Università
degli Studi
di Padova

Dipartimento
di Ingegneria
dell'Informazione

Inference in systems biology: modelling approaches and applications

Ph.D. candidate
Federica Eduati

Advisor
Prof. Gianna Toffolo

School Director
Prof. Matteo Bertocco

Bioengineering Coordinator
Prof. Giovanni Sparacino

Ph.D. School in Information Engineering
Bioengineering section, XXV Series, 2013

Summary

The main topic of this thesis is the study of biological regulatory systems using different computational modelling approaches in order to gain new insights into not yet completely understood biological processes. In ‘systems biology’, mathematical models represent a powerful tool to study biological processes. Models are abstractions of reality always including some degree of simplification: an important ingredient of the modelling process, having a major role in suggesting the appropriate level of abstraction and simplification, is the purpose of the model, that is the question they have to answer.

This thesis is focused on the analysis of how models of different complexity appropriately describe the available data to achieve a given purpose. Such analysis guides the choice of the most appropriate degree of simplification of the system under study that allows neglecting some aspects without compromising the results of the model. Three levels of detail for inference and modelling are analyzed in this thesis depending on the system under consideration. The first level is the network level, where molecules are nodes connected by edges and the interest is in the inference of the topology of connections at large scale. In the second level the network is interpreted as a mean to produce qualitative simulations and predictions which can be compared with experimental data. The third level of detail consist in a more mechanistic dynamic description of the system using ordinary differential equations but limiting the analysis to small subsystems. For each level of detail, appropriate approaches have been developed and applied to in silico and real data of different biological systems. Finally, different modelling appraches have been integrated to analyze insulin signalling pathway on different levels of simplification using a novel experimental dataset collected specifically for this purpose.

Inference of regulatory networks

In this part of the thesis a simple method developed to reconstruct, from single- stimulus/inhibitor protein data, cause-effect networks representing signalling pathways is described. It uses a strictly data driven approach, having Boolean (discrete logic) inference as basic ingredients underlying network reconstruction. Potentialities and limitations of this method have been compared with those of more standard reverse-engineering methods based on mutual information (ARACNe and CLR) and on Bayesian networks.

Qualitative modelling of large networks

A qualitative approach based on logic modelling is introduced in this part of the thesis, and its application to signalling pathways is discussed. In particular a method has

been developed and implemented in the R package CNORfeeder, to integrate prior knowledge with data-driven information obtained using reverse engineering approaches to infer predictive signalling network models from perturbation experiments. This package has been designed to be integrated with an existing software, CellNOpt, which permit to interpret a network as a logic model and to train it against experimental data. This integrated approach has been applied to a network of growth and inflammatory signalling and has been shown to provide a logic model with superior fit to data from human liver cancer cells HepG2, proposing potential missing pathways supported by known interactions among proteins. Performances of CellNOpt in training logic models to high-throughput phosphoproteomic data have been compared to those provided using a declarative problem solving paradigm called Answer Set Programming (ASP).

Quantitative modelling of small sub-networks

A more detailed mechanistic description of small systems has been applied to study biological regulatory networks, using ordinary differential equations focusing on overrepresented patterns that play important functional roles, and in particular on:

- *feed-back loops*: an autoregulatory mechanism has been proposed to explain the adaptation observed at genome-wide scale in the yeast stress response reproducing all the kinetic features of the mRNA time-series;
- *feed-forward loops*: an new approach has been developed to reduce the search space for new interactions to identify potential novel players in mixed regulatory networks, using sequence analysis and model identification criteria to select active miRNA mediated feed-forward loops during adipogenesis.

Multilevel study of insulin signalling pathway

All three levels of detail have been applied in order to analyze insulin signalling pathway. The study is based on data collected during an experiment specifically designed to monitor the dynamic of key proteins after insulin/leucine stimulation. The signalling pathway is reconstructed by retrieving information about known interactions in literature and databases. The pathway is then reduced to identifiable nodes and a semi-qualitative approach based on logic-based differential equations is applied to fit the experimental data. A more detailed description is then developed for a sub-part of the network using ordinary differential equations directly derived from kinetic equations. With this example, we show how the integration of analysis at multiple levels of detail allows to study a problem providing different insights into the system.

Sommario

Questa tesi di dottorato è incentrata sullo studio dei sistemi biologici mediante l'utilizzo di diversi approcci di modellistica computazionale, al fine di esplorare processi biologici non ancora chiari. Nella 'systems biology', i modelli matematici rappresentano un potente mezzo per studiare i processi biologici. I modelli sono astrazioni della realtà e possono includere diversi livelli di semplificazione a seconda del loro scopo.

La tesi è focalizzata sull'analisi di come, modelli di diversa complessità, possono essere utilizzati per raggiungere diversi scopi. Questa analisi guida la scelta del livello di semplificazione della realtà più adatto per trascurare certi dettagli senza però compromettere la sua applicazione. Tre livelli di dettaglio sono analizzati in questa tesi. Il primo livello è la rete, le molecole sono considerate come nodi connessi tra di loro da archi e si è interessati ad inferire la topologia delle connessioni su larga scala. Nel secondo livello, la rete è interpretata come un mezzo per produrre simulazioni qualitative che possono essere confrontate con i dati reali. Il terzo livello di dettaglio consiste in una descrizione dinamica meccanicistica del sistema, mediate l'utilizzo di equazioni differenziali ordinarie ma limitando l'analisi a sottosistemi di dimensioni ridotte. Per ogni livello di dettaglio, sono stati sviluppati approcci adeguati, poi applicati a dati in silico e a dati reali relativi a diversi sistemi biologici. Diversi approcci di modellistica sono stati integrati per l'analisi del pathway del signalling dell'insulina considerando diversi livelli di semplificazione e utilizzando un dataset sperimentale raccolto specificatamente per questo scopo.

Inferenza di reti di regolazione

In questa parte della tesi viene descritto un metodo sviluppato per ricostruire reti causa-effetto da esperimeti sigolo stimolo/singolo inibitore. L'approccio utilizzato è data-driven e ricostruisce la rete basandosi su metodi di inferenza Booleani. Le potenzialità e i limiti di questo metodo sono stati confrontati con quelli di metodi di reverse-engineering standard basati su mutua informazione (ARACNe e CLR) e su reti Bayesiane.

Modellizzazione qualitativa di reti a larga scala

In questa parte della tesi viede introdotto un approccio qualitativo basato su modelli logici applicato a pathway di signalling. Nello specifico, è stato sviluppato un metodo, implementato in un pacchetto R chiamato CNORfeeder, per integrare la conoscenza a priori con informazione di tipo data-driven ricavata con metodi di reverse-engineering, con l'obiettivo di inferire modelli predittivi di reti di signalling a partire da esperimenti di perturbazione. Questo pacchetto è stato pensato per essere integrato con un software

esistente, CellNOpt, che permette di interpretare una rete come un modello logico e di allenarla su dati sperimentali. Questo approccio integrato è stato applicato al pathway responsabile della crescita e della risposta infiammatorio e si è visto che permette di ottenere un modello logico che descrive molto bene i dati misurati su cellule cancerogene di fegato, proponendo anche link potenzialmente mancanti che sono supportati di interazioni note tra proteine. Le performance di CellNOpt nell'allenare i modelli logici a dati di fosfoproteomica, sono state confrontate con quelle ottenute con l'utilizzo di un paradigma dichiarativo per il problem solving chiamato Answer Set Programming (ASP).

Modellizzazione quantitativa di sottoreti

Una descrizione più dettagliata di sottosistemi di dimensioni ridotte è stata applicata allo studio di reti biologiche, mediante l'utilizzo di equazioni differenziali ordinarie per la descrizione di pattern sovrarappresentati che hanno ruoli funzionali importanti:

- *feed-back loops*: un meccanismo di autoregolazione è stato proposto per spiegare il meccanismo di adattamento osservato a livello genome-wide nella risposta allo stress del lievito, questo meccanismo è in grado di riprodurre tutte le caratteristiche cinetiche osservate su serie temporali di mRNA;

- *feed-forward loops*: un nuovo approccio è stato sviluppato per ridurre lo spazio di ricerca di nuove interazioni e per identificare potenziali molecole interessanti conivolte in reti di regolazioni miste trascrizionali e post-trascrizionali. I feed-forward loops mediati da miRNA, attivi durante l'adipogenesi, sono stati selezionati sulla base di analisi di sequenza e di criteri di identificazione di modelli.

Studio multilivello del pathway dei signalling dell'insulina

Tutti e tre i livelli di dettaglio sono stati applicati per l'analisi del pathway del signalling dell'insulina. Lo studio è basato su dati misurati durante un esperimento pensato per monitorare la dinamica di proteine chiave stimolate con insulina e leucina. Il pathway di signalling è stato ricostruito da ricerche bibliografiche e su database ed è poi stato ridotto ai nodi identificabili. Un approccio semi-qualitativo basato su equazioni differenziali derivate da modelli logici è stato applicato per fittare i dati sperimentali. Una descrizione più dettagliata è stata poi sviluppata per una porzione della rete utilizzando equazioni differenziali ordinarie derivate da equazioni cinetiche. Con questo esempio, viene mostrato come l'integrazione di analisi a diversi livelli permette di studiare un problema da diversi punti di vista.

Contents

1 Introduction **1**

1.1 Complexity in cell biology 1

1.2 Modelling biological systems 3

1.3 Overview and contributions 5

2 Inference of networks topology **9**

2.1 Introduction 9

2.2 Inference of signalling networks with FEED 10

2.3 Other methods for reverse-engineering signalling networks 11

Appendix 2.1 Paper: Eduati et al., PLoS one, 2010 12

3 Qualitative modelling of networks **27**

3.1 Introduction 27

3.2 CNORfeeder: network inference for logic modelling 28

3.3 Comparison with ASP method 29

Appendix 3.1 Paper: Eduati et al., Bioinformatics, 2012 30

4 Quantitative modelling of small sub-networks **47**

4.1 Introduction 47

4.2 Feed-back loop to explain yeast stress response 49

4.3 miRNA mediated feed-forward loops 49

Appendix 4.1 Paper: De Palo et al., IET Syst Biol, 2011 50

Appendix 4.2 Paper: Eduati et al., J Comp Biol, 2012 76

5 Multilevel study of insulin signalling pathway **95**

5.1 Introduction 95

5.2 Experimental setup 96

5.3 Insulin signalling network topology 100

5.4 Semi-qualitative modelling using logic models 103

5.5 Quantitative modelling using ODEs 111
Appendix 5.1 Paper: Eduati et al., Submitted 112

6 Conclusions 131

1
Introduction

1.1 Complexity in cell biology

Biological systems are complex and multiscale going from molecules, to cells to organisms and there is increasing interest in the study of those tangled mechanisms that permit to decode the DNA sequence and translate it into structure and function. As stated by the central dogma of molecular biology, genes are lengths of DNA that are transcribed into complementary sequences called mRNAs that are in turn used as templates for proteins (Figure 1.1, left panel). Proteins are the main functional components of the cell as they perform cellular work, they control metabolism and they are responsible for the regulation of DNA transcription. To complete the picture, mechanisms are influenced also by post-transcriptional modifications, that involve molecules called microRNAs, as well as post-translational modifications of proteins in order to adapt the response of the cell to external stimuli (Figure 1.1, right panel). The field of study that deals with the analysis of biological systems and the interactions within them is commonly called 'systems biology': while molecular biology focuses on the understanding of individual cellular components to elucidate how molecules works, systems biology aims at a system level understanding in order to elucidate functional properties. To quote from Ideker, systems biology can be defined as follows:

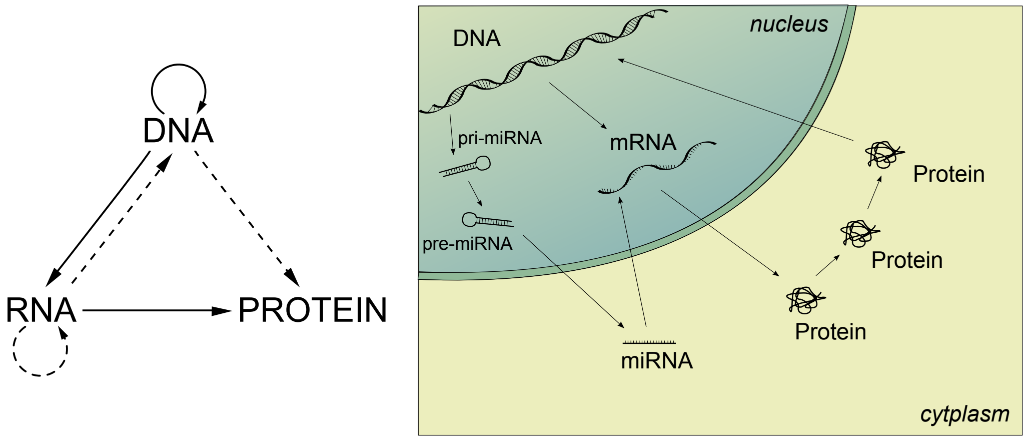


Figure 1.1: *left panel*: central dogma of molecular biology, original cartoon as formulated by Crick in 1970 (Crick, 1970), where solid arrows show general transfers and dotted arrows show special transfer. *right panel*: cartoon schematically representing complexity in biological regulatory systems, post-transcriptional and post-translational complicate the network of regulatory mechanisms.

> *"Systems biology studies biological systems by systematically perturbing them (biologically, genetically, or chemically); monitoring the gene, protein, and informational pathway responses; integrating these data; and ultimately, formulating mathematical models that describe the structure of the system and its response to individual perturbations."* (Ideker et al., 2001)

Even if the concept of 'systems biology' was first introduced in 1948 by Norbert Wiener, the founding father of cybernetics, who attempted to study technical and biological systems using the same scientific approaches (Wiener, 1948), the first models developed to have a system level understanding of biological systems were limited by the inadequacy of data to support formulated hypotheses. The recent rising of systems biology is mainly due to the fact that currently available experimental techniques permit to produce high-throughput genomic and proteomic data that allow to monitor the dynamics of different molecules under different experimental conditions. This large amount of data needs to be handled by mean of computational methods in order to draw information about biological systems. The main aims are the understanding of mechanisms that stay beneath different phenotypes and characterize specific diseases and eventually the identification of drugs able to counterbalance the effects of the disease. However, understanding of biological mechanisms can be done at different systems level; in 2002 Kitano was writing about systems biology:

> *"There is now a golden opportunity for system-level analysis to be grounded in molecular-level understanding, resulting in a continuous spectrum of knowledge.*

> *System-level understanding, the approach advocated in systems biology, requires a shift in our notion of 'what to look for' in biology. While an understanding of genes and proteins continues to be important, the focus is on understanding a system's structure and dynamics."* (Kitano, 2002)

As we will see in the next Paragraph, the choice of the most appropriate computational model should be guided by the biological knowledge and by the available experimental data.

1.2 Modelling biological systems

In the context of systems biology, mathematical models represent a powerful tool to study biological processes. Given a biological system, the development of a model allows to simulate the system, making assumptions which can then be validated or rejected according to the data. Eventually, a valid model can be used to identify the structure and predict the outcome of the system under different experimental conditions. Models are abstractions of reality and may include a different level of simplification depending on the question they have to answer and the available data. A realistic description of all biochemical mechanisms has to be included if the model aims at elucidating a detailed biological mechanism, while a more abstract model is sufficient to predict the output of the system in response to a specific input. However, biological complexity growth exponentially with the number of components of a biological system (Butcher et al., 2004) limiting the level of detail that can be included when dealing with large-scale systems. Another limiting factor is the experimental knowledge of the system: a mechanistic description of the system is unrealistic when individual molecular targets cannot be interrogated and/or when sufficient qualitative data are not available (Bornholdt, 2005).

In this thesis, three different levels of simplification of biological reality are considered, as exemplified in Figure 1.2.

(A) The first level is the *topological* level: complex biological systems are represented as networks of interacting species, where molecules are seen as nodes connected by links. The topology of the network can be inferred from high-troughput data, with the aim of reverse-engineering the relationships between intracellular components responsible for regulating cellular function. Large-scale networks can be considered focusing on the analysis of the structure of the network, for example to understand how a disease affects the topology of the network. However the network itself does not allow to make simulation or predictions unless it is interpreted as a model. For this reason,

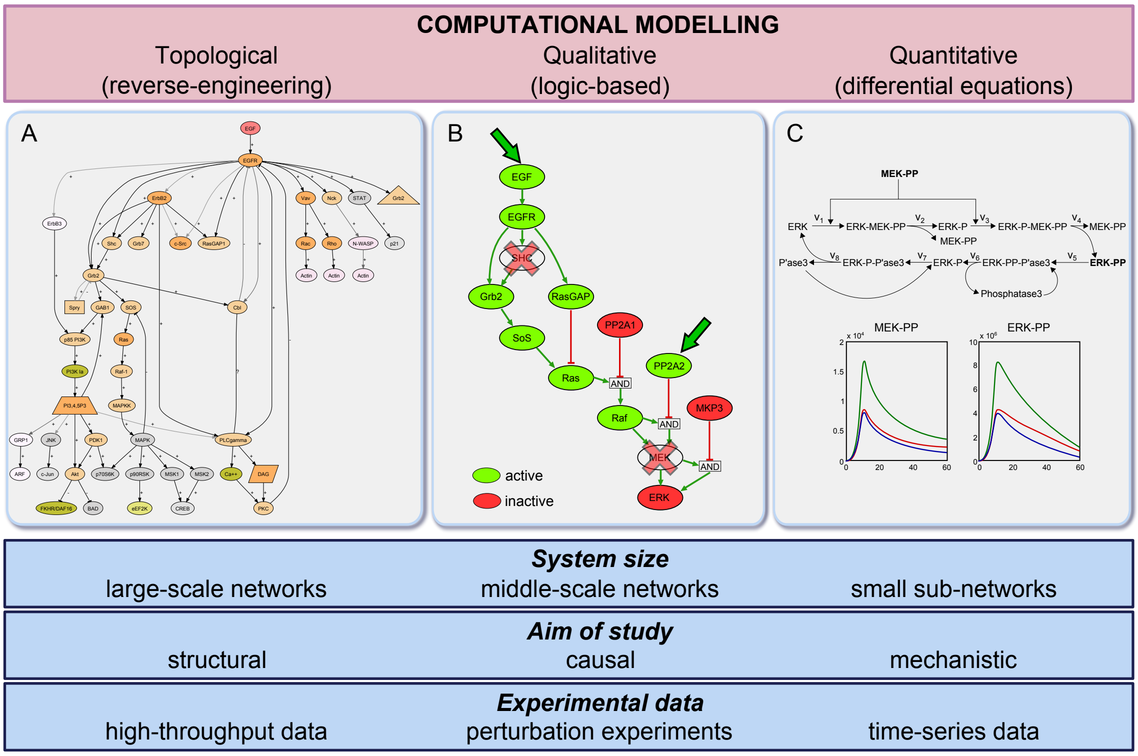


Figure 1.2: Different levels of simplification / modelling approaches. A. *Topological*: network representation of EGFR signalling pathway (taken from (Schlessinger, 2013)) showing connections between key components of the network. B. *Qualitative*: Logic model representing only key molecules of EGFR signalling pathway, where nodes can assume two stated: active (on) or inactive (off) and nodes can be perturbed by stimulation and/or inhibition. C. *Quantitative*: Dynamic representation of MEK-ERK regulation (adapted from (Schoeberl et al., 2002)) including all possible kinetic reactions where every node represent concentration of a chemical species to produce time-couse simulations.

(B) the second level is *qualitative* modelling: networks can be interpreted as causal relationships and encoded in logic models. These models are generally used to interpret middle-scale networks without going too much into the mechanistic detail but allowing to make qualitative simulations and predictions of the outcome of the network under different experimental conditions, which can then be compared with experimental data.

(C) The third level is *quantitative* modelling: detailed biological regulatory mechanisms are considered by modelling mass-action kinetics. Differential equations can be derived to study the dynamics of the system and to simulate (continuous) time-course data. However, this analysis is generally limited to small subsystems, due to the necessity of writing one equation for each molecular species and of having a large number of quantitative dynamic data to estimate kinetic parameters.

This is a general classification thought to define the outline of the thesis as it well categorizes approaches developed during the PhD studies, and it gives an idea of how different modelling approaches are suitable for the solution of different problems. However, different classifications can be formulated (Lawrence et al., 2010) and probably none of them will perfectly categorize all possible computational models. There are many examples, in recent literature, of methods which do not exactly fit into this classification like coarse-grained differential equations models derived to describe middle- and large-scale networks (Feret et al., 2009), or logic-based models able to describe (continuous) dynamic data (Terfve et al., 2012). Nevertheless, the appropriate modelling approach always needs to be chosen or developed based on the considered biological system, the aim of the study and the available experimental data.

1.3 Overview and contributions

This thesis is focused on the analysis of how models of different complexity appropriately describe the available data to achieve a given purpose. Such analysis guides the choice of the most appropriate degree of simplification of reality that allows neglecting some aspects without compromising the results of the model. In the following, one chapter will be devoted to each level of detail and several developed approaches will be described along with their applications to in silico and real data of different biological systems. A separate chapter will be dedicated to the integration of different modelling approaches to analyze insulin signalling pathway on different levels of simplification using a novel experimental dataset collected specifically for this purpose.

In Chapter 2, we present a simple method developed to reconstruct, from single-stimulus/inhibitor protein data, cause-effect networks representing signalling pathways (Eduati et al., 2010). This method, which resulted the best performing in the “Predictive Signaling Network Modeling challenge” of DREAM4 competition, can be used to discover how signalling pathways are altered by diseases and to predict the effect of multiple agents/drugs. It uses a strictly data driven approach, having Boolean (discrete logic) inference as basic ingredients underlying network reconstruction. Boolean inference is appropriate to reconstruct the signalling network structured into input nodes (stimuli) intermediate nodes (inhibitors) and output nodes (phosphoproteins), particularly in situations where the limited number of available samples and the lack of information on the stimulus format prevent the use of more sophisticated modelling approaches. Potentialities and limitations of this method are then compared with those of more standard reverse-engineering methods based on mutual information (ARACNe and CLR) and on Bayesian networks (Eduati et al., 2012a).

In Chapter 3, we introduce a qualitative approach based on logic modelling and its application to signalling pathways. In particular we focus on a method, implemented in the R package CNORfeeder, to integrate prior knowledge with data-driven information obtained using reverse engineering approaches to infer predictive signalling network models from perturbation experiments (Eduati et al., 2012a). This package is designed to be integrated with an existing software, CellNOpt, which permit to interpret a network as a logic model and to train it against experimental data. This integrated approach is applied to a network of growth and inflammatory signalling and is shown to provide a logic model with superior data fit to data from human liver cancer cells HepG2, proposing potential missing pathways supported by known interactions among proteins. Performances of CellNOpt in training logic models to high-throughput phosphoproteomic data are compared to those provided using Answer Set Programming (ASP), a declarative problem solving paradigm, in which a problem is encoded as a logical program such that its answer sets represent solutions to the problem (Videla et al., 2012). On in silico datasets, ASP is shown to be efficient in scalability and computational times. It also guarantees global optimality of solutions providing the complete set of solutions (Guziolowski et al., Submitted).

In Chapter 4, we apply a more detailed mechanistic description of small systems using ordinary differential equations focusing on overrepresented patterns in biological regulatory networks that play important functional roles. We propose a regulatory mechanism able to explain the adaptation observed at genome-wide scale in the yeast stress response reproducing all the kinetic features of the mRNA time-series (De Palo et al., 2011). The gene expression response of yeast to various types of stresses/perturbations shows a common functional and dynamical pattern for the vast majority of genes, characterised by a quick transient peak (affecting primarily short genes) followed by a return to the pre-stimulus level. Kinetically, this process of adaptation following the transient excursion can be modelled using a genome-wide autoregulatory mechanism by means of which yeast aims at maintaining a preferential concentration in its mRNA levels.

A similar modelling approach was also applied to select active miRNA mediated feed-forward loops during adipogenesis, based on sequence analysis and model identification criteria (Eduati et al., 2012b). Different simple models of feed-forward loop circuits are assessed based on their ability to properly reproduce miRNA and mRNA expression data in terms of identification criteria, namely: goodness of fit, precision of the estimates, and comparison with submodels. The work is focused on the development of a new approach to reduce the search space for new interactions and to identify potential novel players. When applied to adipogenic differentiation gene expression data, it provides potential novel players in this regulatory network.

In Chapter 5, all three levels of detail were then applied to a single system in order to analyze the insulin signalling pathway. The study is based on data collected during an experiment specifically designed to monitor the dynamic of key proteins of insulin signalling pathway of human skeletal muscle cells under three different experimental conditions: insulin stimulation, leucine stimulation and insulin stimulation after leucine incubation. The signalling pathway is reconstructed by retrieving information about known interactions in literature and databases. The pathway is then reduced to identifiable nodes and a semi-qualitative approach based on Boolean ODEs is applied to fit the experimental data. A more detailed description is then developed for a sub-part of the network using ordinary differential equations directly derived from kinetic equations (Eduati et al., Submitted). With this example, we show how the integration of analysis at multiple levels of detail allows to study a problem providing different insights into the system.

2

Inference of networks topology

2.1 Introduction

All interactions and control mechanisms between genes, mRNAs, proteins and metabolites are commonly codified in biochemical networks where nodes represent molecules, and links represent interactions or regulatory mechanisms. High-throughput measurement techniques allow simultaneous interrogation of multiple cell components under different perturbations, and reverse-engineering attempts to infer interaction networks topology from these biological data. The aim is the reconstruction of biochemical networks in order to understand how molecules and their interactions determine the function of the cell (Barabási & Oltvai, 2004). Different kind of biochemical networks exists, depending on the system under study, e.g. gene networks, protein networks or metabolic networks. Due to the large diffusion of microarrays, which provides high-throughput measures of gene expression, many reverse-engineering methods have been developed in literature to infer transcriptional networks (Brazhnik et al., 2002) where nodes are genes and connections show how genes (undirectly) affect each other activity. In the last years, increasing attention has been devoted also to protein signalling networks, which govern the transmission of external signals through cells involving protein post-translational modifications.

The DREAM project

The DREAM (Dialogue for Reverse Engineering Assessments and Methods) project (Stolovitzky et al., 2007) is an initiative that bring together different research groups in the effort of developing and discussing reverse-engineering methods to infer networks connectivity in cell biology. In DREAM challenges, simulated or real experimental data are provided to the participants and they are asked to infer the underlying network structure. Performances of different groups are then compare using a known standard.

2.2 Inference of signalling networks with FEED

Signalling pathways are used by cells to respond to environmental changes: a protein or a metabolite binds to transmembrane receptors and, inducing a cascade of signals that involves the activation of different proteins, relay to the nucleus altering the behavior of the cell. Increasing attention has been payed to signalling pathways as it became known that defects in signal transduction cause many disease as cancer, heart disease and diabetes. Large scale protein signaling networks are well studied but there is still not enough information on how those pathways respond to specific stimuli and how diseases and drugs affect the pathways. The topology of the network can be inferred based on the measures of changes of phosphorylation states or activities of some proteins of the network in perturbation experiments. These perturbations can consist in stimulation with hormones or growth factors and/or chemical inhibitions of specific proteins.

In the DREAM 4 “Predictive Signaling Network Challenge” (Prill et al., 2011) participants were asked to reconstruct the topology of interactions between measured and perturbed proteins from single-stimulus/single-inhibitor experimental data and to predict the response to multiple stimuli and/or multiple inhibitors. We developed a method (Eduati et al., 2010) that was used to participate to this challenge resulting the best performer. Network topology is reconstructed in two steps: first, a Boolean table is inferred for each measured protein indicating which stimuli and which inhibitors affect the protein. Then, tables are used to infer a causal network connecting stimuli, inhibited and measured proteins. Since performances were evaluated based on predictions of test data of multiple stimuli/inhibitors (not provided to participants), the reconstructed network was used to linearly combine appropriate single stimulus/inhibitor data. The method is described in detail in Eduati et al. (2010). It was later named FEED in Eduati et al. (2012a), where an improved version of the same algorithm was used for different purposes (see Chapter 3). Full text of the original paper Eduati et al. (2010), is reported in Appendix 2.1.

2.3 Other methods for reverse-engineering signalling networks

Many reverse-engineering methods, originally developed to infer transcriptional networks, were later applied to the inference of the topology of signalling networks from steady-state perturbation experiments. Strictly data-driven methods for reverse-engineering signalling network with no need for prior knowledge, have been used, for example, in Sachs et al. (2005) and Ciaccio et al. (2010). Many developed methods are based on:

- *mutual information*: undirected networks can be inferred based on mutual information estimated between all pairs of measured proteins, considered as random variables. Examples of mutual information based methods are ARACNE (Margolin et al., 2006) and CLR (Faith et al., 2007): in both cases the first step is the inference of the mutual information matrix, but they differs in the apporach used to derive valid interaction from information encoded in the matrix;

- *Bayesian inference*: signalling cascades can be interpreted as Bayesian networks, inferred from data using a statistically founded computational procedure by interpreting statistical effects among molecules as causal relationships. In Bayesian networks, nodes represents measured molecules, seen as random variables, and links represent conditional dependencies: network inference algorithms aim to find a model that closely predicts the observations (Pe'er, 2005).

FEED, Bayesian inference, ARACNE and CLR were used in Eduati et al. (2012a) to infer a benchmark network from in silico data (see 3.1 for additional details). Different methods are known to have specific advantages and limitations, in fact the DREAM experience showed that best results are obtained by combining networks inferred using different approaches (Prill et al., 2011). Bayesian networks have the advantages of being robust to the existence of unobserved variables (not all proteins of the networks can be measured) and of accommodating, in their probabilistic nature, noise that is inherent in biological data. Though, they are acyclic graphs, thus cannon infer feedbacks, that are well known to be important in signalling networks. On the other hand, mutual information based methods only infers unsigned interactions, not providing causality of links that is very important in signalling networks. FEED can infer directed signalling networks with feedbacks, but it derives a cause-effect network with directed links only from stimulated to inhibited or measured proteins, and from inhibited to measured proteins.

Appendix 2.1 Paper: Eduati et al., PLoS one, 2010

The following publication dealing with inference of signalling networks topology has been coauthored by the Ph.D. candidate during her doctoral program.

- F. Eduati, A. Corradin, B. Di Camillo, and G. Toffolo. *A boolean approach to linear prediction for signaling network modeling*. PloS one, 5(9):e12789, 2010.

Full text of the original paper is reported in this Appendix.

A Boolean approach to linear prediction for signaling network modeling

F. Eduati[1], A. Corradin[1], B. Di Camillo[1] and G. Toffolo[1,*]

[1]Department of Information Engineering, University of Padova, Padova, Italy

[*]*Corresponding author:* toffolo@dei.unipd.it

Abstract

The task of the DREAM4 (Dialogue for Reverse Engineering Assessments and Methods) "Predictive signaling network modeling" challenge was to develop a method that, from single-stimulus/inhibitor data, reconstructs a cause-effect network to be used to predict the protein activity level in multi-stimulus/inhibitor experimental conditions. The method presented in this paper, one of the best performing in this challenge, consists of 3 steps: 1. Boolean tables are inferred from single-stimulus/inhibitor data to classify whether a particular combination of stimulus and inhibitor is affecting the protein, 2. a cause-effect network is reconstructed starting from these tables, 3. training data are linearly combined according to rules inferred from the reconstructed network. This method, although simple, permits to achieve a good performance providing reasonable predictions based on a reconstructed network compatible with knowledge from the literature. It can be potentially used to predict how signaling pathways are affected by different ligands and how this response is altered by diseases.

1 Introduction

There is an increasing agreement of the scientific community in attributing complex disease such as cancer, diabetes, heart disease and autoimmunity to defects in signaling trasduction pathways. For instance, in the case of cancer, it is generally acknowledged that genetic mutations are involved in the onset of the disease, but its manifestation is at the pathway functional signaling level (Jones, 2008; Subramanian et al., 2005). Thus, an important step towards a dynamic understanding of the functions and behaviors relevant to a particular system is modeling protein interactions, by integrating available knowledge on signaling pathways with novel high-throughput protein expression data. Development of new therapies would benefit from models and methods able to predict the alterations induced on protein expression levels by different therapeutical agents. Recently, some pioneering efforts were accomplished by Li et al. (Li et al., 2009) who developed a computational framework for a functional input-output description of the Toll-like receptor signaling and the identification of potential targets for its modulation, and by Mitsos et al. (Mitsos et al., 2009) who proposed a computational approach based on the experimental protocol introduced in (Alexopoulos et al., 2010) and a methodology to create cell-specific Boolean models as presented in (Saez-Rodriguez et al., 2009), to evaluate drug actions on signaling pathways.

Evaluation and comparison of the performance of algorithms for network inference and data prediction is still an open issue. The Predictive Signaling Network Modeling challenge of DREAM4 competition provides an important contribution to this topic, by addressing the problem of signaling network inference from single-stimulus/inhibitor data for prediction of multi-stimulus/inhibitor data. The challenge arises from the question of generating a model from a network and data as defined in (Saez-Rodriguez et al., 2009): to this purpose, the organizers provided the topology of a canonical signaling pathway, derived from the literature, and a training set they have published in (Alexopoulos et al., 2010) monitoring the activity of seven phosphoproteins (AKT, $ERK12$, Ikb, $JNK12$, $p38$, $HSP27$, $MEK12$) at three time points (0, 30 minutes and 3 hours) during twenty five different perturbations consisting of combinatorial treatment with zero or one cytokine ($TNFa$, $IL1a$, $IGF1$, TGFa) acting as a stimulus and zero or one inhibitor ($MEKi$, $p38i$, $PI3Ki$, $IKKi$). Participants were asked to a) update the network b) predict the seven phosphoprotein levels in response to twenty pair-wise combinations of stimuli (TGF, $IL1a$, $IGF1$, $TGFa + IGF1$) and inhibitors ($p38i + MEKi$, $PI3Ki + MEKi$, $p38i + IKKi$, $PI3Ki + IKKi$). The corresponding measured levels were available to participants only after the disclosure of the best performing teams and were used by the organizers to evaluate the quality of predictions. Network and data are a subset of those used in (Alexopoulos et al., 2010) and in (Saez-Rodriguez et al., 2009), all measurements were performed using Luminex xMAP sandwich assay as described in (Alexopoulos et al., 2010) and were affected by measurement errors due to technical noise ($SD = 300$), and biological noise ($CV = 8\%$) (Prill et al., 2010).

It was emphasized that the submitted network, specific for the HepG2 cell line, had to include only nodes representing measured or manipulated elements (i.e. stimuli, inhibited proteins and measured proteins) and edges underlying predictions, and that predictions had to be based on the reconstructed network. As anticipated, the challenge was evaluated on the basis of quality of predictions and sparsity of the network. Reliability of predictions was quantified, for each protein p, by the Normalized Squared Error $NSE(p)$:

$$NSE(p) = \sum_{measurements\ of\ P} \frac{prediction - measurement}{measurement\ error} \qquad (1)$$

$NSE(p)$ was compared with a null distribution in which predictions were sampled at random from the measured values of each protein, p-values obtained for each protein were then combined in a Prediction Score: a larger score indicates greater statistical significance of the prediction. Finally, the Overall Score, which also considers the parsimony of the submitted network, was used for team ranking:

$$Overall\ Score = Prediction\ Score - r \cdot (Edge\ Count) \quad (2)$$

where r is a parameter determined empirically by the organizers of the challenge as the minimum, over all teams, of the Prediction Score divided by the Edge Count.

In this paper, a simple data-driven method is presented, that was applied to this DREAM4 challenge. Network topology was reconstructed by inferring Boolean tables from training data, to establish cause-effect relationships characterizing the pathway in terms of links among ligands, inhibitors and proteins. Expression levels of the output proteins during multi-stimulus/inhibitor perturbations were then predicted by a linear combination of training data, in accordance with the reconstructed network.

2 Methods

The method consists of three steps (Fig. 1) based on: 1) inference of Boolean tables from data to classify whether a particular combination of stimulus and inhibitor is affecting the protein, 2) reconstruction of a cause-effect network from Boolean tables, 3) prediction of test data by linear combination of training data, using rules based on the reconstructed network. The three steps are detailed in the following paragraphs by denoting, for a generic protein p ($p = 1, \ldots, 7$):

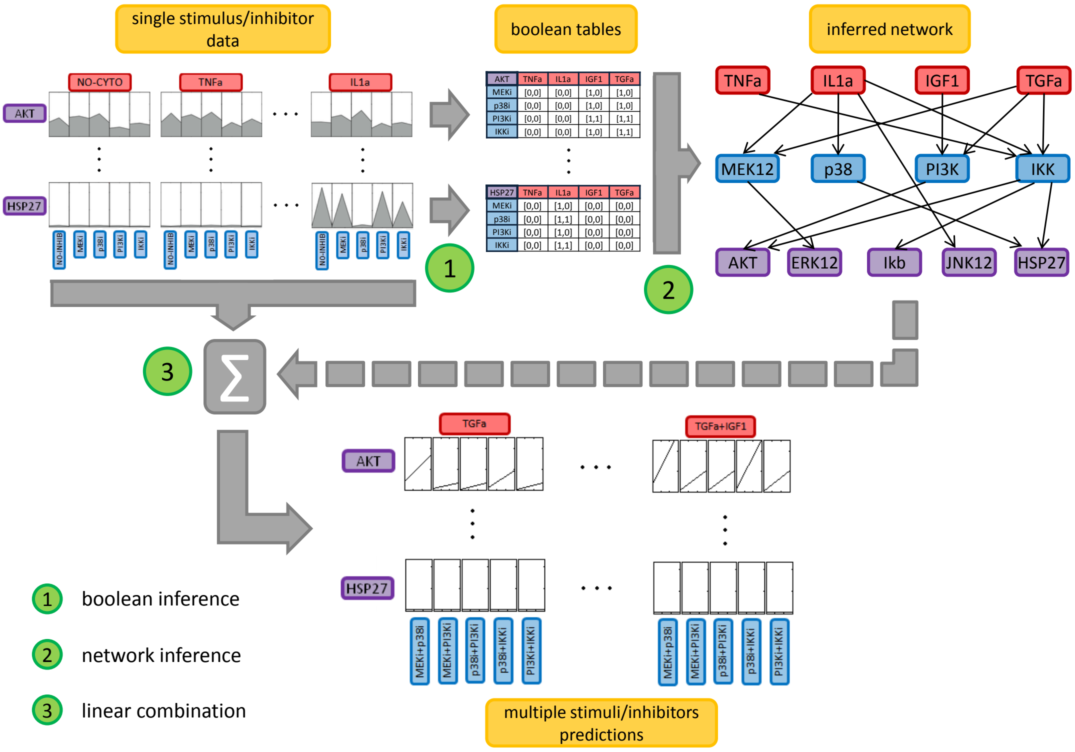
- $x_{i,j}(t)$: protein level at time t collected after perturbation with stimulus i ($i = 0, \ldots, 4$ where $i = 0$ represents the condition without any stimulus) and inhibitor j ($j = 0, \ldots, 4$ where $j = 0$ represents the condition without any inhibitor);
- $\nu_{i,j}(t) = x_{i,j}(t) - x_{i,j\ b}$: protein level with respect to the basal level (indicated by the suffix b);
- $\sigma^2_{x_{i,j}} = 300^2 + (0.08 \cdot x_{i,j})^2$: variance of the measurement error (provided by the organizers) associated to $x_{i,j}$;
- $\sigma^2_{\nu_{i,j}} = \sigma^2_{x_{i,j}} + \sigma^2_{x_{i,j\ b}}$: variance of the measurement error associated to $\nu_{i,j}$.

2.1 Inference of Boolean tables

A table is built for each protein, having a column for each stimulus and a row for each inhibitor and containing in each cell a two-value vector $C_{i,j} = [a_i, b_{i,j}]$ indicating how a particular stimulus/inhibitor combination affects the protein. a_i denotes the action of the stimulus i and $b_{i,j}$ the action of stimulus i/inhibitor j, each quantized in two levels: 1 if action is significant, 0 if not.

Significant increase in protein level in response to a stimulus and significant decrease in response to an inhibitor are tested following (Di Camillo et al., 2005), based on the measurement error distribution.

More precisely, for each stimulus, in absence of inhibitors, the increase of the protein activity ($TEST1_i = \nu_{i,0}(t)$) with respect to the reference, i.e. the condition with no

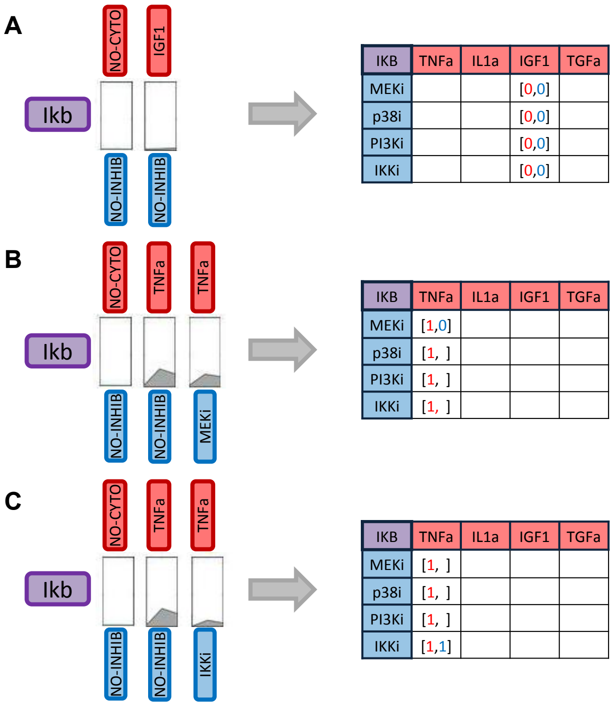


AKT	TNFa	IL1a	IGF1	TGFa
MEKi	[0,0]	[0,0]	[1,0]	[1,0]
p38i	[0,0]	[0,0]	[1,0]	[1,0]
PI3Ki	[0,0]	[0,0]	[1,1]	[1,1]
IKKi	[0,0]	[0,0]	[1,0]	[1,1]

HSP27	TNFa	IL1a	IGF1	TGFa
MEKi	[0,0]	[1,0]	[0,0]	[0,0]
p38i	[0,0]	[1,1]	[0,0]	[0,0]
PI3Ki	[0,0]	[1,0]	[0,0]	[0,0]
IKKi	[0,0]	[1,1]	[0,0]	[0,0]

Figure 1: Workflow representing the 3 steps of the method. (1) Boolean inference of tables from single-stimulus/inhibitor experimental data, (2) network inference from the tables and (3) linear combination of single-stimulus/inhibitor data to predict protein activity level in multi-stimulus/inhibitor conditions, based on network structure.

Eduati et al., PLoS One, 2010, 5(9): e12789



IKB	TNFa	IL1a	IGF1	TGFa
MEKi			[0,0]	
p38i			[0,0]	
PI3Ki			[0,0]	
IKKi			[0,0]	

IKB	TNFa	IL1a	IGF1	TGFa
MEKi	[1,0]			
p38i	[1,]			
PI3Ki	[1,]			
IKKi	[1,]			

IKB	TNFa	IL1a	IGF1	TGFa
MEKi	[1,]			
p38i	[1,]			
PI3Ki	[1,]			
IKKi	[1,1]			

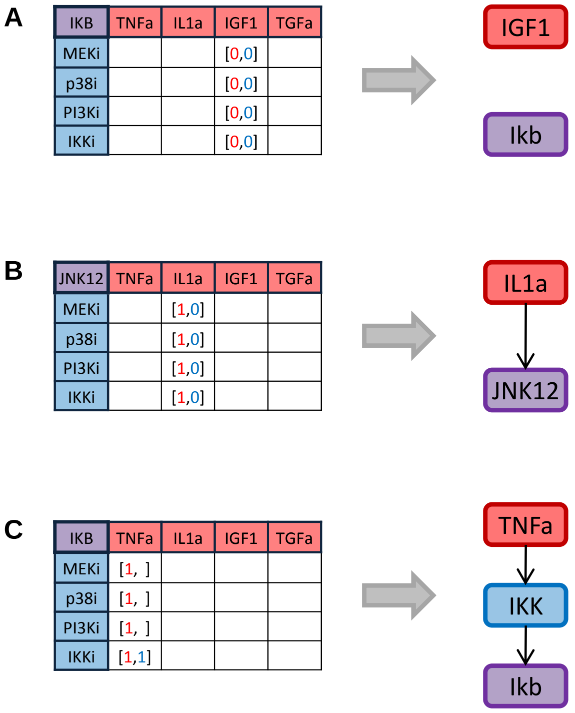
Figure 2: Boolean inference. Three examples are shown: A. stimulus $IGF1$ does not affect protein Ikb, B. stimulus $TNFa$ affects protein Ikb but the presence of MEK inhibitor does not change the protein level, C. stimulus $TNFa$ affects protein Ikb and the presence of IKK inhibitor decreases the protein level.

stimulus and no inhibitor ($REF1 = \nu_{0,0}(t)$)), is considered significant if it exceeds k times the standard deviation of the measurement error, for at least one sample:

$$(TEST1_i - REF1) > k \cdot \sigma_{TEST1_i - REF1} \tag{3}$$

where $\sigma^2_{TEST1_i - REF1} = \sigma^2_{\nu_{0,0}} + \sigma^2_{\nu_{i,0}}$ and k is a parameter to be set. As an example, Fig. 2A reports the activity level of Ikb protein in the condition no stimulus/no inhibitor and in the condition stimulus $IGF1$/no inhibitor. The stimulus does not significantly affect the protein activity level, i.e. condition (3) is not satisfied, thus the first value of cells in the column corresponding to the stimulus $IGF1$ is set to 0, i.e. $a_i = 0$. When condition (3) is not satisfied, as in Fig. 2A, the effect of inhibitors is not considered and the second value of the cell ($b_{i,j}$) is set equal to 0 for all inhibitors j.

As a second example, Fig. 2B shows the activity level of Ikb protein in the condition stimulus $TNFa$/no inhibitor. The stimulus affects the protein level, i.e. condition (3) is satisfied, thus the first value of cells in the column corresponding to stimulus $TNFa$ is set to 1, i.e. $a_i = 1$. When condition (3) is satisfied, the effect of each inhibitor is analyzed. Denoting as reference the condition with the stimulus and no inhibitors ($REF2_i = \nu_{i,0}$) the action of each inhibitor ($TEST2_{i,j} = \nu_{i,j}$) is considered significant



IKB	TNFa	IL1a	IGF1	TGFa
MEKi			[0,0]	
p38i			[0,0]	
PI3Ki			[0,0]	
IKKi			[0,0]	

JNK12	TNFa	IL1a	IGF1	TGFa
MEKi		[1,0]		
p38i		[1,0]		
PI3Ki		[1,0]		
IKKi		[1,0]		

IKB	TNFa	IL1a	IGF1	TGFa
MEKi	[1,]			
p38i	[1,]			
PI3Ki	[1,]			
IKKi	[1,1]			

Figure 3: Network reconstruction. Three examples are shown: A. stimulus $IGF1$ does not affect protein Ikb, B. stimulus $IL1a$ affects protein $JNK12$ but none of the inhibitors exerts a significant effect, C. stimulus $TNFa$ affects protein Ikb and its action is mediated by protein IKK.

if:

$$(REF2_i - TEST2_{i,j}) > k \cdot \sigma_{(TEST2_{i,j} - REF2_i)} \tag{4}$$

where $\sigma^2_{TEST2_{i,j} - REF2_i} = \sigma^2_{\nu_{i,0}} + \sigma^2_{\nu_{i,j}}$. Fig. 2B shows that if protein Ikb is stimulated with stimulus $TNFa$/inhibitor $MEKi$, condition (4) is not satisfied and the second value of the cell corresponding to stimulus $TNFa$ and inhibitor $MEKi$ is set to 0, i.e. $b_{i,j} = 0$. Whereas, with inhibitor $IKKi$ (Fig. 2C) condition (4) is satisfied, thus $b_{i,j}$ is set equal to 1. It is clear from the examples that the number of actions considered as significant is inversely related to the k value.

2.2 Network reconstruction

For each protein, a subnetwork is reconstructed from its Boolean table by adding:

- no links for stimulus/inhibitor combinations corresponding to [0,0] cells (example shown for protein Ikb under stimulation with stimulus $IGF1$ in Fig. 3A);
- a direct link between a cytokine and a phosphoprotein if the column corresponding to that stimulus contained all [1,0] cells (e.g. $IL1a \rightarrow JNK12$ in Fig. 3B);

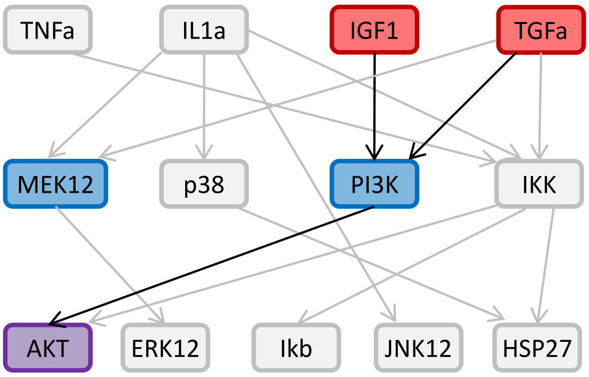


Figure 4: Subnetwork isolation for prediction. Example of the subnetwork considered when AKT value under stimulation with stimuli $IGF1$ and $TGFa$ and inhibitors $MEKi$ and $PI3Ki$ had to be predicted.

- a link passing through and inhibitor if, for that stimulus, there is a [1,1] cell in the row corresponding to that inhibitor (e.g. $TNFa \rightarrow IKK \rightarrow Ikb$ in Fig. 3C).

Subnetworks are then merged, and if, in the resulting network, a cytokine and a protein are connected both directly and indirectly, through an inhibitor, the direct link is pruned and not used for prediction. The Boolean tables are updated consistently.

2.3 Prediction

To predict the phosphorylation level reached by a protein in combinatorial treatments with single or multiple stimuli/multiple inhibitors, the specific subnetwork is isolated. For example, to obtain the prediction of the activity of protein AKT in the condition with stimuli $TGFa$ and $IGF1$ and inhibitors $PI3Ki$ and $MEK12i$, the sub network composed by nodes $TGFa$, $IGF1$, $PI3K$, $MEK12$, AKT and links connecting them are isolated, as shown in Fig. 4.

Depending on the subnetwork configuration, single-stimulus/inhibitor data are linearly combined according to the following formula:

$$\hat{x}_{I,J} = \sum_{i \in I} a_i \sum_{j \in J} b_{i,j} \left(\nu_{i,j} - \nu_{i,0} - \nu_{0,j}\right) + \sum_{i \in I} a_i \nu_{i,0} + \sum_{j \in J} b_{0,j} \nu_{0,j} + x_{I,J\ b} \tag{5}$$

where I and J denote the particular combinations of stimuli (e.g $TGFa + IGF1$) and inhibitors (e.g. $MEKi + PI3Ki$), respectively, for which prediction $\hat{x}$ has to be made, $x_{I,J\ b}$ the basal level under this condition (given by the organizers) and $b_{0,j}$ is assumed equal to 1 if $b_{i,j} = 1$ for at least one $i \in I$. If none of I stimuli are active on the protein, i.e. $a_i = 0$ for all $i \in I$, Equation (5) reduces to:

$$\hat{x}_{I,J} = \nu_{0,0} + x_{I,J\ b} \tag{6}$$

As an example, for the subnetwork shown in Fig. 4, Equation (5) predicts the activity of protein AKT with stimuli $TGFa$ and $IGF1$ and inhibitors $MEKi$ and $PI3Ki$ as

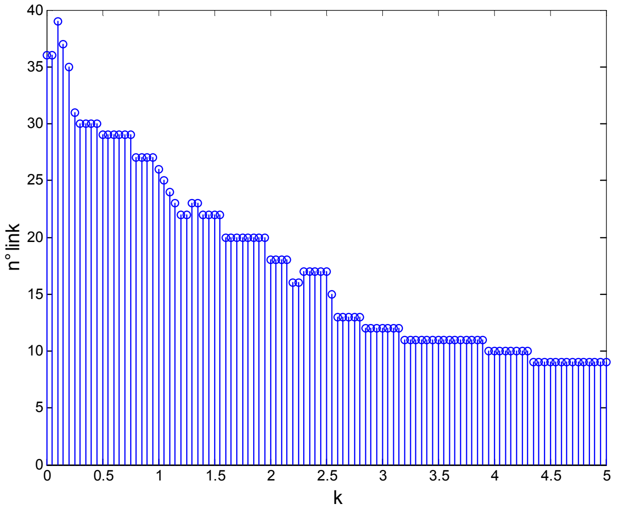


Figure 5: Influence of the parameter k on the number of links.

the sum of the activity level of protein AKT in the condition stimulus $TGFa$/inhibitor $PI3K$ and in the condition stimulus $IGF1$/inhibitor $PI3K$. Since in this sum the effect of the inhibitor is considered twice, the activity level in the condition no stimulus/inhibitor $PI3K$ is then subtracted. If, for a given protein, the reconstructed network predicts that some stimulus/inhibitor combinations do not affect its level, the reference conditions $\nu_{i,0}$ and $\nu_{0,0}$ in Equations (5) and (6) are evaluated by averaging the protein level measured in absence of stimulus/inhibitor with the protein level measured under these conditions.

2.4 Implementation

The algorithm was implemented in Matlab. It requires as input arguments: single-stimulus/inhibitor data, the model of the measurement error, the value of parameter k and multi-stimuli/inhibitors combinations for which predictions are desired and provides as outputs: the reconstructed network with link ranking and predicted values.

3 Results

3.1 Network inference

The choice of parameter k, used in the inference of Boolean tables to define the threshold of significance (Equations (3) and (4)), obviously affects the number of links, as shown in Fig. 5: with a high value of k only few links are included in the network, more are added if k decreases. In Table 1, selected links are ranked, according to the upper limit value of parameter k still allowing the presence of the link, from the most reliable (high value of k) to the less confident. A value of k equal to 2.5 was empirically chosen as

LINK	k	Canonical network
$IL1a \rightarrow IKK \rightarrow Ikb$	10.70	yes
$IL1a \rightarrow p38 \rightarrow HSP27$	9.74	yes
$TGFa \rightarrow MEK12$	8.71	yes
$TGFa \rightarrow PI3K \rightarrow AKT$	5.01	yes
$IGF1 \rightarrow PI3K \rightarrow AKT$	4.38	yes
$TNFa \rightarrow IKK \rightarrow Ikb$	4.34	yes
$IL1a \rightarrow JNK12$	4.31	yes
$IL1a \rightarrow MEK12$	3.91	no
$IL1a \rightarrow IKK \rightarrow HSP27$	3.17	no
$TGFa \rightarrow MEK12 \rightarrow ERK12$	2.76	yes
$IL1a \rightarrow p3$	2.64	yes
$TGFa \rightarrow IKK \rightarrow AKT$	2.55	no

Table 1: Links in the network ranked according to the upper limit value of parameter k allowing the presence of the link.

threshold. It permitted to have an high number of true positives (i.e. links that are both in the canonical and reconstructed network) still limiting the number of false positives (i.e. links that appear in the reconstructed but not in the canonical network). Thus, the canonical network was used only to set a threshold valid for links to be selected, not as a priori information on which links are included in the network.

The cause-effect network (after graph pruning), used for prediction, is shown in Fig. 1. A direct connection between a cytokine (represented in red) and a measured protein (in purple), e.g. $IL1a \rightarrow Ikb$, means that the cytokine stimulation significantly increased the activity level of the target protein. A connection through one of the inhibited proteins (in blue) , e.g. $TNFa \rightarrow IKK \rightarrow Ikb$, means that the cytokine stimulates the target protein level, but if the halfway protein is inhibited the target protein level decreases with respect to the previous condition. All inferred links can be found in the canonical network but three: the one connecting $IL1a$ and $MEK12$ also found by Saez-Rodriguez et al. (Saez-Rodriguez et al., 2009) and the ones connecting IKK to AKT and $HSP27$. From Table 1, the connection between IKK and AKT is the last link is the ranking therefore it is the less confident. On the contrary the connection between IKK and $HSP27$ seems to be quite reliable.

3.2 Prediction

The average Normalized Error (NE), i.e. the square root of NSE for each prediction, was 1.47 corresponding to an average deviation of prediction from measurement equal to 1.47 times the SD of the measurement error. In Fig. 6, an histogram of single prediction NEs reveals that there were some outliers. Thus, the median NE was lower than the average NE and its value, equal to 0.38, indicates that the distance between the prediction and the real value was less than the 38% of the SD of the measurement error for the 50%

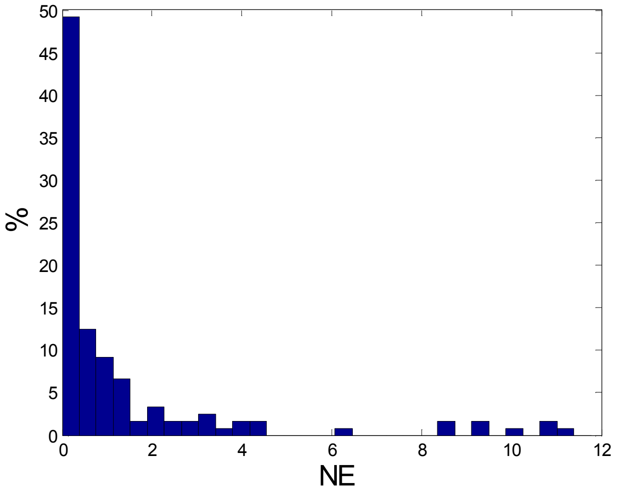


Figure 6: NE histogram. Mean and median NE values were 1.47 and 0.38 respectively.

	TOT	AKT	$ERK12$	Ikb	$JNK12$	$p38$	$HSP27$	$MEK12$
mean NE	1.47	5.03	1.25	0.46	0.45	0.24	0.36	1.96
median NE	0.38	3.84	0.82	0.19	0.33	0.15	0.15	2.05

Table 2: Mean and median NE over all predicted values for each protein.

of the predictions. Results for single proteins (Table 2) show that predictions are more precise for some proteins (e.g. $p38$ and $HSP27$), less precise for others, particularly for Akt, but in most cases the median is lower than the mean, indicating that outliers are distributed among proteins.

In order to evaluate the role of parameter k on the performance, $Prediction\ Score$ and $Overall\ Score$ calculated from Equations (2) by using $r = 0.0827$, which is the value evaluated by the organizers based on the results of all teams, were plotted for different values of k (Fig. 7). Fig. 7A shows that a high $Prediction\ Score$ was obtained only for $1.4 < k < 2.7$ indicating that a reliable network was necessary for the quality of predictions. In fact, low values of k, i.e. networks with many links, and high values of k, i.e. networks with few links, worsened the performance of the method in terms of $Prediction\ Score$. However, the $Overall\ Score$, which favored sparse networks, indicated a good performance even with high k values, as shown in Fig. 7B.

4 Discussion

In this paper, we present a simple method able to reconstruct, from single- stimulus/inhibitor protein data, cause-effect networks representing signaling pathways and to predict protein levels during multi-stimulus/inhibitor perturbations. This method, developed and applied to the Predictive Signaling Network Modeling challenge of DREAM4

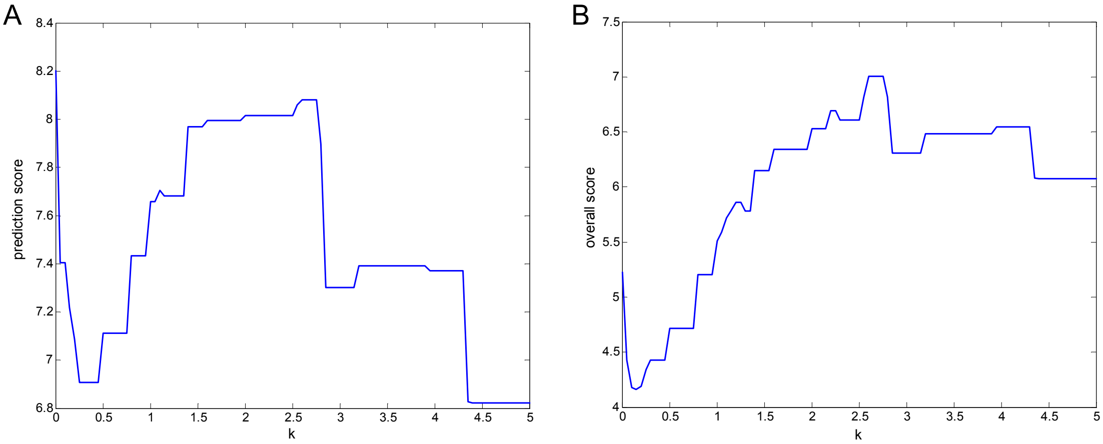


Figure 7: Influence of the parameter k on the performance of the method. (A) The *Prediction Score* is used to evaluate the statistical significance of the prediction, a greater value correspond to a better prediction. (B) The *Overall Score* consider also the parsimony of the network penalizing the Prediction Score with a cost per link. The method is robust to the value of k at least for $2 < k < 2.8$.

competition, can be used to discover how signaling pathways are altered by diseases and to predict the effect of multiple agents/drugs. It uses a data driven approach, having Boolean (discrete logic) inference and linearity assumption as basic ingredients underlying network reconstruction and data prediction.

Boolean inference is appropriate to reconstruct the signaling network structured into input nodes (stimuli) intermediate nodes (inhibitors) and output nodes (phosphoproteins), particularly in situations like the one of the challenge, where the limited number of available samples and the lack of information on the stimulus format prevent the use of more sophisticated modeling approaches, e.g. based on differential equations and model identification. A cause-effect network connecting stimuli, inhibited and measured proteins, was reconstructed by a two step procedure: from single-stimulus/inhibitor data, a table was first built to code significant effects of stimuli and inhibitors on output proteins, which was then translated into links among nodes of the network according to very simple rules. Significance was defined with reference to the measurement error, by exploiting a method used in (Di Camillo et al., 2005) to quantize time series expression data, e.g. a stimulus significantly affects an output protein if it is able to increase its level of a quantity that exceeds the uncertainty associated with the measurement of this quantity. The method needs information about the measurement error: in the case of the challenge a model relating the variance of the error to the expression level was provided by the organizers, in situations where this information is not available, it can be estimated from replicates (Cobelli et al., 2001). A factor k, which multiplies the standard deviation of the errors, was introduced as a threshold to distinguish between not significant (to be explained in terms of measurement errors) and significant effects. The choice of k obviously affects network density, as shown in Fig. 5: low k values favor dense networks and may lead to false positive links; whereas high k values cause sparse networks, potentially associated with false negative links. Thus, k was optimized using the available knowledge built in the canonical network. A value $k = 2.5$ was chosen, able to provide a network with most of the links reproducing direct or indirect connections also present in the canonical pathway. Therefore, a priori information built in the canonical network was only used to set parameter k. Anyhow, the described network reconstruction approach is strictly data-driven and thus usable even when no information is available: to set parameter k we are exploring different solutions, based on either the stability of predictions or the ranking of links. For example, active links can be selected following a method based on a compromise between false positives and false negatives based on a measurement error model, originally proposed to quantize gene expression data (Di Camillo et al., 2005).

The reconstructed network was used to predict protein levels during multi- stimulus/inhibitor perturbations, by linear combination of single-stimulus/inhibitor data which, according to the network, exert significant effects on the proteins. Linearity assumption underlies the predictions, and this can be critical since interferences among different stimuli and/or different inhibitors are likely to occur in the real system. However, no information is available on whether and how interferences take place, therefore linearity is a sort of minimum working assumption, the role of which can be assessed a posteriori, based on the performance of the method in terms of reliability of predictions. Results indicate that the linearity assumption is reasonable, since the median of the deviation between true and predicted values is about 0.38 when normalized to the standard deviation of the measurement error. Performance is reasonably stable with respect to k values, reaching similar prediction scores for k in the range $2 - 2.8$. Choosing a low value, resulting in a dense network, as well as a high value, resulting in a sparse network, deteriorates the quality of predictions. This supports the importance of using a realistic network to select the single-stimulus/inhibitor components to be linearly combined for data prediction.

In conclusion, the method we proposed provides a reliable solution to the problem proposed in the challenge. The method is simple, its implementation in Matlab has very low computational load but, despite of its simplicity, it is very promising and we are currently working on some refinements. As regards the definition of significance of the stimulus/inhibitor effect, we plan to introduce a criterion tailored for time series data, based on the area under the curve like in (Di Camillo et al., 2007), instead of considering single data points. A second aspect regards the choice of parameter k, which is a critical issue of our method, in the situation where a priori knowledge of network density is not available/usable.

References

Alexopoulos, L., Saez-Rodriguez, J., Cosgrove, B., Lauffenburger, D., & Sorger, P. (2010). Networks inferred from biochemical data reveal profound differences in toll-like receptor and inflammatory signaling between normal and transformed hepatocytes. *Molecular & Cellular Proteomics*, *9*(9), 1849–1865.

Cobelli, C., Foster, D., & Toffolo, G. (2001). *Tracer kinetics in biomedical research: from data to model*. Springer.

Di Camillo, B., Sanchez-Cabo, F., Toffolo, G., Nair, S., Trajanoski, Z., & Cobelli, C. (2005). A quantization method based on threshold optimization for microarray short time series. *Bmc Bioinformatics*, *6*(Suppl 4), S11.

Di Camillo, B., Toffolo, G., Nair, S., Greenlund, L., & Cobelli, C. (2007). Significance analysis of microarray transcript levels in time series experiments. *BMC bioinformatics*, *8*(Suppl 1), S10.

Jones, D. (2008). Pathways to cancer therapy. *Nature Reviews Drug Discovery*, *7*(11), 875–876.

Li, F., Thiele, I., Jamshidi, N., & Palsson, B. (2009). Identification of potential pathway mediation targets in toll-like receptor signaling. *PLoS computational biology*, *5*(2), e1000292.

Mitsos, A., Melas, I., Siminelakis, P., Chairakaki, A., Saez-Rodriguez, J., & Alexopoulos, L. (2009). Identifying drug effects via pathway alterations using an integer linear programming optimization formulation on phosphoproteomic data. *PLoS computational biology*, *5*(12), e1000591.

Prill, R., Marbach, D., Saez-Rodriguez, J., Sorger, P., Alexopoulos, L., Xue, X., Clarke, N., Altan-Bonnet, G., & Stolovitzky, G. (2010). Towards a rigorous assessment of systems biology models: the dream3 challenges. *PloS one*, *5*(2), e9202.

Saez-Rodriguez, J., Alexopoulos, L., Epperlein, J., Samaga, R., Lauffenburger, D., Klamt, S., & Sorger, P. (2009). Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. *Molecular systems biology*, *5*(1).

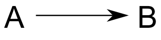
Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, E., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545–15550.

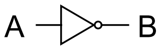
3

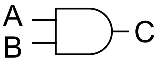
Qualitative modelling of networks

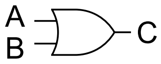
3.1 Introduction

Model are convenient tools to interpret networks in order to study functional interactions between their components. Causal interactions are, intrinsically, qualitative relations between the observed variables; for this reason, biological networks are often interpreted as qualitative logic-based models. Qualitative descriptions are very appealing when it comes to analyze complex biochemical networks, because they are close to the human way of thinking, for example, when A, B and C are molecules, some common interactions that can easily be translated into logic models are:

- B is active when A is active 

(activation)
- B is inactive when A is active 

(NOT gate, inhibition)
- C is active only when A and B are active 

(AND gate)
- C is active when either A or B is active 

(OR gate)

These rules can be combined in order to define more complex behaviors. The use of logic models allows to extend this intuitive properties from subsystems to large scale networks:

the overall behavior of the network might not be intuitive to understand just looking at the network, but can be quickly simulated by a computer (Blinov & Moraru, 2012). In this way, different hypothesis about possible interactions between molecules can be tested by comparing simulations with qualitative behavior observed in experimental data, and the model can be refined based on these tests.

Logic-based models have been largely applied also to signalling networks as reviewed in Morris et al. (2010): in this case, nodes in the graph represent proteins and the edges represent their causal relationship. The state of each component of the network represents the activity of the protein based of the activity of the input nodes, without going into the biochemistry of each interaction.

CellNOptR

Logic gates can be generated by manual curation of the literature, however, when it comes to middle and large scale network, is is not easy to find the combination of gates that best describe all experimental data. CellNOptR (Terfve et al., 2012) is an R software package (www.cellnopt.org) for automatic creation of logic-based models of signal transduction networks using different logic formalisms, based on the method described in Saez-Rodriguez et al. (2009). The method requires two inputs: the prior knowledge network (PKN), that is the state of the art of known interactions between network components, and an experimental dataset. PKNs can be downloaded from public databases, but they are comprehensive networks collected under different experimental conditions and on different cell types. CellNOptR allows to interpret these networks as logic models and to train them to the data in order to find the logic model which best describe the experimental data. The optimized model is, therefore, context and cell line specific and has predictive power. A series of packages have been developed to extend the core package CellNOptR in order to include, in a unique framework, different logic formalisms: Boolean single/multiple steady-state, Boolean discrete time, steady-state fuzzy logic and logic-derived ODEs. These formalisms include a different level of granularity in both time and state: time can go from steady state to discrete to continuous and species state can go from binary (Boolean) to multi-state to continuous.

3.2 CNORfeeder: network inference for logic modelling

CellNOptR is able to handle large amounts of data because the search space for logic models is limited by the prior knowledge. However, this is also a limitation because databases are not complete and there might be missing links that cannot be recovered by the algorithm. For this reason, we developed CNORfeeder (Eduati et al., 2012a), an add-on package, developed to be integrated with CellNOptR, that permits to extend a network derived from literature with links derived strictly from the data via various inference methods, using information on physical interactions of proteins to guide and validate the integration of links. The integrated pipeline include the following steps:

- data are used to infer a data-driven network (DDN) using reverse-engineering methods (as for now FEED, ARACNE, CLR and Bayesian networks are implemented);
- the prior knowledge network (PKN) is compressed according to the data removing non-identifiable nodes;
- the compressed network is integrated with the DDN;
- information derived from protein-protein interaction network (PIN) is used to support and prioritize integrated links;
- the integrated network is used as input for the training.

This methods was applied to real data of growth and inflammatory signalling providing a refined logic-based model with better fit to data with respect to training using only prior knowledge. It also highlighted plausible links that were missing in the PKN and are supported by known interactions among proteins.

The approach and results obtained from its application are described in detail in Eduati et al. (2012a); full text of the original paper is reported in Appendix 3.1.

3.3 Comparison with ASP method

CellNOpt uses a genetic optimization algorithm to optimize the network to the data in order to find the Boolean logic model that best describes the existing data. Being a stochastic algorithm, it cannot guarantee to find the global optimum and it scales poorly since the search space (and thus the computational time) increases exponentially with the network size. A novel method to solve the optimization problem was proposed in Videla et al. (2012) in order to overcome this limitations. This approach encodes the optimization problem in Answere Set Programming (ASP), a declarative problem solving paradigm which guarantee the global optimum by reasoning over the complete solution space. We showed how multiple models can equally describe experimental data mainly for two reasons: not all nodes in the network are measured and not all combinatorial perturbations (stimuli/inhibitors) are tested. This reasoning can be used also in the design of optimal experimental setup (Guziolowski et al., Submitted).

Appendix 3.1 Paper: Eduati et al., Bioinformatics, 2012

The following publication dealing with qualitative modelling of networks has been coauthored by the Ph.D. candidate during her doctoral program.

- F. Eduati, J. De Las Rivas, B. Di Camillo, G. Toffolo, and J. Saez-Rodriguez. *Integrating literature-constrained and data-driven inference of signalling networks*. Bioinformatics, 28(18):2311-2317, 2012.

Full text of the original paper is reported in this Appendix.

Integrating literature-constrained and data-driven inference of signalling networks

F. Eduati[1,2], J. De Las Rivas[3], B. Di Camillo[1], G. Toffolo[1] and J. Saez-Rodriguez[2,*]

[1]Department of Information Engineering, University of Padova, Padova, Italy
[2]European Bioinformatics Institute (EMBL-EBI), Cambridge, UK
[3]Bioinformatics & Functional Genomics Group, CSIC/USAL, Salamanca, Spain

[*]*Corresponding author:* saezrodriguez@ebi.ac.uk

Abstract

Motivation: Recent developments in experimental methods facilitate increasingly larger signal transduction datasets. Two main ap-proaches can be taken to derive a mathematical model from these data: training a network (obtained, for example, from literature) to the data, or inferring the network from the data alone. Purely data-driven methods scale up poorly and have limited interpretability, while literature-constrained methods cannot deal with incomplete networks.

Results: We present an efficient approach, implemented in the R package CNORfeeder, to integrate literature-constrained and data-driven methods to infer signalling networks from perturbation ex-periments. Our method extends a given network with links derived from the data via various inference methods, and uses information on physical interactions of proteins to guide and validate the integration of links. We apply CNORfeeder to a network of growth and inflammatory signalling. We obtain a model with superior data fit in the human liver cancer HepG2 and propose potential missing pathways.

Availability: CNORfeeder is in the process of being submitted to Bioconductor and in the meantime available at www.cellnopt.org.

Contact: saezrodriguez@ebi.ac.uk

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Information about signalling networks is increasingly abundant. Thanks to novel high-throughput methods, large amounts of data about the interactions among proteins is available, which is en-compassed in (unsigned and undirected) protein-protein interaction networks (PINs) (Pieroni et al., 2008). More precise (but with less coverage)

Eduati et al., Bioinformatics, 2012, Vol. 28 no. 18, pages 2311–2317

information is derived from literature and is often described by means of signed and directed causal interactions among proteins. These give rise to what we will call here prior knowledge networks (PKNs). PKNs are partially collected in different databases (e.g. KEGG (Ogata et al., 1999), Reactome (Joshi-Tope et al., 2005), WikiPathways (Pico et al., 2008), and several are ac-cessible via the portal Pathway Commons (Cerami et al., 2011). These databases typically contain literature-derived interactions curated with different degrees of stringency, and based on experimental publications under different experimental conditions using different cell types.

PKNs are, for example, very useful to study topological properties of networks (Ma'ayan et al., 2005) or to map data (Ideker & Sharan, 2008; Terfve & Saez-Rodriguez, 2012). However, they are not functional in the sense that they cannot be used for simulation of a signalling process and therefore prediction of the outcome of a certain experiment, which is fundamental to understand signal transduction and its alterations.

The most common way to model a signalling network is to write down its biochemistry and subsequently translate it to a mathe-matical form, typically a system of differential equations (Aldridge et al., 2006). However, information in PKNs often lacks the required mechanistic detail. In these cases, logic formalisms are a useful approach since all they need is to add logic gates to the existing (signed and directed) interactions.

One can generate logic gates by manual curation based on literature, e.g. (Calzone et al., 2010; Saadatpour et al., 2011; Samaga et al., 2009), reviewed in (Morris et al., 2010; Watterson et al., 2008). An alternative to manual curation consists of generating a logic model from the PKN that is subse-quently trained to experimental data (Saez-Rodriguez et al., 2009). This method, implemented in the Bioconductor package CellNOptR (http://www.ebi.ac.uk/ saezrodriguez/software.html), provides context-specific models with predictive power. It is efficient at handling large amounts of data as the space of possible models is limited by the prior knowledge. This key feature of the approach, however, is also its main limitation: there might be missing links as databases are not complete, and the effect of cross talk between pathways is often not taken into account in the canonical linear representation of the pathways. Hence, adding links to the PKN based on the dedicated data can lead to an improved goodness of fit (Saez-Rodriguez et al., 2009).

With a different and complementary perspective, different "reverse engineering" methods have been used to infer networks from perturbation experiments using data-driven methods that do not rely on prior knowledge of the network (Bansal et al., 2007; Markowetz, 2010). Most of these methods were first developed for transcriptional data but can be applied also to signalling data. For example, in (Ciaccio et al., 2010) Bayesian networks (Pe'er, 2005) were used to infer the connections between 67 proteins with high-throughput data collected using a micro-western array. Two mutual information based approaches, the "algorithm for the reconstruction of accurate cellular networks" (ARACNe) (Margolin et al., 2006) and the "context likelihood of relatedness" (CLR) (Faith et al., 2007), were also applied to the same dataset to corroborate the results. Different methods were also applied in the context of the DREAM initiative (www.the-dream-project.org) for the DREAM4 Predictive Signalling Network Challenge (Prill et al., 2011). 12 research groups inferred signalling networks from perturbation experiments data and were evaluated based on the accuracy of their predictions of the outcome of the network under different experimental conditions. One of the methods that performed best in this task was a simple approach, strictly data-driven, that encodes significant effects of stimuli and inhibitors on measured proteins in a cause-effect network (Eduati et al., 2010).

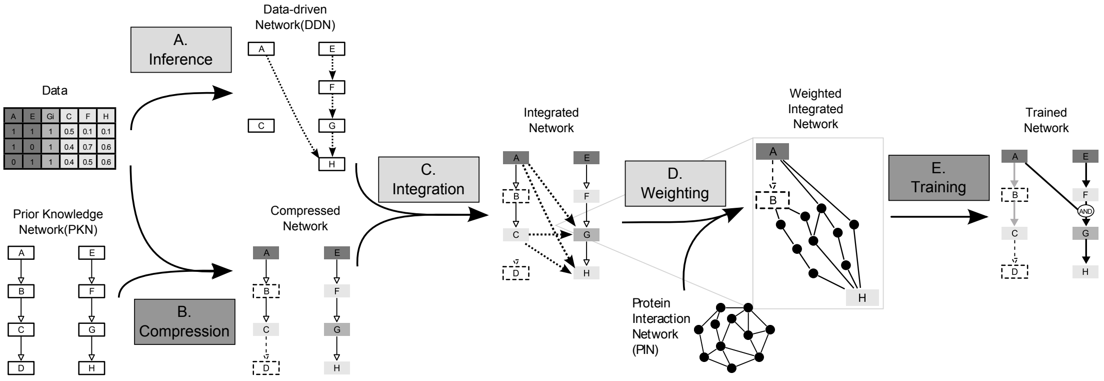
These purely data-driven methods need to consider all possible topologies, and thus in general need more data and scale-up worse than methods that rely on a given topology such as CellNOptR. Furthermore, the resulting data-driven networks (that we will call here DDNs) are limited to interactions between perturbed and measured nodes that are only a subset of the nodes involved in the pathways. Thus, DDNs are not as biologically interpretable as the PKNs and mapping DDNs to PKNs is not simple as one link in the inferred network can generally correspond to multiple links in the PKN. Hence, it is not trivial how to correctly map this relationship. In this paper we attempt to combine the strengths of literature-based and data-driven inference methods. We describe a procedure (implemented in the R package CNORfeeder), to integrate prior knowledge encoded in the PKN with data-driven information obtained using reverse engineering approaches. PINs are used to prioritize links and to provide experimental support for them, and thus help to discriminate among options and add information on integrated links. The resulting network is then trained against experimental data to obtain a final refined model that has a better fit to data with respect to the PKN, highlighting plausible links that were missing in the PKN. We illustrate its application with a signalling network encompassing multiple pathways and readouts trained with data from the liver cancer cell HepG2. We show how CNORfeeder provides a significantly improved fit based on links supported by known interactions among proteins.

2 Methods

We implemented CNORfeeder, an R package designed to be integrated with methods based on prior-knowledge such as CellNOptR as shown in Fig. 1. The integrated pipeline can be summarized in the following steps:

A. Inference (CNORfeeder) A strictly data-driven network (DDN) is inferred from available data using different reverse engineering methods (so far FEED, Bayesian networks, ARACNe and CLR). This network is specific for the experiments under study, thus it only includes perturbed and measured nodes and does not exploit information available in literature.

B. Compression (CellNOptR) The prior knowledge network (PKN) is com-pressed according to the procedure detailed in (Saez-Rodriguez et al., 2009). First, if a node has no readout downstream of it (such as D in Fig. 1), its state cannot be inferred (it is non-observable), and is not considered. Similarly, if a node has no perturbation upstream, it is not included as it will not be affected. Then, nodes that are neither perturbed nor measured are bypassed so that their compression does



A	E	Gi	C	F	H
1	1	1	0.5	0.1	0.1
1	0	1	0.4	0.7	0.6
0	1	1	0.4	0.5	0.6

Figure 1: Integrated pipeline of CNORfeeder (light boxes) and CellNOptR (dark boxes). A. Data are used to infer, using reverse-engineering methods, a strictly data-driven network (DDN); B. the prior knowledge network (PKN) is compressed according to the data (dark, middle and light grey nodes are respectively stimulated, inhibited and measured), removing non-identifiable nodes (dashed); C. the compressed network is integrated with the DDN (dotted links are obtained from the DDN and continuous links from the PKN); D. information derived from protein-protein interaction networks (PINs) are used to support and prioritize integrated links; E. The integrated network is used as input for the training: in the trained model, thick black lines denote interactions (and gates) in the trained model, and light-grey links denote presence in the integrated network but not in the trained model.

not change the logic of the remaining nodes (e.g. B which is between A and C in Fig. 1).

C. Integration (CNORfeeder) The compressed network is expanded using the DDN in order to include links that are missing in the a priori information but that seem to be supported by data.

D. Weighting (CNORfeeder) Protein interaction networks (PINs) are used to support and prioritize the integrated links.

E. Training (CellNOptR) The integrated network is finally converted in to a superstructure containing all possible logic gates compatible with the net-work. If a node (such as G) is affected by multiple nodes (A and F), then both an OR and an AND gate are created. Then a genetic algorithm is used to search for the model contained in the superstructure which best describes the data (as determined by a score based on the mean squared error) with the minimum number of links. The objective function is modified with respect to that introduced in (Saez-Rodriguez et al., 2009) to include additional penalization for the integrated links using weights derived from PINs (Equation (2)).

Steps performed by CNORfeeder will be detailed in the following sections.

2.1 Inference using reverse-engineering methods

CNORfeeder can in principle leverage any network inference method. So far, we have integrated the following:

FEED inference

It is the R implementation of an improved version of the algorithm described in (Eduati et al., 2010). The inference of the network can be divided in two steps. Fist, perturbation experiments are used to infer a Boolean table for each measured protein, codifying if a particular stimulus inhibitor combination affects the protein. A stimulus or an inhibitor significantly affects an output protein if it is able to modify its activity level by a quantity that exceeds the uncertainty associated with its measurement. These Boolean tables are than translated into links among stimulated, inhibited and measured nodes, giving rise to the inferred network (see Supplementary Material for more details).

Bayesian Network inference

There are different approaches to derive causal influences between measured proteins using Bayesian networks. We have used the “catnet” R package (available from http://cran.r-project.org/web/packages/catnet/index.html) to derive categorical Bayesian networks from static data (see Supplementary Material for more details).

Mutual information networks

This class of methods computes the mutual information matrix between the measurements associated with different proteins and, based on that, infers an undirected network. In particular, ARACNe and CLR algorithms as implemented in the “minet” R package (Meyer et al., 2008) (see Supplementary Material for more details), are included in CNORfeeder.

In silico data were generated using a “Gold Standard” or true network, depicted in Fig. 2E, to compare the four algorithms. The “Gold Standard” was randomly generated and interpreted as a logic Boolean model to simu-late perturbation experiments using CellNOptR. This was performed by stimulating (nodes in dark grey), inhibiting (nodes in middle grey) and measuring (nodes in light grey) the specified proteins. These in silico data were then given as input to the inference methods; resulting networks (DDNs) are shown in Fig. 2A-D. The advantage of this approach, with respect to the use of real data, is that the Gold Standard can be used to compare the performances of the different methods.

In Fig. 2A-D dark grey links are those that are perfectly reconstructed being present both in the Gold Standard and in the inferred network. Some of the links in the

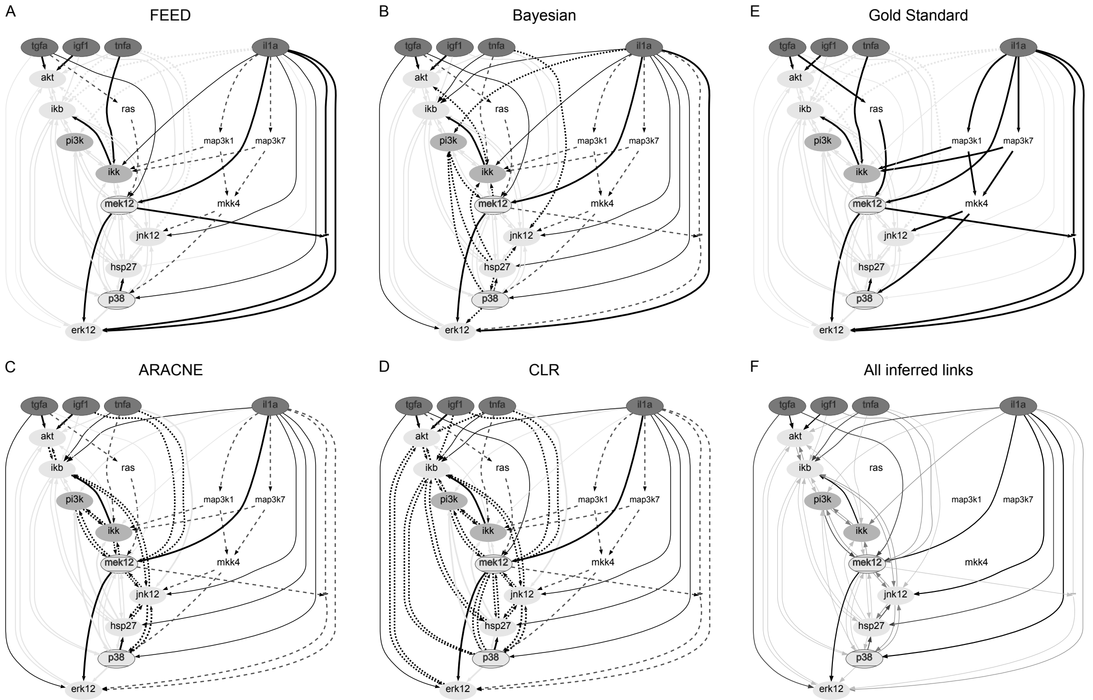


Figure 2: Reverse engineering of a Gold Standard network (E) using four different inference methods (A-D). Dark, middle and light grey nodes are respectively stimulated, inhibited and measured. Link styles represent the comparison of the inferred networks (DDNs) with the Gold Standard: dark thick continuous for links in both networks, dark thin continuous for links in the DDN that correspond to a path in the Gold Standard, dashed for links in the Gold Standard not present in the DDN and dotted for links in the DDN that are not in the Gold Standard, light grey for links that are not in the network under examination but are in one of the other networks. In panel F the gradation of grey represent the consistency between DDNs in panels A-D.

Gold Standard are not inferred by the algorithm (dashed ones), e.g. $tgfa \rightarrow ras$ or $ras \rightarrow mek12$ for FEED. However in some cases the algorithm is still able to infer at least an indirect link (light grey ones), e.g. $tgfa \rightarrow mek12$ for FEED. Dotted links are those that are inferred by the algorithm but do not correspond to links in the Gold Standard even as indirect links. In this example, FEED is able to infer all links in the Gold standard, at least as indirect ones, without including any false positive links. It is important to notice that mutual information approaches do not allow for determining the directionality of the links; for light and dark grey links the directionality was assessed based on comparison with the Gold Standard to simplify the Figure. A similar approach can be used in real cases by comparison with the PKN, but there is no way to assess the directionality of missing links. In Fig. 2F, links are represented with different gradations of grey according to the consistency between the analyzed inference methods: black links are reconstructed by all methods. As expected, links involving proteins that are neither perturbed nor measured (white nodes) can-not be reconstructed by any inference algorithm. However, those nodes can be important for the signalling network and often there is available literature derived information about their role. This is one of the reasons why it is fundamental to integrate the information derived from data-driven inference methods with the prior knowledge obtained from other resources.

2.2 Integration with the PKN

Some of the links included in the DDN might be missing in the PKN, and are thus candidates to be integrated with it. However, the PKN generally includes more nodes with respect to the DDN and a link in the DDN could, in some cases, correspond to more than one link in the PKN. As shown in Fig. 1C, if there is a connection between a cue (i.e. a stimulated or an inhibited protein) and a measured protein in the DDN (e.g. from A to H), we have to connect all nodes in the different paths corresponding to that link. This means adding a link not only from the cue to the measured protein, but also from all nodes downstream of the cue, until the following cue is reached, to all nodes upstream of the measured proteins, until the previous measured protein is reached.

2.3 Protein-protein interaction network

The human protein-protein interaction network was built using a unified PPI dataset obtained as APID (Prieto & Rivas, 2006), by the combination of interactions coming from six source databases. The starting whole dataset was composed of 68488 human physical protein-protein interactions validated by at least one experimental method and reported in one article published in PubMed. From this dataset we obtained two PPI subsets with increasing confidence: a set of 28971 interactions validated by at least one "binary" experimental method (binary as defined in (Rivas & Fontanillo, 2010)); a set 6033 interactions validated by at least two experimental methods, one of them binary.

2.4 Weighting and training of integrated network

The integrated network is then optimized using CellNOptR to find the model which best describes the data using information from PINs to differently prioritize integrated links. As described in (Saez-Rodriguez et al., 2009), a bipartite objective function is used to balance fit and size i.e. to find models with good fit to the data but with the minimum number of links. Defining P as a Boolean vector encoding the candidate solution model (value 1 or 0 is assigned depending if the link is included or not in the model), the function that is minimized during the optimization process is the following:

$$\vartheta(P) = \vartheta_f(P) + \alpha \cdot \vartheta_s(P) \tag{1}$$

where $\vartheta_f(P) = \frac{1}{N}\sum_{n=1}^{N}(data_n - pred_n)$ is the mean squared error (MSE) deviation between the normalized experimental data (continuous values between 0 and 1), and the model prediction (binary values 0 or 1), for all N measured data points. $\vartheta_s(P) = \sum_{m=1}^{M}(\nu_m - P_m)^2$ is a term to penalize increasing model size according to a tunable parameter α. The size penalty $\vartheta_s(P)$ is computed as the weighted sum of the M links, which are mathematically hyperedges in the hypergraph that defines the model; see (Saez-Rodriguez et al., 2009) for details. The weight (ν) is given by the number of starting nodes, e.g. hyperedge $A\ AND\ B \rightarrow C$ is weighted twice compared to $A \rightarrow C$. In Equation (1) it is possible to include a tunable parameter β to allow a stronger penalization of links integrated to the PKN leading to

$$\vartheta_s(P) = \vartheta_{pkn}(P) + \beta \cdot \vartheta_{add}(P) \tag{2}$$

where the size penalty $\vartheta_s(P)$ is the sum of two terms: one for the links in the PKN ($\vartheta_{pkn}(P)$) and one for integrated links ($\vartheta_{add}(P)$). This is motivated by the fact that, being supported by literature, links in the PKN are more reliable with respect to links integrated using data-driven approaches and they should be prioritized in the training.

Additionally, integrated links can be differently prioritized based on information derived from PINs: the basic idea is that if, for a directed link $A \rightarrow B$ integrated in the PKN, there is a corresponding path in the PIN, it is more plausible that there is a molecular pathway $A \rightarrow B$. Because shorter paths are more feasible, as a first approximation the shortest path length between A and B in the PIN can be used as a reliability score for the integrated link. Since the optimization is performed on a compressed version of the PKN, one link integrated in the compressed network generally corresponds to multiple possible links integrated in the PKN (Fig. 1E). Thus, the reliability score for each integrated link i is given by $\omega_i = \sum_{j_i=1}^{J_i}\frac{1}{d_{j_i}}$, where $j_i = 1, \ldots, J_i$ are the links in the PKN corresponding to the integrated link i in the compressed network. The shortest path d is computed using the Dijkstra's algorithm implemented in the igraph R package (Csardi & Nepusz, 2006) considering the PIN as a graph where the weight of the edges is the inverse of the number of experiments (experimental evidences) that validate it.

Thus, the penalty for all A integrated links into the compressed network P, can be defined as

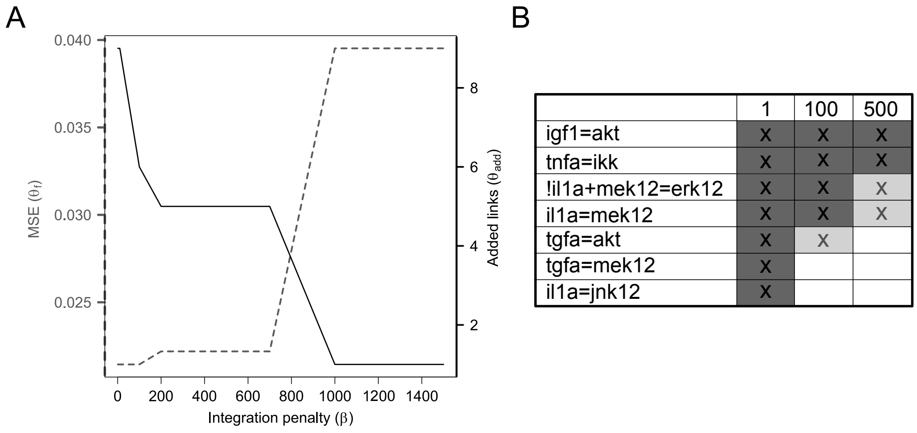
$$\vartheta_{add}(P) = \sum_{i=1}^{A} \nu_i \left(\frac{1}{\omega_i} + 1 \right) P_i \tag{3}$$

The training step, to find the P that minimizes $\vartheta_(P)$ in Equation (1), is performed with CellNOptR using a genetic algorithm that explores the P-space. The genetic algorithm is run multiple times, and in each run the values for the explored models is recorded, so that at the end a family of models is reported.

3 Results

The method was applied to a dataset of a human liver cell line (HepG2) from the DREAM 4 challenge (Prill et al., 2011), where the phosphorylation of seven proteins (akt, $erk12$, ikb, $jnk12$, $p38$, $hsp27$, $mek12$) is measured 30 minutes after combinatorial stimulation with four ligands (tnfa, il1a, igf1, tgfa) and four inhibitors (pi3k, ikk, p38, mek12). The level of phosphorylation of proteins is measured using the Luminex xMAP assay and provides a value of the phosphorylation in arbitrary units, that can be used to compare values at two conditions. In our case we compare the values between 0 and 30, and this change is a proxy of the induced activation of the corresponding protein. The normalization of this data to a value between 0 and 1 is achieved using a method based on a set of thresholds as described in (Saez-Rodriguez et al., 2009). According to the CellNOptR pipeline, the PKN was first com-pressed removing all non-observable and non-controllable nodes and then expanded as described in (Saez-Rodriguez et al., 2009) to include all possible combinations of AND and OR gates compatible with the network obtaining a total of 62 hyperedges. Additionally, 18 links inferred using FEED were integrated in the network according to the procedure previously described and the integrated network was used for optimization using CellNOptR.

Fixing $\alpha = 0.001$ the influence of the integration penalty (β) on the number of integrated links selected by the optimization process on the fit of the optimal model to the data (in terms of MSE) was tested as shown in Figure 3. As expected, a low value of β obtains the best fit but at the price of a high number of integrated links included in the optimal model (9 with $\beta = 1$). An increase of the value of β decreases the number of selected integrated links but worsens the fit to the data. With $\beta = 1000$ only the integrated link $tnfa \rightarrow ikk$ is included in the optimal model: the presence of this link is well supported by the data since it lowers the MSE from 0.064 (the optimal fit obtained with CellNOptR using as input the non-integrated network) to 0.040. The integrated links can be ranked as shown in Fig. 3B according to the highest value of β allowing their selection and thus according to their effect on the improvement of the fit. A lower number of links is selected when using the PIN to additionally penalize unsupported links (highlighted in dark grey in Fig. 3B). Those links, combined with the information from the PIN, suggest possible missing connections in the PKN. For example, in the PIN there is an interaction between the adaptor $irs1$ and the kinase pdk that would justify the link $igf1 \rightarrow akt$ in the compressed network since, in the PKN (Fig.



	1	100	500
igf1=akt	X	X	X
tnfa=ikk	X	X	X
!il1a+mek12=erk12	X	X	X
il1a=mek12	X	X	X
tgfa=akt	X	X	
tgfa=mek12	X		
il1a=jnk12	X		

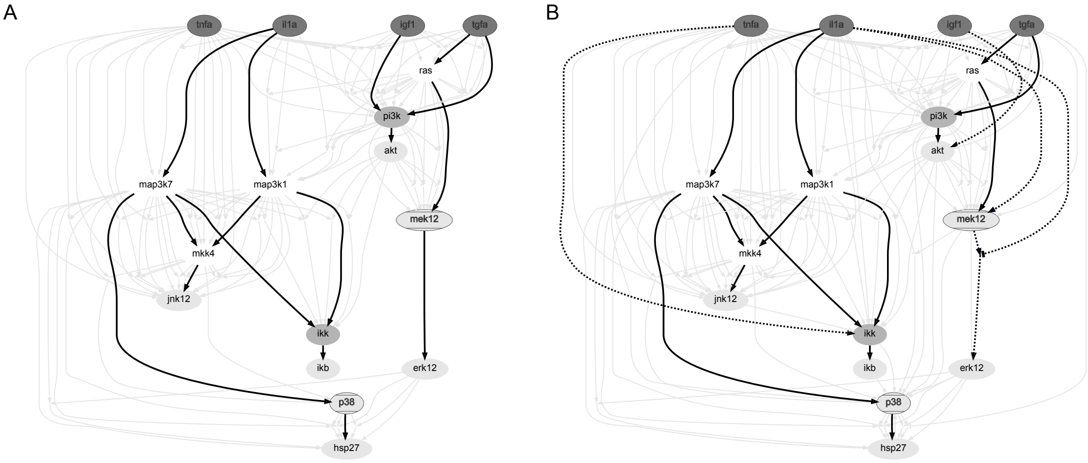
Figure 3: Panel A: effect of tuning parameter β on the number of integrated links (continuous line) and on the fit (MSE, dashed line). Panel B: links integrated for different values of β (1, 100, 500); a reduced number of links is selected when using PIN to prioritize links (dark grey).

S2 in Supplementary Material), $igf1$ binds to its receptor and pdk regulates $akt4$ (links $igf1 \rightarrow igfr$ and $pdk1 \rightarrow akt$ in Fig. S2; note that in Fig. 4 the compressed networks are shown and thus intermediates $igfr$, $irs1s$ and $pdk1$ are not present). Therefore the path $igf1 \rightarrow igfr \rightarrow irs1s \rightarrow pdk1 \rightarrow akt$ is supported by a combination of literature and interaction data. Similarly, to support the link $tnfa \rightarrow ikk$ there is a validated interaction between the $tnfa$ receptor and cot, a protein that activate ikk, leading to the combined pathway $tnf \rightarrow tnfr \rightarrow cot \rightarrow ikk$.

In Fig. 4A-B the results of CellNOptR optimization (with $\beta = 700$) are shown using as input the compressed network and the integrated network respectively. In the upper panels, optimal models are shown: links selected by the optimization algorithm are represented with continuous line if derived from the PKN and dotted line if integrated using FEED. In the lower panel the improvement in the fit is shown (from 0.064 to 0.022), which is particularly large for proteins ikb, $mek12$ and akt. In this case study, using the same parameter setting ($alpha = 0.001$, $\beta = 700$), networks integrated using ARACNe and CLR do not provide an improvement of the fit, while Bayesian networks obtain an MSE of 0.040 (see Supplementary Information). As for the computational times, FEED, CLR, and ARACNe inferred the network in $\tilde{1}$ second while Bayesian inference took $\tilde{1}$ hour on a cluster.

To evaluate the scalability of our method, we applied CNORfeeder to a larger dataset obtained also in the cell lines HepG2, comprising 7 stimuli, 7 inhibitors and 15 readouts (Saez-Rodriguez et al., 2009). We obtained comparably good results (see Supplementary Information).

Furthermore, we investigated the ability of our method to capture feedback loops, which are fundamental in the regulation of signal transduction. We constructed a toy model containing a negative feedback loops and simulated data at 2 different time points (10 and 30 minutes). We used FEED as a reverse-engineering method to retrieve, from



	akt	erk12	ikb	jnk12	p38	hsp27	mek12	**TOTAL**
MSE	0.119	0.002	0.159	0.000	0.050	0.000	0.125	**0.064**

	akt	erk12	ikb	jnk12	p38	hsp27	mek12	**TOTAL**
MSE	0.087	0.002	0.000	0.000	0.050	0.000	0.023	**0.022**

Figure 4: Results of the training of the compressed model (A) and of the integrated network (B) against data using CellNOptR. Dark, middle and light grey nodes are respectively stimulated, inhibited and measured. Selected links are represented with continuous line if derived from the PKN and dotted line if integrated, links not selected are in light grey. In the tables the fit (in terms of MSE) is reported for each measured protein along with the sum for all proteins.

the data, a link of the feedback that was missing in the PKN and then applied a recently implemented package of CellNOptR (www.cellnopt.org) that, looking also at the second time point, was able to select all links of the feedback loop (see Supplementary Information for further details).

4 Discussion

In this paper we present an approach that integrates literature-constrained and data-driven methods to efficiently infer signalling networks from experimental data collected under perturbation experiments with different stimuli and inhibitors. The procedure is implemented in the R package CNORfeeder and consists of (i) inference of a data-derived network (DDN) using strictly data-driven reverse-engineering methods (so far FEED, Bayesian networks, and mutual information approaches), (ii) integration of the DDN with a literature-derived prior knowledge network (PKN), using protein interaction networks (PINs) to prioritize and validate integrated links, and (iii) training of the integrated network against data using CellNOptR to obtain a logic model that best describe the data with the minimum number of links.

Links that improve the fit to data with respect to the PKN alone may be missing due to the difficulty assembling all available pathway information or because of incomplete knowledge of the biology. Protein interaction networks (PINs) are used as a complementary source of information to tackle this problem. PINs contain physical interactions between proteins, including those that potentially lead to protein activations, and they typically include more nodes and many more links than those based on literature-derived pathways. For this reason they have been proposed to extend pathways (Glaab et al., 2010) but they have the main limitation of a lack of directionality. PINs are also known to have high false positive and false negative rates, and we therefore used a highly curated PIN that integrates different sources and experimental techniques. This PIN seems to be quite complete for the pathways we studied (canonical pro-growth and inflammatory pathways) since we verified that for links in the PKN there is generally also a direct connection in the PIN (Fig. 5). Interestingly, when mapping to the PIN the links integrated in the PKN, we found a corresponding short path that does not pass through other nodes of the PKN. To limit the effect of false positive links in the PIN when searching for the shortest path, we weighted the edges according to the number of experimental evidences that support them. The length of the shortest path is then used to differently prioritize the integrated links in the training of the network, but other metrics could be used to discriminate between links. PINs were previously shown to be potentially useful to find previously unknown modulators of signalling pathways in (Vinayagam et al., 2011), where a Bayesian learning strategy was applied to assign directionality to a comprehensive PIN exploiting information on the shortest path from membrane receptors to transcription factors. In our method, we can take advantage of the directed links inferred via reverse engineering to limit the paths present in the PIN we integrate, and limiting the search space for the optimization algorithm.

We have used different data-driven inference methods, and applied them to both in silico (to reverse-engineer a benchmark network with known topology) and real data (to integrate links missing in the PKN that improve the fit of the model to the data in the liver cancer cell line HepG2). Each reverse-engineering method has specific features and can be suitable for different needs: for example Bayesian networks can provide statistically rigorous results but at the price of high computational costs, while mutual information approaches are computationally fast but are limited mostly by the lack of directionality of the inferred links. FEED seems to be particularly suitable to infer causal networks from single-stimulus/single-inhibitor experiments with low computational costs but, as for now, does not exploit data from all multiple combinatorial perturbation experiments.

It is not the purpose of this study to compare reverse-engineering methods (which would require a larger set of benchmark networks with known topology and a more realistic simulation of experimental data). The spirit of the paper is more in line with the lesson derived from the DREAM challenges (Marbach et al., 2010; Prill et al., 2011) that different approaches can provide complementary insights into the same problem. We have thus employed various approaches and we plan to extend it to others in the future. Furthermore, some reverse-engineering methods can use prior knowledge, in particular Bayesian inference methods (Bender et al., 2011; Mukherjee & Speed, 2008), so that we could use the PKN or results from the training with CellNOptR to guide a further search for novel links.

To conclude, the integration of literature-constrained and data-driven inference methods overcomes the limitations of both: for purely data-driven inference methods, the poor scalability (as the search space increases exponentially) and limited biological in-

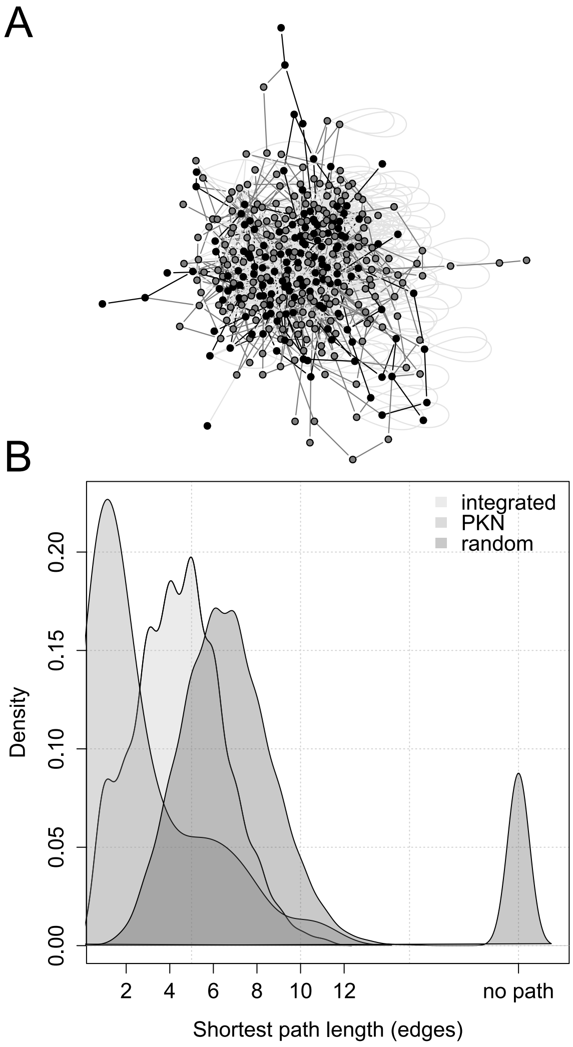


Figure 5: Mapping of the prior knowledge network (PKN) to the protein-protein interaction network (PIN). Panel A represents the subgraph of the PIN that include only nodes belonging to the PKN (dark grey) and nodes used in the mapping of integrated links (light grey); the network was plotted with the R package igraph. The same colour code is used for the edges: as expected, shortest paths between nodes in the PKN (dark grey) are generally shorter than paths used to map integrated links (light grey). This is highlighted also in panel B where the density of the shortest path length (in terms of number of edges) is plotted for integrated links, for links in the PKN and for random links.

terpretability (since they are limited to measured and perturbed proteins excluding intermediate ones), and for methods constrained to prior knowledge their inability to overcome incompleteness in the networks. We propose here an approach (and software package) to combine them that is effective and extendable to include other methods.

Acknowledgements

The authors thank T. Cokelaer for help coding the R package, A. MacNamara for critical reading of the manuscript and assistance with the figures, and L. Alexopoulos for useful discussions.

Funding: JSR thanks funding from EU-7FP-BioPreDyn, JdlR from EU FP7-HEALTH-2007-B (ref. 223411), Spanish ISCiii (ref. PS09/00843), and Junta Castilla y Leon (ref. CSI07A09). FE was partially supported by the “Borsa Gini” scholarship, awarded by “Fondazione Aldo Gini”, Padova, Italy.

References

Aldridge, B. B., Burke, J. M., Lauffenburger, D. A., & Sorger, P. K. (2006). Physicochemical modelling of cell signalling pathways. *Nature cell biology*, *8*(11), 1195–1203.

Bansal, M., Belcastro, V., Ambesi-Impiombato, A., & di Bernardo, D. (2007). How to infer gene networks from expression profiles. *Molecular systems biology*, *3*, 78.

Bender, C., Heyde, S., Henjes, F., Wiemann, S., Korf, U., & Beissbarth, T. (2011). Inferring signalling networks from longitudinal data using sampling based approaches in the r-package ’ddepn’. *BMC bioinformatics*, *12*, 291.

Calzone, L., Tournier, L., Fourquet, S., Thieffry, D., Zhivotovsky, B., Barillot, E., & Zinovyev, A. (2010). Mathematical modelling of cell-fate decision in response to death receptor engagement. *PLoS computational biology*, *6*(3), e1000702.

Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., Schultz, N., Bader, G. D., & Sander, C. (2011). Pathway commons, a web resource for biological pathway data. *Nucleic acids research*, *39*(Database issue), D685–90.

Ciaccio, M. F., Wagner, J. P., Chuu, C. P., Lauffenburger, D. A., & Jones, R. B. (2010). Systems analysis of egf receptor signaling dynamics with microwestern arrays. *Nature methods*, *7*(2), 148–155.

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *Int. J. Complex Syst.*, *1695*.

Eduati, F., Corradin, A., Camillo, B. D., & Toffolo, G. (2010). A boolean approach to linear prediction for signaling network modeling. *PloS one*, *5*(9), e12789.

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

Glaab, E., Baudot, A., Krasnogor, N., & Valencia, A. (2010). Extending pathways and processes using molecular interaction networks to analyse cancer genome data. *BMC bioinformatics*, *11*, 597.

Ideker, T., & Sharan, R. (2008). Protein networks in disease. *Genome research*, *18*(4), 644–652.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G. R., Wu, G. R., Matthews, L., Lewis, S., Birney, E., & Stein, L. (2005). Reactome: a knowledgebase of biological pathways. *Nucleic acids research*, *33*(Database issue), D428–32.

Ma'ayan, A., Jenkins, S. L., Neves, S., Hasseldine, A., Grace, E., Dubin-Thaler, B., Eungdamrong, N. J., Weng, G., Ram, P. T., Rice, J. J., Kershenbaum, A., Stolovitzky, G. A., Blitzer, R. D., & Iyengar, R. (2005). Formation of regulatory patterns during signal propagation in a mammalian cellular network. *Science (New York, N.Y.)*, *309*(5737), 1078–1083.

Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Floreano, D., & Stolovitzky, G. (2010). Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(14), 6286–6291.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., & Califano, A. (2006). Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, *7 Suppl 1*, S7.

Markowetz, F. (2010). How to understand the cell by breaking it: network analysis of gene perturbation screens. *PLoS computational biology*, *6*(2), e1000655.

Meyer, P. E., Lafitte, F., & Bontempi, G. (2008). minet: A r/bioconductor package for inferring large transcriptional networks using mutual information. *BMC bioinformatics*, *9*, 461.

Morris, M. K., Saez-Rodriguez, J., Sorger, P. K., & Lauffenburger, D. A. (2010). Logic-based models for the analysis of cell signaling networks. *Biochemistry*, *49*(15), 3216–3224.

Mukherjee, S., & Speed, T. P. (2008). Network inference using informative priors. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(38), 14313–14318.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., & Kanehisa, M. (1999). Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, *27*(1), 29–34.

Pe'er, D. (2005). Bayesian network analysis of signaling networks: a primer. *Science's STKE : signal transduction knowledge environment*, *2005*(281), pl4.

Pico, A. R., Kelder, T., van Iersel, M. P., Hanspers, K., Conklin, B. R., & Evelo, C. (2008). Wikipathways: pathway editing for the people. *PLoS biology*, *6*(7), e184.

Pieroni, E., de la Fuente van Bentem, S., Mancosu, G., Capobianco, E., Hirt, H., & de la Fuente, A. (2008). Protein networking: insights into global functional organization of proteomes. *Proteomics*, *8*(4), 799–816.

Prieto, C., & Rivas, J. D. L. (2006). Apid: Agile protein interaction dataanalyzer. *Nucleic acids research*, *34*(Web Server issue), W298–302.

Prill, R. J., Saez-Rodriguez, J., Alexopoulos, L. G., Sorger, P. K., & Stolovitzky, G. (2011). Crowdsourcing network inference: the dream predictive signaling network challenge. *Science signaling*, *4*(189), mr7.

Rivas, J. D. L., & Fontanillo, C. (2010). Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS computational biology*, *6*(6), e1000807.

Saadatpour, A., Wang, R. S., Liao, A., Liu, X., Loughran, T. P., Albert, I., & Albert, R. (2011). Dynamical and structural analysis of a t cell survival network identifies novel candidate therapeutic targets for large granular lymphocyte leukemia. *PLoS computational biology*, *7*(11), e1002267.

Saez-Rodriguez, J., Alexopoulos, L. G., Epperlein, J., Samaga, R., Lauffenburger, D. A., Klamt, S., & Sorger, P. K. (2009). Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. *Molecular systems biology*, *5*, 331.

Samaga, R., Saez-Rodriguez, J., Alexopoulos, L. G., Sorger, P. K., & Klamt, S. (2009). The logic of egfr/erbb signaling: theoretical properties and analysis of high-throughput data. *PLoS computational biology*, *5*(8), e1000438.

Terfve, C., & Saez-Rodriguez, J. (2012). Modeling signaling networks using high-throughput phospho-proteomics. *Advances in Experimental Medicine and Biology*, *736*, 19–57.

Vinayagam, A., Stelzl, U., Foulle, R., Plassmann, S., Zenkner, M., Timm, J., Assmus, H. E., Andrade-Navarro, M. A., & Wanker, E. E. (2011). A directed protein interaction network for investigating intracellular signal transduction. *Science signaling*, *4*(189), rs8.

Watterson, S., Marshall, S., & Ghazal, P. (2008). Logic models of pathway biology. *Drug discovery today*, *13*(9-10), 447–456.

4

Quantitative modelling of small sub-networks

4.1 Introduction

The most common computational technique in systems biology is kinetic modelling of chemical reactions. Biochemical reactions are catalyzed by enzymes, which are highly specific and remain unchanged after reaction. For the mass-action low, reaction rate is proportional to probability of collision of reactants, thus to the concentration of reactants and to the power of molecularity. For the enzymatic reaction:

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} ES \overset{k_2}{\rightarrow} E + P \tag{4.1}$$

the ODE which describe the dynamic of the complex ES is:

$$\dot{ES} = k_1 \cdot E \cdot S - (k_{-1} + k_2) \cdot ES \tag{4.2}$$

In this way, the continuous time behavior of a biological system can be described by a set of differential equations where variables are concentrations of biochemical species, and parameters are the kinetic constants; the solution of these equations provides insights into the studied process. Kinetic models could, in principle, be used to describe any biological process. However, when it comes to the description of large biochemical networks, two main problems arise. The first is that the definition of kinetic models requires an advanced knowledge of all processes taking part into the described system, which is not always available. The second is that kinetic modelling implies the description of all biochemical species involved, and many parameters are required to describe a large system. In order to estimate these parameters a large amount of time-course experimental data are required; this is often a problem because data are available only for a subset of network components and there are rather large time intervals between successive measurements preventing from a complete description of the dynamic of the system (Lawrence et al., 2010). To overcome these problems, a coarse-grained description of the system can be adopted in order to reduce the system of differential equations: model reduction techniques are used to eliminate variables based on constrains (e.g. conservation equations, steady-state conditions), avoiding the explicit representation of the complete reaction network involving all possible molecular species (Feret et al., 2009). Under certain conditions, it is also possible to divide the system in smaller sub-systems of particular interest, and to analyze separately their dynamic behavior. In particular, network motifs are known to play important functional roles in many regulatory networks.

Network motifs

Biochemical networks are known to contain recurrent local patterns that are defined network motifs (Milo et al., 2002). They are considered to be basic building blocks of complex networks as they are characterized by typical dynamic behavior and they carry out specific information-processing functions (Alon, 2007). Some examples of recurrent motifs in transcriptional and signalling networks are shown in Figure 4.1. Autoregulation, for example, speeds-up the response if it has negative sign or has the opposite effect if positive, negative feedback loop contribute to stability and feed-forward loops can act as filters, sign-sensitive delay or pulse generators depending on the signs of interactions.

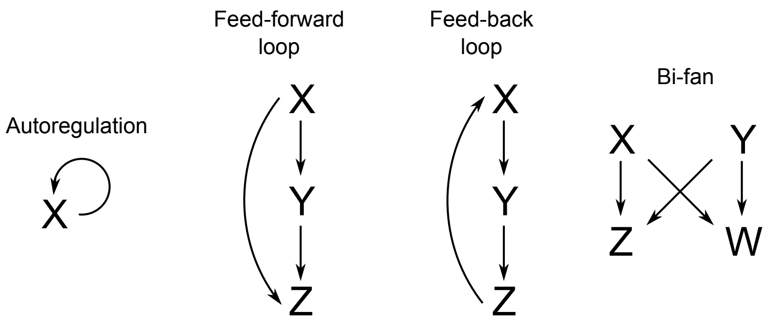


Figure 4.1: Examples of recurrent network motifs.

4.2 Feed-back loop to explain yeast stress response

It is well known that cells are able to rapidly respond to the change of environmental conditions adjusting their genomic expression. When yeast is stimulated with different types of stresses or stimulations most of its genes show a common expression pattern characterized by a quick transient peak followed by a return to a pre-stimulus level, namely the "Environmental Stress Response". In this work we propose a model of autoregulation that is able to explain the yeast environmental stress response reproducing all the features shown by experimental data. We use a negative integrative feedback, where the variable being integrated is the relative mRNA abundance and its integral is the corresponding protein. Gene specific parameters are estimated by fitting gene and protein expression time series and the resulting model is shown to reproduce the known experimental features (e.g. robust adaptation, graded and reciprocal response to stimulation).

The approach and results obtained are described in detail in De Palo et al. (2011); full text of the original paper is reported in Appendix 4.1.

4.3 miRNA mediated feed-forward loops

miRNAs are small non coding molecules that post-transcriptionally regulate gene expression: mature miRNAs bind to the 3' UTR of target mRNAs decreasing the frequency of translation and increasing mRNA degradation rate. They are known from the literature to be involved in one of the main recurrent regulatory circuits: the feed-forward loops (FFL) where a TF regulates a miRNA and they both regulate a target mRNA. In this work we postulate three slightly different models with reasonable hypothesis, and we use them to select FFLs that are active in a particular experimental condition, based on model identification criteria, namely: goodness of fit (whiteness and amplitude of the residuals), precision of the estimates and comparison with submodels with one missing regulatory link to verify that the whole topology is necessary to describe data. As a case study, the approach was applied to select plausibly active FFLs during human adipogenesis, from a large list of putative FFLs previously identified based on sequence analysis. Interesting biological results were provided among selected FFLs: some of them were positive controls, including miRNAs and TFs known from the literature to be regulators in adipogenesis, while others included potential novel players.

The approach and results obtained are described in detail in Eduati et al. (2012b); full text of the original paper is reported in Appendix 4.2.

Appendix 4.1 Paper: De Palo et al., IET Syst Biol, 2011

The following publication dealing with qualitative modelling of small subnetworks have been coauthored by the Ph.D. candidate during her doctoral program.

- G. De Palo, F. Eduati, M. Zampieri, B. Di Camillo, G. Toffolo, and C. Altafini. *Adaptation as a genome-wide autoregulatory principle in the stress response of yeast*. Systems Biology, IET, 5(4):269-279, 2011.

Full text of the original paper is reported in this Appendix.

Adaptation as a genome-wide autoregulatory principle in the stress response of yeast

G. De Palo[1], F. Eduati[2], M. Zampieri[1], B. Di Camillo[2], G. Toffolo[2] and C. Altafini[1,*]

[1]SISSA Int. School for Advanced Studies, Trieste, Italy
[2]Department of Information Engineering, University of Padova, Padova, Italy

[*]*Corresponding author:* altafini@sissa.it

Abstract

The gene expression response of yeast to various types of stresses/perturbations shows a common functional and dynamical pattern for the vast majority of genes, characterized by a quick transient peak (affecting primarily short genes) followed by a return to the pre-stimulus level. Kinetically, this process of adaptation following the transient excursion can be modeled using a genome-wide autoregulatory mechanism by means of which yeast aims at maintaining a preferential concentration in its mRNA levels. The resulting feedback system explains well the different time constants observable in the transient response, while being in agreement with all the known experimental dynamical features. For example it suggests that a very rapid transient can be induced also by a slowly varying concentration of the gene products.

1 Introduction

Typically, at the level of gene expression, the response to a stimulus, or to a change in some environmental condition, or even to the substrate composition can be decomposed into a rapid adaptation phase, occurring with a typical time constant of the order of the tens of minutes (Causton et al., 2001; Foat et al., 2005; Gasch et al., 2000; Levy et al., 2007; Ronen & Botstein, 2006), superimposed to a long term permanent modification of the gene expression steady state (occurring, for example, at the diauxic shift, (DeRisi et al., 1997)). For *S.cerevisiae*, the rapid adaptation described in (Gasch et al., 2000; Chechik et al., 2008; Yoshimoto et al., 2002; Causton et al., 2001; Ronen & Botstein, 2006) consists essentially of a transient change in the mRNA concentration followed by a return to the basal pre-stimulus level for almost the entire population of genes. This massive adaptation phenomenon is observed in response to both temporary (such as the glucose pulses of (Ronen & Botstein, 2006)) and permanent (such as the environmental stresses of (Gasch et al., 2000; Yoshimoto et al., 2002; Causton et al., 2001)) stimuli. A simple correlation analysis reveals that the responses to different types of stimuli have consistent similarities and are strongly correlated with the Half Life (HL) of the corresponding genes, as shown in Fig. 1 (a).

The aim of this paper is to propose a kinetic model for the genome-wide rapid adaptation of the transient response to stimuli able to explain the following features:

(a) the response is highly stereotypical across many stimulations (Gasch, 2002);

(b) the response is to some extent graded (i.e., proportional to the magnitude of the stimulus, see e.g. the heat shock responses of (Gasch et al., 2000)), and reciprocal (a temperature upshift induces a transient pattern which is roughly similar, except for the sign, to that of a temperature downshift) (Gasch et al., 2000);

(c) robust adaptation is observed to stimuli of various “order” i.e., both vanishing and persistent stimuli are reabsorbed;

(d) whenever for a gene adaptation is not perfect, the new steady state reached has the same sign of the peak of the transient excursion.

(e) the transcriptional transient seems to be faster than the changes in other cellular quantities (such as growth rate (Levy et al., 2007) or protein concentration (Chechik et al., 2008));

(f) the rise time of the transient peak is shorter than its decay time (i.e., the time needed to return to the pre-stimulus level);

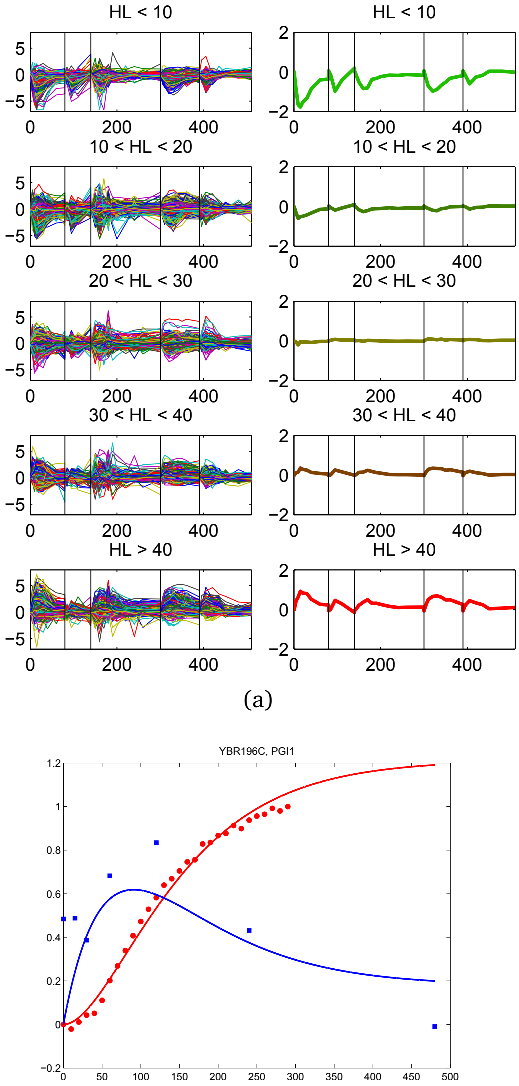
(g) the decay time constant is roughly of the same magnitude as the degradation time (HL) of the genes inferred from experimental data (Grigull et al., 2004; Kuai et al., 2005; Wang et al., 2002) although, if transcription is blocked after the onset of the stress response, the upregulated genes seem to degrade faster than expected, while repressed genes seem to degrade slower than expected (Shalem et al., 2008);

(h) the transient response typically does not induce oscillations (noticeable above the noise level);

(i) both the maximal amplitude of the transient peak and the area under the transient response are roughly proportional to the HL of the genes, while the peaking time (i.e., instant at which a gene has its maximal excursion during the transient) is not significantly correlated with HL, as shown in Fig. 1 (a);

(j) as described below, a large transient excursion is induced primarily on short genes;

(k) for what is known, changes in the transcriptional response are maintained at the level of translation (Preiss et al., 2003; Chechik et al., 2008).



(a)

(b)

Figure 1: In (a) five time series from (Gasch et al., 2000) showing the "step response" to different environmental stresses are plotted. For visualization purposes, the 5 series are shown sequentially one after the other. The time axis is in minutes, the relative mRNA abundance is in log_2 basis. In the left panels, the 5153 genes are clustered in five groups according to the respective HL (in min). In the right panels the averages of the profiles in each cluster are computed. In all responses the trend followed during the transient is highly correlated, i.e. genes with similar HL behave similarly in the various responses. In particular, genes with short HL tend to be downregulated while genes with longer HL upregulated. In (b) the linear model (2) is fitted to the experimental time series of a pair gene/protein in response to treatement with DTT (data from (Chechik et al., 2008), scales are normalized). For this class of redox stimulations, the time constants of the response are higher than in (a). Nevertheless, as predicted by the model, the protein time series resembles the integral of the gene time series, which corresponds to the adaptation scenario by means of integral feedback autoregulation. Parameter values estimated by fitting model (2) to these data, as well as to those of the other eight genes, are listed in Table 1.

Adaptation, intended as the mechanism by means of which a biological system is able to recover the "optimal" working level of a variable in spite of a persistent stimulus, is common to many biological systems. Examples are numerous: various signal transduction pathways (Behar et al., 2007), bacterial chemotaxis (Alon et al., 1999), sensory transduction (Torre et al., 1995). This property is typically characterized by means of a negative feedback loop. In the present context, as the gene expression returns to its pre-stimulus level regardless of the amplitude of the stimulus, the system has to have an encoded robust regulatory mechanism as well as a memory of the nominal pre-stimulus concentration value for each gene.

Following (Yi et al., 2000), a control-theoretic interpretation of adaptation involves an integral feedback loop. In this scheme, the integral of the displacement from a nominal level (i.e., of the error) of a variable is fed back with negative sign. Adaptation is achieved as this variable returns to the nominal level (that is as the error tends to zero) in spite of a persistent stimulus that, in absence of feedback, would alter the steady state value. In the context of the present paper, the variable being integrated is the relative mRNA abundance, and its integral may be taken to represent the relative abundance of the corresponding "gene product". We assume here that this quantity acts homeostatically on the mRNA transcription rate, reequilibrating the gene expression to the nominal level of concentration. Negative autoregulation of transcription or "autogenous control" (Savageau, 1974; Goldberger, 1974) is a mechanisms that allows to reduce fluctuations around the steady state (Becskei & Serrano, 2000) and to decrease the rise time of a response (Rosenfeld et al., 2002), although it is often invoked for specific transcription factors repressing their own transcription. In our case, we shall assume that negative autoregulation works as a general ubiquitous homeostatic principle, opposing (permanent) changes into the mRNA levels with respect to an "optimal" working concentration for the transcripts, and affecting preferentially the short ORF for which protein levels seem to fluctuate more. In the out of equilibrium scenario represented by the transient response, the autoregulatory action is meant to represent homeostasis in both the synthesis and the degradation components of the rate law. As such, the feedback action can modulate the "effective" degradation rate in presence of transcriptional blockage, coherently with data reported in (Shalem et al., 2008).

Autoregulation, in practice, couples the dynamics of a gene and its product. On the dynamical model, this coupling results into a second time constant, which for each gene can be used in the description of the transient response. The aim of this paper is to show that using a negative integral feedback to describe this coupling can explain not only adaptation but also the very rapid transcriptional response to changes in other slower cellular variables (such as the gene products). As a result, even a simple linear ODE model can reproduce all the features listed above, provided that the modes (eigenvalues) of the system and the sign of the input response are appropriately chosen.

2 Results

The experiments here analyzed (a list of the time series used is in the Supplementary Notes) consist of two-channel microarrays in which the mRNA abundance during the transient is hybridized against a basal pre-stimulus mRNA concentration, so that a value approaching 1 (or 0 if a log scale is considered, as in Fig. 1) corresponds to a return to the pre-existing steady state. As can be seen from Fig. 1 (a), in each of the 5 time series of (Gasch et al., 2000) (chosen among those providing a sufficiently fast kinetics, see Supplementary Notes for details), we observe that for almost 90% of the genes, the relative expression ranges within $[-\log_2(1.5),\, \log_2(1.5)]$ at the end of each transient (the percentage goes up to 95% if we consider an interval of $[-1,\, 1]$), while during the transient only $\sim\, 50\%$ of the genes remain inside the interval $[-1,\, 1]$ on each time series. Hence we can assume that globally the system undergoes a transient excursion in response to each stimulus, and that such an excursion is reabsorbed in a time scale of the order of the hour, meaning that the system has adapted in spite of a persistent stimulation.

Expanding on the concept of autogenous control (Savageau, 1974; Goldberger, 1974), the basic assumption underlying our model is that an increase of the abundance of a certain protein well above (resp. below) the normal “working” level disfavors (resp. favors) the transcription of the corresponding gene. Under such negative autoregulation, a basic model for transcription and translational kinetics (Hargrove & Schmidt, 1989; Savageau, 1974; Simpson et al., 2003; Rosenfeld et al., 2002), derived in Material and Methods, is the following:

$$\begin{aligned} \frac{dm_i}{dt} &= -\delta_i(m_i-1) - a_i(p_i-1) + b_i u \\ \frac{dp_i}{dt} &= -\lambda_i(p_i-1) + r(m_i-1), \end{aligned} \tag{1}$$

where: $m_i = \frac{[\text{mRNA}]_i^{\text{red}}}{[\text{mRNA}]_i^{\text{green}}}$ is the mRNA concentration of the i-th gene relative to the basal level, equal to the ratio between the “red” and “green” channels associated to the stimulus response and the basal mRNA level, respectively; p_i is the relative concentration of the corresponding gene product; δ_i and λ_i are degradation rate constants; a_i is the strength of the negative feedback; r is a translational rate constant, assumed equal for all genes. The expressions $m_i - 1$ and $p_i - 1$ are meant to represent displacements from the basal levels in response to a stimulus u whose amplitude and sign (i.e., role as activator or repressor of m_i) are given by b_i. Model (1) assumes linear kinetics, but this does not affect the qualitative conclusions of the study, as shown in Supplementary Materials. Moreover, model (1) does not account for translational delays, e.g. due to the export and localization of the mRNA, and/or the limited rates of translation initiation and peptidic chain elongation (see Supplementary material for a model with delay).

For the limited time horizon considered here, a couple of hours, the model can be further simplified. In fact, unlike the gene degradation rate δ_i, for which knowledge from several genome-wide datasets is available in the literature (Grigull et al., 2004;

Kuai et al., 2005; Wang et al., 2002), knowledge of protein degradation rate λ_i is quite limited at a proteome-wide scale (Belle et al., 2006), but it is commonly accepted (Hargrove & Schmidt, 1989; Belle et al., 2006) that the protein degradation dynamics are slower (or much slower) than the corresponding mRNA dynamics. For the stress response of yeast we have quantified this difference using the recent data of (Chechik et al., 2008) consisting of measurements of nine genes and their corresponding proteins in response to redox stress (DTT). The dynamical response to this type of stimulation is known to be slower than for example the heat shock response (Gasch et al., 2000; Chechik & Koller, 2009). However, it is plausible to assume that the ratios δ_i/λ_i are similarly related. By fitting model (1) on these data, all model parameters where precisely estimated (Table 1). In particular it turns out that on average the ratio δ_i/λ_i is 20: while the genes have a HL ($= \ln(2)/\delta_i$) of approximately 35 min, in line with the known degradation time constants (Grigull et al., 2004; Kuai et al., 2005; Wang et al., 2002), the HL of the corresponding proteins is more than 11 hours. This means that in the time horizon considered here (a couple of hours), the contribution of λ_i is totally irrelevant. To confirm this observation, it is worth observing that the protein time series of (Chechik et al., 2008) show a growing front which is much slower than the corresponding genes, and no decline (i.e., the “transient” growth is not yet exhausted at the end of the recorded time series), see Fig. 1 (b). Thus, the model equations become:

$$\begin{aligned}\frac{dm_i}{dt} &= -\delta_i(m_i - 1) - a_i(p_i - 1) + b_i u \\ \frac{dp_i}{dt} &= r(m_i - 1).\end{aligned} \tag{2}$$

The second equation of (2) can be integrated and the resulting integral, i.e., the area under the mRNA profile, represents an estimate of the protein abundance $p_i(t)$. When

Table 1: Gene and protein parameters estimated by fitting model (1) to the time series of (Chechik et al., 2008). In this case, the translational rate constant r is also considered a gene specific parameter. The average value obtained for r corresponds to our choice in the rest of the paper.

ORF	a_i	b_i	δ_i	λ_i	r
YBR001C	0.008	0.011	0.016	0.000	0.008
YBR196C	0.012	0.019	0.023	0.001	0.010
YDR074W	0.005	0.010	0.013	0.000	0.008
YDR261C	0.001	0.004	0.006	0.000	0.011
YER003C	0.013	0.019	0.023	0.001	0.009
YGL253W	0.013	0.016	0.023	0.001	0.010
YHR163W	0.005	0.012	0.033	0.000	0.013
YKL127W	0.009	0.012	0.017	0.000	0.008
YNL241C	0.010	0.016	0.020	0.001	0.009
mean	0.008	0.013	0.020	0.001	0.010

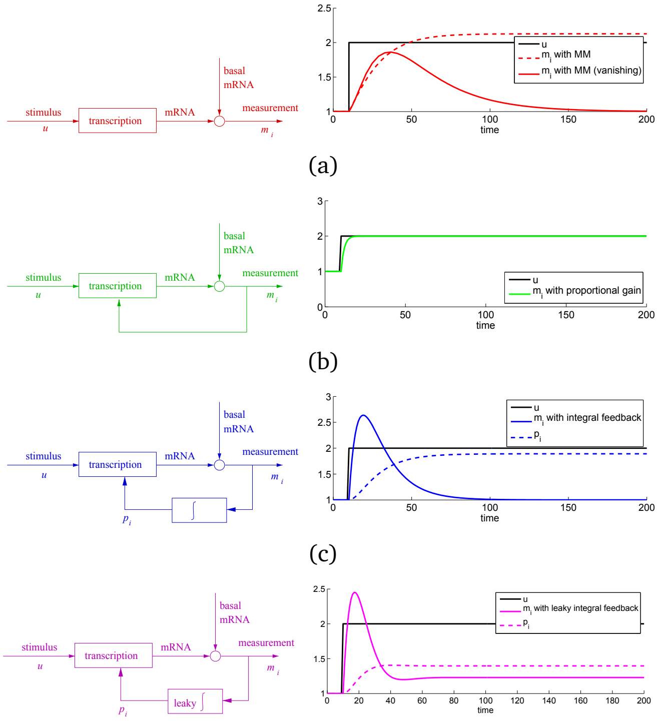


Figure 2: Four different schemes for the step response in (4). (a): open-loop scheme with nonvanishing/vanishing transcriptional synthesis term $f(\cdot)$. When $f(w(u)) = f(w)u$ with $f(w)$ a Michaelis-Menten (MM) function of the transcription factors w, see Supplementary Notes for details, then the dotted red line is obtained for m_i. When instead $f(w(u))$ is a vanishing function then the response also vanishes (shown in red, solid). This model is equivalent to the so-called impulse model of (Chechik & Koller, 2009). However, if we think in absolute (rather than relative) terms, it entails an exact knowledge of the nominal level (hence, implicitely, a form of memory, like the one obtained here by means of integral feedback). Notice further how for a synthesis term f which is zero-order in m_i, the rising front has a limited slope regardless of the form of f (see Supplementary Notes for a more detailed analysis of open-loop time constants). (b): a regulation scheme with a proportional feedback, i.e., a feedback directly on the relative mRNA abundance. Adaptation is not achieved, rather, the mRNA level tends to "track" exactly the amplitude of the stimulus u. (c): regulation with integral feedback. Adaptation is achieved for any value of nominal concentration. Both feedback schemes (b) and (c) decrease the rise time of the response. In (c) this is achieved via the much slower dynamical variable p_i. (d): "Quasi-adaptation" in presence of protein degradation terms. When a protein degradation term is added to the equations as in (1), then perfect adaptation is lost. However, for reasonable values of protein degradation rates, the new steady state is still close enough to the full recovery of the pre-stimulus level and the shape of the transient is essentially unchanged. Hence we can talk about "quasi-adaptation". Notice how in this case the new steady state reached has the same sign of the transient excursion.

fed back in the first equation of (2) with negative sign, it has the effect of achieving perfect adaptation in m_i, i.e. the mRNA abundance returns exactly to its basal level in spite of a persistent stimulus u, (Fig. 2). A second effect of the negative feedback is to speed up the transient response (Rosenfeld et al., 2002). In order to understand how this is achieved, consider the state and input matrices A_i and B_i of the linear system (2) (see Material and Methods for details). A_i has two eigenvalues $s_{i,1}$ and $s_{i,2}$, meaning that the effect of coupling the gene m_i with the protein p_i is to introduce a second dynamical mode into the evolution of the system. These two eigenvalues are always stable and, if they are chosen real so that damped oscillations are excluded, the explicit solution for (2) to a step stimulus u is:

$$\begin{aligned} m_i(t) &= 1 + \frac{b_i}{\gamma_i}\left(e^{s_{i,2}t} - e^{s_{i,1}t}\right) \\ p_i(t) &= 1 + \frac{b_i r}{\gamma_i}\left(\frac{e^{s_{i,2}t}-1}{s_{i,2}} - \frac{e^{s_{i,1}t}-1}{s_{i,1}}\right). \end{aligned} \tag{3}$$

where $\gamma_i = \sqrt{\delta_i^2 - 4ra_i}$ (see Materials and methods). The two eigenvalues give rise to two exponential modes: if $s_{i,2}$ is of the same order as the natural mRNA degradation time constant ($s_{i,2} \sim \ln(2)/\mathrm{HL}_i$), and $s_{i,1} < s_{i,2}$, then, from the first equation in (3), the fast mode $s_{i,1}$ induces a sharp rising front in the transient but is rapidly exhausted, and is then followed by a more gentle decay to the pre-stimulus level which resembles a typical first order degradation, governed by the slow mode $s_{i,2}$ which is more long-lived. This is exemplified in Fig. 3 for a specific stress-inhibited gene and then extended to all genes: the difference in the two eigenvalues $s_{i,1}$ and $s_{i,2}$ induces a transient response which well reproduces the observed time courses.

In the case of real eigenvalues, the lack of oscillatory behavior implies that $p_i(t)$ is typically monotonic and of sign equal to that of the m_i transient (an effect similar to the “potentiation” described in (Preiss et al., 2003)). Coherently with the experimental data of Fig. 1 (b), the dynamics for p_i are much slower than those of the corresponding m_i (Fig. 3, left). Nevertheless, this slow dynamics is crucial to speed up the transient response of m_i, see Supplementary notes for a further discussion on open-loop time constants. It is worth noticing that the embedding of a fast regulation loop into a slower one is a universal rule of thumb of an engineering control design requiring nested loops, because it minimizes the cross talk between the two loops and therefore also the possibility of spurious dynamical behaviors. Values of eigenvalues shown in Fig. 4 (top left panel) indicate that the fastest mode of A_i (dominating the rising front of the transient) is always much more negative than the slowest mode (dominating the decaying front): $s_{i,1} \ll s_{i,2} < 0$. For all 5 responses of Fig. 1 (a), the time at which the transient gene expression peaks, t_{peak}, is approximately 25 min. In (3), if $t_{\text{peak}} \sim 25\text{min}$, then $e^{s_{i,1}t_{\text{peak}}} < 0.2$ for 87% of the genes (while $e^{s_{i,2}t_{\text{peak}}} < 0.2$ for only 7%) meaning that indeed the transient response declines due to the exhaustion of the fast mode. Since the mean HL is $\approx 25 \pm 15$ min, in the time horizon of the 5 series ($t_{\text{end}} = 80, 60, 160, 90, 120$ min for the 5 time courses), the transient has sufficient time to decay back at almost basal level for most genes. In order to evaluate the effect of different stimuli on a single gene, it is of interest to compare the sign of the parameter b_i across the five time series. Our results (Fig. 4 (top right panel)) say that for at least 50% of the genes the sign assignement is unanimous in the 5 series.

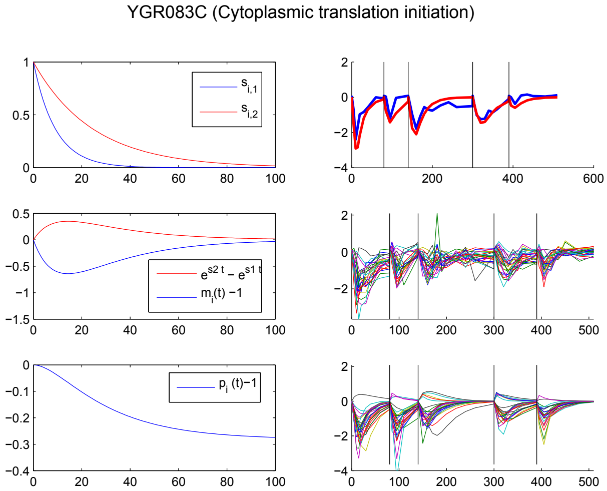


Figure 3: Time response of the system (2) to a step-like input for the category of cytoplasmatic transcription initiation genes and for gene GCD2/YGR083C (subunit of the translation initiation factor eIF2B) in particular. The two modes have different real parts ($s_{i,1} = -0.11 < s_{i,2} = -0.04$), thus their difference typically shows a profile like that reproduced in the middle left plot. The sign of b_i then determines whether the gene is classified as up- or down-regulated by the stimulus (still middle plot). The area under the $m_i(t)$ time course, proportional to the gene product $p_i(t)$ shown in the bottom left plot, is monotonically growing with a much slower time constant, as expected. For the gene considered here, the experimental and reconstructed profiles are shown in the top right panel (blue and red respectively, both in $\log_2$ scale) while profiles and model-based predictions of the entire category of cytoplasmic transcription initiation genes are shown in the middle and bottom plots of the right column, respectively.

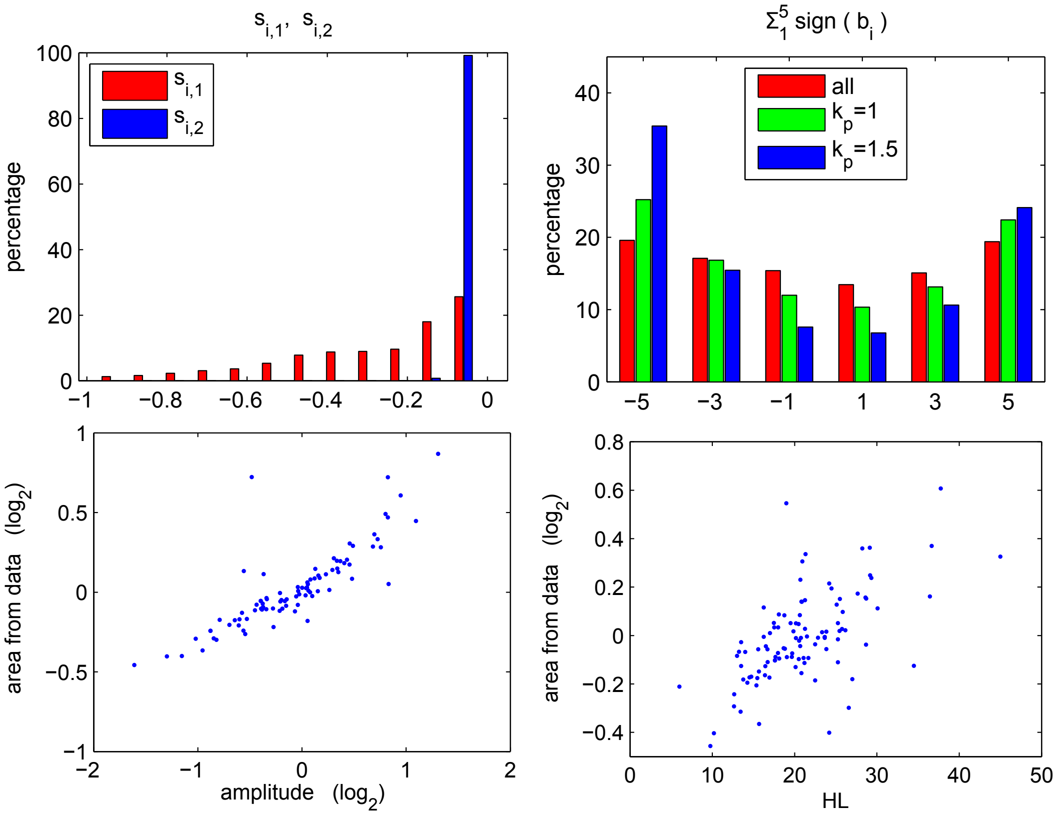


Figure 4: Relationship between eigenvalues $s_{i,1}$ and $s_{i,2}$ for all genes in the five time series of Fig. 1 (top left panel). Sum of the signs of the forcing term b_i (top right panel) in the five time series for all genes and for the most perturbed genes, selected according to a threshold k_p, i.e. satisfying $\max |log_2(m_i(t))| > k_p$. Notice that, choosing $k_p = 1$, for at least 50% of the genes the sign assignement is unanimous in the 5 series of (Gasch et al., 2000) (more than 60% for $k_p = 1.5$). In the bottom row, for the PC complexes of Fig. 5 the area under the mRNA response measured on the data is compared with the corresponding maximal signed amplitude observed during the transient (left) and with the HL (right). The high agreement between area (i.e., $\log_2(p_i(t_{\text{end}}))$) and the sign of the peak of mRNA during the transient is confirming that most transient excursions are not oscillatory.

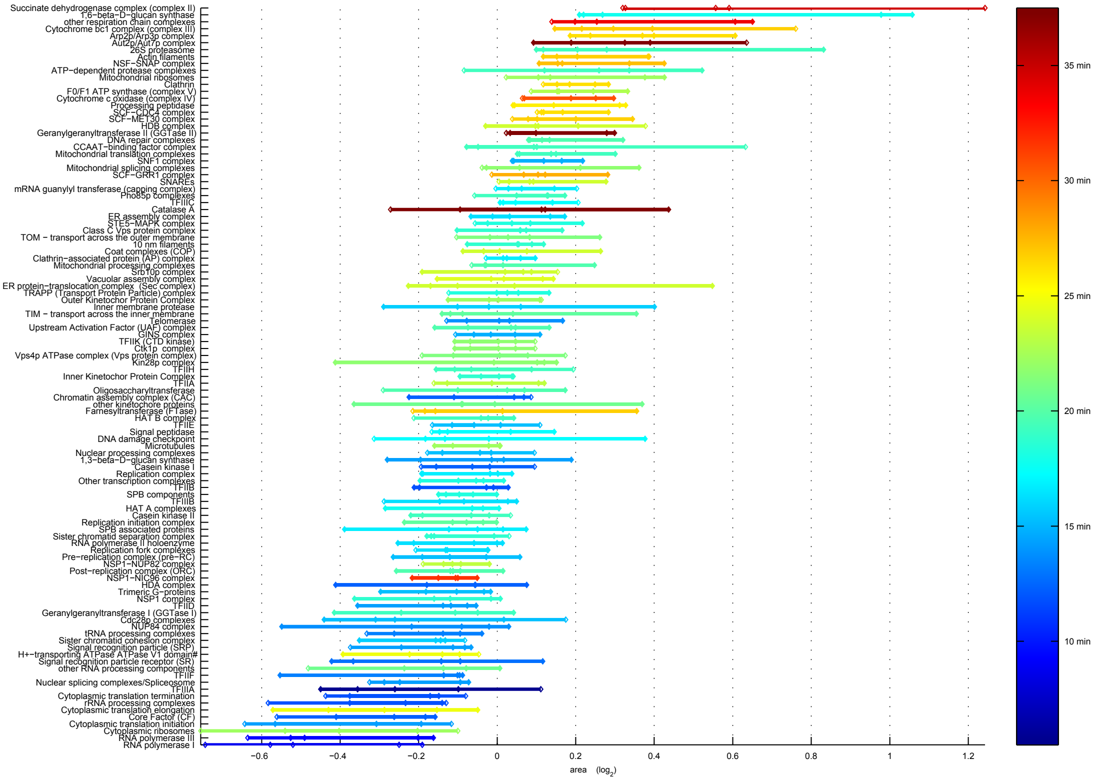


Figure 5: List of significant PC and corresponding areas for the 5 time series of Fig. 1 (a). The solid markers, representing the values in the 5 experiments, indicate that for most of the neatly up or down-regulated categories, the 5 values have identical signs. The color scale represents the value of HL associated to the protein complex: blue means short HL (minimum HL is 6 min) and red long HL ($\geqslant 40$ min).

The model permits to predict the value of gene product p_i at the end of the observation period t_{end}, from the area under gene expression m_i. To reduce the effect of noise in the mRNA time series, it is convenient to lump together genes whose products form a protein complex (PC). In fact these genes are known to have similar dynamics (Wang et al., 2002), observation largely confirmed by our analysis. In Fig. 5 areas for the 5 time series of Fig. 1 are shown: it turns out that for most of the neatly up or down-regulated PCs the 5 values have identical sign. A good agreement is also found across different data sets (Fig. 6) still showing correlation in the PCs and KEGG pathways areas. The comparison between numerically computed areas and model-based estimated areas (p_i) is shown in Fig. 6 (values are averaged over all genes forming the complex). Panels in the main diagonal show the correlation between computed and estimated areas for all datasets which is good, as expected. Not only a strong degree of correlation in the responses to various inhibitory stimuli (such as thermal, oxidative, osmotic, acid stresses) can be reproduced by the model, but also the anticorrelation between responses to inhibitory and excitatory stimuli (such as the reciprocal stresses discussed in (Gasch et al.,

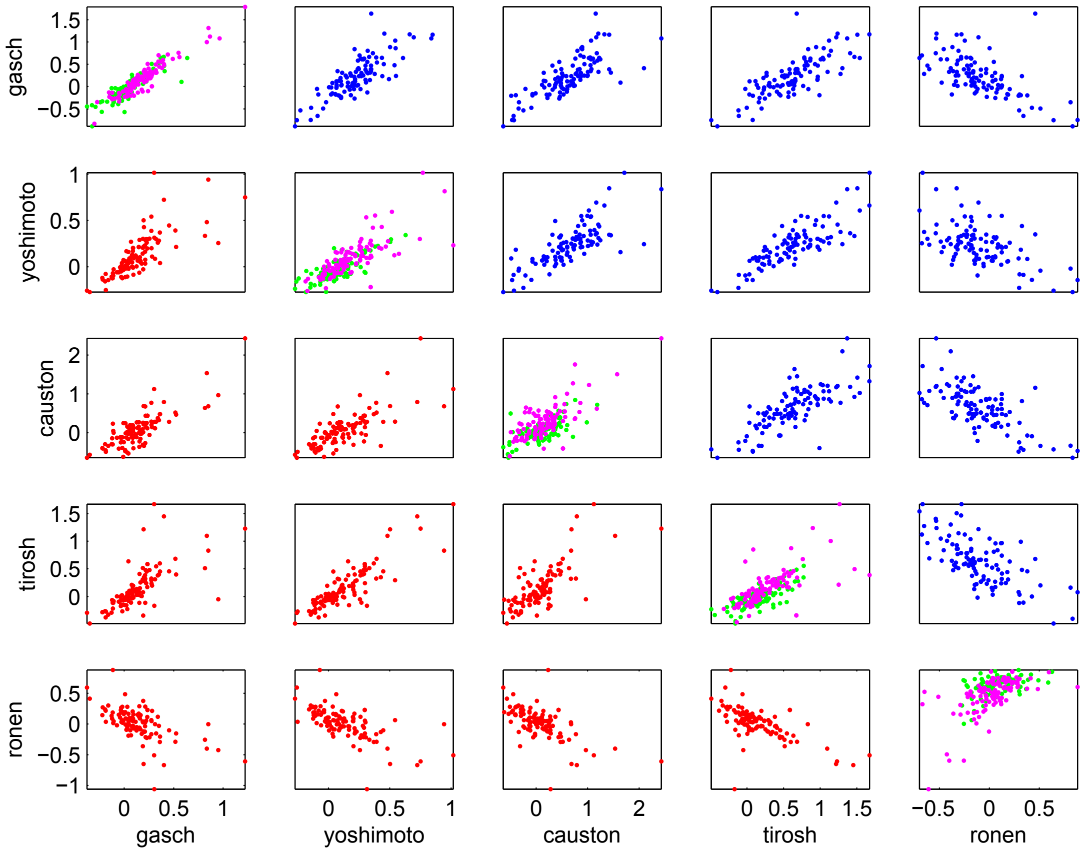


Figure 6: Comparison of the average area under the curve for the time series of (Gasch et al., 2000) (labeled "gasch"), (Yoshimoto et al., 2002) ("yoshimoto"), (Causton et al., 2001) ("causton"), (Tirosh et al., 2006) ("tirosh") and (Ronen & Botstein, 2006) ('ronen"). In the upper triangular part, in blue, the scatter plots represent the area of one set of data against any of the other sets for the PC complexes of Fig. 5. In the bottom triangular part, in red, the scatter plots are for the KEGG pathways represented in Fig. S2. While "gasch", "yoshimoto", "causton" and "tirosh" are inhibitory stimuli (stresses), "ronen" are activatory pulses of nutrient. Hence the antidiagonal pattern in the areas shown in the bottom row and column. In the diagonal plots, the area predicted by the model (with parameters tuned on the 5 series of (Gasch et al., 2000)) is shown against the corresponding measured area for PC (in green) and for the KEGG pathways (in magenta). In the simulation of the activatory stimuli of (Ronen & Botstein, 2006), the signs of the b_i are exchanged. This, together with the anticorrelated plots of the bottom row and column, validates the reciprocity property already observed in (Gasch et al., 2000) for some classes of stresses.

2000) or the nutrient inputs of (Ronen & Botstein, 2006)), see last row and column of Fig. 6.

The basic autoregolation mechanisms discussed so far neglects any gene-gene regulatory mechanism beyond co-participation in a protein complex or metablic pathway. A popular example of such regulatory mechanisms is the causal relationship between a transcription factor and its target genes. Having only mRNA profiles available, the only statistical test we can perform to evaluate this type of regulation is a significance analysis of the corresponding correlation coefficients on the transcriptional regulatory map of (Luscombe et al., 2004; Pilpel et al., 2001). For all of our time series, however, the correlation between transcription factors and corresponding target genes is always insignificant (Z-score test), see Fig. S2, Table S2 and the Supplementary Notes for more details. This suggests that the transient excursion must be triggered by post-transcriptional or post-translational modifications of the transcription factors which for the time being are largely unknown. However, even when we look at genes co-transcribed by the same transcription factor, we obtain that the correlation is still comparable to that of a random choice of genes, unlike for example the co-participation in a PC, see again Fig. S2. The approach of modeling the stress response in an “open-loop” fashion as the causal action of the transcription factors on their target genes is pursued for example in (Pilpel et al., 2001). It has several drawbacks, like for example that the map of transcription factor-target genes, though largely incomplete, is already combinatorially complex, condition-specific (Luscombe et al., 2004; Pilpel et al., 2001; Gasch, 2002), and the sign of the interactions (activator/repressor) is often unavailable. More importantly for us, this open-loop approach is unable to satisfy all of the kinetic constraints on the time series, like the sharpness of the rise front of the transient, and does not provide an explanation for the adaptation observed for which a form of feedback is required.

3 Discussion

In this work we have proposed a kinetic model aimed at describing different features characteristic of yeast transient response to stimuli, as listed in the Introduction. Here we discuss these features versus model behaviour in details.

3.1 Stereotypical response ((a) and (b))

Analysis of the 5 time series of Fig. 1 confirms that the similarities in their pattern are much more abundant than the stimulus-specific differences. This can be deduced from the sign concordances of the b_i values, see Fig. 4 (top right panel). If the 5 time series are compared with others from (Yoshimoto et al., 2002; Causton et al., 2001; Tirosh et al., 2006) also representing responses to prolonged stimuli, the pattern of up/down regulation is very similar. For the same PC as in Fig. 5, the comparison of average areas is shown in Fig. 6. In (Ronen & Botstein, 2006) instead, yeast is fed with pulses of glucose of different magnitude. In this case the sign of the responses is reciprocal for most genes, as can be seen in the scatter plots in the last row and column of Fig. 6

(see Supplementary Material). Sorting the PC complexes by the corresponding $p_i(t_{\text{end}})$, see Fig. 5, reveals that the downregulated categories (negative areas in the log scale of Fig. 5) are essentially all involved into transcriptional and translational processes, while in the most upregulated categories are respiratory metabolism and proteolysis. Notice that, coherently, also the ribosomal biogenesis is very different between the cytoplasmic and mitochondrial compartments. Analogous results are obtained for the KEGG pathways, see Fig. S2. Grouping the genes further according to KEGG hierarchy, we obtain the 15 macrocategories shown in Fig. S3 which give a general overview of the environmental stress response strategy in agreement with e.g. (Gasch et al., 2000; Gasch, 2002). This consists in a reduction of the energy-consuming (cytoplasmic) ribosomal biosynthesis and RNA processing machinery in favor of energy-producing components such as the respiratory chain complexes and the mitochondrial compartment in general.

Notice on Fig. 5 and S3 the correlation with the empirical values of HL assignable to these categories. It is worth observing how the ordering of the categories found here resembles e.g. the ordering of the phases of the peaks in the so-called yeast metabolic cycle (Tu et al., 2005), suggesting the unfolding of a common gene expression program. See Fig. S4 and (Soranzo et al., 2009) for more details.

3.2 Robust adaptation ((c) and (d))

While a few genes required for reacting to a specific cellular stress might show a permanent change in the gene expression steady state (Gasch, 2002), the vast majority of genes returns to their basal pre-stimulus level. It is for this category that we talk about adaptation. The rapid and massive transient excursion characterizing stress responses might help in activating immediate cellular reaction mechanisms (such as redistribution of energetic resources) while adaptation might be a mean to resume a mode of action as close as possible to optimal in spite of permanent environmental changes. The slower degradation rates for the gene products observed in real data (Chechik et al., 2008) and confirmed by our observation (Table 1) guarantee that the proteins are long-lived and that the modified cellular response is sustainable for a long period. As sketched in Fig. 2 (d), when the protein degradation is not neglected, its effect on the model is to alter the steady state value of m_i. The sign of this modification agrees with that of the transient excursion.

3.3 Fast transcriptional response from slow feedback ((e) and (f))

The transcriptional response to stresses can be activated directly by the external perturbation through signaling mechanisms in an essentially open-loop fashion, or through changes in the cellular state (e.g. amount of biomass or energy or metabolite composition) that induce feedback reactions (Levy et al., 2007). Some of these “internal variables”, such as growth rate, have been shown to happen at a slower pace than the transcriptional response (Levy et al., 2007). One of the characteristics of our model is that slower dynamics are instrumental in inducing the fast transcriptional response, provided that there is coupling between quantities as happens in presence of feedback.

While a feedback coupling between genes (faster) and gene products (slower) can optimize the stress recovery by speeding up the system response, the signs of the transcriptional transient excursions are instead directly correlated with the external perturbation and determine the strategy of the cellular response. Once the signature of the transient is identified, any cellular variable (for example the already mentioned growth rate) could in principle be used for the integral feedback in place of the gene products we use here. The fact that most genes show a synchronized peaking time for their transient may be a sign that a coordinated cellular mechanism is indeeed responsible for the feedback. Experimental data for cellular quantities such as the growth rate are however too few and difficult to obtain (Airoldi et al., 2009). See Supplementary Notes for more details and examples.

3.4 Stress response with transcriptional blockage ((g))

In (Shalem et al., 2008), the stress response is studied in conjunction with a blockage of transcription (delayed in time with respect to the begin of the stressful stimulation). The mRNA profiling reveals that genes activated during the transient response seem to be destabilized when the stress is followed by the transcriptional arrest (i.e., the genes seem to degrade faster than expected by the known HL values) and, viceversa, genes that are repressed in the transient seem to be stabilized by the combination stress + transcriptional blockage. In our model, we assume that the transcriptional arrest occurs at the peaking time of the transient, t_{peak}. Blocking transcription means putting $c = 0$ in (5) and $b_i = 0$ in (2), i.e., the ODE for the system reduces to

$$\begin{aligned} \frac{dm_i}{dt} &= -\delta_i m_i - a_i(p_i - 1) \\ \frac{dp_i}{dt} &= r(m_i - 1). \end{aligned}$$

In our scheme, the negative autoregulatory feedback term has a dual role, influencing both the synthesis and the degradation rate, and predicts correctly the altered degradation rates in the perturbed system of (Shalem et al., 2008). Consider first the case of an upregulated gene. For it $m_i(t_{\text{peak}}) > 1$ and, from (2), $p_i(t_{\text{peak}}) > 1$, implying $-a_i(p_i(t_{\text{peak}}) - 1) < 0$. Hence we have

$$\underbrace{-\delta_i m_i - a_i(p_i - 1)}_{\text{degradation with negative autoregulation}} \quad < \quad \underbrace{-\delta_i m_i}_{\text{reference degradation}} \quad < 0$$

i.e., when p_i is different from the nominal concentration ($p_i \neq 1$) the rate dm_i/dt is more negative than expected, meaning that the upregulated gene is destabilized. On the contrary, for a repressed gene we have $0 < m_i(t_{\text{peak}}) < 1$ and $0 < p_i(t_{\text{peak}}) < 1$, implying $-a_i(p_i(t_{\text{peak}}) - 1) > 0$, which leads to

$$\underbrace{-\delta_i m_i}_{\text{reference degradation}} \quad < \quad \underbrace{-\delta_i m_i - a_i(p_i - 1)}_{\text{degradation with negative autoregulation}} \quad < 0.$$

In this case, the repressed gene is stabilized by the autoregulation.

3.5 Proportionality of transient peak amplitude and area of m_i with HL ((i))

From (3) we can obtain the following expression:

$$p_i(t_{\text{end}}) - 1 \simeq \frac{b_i r}{\gamma_i}\left(\frac{1}{s_{i,1}} - \frac{1}{s_{i,2}}\right) \simeq -\frac{b_i r}{\gamma_i s_{i,2}} = \frac{b_i r}{\gamma_i}\text{HL}_i.$$

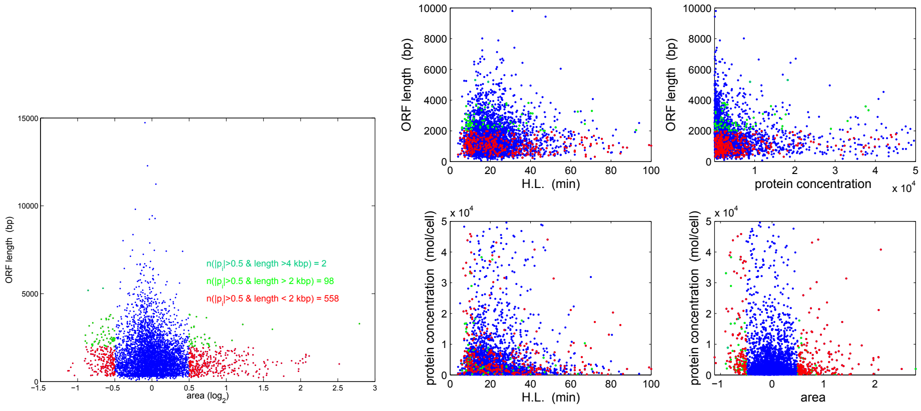
This expression provides an explanation in terms of the model (2) of the roughly direct proportionality observed between the area and the values of HL, shown in Fig. 4 (bottom right panel) for the PCs of Fig. 5. Notice on the same Fig. 4 (bottom left panel) the tight relationship between the amplitude (i.e., the signed peak in m_i) and the area (computed either via the model or from the data). From (3), model-based area and amplitude share the same gene-specific multiplicative constant b_i/γ_i.

3.6 Transiently perturbed genes are short ((j))

The transient excursion is induced (for both upregulated and repressed genes) to a very large extent on short genes: 85% of the genes labeled as transiently perturbed ($|\log_2(p_i)| > 0.5$) have length $\leqslant$ 2kbp (p-value 10^{-5}, hyergeometric test; genes with ORF (acronym for Open Reading Frame, i.e., gene-encoding DNA sequences) length $\leqslant$ 2kbp form $\sim 70\%$ of the total), see Fig. 7 (a). The fact that also downregulated genes are relatively short excludes the scenario in which a marked transient excursion is only the consequence of an increased synthesis rate affecting more the short ORF than the long ones. It is interesting to compare the behavior observed on the transient with the average absolute abundance of the corresponding proteins, as estimated in (Ghaemmaghami et al., 2003) for non-stressed yeast. From Fig. 7 (b) (right lower plot), the induced genes (again, both up- and down-regulated) seem also to correspond to gene products having a low concentration in the "ordinary, stationary" conditions of (Ghaemmaghami et al., 2003) (p-value 0.05, t-test). The short length of the mRNAs certainly favours more rapid fluctuations which could induce more easily changes also at the level of gene products. More marked changes in protein abundances favour the feedback regulation we are hypothesizing. The model (2) does not explicitly include the length of a gene in its parameters. However, as can be seen on the lower left plot of Fig. 7 (b), it tends to associate to a consistent fraction of short genes a high value of the forcing parameter b_i, meaning that the impact of the stimulation on the kinetics on these genes is more pronounced.

4 Conclusions

Yeast reacts to a change of environmental conditions by means of a highly coordinated transcriptional response which is faster than it would be expected from the "natural" degradation time constant but which is only transient. In this paper we propose a model able to explain this quick response by means of a feedback mechanism aiming



(a) (b)

Figure 7: In (a), the average area under the transient response for the 5 time series is plotted against the length of the corresponding ORFs. Longer ORFs clearly tend to be perturbed less, while the genes significatively perturbed (both up- and down-regulated) are for the vast majority shorter than 2kbp (shown in red). There seems to be some degree of inverse correlation also between ORF length and (absolute) protein abundances estimated in (Ghaemmaghami et al., 2003) in ordinary (unperturbed) growth conditions, see (b) (top right), with, in particular, the really abundant proteins corresponding to short ORFs. Likewise, the correlations of both ORF lengths and protein abundances with HL seem to be to some extent skewed, with long lived mRNAs corresponding to short genes and abundant gene products. The color code in (b) is the same as (a). From it (in particular the two plots on the right) we can deduce that most genes perturbed during the transient stress response (red dots) correspond to products having low/medium abundances.

at adapting the system to the new condition. From a dynamical point of view, this can be formulated in terms of a second mode, faster than degradation, which dominates the transient excursion but which, being quickly exhausted, is not observable on standard turnover experimental curves. It is shown that this second mode can be induced by a feedback mechanism from a much slower dynamical variable, which could correspond to the concentration of gene products.

5 Materials and Methods

5.1 Model construction

The changes in the relative concentration of mRNA with respect to its basal level can be described with a typical model for the transcription kinetics (Hargrove & Schmidt, 1989; Alon, 2006; José E. Pérez-Ortin & Moreno, 2007; Ronen et al., 2002; Ronen & Botstein, 2006; Foat et al., 2005; Farina et al., 2008):

$$\frac{dm_i}{dt} = -\delta_i m_i + f_i, \tag{4}$$

where the function f_i describes the transcription synthesis rate for the i-th gene and is usually zero-order in m_i, i.e., independent of the concentration of m_i. In the literature, f_i is often expressed as a function of the transcription factor(s) w_i governing the expression of m_i, with various types of functional dependence like linear, Michaelis-Menten or of Hill type (José E. Pérez-Ortin & Moreno, 2007; Ronen et al., 2002; Khanin et al., 2006; Ronen & Botstein, 2006; Foat et al., 2005; Buchler et al., 2005), see Supplementary Notes for examples. Following this approach requires the knowledge of the transcription factor w_i acting on each gene. Even if this information is partially available for *S.cerevisiae* (Luscombe et al., 2004; Pilpel et al., 2001), predicting the kinetics of the transient response from them is troublesome for the reasons explained at the end of Section 2, and also because in the literature the kinetic models mentioned above are mostly used for describing variations in the steady state following a perturbation, not for the transient dynamics itself. Moreover, under the assumption that the transcription synthesis rate f_i is of zero-order in m_i (José E. Pérez-Ortin & Moreno, 2007), the fast rising front of the transient cannot be explained in terms of a model like (4) at least for reasonable values of the degradation time constants δ_i (Fig. 2, see Supplementary Notes for a thorough analysis). On top of all these complications, modeling the effect of an external stimulus u on the transcriptional regulation means expressing w_i as a function of u. Nothing is known in general about this further functional dependence $w_i = w_i(u)$. Bypassing the transcription factors, the $f(u)$ can for example be represented as an open-loop impulse like in (Chechik & Koller, 2009) or, more generally, as a finite width kernel, vanishing after some time, see Fig. 2 (a). As the m_i represent relative concentrations, these open loop models entail (without explicitely explaining) a form of memory of the "ideal" pre-stimulus absolute concentration, as well as a form of adaptation if one considers the stimulation u as a step (e.g. a permanent increase in temperature). Both elements are characterized in our model by means of a feedback term. In absence of such feedback, the transcription synthesis rate f_i consists for us only of a basal (constant) term plus a term linear in the stimulus u, of the form of a zero order kinetics in m_i:

$$\frac{dm_i}{dt} = -\delta_i m_i + c_i + b_i u. \tag{5}$$

The parameter c_i corresponds to the basal rate of transcription in absence of external stimuli ($u = 0$). Therefore, since for the unperturbed system the steady state must be $\bar{m}_i = c_i/\delta_i = 1$, we have $c_i = \delta_i$. The parameter b_i instead carries information about the activator/inhibitor effect of u on the mRNA concentration. When u is a persistent stimulus, e.g. $u(t) = 1$, $t \geqslant 0$, then in (5) the steady state value is modified to $m_i = (c_i + b_i u)/\delta_i = 1 + b_i u/\delta_i \neq 1$, i.e., the system (5) is not adapted to step-like inputs u and cannot recover its pre-stimulus mRNA level.

An increase in the transcription rate of the i-th gene induces an increase in the total quantity of mRNA produced over time

$$p_i(t) - \bar{p}_i = r\int_0^t (m_i(\tau) - \bar{m}_i)d\tau = r\int_0^t (m_i(\tau) - 1)d\tau \tag{6}$$

where, as above, $\bar{m}_i = 1$ is the pre-stimulus relative mRNA abundance, r is a rate constant (representing for example the ribosome density, and assumed to be the same for all genes) and $\bar{p}_i$ is an integration constant (representing the basal level of p_i, see below). Differentiating this expression,

$$\frac{dp_i}{dt} = r(m_i - 1), \tag{7}$$

we see that the variable p_i represents a dynamical quantity "downstream" of transcription. In this paper p_i is taken to describe the concentration of the corresponding gene product relative to the basal level, hereafter fixed as $\bar{p}_i = 1$. This (very common (Hargrove & Schmidt, 1989; Belle et al., 2006; Simpson et al., 2003)) choice is a simplification of the complex mechanisms characterizing translation and protein synthesis, involving for example changes in the translation initiation, in the ribosomal density or in the polysomal association (Kuhn et al., 2001; Preiss et al., 2003), all steps not well-characterized dynamically. From what is known experimentally, the dynamics at the polysomes level for example seems to be correlated with the transcriptional perturbation of the mRNAs (in (Preiss et al., 2003) it is shown that the frequency of association with polysomes increases for upregulated genes and decreases for downregulated genes). An ODE like (7) for a gene product usually contains a degradation term. Given that the transcriptional perturbation propagates through the protein synthesis process with a time delay and that the protein turnover rate is typically considered slower than the corresponding mRNA turnover rate (Hargrove & Schmidt, 1989; Belle et al., 2006), the influence of the protein degradation term on the dynamics becomes negligible for the time horizon of interest here, see Fig. 1 (b).

The homeostatic effect assumed in the paper consists of a feedback autoregulation acting in correspondence of a displacement from the basal level (i.e., for $p_i \neq 1$) and can be modeled as in the system (2). The model (2) predicts that the equilibrium is reached for p_i corresponding to $\tilde{p}_i = 1 + b/a$. In order to have $\tilde{p}_i > 0$, the parameters must therefore satisfy the consistency condition $b_i > -a_i$. *De facto*, the amplitude of p_i depends on the rate constant r. For all time series considered, a choice of $r = 0.01$ (motivated by the experimental data rather than by the dynamical model chosen, see Supplementary Notes) is sufficient to have biologically consistent values of $\tilde{p}_i$ for the range of a_i, b_i required by the fitting procedure.

Since the model misses a degradation term in p_i, the protein concentration changes in response to the persistent stimulus from $\bar{p}_i$ to $\tilde{p}_i$ without ever returning to the basal level. Introducing such a term as in (1) typically leads to only minor differences, although exact adaptation in m_i and monotonicity in p_i are lost. For sufficiently high ratios of δ_i/λ_i (~ 5 or larger), the differences with respect to (2) are minimal, and we can talk about "quasi-adaptation" and of an autoregulatory feedback which behaves like a "leaky" integral, see Fig. 2 (d). The system still has two modes with distinct time constants and the dominant mode still affects primarily only the rising front of m_i.

5.2 Model identification and analysis

To simplify calculations, it is convenient to change variables, shifting the steady state to the origin. Letting $x_i = \begin{bmatrix} m_i - 1 \\ p_i - 1 \end{bmatrix}$, and denoting the state and input matrices for the systems as

$$A_i = \begin{bmatrix} -\delta_i & -a_i \\ r & 0 \end{bmatrix}, \qquad \text{and} \qquad B_i = \begin{bmatrix} b_i \\ 0 \end{bmatrix}, \tag{8}$$

then for each gene we have the linear system (with input)

$$\dot{x}_i = A_i x_i + B_i u \tag{9}$$

whose solution for the step response is

$$x_i(t) = e^{A_i t} x_i(0) + \int_0^t e^{A_i(t-\tau)} B_i u(\tau) d\tau. \tag{10}$$

Since $\mathrm{tr}(A_i) = -\delta_i < 0$ and $\det(A_i) = r a_i > 0$, the system is always stable and its eigenvalues are:

$$s_{i,1} = -\frac{\delta_i}{2} - \frac{\gamma_i}{2}, \qquad \text{and} \qquad s_{i,2} = -\frac{\delta_i}{2} + \frac{\gamma_i}{2} \tag{11}$$

where $\gamma_i = \sqrt{\delta_i^2 - 4 r a_i}$. A visual inspection of the time series shows that for the vast majority of genes the large excursion corresponding to the transient is damped without inducing oscillatory behavior (at least above what can be considered measurement noise). Hence in the model fitting we assumed:

1. the two eigenvalues are real, i.e., $\delta_i^2 - 4 r a_i > 0$;

2. the time constant of the fastest eigenvalue is shorter than that of the "free degradation" given by the HL alone.

The two conditions are compatible with each other and with the model structure. In order to agree also with the available HL measures, we shall assume the following:

$$s_{i,1} < s_{i,2} \sim -\frac{\ln(2)}{\mathrm{HL}_i} < 0.$$

If we choose $s_{i,2} = -\frac{\ln(2)}{\mathrm{HL}_i}$, then we obtain the following conditions:

$$\begin{cases} a_i = -s_{i,2}(\delta_i + s_{i,2})/r > 0 \\ \delta_i > -2s_{i,2} > 0 \end{cases} \tag{12}$$

In correspondence of a persistent stimulus, $u(t) = 1$ for $t \geqslant 0$, the system (10) can be solved explicitly. Since at $t = 0$ the system is at rest (i.e., in the basal state $x_i(0) = [0\ 0]^T$ for all i, corresponding to $m_i(0) = 1$ and $p_i(0) = 1$), only the forced evolution (second term in (10)) matters and we obtain:

$$x_i(t) = \frac{b_i}{\gamma_i} \begin{bmatrix} e^{s_{i,2}t} - e^{s_{i,1}t} \\ r\left(\frac{e^{s_{i,2}t}-1}{s_{i,2}} - \frac{e^{s_{i,1}t}-1}{s_{i,1}}\right) \end{bmatrix} \tag{13}$$

i.e., Equation (3) for $m_i(t)$ and $p_i(t)$.

Notice that as $t \to \infty$ from the second equation of (3) we obtain that $p_i(t) > 0$ if $rb_i\left(1/s_{i,1} - 1/s_{i,2}\right)/\gamma_i > -1$, i.e., for $b_i > -a_i$ as mentioned above.

The first equation of (13) can be used to fit the parameters in the dynamical model (9). For each gene, this corresponds to identifying the values of δ_i and b_i that optimize the fit of $m_i(t)$ to the experimental time series. With these parameters, (9) is completely determined. The second equation of (13) can then be used to compare the area predicted by the model with the area computed from the experimental data.

References

Airoldi, E. M., Huttenhower, C., Gresham, D., Lu, C., Caudy, A. A., Dunham, M. J., Broach, J. R., Botstein, D., & Troyanskaya, O. G. (2009). Predicting cellular growth from gene expression signatures. *PLoS Comput Biol*, *5*(1).

Alon, U. (2006). *An Introduction to Systems Biology - Design Principles of Biological Circuits*. Chapman & Hall/CRC;.

Alon, U., Surette, M. G., Barkai, N., & Leibler, S. (1999). Robustness in bacterial chemotaxis. *Nature*, *397*(6715), 168–171.

Becskei, A., & Serrano, L. (2000). Engineering stability in gene networks by autoregulation. *Nature*, *405*(6786), 590–593.

Behar, M., Hao, N., Dohlman, H. G., & Elston, T. C. (2007). Mathematical and computational analysis of adaptation via feedback inhibition in signal transduction pathways. *Biophys J*, *93*(3), 806–821.

Belle, A., Tanay, A., Bitincka, L., Shamir, R., & O'Shea, E. K. (2006). Quantification of protein half-lives in the budding yeast proteome. *Proc. Natl. Acad. Sci. U.S.A.*, *103*(35), 13004–13009.

Buchler, N. E., Gerland, U., & Hwa, T. (2005). Nonlinear protein degradation and the function of genetic circuits. *Proc. Natl. Acad. Sci. U.S.A.*, *102*(27), 9559–9564.

Causton, H. C., Ren, B., Koh, S. S., Harbison, C. T., Kanin, E., Jennings, E. G., Lee, T. I., True, H. L., Lander, E. S., & Young, R. A. (2001). Remodeling of yeast genome expression in response to environmental changes. *Mol Biol Cell*, *12*(2), 323–337.

Chechik, G., & Koller, D. (2009). Timing of gene expression responses to environmental changes. *Journal of Computational Biology*, *16*(2), 279–290.

Chechik, G., Oh, E., Rando, O., Weissman, J., Regev, A., & Koller, D. (2008). Activity motifs reveal principles of timing in transcriptional control of the yeast metabolic network. *Nat. Biotechnol.*, *26*(11), 1251–1259.

DeRisi, J. L., Iyer, V. R., & Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, *278*(5338), 680–686.

Farina, L., De Santis, A., Salvucci, S., Morelli, G., & Ruberti, I. (2008). Embedding mrna stability in correlation analysis of time-series gene expression data. *PLoS Comput Biol*, *4*(8).

Foat, B. C., Houshmandi, S. S., Olivas, W. M., & Bussemaker, H. J. (2005). Profiling condition-specific, genome-wide regulation of mRNA stability in yeast. *Proc. Natl. Acad. Sci. U.S.A.*, *102*(49), 17675–17680.

Gasch, A. P. (2002). The environmental stress response: a common yeast response to environmental stresses. In S. Hohmann, & P. Mager (Eds.) *Yeast Stress Responses*, (pp. 11–70). Springer-Verlag, Heidelberg.

Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., & Brown, P. O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, *11*(12), 4241–4257.

Ghaemmaghami, S., Huh, W. K., Bower, K., Howson, R. W., Belle, A., Dephoure, N., O'Shea, E. K., & Weissman, J. S. (2003). Global analysis of protein expression in yeast. *Nature*, *425*(6959), 737–741.

Goldberger, R. F. (1974). Autogenous regulation of gene expression. *Science*, *183*, 810 – 816.

Grigull, J., Mnaimneh, S., Pootoolal, J., Robinson, M. D., & Hughes, T. R. (2004). Genome-wide analysis of mRNA stability using transcription inhibitors and microarrays reveals posttranscriptional control of ribosome biogenesis factors. *Mol. Cell. Biol.*, *24*(12), 5534–5547.

Hargrove, J. l., & Schmidt, F. H. (1989). The role of mRNA and protein stability in gene expression. *FASEB J.*, *3*(12), 2360–2370.

José E. Pérez-Ortin, P. M. A., & Moreno, J. (2007). Genomics and gene transcription kinetics in yeast. *Trends in Genetics*, *23*(5), 250–257.

Khanin, R., Vinciotti, V., & Wit, E. (2006). Reconstructing repressor protein levels from expression of gene targets in Escherichia coli. *Proc. Natl. Acad. Sci. U.S.A.*, *103*(49), 18592–18596.

Kuai, L., Das, B., & Sherman, F. (2005). A nuclear degradation pathway controls the abundance of normal mRNAs in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U.S.A.*, *102*(39), 13962–13967.

Kuhn, K. M., DeRisi, J. L., Brown, P. O., & Sarnow, P. (2001). Global and specific translational regulation in the genomic response of saccharomyces cerevisiae to a rapid transfer from a fermentable to a nonfermentable carbon source. *Mol. Cell. Biol.*, *21*(3), 916–927.

Levy, S., Ihmels, J., Carmi, M., Weinberger, A., Friedlander, G., & Barkai, N. (2007). Strategy of transcription regulation in the budding yeast. *PLoS ONE*, *2*(2).

Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A., & Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, *431*, 308–312.

Pilpel, Y., Sudarsanam, P., & Church, G. M. (2001). Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat Genet*, *29*(2), 153–159.

Preiss, T., Baron-Benhamou, J., Ansorge, W., & Hentze, M. W. (2003). Homodirectional changes in transcriptome composition and mRNA translation induced by rapamycin and heat shock. *Nat. Struct. Biol.*, *10*(12), 1039–1047.

Ronen, M., & Botstein, D. (2006). Transcriptional response of steady-state yeast cultures to transient perturbations in carbon source. *Proc. Natl. Acad. Sci. U.S.A.*, *103*(2), 389–394.

Ronen, M., Rosenberg, R., Shraiman, B. I., & Alon, U. (2002). Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proc Natl Acad Sci U S A*, *99*(16), 10555–10560.

Rosenfeld, N., Elowitz, M. B., & Alon, U. (2002). Negative autoregulation speeds the response times of transcription networks. *J Mol Biol*, *323*(5), 785–793.

Savageau, M. A. (1974). Comparison of classical and autogenous systems of regulation in inducible operons. *Nature*, *252*, 546–549.

Shalem, O., Dahan, O., Levo, M., Martinez, M. R., Furman, I., Segal, E., & Pilpel, Y. (2008). Transient transcriptional responses to stress are generated by opposing effects of mRNA production and degradation. *Mol Syst Biol*, *4*, 223–223.

Simpson, M. L., Cox, C. D., & Sayler, G. S. (2003). Frequency domain analysis of noise in autoregulated gene circuits. *Proc. Natl. Acad. Sci. U.S.A.*, *100*(8), 4551–4556.

Soranzo, N., Zampieri, M., Farina, L., & Altafini, C. (2009). mRNA stability and the unfolding of gene expression in the long-period yeast metabolic cycle. *BMC Systems Biology*, *3*, 18.

Tirosh, I., Weinberger, A., Carmi, M., & Barkai, N. (2006). A genetic signature of interspecies variations in gene expression. *Nat. Genet.*, *38*(7), 830–834.

Torre, V., Ashmore, J. F., Lamb, T. D., & Menini, A. (1995). Transduction and adaptation in sensory receptor cells. *J Neurosci*, *15*(12), 7757–7768.

Tu, B. P., Kudlicki, A., Rowicka, M., & McKnight, S. L. (2005). Logic of the yeast metabolic cycle: Temporal compartmentalization of cellular processes. *Science*, *310*(5751), 1152–1158.

Wang, Y., Liu, C. L., Storey, J. D., Tibshirani, R. J., Herschlag, D., & Brown, P. O. (2002). Precision and functional specificity in mRNA decay. *Proc. Natl. Acad. Sci. U.S.A.*, *99*(9), 5860–5865.

Yi, T. M., Huang, Y., Simon, M. I., & Doyle, J. (2000). Robust perfect adaptation in bacterial chemotaxis through integral feedback control. *Proc. Natl. Acad. Sci. U.S.A.*, *97*(9), 4649–4653.

Yoshimoto, H., Saltsman, K., Gasch, A. P., Li, H. X., Ogawa, N., Botstein, D., Brown, P. O., & Cyert, M. S. (2002). Genome-wide analysis of gene expression regulated by the calcineurin/crz1p signaling pathway in saccharomyces cerevisiae. *J Biol Chem*, *277*(34), 31079–31088.

Appendix 4.2 Paper: Eduati et al., J Comp Biol, 2012

The following publication dealing with qualitative modelling of small subnetworks have been coauthored by the Ph.D. candidate during her doctoral program.

- F. Eduati, B. Di Camillo, M. Karbiener, M. Scheideler, D. Cora, M. Caselle, and G. Toffolo. *Dynamic modeling of miRNA-mediated feed-forward loops*. Journal of Computational Biology, 19(2):188-199, 2012.

Full text of the original paper is reported in this Appendix.

Dynamic modeling of miRNA-mediated feed-forward loops

F. Eduati[1], B. Di Camillo[1], M. Karbiener[2], M. Scheideler[2], D. Corà[3], M. Caselle[4,5]
and G. Toffolo[1,*]

[1]Department of Information Engineering, University of Padova, Padova, Italy
[2]Institute for Genomics and Bioinformatics, Graz University, Graz, Austria
[3]IRCC, School of Medicine, University of Torino, Torino, Italy
[4]Department of Theoretical Physics, University of Torino, Torino, Italy
[5]Center for Complex Systems in Molecular Biology and Medicine, Torino, Italy

[*]*Corresponding author:* `toffolo@dei.unipd.it`

Abstract

Given the important role of microRNAs (miRNAs) in genome-wide regulation of gene expression, increasing interest is devoted to mixed transcriptional and post-transcriptional regulatory networks analyzing the combinatorial effect of transcription factors (TFs) and miRNAs on target genes. In particular, miRNAs are known to be involved in feed-forward loops (FFLs) where a TF regulates a miRNA and they both regulate a target gene. Different algorithms have been proposed to identify miRNA targets, based on pairing between the 5' region of the miRNA and the 3'UTR of the target gene and correlation between miRNA host genes and target mRNA expression data. Here we propose a quantitative approach integrating an existing method for mixed FFL identification based on sequence analysis with differential equation modeling approach that permits to select active FFLs based on their dynamics. Different models are assessed based on their ability to properly reproduce miRNA and mRNA expression data in terms of identification criteria, namely: goodness of fit, precision of the estimates and comparison with submodels. In comparison with standard approach based on correlation, our method improves in specificity. As a case study, we applied our method to adipogenic differentiation gene expression data providing potential novel players in this regulatory network.

1 Introduction

MicroRNAs (miRNAs) are small ($\sim$ 22 nt) non-coding RNAs that post-transcriptionally regulate gene expression. They are transcribed as pri-miRNAs, then processed and exported from the nucleus to the cytoplasm in the form of pre-miRNA hairpins where they are cleaved by Dicer enzyme and incorporated in the RNA-induced silencing complex

Eduati et al., J. Comp. Biol., 2012, 19(2): 188-199

(RISC) to allow the interaction with target mRNAs via base pairing: binding to mRNA 3' UTR causes the decrease of the frequency of translation and the increase of mRNA degradation rate (Du & Zamore, 2005; Bartel, 2004; Baek et al., 2008; Selbach et al., 2008). MiRNAs are known to be involved in different biological processes, e.g. cell cycle control, cellular growth, differentiation, apoptosis and embryogenesis, and to play critical roles in human diseases (Jiang et al., 2009). Their important regulatory role has come into focus in the last few years and main attention has been paid to miRNAs and their target genes identification (Lagos-Quintana et al., 2003; Bentwich et al., 2005; Jung et al., 2010; Lagos-Quintana et al., 2001). Different algorithms have been developed at this purpose, based on sequence data, looking for evolutionarily conserved Watson-Crick pairing between the 5' region of the miRNA and the 3'UTR of the target gene (Griffiths-Jones et al., 2006; Bartel, 2009; Friedman et al., 2009; Lewis et al., 2003, 2005). There is also increasing interest in the dynamic description and the quantification of the regulation of gene expression by miRNAs and several scientific studies have characterized miRNA mediated degradation rates using models based on ordinary differential equation (Khanin & Vinciotti, 2008; Shimoni et al., 2007; Levine et al., 2007b,a; Vohradsky et al., 2010).

Given the important role of miRNAs in genome-wide regulation of gene expression, increasing interest is devoted to mixed transcriptional and post-transcriptional regulatory networks analyzing the combinatorial effect of transcription factors (TFs) and miRNAs on target genes. In particular, miRNAs are known to be involved in feed-forward loops (FFLs) where a TF regulates a miRNA and they both regulate a target gene (Shimoni et al., 2007; Shalgi et al., 2007; Tsang et al., 2007; Re et al., 2009). The dynamic of FFL has been extensively studied in transcriptional networks (Mangan & Alon, 2003; Kalir et al., 2005; Kaplan et al., 2008; Macia et al., 2009; Alon, 2007) since this regulatory pattern is overrepresented in biological networks with respect to random networks (Milo et al., 2002; Shen-Orr et al., 2002) and thus represents a basic building block, favored by evolution and playing important functional roles. For example, FFLs involving miRNAs permit to accomplish target gene fine tuning and noise buffering (Li et al., 2009; Wu et al., 2009). In (Tsang et al., 2007) Correlation between miRNA host genes and target mRNA has been assessed together with conserved 3'UTR motifs to define putative regulatory relationships between a miRNA and a set of target genes sharing the same TF. A quantitative description of the regulatory interactions, e.g. based on differential equation models, could be helpful to characterize putative miRNA mediated FFLs. A similar approach has been adopted in (Vu & Vohradsky, 2007; Chen et al., 2005, 2004), where differential equations were fitted to expression data for transcriptional networks not involving miRNAs. As regards small RNA mediated FFL, a differential equation based model has been used in (Shimoni et al., 2007) only to simulate the dynamic of a generic circuit using plausible parameter values derived from literature.

In this work we propose a general analytical framework based on the use of differential equations to extensively characterize a list of putative miRNA mediated FFLs. Our approach, when applied to a list of putative FFLs, provides some criteria to select active FFLs based on their ability to reproduce dynamic expression data. In this context, we do

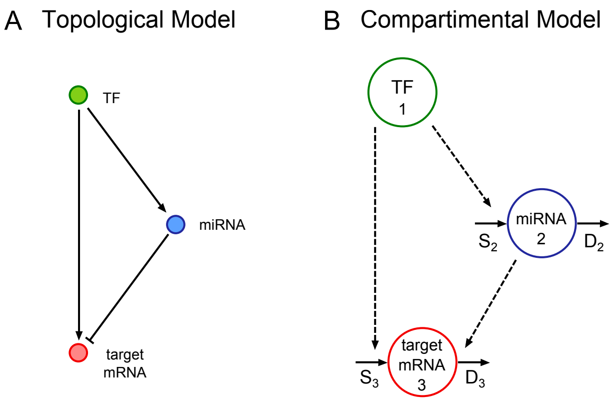


Figure 1: MiRNA mediated FFL. (A) Topologial model of the FFL where a TF regulates a miRNA and they both regulate the target mRNA: TF regulations can be positive or negative while miRNA regulation of the target gene is negative; (B) compartmental model of the FFL where S and D represents synthesis and degradation, respectively and dotted arrows are the regulation processes affecting S and D.

not use the data to validate the models, but, on the opposite, three models are used to fit the data and select active FFLs based on the goodness of fit. The first model M1 is borrowed from previous literature (Khanin & Vinciotti, 2008; Shimoni et al., 2007; Levine et al., 2007b,a) . Models M2 and M3 are linear simplifications of model M1 since, as shown in the following, the choice of the most appropriate model strictly depends on the available dataset.

We estimate the significance of our method in comparison with random FFLs obtained by randomly selecting links between miRNAs, TFs and target mRNA and in comparison with a more standard approach, based on correlation between TF, miRNA and target mRNA.

2 Models

In the miRNA mediated FFL circuit (Fig. 1A) a transcription factor TF (X_1) regulates a miRNA (X_2) and they both regulate a target mRNA (X_3). Three models based on ordinary differential equations (ODEs) are examined to describe the miRNA and target mRNA expression kinetics. All models consider X_1 as forcing function and describe the rate of change of X_2 and X_3 as the balance between their synthesis/transcription (S_i) and degradation (D_i) with the basal expression level (X_{ib}) as initial condition, the correspondent compartmental model is shown in Fig. 1B. Thus, for $i = 2, 3$, the differential equation describing the variables is

$$\dot{X}_i(t) = S_i(t) - D_i(t), \qquad X_i(0) = X_{ib} \tag{1}$$

The synthesis is expressed as the sum of a basal term (S_{ib}) plus a positive (activation) or negative (repression) term (ΔS_i) encoding the effect of the specific TF on the transcription of miRNA and target mRNA. As regards degradation (D_i), for miRNA it is assumed to be a function only of its expression while for the target mRNA the effect of the miRNA level is also modeled.

$$\begin{aligned}\dot{X}_2(t) &= S_{2b}(t) + \Delta S_2\left[X_2(t)\right] - D_2\left[X_2(t)\right] \\ \dot{X}_3(t) &= S_{3b}(t) + \Delta S_3\left[X_1(t)\right] - D_3\left[X_2(t), X_3(t)\right]\end{aligned} \tag{2}$$

The three models adopt the same description for miRNA degradation, i.e. a first order process with constant rate d_2 , while they differ in the functional description assumed for ΔS_2, ΔS_3 and D_3.

1. **Model M1** describes the TF regulation on the miRNA (ΔS_2) and the target mRNA (ΔS_3) by a saturative Michaelis-Menten function, and the miRNA mediated degradation of the target mRNA (D_3) as the sum of a first order process, with constant rate, with respect to X_3 and a nonlinear term that depends also on X_2 as in (Khanin & Vinciotti, 2008; Shimoni et al., 2007; Levine et al., 2007a,b).

2. **Model M2** assumes TF regulation (ΔS_2,ΔS_3) to be linearly dependent on its level, while the functional description of target mRNA degradation (D_3) has nonlinear dynamics as in M1.

3. **Model M3** is derived from M2 linearizing the miRNA mediated degradation model (D_3), thus the kinetics of the whole model is linear.

Since in log scale spot array data are expressed as differences with respect to a basal pre-differentiation state, it is convenient to consider as state variables $x_i = X_i - X_{ib}$ for $i = 1, 2, 3$ where X_{ib} is the reference, collected at day -3. Considering that at the basal state $\dot{X}_i(t) = 0$ for $i = 2, 3$ it is possible to express the basal transcriptions S_i as function of the regulation parameters and the basal expression levels. After some passages, models M1, M2 and M3 turn out to be:

1. **Model M1**

$$\begin{aligned}\dot{x}_2(t) &= \frac{\alpha_2 x_1(t)}{\beta_2 + x_1(t)} - d_2 x_2(t) & x_2(0) &= 0 \\ \dot{x}_3(t) &= \frac{\alpha_3 x_1(t)}{\beta_3 + x_1(t)} - p x_3(t) - q x_2(t) - r x_2(t) x_3(t) & x_3(0) &= 0\end{aligned} \tag{3}$$

2. **Model M2**

$$\begin{aligned}\dot{x}_2(t) &= a_2 x_1(t) - d_2 x_2(t) & x_2(0) &= 0 \\ \dot{x}_3(t) &= a_3 x_1(t) - p x_3(t) - q x_2(t) - r x_2(t) x_3(t) & x_3(0) &= 0\end{aligned} \tag{4}$$

3. **Model M3**

$$\begin{aligned} \dot{x}_2(t) &= a_2 x_1(t) - d_2 x_2(t) & x_2(0) &= 0 \\ \dot{x}_3(t) &= a_3 x_1(t) - d_3 x_3(t) - s x_2(t) & x_3(0) &= 0 \end{aligned} \tag{5}$$

The mathematical derivation of Equations (3), (4) and (5) and the meaning of each parameter in terms of synthesis and degradation rate are detailed in the Supplementary Material.

2.1 Model identification

A priori identifiability analysis of M1, M2 and M3 (Equations (3), (4) and (5)) tested using the software DAISY (Bellu et al., 2007), indicates that all three models are a priori globally identifiable, i.e. it is theoretically possible to estimate the set of unknown parameters $\boldsymbol{\vartheta}$ from the data, at least under ideal conditions (noise-free data, continuous time observations and error-free model structure).

$\hat{\boldsymbol{\vartheta}}$ can be estimated by Weighted Least Square, i.e. minimizing the Weighted Residual Sum of Squares (WRSS)

$$WRSS = \sum_{i=2,3} \sum_{j=1}^{N_i} \omega_i(t_j) \left[z_i(t_j) - x_i(t_j, \boldsymbol{\vartheta}) \right]^2 \tag{6}$$

where $z_i(t_j)$ is the observed datum at time j, $x_i(t_j, \boldsymbol{\vartheta})$ is the predicted datum at time j computed using the model Equations (3), (4) and (5), $\omega_i(t_j)$ is the weight assigned to datum j (inverse of the variance of the measurement error) and N_i is the number of time points. The external summation takes into account that residuals for both miRNA and target mRNA are simultaneously minimized, thus miRNA e mRNA time series collected under the same experimental conditions are required for model identification. The measurement error is assumed to be Gaussian with zero mean and a known variance. The variance can be experimentally determined by analyzing replicates of each measure. A general model for the error variance is

$$v_i(t_j) = \alpha + \beta \left[z_i(t_j) \right]^{\gamma} \tag{7}$$

where α, β and γ are parameters to be estimated from replicates, e.g. by plotting the mean of each replicate against its variance and fitting on these data the unknown parameters of the error model Equation (7), as described in (Cobelli et al., 2000).

Since data are affected by a measurement error, also $\hat{\boldsymbol{\vartheta}}$ is affected by an error and the a posteriori identifiability of the models assesses the precision with which the parameters are estimated in terms of percentage coefficient of variation (CV)

$$CV(\hat{\boldsymbol{\vartheta}}) = \frac{SD(\hat{\boldsymbol{\vartheta}})}{\hat{\boldsymbol{\vartheta}}} \cdot 100 \tag{8}$$

where $SD(\hat{\boldsymbol{\vartheta}})$ is the standard deviation of the estimate.

2.2 FFLs selection

For each model, selection of active FFLs from a large set of putative ones exploits identification results in terms of consistency with the three following criteria:

1. **Goodness of fit.** A valid model should provide an adequate fit to the data. The goodness of fit can be evaluated on residuals, based both on their whiteness, i.e. residuals should be uncorrelated, and on their amplitude, i.e. deviation between predicted and observed values should be comparable to the measurement error. To evaluate the whiteness of the residuals, the number of runs, i.e. subsequences of residuals having the same sign, are analyzed for both miRNA and mRNA residual patterns. For the amplitude property, a global measure is provided by WRSS divided by the degree of freedom, i.e. difference between the number of data and the number of parameters: since weighted residuals should be independent with unit variance, WRSS should be the outcome of a random variable with Chi-Square distribution.

2. **Precision of the estimates.** FFLs having all parameters estimates with $CV < 100$ are considered reliable.

3. **Comparison with submodels.** In order to verify that the FFL model (Fig. 1B) is the optimal description of the circuit, its performance is compared with that of two submodels (Fig. 2) with missing regulatory links: in Submodel 1 the regulatory link between the TF and the target mRNA is missing, while in Submodel 2 the effect of miRNA on target mRNA degradation rate is not considered. Once the two submodels are identified, their performance is assessed versus the original one based on the Akaike Information Criterion (AIC) that implements the principle of parsimony, i.e. selects the model best able to fit the data with the minimum number of parameters:

$$AIC = WRSS + 2L \tag{9}$$

   The FFL model is selected if its AIC is the lowest compared with submodels.

Summing up, if criteria 1 and 2 are satisfied for a dataset of putative FFL data, i.e. the model satisfactorily reproduces the data with all parameters precisely estimated from them, criteria 3 is applied and the FFL topology is finally selected as active provided that the complete model results to be the optimal model according to the AIC.

3 A case study in adipogenesis

To discuss a practical application of the proposed method, we applied it to miRNA and mRNA expression time series of human multipotent adipose-derived stem cells (hMADS) upon adipogenic differentiation. The initial panel of putative FFLs was selected based on sequence analysis; therefore it includes also false positive matches and/or FFLs non active during adipogenesis.

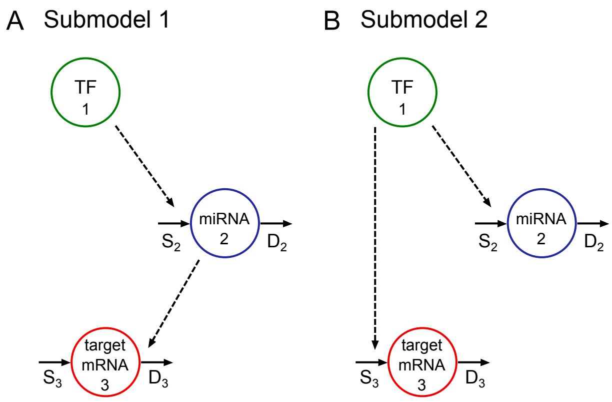


Figure 2: Submodels with missing regulatory links with respect to the FFL. (A) No effect of TF regulation on target gene; (B) no effect of miRNA on mRNA degradation rate.

3.1 Data

Two independent cell culture experiments were performed as biological replicates during adipogenic differentiation of human mesenchymal stem cells as previously described in (Scheideler et al., 2008; Karbiener et al., 2009). Cells were harvested at the preconfluent stage as reference (day -3) and at seven subsequent time points during human adipogenic differentiation: day -2 and 0 before, and 1, 2, 5, 10, 15 days after induction of differentiation. All hybridizations were repeated with reversed dye assignment (dye-swap). Background subtraction as well as global mean and dye swap normalization were applied. The resulting ratios were $\log 2$ transformed and the independent experiments were averaged. Complete miRNA and mRNA time-series expression data used for this study conform to the MIAME guidelines and are available in GEO database (GSE29186).

A list of mixed TF / miRNA FFLs was generated by means of a bioinformatic pipeline mainly based on an ab-initio sequence analysis of human and mouse regulatory regions as described in (Re et al., 2009) using CircuitsDB (Friard et al., 2010). Briefly, in CircuitsDB a catalogue of non-redundant promoter regions for protein-coding and miRNA genes in the human and mouse genomes were first constructed (see Supplementary Material for additional details). In parallel to that, a catalogue of non-redundant human and mouse 3'-UTR regions for protein-coding genes was defined. A transcriptional regulatory network and, separately, a list of post-transcriptionally regulated genes was then generated for human by looking for conserved overrepresented motifs in the human and mouse promoters and 3'-UTRs previously assembled. The two networks were subsequently combined looking for mixed feed-forward regulatory loops, i.e. all the possible instances in which a master transcription factor regulates a miRNA and together with it a set of joint target coding genes.

Associating the list of 474 miRNA-mediated FFLs obtained using CircuitsDB with the available miRNA and mRNA time series data, the final dataset consisted of 329 putative FFLs (Supplementary Table S1) including 33 TFs, 35 miRNAs and 184 target mRNAs.

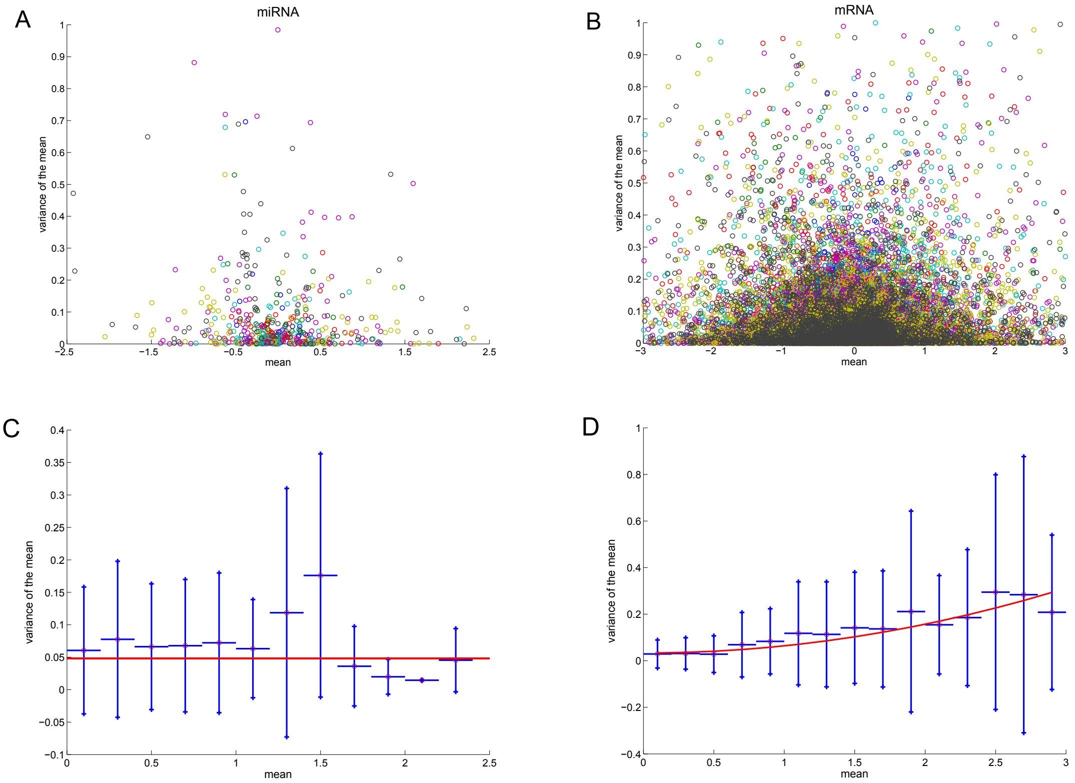


Figure 3: Measurement error variance against expression estimated from the replicates for (A) miRNA and (B) mRNA datasets. In (C) and (D) these data are binned and, for each interval, the mean ± standard deviation is represented; the red line shows the fitted measurement error models, Equation (10).

3.2 Measurement error

The measurement error models for miRNA and mRNA expression data were derived from the replicates, shown in Fig. 3A-B, respectively, as mean of the intensities versus their variance. To better define the dependence of the variance on the intensity, the positive x-axis was divided in intervals and, for each interval, the variance mean values were averaged as shown in Fig. 3C-D. By fitting Equation (7) on these data, the resulting models are

$$\begin{aligned} \nu_2(t_j) &= 0.0484 \\ \nu_i(t_j) &= 0.033 + 0.031 \cdot z_i(t_j)^2, \qquad i = 1, 3 \end{aligned} \tag{10}$$

where ν_2 and ν_i in Equation (10) are referred to the miRNA and to the mRNA (valid for both TFs, and target mRNAs) datasets, respectively.

3.3 Implementation

To assess criterion 1, i.e. whiteness and amplitude of the residuals, statistical tests could not be applied due to the low number (seven) of samples. Thus, conservative empirical thresholds were set to satisfy criterion 1: both miRNA and target mRNA residuals time series must have at least 3 runs and WRSS divided by the degree of freedom lower than 2. All computations were performed in the Matlab environment (Matlab R2010a), further details are supplied in the Supplementary Material.

4 Results

When the three criteria were applied to M1, no FFLs were selected as active, essentially because criterion 2 failed, indicating that the functional descriptions built in the model were too complex to be resolved from the available data. Conversely, 3 FFLs were selected with M2 and 23 with M3 as summarized in Table 1 and Table 2 respectively, where estimated parameters and their precision are reported. Two out of the three FFLs selected using M2 were identified also with M3, thus the total number of active FFLs is 24. It is interesting to notice that most of selected FFLs (21 out of 24) are incoherent. This type of FFL is known to play a significant role in biological regulation conferring precision and stability to gene expression regulation (Mangan & Alon, 2003; Wu et al., 2009; Hornstein & Shomron, 2006; Osella et al., 2011). As discussed in (Macia et al., 2009), the target gene of incoherent FFLs generally shows a pulser response characterized by a rapid increase/decrease of its concentration followed by the return to a new basal level, while the target gene of coherent FFLs tends to exhibit a grader response characterized by a transient increase/decrease from the initial to the final state. These behaviors were confirmed by our data, as evident from Fig. 4, where expression profiles of two incoherent (A) and two coherent (B) FFLs are shown along with the mean target gene expression levels (considering absolute values) between selected incoherent (C) and coherent (D) FFLs.

Analyzing the active FFLs from a biological point of view, it was found that out of the 24 selected FFLs, 9 FFLs involve TFs and 6 involve miRNAs (marked with an x in Table 1 and Table 2) that are already known from the literature to be regulators of adipogenesis and adipocyte-related functions. A discussion of the results in comparison with the biological literature is available as Supplementary Material.

To estimate the significance of the proposed method, ten sets of 329 random FFLs were generated choosing one random miRNA and 2 random mRNA to play the role of the TF and the target gene respectively. Applying the previously described selection procedure, 0 FFLs were selected using M2 and an average of 15.6 FFLs, with a standard deviation of 1.5, were selected using M3. Instead, using a simple correlation analysis to choose FFLs having a correlation coefficient above 0.75 in absolute values for all three links, 12 FFLs were selected on the list of putative FFLs, and 18.6 ± 4.6 were selected on the randomized datasets.

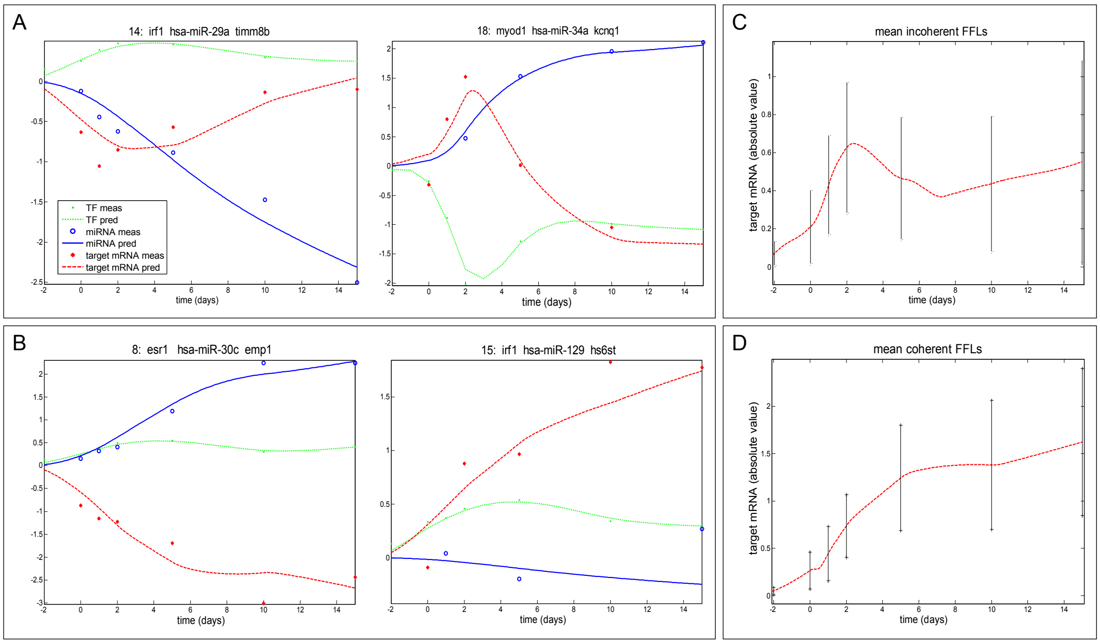


Figure 4: Expression profiles of selected FFLs. TF (green), miRNA (blue) and target mRNA (red) for (A) 2 incoherent and (B) 2 coherent FFLs: spots represent experimental data while lines represent the predicted/reconstructed profiles. In (C) and (D) the average absolute value of predicted target mRNA expression for incoherent and coherent FFLs.

Model M2												
	TF		miRNA		target mRNA	$a_2\ (CV)$	$a_3\ (CV)$	$d_2\ (CV)$	$p\ (CV)$	$q\ (CV)$	$r\ (CV)$	C/I
1	hif1a	x	hsa-miR-24		h41	1.22 (74)	2.83 (44)	0.96 (79)	1.10 (67)	2.17 (52)	0.87 (78)	I
2	srf		hsa-miR-100		impdh1	1.40 (41)	0.68 (59)	0.57 (54)	0.91 (27)	0.34 (63)	0.84 (30)	I
3	tcf4		hsa-miR-23a		ndufa7	0.47 (62)	0.77 (44)	0.30 (83)	1.04 (21)	0.66 (42)	1.29 (69)	I
mean (absolute values)						1.03 (59)	1.43 (49)	0.61 (72)	1.02 (38)	1.06 (52)	0.85 (54)	
SE (absolute values)						0.49 (17)	1.22 (9)	0.33 (16)	0.10 (25)	0.98 (11)	0.25 (26))	

Table 1: Summary of selected FFLs and their estimated parameters using Model M2. TF, miRNA and target mRNA names of selected FFLs using model M2 are reported along with the estimated parameters, their precision in terms of CV and a flag to distinguish between coherent (C) and incoherent (I) FFLs. TF and miRNA already known to be key regulators of adipogenesis and adipocyte-related functions are marked with an x.

Model M3											
	TF		miRNA		target mRNA	a_2 (CV)	a_3 (CV)	d_2 (CV)	d_3 (CV)	s (CV)	C/I
1	runx1		hsa-miR-148b		tnfrsf6b	-0.07 (17)	-1.73 (50)	-	8.28 (25)	4.49 (43)	I
2	runx1		hsa-miR-148b		loc51026	-0.07 (17)	2.70 (9)	-	2.85 (1)	2.70 (76)	C
3	runx1		hsa-miR-148b		tmod	-0.07 (17)	-1.79 (37)	-	3.70 (13)	1.44 (76)	I
4	esr1	x	hsa-miR-148b		map1b	0.14 (18)	1.69 (60)	-	2.49 (6)	4.49 (29)	I
5	esr1	x	hsa-miR-148b		tparl	0.15 (18)	-2.00 (48)	-	2.89 (2)	2.78 (42)	C
6	esr1	x	hsa-miR-148b		apt6m8-9	0.15 (18)	5.37 (33)	-	2.04 (19)	2.41 (41)	I
7	esr1	x	hsa-miR-152		apt6m8-9	0.42 (39)	3.94 (19)	0.16 (67)	1.12 (7)	1.55 (35)	I
8	esr1	x	hsa-miR-30c	x	emp1	0.65 (21)	-4.58 (28)	0.09 (46)	2.16 (5)	1.76 (40)	C
9	ets1		hsa-miR-199a*		hke2	-0.22 (17)	-1.60 (80)	0.90 (17)	0.35 (99)	7.52 (73)	I
10	hif1a	x	hsa-miR-199b		crtl1	-2.32 (59)	-4.99 (24)	1.26 (63)	0.97 (96)	2.47 (27)	I
11	hif1a	x	hsa-miR-24		h41	1.21 (34)	6.02 (35)	0.93 (35)	1.87 (16)	4.21 (46)	I
12	hif1a	x	hsa-miR-199a	x	crtl1	-2.17 (30)	-3.05 (45)	1.19 (25)	1.27 (19)	1.25 (63)	I
13	foxm1		hsa-let-7a		nap1l1	-0.02 (3)	-0.89 (29)	0.14 (58)	1.70 (3)	14.10 (52)	I
14	irf1		hsa-miR-29a	x	timm8b	-0.40 (7)	-5.00 (34)	-	2.23 (34)	0.60 (37)	I
15	irf7		hsa-miR-129		hs6st	-0.04 (82)	2.53 (37)	-	2.38 (6)	13.93 (89)	C
16	irf2		hsa-miR-125b		bcl2	-0.33 (16)	6.07 (23)	-	3.27 (3)	1.55 (45)	C
17	myc	x	hsa-miR-202		tnfrsf4	0.08 (42)	0.34 (92)	-	1.20 (95)	4.44 (87)	I
18	myod1		hsa-miR-34a	x	kcnq1	-0.30 (22)	-2.03 (27)	0.14 (35)	1.49 (24)	2.04 (29)	I
19	myod1		hsa-miR-34a	x	scn2b	-0.28 (21)	-2.00 (27)	0.12 (37)	4.61 (5)	0.58 (88)	I
20	ncx		hsa-let-7e	x	nap1l1	0.13 (67)	8.38 (14)	0.08 (87)	3.17 (3)	10.61 (76)	I
21	nfya		hsa-miR-148b		p3	-0.17 (18)	-1.66 (88)	-	4.05 (6)	4.73 (29)	I
22	tcf4		hsa-miR-23a		ndufa7	0.50 (66)	0.84 (54)	0.33 (89)	1.05 (28)	0.60 (75)	I
23	tel2		hsa-miR-199a*		hke2	0.65 (37)	6.91 (22)	0.34 (41)	1.34 (5)	4.41 (30)	I
mean (absolute values)						0.46 (30)	3.31 (40)	0.47 (50)	2.46 (23)	4.12 (53)	
SE (absolute values)						0.63 (21)	2.19 (22)	0.46 (23)	1.66 (31)	3.91 (22)	

Table 2: TF, miRNA and target mRNA names of selected FFLs using model M3 are reported along with the estimated parameters, their precision in terms of CV and a flag to distinguish between coherent (C) and incoherent (I) FFLs. TF and miRNA already known to be key regulators of adipogenesis and adipocyte-related functions are marked with an x. When the estimated degradation parameter (d_2) was small and with low precision, i.e. the process was too slow to be determined in the time horizon of the experiment, it was set to 0 and model identification was repeated.

5 Discussion

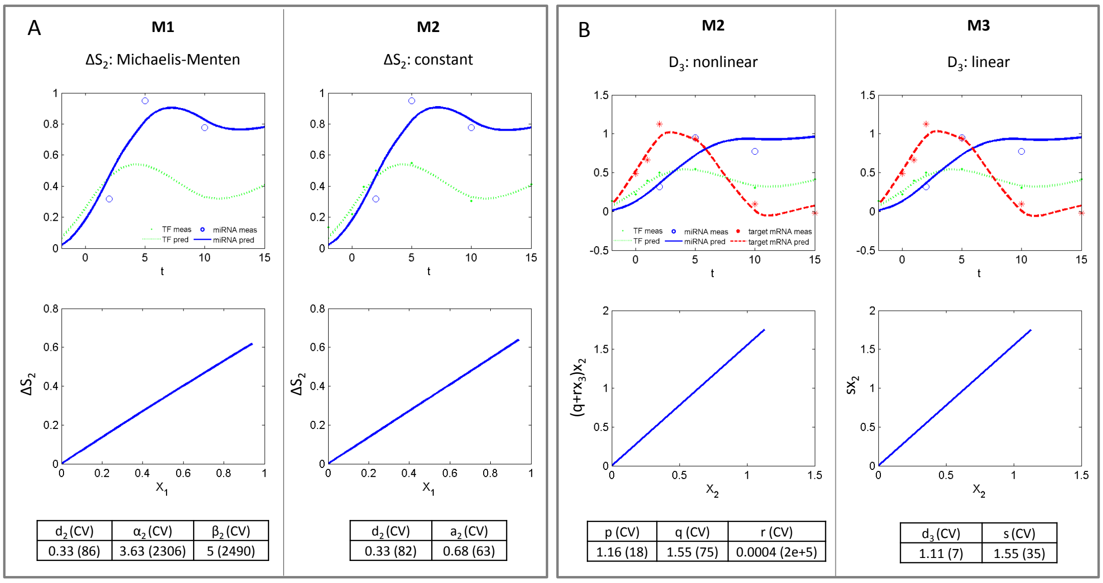
In this work we propose a method to select active FFLs from a large set of putative ones based on miRNA and mRNA expression time series, using differential equation based models and identification criteria. A list of putative mixed transcriptional and post-trancriptional FFLs is generated on the basis of conserved overrepresented motifs in human and mouse promoters and 3' UTR. Identification of three alternative dynamic models, able to describe the miRNA and target mRNA dynamic data based on ordinary differential equations (ODEs) using the TF profile as forcing function, provides the basis for the selection of active FFLs. A putative FFL is selected as active if the feed-forward topology (Fig. 1A), associated with a plausible dynamic description, is necessary and sufficient to reproduce the available gene expression profiles, i.e. the model is able to reproduce data (criterion 1), outperforming with respect to submodels in terms of principle of parsimony (criterion 3) and its parameters can be estimated with acceptable precision from available data (criterion 2).

5.1 Comparison of dynamic models

Instead of postulating a univocal description for miRNA and mRNA expression kinetics, three models of increasing complexity are proposed. Model M1 assumes Michaelis-Menten kinetics for miRNA and target mRNA regulation accomplished by the TF and models miRNA mediated degradation of the target mRNA as a first order process with constant rate plus a nonlinear term dependent on miRNA and target mRNA expression. In model M2 linearity is assumed for TF regulation on miRNA and target mRNA, whereas nonlinearity is maintained for miRNA mediated degradation of the target mRNA. In M3 also the miRNA mediated degradation of the target mRNA is linearized, thus the whole model is described by a linear kinetics. The increasing complexity of the models adapts to different type of gene expression data. The choice of the most appropriate model depends on the range and on the number of time points of the available time series and can be made using the same criteria described for the selection of active FFLs: goodness of fit, precision of the estimates and principle of parsimony. In particular, to estimate the Michaelis-Menten parameters of model M1 the whole Michaelis-Menten curve should be observable requiring expression data in an adequate range and sufficiently detailed. If these criteria are not satisfied by the available data, the linearization of the model still provide an adequate fit, allowing also a more precise estimation of the parameters. That does not mean that the more complex model is invalid, but only that the linearized one is more suitable for the available dataset.

5.2 Case study

In our case study, we used the three models on gene expression time series to select active FFLs during human adipogenesis. Since they showed a comparable ability to reproduce the data, the simplest model M3 was selected based on the principle of parsimony in 251 out of the 329 analyzed FFLs. Moreover, parameter estimates of model M1



d$_2$ (CV)	α$_2$ (CV)	β$_2$ (CV)
0.33 (86)	3.63 (2306)	5 (2490)

d$_2$ (CV)	a$_2$ (CV)
0.33 (82)	0.68 (63)

p (CV)	q (CV)	r (CV)
1.16 (18)	1.55 (75)	0.0004 (2e+5)

d$_3$ (CV)	s (CV)
1.11 (7)	1.55 (35)

Figure 5: Comparison between the candidate models. (A) upper panels: similarity between models M1 and M2 predictions for miRNA (blue) profile indicates that the Michaelis-Menten function is not necessary; lower panels: confirmation that the model prediction of the link between TF and miRNA, postulated as linear for M2, is operating in the linear range for M1; (B) upper panels: similarity between M2 and M3 predictions for target mRNA (red) profile suggests that a linear description of target mRNA degradation is sufficient; lower panel: confirmation that the miRNA mediated degradation rate, postulated as linear for M3, is operating in the linear range for M2.

were affected by very high CVs in all FFLs and those of M2 in all FFLs but 3, indicating that nonlinear models M1 and M2 were not a posteriori identifiable. Fig. 5 shows the effect of the linearization of the synthesis mediated by the TF (ΔS_2), i.e. of using model M2 instead of M1 (panel A), and of the subsequent linearization of the degradation of the target mRNA (D_3), i.e. of using model M3 instead of M2 (panel B). In particular, using the analyzed dataset, the Michaelis-Menten curve is in the linear range (Fig. 5A left panel) and model M1 is not a posteriori identifiable (α_2 and β_2 show high CVs). In this case, X_1 is much lower than the half saturation constant β_2, then parameters α_2 and β_2 cannot be separately resolved but only the ratio between the two can be essentially estimated. Conversely, using M2 the parameter related to the synthesis mediated by the TF (ΔS_2) is a posteriori identifiable (Fig. 5A right panel). Similarly, for the nonlinear description of the miRNA mediated degradation rate (Fig. 5B left panel) parameter r shows high CV and thus model M2 is not a posteriori identifiable. However, since rx_3 is much lower than q, the miRNA mediated degradation rate can be reasonably linearized as in M3 (Fig. 5B right panel) providing a simplification of the model with a reduced number of parameters and fit comparable to M2.

Analyzing the active FFLs from a biological point of view, it was found that out of the 24 selected FFLs, 9 FFLs involve TFs and 6 involve miRNAs (marked with an x in Table 1 and Table 2)) that are already known from the literature to be regulators of adipogenesis and adipocyte-related functions. A discussion of the results in comparison with the biological literature is available as Supplementary Material; however, few information is available in the literature regarding miRNA mediated FFLs involved in adipogenesis and most datasets such the ones presented in (Baroudi et al., 2011) contain mainly information related to cancer. The limited available knowledge about human transcription networks and miRNA-mediated regulations in adipogenesis makes biological validation of regulatory links difficult and, at the same time, highlights the importance of the development of algorithms, like the one presented in this work, to predict testable regulation processes.

The significance of our method was estimated in comparison with random FFLs obtained by randomly selecting links between miRNAs, TFs and target mRNA. 329 random FFLs (equal to the number of putative FFLs estimated by pairing between the 5' region of the miRNA and the 3' UTR of the target gene) were generated ten times choosing one random miRNA and two random mRNAs to play the role of the TF and the target gene respectively. The previously described selection procedure was then applied to the randomized set of FFLs obtaining an average of 15.6 selected FFLs, with a standard deviation of 1.5. This can represent a rough estimation of the number of False Positive FFLs among the 24 selected by our method. Let's note that, if instead of using differential equation based modeling, we select FFLs based on correlation between TF, miRNA and target mRNA, we select 12 FFLs on the original dataset and 18.6 ± 4.6 on the randomized datasets, thus showing the increased specificity achieved by our approach.

The presented method selects triplets that can be explained by a simple FFL, whose effect can be isolated from the rest of the network, and described by one of the three proposed models. Thus, the presence of possible additional regulatory links is not excluded by our analysis, but we can say that, for the selected FFLs, this scheme provides a minimal plausible description of the regulatory interactions. The approach presented here does not allow identifying topologies incorporating more than one TF and/or miRNA. More complex topologies will be studied in future work by extending the approach here developed; moreover, we plan to analyze dynamic descriptions that will require a tighter sampling schedule.

Acknowledgments

We thank Prof. Gérard Ailhaud, Dr. Christian Dani, and Dr. Ez-Zoubir Amri for hMADS cells.

References

Alon, U. (2007). Network motifs: theory and experimental approaches. *Nature reviews.Genetics*, *8*(6), 450–461.

Baek, D., Villen, J., Shin, C., Camargo, F. D., Gygi, S. P., & Bartel, D. P. (2008). The impact of micrornas on protein output. *Nature*, *455*(7209), 64–71.

Baroudi, M. E., Cora, D., Bosia, C., Osella, M., & Caselle, M. (2011). A curated database of mirna mediated feed-forward loops involving myc as master regulator. *PloS one*, *6*(3), e14742.

Bartel, D. P. (2004). Micrornas: genomics, biogenesis, mechanism, and function. *Cell*, *116*(2), 281–297.

Bartel, D. P. (2009). Micrornas: target recognition and regulatory functions. *Cell*, *136*(2), 215–233.

Bellu, G., Saccomani, M. P., Audoly, S., & D'Angio, L. (2007). Daisy: a new software tool to test global identifiability of biological and physiological systems. *Computer methods and programs in biomedicine*, *88*(1), 52–61.

Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., Barzilai, A., Einat, P., Einav, U., Meiri, E., Sharon, E., Spector, Y., & Bentwich, Z. (2005). Identification of hundreds of conserved and nonconserved human micrornas. *Nature genetics*, *37*(7), 766–770.

Chen, H. C., Lee, H. C., Lin, T. Y., Li, W. H., & Chen, B. S. (2004). Quantitative characterization of the transcriptional regulatory network in the yeast cell cycle. *Bioinformatics (Oxford, England)*, *20*(12), 1914–1927.

Chen, K. C., Wang, T. Y., Tseng, H. H., Huang, C. Y., & Kao, C. Y. (2005). A stochastic differential equation model for quantifying transcriptional regulatory network in saccharomyces cerevisiae. *Bioinformatics (Oxford, England)*, *21*(12), 2883–2890.

Cobelli, C., Foster, D., & Toffolo, G. (2000). *Tracer kinetics in biomedical research: from data to model*. New York: Kluwer Academic/Plenum.

Du, T., & Zamore, P. D. (2005). microprimer: the biogenesis and function of microrna. *Development (Cambridge, England)*, *132*(21), 4645–4652.

Friard, O., Re, A., Taverna, D., Bortoli, M. D., & Cora, D. (2010). Circuitsdb: a database of mixed microrna/transcription factor feed-forward regulatory circuits in human and mouse. *BMC bioinformatics*, *11*, 435.

Friedman, R. C., Farh, K. K., Burge, C. B., & Bartel, D. P. (2009). Most mammalian mrnas are conserved targets of micrornas. *Genome research*, *19*(1), 92–105.

Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A., & Enright, A. J. (2006). mirbase: microrna sequences, targets and gene nomenclature. *Nucleic acids research*, *34*(Database issue), D140–4.

Hornstein, E., & Shomron, N. (2006). Canalization of development by micrornas. *Nature genetics*, *38 Suppl*, S20–4.

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., Li, M., Wang, G., & Liu, Y. (2009). mir2disease: a manually curated database for microrna deregulation in human disease. *Nucleic acids research*, *37*(Database issue), D98–104.

Jung, C. H., Hansen, M. A., Makunin, I. V., Korbie, D. J., & Mattick, J. S. (2010). Identification of novel non-coding rnas using profiles of short sequence reads from next generation sequencing data. *BMC genomics*, *11*, 77.

Kalir, S., Mangan, S., & Alon, U. (2005). A coherent feed-forward loop with a sum input function prolongs flagella expression in escherichia coli. *Molecular systems biology*, *1*, 2005.0006.

Kaplan, S., Bren, A., Dekel, E., & Alon, U. (2008). The incoherent feed-forward loop can generate non-monotonic input functions for genes. *Molecular systems biology*, *4*, 203.

Karbiener, M., Fischer, C., Nowitsch, S., Opriessnig, P., Papak, C., Ailhaud, G., Dani, C., Amri, E., & Scheideler, M. (2009). microrna mir-27b impairs human adipocyte differentiation and targets pparγ. *Biochemical and biophysical research communications*, *390*(2), 247–251.

Khanin, R., & Vinciotti, V. (2008). Computational modeling of post-transcriptional gene regulation by micrornas. *Journal of computational biology : a journal of computational molecular cell biology*, *15*(3), 305–316.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., & Tuschl, T. (2001). Identification of novel genes coding for small expressed rnas. *Science (New York, N.Y.)*, *294*(5543), 853–858.

Lagos-Quintana, M., Rauhut, R., Meyer, J., Borkhardt, A., & Tuschl, T. (2003). New micrornas from mouse and human. *RNA (New York, N.Y.)*, *9*(2), 175–179.

Levine, E., Jacob, E. B., & Levine, H. (2007a). Target-specific and global effectors in gene regulation by microrna. *Biophysical journal*, *93*(11), L52–4.

Levine, E., Zhang, Z., Kuhlman, T., & Hwa, T. (2007b). Quantitative characteristics of gene regulation by small rna. *PLoS biology*, *5*(9), e229.

Lewis, B. P., Burge, C. B., & Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. *Cell*, *120*(1), 15–20.

Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P., & Burge, C. B. (2003). Prediction of mammalian microrna targets. *Cell*, *115*(7), 787–798.

Li, X., Cassidy, J. J., Reinke, C. A., Fischboeck, S., & Carthew, R. W. (2009). A microrna imparts robustness against environmental fluctuation during development. *Cell*, *137*(2), 273–282.

Macia, J., Widder, S., & Sole, R. (2009). Specialized or flexible feed-forward loop motifs: a question of topology. *BMC systems biology*, *3*, 84.

Mangan, S., & Alon, U. (2003). Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(21), 11980–11985.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., & Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science (New York, N.Y.)*, *298*(5594), 824–827.

Osella, M., Bosia, C., Corà, D., & Caselle, M. (2011). The role of incoherent microrna-mediated feedforward loops in noise buffering. *PLoS computational biology*, *7*(3), e1001101.

Re, A., Cora, D., Taverna, D., & Caselle, M. (2009). Genome-wide survey of microrna-transcription factor feed-forward regulatory circuits in human. *Molecular bioSystems*, *5*(8), 854–867.

Scheideler, M., Elabd, C., Zaragosi, L. E., Chiellini, C., Hackl, H., Sanchez-Cabo, F., Yadav, S., Duszka, K., Friedl, G., Papak, C., Prokesch, A., Windhager, R., Ailhaud, G., Dani, C., Amri, E. Z., & Trajanoski, Z. (2008). Comparative transcriptomics of human multipotent stem cells during adipogenesis and osteoblastogenesis. *BMC genomics*, *9*, 340.

Selbach, M., Schwanhausser, B., Thierfelder, N., Fang, Z., Khanin, R., & Rajewsky, N. (2008). Widespread changes in protein synthesis induced by micrornas. *Nature*, *455*(7209), 58–63.

Shalgi, R., Lieber, D., Oren, M., & Pilpel, Y. (2007). Global and local architecture of the mammalian microrna-transcription factor regulatory network. *PLoS computational biology*, *3*(7), e131.

Shen-Orr, S. S., Milo, R., Mangan, S., & Alon, U. (2002). Network motifs in the transcriptional regulation network of escherichia coli. *Nature genetics*, *31*(1), 64–68.

Shimoni, Y., Friedlander, G., Hetzroni, G., Niv, G., Altuvia, S., Biham, O., & Margalit, H. (2007). Regulation of gene expression by small non-coding rnas: a quantitative view. *Molecular systems biology*, *3*, 138.

Tsang, J., Zhu, J., & van Oudenaarden, A. (2007). Microrna-mediated feedback and feedforward loops are recurrent network motifs in mammals. *Molecular cell*, *26*(5), 753–767.

Vohradsky, J., Panek, J., & Vomastek, T. (2010). Numerical modelling of microrna-mediated mrna decay identifies novel mechanism of microrna controlled mrna downregulation. *Nucleic acids research*, *38*(14), 4579–4585.

Vu, T. T., & Vohradsky, J. (2007). Nonlinear differential equation model for quantification of transcriptional regulation applied to microarray data of saccharomyces cerevisiae. *Nucleic acids research*, *35*(1), 279–287.

Wu, C. I., Shen, Y., & Tang, T. (2009). Evolution under canalization and the dual roles of micrornas: a hypothesis. *Genome research*, *19*(5), 734–743.

5

Multilevel study of insulin signalling pathway

5.1 Introduction

Defects in signalling pathways involved in insulin action are associated with insulin resistance and thus with obesity and type 2 diabetes (T2D). T2D is characterized in by a defective responsiveness of tissues to insulin stimulation and is one of the main causes of mortality worldwide (Saltiel & Kahn, 2001). 346 million people are estimated to have T2D as of 2012: these people are likely to have long-term complications including heart diseases, strokes, diabetic retinopathy, kidney failure and poor circulation of limbs which may lead to amputations. Insulin signalling pathway is very important for the cell because, when insulin reaches the cell surface, it binds to the insulin receptor (IR) triggering a complex cascade of signals that involves different proteins and culminate in several important biological responses, as: protein synthesis, fatty acids synthesis, glucose utilization, glycogen synthesis, cell growth, proliferation and differentiation. Thus, the understanding of insulin signalling pathway could lead to a better understanding of the pathophysiology of insulin resistance, and the identification of key molecules and processes could lead to newer and more effective therapeutic agents for treating these common disorders that are already an uprising epidemic (LeRoith et al., 2003).

The aim of this work is the design of an experimental study (Section 5.2) and the application of modeling techniques to improve the knowledge of insulin signaling pathway in order to have a better understanding of its underlying mechanism. In particular, we focus on the analysis of the effects of branched-chain aminoacids (as leucine), that can be orally supplied and are know to interact with insulin stimulated signal transduction. However, insulin signalling is complex, involving post-transcriptional modification of many molecules on multiple sites and including different feedback loops and cross-talks with other pathways. For this reasons, we analyze the pathway on different levels of detail: we start retrieving information from literature about the topology of signalling network (Section 5.3) and we then apply two different modelling techniques (Sections 5.4 and 5.5):

- *semi-qualitative*: the first approach is based on logic models and is used to have a comprehensive view of the pathway and its behavior;
- *quantitative*: in this second approach we derive ODEs from mass-action kinetic rules limiting the analysis to a particular regulatory mechanism.

These approaches will be explained more in detail in Section 5.4 and Section 5.5 respectively.

5.2 Experimental setup

Three independent cell culture experiments (set A, B and C) were performed as biological replicates on a commercially available cell line of human skeletal muscle myoblasts that had been differentiated into myotubes. Cells were grown to 85% confluence using DMEM 10% FBS as growth medium. The medium was switched to serum-free DMEM overnight before stimulus and replaced with amino acid-free medium (EBSS) 1 hour before stimulus. Cells were harvested by addition of lysis buffer and scraping at 6 time points (0, 2, 5, 10, 30 and 60 minutes) after stimulus to monitor both the early and the late insulin signaling effects. Cell lines were exposed to three different stimuli, as shown in Figure 5.1:

1. INS: cells were stimulated with insulin (EBSS $+100nM$ insulin) at time 0;
2. LEU: cells were stimulated with leucine (EBSS $+400\mu M$ leucine) at time 0;
3. LEU+INS: cells were pre-incubated with leucine (EBSS $+400\mu M$ leucine) for one hour and then stimulated with insulin (EBSS $+400\mu M$ leucine $+100nM$ insulin) at time 0.

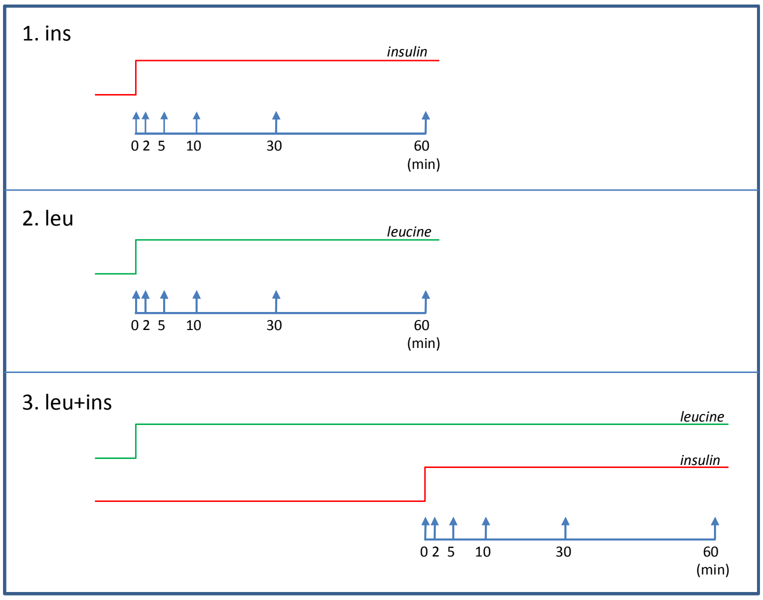


Figure 5.1: Experimental setup

Protein	Antibody
$4EBP1$	$p4EBP1$ (T37,T46)
AKT	$pAKT$ (S473)
$ERK12$	$ppERK12$ (T202,Y204)
$GSK3\beta$	$pGSK3\beta$ (S9)
$mTOR$	$pmTOR$ (S2448)
$P70S6K$	$pP70S6K$ (T389)
$FOXO1$	$pFOXO1$ (S256)

Table 5.1: Measured proteins and detected phosphorylation sites

Leucine and insulin levels were monitored to ensure they remain constant during the experiments.

Western blot

Measured proteins are: $4EBP1$, AKT, $ERK12$, $GSK3\beta$, $mTOR$, $P70S6K$, $FOX01$. In the Table 5.1, measured proteins along with the specific phosphorylation site that is detected by the used antibody are listed.

For each biological replicate (set A, B and C) of each protein, the complete time course for all three stimuli was run on the same blot when possible. Measures were then repeated on different days and different blots to check technical reproducibility.

Band intensities were analyzed by using the Odyssey infrared image system (LiCor). This system, based on infrared detection, permits probing the lysate with the antibodies for both the total and the phosphorylated amount of the protein using secondary antibodies with two different dyes. The housekeeping protein, Beta-actin was measured and used as a loading control. A commercially available lysate was loaded onto each gel as an internal standard but did not always provide a detectable signal for the target proteins. Therefore band densities in each gel were normalized to that of the pre-stimulus 0' time point in order to be able to compare data from different gels and replicates were then mediated. Both the total and phosphorylated time-series data are shown in Figure 5.2 for all proteins under all three experimental conditions.

Linear relationship between Western blot signals and protein concentrations for selected antibodies was experimentally verified. Experimental error model was derived from data providing an estimated standard deviation equal to 0.25 times the associated measure.

Glycogen synthesis assay

For all experimental conditions the newly synthesized glycogen was measured as the amount of 14C-labeled glucose incorporated into glycogen over 90 minutes. Differentiated myotubes were cultured in six-well plates. They were serum-starved with DMEM O/N and then switched to EBSS for 1 hour, after which the cultures were exposed to one of the three media: 1) EBSS $+100nM$ insulin, 2) EBSS $+400\mu M$ leucine, or 3) EBSS $+100\mu M$ insulin $+400\mu M$ leucine for 1 hour priming. Priming media were removed and media a-c containing uniformly labeled 14C-glucose was added to the cultures. After 90 minutes, medium was removed, cells were washed with PBS and cells were extracted with 1 ml 0.03% SDS: $0.15ml$ was used for protein determination and $0.85ml$ was used for glycogen extraction. Carrier glycogen was added, samples were heated to $95°C$ for 30 minutes and glycogen pellets were collected by centrifugation, washed with 70% ethanol, and resuspended in $200\mu l$ distilled water. The incorporation of 14C-labeled glucose into glycogen was determined by 15 minute counts in a beta counter, using $10ml$ optifluor liquid scintillation cocktail per sample. Each condition was run at least in triplicate and are expressed relative to the basal value (no insulin/leucine stimulus).

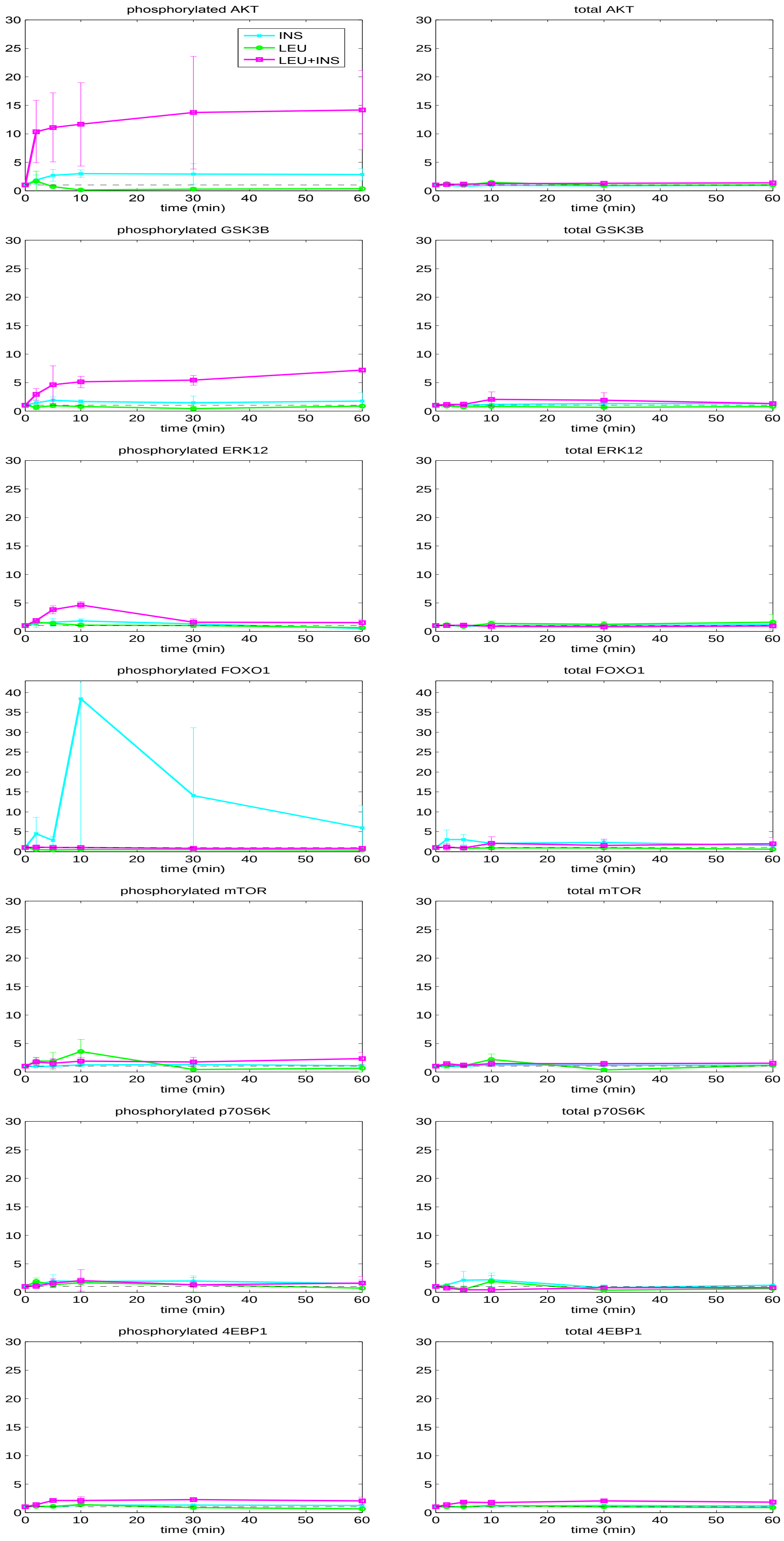


Figure 5.2: Plot of the experimental time series data for all three experimental conditions: insulin stimulation, leucine stimulation and insulin stimulation after perincubation with leucine.

5.3 Insulin signalling network topology

Signalling pathways are collected in public manually curated databases like KEGG (Kanehisa et al., 2010), Reactome (Joshi-Tope et al., 2005), WikiPathways (Pico et al., 2008), PID (Schaefer et al., 2009), which collect pathway data from multiple organisms and tissues; it is not easy to have a complete view of the pathway since comprehensive knowledge is fragmented among multiple sources. Moreover, it is important to consider that signalling pathways are not isolates since they are involved in important interactions, called cross-talks, with other pathways. In order to have an overview of information available about insulin related pathways, we used Pathway Commons (Cerami et al., 2011a), a repository that integrates publicly available biological pathway and molecular interaction data from nine public databases. For insulin pathway, Pathway Commons refers to the Reactome network shown in Figure 5.3. This network includes detailed information about known kinetic reactions between molecules; for all species, all possible (or interesting) states are included (e.g. phosphorylated/unphosphorylated) and complexes that can be formed are shown as different molecular species. Another schematic representations of the pathway, which include a higher level of simplification, is the one shown in Figure 5.4. This network does not go into the mechanistic detail of the interaction and only one node is considered for each species, but it gives a better idea of causal relationships between nodes of the network being more intuitive.

As shown in Figure 5.4, signalling response to insulin is activated by the insulin receptor (IR) and involves two main pathways: the *PI3K-AKT pathway* (in red in the Figure) and the *MAPK pathway* (in blue). The *PI3k-AKT pathway* is aided in its action by another pathway, the *Cbl/CAP pathway* (in purple). Their cooperation regulates the main insulin action, such as glycogen synthesis and glucose transport, while *MAPK pathway* is a more general pathway that can be activated by different growth factors, all leading to enhanced cell growth. Insulin signalling response is particularly complex as it includes many cross-talks, where signalling molecules are shared among pathways, and a multitude of negative and positive feedback loops which play a key role in the control of insulin sensitivity. The most important action is played by the negative feedback which emanates from AKT and, throught the formation of mTOR complex and the activation of P70S6K, results in serine phosphorylation and inactivation of IRS signalling.

It is important to notice that literature-derived interactions collected in these databases are typically derived from publications using different experimental conditions and cell types. However not all interactions are equally important or present in all cell types; this is taken into account when the network is interpreted as a model and fitted to our experimental data in the following Paragraphs.

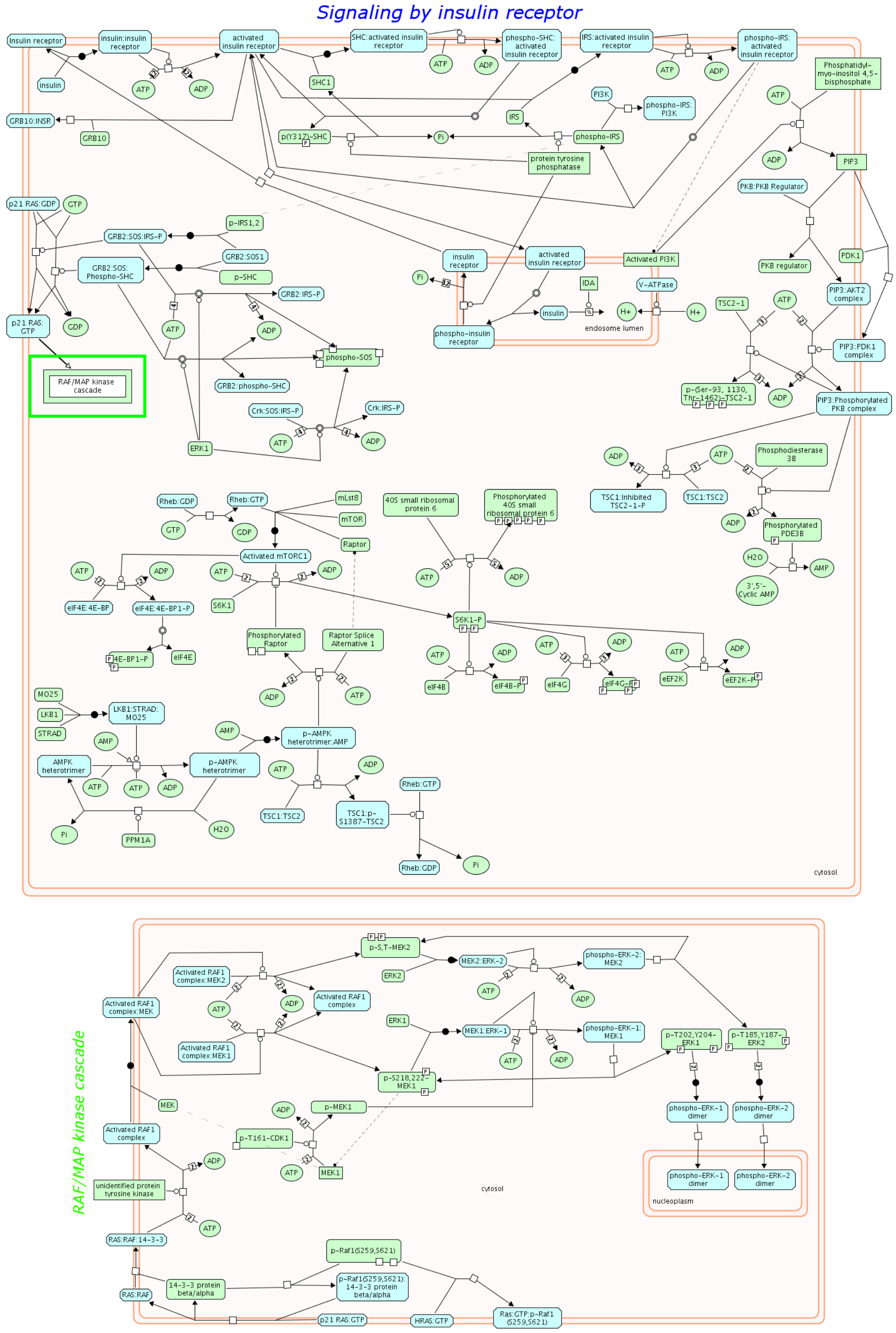


Figure 5.3: REACTOME pathways (Joshi-Tope et al., 2005): *Signaling by insulin receptor* and *RAF/MAPK cascade*

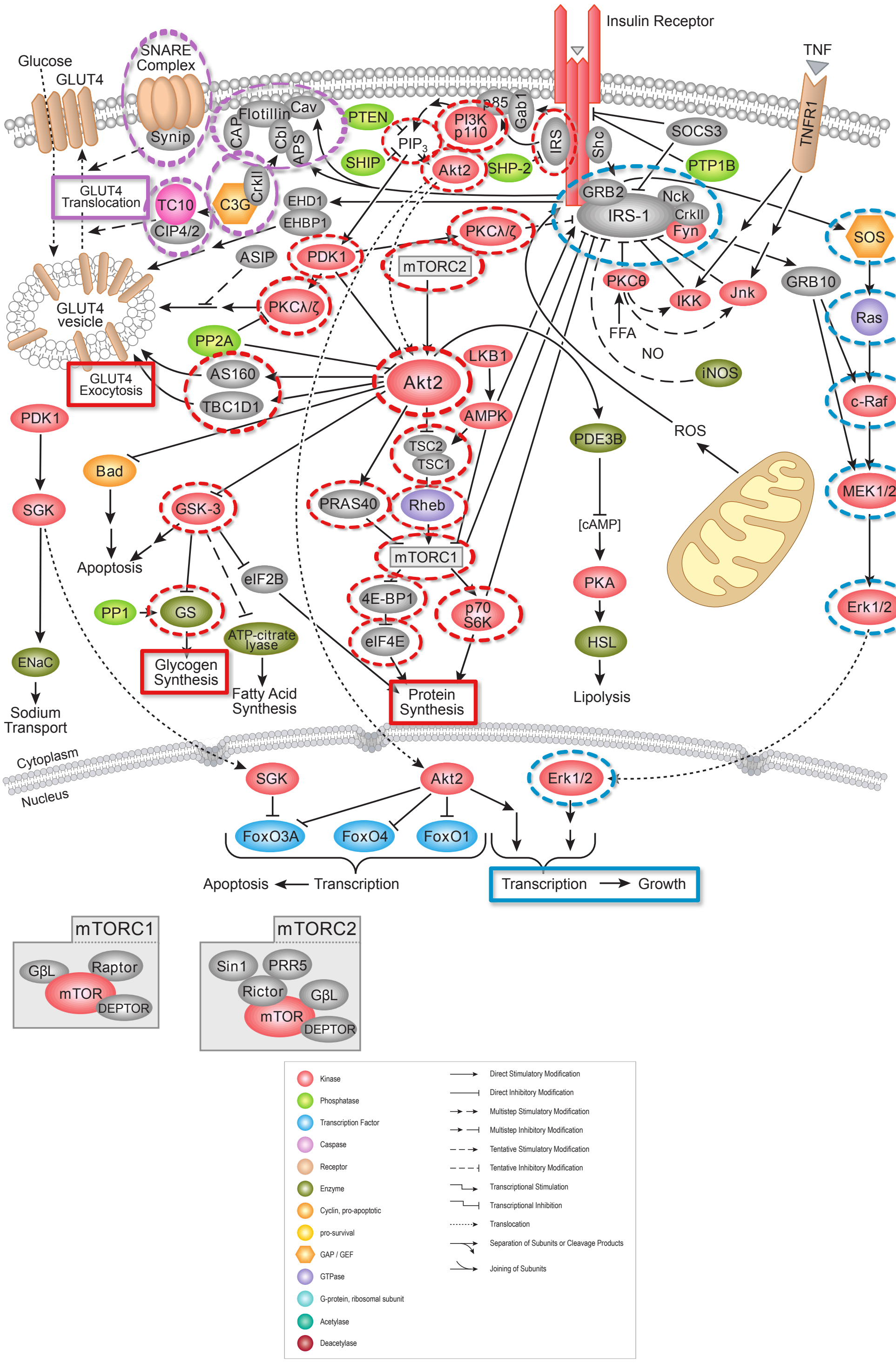


Figure 5.4: Cell Signaling Technology® pathway (CellSignalingTechnology, 2013): *Insulin receptor Signaling* where different sub-pathway are highlighted: *PI3K-AKT* pathway (in red), *MAPK* pathway (in blue), *CBL-Cap* pathway (in purple)

5.4 Semi-qualitative modelling using logic models

Basic theory of logic-based ordinary differential equations

The formalism used to derive the logic based ODEs was developed in (Wittmann et al., 2009) in order to obtain a compromise between:

- *Boolean models*: which cannot describe the continuous behavior of biochemical processes, but permit to describe large scale signalling networks
- *Mass action based models*: which accurately describe the underlying biochemistry but are typically limited to a small well studied system

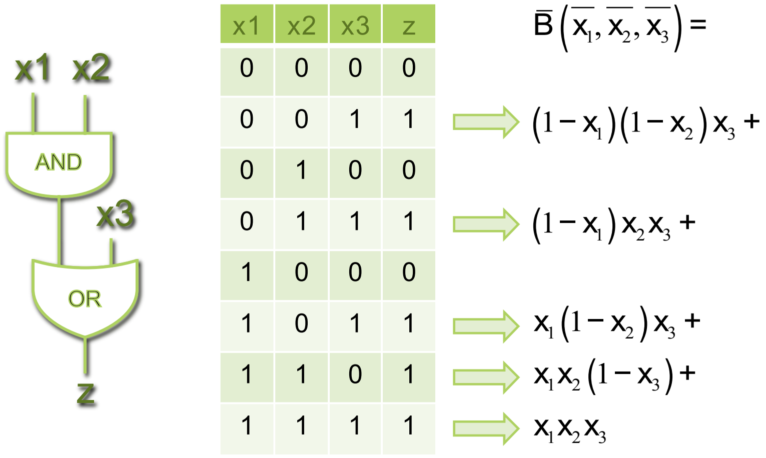
Signalling networks, as described in the previous section, can be easily interpreted as logic Boolean models simply by encoding causal relationships in states of individual components (state variables): species can assume an active (TRUE or 1) or inactive states (FALSE or 0) and links can then be interpreted as logical operator AND, OR and NOT (Blinov & Moraru, 2012). However, Boolean models are generally limited to discrete (typically two) states and discrete time, since state of species at time $t+1$ is a function of states at time t. Logic-based ODEs overcome this limitations by allowing the construction of continuous model from a qualitative (Boolean) knowledge, being suitable for application to our case study where time course measures are available, with no need for additional knowledge on the biochemestry.

The formalism commonly use in boolean models is the following:

- $X_1, X_2, \ldots, X_N$ are the N species (proteins in our case) each represented by a variable x_i taking values in $\{0,1\}$
- $R_i := \{X_{i1}, X_{i2}, \ldots X_{iN_i}\} \subset \{X_1, X_2, \ldots X_N\}$ is, for each species X_i, the set of species that influence x_i
- $B_i : \{0,1\}^{N_i} \to \{0,1\}$ is, for each species X_i, the update function giving the value of x_i at the next time step for every possible combination of $(x_{i1}, x_{i2}, \ldots, x_{iN_i}) \in \{0,1\}^{N_i}$

Logic-based ODE models are derive from Boolean models by considering:

- continuous variables $\overline{x_i}$, taking values in $[0,1]$, instead of the discrete variables x_i
- functions $\overline{B_i} : [0,1]^N \to [0,1]$ as continuous homologues of Boolean functions B_i



x1	x2	x3	z
0	0	0	0
0	0	1	1
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1

$$\overline{B}\left(\overline{x_1},\overline{x_2},\overline{x_3}\right) = \left(1-x_1\right)\left(1-x_2\right)x_3 + \left(1-x_1\right)x_2x_3 + x_1\left(1-x_2\right)x_3 + x_1x_2\left(1-x_3\right) + x_1x_2x_3$$

Figure 5.5: Example of derivation of the continuous update function $\overline{B_i}$ for a simple logic model using a Boolean table encoding the discrete update function B_i via multilinear interpolation

At this point the ODE can be defined as:

$$\dot{\overline{x_i}} = \frac{1}{\tau_i}\left(\overline{B_i}\left(\overline{x_{i1}},\overline{x_{i2}},\ldots,\overline{x_{iN_i}}\right) - \overline{x_i}\right) \tag{5.1}$$

where τ_i can be interpreted as the life-time of species X_i, and $\overline{B_i}$ can be computed as a multilinear interpolation of the Boolean function B:

$$\overline{B_i}\left(\overline{x_1},\overline{x_2},\ldots,\overline{x_N}\right) := \sum_{x_1=0}^{1}\sum_{x_2=0}^{1}\cdots\sum_{x_N=0}^{1}\left[B\left(x_1,x_2,\ldots,x_N\right)\cdots\prod_{i=1}^{N}\left(x_i\overline{x_i}+(1-x_i)(1-\overline{x_i})\right)\right] \tag{5.2}$$

An example of how the continuous homologues $\overline{B_i}$ is derived from the corresponding Boolean function B_i (interpreted as a Boolean table) via multilinear interpolation, is shown in Figure 5.5 for a simple logic model where the species represented by the state variable z is regulated as a function of x_1, x_2 and x_3 thrugh AND and OR logic gates.

Since molecular interactions are known to show a switch-like behavior, sigmoid shaped Hill functions are generally preferred considering, instead of each state variable x, the corresponding function:

$$f(\overline{x}) = \frac{\overline{x}^n}{\overline{x}^n + k^n} \tag{5.3}$$

where parameters n and k have clear biological meanings:

- n: is the Hill coefficient the slope of the curve, representing the cooperativity of the interactions;

- k: is the value at which the activation is half maximal, representing the threshold at which the state of a species is “on” or “off” in the Boolean model.

This paragraph is meant only as a basic introduction of logic-based ODEs, for complete understandig of the formalism reading of (Wittmann et al., 2009) and (Krumsiek et al., 2010) is recommended. The formalism was also implemented in the MATLAB toolbox Odefy (Krumsiek et al., 2010).

For the following analysis, we will use the formalism as implemented in the CNORode package. CNORode, developed as an add-on to the core package CellNOptR as explained in Terfve et al. (2012), was used to perform model training. As previously mentioned, CellNOptR is an open source R/Bioconductior package that extends the method to train logic models to signalling networks presented in Saez-Rodriguez et al. (2009) to different logic formalism, one of which is the logic-based ODEs approach. When the structure of the network and the available time-course data are give as input, CNORode interprets the network as logic-based ODEs model ad trains the parameters to fit the data using global optimization algorithms.

Logic-based ODE model of insulin signalling pathway

The network structure derived from a priori information retrieved from literature, as explained in Section 5.3, is then compressed and interpreted as a logic model in order to obtain the model shown in Figure 5.5. The compression step consists in the removal of nodes that are not identifiable from the available experimental setup, in order to obtain a simplified network. For example, the causal relationship $AKT \rightarrow TSC \rightarrow Rheb \rightarrow mTOR$ is simplified to $AKT \rightarrow mTOR$ because no intermediate measures are available and $mTOR$ has no other input nodes that could change the logic of the network. IRS node is preserved even if not measured because it represents an important crossroad of signals. The network is then interpreted as a logic model including a NOT operator when there is a negative regulation (e.g. $P70S6K$ negatively regulates IRS). To understand the reason of the AND gate in IRS feedback, it is convenient to explain the basic biology that stays beneath this mechanism: it is known that protein IRS is phosphorylated on a tyrosine residue as a consequence of insulin stimulation of the insulin receptor, this phosphorylation causes its activation starting the downstream signalling cascade. This signal eventually reaches $P70S6K$ that, when activated, promotes serine phosphorylation of IRS contrasting its activation by obstructing tyrosine phosphorylation. This effect is well described by the fourth column of the truth table in Table 5.2 where, considering the AND ($\wedge$) gate, protein AKT is active only when ins is active ($= 1$) and $P7S6K$ is

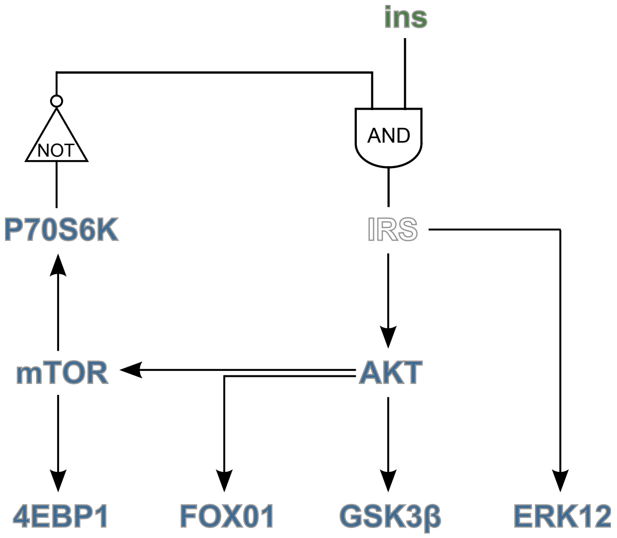


Figure 5.6: Logic model of insulin signalling pathway: insulin stimulus is represented in green and measured proteins in blue

not $(= 0)$. A different behavior would be obtained by considering an OR $(\vee)$ gate (fifth column in the truth table).

Logic model in Figure 5.5 is interpreted, using the logic-based ODEs formalism, by the following differential equations:

$$\begin{aligned}
&ins = u && \\
&IRS = x_1 && \dot{x}_1 = \tau_1 \left[(f_{u1}(u))(1 - f_{71}(x_7)) - x_1\right] \\
&AKT = x_2 && \dot{x}_2 = \tau_2 \left[(f_{12}(x_1)) - x_2\right] \\
&FOXO1 = x_3 && \dot{x}_3 = \tau_3 \left[(f_{23}(x_2)) - x_3\right] \\
&GSK3\beta = x_4 && \dot{x}_4 = \tau_4 \left[(f_{24}(x_2)) - x_4\right] \\
&mTOR = x_5 && \dot{x}_5 = \tau_5 \left[(f_{25}(x_2)) - x_5\right] \\
&4EBP1 = x_6 && \dot{x}_6 = \tau_6 \left[(f_{56}(x_5)) - x_6\right] \\
&P70S6K = x_7 && \dot{x}_7 = \tau_7 \left[(f_{57}(x_5)) - x_7\right] \\
&ERK12 = x_8 && \dot{x}_8 = \tau_8 \left[(f_{18}(x_1)) - x_8\right]
\end{aligned} \tag{5.4}$$

where

$$f_{ij}(x_j) = \frac{x_j^{n_{ij}}}{x_j^{n_{ij}} + k_{ij}^{n_{ij}}} \tag{5.5}$$

and n_{ij} and k_{ij} are the unknown parameters.

Data were normalized between 0 and 1 using the procedure described in (Saez-

$\boldsymbol{ins}$	$\boldsymbol{P70S6K}$	$\neg P70S6K$	$\boldsymbol{IRS} = ins \wedge (\neg P70S6K)$	$\boldsymbol{IRS} = ins \vee (\neg P70S6K)$
0	0	1	0	1
0	1	0	0	0
1	0	1	1	1
1	1	0	0	1

Table 5.2: Truth table of IRS with inputs ins and $P70S6K$. Both the AND ($\wedge$) and the OR ($\vee$) case are show. $\neg$ represents the NOT operator.

Rodriguez et al., 2009) and implemented in the DataRail MATLAB toolbox (Saez-Rodriguez et al., 2008): a set of thresholds is used to obtain a non-linear data normalization aimed at providing a balanced normalization of small highly reproducible differences without overemphasizing outliers. Since phosphorylated proteins stimulated only with leucine (LEU experimental condition) remain constant over time, the comparison is focused on the differences between the estimated parameters in the conditions with ($LEU + INS$) and without (INS) leucine preincubation.

CNORode (Terfve et al., 2012) was used to interpret the model and estimate the unknown parameters form data using Scatter Search, an evolutionary optimization algorithm described in (Egea et al., 2010). For both experimental conditions (INS and $LEU + INS$), the estimation procedure was repeated 10 times with different initial values for the parameters in order to have an idea of the confidence of the estimated parameters.

Results

Results of model identification are shown is Figure 5.7 where, for each of the 26 parameters of the logic-based ODE model, in each experimental condition (INS and $LEU + INS$) estimates over 10 runs are represented as box plots. The height of each box can be interpreted as a qualitative index of the precision of the parameters: a smaller box corresponds to a parameter for which a similar value is estimated over different runs, thus we have more confidence in the estimated value. A larger box correspond to a parameter which was not estimated with good precision (not a posteriori identifiable). Most parameters are estimated with reasonable precision, but there are some parameters, mainly the time constants that influence both the rate of production and degradation (τ), which cannot be precisely estimated from the available data.

These sets of parameters are able to well describe the experimental data as shown in Figure 5.8. For effect of the negative feedback, we can notice that IRS, and consequently all the downstream proteins, has a pick of activity followed by the return to a new

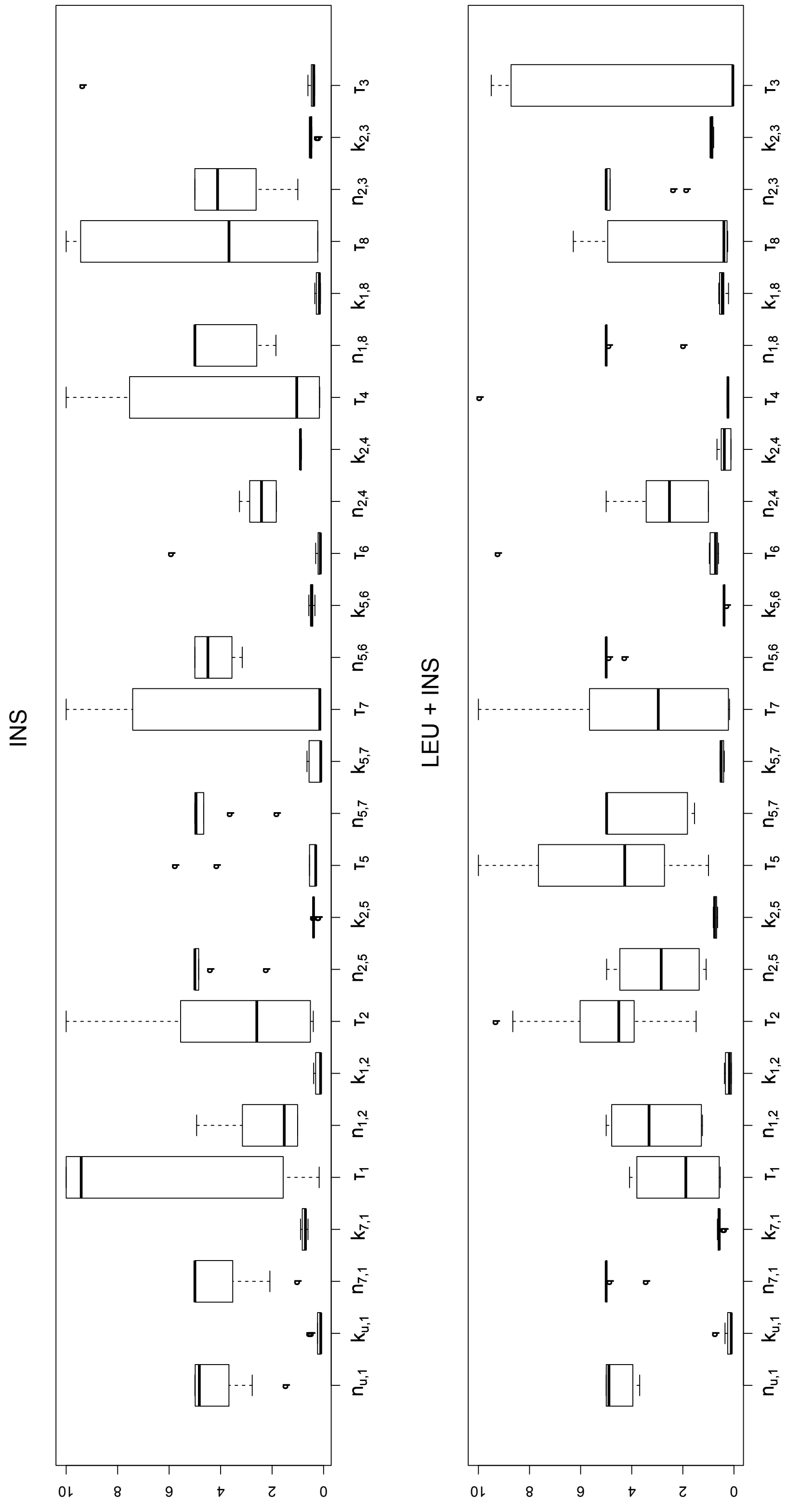


Figure 5.7: Box plots of parameters estimates for INS and $LEU + INS$ cases over 10 runs.

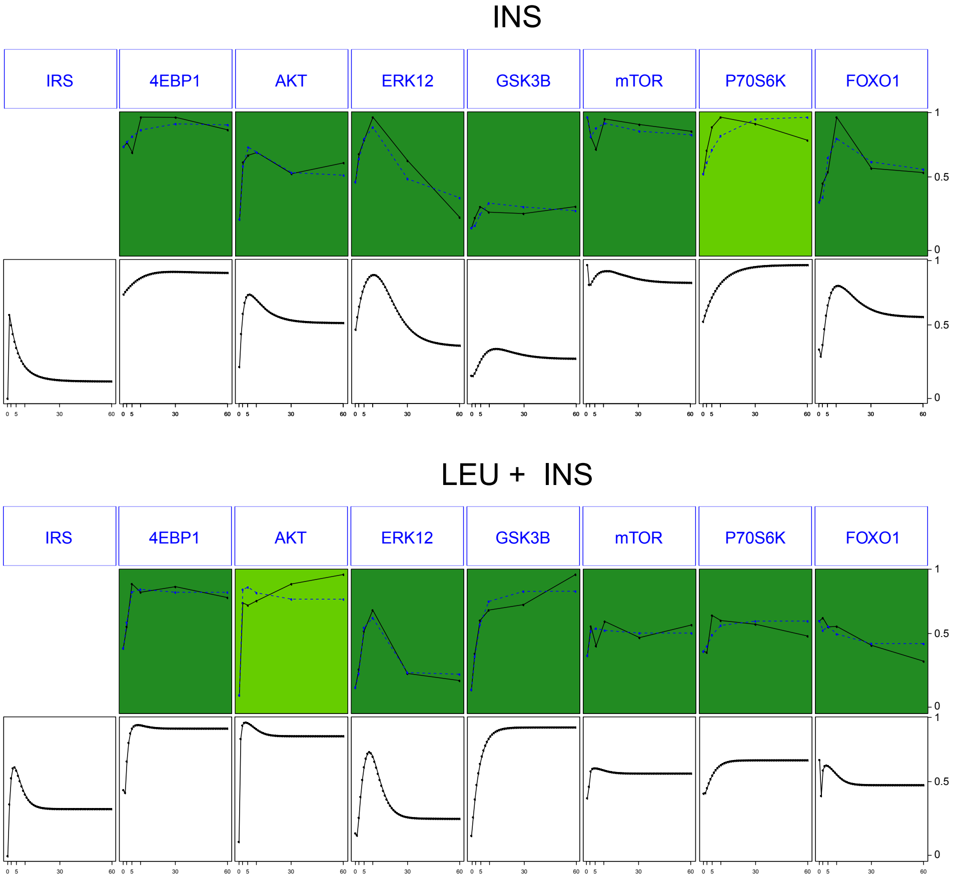


Figure 5.8: Fit of the logic-based ODE model of insulin signalling to experimental data for INS and $LEU + INS$ cases: upper panels show the normalized experimental data used for training along with the corresponding simulated data, lower panels show the time course of simulations.

basal value when that active signal reaches $P70S6K$, that consequently deactivates IRS. This qualitative behavior was observed also in Sedaghat et al. (2002) where a mathematical model of the insulin responsive glucose transport was implemented and used for quantitative simulation. After leucine preincubation, the new basal level reached by IRS seems to be slightly higher than the one reached only with insulin stimulation and this is reasonable since leucine is know to act on the feedback mechanisms, even if with unclear effects (Zeanandin et al., 2012; Macotela et al., 2011; Tremblay et al., 2007). Another protein which behavior seems to be particularly affected by leucine preicubation is $GSK3\beta$, we will focus on the analysis of its regulation with a more mechanistic approach in the following paragraph. To better understand which parameters are mainly affected by leucine preincubation, a t-test was performed for all parameters to verify if their distributions are significantly different in the two analyzed experimental

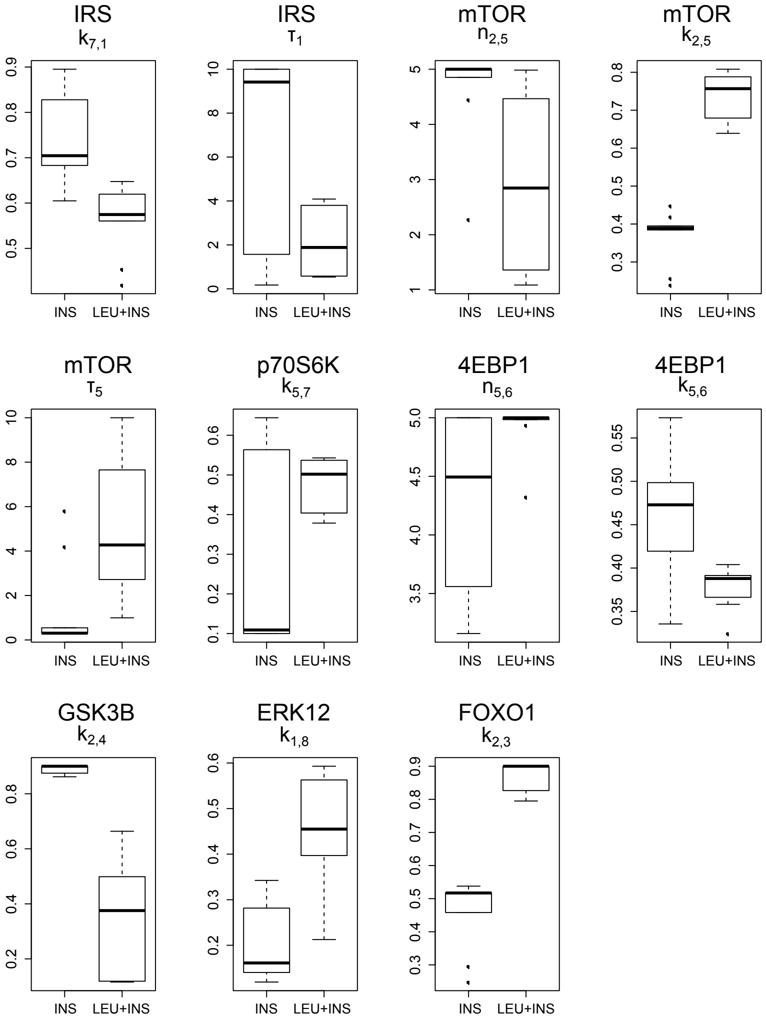


Figure 5.9: Comparison of parameters significantly different between the two experimental condition INS and $LEU + INS$ ($\text{p}-\text{val} < 0.05$)

conditions. Figure 5.9 shows parameters for which the test reported a p-value smaller than 0.05.

Conclusions

Logic-based models have been applied to analyze the insulin signalling pathway and they proved to be a suitable tool to provide a global view of the behavior of the network, without going too much into the mechanistic detail. Logic-based ODEs were used to provide continuous simulations that are able to well describe the experimental data, and to show patters that are in agreement with the state of the art literature (Sedaghat et al., 2002). In particular, the model was used to compare the pathway under two different experimental conditions: insulin stimulation and preincubation with leucine followed by insulin stimulation. The estimation of parameters in both conditions allows to highlight affected regulation processes; some of the parameters that are observed to change between the two conditions are related to processes already connected with defects in insulin signal transduction, and in particular with insulin resistance (as $P70S6K$ mediated regulation of IRS), while others correspond to regulatory mechanisms which have not been studied in detail in this context, but could be of interest (as AKT mediated regulation of $GSK3\beta$). Not all parameters where shown to be estimated with good precision and additional tests could be done, for example, by fixing some parameters. However, a quantitative analysis is out of the purpose of this work, that is meant to be only a qualitative screening of the behavior of the network with and without leucine preincubation, in order to have an idea of which mechanisms could be studied further going more into the mechanistic detail (see Section 5.5 and Appendix 5.1).

5.5 Quantitative modelling using ODEs

From the previous qualitative analysis of the global insulin signalling network, the $AKT - GSK3\beta$ regulation, was shown to be potentially important for the understanding of leucine effect. We will now focus on a more accurate description of this portion of the pathway using a more mechanistic approach based on ODEs. This interaction plays a particularly important role in signalling response to insulin stimulation, being involved in the regulation of glycogen synthesis. Glycogen synthesis, or glycogenesis, is the process in which glucose is converted into glycogen for storage, in order to decrease the glucose level in the blood after meal. In this work we derive an ODE model from mass action kinetics and apply it to the study of the dynamic response of $GSK3\beta$. Two descriptive indices are defined to characterize the sensitivity and the swiftness of the response of the system, considered as a block isolated from the upstream signalling pathway, and are used to quantitatively evaluate the effects of leucine. $GSK3\beta$ activity is then connected to glycogen synthesis at steady state.

The method and results obtained from its application are described in detail in Eduati et al. (Submitted); full text of the original manuscript is reported in Appendix 5.1.

Appendix 5.1 Paper: Eduati et al., Submitted

The following manuscript dealing with quantitative modelling of insulin signalling have been coauthored by the Ph.D. candidate during her doctoral program.

- F. Eduati, B. Di Camillo, G. Toffolo. *Dynamic analysis of leucine effects on insulin activated Akt/GSK3β signalling pathway in human skeletal muscle cells*. Submitted.

Full text of the original manuscript is reported in this Appendix formatted as submitted to the journal.

Modelling leucine effects on insulin activated Akt/GSK3β signalling dynamics in human skeletal muscle cells

F. Eduati[1], B. Di Camillo[1], E. Murphy[2,3], S. Nair[2], A. Avogaro[3] and G. Toffolo[1,*]

[1]Department of Information Engineering, University of Padova, Padova, Italy
[2]Endocrine Research Unit, Mayo Clinic, Rochester, MN, USA
[3]Department of Clinical and Experimental Medicine, University of Padova, Padova, Italy

[*]*Corresponding author:* toffolo@dei.unipd.it

Abstract

$AKT/GSK3\beta$ signalling pathway plays an important functional role in the transduction of insulin signalling and, in particular, in the regulation of glycogen synthesis, in facts dysfunctions in insulin stimulated pathways are associated with metabolic disorders and insulin resistance. Branched-chain aminoacids, as leucine, are known to interact with this pathways, but their role on glucose metabolism is still unclear. In this paper, we study of effects of leucine on the dynamics of $GSK3\beta$ phosphorylation using experimental data of AKT and $GSK3\beta$ phosphorylation measured on skeletal muscle cells in three different experimental conditions: insulin stimulation, leucine stimulation or insulin stimulation after preincubation with leucine. To this purpose, we derive a simple mathematical model of protein phosphorylation, aimed to measure descriptive indices, useful to compare the dynamics of the pathway under different experimental conditions. The model is derived from mass action kinetics and refined based on the experimental data. Two quantitative indices, namely steady state and rise time, are defined related to sensitivity and swiftness of the response of the system. Our analysis reveals that leucine preincubation affects both $GSK3\beta$ phosphorylation and dephosphorylation, amplifying the response of the system to insulin stimulation (the reached steady state goes from 1.63 to 5.82), but slowing the dynamics (the rise time goes from 0.10 to 2.99). The increase in $GSK3\beta$ phosphorylation is shown to lead to improved glycogen synthesis being of potential interest for insulin-resistant states.

1 Introduction

Insulin signalling pathway plays an important role in cellular homeostasis as it modulates the response to insulin stimulation promoting glucose uptake from the blood. In particular, glycogen synthesis, or glycogenesis, is the process of converting glucose into glycogen for storage decreasing the blood glucose level after meal. Glycogen stored in liver is then available for liver itself as well as for the rest of the body, while glycogen stored in a muscle is available only for the muscle itself. As schematized in Figure 1, this process is controlled by the enzyme *glycogen synthase* which, in its active form *glycogen synthase a*, is responsible for adding UDP-glucose to a growing chain of glycogen. The phosphorylation of this enzyme transforms it in its inactive form *glycogen synthase b*. There are two different kinases that phosphorylate this enzyme: *protein kinase A (PKA)*, which is regulated by glucagon (and not affected by insulin), and *GSK3β* which is part of the *Akt* (or *protein kinase B*) mediated insulin signalling pathway (LeRoith et al., 2003). Insulin is secreted when glucose concentration in the blood increases and it stimulates glycogen synthesis through insulin signalling pathway in order to store it as glycogen. *GSK3* is constitutively active being phosphorylated in the tyrosine residue (Tyr^{279} in the α-isoform and Tyr^{216} in the β-isoform) but it is dephosphorylated (thus, deactivated/inhibited) by serine phosphorylation of a single residue the N-terminus (Ser^{21} in the α-isoform and Ser^{9} in the β-isoform) when stimulated with insulin (Cohen, 1999).

Type 2 diabetes (T2D), that affect 90% of diabetics, may arise from defects in signal transduction (Cozzone et al., 2008). Dysfunctions in components of insulin signalling are associated with many metabolic disorders and are linked to insulin resistance and T2D, as reviewed in (Bjrnholm & Zierath, 2005). People affected by this disease at early stage can produce and secrete insulin, but are resistant to this hormone. This means that a higher production of insulin is needed to sort the desired glucose uptake and storage and, as a long term effect, this leads to complications that culminate in the failure of pancreas. One of the main potential target for treating diabetes is *GSK3* (Cohen & Goedert, 2004; Meijer et al., 2004; McManus et al., 2005) since its inhibition allows to promote activation of glycogen synthase mimicking the effect of insulin. Another particularly important kinase in this context, is *Akt* that is known to play a central role in the physiology and pathology of different signalling pathways (Hers et al., 2011; Taniguchi et al., 2006).

Strategies to treat insulin resistance include both pharmacologic interventions (Moller et al., 2001; Inzucchi, 2002) and lifestyle (dietary) modifications. In this context, increasing interest is devoted to the study of the effects of branched-chain amino acids (BCAAs), essential aminoacids that cannot be synthesized by the human body thus they must be supplied in the diet (generally as component of proteins). In particular, leucine, a BCAA, was shown to play a regulatory role on intracellular signalling and insulin sensitivity (Kimball & Jefferson, 2006; Nair & Short, 2005; Macotela et al., 2011); however, leucine effect on insulin signalling pathway is controversial. On one hand, leucine is known to activate *mTOR* pathway, promoting protein synthesis but affecting also the *mTOR* → *S6K* → *IRS* feedback leading to impaired activation of *PI3K-Akt* that promotes insulin resistance (Tremblay et al., 2007; Um et al., 2006; Newgard et al., 2009). On the other hand, there are evidences of leucine induced improvement of glucose metabolism and glucose tolerance (Layman & Walker, 2006; Zhang et al., 2007; Macotela et al., 2011). A recent study (Zeanandin et al., 2012) shows how these alternative responses

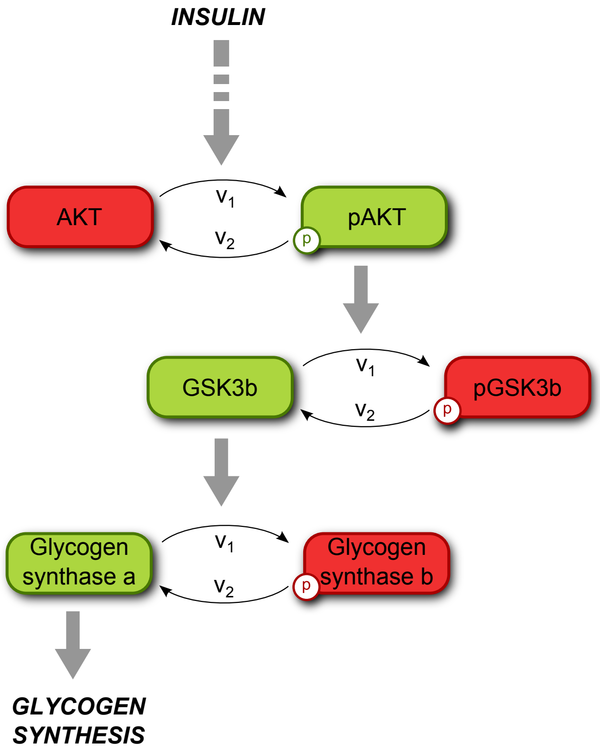


Figure 1: Outline of insulin promoted glycogen synthesis. Insulin indirectly promotes *AKT* phosphorylation and activation, which in turn promotes *GSK3β* phosphorilation inactivating it and allowing *GSK3β* substrate, *Glycogen synthase*, to promote glycogen synthesis. All active forms are in geen and inactive forms in red, p denotes the phosphorylated state.

depend on the tissue under study. In particular, adipose and muscle tissue are studied, revealing that leucine mainly affects *mTOR/S6K1* signalling pathway and insulin sensitivity in adipose tissue, while a similar effect is not induced in skeletal muscle. On the contrary, dietary leucine promotes Akt phosphorylation and improves insulin stimulated glucose transport in skeletal muscle.

Mathematical models based on ordinary differential equations (ODEs) are commonly used to analyze the dynamics of signalling networks. They are written from mass action kinetics allowing a detailed mechanistic description of proteins regulation, characterized by a large number of kinetic parameters that are inferred almost exclusively from literature. In particular, published mathematical models developed to study insulin signalling pathway are described in (Sedaghat et al., 2002; Dalle Pezze et al., 2012).

While the above models aimed essentially to simulate the behaviour of the system, in this paper we address the problem of using ODE based model as a tool to measure descriptive indices from experimental data, able to characterize insulin signalling modules in a specific tissue under specific conditions. Here we focus on the effect of leucine on insulin induced dynamic response of *Akt/GSK3β* deriving an ODE based model and defining descriptive indices to characterize the system. The focus is on the pathway in skeletal muscle cells, as skeletal muscle is a tissue that plays a major role accounting for approximately 75% of whole body insulin-stimulated glucose uptake. The dataset consists of protein level and phosphorylation monitored under three different experimental conditions: insulin stimulation, leucine stimulation and insulin stimulation after leucine preincubation. A general mathematical model of protein phosphorylation is derived from biochemistry with reasonable assumptions. Particular attention has been paid to the identifiability of parameters both a priori (if a unique solution theoretically exist) and a posteriori (if this solution can be derived from available experimental data). For this reason, four different hypotheses are assessed and compared using the experimental data. The best model is then used to further evaluate the effects of leucine on the dynamics of *GSK3β* inactivation analyzing this system as a block isolated from the upstream signalling pathway and defining two indices to characterize sensitivity and swiftness of the response of the system to a given input. *GSK3β* activity is then connected to glycogen synthesis at steady state.

2 Material and methods

2.1 Data

Three independent cell culture experiments were performed as biological replicates on a commercially available cell line of human skeletal muscle myoblasts that had been differentiated into myotubes as described in [add ref]. Cells were harvested at 6 time points (0, 2, 5, 10, 30 and 60 minutes) after stimulus to monitor both the early and the late insulin signaling effects. Cell lines were exposed to three different stimuli:

1. INS: cells were stimulated with $+100nM$ insulin at time 0;
2. LEU: cells were stimulated with $+400\mu M$ leucine at time 0;
3. LEU+INS: cells were pre-incubated with $+400\mu M$ leucine for one hour and then stimulated with $+100nM$ insulin at time 0.

Total proteins and their phosphorylated quote were measured using western blot. For each biological replicate of each protein, the complete time course for all three stimuli was run on the same blot. Measures were then repeated on different days and different blots to check technical reproducibility. Band intensities were analyzed by using the Odyssey infrared image system (LiCor). This system, based on infrared detection, permits probing the lysate with the antibodies for both the total and the phosphorylated amount of the protein using secondary antibodies with two different dyes. For the phosphorylated proteins antibody recognizing Ser^{473} phosphorilated *AKT* and antibody recognizing Ser^{9} phosphorylated *GSK3β* were used. The total amount of protein was verified to remain constant throughout the experiment for all experimental conditions. Linear relationship between western blot signals and protein concentration for selected antibodies was experimentally verified. Replicates, referred to their respective

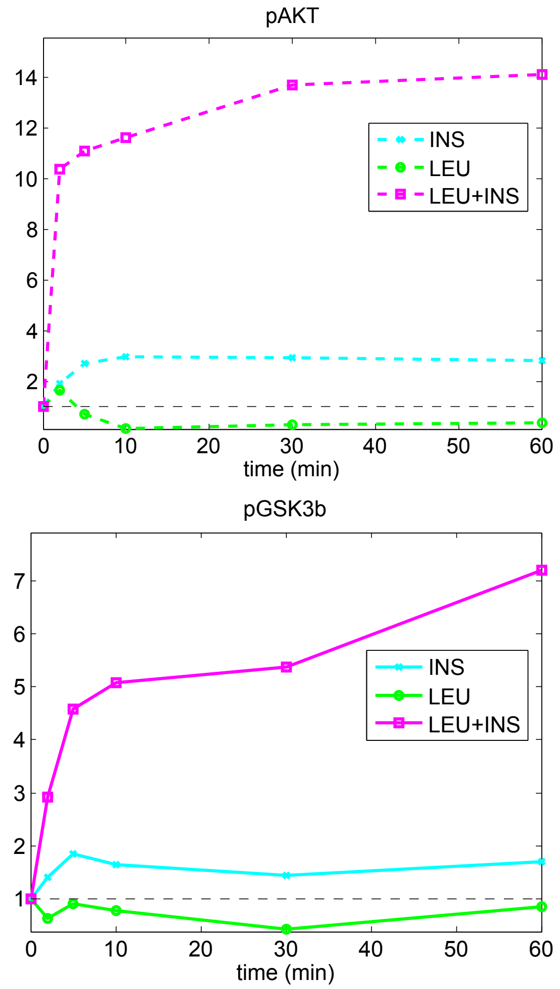


Figure 2: Experimental data. Phosphorylated $GSK3\beta$ and phosphorylated AKT time series data are shown for all three experimental conditions: insulin stimulation, leucine stimulation and insulin stimulation after perincubation with leucine.

basal and then mediated, provided an estimate of the standard deviation of the experimental error, equal to the 25% of the associated measure. In Figure 2, time series data of phosphorylated *GSK3*β and *AKT* are shown for all three experimental conditions: insulin stimulation, leucine stimulation and insulin stimulation after preincubation with leucine.

For all experimental conditions, the newly synthesized glycogen was measured as the amount of 14C-labeled glucose incorporated into glycogen over 90 minutes. All measures are in triplicates and were mediated and expressed relative to the basal value (no insulin/leucine stimulus); values are 1.53 and 2.25 for the conditions without (INS) and with leucine preincubation (LEU+INS), respectively.

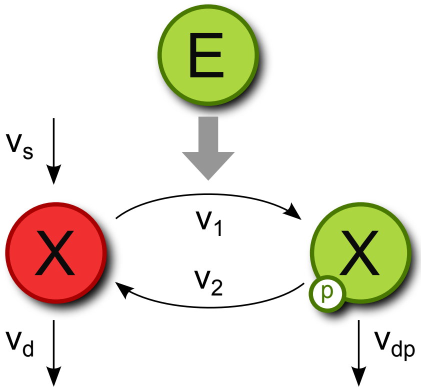


Figure 3: Schematic representation of protein phosphorylation and dephosphorylation. Protein X in the unphosphorylated/inactive form (red) can be phosphorylated by the enzyme kinase $E(t)$ assuming a phosphorylated/active form (green), where p denotes the phosphorylated state. $X(t)$ is synthesized from precursor and both $X(t)$ and $pX(t)$ are degraded.

2.2 Model definition

As shown in Figure 3, a generic protein $X(t)$ in the unphosphorylated/inactive form can interact with the enzyme kinase $E(t)$ which regulates its activity favouring the formation of the phosphorylated/active state $pX(t)$. Mass balance of the phosphorylated and the unphosphorylated forms written in terms of rate of phosphorylation (v_1) and dephosphorylation (v_2), unphosphorylated production from precursor (v_s), and degradation of both forms (v_d, v_{dp}), consists of the following set of differential equations:

$$\begin{aligned} \dot{X}(t) &= v_s + v_2(t) - v_1(t) - v_d(t) \\ p\dot{X}(t) &= v_1(t) - v_2(t) - v_{dp}(t) \end{aligned} \tag{1}$$

Some commonly used assumptions are made for the rates of production, degradation and phosphorylation/dephosphorylation (Klipp et al., 2011). A constant value is assumed for the production term, since it is well known that phosphorylation processes work in a much faster scale with respect to transcriptional and translational regulation. Degradation terms are assumed to be linearly dependent on the concentration of their substrates as derived from mass action kinetics. The same assumption is made for the rate of inactivation: dephosphorylation can be caused by the protein itself (autodephosphorylation) or by independent proteins, called phosphatases, that in most cases are abundant and not specific thus are not a limiting factor. Therefor, rates are defined as follow:

$$\begin{aligned} v_d(t) &= k_d \cdot X(t) \\ v_{dp}(t) &= k_d \cdot pX(t) \\ v_2(t) &= k_2 \cdot pX(t) \end{aligned} \tag{2}$$

where $X(t)$ and $pX(t)$ degradation rates are assumed to have the same kinetic constant k_d.

The activation term (v_1) is dependent on the amount of substrate and active kinase that regulate this process. The amount of catalyst and target protein are in the same order of magnitude and, in most cases, the reaction may be considered monomeric. Thus, a fair linear approximation of the rate of activation is given by the sum of a kinase independent term plus a term regulated by the kinase:

$$v_1(t) = (k_1 + \alpha \cdot E(t)) \cdot X(t) \tag{3}$$

Under all experimental conditions, the total amount of observed proteins ($X_{TOT}(t) = X(t) + pX(t)$) remains constant for the duration of the experiment. Therefore, assuming the derivative of the total protein equal to zero, the sum of Equations (1) provides a link between synthesis and degradation:

$$v_s(t) = v_d(t) + v_{dp}(t) \tag{4}$$

It is now possible to exploit the conservation relation $X_{TOT} = X(t) + pX(t)$ to express the system only in function of the state variable pX. From Equation (1), using Equation (3) for the expression of v_1, and considering that inactivation process is much faster than protein degradation ($k_2 \gg k_d$), thus $k_2 + k_d \approx k_2$, the dynamic of pX is described by the following equation:

$$p\dot{X}(t) = (k_1 + \alpha \cdot E(t)) \cdot X_{TOT} - (k_1 + \alpha \cdot E(t) + k_2) \cdot pX(t) \tag{5}$$

associated with measurement equations:

$$\begin{aligned} y_1(t) &= \frac{pX(t)}{pX_b} \triangleq px(t) \\ y_2(t) &= \frac{E(t)}{E_b} \triangleq e(t) \end{aligned} \tag{6}$$

where pX_b and E_b are the basal pre-stimulus levels of the phosphorylated protein and the substrate, respectively.

The model described by Equations (5) and (6) is not a priori identifiable, meaning that parameters cannot be uniquely determined from the observation, assuming perfect experimental data (Cobelli & Carson, 2007; Chis et al., 2011). However, it is possible to derive a uniquely identifiable parametrization by considering $px(t)$ as the state variable, and $e(t)$ as the controlling input.

$$p\dot{x}(t) = (k_1 + \alpha \cdot E_b \cdot e(t)) \cdot \frac{X_{TOT}}{pX_b} - (k_1 + \alpha \cdot E_b \cdot e(t) + k_2) \cdot px(t) \tag{7}$$

The ratio X_{TOT}/pX_b can be derived as function of other parameters by exploiting that, at basal stationary state, the system is at equilibrium, i.e. $p\dot{X}(0) = 0$, thus from Equation (5):

$$\frac{X_{TOT}}{pX_b} = \frac{k_1 + \alpha \cdot E_b + k_2}{k_1 + \alpha \cdot E_b} \tag{8}$$

Substituting (8) in (7) the following differential equation is derived:

$$\dot{px}(t) = f_1\left[e(t)\right] - f_2\left[e(t)\right] \cdot px(t), \qquad px(0) = 1 \tag{9}$$

where

$$\begin{aligned} f_1\left[e(t)\right] &= \left(\frac{k_1 + \alpha' + k_2}{k_1 + \alpha'}\right) \cdot \left(\alpha' \cdot e(t) + k_1\right) \\ f_2\left[e(t)\right] &= k_1 + k_2 + \alpha' \cdot e(t) \end{aligned} \tag{10}$$

Using the software DAISY (Bellu et al., 2007), parameters $\alpha' = \alpha \cdot E_b$, k_1 and k_2 were verified to be a priori identifiable from data measured during a single experiment.

The phosphorylated protein normalized to its basal value, px, evolves as the solution of a linear, time-variant, first order system, with macroparameters $f1$ and $f2$ dependent upon kinase expression e(t). If $e(t) = const = q$, $f1$ and $f2$ are constant in time and the solution of Equation (9), shown in Figure 4, is simply:

$$px(t) = \frac{f_1(q)}{f_2(q)} + \left(1 - \frac{f_1(q)}{f_2(q)}\right) \cdot e^{-f_2(q)t} \tag{11}$$

Thus, the output $px(t)$ growth exponentially with time t reaching asymptotically a steady state condition. Both the steady state and the time to reach it depend on parameters k_1, k_2 and α' and on the amplitude q of the input function. We can define the following two descriptive indices, shown in Figure 4, as:

1. steady state (SS): px phosphorylation level for $t \to \infty$

   $$SS(q) = \frac{f_1(q)}{f_2(q)} = \frac{\left(\frac{k_1+\alpha'+k_2}{k_1+\alpha'}\right) \cdot (\alpha' \cdot q + k_1)}{k_1 + k_2 + \alpha' \cdot q} \tag{12}$$

   it characterizes the response of the system at equilibrium, quantifying the maximum amplitude that can be reached by the output.

2. rise time (t_r): time to go from 10% to 90% of the final steady state value

   $$\tau_r(q) = \frac{2.197}{f_2(q)} = \frac{2.197}{k_1 + k_2 + \alpha' \cdot q} \tag{13}$$

   it characterize the dynamics of the response, assessing the ability of the system to respond to a fast input signal.

2.3 Parameters estimation

Parameters are estimated from the experimental data using non-linear least square optimization (solved with the trust-region-reflective algorithm implemented in Matlab) to find the set of parameters that allows minimizing the weighted sum of squared residuals

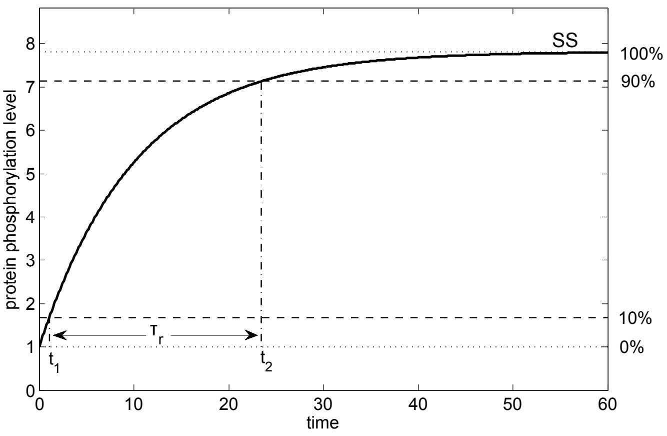


Figure 4: Descriptive indices. Illustration of the indices which caracterize the response of the system to a step function: rise time (τ_r) and steady state (SS).

($WRSS$). In order to reduce the risk of finding local minima the identification procedure is repeated 100 times choosing random initial estimates from a uniform distribution $\mathcal{U}\,(0, 15)$. All parameters are constrained within the interval $[0, 20]$. The optimal solution is defined as the set of estimates which provides the best compromise between good fit and a posteriori identifiability. Thus, the set of parameters that is estimated with the best precision (providing smaller confidence intervals in terms of coefficient of variation (CV) derived from Fisher matrix), is selected as optimal solution among those sets of estimates which provide a fit (in terms of $WRSS$) in the range within one standard deviation from the best fit to data.

In addition to the complete model, three simplified versions (Table 1) are tested, based on the assumption that some parameters cannot be distinguished from zero and/or are not affected by experimental conditions:

case A. all three parameters (Equations (9) and (10)) are estimated separately for each experimental condition;

case B. since phosphorylation is likely to be equal in all experimental conditions when the stimulus is absent, the number of parameters can be reduced by estimating

	α'	k_1	k_2
A.	Different	Different	Different
B.	Different	Equal	Different
C.	Different	0	Different
D.	Different	0	Equal

Table 1: The four different tested model variants: model variants regards the presence ('different') or absence ('equal') of an effect of leucine/insulin on model parameters, or the absence of basal phosphorylation ($k_1 = 0$).

only one basal phosphorylation, i.e. k_1 is the same in all experimental conditions;

case C. it is reasonable to assume that basal phosphorylation can be neglected and fixed equal to zero, further reducing the number of estimated parameters;

case D. in order to verify that the differences between the two experimental conditions are due to both kinase regulation α' and dephosphorylation k_2 also the case in which k_2 is equal for all experimental conditions is tested.

Performances of the four models derived from Equations (9) and (10) by including the four sets of assumptions, are compared based on Akaike information criterion (*AIC*) computed as $AIC = WRSS + 2 \cdot N$, where N is the number of estimated parameters. *AIC* is a selection criterion to find the model which best explains data with the minimum number of parameters in order to avoid overfitting.

Results

The model of protein phosphorylation was applied to study *AKT/GSK3*β signalling pathway in the context of insulin signalling. First, the model is used to characterize *AKT* mediated phosphorylation of *GSK3*β (see Figure 1) selecting the assumptions that are best supported by the experimental data according to the model identification criteria. Then, the calibrated model is exploited to give deeper insights into the dynamics of *GSK3*β phosphorylation and mechanisms affected by leucine preincubiation, using the previously described descriptive indices. Finally, *GSK3*β phosphorylation is linked to glycogen synthesis at steady state.

AKT mediated regulation of *GSK3*β phosphorylation

Time series data shown in Figure 2 suggest that insulin stimulation always affects the level of phosphorylation of both *AKT* and *GSK3*β while leucine alone has no effect on the observed time series. This means that leucine does not directly promote phosphorylation but only affects the regulatory mechanisms that stay beneath it, thus only the conditions without and with leucine preincubation were assessed.

The model of phosphorylation and inactivation of *GSK3*β by *AKT* can be easily derived from the general model in the Methods section by considering:

$$\begin{aligned} E(t) &= pAKT(t) \\ X(t) &= GSK3\beta(t) \\ pX(t) &= pGSK3\beta(t) \end{aligned} \tag{14}$$

and

$$\begin{aligned} e(t) &= \frac{pAKT(t)}{pAKT_b} \\ px(t) &= \frac{pGSK3\beta(t)}{pGSK3\beta_b} \end{aligned} \tag{15}$$

Parameters k_1, $\alpha' = \alpha \cdot pAKT_b$ and k_2 are estimated by fitting $px(t)$ time series data and using $e(t)$ as forcing function.

The four hypotheses described in Table 1 were tested and results of model identification are shown in Table 2: each estimated parameter is reported along with the respective *CV*, an index of the precision of the estimates that can be considered reasonable if lower than 100, the *WRSS* and the *AIC* index. Results can be summarized as follow.

case A. All six parameters are estimated with poor precision, even if data are well fitted.

case B. The estimation of a common value for the basal phosphorylation (k_1) in all experimental conditions, reduces to five the total number of estimated parameters, partially improving the precision of their estimates, but not solving the problem of overall a posteriori identifiability of the system.

case C. When basal phosphorylation is fixed equal to zero, the number of estimated parameters is reduced to four. This choice is supported also by the fact that a low value was estimated for k_1 in both previous cases. The resulting model is a posteriori identifiable with slightly worse fit (1.42 instead of 1.19) with respect to case B., but reduced *AIC* index (9.94 instead of 11.19) attesting that this is the best model according to the defined principle of parsimony.

case D. This hypothesis provides an increased *AIC* value, and can thus be discarded, proving that the differences between the two experimental conditions are due to both kinase regulation α' and dephosphorylation k_2.

As expected, model simplification with the reduction of the number of parameters to be estimated, allows to improve the a posteriori identifiability of the model at the price of a slight worsening of the fit. Case C. is the best supported by the available data among the four testes hypotheses, being the one which provides a posteriori identifiability with the lowest *AIC* index. As shown in Figure 5, this model well describes observed data for both experimental conditions. This model will be used in the next paragraph to analyze the effects of leucine preincubation on the response of the system.

	α'		k_1		k_2		$WRSS$	AIC
	INS	$LEU+INS$	INS	$LEU+INS$	INS	$LEU+INS$		
A.	6.62 (571)	0.01 (1788)	0.00 (10^8)	0.01 (1479)	9.77 (639)	0.75 (219)	1.19	13.19
B.	9.59 (85)	0.01 (976)	0.01 (752)		14.21 (71)	0.73 (133)	1.19	11.19
C.	**5.09 (35)**	**0.03 (77)**	**0**		**7.54 (27)**	**0.27 (58)**	**1.42**	**9.94**
D.	0.21 (76)	0.04 (64)	0		0.31 (51)		2.10	10.10

Table 2: Results of model identification. For the four hypothesis described in Table 1, and for the two analyzed experimental conditions (INS and LEU+INS), each estimated parameter is reported along with the respective CV (an index of the precision of the estimates, reasonable if lower than 100), the WRSS (weighted sum of the square of the residuals) and the AIC index (index of parsimony, lower index corresponds to the best model.

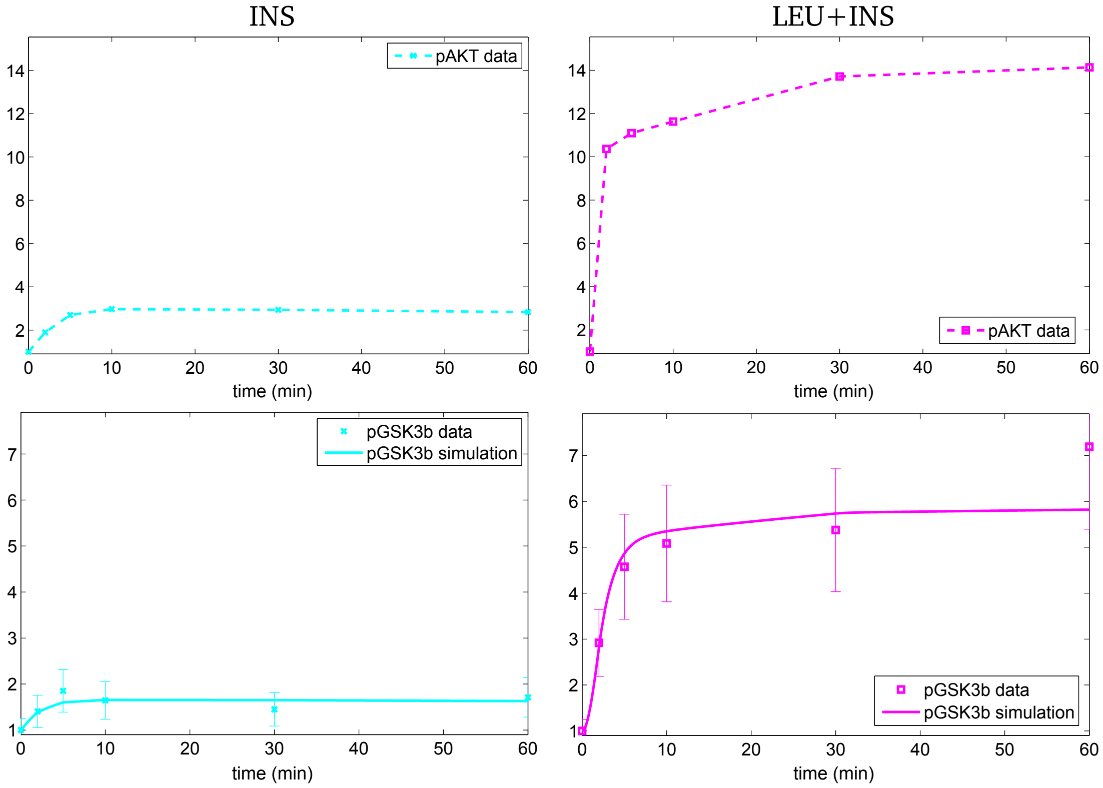


Figure 5: Comparison between simulated and real data. For both INS (left panels) and LEU+INS (right panels) experimental conditions, phosphorylated AKT (upper panels) and phosphorylated $GSK3\beta$ (lower panels) time courses data (symbol and error bars) are shown. Models predictions (continuous line) for phosphorylated $GSK3\beta$ are compared to the real data.

2.4 Effects of leucine on the dynamics of *GSK3*β inactivation

Visual inspection of data in Figure 5 reveals an increase of *GSK3*β phosphorylation after leucine preincubation, but a qualitative analysis does not allow to determine if this effect is only due to the increased level of *AKT* phosphorylation or to an additional leucine effect on regulatory mechanisms beneath *GSK3*β phosphorylation. Results of model identification reported in Table 2, suggest that leucine preincubation slowers the dynamics of the system by decreasing both *AKT* dependent phosphorylation (α' from 5.09 to 0.03) and independent dephosphorylation (k_2 from 7.54 to 0.27).

Under the assumption of an input step function of amplitude q, steady state (SS) and rise time (τ_r), defined in the Methods sections, can be used to analyze the dynamics of *GSK3*β inactivation and its dependence on phosphorylated AKT, providing a measure of sensitivity and swiftness of the response of the system to the *AKT* signal as a function of the the value of q. In Figure 6, SS and τ_r are shown as a function of q, in the two analyzed experimental conditions. Figure 6, left panel, shows that, with insulin stimulation

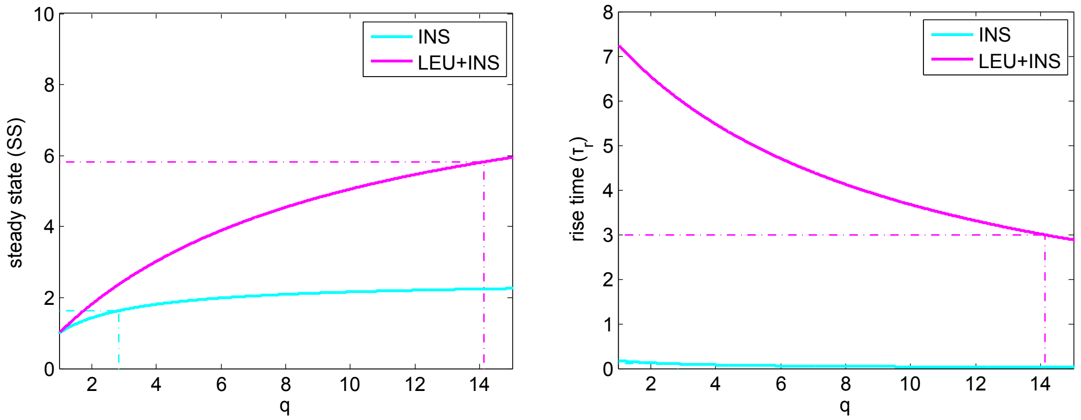


Figure 6: Characterization of the dynamic response of $GSK3\beta$. Steady state (SS) and rise time (τ_r) are shown for increasing values of q, for the two experimental conditions: without (*INS*, cyan blue) and with (*LEU+INS*, magenta) leucine preincubation. Dotted lines represent values of q corresponding to experimental data of phosphorylated AKT at steady state.

(INS), for increasing values of q the SS rapidly growth until reaching a saturation level equal to 2.48 while, when cells are preincubated with leucine ($LEU + INS$), SS growth slightlier with q but it can reach a higher level (9.20). For example, with $q = 14.13$ (the experimental value of phosphorylated AKT at steady state in the condition $LEU+INS$) a higher value of SS is reached with respect to $q = 2.83$ (experimental value in INS condition), whichever experimental curve is considered. But, even with the same value of q (for example $q = 14.13$), leucine preincubation would always provide a higher SS. Figure 6, right panel, shows that for increasing values of q, the τ_r decreases until reaching 0 and that the response of the system is always faster in the condition without leucine preincubation. Thus, a stronger stimulus (higher q) always induce a stronger and faster response, while leucine preincubation amplifies the response of the system but slows down the dynamics.

To better quantify the effect of leucine preincubation on the behaviour of the system, let's consider the relationship between the kinase level needed in the two experimental conditions to reach the same steady state level. The equivalency between the SS value in the two experimental conditions is derived substituting the estimated parameters in Equation (12), as follow:

$$\frac{(7.54 + 5.09) \cdot q_{INS}}{(7.54 + q_{INS} \cdot 5.09)} = \frac{(0.27 + 0.03) \cdot q_{LEU+INS}}{(0.27 + q_{LEU+INS} \cdot 0.03)} \tag{16}$$

that brings to:

$$q_{LEU+INS} = \frac{3.41}{2.26 + 1.15 \cdot q_{INS}} \cdot q_{INS} \tag{17}$$

thus, for $q_{INS} = 1$ (that is basal state), $q_{LEU+INS} = q_{INS} = 1$, but as soon as $q_{INS} > 1$, $q_{INS} > q_{LEU+INS}$. Meaning that, after leucine preincubation, the system is more efficient since a lower amount of phosphorylated *AKT* is able to activate the same level of *GSK3β* phosphorylation. However, this always comes at the price of a higher rise time: for $q_{INS} > 1$ we have that $(7.54 + q_{INS} \cdot 5.09) > (0.27 + q_{LEU+INS} \cdot 0.03)$, thus $\tau_{r,LEU+INS} > \tau_{r,INS}$.

Regulation of glycogen synthesis

Following the outline in Figure 1, we applied the model described in the Methods section to study how $GSK3\beta$ inactivation by AKT promotes glycogen synthesis. Having measures of glycogen synthesis for both conditions (insulin alone and with leucine preincubation) only at 90 minutes after stimulation with insulin, we focus on the relationship between phosphorylated $GSK3\beta$ (px) and glycogen synthesis (gly) at steady state. This relationship can be derived from the following three equations:

1.
$$X(t) = X_{TOT} - pX(t) = X_{TOT} - px(t) \cdot GSK3\beta_b \tag{18}$$
where $X(t)$ represents unphosphorylated $GSK3\beta$. X_{TOT} (total $GKS3\beta$) and $GSK3\beta_b$ (basal $GKS3\beta$) are unknown constants.

2.
$$\dot{GS_a}(t) = a \cdot GS_b(t) - b \cdot X(t) \cdot GS_a(t) = a \cdot (GS_{TOT} - GS_a(t)) - b \cdot X(t) \cdot GS_a(t) \tag{19}$$
where GSa is the unphosphorylated and active form *glycogen synthase a* and GSb is the phosphorylated and inactive form *glycogen synthase b*. a, b and GS_{TOT} (total *glycogen synthase*) are unknown constants.

3.
$$gly(t) = c \cdot GS_a(t) \tag{20}$$
Glycogen synthesis (gly) is assumed to be linearly dependent on GSa with unknown constant c.

Reasoning in terms of steady state (where $\dot{GS_a} = 0$), it is possible to derive the following relationship between glycogen synthesis (SS_{gly}) and phosphorylated $GSK3\beta$ (SS_{px}) at steady state:

$$SS_{gly} = \frac{\lambda}{\gamma - SS_{px}} \tag{21}$$

where $\lambda = \frac{c \cdot a \cdot GS_{TOT}}{b \cdot GSK3\beta_b}$ and $\gamma = \frac{a + b \cdot X_{TOT}}{b \cdot GSK3\beta_b}$ are unknown parameters.

Reminding that all data are expressed relative to the time 0' pre-stimulus sample as described in the Data paragraph, both px and gly are equal to 1 at basal steady state level. Thus, from Equation (21): $\lambda = \gamma - 1$. One curve can be derived for each experimental condition, valid in the range of values under consideration and passing through the points corresponding to the measured steady state value of gly and px,

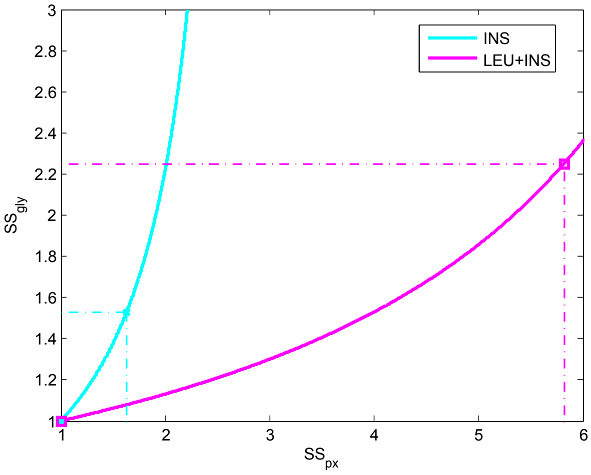


Figure 7: Regulation of glycogen synthesis at steady state. The relationship between glycogen synthesis (SS_{gly}) and phosphorylated $GSK3\beta$ (SS_{px}) at steady state is shown for both experimental conditions: INS (cyan blue) and LEU+INS (magenta). Symbols represent experimental data.

as shown in Figure 7. An increase in $GSK3\beta$ phosphorylation always corresponds to an increase in glycogen synthesis as expected, however Figure 7 shows that without leucine preincubation, glycogen synthesis growth faster as a function of phosphorylated $GSK3\beta$.

3 Discussion

In this paper, a mathematical model to describe protein phosphorylation was developed and applied to study *AKT/GSK3β* signalling pathway in the context of insulin signalling, focusing on the regulatory mechanisms that bring to the control of glycogen synthesis as shown in Figure 1 where, as a global effect, the inactivation of *GSK3β* by *AKT* promotes glycogen synthesis.

Data show that leucine alone has no effect on *AKT* and *GSK3β* activity, but it improves the effect of insulin, meaning that leucine does not directly promote *AKT* and *GSK3β* phosphorylation but only affects the regulatory mechanisms that stay beneath it. In fact, insulin and leucine are known to act through independent signalling mechanisms (Greiwe et al., 2001) stimulating different pathways mediated, in part, through the same proteins. One example of verified crosstalk between these pathways is the binding protein *mTOR*, where insulin and leucine signals converge to regulate protein synthesis (Anthony et al., 2001), but other unknown crosstalk proteins might be involved in the regulation of glycogen synthesis. However, from qualitative inspection of the data it is possible to infer only that both *AKT* and *GSK3β* phosphorylations are enhanced after leucine preincubation, but it is not possible to distinguish if phosphorylated *GSK3β* is higher only as a direct effect of increased *AKT* kinase activity or if kinetic parameters are also affected. The use of mathematical models allowed to gain a quantitative insight on the dynamics of the system: in particular, we tested 4 different hypothesis and results of model identification showed that leucine preincubation changes kinetic parameters involved in both protein phosphorylation and dephosphorylation. It seems, thus, reasonable to hypothesize that other kinases and/or phosphatases directly and/or indirectly involved in this regulatory mechanisms might be part of leucine signalling pathway, affecting *GSK3β* inactivation only when both stimuli are present. Data also shows that increased *GSK3β* phosphorylation corresponds to increased glycogen synthesis, thus it is possible that leucine (or other BCAA) administration might improve glycogen synthesis also in insulin-resistant states. However this analysis was limited to few steady state values, thus more (possibly dynamic) experimental data would be needed for further conclusions.

Particular attention was payed in model definition to find an adequate compromise in the level of detail included in the model: convenient biological assumptions were made in order to guarantee a priori identifiability of the parameters under ideal conditions (noise-free data, continuous time observation and error-free model structure) and different hypothesis were tested in order to obtain a set of parameters a posteriori identifiable, thus estimated with reasonable confidence from available data. In order to give a more readable description of the dynamics of the identified model, two descriptive indices were defined to quantitatively characterize sensitivity and swiftness of the response of the system. This indices allow to perform a partition analysis of the signalling pathway dividing it in different blocks isolated from the upstream and downstream pathway; each block can be characterized by two values (steady state and rise time) that describe its dynamic response. This description used here to study *GSK3β/AKT* signalling could be extended to study pathways at larger scale simplifying the comparison of the pathway under different experimental conditions, on different cell types or in different states.

References

Anthony, J., Anthony, T., Kimball, S., & Jefferson, L. (2001). Signaling pathways involved in translational control of protein synthesis in skeletal muscle by leucine. *The Journal of nutrition*, *131*(3), 856S–860S.

Bellu, G., Saccomani, M., Audoly, S., & D'Angiò, L. (2007). Daisy: A new software tool to test global identifiability of biological and physiological systems. *Computer methods and programs in biomedicine*, *88*(1), 52.

Bjrnholm, M., & Zierath, J. (2005). Insulin signal transduction in human skeletal muscle: identifying the defects in type ii diabetes. *Biochemical Society Transactions*, *33*, 354–357.

Chis, O., Banga, J., & Balsa-Canto, E. (2011). Structural identifiability of systems biology models: A critical comparison of methods. *PloS One*, *6*(11), e27755.

Cobelli, C., & Carson, E. (2007). *Introduction to modeling in physiology and medicine*. Academic Press.

Cohen, P. (1999). The croonian lecture 1998. identification of a protein kinase cascade of major importance in insulin signal transduction. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *354*(1382), 485–495.

Cohen, P., & Goedert, M. (2004). Gsk3 inhibitors: development and therapeutic potential. *Nature reviews Drug discovery*, *3*(6), 479–487.

Cozzone, D., Fröjdö, S., Disse, E., Debard, C., Laville, M., Pirola, L., & Vidal, H. (2008). Isoform-specific defects of insulin stimulation of akt/protein kinase b (pkb) in skeletal muscle cells from type 2 diabetic patients. *Diabetologia*, *51*(3), 512–521.

Dalle Pezze, P., Sonntag, A., Thien, A., Prentzell, M., Godel, M., Fischer, S., Neumann-Haefelin, E., Huber, T., Baumeister, R., Shanley, D., et al. (2012). A dynamic network model of mtor signaling reveals tsc-independent mtorc2 regulation. *Science Signalling*, *5*(217), ra25.

Greiwe, J., Kwon, G., McDaniel, M., & Semenkovich, C. (2001). Leucine and insulin activate p70 s6 kinase through different pathways in human skeletal muscle. *American Journal of Physiology-Endocrinology And Metabolism*, *281*(3), E466–E471.

Hers, I., Vincent, E., & Tavaré, J. (2011). Akt signalling in health and disease. *Cellular signalling*, *23*(10), 1515–1527.

Inzucchi, S. (2002). Oral antihyperglycemic therapy for type 2 diabetes. *JAMA: the journal of the American Medical Association*, *287*(3), 360–372.

Kimball, S., & Jefferson, L. (2006). Signaling pathways and molecular mechanisms through which branched-chain amino acids mediate translational control of protein synthesis. *The Journal of nutrition*, *136*(1), 227S–231S.

Klipp, E., Liebermeister, W., Wierling, C., Kowald, A., Lehrach, H., & Herwig, R. (2011). *Systems biology*. Wiley-VCH.

Layman, D., & Walker, D. (2006). Potential importance of leucine in treatment of obesity and the metabolic syndrome. *The Journal of nutrition*, *136*(1), 319S–323S.

LeRoith, D., Olefsky, J., & Taylor, S. (2003). *Diabetes mellitus: a fundamental and clinical text*. Lippincott Williams & Wilkins.

Macotela, Y., Emanuelli, B., Bång, A., Espinoza, D., Boucher, J., Beebe, K., Gall, W., & Kahn, C. (2011). Dietary leucine-an environmental modifier of insulin resistance acting on multiple levels of metabolism. *PLoS One*, *6*(6), e21187.

McManus, E., Sakamoto, K., Armit, L., Ronaldson, L., Shpiro, N., Marquez, R., & Alessi, D. (2005). Role that phosphorylation of gsk3 plays in insulin and wnt signalling defined by knockin analysis. *The EMBO journal*, *24*(8), 1571–1583.

Meijer, L., Flajolet, M., & Greengard, P. (2004). Pharmacological inhibitors of glycogen synthase kinase 3. *Trends in pharmacological sciences*, *25*(9), 471–480.

Moller, D., et al. (2001). New drug targets for type 2 diabetes and the metabolic syndrome. *Nature*, *414*(6865), 821–827.

Nair, K., & Short, K. (2005). Hormonal and signaling role of branched-chain amino acids. *The Journal of nutrition*, *135*(6), 1547S–1552S.

Newgard, C., An, J., Bain, J., Muehlbauer, M., Stevens, R., Lien, L., Haqq, A., Shah, S., Arlotto, M., Slentz, C., et al. (2009). A branched-chain amino acid-related metabolic signature that differentiates obese and lean humans and contributes to insulin resistance. *Cell metabolism*, *9*(4), 311–326.

Sedaghat, A., Sherman, A., & Quon, M. (2002). A mathematical model of metabolic insulin signaling pathways. *American Journal of Physiology-Endocrinology and Metabolism*, *283*(5), E1084–E1101.

Taniguchi, C., Emanuelli, B., & Kahn, C. (2006). Critical nodes in signalling pathways: insights into insulin action. *Nature Reviews Molecular Cell Biology*, *7*(2), 85–96.

Tremblay, F., Lavigne, C., Jacques, H., & Marette, A. (2007). Role of dietary proteins and amino acids in the pathogenesis of insulin resistance. *Annu. Rev. Nutr.*, *27*, 293–310.

Um, S., D'Alessio, D., & Thomas, G. (2006). Nutrient overload, insulin resistance, and ribosomal protein s6 kinase 1, s6k1. *Cell metabolism*, *3*(6), 393–402.

Zeanandin, G., Balage, M., Schneider, S., Dupont, J., Hébuterne, X., Mothe-Satney, I., & Dardevet, D. (2012). Differential effect of long-term leucine supplementation on skeletal muscle and adipose tissue in old rats: an insulin signaling pathway approach. *Age*, *34*(2), 371–387.

Zhang, Y., Guo, K., LeBlanc, R., Loh, D., Schwartz, G., & Yu, Y. (2007). Increasing dietary leucine intake reduces diet-induced obesity and improves glucose and cholesterol metabolism in mice via multimechanisms. *Diabetes*, *56*(6), 1647–1654.

6
Conclusions

In this thesis, the problem of inference in systems biology is studied from several points of view, and different methods are developed and applied to analyze biological processes using the appropriate level of simplification of reality. We describe methods aimed at elucidating mechanisms underlying observed data, showing how mathematical modelling can be a powerful tool to gain insights into biological systems. Even if in different biological contexts and considering different levels of abstraction of reality, all developed approaches allow to make hypothesis that can be tested further. For example, we can make testable hypotheses about plausible effects of diseases and drugs on the structure of a system or about the functional role of unknown regulatory links. These methods are thought as general tools to solve classes of problems with different but complementary approaches, and their application to real case studies is used to prove their efficiency and their potentialities.

In Chapter 2, a method to infer large-scale signalling networks strictly from perturbation experiment data was described and applied to a real case study. In this case, the interest is in the analysis of the topology of the network, for example to understand how diseases or drugs affect the structure of the network. This approach is strictly data-driven and allows to infer cell type specific networks even when only static data are available, with no need for prior knowledge and being computationally very fast. However, as for most of reverse-engineering methods, reconstructed networks have poor biological interpretability being limited to perturbed and measured nodes. Clearly, since we are dealing with static networks, simulations and predictions cannot be performed unless we associate networks with suitable mathematical models.

The most simple and intuitive way of modelling networks without describing the biochemistry of interactions, are logic-based models. In Chapter 3, the previously descibed approach, along with other reverse-engineering methods, were integrated with a logic-based method (called CellNOpt (Saez-Rodriguez et al., 2009)) to obtain refined models from prior knowledge about the network structure and from experimental data. Literature constrained and data-driven approaches were integrated to obtain a logic model that is able to describe the given data. When applied to the signalling pathway of growth and inflammatory signalling, this integrated approach provided a logic model that well describe data highlighting links that were missing in the prior knowledge but are useful to explain data and are supported by information derived from protein interaction networks.

When the interest is in making hypothesis about mechanisms underlying specific regulatory circuits, a more realistic and quantitative description of chemical reactions dynamics is required. In Chapter 4 we focused on quantitative modelling of small systems using ordinary differential equations, exploiting the fact that some recurrent regulatory motifs, called network motifs, are known to play important functional roles in cell biology. In a first study, we aimed at explaining the adaptation observed at genome-wide scale in the yeast stress response showing how an autoregulatory mechanism based on feedback loops is able to reproduce all the kinetic features observed in mRNA time-series data. In a second study, we aimed at reducing the search space for new interactions in mixed transcriptional and post-transcriptional regulatory networks in order to identify unknown regulatory mechanisms involving miRNA. Starting form a list of putative feed-forward loops selected based on sequence analysis, we used model identification criteria to select feed-forward loops supported by experimental data identifying potential novel players which might play a functional role.

In Chapter 5, approaches discussed in previous Chapters were integrated in the study of insulin signalling pathway, in order to analyze the systems from different points of view. Key molecules of the pathway were measured to study effects of leucine on insulin stimulated pathway. The network was first derived from public databases and interpreted using a logic-based modelling approach to study the entire pathway in a qualitative way. This choice was motivated by the fact that only few proteins of the network could be measured and, for some of them, time-course measures resulted to be more reliable as pattern then an quantitative measures. The main aim of this analysis was to highlight interesting regulatory mechanisms that were of particular interest for further study. In this way, the $AKT - GSK3\beta$ pathway was identified to be affected by leucine and was thus analyzed more in detail using ordinary differential equations derived from mass action kinetics, in order to make hypotheses about how leucine affects the dynamics of this regulatory mechanism and it potential role in insulin resistant cases.

In this thesis we have shown how network inference and mathematical modelling can be very useful tools to gain insights into biological systems and that the choice of the appropriate level of simplification to be included in the model is of paramount importance. The knowledge of different approaches allows to chose the best one depending on the purpose of the research, (e.g. analyze the effect of a disease, reveal an unknown regulatory mechanism), on the analyzed biological system (e.g. network level, mechanistic level) and on the available data (e.g. time series, static data). The development of a model has to start from a problem of biological interest which guides the choice of the best approach to face the specific question. Very often, the integration of different modelling approaches is the best way for studying a biological system, since it permits to have different and complementary points of view.

Bibliography

Airoldi, E., Huttenhower, C., Gresham, D., Lu, C., Caudy, A., Dunham, M., Broach, J., Botstein, D., & Troyanskaya, O. (2009). Predicting cellular growth from gene expression signatures. *PLoS Computational Biology*, *5*(1), e1000257.

Aldridge, B. B., Burke, J. M., Lauffenburger, D. A., & Sorger, P. K. (2006). Physicochemical modelling of cell signalling pathways. *Nature cell biology*, *8*(11), 1195–1203.

Alexopoulos, L., Saez-Rodriguez, J., Cosgrove, B., Lauffenburger, D., & Sorger, P. (2010). Networks inferred from biochemical data reveal profound differences in toll-like receptor and inflammatory signaling between normal and transformed hepatocytes. *Molecular & Cellular Proteomics*, *9*(9), 1849–1865.

Alon, U. (2006). *An Introduction to Systems Biology - Design Principles of Biological Circuits*. Chapman & Hall/CRC;.

Alon, U. (2007). Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, *8*(6), 450–461.

Alon, U., Surette, M. G., Barkai, N., & Leibler, S. (1999). Robustness in bacterial chemotaxis. *Nature*, *397*(6715), 168–171.

Anthony, J., Anthony, T., Kimball, S., & Jefferson, L. (2001). Signaling pathways involved in translational control of protein synthesis in skeletal muscle by leucine. *The Journal of nutrition*, *131*(3), 856S–860S.

Baek, D., Villen, J., Shin, C., Camargo, F. D., Gygi, S. P., & Bartel, D. P. (2008). The impact of micrornas on protein output. *Nature*, *455*(7209), 64–71.

Bansal, M., Belcastro, V., Ambesi-Impiombato, A., & di Bernardo, D. (2007). How to infer gene networks from expression profiles. *Molecular systems biology*, *3*, 78.

Barabási, A., & Oltvai, Z. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, *5*(2), 101–113.

Baroudi, M. E., Cora, D., Bosia, C., Osella, M., & Caselle, M. (2011). A curated database of mirna mediated feed-forward loops involving myc as master regulator. *PloS one*, *6*(3), e14742.

Bartel, D. P. (2004). Micrornas: genomics, biogenesis, mechanism, and function. *Cell*, *116*(2), 281–297.

Bartel, D. P. (2009). Micrornas: target recognition and regulatory functions. *Cell*, *136*(2), 215–233.

Becskei, A., & Serrano, L. (2000). Engineering stability in gene networks by autoregulation. *Nature*, *405*(6786), 590–593.

Behar, M., Hao, N., Dohlman, H., & Elston, T. (2007). Mathematical and computational analysis of adaptation via feedback inhibition in signal transduction pathways. *Biophysical journal*, *93*(3), 806–821.

Belle, A., Tanay, A., Bitincka, L., Shamir, R., & O'Shea, E. (2006). Quantification of protein half-lives in the budding yeast proteome. *Proceedings of the National Academy of Sciences*, *103*(35), 13004–13009.

Bellu, G., Saccomani, M. P., Audoly, S., & D'Angio, L. (2007). Daisy: a new software tool to test global identifiability of biological and physiological systems. *Computer methods and programs in biomedicine*, *88*(1), 52–61.

Bender, C., Heyde, S., Henjes, F., Wiemann, S., Korf, U., & Beissbarth, T. (2011). Inferring signalling networks from longitudinal data using sampling based approaches in the r-package 'ddepn'. *BMC bioinformatics*, *12*, 291.

Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., Barzilai, A., Einat, P., Einav, U., Meiri, E., Sharon, E., Spector, Y., & Bentwich, Z. (2005). Identification of hundreds of conserved and nonconserved human micrornas. *Nature genetics*, *37*(7), 766–770.

Bjrnholm, M., & Zierath, J. (2005). Insulin signal transduction in human skeletal muscle: identifying the defects in type ii diabetes. *Biochemical Society Transactions*, *33*, 354–357.

Blinov, M., & Moraru, I. (2012). Logic modeling and the ridiculome under the rug. *BMC biology*, *10*(1), 92.

Bornholdt, S. (2005). Systems biology: less is more in modeling large genetic networks. *Science Signalling*, *310*(5747), 449.

Bouskila, M., Hirshman, M., Jensen, J., Goodyear, L., & Sakamoto, K. (2008). Insulin promotes glycogen synthesis in the absence of gsk3 phosphorylation in skeletal muscle. *American Journal of Physiology-Endocrinology And Metabolism*, *294*(1), E28–E35.

Brazhnik, P., de la Fuente, A., & Mendes, P. (2002). Gene networks: how to put the function in genomics. *TRENDS in Biotechnology*, *20*(11), 467–472.

Buchler, N., Gerland, U., & Hwa, T. (2005). Nonlinear protein degradation and the function of genetic circuits. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(27), 9559–9564.

Butcher, E., Berg, E., & Kunkel, E. (2004). Systems biology in drug discovery. *Nature biotechnology*, *22*(10), 1253–1259.

Calzone, L., Tournier, L., Fourquet, S., Thieffry, D., Zhivotovsky, B., Barillot, E., & Zinovyev, A. (2010). Mathematical modelling of cell-fate decision in response to death receptor engagement. *PLoS computational biology*, *6*(3), e1000702.

Causton, H., Ren, B., Koh, S., Harbison, C., Kanin, E., Jennings, E., Lee, T., True, H., Lander, E., & Young, R. (2001). Remodeling of yeast genome expression in response to environmental changes. *Molecular biology of the cell*, *12*(2), 323–337.

CellSignalingTechnology. Insulin receptor signaling. `http://www.cellsignal.com`.

Cerami, E., Gross, B., Demir, E., Rodchenkov, I., Babur, Ö., Anwar, N., Schultz, N., Bader, G., & Sander, C. (2011a). Pathway commons, a web resource for biological pathway data. *Nucleic Acids Research*, *39*(suppl 1), D685–D690.

Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., Schultz, N., Bader, G. D., & Sander, C. (2011b). Pathway commons, a web resource for biological pathway data. *Nucleic acids research*, *39*(Database issue), D685–90.

Chechik, G., & Koller, D. (2009). Timing of gene expression responses to environmental changes. *Journal of Computational Biology*, *16*(2), 279–290.

Chechik, G., Oh, E., Rando, O., Weissman, J., Regev, A., & Koller, D. (2008). Activity motifs reveal principles of timing in transcriptional control of the yeast metabolic network. *Nature biotechnology*, *26*(11), 1251–1259.

Chen, H. C., Lee, H. C., Lin, T. Y., Li, W. H., & Chen, B. S. (2004). Quantitative characterization of the transcriptional regulatory network in the yeast cell cycle. *Bioinformatics (Oxford, England)*, *20*(12), 1914–1927.

Chen, K. C., Wang, T. Y., Tseng, H. H., Huang, C. Y., & Kao, C. Y. (2005). A stochastic differential equation model for quantifying transcriptional regulatory network in saccharomyces cerevisiae. *Bioinformatics (Oxford, England)*, *21*(12), 2883–2890.

Chis, O., Banga, J., & Balsa-Canto, E. (2011). Structural identifiability of systems biology models: A critical comparison of methods. *PloS One*, *6*(11), e27755.

Ciaccio, M. F., Wagner, J. P., Chuu, C. P., Lauffenburger, D. A., & Jones, R. B. (2010). Systems analysis of egf receptor signaling dynamics with microwestern arrays. *Nature methods*, *7*(2), 148–155.

Cobelli, C., & Carson, E. (2007). *Introduction to modeling in physiology and medicine*. Academic Press.

Cobelli, C., Foster, D., & Toffolo, G. (2000). *Tracer kinetics in biomedical research: from data to model*. New York: Kluwer Academic/Plenum.

Cohen, P. (1999). The croonian lecture 1998. identification of a protein kinase cascade of major importance in insulin signal transduction. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *354*(1382), 485–495.

Cohen, P., & Goedert, M. (2004). Gsk3 inhibitors: development and therapeutic potential. *Nature reviews Drug discovery*, *3*(6), 479–487.

Cozzone, D., Fröjdö, S., Disse, E., Debard, C., Laville, M., Pirola, L., & Vidal, H. (2008). Isoform-specific defects of insulin stimulation of akt/protein kinase b (pkb) in skeletal muscle cells from type 2 diabetic patients. *Diabetologia*, *51*(3), 512–521.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, *227*(5258), 561–563.

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *Int. J. Complex Syst.*, *1695*.

Dalle Pezze, P., Sonntag, A., Thien, A., Prentzell, M., Godel, M., Fischer, S., Neumann-Haefelin, E., Huber, T., Baumeister, R., Shanley, D., et al. (2012). A dynamic network model of mtor signaling reveals tsc-independent mtorc2 regulation. *Science Signalling*, *5*(217), ra25.

De Palo, G., Eduati, F., Zampieri, M., Di Camillo, B., Toffolo, G., & Altafini, C. (2011). Adaptation as a genome-wide autoregulatory principle in the stress response of yeast. *Systems Biology, IET*, *5*(4), 269–279.

DeRisi, J., Iyer, V., & Brown, P. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, *278*(5338), 680–686.

Di Camillo, B., Sanchez-Cabo, F., Toffolo, G., Nair, S., Trajanoski, Z., & Cobelli, C. (2005). A quantization method based on threshold optimization for microarray short time series. *Bmc Bioinformatics*, *6*(Suppl 4), S11.

Di Camillo, B., Toffolo, G., Nair, S., Greenlund, L., & Cobelli, C. (2007). Significance analysis of microarray transcript levels in time series experiments. *BMC bioinformatics*, *8*(Suppl 1), S10.

Du, T., & Zamore, P. D. (2005). microprimer: the biogenesis and function of microrna. *Development (Cambridge, England)*, *132*(21), 4645–4652.

Eduati, F., Corradin, A., Di Camillo, B., & Toffolo, G. (2010). A boolean approach to linear prediction for signaling network modeling. *PloS one*, *5*(9), e12789.

Eduati, F., De Las Rivas, J., Di Camillo, B., Toffolo, G., & Saez-Rodriguez, J. (2012a). Integrating literature-constrained and data-driven inference of signalling networks. *Bioinformatics*, *28*(18), 2311–2317.

Eduati, F., Di Camillo, B., Karbiener, M., Scheideler, M., Corà, D., Caselle, M., & Toffolo, G. (2012b). Dynamic modeling of mirna-mediated feed-forward loops. *Journal of Computational Biology*, *19*(2), 188–199.

Eduati, F., Di Camillo, B., & Toffolo, G. (Submitted). Dynamic analysis of leucine effects on insulin activated akt/gsk3β signalling pathway in human skeletal muscle cells. .

Egea, J., Martí, R., & Banga, J. (2010). An evolutionary method for complex-process optimization. *Computers & Operations Research*, *37*(2), 315–324.

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

Farina, L., De Santis, A., Salvucci, S., Morelli, G., & Ruberti, I. (2008). Embedding mrna stability in correlation analysis of time-series gene expression data. *PLoS computational biology*, *4*(8), e1000141.

Feret, J., Danos, V., Krivine, J., Harmer, R., & Fontana, W. (2009). Internal coarse-graining of molecular systems. *Proceedings of the National Academy of Sciences*, *106*(16), 6453–6458.

Foat, B., Houshmandi, S., Olivas, W., & Bussemaker, H. (2005). Profiling condition-specific, genome-wide regulation of mrna stability in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(49), 17675–17680.

Friard, O., Re, A., Taverna, D., Bortoli, M. D., & Cora, D. (2010). Circuitsdb: a database of mixed microrna/transcription factor feed-forward regulatory circuits in human and mouse. *BMC bioinformatics*, *11*, 435.

Friedman, R. C., Farh, K. K., Burge, C. B., & Bartel, D. P. (2009). Most mammalian mrnas are conserved targets of micrornas. *Genome research*, *19*(1), 92–105.

Gasch, A., Spellman, P., Kao, C., Carmel-Harel, O., Eisen, M., Storz, G., Botstein, D., & Brown, P. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell*, *11*(12), 4241.

Gasch, A. P. (2002). The environmental stress response: a common yeast response to environmental stresses. In S. Hohmann, & P. Mager (Eds.) *Yeast Stress Responses*, (pp. 11–70). Springer-Verlag, Heidelberg.

Ghaemmaghami, S., Huh, W., Bower, K., Howson, R., Belle, A., Dephoure, N., O'Shea, E., Weissman, J., et al. (2003). Global analysis of protein expression in yeast. *Nature*, *425*(6959), 737–741.

Glaab, E., Baudot, A., Krasnogor, N., & Valencia, A. (2010). Extending pathways and processes using molecular interaction networks to analyse cancer genome data. *BMC bioinformatics*, *11*, 597.

Goldberger, R. F. (1974). Autogenous regulation of gene expression. *Science*, *183*, 810 – 816.

Greiwe, J., Kwon, G., McDaniel, M., & Semenkovich, C. (2001). Leucine and insulin activate p70 s6 kinase through different pathways in human skeletal muscle. *American Journal of Physiology-Endocrinology And Metabolism*, *281*(3), E466–E471.

Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A., & Enright, A. J. (2006). mirbase: microrna sequences, targets and gene nomenclature. *Nucleic acids research*, *34*(Database issue), D140–4.

Grigull, J., Mnaimneh, S., Pootoolal, J., Robinson, M., & Hughes, T. (2004). Genome-wide analysis of mrna stability using transcription inhibitors and microarrays reveals posttranscriptional control of ribosome biogenesis factors. *Molecular and cellular biology*, *24*(12), 5534–5547.

Guziolowski, C., Videla, S., Eduati, F., Thiele, S., Cokelaer, T., Siegel, A., & Saez-Rodriguez, J. (Submitted). Exaustively characterizing feasible logic models of a signaling network using answer set programming. .

Hargrove, J., & Schmidt, F. (1989). The role of mrna and protein stability in gene expression. *The FASEB Journal*, *3*(12), 2360–2370.

Hers, I., Vincent, E., & Tavaré, J. (2011). Akt signalling in health and disease. *Cellular signalling*, *23*(10), 1515–1527.

Hornstein, E., & Shomron, N. (2006). Canalization of development by micrornas. *Nature genetics*, *38 Suppl*, S20–4.

Ideker, T., Galitski, T., & Hood, L. (2001). A new approach to decoding life: systems biology. *Annual review of genomics and human genetics*, *2*(1), 343–372.

Ideker, T., & Sharan, R. (2008). Protein networks in disease. *Genome research*, *18*(4), 644–652.

Inzucchi, S. (2002). Oral antihyperglycemic therapy for type 2 diabetes. *JAMA: the journal of the American Medical Association*, *287*(3), 360–372.

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., Li, M., Wang, G., & Liu, Y. (2009). mir2disease: a manually curated database for microrna deregulation in human disease. *Nucleic acids research*, *37*(Database issue), D98–104.

Jones, D. (2008). Pathways to cancer therapy. *Nature Reviews Drug Discovery*, *7*(11), 875–876.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G. R., Wu, G. R., Matthews, L., Lewis, S., Birney, E., & Stein, L. (2005). Reactome: a knowledgebase of biological pathways. *Nucleic acids research*, *33*(Database issue), D428–32.

Jung, C. H., Hansen, M. A., Makunin, I. V., Korbie, D. J., & Mattick, J. S. (2010). Identification of novel non-coding rnas using profiles of short sequence reads from next generation sequencing data. *BMC genomics*, *11*, 77.

Kalir, S., Mangan, S., & Alon, U. (2005). A coherent feed-forward loop with a sum input function prolongs flagella expression in escherichia coli. *Molecular systems biology*, *1*, 2005.0006.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., & Hirakawa, M. (2010). Kegg for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids research*, *38*(suppl 1), D355–D360.

Kaplan, S., Bren, A., Dekel, E., & Alon, U. (2008). The incoherent feed-forward loop can generate non-monotonic input functions for genes. *Molecular systems biology*, *4*, 203.

Karbiener, M., Fischer, C., Nowitsch, S., Opriessnig, P., Papak, C., Ailhaud, G., Dani, C., Amri, E., & Scheideler, M. (2009). microrna mir-27b impairs human adipocyte differentiation and targets pparγ. *Biochemical and biophysical research communications*, *390*(2), 247–251.

Khanin, R., & Vinciotti, V. (2008). Computational modeling of post-transcriptional gene regulation by micrornas. *Journal of computational biology : a journal of computational molecular cell biology*, *15*(3), 305–316.

Khanin, R., Vinciotti, V., & Wit, E. (2006). Reconstructing repressor protein levels from expression of gene targets in escherichia coli. *Proceedings of the National Academy of Sciences*, *103*(49), 18592–18596.

Kimball, S., & Jefferson, L. (2006). Signaling pathways and molecular mechanisms through which branched-chain amino acids mediate translational control of protein synthesis. *The Journal of nutrition*, *136*(1), 227S–231S.

Kitano, H. (2002). Systems biology: a brief overview. *Science*, *295*(5560), 1662–1664.

Klipp, E., Liebermeister, W., Wierling, C., Kowald, A., Lehrach, H., & Herwig, R. (2011). *Systems biology*. Wiley-VCH.

Krumsiek, J., Pölsterl, S., Wittmann, D., & Theis, F. (2010). Odefy-from discrete to continuous models. *BMC bioinformatics*, *11*(1), 233.

Kuai, L., Das, B., & Sherman, F. (2005). A nuclear degradation pathway controls the abundance of normal mrnas in saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(39), 13962–13967.

Kuhn, K., DeRisi, J., Brown, P., & Sarnow, P. (2001). Global and specific translational regulation in the genomic response of saccharomyces cerevisiae to a rapid transfer from a fermentable to a nonfermentable carbon source. *Molecular and cellular biology*, *21*(3), 916–927.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., & Tuschl, T. (2001). Identification of novel genes coding for small expressed rnas. *Science (New York, N.Y.)*, *294*(5543), 853–858.

Lagos-Quintana, M., Rauhut, R., Meyer, J., Borkhardt, A., & Tuschl, T. (2003). New micrornas from mouse and human. *RNA (New York, N.Y.)*, *9*(2), 175–179.

Lawrence, N., Girolami, M., Rattray, M., & Sanguinetti, G. (2010). *Learning and inference in computational systems biology*. MIT Press.

Layman, D., & Walker, D. (2006). Potential importance of leucine in treatment of obesity and the metabolic syndrome. *The Journal of nutrition*, *136*(1), 319S–323S.

LeRoith, D., Olefsky, J., & Taylor, S. (2003). *Diabetes mellitus: a fundamental and clinical text*. Lippincott Williams & Wilkins.

Levine, E., Jacob, E. B., & Levine, H. (2007a). Target-specific and global effectors in gene regulation by microrna. *Biophysical journal*, *93*(11), L52–4.

Levine, E., Zhang, Z., Kuhlman, T., & Hwa, T. (2007b). Quantitative characteristics of gene regulation by small rna. *PLoS biology*, *5*(9), e229.

Levy, S., Ihmels, J., Carmi, M., Weinberger, A., Friedlander, G., & Barkai, N. (2007). Strategy of transcription regulation in the budding yeast. *PLoS One*, *2*(2), e250.

Lewis, B. P., Burge, C. B., & Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. *Cell*, *120*(1), 15–20.

Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P., & Burge, C. B. (2003). Prediction of mammalian microrna targets. *Cell*, *115*(7), 787–798.

Li, F., Thiele, I., Jamshidi, N., & Palsson, B. (2009a). Identification of potential pathway mediation targets in toll-like receptor signaling. *PLoS computational biology*, *5*(2), e1000292.

Li, X., Cassidy, J. J., Reinke, C. A., Fischboeck, S., & Carthew, R. W. (2009b). A microrna imparts robustness against environmental fluctuation during development. *Cell*, *137*(2), 273–282.

Luscombe, N., Babu, M., Yu, H., Snyder, M., Teichmann, S., & Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, *431*(7006), 308–312.

Ma'ayan, A., Jenkins, S. L., Neves, S., Hasseldine, A., Grace, E., Dubin-Thaler, B., Eungdamrong, N. J., Weng, G., Ram, P. T., Rice, J. J., Kershenbaum, A., Stolovitzky, G. A., Blitzer, R. D., & Iyengar, R. (2005). Formation of regulatory patterns during signal propagation in a mammalian cellular network. *Science (New York, N.Y.)*, *309*(5737), 1078–1083.

Macia, J., Widder, S., & Sole, R. (2009). Specialized or flexible feed-forward loop motifs: a question of topology. *BMC systems biology*, *3*, 84.

Macotela, Y., Emanuelli, B., Bång, A., Espinoza, D., Boucher, J., Beebe, K., Gall, W., & Kahn, C. (2011). Dietary leucine-an environmental modifier of insulin resistance acting on multiple levels of metabolism. *PLoS One*, *6*(6), e21187.

Mangan, S., & Alon, U. (2003). Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(21), 11980–11985.

Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Floreano, D., & Stolovitzky, G. (2010). Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(14), 6286–6291.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., & Califano, A. (2006). Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, *7 Suppl 1*, S7.

Markowetz, F. (2010). How to understand the cell by breaking it: network analysis of gene perturbation screens. *PLoS computational biology*, *6*(2), e1000655.

McManus, E., Sakamoto, K., Armit, L., Ronaldson, L., Shpiro, N., Marquez, R., & Alessi, D. (2005). Role that phosphorylation of gsk3 plays in insulin and wnt signalling defined by knockin analysis. *The EMBO journal*, *24*(8), 1571–1583.

Meijer, L., Flajolet, M., & Greengard, P. (2004). Pharmacological inhibitors of glycogen synthase kinase 3. *Trends in pharmacological sciences*, *25*(9), 471–480.

Meyer, P. E., Lafitte, F., & Bontempi, G. (2008). minet: A r/bioconductor package for inferring large transcriptional networks using mutual information. *BMC bioinformatics*, *9*, 461.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., & Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science (New York, N.Y.)*, *298*(5594), 824–827.

Mitsos, A., Melas, I., Siminelakis, P., Chairakaki, A., Saez-Rodriguez, J., & Alexopoulos, L. (2009). Identifying drug effects via pathway alterations using an integer linear programming optimization formulation on phosphoproteomic data. *PLoS computational biology*, *5*(12), e1000591.

Moller, D., et al. (2001). New drug targets for type 2 diabetes and the metabolic syndrome. *Nature*, *414*(6865), 821–827.

Morange, P. E., Alessi, M. C., Verdier, M., Casanova, D., Magalon, G., & Juhan-Vague, I. (1999). Pai-1 produced ex vivo by human adipose tissue is relevant to pai-1 blood level. *Arteriosclerosis, Thrombosis, and Vascular Biology*, *19*(5), 1361–1365.

Morris, M. K., Saez-Rodriguez, J., Sorger, P. K., & Lauffenburger, D. A. (2010). Logic-based models for the analysis of cell signaling networks. *Biochemistry*, *49*(15), 3216–3224.

Mukherjee, S., & Speed, T. P. (2008). Network inference using informative priors. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(38), 14313–14318.

Nair, K., & Short, K. (2005). Hormonal and signaling role of branched-chain amino acids. *The Journal of nutrition*, *135*(6), 1547S–1552S.

Newgard, C., An, J., Bain, J., Muehlbauer, M., Stevens, R., Lien, L., Haqq, A., Shah, S., Arlotto, M., Slentz, C., et al. (2009). A branched-chain amino acid-related metabolic signature that differentiates obese and lean humans and contributes to insulin resistance. *Cell metabolism*, *9*(4), 311–326.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., & Kanehisa, M. (1999). Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, *27*(1), 29–34.

Osella, M., Bosia, C., CorÃ , D., & Caselle, M. (2011). The role of incoherent microrna-mediated feedforward loops in noise buffering. *PLoS computational biology*, *7*(3), e1001101.

Pe'er, D. (2005). Bayesian network analysis of signaling networks: a primer. *Science's STKE : signal transduction knowledge environment*, *2005*(281), pl4.

Pérez-Ortín, J., Alepuz, P., & Moreno, J. (2007). Genomics and gene transcription kinetics in yeast. *TRENDS in Genetics*, *23*(5), 250–257.

Pico, A., Kelder, T., Van Iersel, M., Hanspers, K., Conklin, B., & Evelo, C. (2008). Wikipathways: pathway editing for the people. *PLoS biology*, *6*(7), e184.

Pieroni, E., de la Fuente van Bentem, S., Mancosu, G., Capobianco, E., Hirt, H., & de la Fuente, A. (2008). Protein networking: insights into global functional organization of proteomes. *Proteomics*, *8*(4), 799–816.

Pilpel, Y., Sudarsanam, P., & Church, G. (2001). Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature genetics*, *29*(2), 153–159.

Preiss, T., Baron-Benhamou, J., Ansorge, W., & Hentze, M. (2003). Homodirectional changes in transcriptome composition and mrna translation induced by rapamycin and heat shock. *Nature Structural & Molecular Biology*, *10*(12), 1039–1047.

Prieto, C., & Rivas, J. D. L. (2006). Apid: Agile protein interaction dataanalyzer. *Nucleic acids research*, *34*(Web Server issue), W298–302.

Prill, R., Marbach, D., Saez-Rodriguez, J., Sorger, P., Alexopoulos, L., Xue, X., Clarke, N., Altan-Bonnet, G., & Stolovitzky, G. (2010). Towards a rigorous assessment of systems biology models: the dream3 challenges. *PloS one*, *5*(2), e9202.

Prill, R., Saez-Rodriguez, J., Alexopoulos, L., Sorger, P., & Stolovitzky, G. (2011). Crowdsourcing network inference: the dream predictive signaling network challenge. *Science Signalling*, *4*(189), mr7.

Re, A., Cora, D., Taverna, D., & Caselle, M. (2009). Genome-wide survey of microrna-transcription factor feed-forward regulatory circuits in human. *Molecular bioSystems*, *5*(8), 854–867.

Rivas, J. D. L., & Fontanillo, C. (2010). Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS computational biology*, *6*(6), e1000807.

Ronen, M., & Botstein, D. (2006). Transcriptional response of steady-state yeast cultures to transient perturbations in carbon source. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(2), 389–394.

Ronen, M., Rosenberg, R., Shraiman, B., & Alon, U. (2002). Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proceedings of the National Academy of Sciences*, *99*(16), 10555–10560.

Rosenfeld, N., Elowitz, M., Alon, U., et al. (2002). Negative autoregulation speeds the response times of transcription networks. *Journal of molecular biology*, *323*(5), 785–793.

Saadatpour, A., Wang, R. S., Liao, A., Liu, X., Loughran, T. P., Albert, I., & Albert, R. (2011). Dynamical and structural analysis of a t cell survival network identifies novel candidate therapeutic targets for large granular lymphocyte leukemia. *PLoS computational biology*, *7*(11), e1002267.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D., & Nolan, G. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science Signalling*, *308*(5721), 523.

Saez-Rodriguez, J., Alexopoulos, L., Epperlein, J., Samaga, R., Lauffenburger, D., Klamt, S., & Sorger, P. (2009). Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. *Molecular systems biology*, *5*(1).

Saez-Rodriguez, J., Goldsipe, A., Muhlich, J., Alexopoulos, L., Millard, B., Lauffenburger, D., & Sorger, P. (2008). Flexible informatics for linking experimental data to mathematical models via datarail. *Bioinformatics*, *24*(6), 840–847.

Saltiel, A., & Kahn, C. (2001). Insulin signalling and the regulation of glucose and lipid metabolism. *Nature*, *414*(6865), 799–806.

Samaga, R., Saez-Rodriguez, J., Alexopoulos, L. G., Sorger, P. K., & Klamt, S. (2009). The logic of egfr/erbb signaling: theoretical properties and analysis of high-throughput data. *PLoS computational biology*, *5*(8), e1000438.

Savageau, M. (1974). Comparison of classical and autogenous systems of regulation in inducible operons. *Nature*, *252*, 546–549.

Schaefer, C., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., & Buetow, K. (2009). Pid: the pathway interaction database. *Nucleic acids research*, *37*(suppl 1), D674–D679.

Scheideler, M., Elabd, C., Zaragosi, L. E., Chiellini, C., Hackl, H., Sanchez-Cabo, F., Yadav, S., Duszka, K., Friedl, G., Papak, C., Prokesch, A., Windhager, R., Ailhaud, G., Dani, C., Amri, E. Z., & Trajanoski, Z. (2008). Comparative transcriptomics of human multipotent stem cells during adipogenesis and osteoblastogenesis. *BMC genomics*, *9*, 340.

Schlessinger, J. Epidermal growth factor receptor pathway. sci. signal.(connections map in the database of cell signaling, as seen 3 january 2013). `http://stke.sciencemag.org/cgi/cm/stkecm;CMP_1498`.

Schoeberl, B., Eichler-Jonsson, C., Gilles, E., & Muller, G. (2002). Computational modeling of the dynamics of the map kinase cascade activated by surface and internalized egf receptors. *Nature biotechnology*, *20*(4), 370–375.

Sedaghat, A., Sherman, A., & Quon, M. (2002). A mathematical model of metabolic insulin signaling pathways. *American Journal of Physiology-Endocrinology and Metabolism*, *283*(5), E1084–E1101.

Selbach, M., Schwanhausser, B., Thierfelder, N., Fang, Z., Khanin, R., & Rajewsky, N. (2008). Widespread changes in protein synthesis induced by micrornas. *Nature*, *455*(7209), 58–63.

Shalem, O., Dahan, O., Levo, M., Martinez, M., Furman, I., Segal, E., & Pilpel, Y. (2008). Transient transcriptional responses to stress are generated by opposing effects of mrna production and degradation. *Molecular systems biology*, *4*(1).

Shalgi, R., Lieber, D., Oren, M., & Pilpel, Y. (2007). Global and local architecture of the mammalian microrna-transcription factor regulatory network. *PLoS computational biology*, *3*(7), e131.

Shen-Orr, S. S., Milo, R., Mangan, S., & Alon, U. (2002). Network motifs in the transcriptional regulation network of escherichia coli. *Nature genetics*, *31*(1), 64–68.

Shimoni, Y., Friedlander, G., Hetzroni, G., Niv, G., Altuvia, S., Biham, O., & Margalit, H. (2007). Regulation of gene expression by small non-coding rnas: a quantitative view. *Molecular systems biology*, *3*, 138.

Simpson, M., Cox, C., & Sayler, G. (2003). Frequency domain analysis of noise in autoregulated gene circuits. *Proceedings of the National Academy of Sciences*, *100*(8), 4551–4556.

Soranzo, N., Zampieri, M., Farina, L., & Altafini, C. (2009). mrna stability and the unfolding of gene expression in the long-period yeast metabolic cycle. *BMC systems biology*, *3*(1), 18.

Stolovitzky, G., Monroe, D., & Califano, A. (2007). Dialogue on reverse-engineering assessment and methods. *Annals of the New York Academy of Sciences*, *1115*(1), 1–22.

Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, E., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545–15550.

Taniguchi, C., Emanuelli, B., & Kahn, C. (2006). Critical nodes in signalling pathways: insights into insulin action. *Nature Reviews Molecular Cell Biology*, *7*(2), 85–96.

Terfve, C., Cokelaer, T., Henriques, D., MacNamara, A., Goncalves, E., Morris, M., van Iersel, M., Lauffenburger, D., & Saez-Rodrigues, J. (2012). Cellnoptr: a flexible toolkit to train protein signaling networks to data using multiple logic formalisms. *BMC Syst Biol*, *6*, 133.

Terfve, C., & Saez-Rodriguez, J. (2012). Modeling signaling networks using high-throughput phospho-proteomics. *Advances in Experimental Medicine and Biology*, *736*, 19–57.

Tirosh, I., Weinberger, A., Carmi, M., & Barkai, N. (2006). A genetic signature of interspecies variations in gene expression. *Nature genetics*, *38*(7), 830–834.

Torre, V., Ashmore, J., Lamb, T., & Menini, A. (1995). Transduction and adaptation in sensory receptor cells. *The Journal of neuroscience*, *15*(12), 7757–7768.

Tremblay, F., Lavigne, C., Jacques, H., & Marette, A. (2007). Role of dietary proteins and amino acids in the pathogenesis of insulin resistance. *Annu. Rev. Nutr.*, *27*, 293–310.

Tsang, J., Zhu, J., & van Oudenaarden, A. (2007). Microrna-mediated feedback and feedforward loops are recurrent network motifs in mammals. *Molecular cell*, *26*(5), 753–767.

Tu, B., Kudlicki, A., Rowicka, M., & McKnight, S. (2005). Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. *Science*, *310*(5751), 1152–1158.

Um, S., D'Alessio, D., & Thomas, G. (2006). Nutrient overload, insulin resistance, and ribosomal protein s6 kinase 1, s6k1. *Cell metabolism*, *3*(6), 393–402.

Videla, S., Guziolowski, C., Eduati, F., Thiele, S., Grabe, N., Saez-Rodriguez, J., & Siegel, A. (2012). Revisiting the training of logic models of protein signaling networks with asp. In *Computational Methods in Systems Biology*, vol. 0 of *Lecture Notes in Computer Science*, (pp. 342–361). Springer Berlin / Heidelberg.

Vinayagam, A., Stelzl, U., Foulle, R., Plassmann, S., Zenkner, M., Timm, J., Assmus, H. E., Andrade-Navarro, M. A., & Wanker, E. E. (2011). A directed protein interaction network for investigating intracellular signal transduction. *Science signaling*, *4*(189), rs8.

Vohradsky, J., Panek, J., & Vomastek, T. (2010). Numerical modelling of microrna-mediated mrna decay identifies novel mechanism of microrna controlled mrna downregulation. *Nucleic acids research*, *38*(14), 4579–4585.

Vu, T. T., & Vohradsky, J. (2007). Nonlinear differential equation model for quantification of transcriptional regulation applied to microarray data of saccharomyces cerevisiae. *Nucleic acids research*, *35*(1), 279–287.

Wang, Y., Liu, C., Storey, J., Tibshirani, R., Herschlag, D., & Brown, P. (2002). Precision and functional specificity in mrna decay. *Proceedings of the National Academy of Sciences*, *99*(9), 5860–5865.

Watterson, S., Marshall, S., & Ghazal, P. (2008). Logic models of pathway biology. *Drug discovery today*, *13*(9-10), 447–456.

Wiener, N. (1948). Cybernetics; or control and communication in the animal and the machine. *John Wiley*.

Wittmann, D., Krumsiek, J., Saez-Rodriguez, J., Lauffenburger, D., Klamt, S., & Theis, F. (2009). Transforming boolean models to continuous models: methodology and application to t-cell receptor signaling. *BMC systems biology*, *3*(1), 98.

Wu, C. I., Shen, Y., & Tang, T. (2009). Evolution under canalization and the dual roles of micrornas: a hypothesis. *Genome research*, *19*(5), 734–743.

Yi, T., Huang, Y., Simon, M., & Doyle, J. (2000). Robust perfect adaptation in bacterial chemotaxis through integral feedback control. *Proceedings of the National Academy of Sciences*, *97*(9), 4649–4653.

Yoshimoto, H., Saltsman, K., Gasch, A., Li, H., Ogawa, N., Botstein, D., Brown, P., & Cyert, M. (2002). Genome-wide analysis of gene expression regulated by the calcineurin/crz1p signaling pathway in saccharomyces cerevisiae. *Journal of Biological Chemistry*, *277*(34), 31079–31088.

Zeanandin, G., Balage, M., Schneider, S., Dupont, J., Hébuterne, X., Mothe-Satney, I., & Dardevet, D. (2012). Differential effect of long-term leucine supplementation on skeletal muscle and adipose tissue in old rats: an insulin signaling pathway approach. *Age*, *34*(2), 371–387.

Zhang, Y., Guo, K., LeBlanc, R., Loh, D., Schwartz, G., & Yu, Y. (2007). Increasing dietary leucine intake reduces diet-induced obesity and improves glucose and cholesterol metabolism in mice via multimechanisms. *Diabetes*, *56*(6), 1647–1654.