Zukerberg, C. Ngwu, E. Harlow, J. A. Lees, In vivo association of E2F and DP family proteins. *Molecular and cellular biology* **15**, 2536 (1995).

164. F. Chan, J. Zhang, L. Cheng, D. N. Shapiro, A. Winoto, Identification of human and mouse p19, a novel CDK4 and CDK6 inhibitor with homology to p16ink4. *Molecular and cellular biology* **15**, 2682 (1995).

165. D. O. Morgan, *The cell cycle: principles of control*. (New Science Press, 2007).

166. M. F. Roussel, The INK4 family of cell cycle inhibitors in cancer. *Oncogene* **18**, 5311 (1999).

167. V. Masciullo, K. Khalili, A. Giordano, The Rb family of cell cycle regulatory factors: clinical implications. *International journal of oncology* **17**, 897 (2000).

168. R. L. Beijersbergen, L. Carlée, R. M. Kerkhoven, R. Bernards, Regulation of the retinoblastoma protein-related p107 by G1 cyclin complexes. *Genes & development* **9**, 1340 (1995).

169. X. Leng, M. Noble, P. D. Adams, J. Qin, J. W. Harper, Reversal of growth suppression by p107 via direct phosphorylation by cyclin D1/cyclin-dependent kinase 4. *Molecular and cellular biology* **22**, 2242 (2002).

170. J. Stanelle, T. Stiewe, C. C. Theseling, M. Peter, B. M. Pützer, Gene expression changes in response to E2F1 activation. *Nucleic acids research* **30**, 1859 (2002).

171. S. Inoshita, Y. Terada, O. Nakashima, M. Kuwahara, S. Sasaki, F. Marumo, Roles of E2F1 in mesangial cell proliferation in vitro. *Kidney international* **56**, 2085 (1999).

172. T. C. Ko, H. M. Sheng, D. Reisman, E. A. Thompson, R. D. Beauchamp, Transforming growth factor-beta 1 inhibits cyclin D1 expression in intestinal epithelial cells. *Oncogene* **10**, 177 (1995).

173. M. Li, P. Zhang, The function of APC/CCdh1 in cell cycle and beyond. *Cell division* **4**, 2 (2009).

174. Y. Geng, W. Whoriskey, M. Y. Park, R. T. Bronson, R. H. Medema, T. Li, R. A. Weinberg, P. Sicinski, Rescue of cyclin D1 deficiency by knockin cyclin E. *Cell* **97**, 767 (1999).

175. T. J. Takara, S. P. Bell, Multiple Cdt1 molecules act at each origin to load replication□ competent Mcm2–7 helicases. *The EMBO journal* **30**, 4885 (2011).

176. M. Kneissl, V. Pütter, A. A. Szalay, F. Grummt, Interaction and assembly of murine pre-replicative complex proteins in yeast and mouse cells. *Journal of molecular biology* **327**, 111 (2003).

177. P. Saha, J. Chen, K. C. Thome, S. J. Lawlis, Z.-h. Hou, M. Hendricks, J. D. Parvin, A. Dutta, Human CDC6/Cdc18 associates with Orc1 and cyclin-cdk and is selectively eliminated from the nucleus at the onset of S phase. *Molecular and Cellular Biology* **18**, 2758 (1998).

178. G. Abdurashidova, M. B. Danailov, A. Ochem, G. Triolo, V. Djeliova, S. Radulescu, A. Vindigni, S. Riva, A. Falaschi, Localization of proteins bound to a replication origin of human DNA along the cell cycle. *The EMBO journal* **22**, 4294 (2003).

179. M. J. Ravitz, S. Yan, C. Dolce, A. J. Kinniburgh, C. E. Wenner, Differential regulation of p27 and cyclin D1 by TGF□β and EGF in C3H 10T1/2 mouse fibroblasts. *Journal of cellular physiology* **168**, 510 (1996).

180. H. Xu, Z. Zhang, M. Li, R. Zhang, MDM2 promotes proteasomal degradation of p21Waf1 via a conformation change. *Journal of Biological Chemistry* **285**, 18407 (2010).

181. Z. Zhang, H. Wang, M. Li, S. Agrawal, X. Chen, R. Zhang, MDM2 is a negative regulator of p21WAF1/CIP1, independent of p53. *Journal of Biological Chemistry* **279**, 16000 (2004).

182. J. Pomerantz, N. Schreiber-Agus, N. J. Liégeois, A. Silverman, L. Alland, L. Chin, J. Potes, K. Chen, I. Orlow, H.-W. Lee, The Ink4a tumor suppressor gene product, p19 Arf, interacts with MDM2 and neutralizes MDM2's inhibition of p53. *Cell* **92**, 713 (1998).

183. P. A. Clark, S. Llanos, G. Peters, Multiple interacting domains contribute to p14^ A^ R^ F mediated inhibition of MDM2. *Oncogene* **21**, 4498 (2002).

184. Y. Zhang, Y. Xiong, W. G. Yarbrough, ARF promotes MDM2 degradation and stabilizes p53: ARF-INK4a locus deletion impairs both the Rb and p53 tumor suppression pathways. *Cell* **92**, 725 (1998).

185. T. Kamijo, J. D. Weber, G. Zambetti, F. Zindy, M. F. Roussel, C. J. Sherr, Functional and physical interactions of the ARF tumor suppressor with p53 and Mdm2. *Proceedings of the National Academy of Sciences* **95**, 8292 (1998).

186. K. Macdonald, M. R. Bennett, cdc25A is necessary but not sufficient for optimal c-myc–induced apoptosis and cell proliferation of vascular smooth muscle cells. *Circulation research* **84**, 820 (1999).

187. J. Seoane, C. Pouponnot, P. Staller, M. Schader, M. Eilers, J. Massagué, TGFβ influences Myc, Miz-1 and Smad to control the CDK inhibitor p15INK4b. *Nature cell biology* **3**, 400 (2001).

188. P. Staller, K. Peukert, A. Kiermaier, J. Seoane, J. Lukas, H. Karsunky, T. Möröy, J. Bartek, J. Massagué, F. Hänel, Repression of p15INK4b expression by Myc through association with Miz-1. *Nature cell biology* **3**, 392 (2001).

189. W. Lutz, J. Leon, M. Eilers, Contributions of Myc to tumorigenesis. *Biochimica Et Biophysica Acta (BBA)-Reviews on Cancer* **1602**, 61 (2002).

190. B. M. Sicari, R. Troxell, F. Salim, M. Tanwir, K. K. Takane, N. Fiaschi-Taesch, c-myc and skp2 coordinate p27 degradation, vascular smooth muscle proliferation, and neointima

formation induced by the parathyroid hormone-related protein. *Endocrinology* **153**, 861 (2011).

191. F. Faiola, X. Liu, S. Lo, S. Pan, K. Zhang, E. Lymar, A. Farina, E. Martinez, Dual regulation of c-Myc by p300 via acetylation-dependent control of Myc protein turnover and coactivation of Myc-induced transcription. *Molecular and cellular biology* **25**, 10220 (2005).

192. H. N. Rajabi, S. Baluchamy, S. Kolli, A. Nag, R. Srinivas, P. Raychaudhuri, B. Thimmapaya, Effects of depletion of CREB-binding protein on c-Myc regulation and cell cycle G1-S transition. *Journal of Biological Chemistry* **280**, 361 (2005).

193. X.-H. Feng, Y.-Y. Liang, M. Liang, W. Zhai, X. Lin, Direct interaction of c-Myc with Smad2 and Smad3 to inhibit TGF-β-mediated induction of the CDK inhibitor p15 Ink4B. *Molecular cell* **9**, 133 (2002).

194. D. Koinuma, S. Tsutsumi, N. Kamimura, H. Taniguchi, K. Miyazawa, M. Sunamura, T. Imamura, K. Miyazono, H. Aburatani, Chromatin immunoprecipitation on microarray analysis of Smad2/3 binding sites reveals roles of ETS1 and TFAP2A in transforming growth factor β signaling. *Molecular and cellular biology* **29**, 172 (2009).

195. M. Wang, H. Huang, Y. Chen, Smad2/3 is involved in growth inhibition of mouse embryonic palate mesenchymal cells induced by all□trans retinoic acid. *Birth Defects Research Part A: Clinical and Molecular Teratology* **85**, 780 (2009).

196. B. Zhang, X. Chen, S. Bae, K. Singh, M. Washington, P. Datta, Loss of Smad4 in colorectal cancer induces resistance to 5-fluorouracil through activating Akt pathway. *British journal of cancer* **110**, 946 (2014).

197. M. Alotaibi, Y. Kitase, C. Shuler, smad2 Overexpression reduces the proliferation of the Junctional Epithelium. *Journal of dental research* **93**, 898 (2014).

198. W. Liu, G. Wu, W. Li, D. Lobur, Y. Wan, Cdh1-anaphase-promoting complex targets Skp2 for destruction in transforming growth factor β-induced growth inhibition. *Molecular and cellular biology* **27**, 2967 (2007).

199. A. Koff, A. Giordano, D. Desai, K. Yamashita, J. W. Harper, S. Elledge, T. Nishimoto, D. O. Morgan, B. R. Franza, J. M. Roberts, Formation and activation of a cyclin E-cdk2 complex during the G1 phase of the human cell cycle. *Science* **257**, 1689 (1992).

200. J. Li, M. Deng, Q. Wei, T. Liu, X. Tong, X. Ye, Phosphorylation of MCM3 protein by cyclin E/cyclin-dependent kinase 2 (Cdk2) regulates its function in cell cycle. *Journal of Biological Chemistry* **286**, 39776 (2011).

201. R. L. Ferguson, J. L. Maller, Centrosomal localization of cyclin E-Cdk2 is required for initiation of DNA synthesis. *Current Biology* **20**, 856 (2010).

202. W. Jiang, N. J. Wells, T. Hunter, Multistep regulation of DNA replication by Cdk phosphorylation of HsCdc6. *Proceedings of the National Academy of Sciences* **96**, 6193 (1999).

203. H. Zalzali, B. Nasr, M. Harajly, H. Basma, F. Ghamloush, S. Ghayad, N. Ghanem, G. I. Evan, R. Saab, CDK2 Transcriptional Repression Is an Essential Effector in p53-Dependent Cellular Senescence—Implications for Therapeutic Intervention. *Molecular Cancer Research* **13**, 29 (2015).

204. K. R. Nevis, M. Cordeiro-Stone, J. G. Cook, Origin licensing and p53 status regulate Cdk2 activity during G1. *Cell Cycle* **8**, 1952 (2009).

205. M. Vairapandi, N. Azam, A. G. Balliet, B. Hoffman, D. A. Liebermann, Characterization of MyD118, Gadd45, and proliferating cell nuclear antigen (PCNA) interacting domains PCNA impedes MyD118 and Gadd45-mediated negative growth control. *Journal of Biological Chemistry* **275**, 16810 (2000).

206. P. A. Hall, J. M. Kearsey, P. J. Coates, D. G. Norman, E. Warbrick, L. S. Cox, Characterisation of the interaction between PCNA and Gadd45. *Oncogene* **10**, 2427 (1995).

207. I.-T. Chen, M. L. Smith, P. M. O'Connor, A. J. Fornace Jr, Direct interaction of Gadd45 with PCNA and evidence for competitive interaction of Gadd45 and p21Waf1/Cip1 with PCNA. *Oncogene* **11**, 1931 (1995).

208. M. L. Smith, I.-T. Chen, Q. Zhan, I. Bae, C.-Y. Chen, T. M. Gilmer, M. B. Kastan, P. M. O'Connor, A. J. Fornace, Interaction of the p53-regulated protein Gadd45 with proliferating cell nuclear antigen. *Science* **266**, 1376 (1994).

209. G. F. Morris, J. R. Bischoff, M. B. Mathews, Transcriptional activation of the human proliferating-cell nuclear antigen promoter by p53. *Proceedings of the National Academy of Sciences* **93**, 895 (1996).

210. C. V. Shivakumar, D. R. Brown, S. Deb, S. P. Deb, Wild-type human p53 transactivates the human proliferating cell nuclear antigen promoter. *Molecular and Cellular Biology* **15**, 6785 (1995).

211. J. Xu, G. F. Morris, p53-mediated regulation of proliferating cell nuclear antigen expression in cells exposed to ionizing radiation. *Molecular and cellular biology* **19**, 12 (1999).

212. Z. Goldberg, R. V. Sionov, M. Berger, Y. Zwang, R. Perets, R. A. Van Etten, M. Oren, Y. Taya, Y. Haupt, Tyrosine phosphorylation of Mdm2 by c□Abl: implications for p53 regulation. *The EMBO journal* **21**, 3715 (2002).

213. R. V. Sionov, E. Moallem, M. Berger, A. Kazaz, O. Gerlitz, Y. Ben-Neriah, M. Oren, Y. Haupt, c-Abl neutralizes the inhibitory effect of Mdm2 on p53. *Journal of Biological Chemistry* **274**, 8371 (1999).

214. V. Zuckerman, K. Lenos, G. M. Popowicz, I. Silberman, T. Grossman, J.-C. Marine, T. A. Holak, A. G. Jochemsen, Y. Haupt, c-Abl phosphorylates Hdmx and regulates its interaction with p53. *Journal of Biological Chemistry* **284**, 4031 (2009).

215. Y.-M. Yoon, K.-H. Baek, S.-J. Jeong, H.-J. Shin, G.-H. Ha, A.-H. Jeon, S.-G. Hwang, J.-S. Chun, C.-W. Lee, WD repeat□containing mitotic checkpoint proteins act as transcriptional repressors during interphase. *FEBS letters* **575**, 23 (2004).

216. D. M. Goudelock, K. Jiang, E. Pereira, B. Russell, Y. Sanchez, Regulatory interactions between the checkpoint kinase Chk1 and the proteins of the DNA-dependent protein kinase complex. *Journal of Biological Chemistry* **278**, 29940 (2003).

217. H. Tian, A. T. Faje, S. L. Lee, T. J. Jorgensen, Radiation-induced phosphorylation of Chk1 at S345 is associated with p53-dependent cell cycle arrest pathways. *Neoplasia* **4**, 171 (2002).

218. S.-Y. Shieh, J. Ahn, K. Tamai, Y. Taya, C. Prives, The human homologs of checkpoint kinases Chk1 and Cds1 (Chk2) phosphorylate p53 at multiple DNA damage-inducible sites. *Genes & development* **14**, 289 (2000).

219. N. Nitani, K.-i. Nakamura, C. Nakagawa, H. Masukata, T. Nakagawa, Regulation of DNA replication machinery by Mrc1 in fission yeast. *Genetics* **174**, 155 (2006).

220. L. Yin, A. M. Locovei, G. D'Urso, Activation of the DNA damage checkpoint in mutants defective in DNA replication initiation. *Molecular biology of the cell* **19**, 4374 (2008).

221. A. M. Locovei, L. Yin, G. D'Urso, A genetic screen for replication initiation defective (rid) mutants in Schizosaccharomyces pombe. *Cell division* **5**, 1 (2010).

222. P. Liu, L. R. Barkley, T. Day, X. Bi, D. M. Slater, M. G. Alexandrow, H.-P. Nasheuer, C. Vaziri, The Chk1-mediated S-phase checkpoint targets initiation factor Cdc45 via a Cdc25A/Cdk2-independent mechanism. *Journal of Biological Chemistry* **281**, 30631 (2006).

223. R. G. Syljuåsen, C. S. Sørensen, L. T. Hansen, K. Fugger, C. Lundin, F. Johansson, T. Helleday, M. Sehested, J. Lukas, J. Bartek, Inhibition of human Chk1 causes increased initiation of DNA replication, phosphorylation of ATR targets, and DNA breakage. *Molecular and cellular biology* **25**, 3553 (2005).

224. D. Sampath, Z. Shi, W. Plunkett, Inhibition of cyclin-dependent kinase 2 by the Chk1-Cdc25A pathway during the S-phase checkpoint activated by fludarabine: dysregulation by 7-hydroxystaurosporine. *Molecular pharmacology* **62**, 680 (2002).

225. Y. Zhu, C. Alvarez, R. Doll, H. Kurata, X. M. Schebye, D. Parry, E. Lees, Intra-S-phase checkpoint activation by direct CDK2 inhibition. *Molecular and cellular biology* **24**, 6268 (2004).

226. J. Lee, A. Kumagai, W. G. Dunphy, Positive regulation of Wee1 by Chk1 and 14-3-3 proteins. *Molecular biology of the cell* **12**, 551 (2001).

227. M. J. O'Connell, J. M. Raleigh, H. M. Verkade, P. Nurse, Chk1 is a wee1 kinase in the G2 DNA damage checkpoint inhibiting cdc2 by Y15 phosphorylation. *The EMBO journal* **16**, 545 (1997).

228. P. Saini, Y. Li, M. Dobbelstein, Wee1 is required to sustain ATR/Chk1 signaling upon replicative stress. *Oncotarget* **6**, 13072 (2015).

229. E. Berkovich, D. Ginsberg, ATM is a target for positive regulation by E2F-1. *Oncogene* **22**, 161 (2003).

230. J. Wu, X. Zhang, L. Zhang, C.-Y. Wu, A. H. Rezaeian, C.-H. Chan, J.-M. Li, J. Wang, Y. Gao, F. Han, Skp2 E3 ligase integrates ATM activation and homologous recombination repair by ubiquitinating NBS1. *Molecular cell* **46**, 351 (2012).

231. J. Kim, N. Kakusho, M. Yamada, Y. Kanoh, N. Takemoto, H. Masai, Cdc7 kinase mediates Claspin phosphorylation in DNA replication checkpoint. *Oncogene* **27**, 3475 (2008).

232. M. Scian, E. Carchman, L. Mohanraj, K. Stagliano, M. Anderson, D. Deb, B. Crane, T. Kiyono, B. Windle, S. Deb, Wild-type p53 and p73 negatively regulate expression of proliferation related genes. *Oncogene* **27**, 2583 (2008).

233. A. Duursma, R. Agami, p53-Dependent regulation of Cdc6 protein stability controls cellular proliferation. *Molecular and cellular biology* **25**, 6937 (2005).

234. D. M. Price, Z. Jin, S. Rabinovitch, S. D. Campbell, Ectopic expression of the Drosophila Cdk1 inhibitory kinases, Wee1 and Myt1, interferes with the second mitotic wave and disrupts pattern formation during eye development. *Genetics* **161**, 721 (2002).

235. W. Hu, Z. Feng, A. J. Levine, The regulation of multiple p53 stress responses is mediated through MDM2. *Genes & cancer* **3**, 199 (2012).

236. Z.-C. Yu, Y.-F. Huang, S.-Y. Shieh, Requirement for human Mps1/TTK in oxidative DNA damage repair and cell survival through MDM2 phosphorylation. *Nucleic acids research* **44**, 1133 (2016).

237. M. Maric, T. Maculins, G. De Piccoli, K. Labib, Cdc48 and a ubiquitin ligase drive disassembly of the CMG helicase at the end of DNA replication. *Science* **346**, 1253596 (2014).

238. I. Bruck, D. L. Kaplan, GINS and Sld3 compete with one another for Mcm2-7 and Cdc45 binding. *Journal of Biological Chemistry* **286**, 14157 (2011).

239. C. F. Hardy, Identification of Cdc45p, an essential factor required for DNA replication. *Gene* **187**, 239 (1997).

240. O. M. Aparicio, D. M. Weinstein, S. P. Bell, Components and dynamics of DNA replication complexes in S. cerevisiae: redistribution of MCM proteins and Cdc45p during S phase. *Cell* **91**, 59 (1997).

241. T. Tanaka, D. Knapp, K. Nasmyth, Loading of an Mcm protein onto DNA replication origins is regulated by Cdc6p and CDKs. *Cell* **90**, 649 (1997).

242. L. Zou, B. Stillman, Formation of a preinitiation complex by S-phase cyclin CDK-dependent loading of Cdc45p onto chromatin. *Science* **280**, 593 (1998).

243. T. Masuda, S. Mimura, H. Takisawa, CDK□and Cdc45□dependent priming of the MCM complex on chromatin during S□phase in Xenopus egg extracts: possible activation of MCM helicase by association with Cdc45. *Genes to Cells* **8**, 145 (2003).

244. C. L. Lunn, J. C. Chrivia, J. J. Baldassare, Activation of Cdk2/Cyclin E complexes is dependent on the origin of replication licensing factor Cdc6 in mammalian cells. *Cell Cycle* **9**, 4533 (2010).

245. B. O. Petersen, J. Lukas, C. S. Sørensen, J. Bartek, K. Helin, Phosphorylation of mammalian CDC6 by cyclin A/CDK2 regulates its subcellular localization. *The EMBO Journal* **18**, 396 (1999).

246. J. Knockleby, B. J. Kim, H. Lee, Cdk1 prevents DNA rereplication in G2/M by phosphorylating and facilitating the removal of Cdc7 from chromatin at the end of S phase. *Cancer Research* **73**, 575 (2013).

247. R. Nougarède, F. Della Seta, P. Zarzov, E. Schwob, Hierarchy of S-phase-promoting factors: yeast Dbf4-Cdc7 kinase requires prior S-phase cyclin-dependent kinase activation. *Molecular and Cellular Biology* **20**, 3795 (2000).

248. Y. Yamagishi, C.-H. Yang, Y. Tanno, Y. Watanabe, MPS1/Mph1 phosphorylates the kinetochore protein KNL1/Spc7 to recruit SAC components. *Nature cell biology* **14**, 746 (2012).

249. N. London, S. Biggins, Mad1 kinetochore recruitment by Mps1-mediated phosphorylation of Bub1 signals the spindle checkpoint. *Genes & development* **28**, 140 (2014).

250. A. R. Tipton, W. Ji, B. Sturt-Gillespie, M. E. Bekier, K. Wang, W. R. Taylor, S.-T. Liu, Monopolar spindle 1 (MPS1) kinase promotes production of closed MAD2 (C-MAD2) conformer and assembly of the mitotic checkpoint complex. *Journal of Biological Chemistry* **288**, 35149 (2013).

251. H. Huang, J. Hittle, F. Zappacosta, R. S. Annan, A. Hershko, T. J. Yen, Phosphorylation sites in BubR1 that regulate kinetochore attachment, tension, and mitotic exit. *The Journal of cell biology* **183**, 667 (2008).

252. E. Chiroli, G. Rancati, I. Catusi, G. Lucchini, S. Piatti, Cdc14 inhibition by the spindle assembly checkpoint prevents unscheduled centrosome separation in budding yeast. *Molecular biology of the cell* **20**, 2626 (2009).

253. G. Fang, Checkpoint protein BubR1 acts synergistically with Mad2 to inhibit anaphase-promoting complex. *Molecular biology of the cell* **13**, 755 (2002).

254. A. R. Tipton, K. Wang, L. Link, J. J. Bellizzi, H. Huang, T. Yen, S.-T. Liu, BUBR1 and closed MAD2 (C-MAD2) interact directly to assemble a functional mitotic checkpoint complex. *Journal of Biological Chemistry* **286**, 21173 (2011).

255. A. De Antoni, C. G. Pearson, D. Cimini, J. C. Canman, V. Sala, L. Nezi, M. Mapelli, L. Sironi, M. Faretta, E. D. Salmon, The Mad1/Mad2 complex as a template for Mad2 activation in the spindle assembly checkpoint. *Current Biology* **15**, 214 (2005).

256. D. M. Brady, K. G. Hardwick, Complex formation between Mad1p, Bub1p and Bub3p is crucial for spindle checkpoint function. *Current Biology* **10**, 675 (2000).

257. R. Fraschini, A. Beretta, L. Sironi, A. Musacchio, G. Lucchini, S. Piatti, Bub3 interaction with Mad2, Mad3 and Cdc20 is mediated by WD40 repeats and does not require intact kinetochores. *The EMBO journal* **20**, 6648 (2001).

258. V. Rossio, E. Galati, M. Ferrari, A. Pellicioli, T. Sutani, K. Shirahige, G. Lucchini, S. Piatti, The RSC chromatin-remodeling complex influences mitotic exit and adaptation to the

spindle assembly checkpoint by controlling the Cdc14 phosphatase. *The Journal of cell biology* **191**, 981 (2010).

259. L. Yu, W. Guo, S. Zhao, J. Tang, J. Liu, Knockdown of Mad2 induces osteosarcoma cell apoptosis-involved Rad21 cleavage. *Journal of Orthopaedic Science* **16**, 814 (2011).

260. M. J. Kallio, V. A. Beardmore, J. Weinstein, G. J. Gorbsky, Rapid microtubule-independent dynamics of Cdc20 at kinetochores and centrosomes in mammalian cells. *The Journal of cell biology* **158**, 841 (2002).

261. J. Nilsson, M. Yekezare, J. Minshull, J. Pines, The APC/C maintains the spindle assembly checkpoint by targeting Cdc20 for destruction. *Nature cell biology* **10**, 1411 (2008).

262. V. Sudakin, G. K. Chan, T. J. Yen, Checkpoint inhibition of the APC/C in HeLa cells is mediated by a complex of BUBR1, BUB3, CDC20, and MAD2. *The Journal of cell biology* **154**, 925 (2001).

263. L. A. Malureanu, K. B. Jeganathan, M. Hamada, L. Wasilewski, J. Davenport, J. M. van Deursen, BubR1 N terminus acts as a soluble inhibitor of cyclin B degradation by APC/C Cdc20 in interphase. *Developmental cell* **16**, 118 (2009).

264. H. Izumi, Y. Matsumoto, T. Ikeuchi, H. Saya, T. Kajii, S. Matsuura, BubR1 localizes to centrosomes and suppresses centrosome amplification via regulating Plk1 activity in interphase cells. *Oncogene* **28**, 2806 (2009).

265. S. Matsumura, F. Toyoshima, E. Nishida, Polo-like kinase 1 facilitates chromosome alignment during prometaphase through BubR1. *Journal of Biological Chemistry* **282**, 15217 (2007).

266. O. K. Wong, G. Fang, Cdk1 phosphorylation of BubR1 controls spindle checkpoint arrest and Plk1-mediated formation of the 3F3/2 epitope. *The Journal of cell biology* **179**, 611 (2007).

267. A. De Luca, V. Esposito, A. Baldi, P. P. Claudio, Y. Fu, M. Caputi, M. M. Pisano, F. Baldi, A. Giordano, CDC2□related kinase PITALRE phosphorylates pRb exclusively on serine and is widely expressed in human tissues. *Journal of cellular physiology* **172**, 265 (1997).

268. J. Lees, K. Buchkovich, D. Marshak, C. Anderson, E. Harlow, The retinoblastoma protein is phosphorylated on multiple sites by human cdc2. *The EMBO journal* **10**, 4279 (1991).

269. M. Jackman, M. Firth, J. Pines, Human cyclins B1 and B2 are localized to strikingly different structures: B1 to microtubules, B2 primarily to the Golgi apparatus. *The EMBO journal* **14**, 1646 (1995).

270. K. L. Gould, S. Moreno, D. Owen, S. Sazer, P. Nurse, Phosphorylation at Thr167 is required for Schizosaccharomyces pombe p34cdc2 function. *The EMBO Journal* **10**, 3297 (1991).

271. T. Enoch, P. Nurse, Mutation of fission yeast cell cycle control genes abolishes dependence of mitosis on DNA replication. *Cell* **60**, 665 (1990).

272. L. L. Parker, H. Piwnica-Worms, Inactivation of the p34cdc2-cyclin B complex by the human WEE1 tyrosine kinase. *Science* **257**, 1955 (1992).

273. C. Featherstone, P. Russell, Fission yeast p107wee1 mitotic inhibitor is a tyrosine/serine kinase. (1991).

274. R. N. Booher, P. S. Holman, A. Fattaey, Human Myt1 is a cell cycle-regulated kinase that inhibits Cdc2 but not Cdk2 activity. *Journal of Biological Chemistry* **272**, 22300 (1997).

275. G. Den Haese, N. Walworth, A. Carr, K. Gould, The Wee1 protein kinase regulates T14 phosphorylation of fission yeast Cdc2. *Molecular biology of the cell* **6**, 371 (1995).

276. T. M. Yamamoto, M. Iwabuchi, K. Ohsumi, T. Kishimoto, APC/C–Cdc20-mediated degradation of cyclin B participates in CSF arrest in unfertilized Xenopus eggs. *Developmental biology* **279**, 345 (2005).

277. D. Izawa, J. Pines, How APC/C-Cdc20 changes its substrate specificity in mitosis. *Nature cell biology* **13**, 223 (2011).

278. B. Sebastian, A. Kakizuka, T. Hunter, Cdc25M2 activation of cyclin-dependent kinases by dephosphorylation of threonine-14 and tyrosine-15. *Proceedings of the National Academy of Sciences* **90**, 3521 (1993).

279. J. Gautier, M. J. Solomon, R. N. Booher, J. F. Bazan, M. W. Kirschner, cdc25 is a specific tyrosine phosphatase that directly activates p34 cdc2. *Cell* **67**, 197 (1991).

280. W. G. Dunphy, A. Kumagai, The cdc25 protein contains an intrinsic phosphatase activity. *Cell* **67**, 189 (1991).

281. N. Shindo, K. Kumada, T. Hirota, Separase sensor reveals dual roles for separase coordinating cohesin cleavage and cdk1 inhibition. *Developmental cell* **23**, 112 (2012).

282. I. H. Gorr, D. Boos, O. Stemmann, Mutual inhibition of separase and Cdk1 by two-step complex formation. *Molecular cell* **19**, 135 (2005).

283. O. Stemmann, H. Zou, S. A. Gerber, S. P. Gygi, M. W. Kirschner, Dual inhibition of sister chromatid separation at metaphase. *Cell* **107**, 715 (2001).

284. S. Hauf, I. C. Waizenegger, J.-M. Peters, Cohesin cleavage by separase required for anaphase and cytokinesis in human cells. *Science* **293**, 1320 (2001).

285. B. E. McGuinness, M. Anger, A. Kouznetsova, A. M. Gil-Bernabé, W. Helmhart, N. R. Kudo, A. Wuensche, S. Taylor, C. Hoog, B. Novak, Regulation of APC/C activity in oocytes by a Bub1-dependent spindle assembly checkpoint. *Current Biology* **19**, 369 (2009).

286. T. Kim, M. W. Moyle, P. Lara-Gonzalez, C. De Groot, K. Oegema, A. Desai, Kinetochore-localized BUB-1/BUB-3 complex promotes anaphase onset in C. elegans. *The Journal of cell biology* **209**, 507 (2015).

287. G. M. Cooper, R. E. Hausman, *The cell: a molecular approach*. (ASM Press, 2009).

288. A. T. Hahn, J. T. Jones, T. Meyer, Quantitative analysis of cell cycle phase durations and PC12 differentiation using fluorescent biosensors. *Cell Cycle* **8**, 1044 (2009).

289. J. White, S. Dalton, Cell cycle control of embryonic stem cells. *Stem cell reviews* **1**, 131 (2005).

290. V. C. Li, A. Ballabeni, M. W. Kirschner, Gap 1 phase length and mouse embryonic stem cell self-renewal. *Proceedings of the National Academy of Sciences* **109**, 12550 (2012).

291. S. P. Bell, A. Dutta, DNA replication in eukaryotic cells. *Annual review of biochemistry* **71**, 333 (2002).

292. D. Y. Takeda, A. Dutta, DNA replication and progression through S phase. *Oncogene* **24**, 2827 (2005).

293. L. Zou, B. Stillman, Assembly of a complex containing Cdc45p, replication protein A, and Mcm2p at replication origins controlled by S-phase cyclin-dependent kinases and Cdc7p-Dbf4p kinase. *Molecular and Cellular Biology* **20**, 3086 (2000).

294. R. Chilà, C. Celenza, M. Lupi, G. Damia, L. Carrassa, Chk1-Mad2 interaction: a crosslink between the DNA damage checkpoint and the mitotic spindle checkpoint. *Cell Cycle* **12**, 1083 (2013).

295. S.-Y. Hyun, B. Sarantuya, H.-J. Lee, Y.-J. Jang, APC/C Cdh1-dependent degradation of Cdc20 requires a phosphorylation on CRY-box by Polo-like kinase-1 during somatic cell cycle. *Biochemical and biophysical research communications* **436**, 12 (2013).

296. V. Lobjois, D. Jullien, J.-P. Bouché, B. Ducommun, The polo-like kinase 1 regulates CDC25B-dependent mitosis entry. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* **1793**, 462 (2009).

297. R. de Sousa Abreu, L. O. Penalva, E. M. Marcotte, C. Vogel, Global signatures of protein and mRNA expression levels. *Molecular BioSystems* **5**, 1512 (2009).

298. C. Vogel, E. M. Marcotte, Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics* **13**, 227 (2012).

299. D. Greenbaum, C. Colangelo, K. Williams, M. Gerstein, Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome biology* **4**, 1 (2003).

300. D. L. Young, S. Michelson, *Systems biology in drug discovery and development*. (John Wiley & Sons, 2011), vol. 9.

301. A. Koussounadis, S. P. Langdon, I. H. Um, D. J. Harrison, V. A. Smith, Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system. *Scientific reports* **5**, 10775 (2015).

302. T. Ly, Y. Ahmad, A. Shlien, D. Soroka, A. Mills, M. J. Emanuele, M. R. Stratton, A. I. Lamond, A proteomic chronology of gene expression through the cell cycle in human myeloid leukemia cells. *Elife* **3**, e01630 (2014).

303. C. M. Bishop, *Neural networks for pattern recognition*. (Oxford university press, 1995).

304. Y. Wu, H. Wang, B. Zhang, K.-L. Du, Using radial basis function networks for function approximation and classification. *ISRN Applied Mathematics* **2012**, (2012).

305. A. Hastings, T. Powell, Chaos in a three□species food chain. *Ecology* **72**, 896 (1991).

306. B. Dennis, R. A. Desharnais, J. Cushing, S. M. Henson, R. Costantino, Estimating chaos and complex dynamics in an insect population. *Ecological Monographs* **71**, 277 (2001).

307. G. Sugihara, R. May, H. Ye, C.-h. Hsieh, E. Deyle, M. Fogarty, S. Munch, Detecting causality in complex ecosystems. *science* **338**, 496 (2012).

308. E. Benincà, K. D. Jöhnk, R. Heerkloss, J. Huisman, Coupled predator–prey oscillations in a chaotic food web. *Ecology letters* **12**, 1367 (2009).

309. H. Leung, T. Lo, S. Wang, Prediction of noisy chaotic time series using an optimal radial basis function neural network. *IEEE Transactions on Neural Networks* **12**, 1163 (2001).

310. T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *Biometrics*, (2002).