UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Ingegneria dell'Informazione

Scuola di Dottorato di Ricerca in Ingegneria dell'Informazione

Indirizzo: Bioingegneria

CICLO XXIII

INFERENCE OF GENE REGULATION FROM EXPRESSION DATA. MATHEMATICAL MODELING AND THE DESIGN OF A GENOMIC STUDY TO INVESTIGATE IFN-α TRANSCRIPTIONAL RESPONSE MODULATORS.

Direttore della Scuola: Ch.mo Prof. Matteo Bertocco
Supervisore: Ch.mo Prof. Gianna Maria Toffolo

Dottorando: Angela Grassi

July 2011

To My Father
for having passed to me a bit of his *ingegno* and curiosity

To My Mother
for her engagement and love for her work and family

To Gianluca
for having opened my eyes and my mind

Acknowledgements

Many people deserve my warmest thanks, for having helped and supported me during my PhD.

Gianna Toffolo, for being my supervisor and for having taught to me important things for my work and my life.

Stefano Indraccolo, for being my *biological supervisor*, for the exciting joint work of the last period of my thesis, in the true spirit of systems biology. Many novel findings, including my idea of using FFLs to reconstruct regulatory modules, stemmed from our discussions.

Ernst Wit, for having introduced me to Bayesian modelling and to the problem of inferring gene regulatory networks. He is a great statistician and his guide has been fundamental to the first part of my thesis.

Lorenzo Finesso, for his unconditional support, his friendship and his bad temper. He is a great theoretical scientist and it was an honor for me working with him.

Paola Zanovello, for giving me the opportunity to continue my work and find new collaborations at IOV.

Barbara Di Camillo, for the helpful discussions on the selection procedure and for bringing enthusiasm and innovation to the genomic group at DEI.

Francesco Ciccarese, for having performed all the biological experiment presented in the second part of the thesis. It was a pleasure following him in the several phases of his work and sharing together the endless days of kinetics (more than 12 hour in lab). This taught me a lot and, above all, the respect for the data.

All the people with whom I have collaborated during my PhD for different projects. In order of appearance: Lorenzo Finesso and Peter Spreij on approximate realization of hidden Markov models, Giovanni Pacini and Andrea Tura on a longitudinal study on gestational diabetes mellitus, Barbara Di Camillo and Luisa Barzon on the analysis of an HPV miRNA dataset, Rita Zamarchi and Elisabetta Rossi on a study of circulating tumor cells in renal cancer.

A special thank to my office mates in the last period of my thesis: Tiziana Sanavia at DEI and Maria Chiara Scaini at IOV. It would have not been the same without their support. My PhD sisters: Elena Ceseracciu, Costanza D'Avanzo, and Sara Nasso. Thanks to Federica Eduati, Marco Falda, Subhamoy Mukherjee, and Aichi Msaki for working, and not working, discussions in the last years.

I am indebted to three Institutions and specifically to the people who direct them and financially supported my work towards my PhD. Dr Ferdinando Grandori, director of ISIB-CNR, Prof Gianna Toffolo, at DEI, and Prof Paola Zanovello at IOV.

I am also indebted to the Centro Studio Invecchiamento - CNR Padova for their financial effort in supporting the Project on IFN-α developed in this thesis (PRIL 2008-2010).

Finally my family and friends for having understood and accepted the reasons of my absence.

Sommario

L'oggetto principale di questa tesi è l'inferenza di regolazioni geniche a partire da dati quantitativi di espressione genica. Questo obiettivo è di centrale importanza in ambito sanitario poiché malattie genetiche complesse come il cancro sono causate dalla deregolazione o dalla regolazione aberrante di geni.

La tesi è strutturata in due parti principali corrispondenti, rispettivamente, ad un approccio teorico e pratico all'inferenza di regolazioni geniche.

Nella la prima parte della tesi viene presentato un modello gerarchico bayesiano per la ricostruzione di reti di regolazione da dati di microarray. Le interazioni trascrizionali sono descritte attraverso un semplice modello lineare su scala logaritmica. A causa dell'elevata dimensionalità dei dati, viene imposto un vincolo topologico di tipo scale-free sulla distribuzione dell'outdegree, in accordo con le principali caratteristiche esibite da reti geniche. Per inferire la struttura della rete è stato messo a punto un algoritmo MCMC. La novità della procedura proposta risiede nell'introduzione del vincolo topologico scale-free sulla struttura della rete direttamente all'interno dell'aggiornamento MCMC della matrice di connettività. Questo nuovo aggiornamento prevede l'introduzione di una statistica χ^2 ed è direttamente collegato alle tecniche di Approximate Bayesian Computation. L'algoritmo proposto introduce in modo efficace il vincolo topologico e le prestazioni, valutate sia su dati simulati che reali, sono allineate a quelle di altri metodi di ricostruzione di reti.

Nella seconda parte della tesi viene presentato un nuovo design sperimentale per inferire moduli di regolazione genica da dati di real-time

PCR. L'obiettivo è quello di caratterizzare la risposta trascrizionale di IFN-α in cellule endoteliali umane, attraverso l'individuazione di modulatori chiave e dei moduli regolatori in cui sono coinvolti. Questo è un obiettivo molto innovativo ed ambizioso, poiché studi precedenti si sono limitati ad una descrizione 'statica' della risposta indotta da IFN-α in cellule endoteliali. Abbiamo progettato array TaqMan personalizzati che ci hanno permesso di misurare l'espressione genica di circa 90 trascritti regolati dall'IFN-α, mediante tecniche di PCR quantitativa. I geni monitorati comprendono la parte più alta della signature indotta da IFN-α a 5h e i trascritti del pathway di signalling dell'interferone. Il set-up sperimentale progettato è informativo e prevede una doppia stimolazione (IFN-α e wash-out) e perturbazioni del sistema mediante silenziamento genico di alcuni candidati modulatori. Le modulazioni significative sono state evidenziate attraverso una procedura di selezione in due fasi: viene prima applicato un filtro basato sulla varianza e poi un test statistico sulle singole variabili, con una correzione di Bonferroni per test multipli. Dai risultati dell'analisi di significatività, sono stati ricostruiti moduli regolatori a partire da feed-forward loop, i più piccoli moduli ricorrenti nelle reti biologiche.

L'approccio riduzionista utilizzato in questa tesi, passando da una prima esperienza di inferenza in un problema su larga scala, ad un problema più ristretto focalizzato su una novantina di trascritti, tra cui i geni della pathway di signalling centrale per il nostro problema, ha avuto successo. I risultati ottenuti dall'analisi del dataset generato sono di grande interesse biologico ed arricchiscono l'attuale conoscenza sull'IFN-α. Questo lavoro ha raggiunto l'obiettivo di generare diverse nuove ipotesi biologiche che sono degne di essere validate sperimentalmente. Il nuovo design sperimentale e il nuovo metodo di analisi sviluppato per ricostruire moduli di regolazione sono trasportabili e largamente applicabili a diversi problemi biologici.

Abstract

The main object of this thesis is the inference of gene regulation from quantitative gene expression data. This goal is of central importance for health care as complex genetic diseases such as cancer are caused by deregulation or aberrant regulation of genes.

The thesis is structured into two main parts corresponding to a theoretical and a practical approach to the inference of gene regulation.

In the first part of the thesis a Bayesian hierarchical model for the reconstruction of regulatory networks from gene expression microarray data is presented. Transcriptional interactions are described via a simple linear model on the logarithmic scale. Due to the typically high dimensionality of the data, a scale-free topological constraint on the outdegree distribution, in agreement with the main feature exhibited by gene networks, has been imposed. An MCMC algorithm for inferring the structure of the network was developed. The novelty of the procedure resides in the introduction of a scale-free topological constraint on the network structure directly within the MCMC update of the connectivity matrix. This new update involves the introduction of a χ^2-statistic and is connected with Approximate Bayesian Computation techniques. The proposed algorithm is effective in the introduction of the topological constraint and its performances, assessed on simulated and real data, are in line with other reconstruction methods.

In the second part of the thesis a novel experimental design to infer gene regulatory modules from real-time PCR data is presented. The aim is to characterize the IFN-α-transcriptional response in human endothelial cells, by identifying key modulators and regulatory

modules in which they are involved. This is a very innovative and ambitious goal as previously published studies have so far been limited to a 'static' description of the IFN-α response in endothelial cells. We designed customized TaqMan arrays which allowed us to measure the gene expression of about 90 validated transcripts regulated by IFN-α, by state-of-the-art quantitative PCR techniques. The monitored genes include the top of the transcriptional signature induced by IFN-α at 5h and transcripts from the IFN-α signaling pathway. We designed an informative experimental setup, with a double stimulation (IFN-α and wash-out) and perturbation of the system by RNAi silencing of few candidate modulators. The significative modulations were elicited through a two-stage selection procedure, that first filters observations by a variance based criterion and then applies a variable-by-variable statistical test procedure, with a Bonferroni multiple testing correction. Regulatory modules were reconstructed from the results of the significance analysis, starting from feed-forward loops, the smallest building modules, which are recurring throughout biological networks.

The reductionist approach used in this thesis, passing from a first experience of inference in large scale problem, to a smaller one focused on about ninety transcripts, including genes from the signal transduction pathway central to our problem, proved successful. Results obtained from the analysis of the generated dataset are of great biological interest and enrich the state-of-the-art knowledge on IFN-α. A reasonable number of new biological hypotheses that are worth to be validated has been generated by this work. The novel experimental design and the novel method of analysis developed to reconstruct regulatory modules are transportable and widely applicable to various biological problems.

Contents

List of Figures

List of Tables

Glossary

Housekeeping genes (HK) Genes with stable expression that are employed as controls in gene expression assays.

HUVEC Human Umbilical Vein Endothelial Cells, endothelial cells isolated from normal human umbilical vein.

ISGs Interferon stimulated genes, see also ISREs.

ISREs Interferon stimulated response elements are specific nucleotide sequences in the promoters of certain genes, known as ISGs.

Ligand Molecule that specifically binds the binding site of a receptor.

PCR Polymerase Chain Reaction

Real-time PCR also called quantitative PCR or qPCR is a method for determining the amount of a target sequence or gene present in a sample.

RNAi RNA interference, term coined by Fire et al. [1998]. It's a phenomenon in which small double-stranded RNA (referred as siRNA) can induce efficient sequence-specific silence of gene expression.

siRNAs Small Interfering RNAs are 21∼23-nt double-stranded RNA molecules that guide the cleavage and degradation of their cognate RNA.

Introduction

This thesis deals with the problem of inferring gene regulation from expression data. We face this problem from two different perspectives: a theoretical one and a practical one, corresponding to the two main parts of the thesis. In the first part we develop a novel statistical method to infer gene regulatory networks with topological constraints, while in the second one we describe the design, the realization and the analysis of a new genomic experiment to infer IFN-α transcriptional response modulators and some regulatory modules involving them. In this introduction the basic concepts needed for the understanding of the framework of the thesis are introduced: gene expression and gene regulation, gene regulatory networks and their topological properties, network inference methods, RNA interference silencing data. Finally, we give an overview of the organization of the thesis.

0.1 Introduction

Over the past twenty years, advances in molecular biology have led to the identification and characterization of the functional components of cells. The completion of genome sequencing projects and the availability of high-throughput data at different levels, have allowed to collect a large amount of information of information about genes, small RNAs, transcripts, proteins and metabolites in dedicated databases. The great interest in disclosing the meaning of this information for possible application in health care has led to the birth of *systems biology*, Kitano [2002], Pennisi [2003], aimed at the 1) investigation of regulation among components of cellular networks, 2) integration of statistical and computational

methods with experimental efforts. Despite many multidisciplinary researchers are working in this new area, we are still far from understanding the functioning of the *system cell* and the way in which its components interact and regulate each other. The work described in this thesis resides in the general field of *systems biology* and addresses the problem of inferring gene regulation from quantitative gene expression data.

0.2 Gene expression and gene regulation

We summarize here some basic biological concepts to introduce the reader to the concept of gene regulation; for a more extensive account see Alberts [2002].

The cell

The cell is the fundamental unit of life. All living organisms, from the simplest ones to the most organized, are made of cells. All those cells, whether from a uni-cellular creature (such as a bacterium or a yeast) or a human being, share the same basic building blocks, *proteins*. The proteins fulfill a wide range of tasks in a cell: they act as regulators of almost all intracellular processes, as selective porters on the cell membrane, as accelerators of chemical reactions and many more. The main difference between cells of the same organism performing different functions and forming different tissues is in their protein content, or proteome. Each cell of a given organism contains the same genetic information (operative instructions for manufacturing proteins), this information is encoded by the deoxyribonucleic acid (DNA) and is replicated each time a cell divides.

DNA, genes and proteins

DNA molecules are organized in one or more chromosomes. A DNA molecule consists of two long polynucleotide chains known as DNA strands, held together by hydrogen-bonds to form a double helix. Each nucleotide is composed by a sugar (deoxyribose), a phosphate group and a base that may be either adenine (A), cytosine (C), guanine (G), or thymine (T). The specific base characterizes the nucleotide and so the same capital letters (A, C, G, T) are also commonly used to denote the four different nucleotides. The two strands complement each

other in a very specific way: if A occurs at some position on one of the two strands then T will occur in the matching position on the other strand, and similarly, C will be always paired with G. One of the two strands is sufficient to characterize both strands, as an example ACGTTACCG is matched by the sequence of nucleotides TGCAATGGC. Along the DNA strands there are regions of distinguished subsequences called *genes*, typically hundreds or thousands of nucleotide long, which encode the proteins. Proteins constitute the building blocks of cells and tissues, determining the structure, function, evolution and reproduction of an organism.

Beside the DNA, in a cell we find many forms of ribonucleic acid (RNA), a single polynucleotide chain made of four different nucleotides. It differs from DNA both for the fact that RNA molecules are single stranded, and for their chemical structure. Each nucleotide contains the sugar ribose rather than deoxyribose, and the four bases are adenine (A), cytosine (C), guanine (G) and uracil(U), rather than thymine. The complementary base pairing property still holds between DNA and RNA, with the pair A-U in which U replaces T.

The vast majority of genes specify the amino acid sequence of proteins, and the RNA molecules that are copied from these genes (and ultimately direct the synthesis of proteins) are collectively called *messenger* RNA (mRNA). The final product of the remaining genes however is the RNA itself: *ribosomal* (rRNA) forms the core of ribosomes and *transfer RNA* (tRNA) forms the adaptors that select amino acids and hold them in place on a ribosome for their incorporation into proteins.

The central dogma of molecular biology

The central dogma of molecular biology states that there are two main steps in the production of a protein from the corresponding gene. In the first step, called *transcription*, a messenger RNA (mRNA) is copied from a gene using the DNA sequence as template. At the second step, called *translation*, the mRNA is translated by ribosomes to the corresponding protein.

Transcription (DNA $\rightarrow$ RNA)

An enzymatic complex called RNA polymerase binds to a specific location in

the DNA, very close to the nucleotide sequence of the target gene, and unwinds the DNA's double helix forming a local gap between the two strands. It then moves stepwise on the sense strand of DNA assembling the complementary RNA sequence, one nucleotide after the other. The transcription continues until a stop signal (a short nucleotide sequence in the genetic code) is encountered.

Eucaryotic cells, in general, are bigger and more elaborate than bacteria. Some live as single-cell organisms like yeasts; others form multicellular complex organisms, including plants, animals and fungi. All eucaryotic cells have a nucleus in which is enclosed the genetic information and a variety of organelles with specific functions in the cytoplasm. In eucaryotes, the process of transcription takes place in the nucleus and before moving to the cytoplasm mRNA undergoes several transformations, in particular is subjected to a process called *splicing*. In this process, the mRNA sequence is clipped of its non-coding regions, called *introns*. The mRNA then moves to the cytoplasm where translation takes place.

Translation (RNA $\rightarrow$ proteins)
Ribosomes attach to the mRNA and move stepwise, reading the mRNA sequence in sets of three nucleotides, called codons. Each codon specifies a particular amino acid that is needed to make the protein. Usually, 20 different types of amino acids are found in proteins but the possible codons are 4^3, meaning that more nucleotide triplets can specify the same amino acid. Specific codons in mRNA signal where to start and stop protein synthesis. Following the instructions in the genetic code, the ribosome stepwise links the amino acids into a polypeptide chain and at the stop signal releases the complete protein.

Gene expression
Together transcription and translation constitute *gene expression*. Many identical RNA copies can be made from the same gene, and each RNA molecule can instruct the synthesis of many identical protein molecules. The protein and mRNA molecule quantities of a specific gene are referred to as the gene's *protein* and *mRNA expression levels*, respectively. Gene expression is a highly regulated process by which a cell can answer to external signals and adapt to environmental

changes.

Gene regulation

Although regulation of gene expression may occur at multiple steps, for most genes initiation of transcription is the most relevant site of control. The different levels of control are sketched in Figure 1, Alberts [2002], besides them, at the RNA level, acts also the control due to microRNAs that may silence the transcript of a target gene, Ambros and Chen [2007]. Proteins may interact with each other or with signalling molecules, modifying their functionality. Transcription of each gene is switched on or off in cells by *gene regulatory proteins*, also called *Transcription Factors (TF)*, that can act either as activators (increasing the transcription rate of a gene) or repressors (reducing its transcription rate).

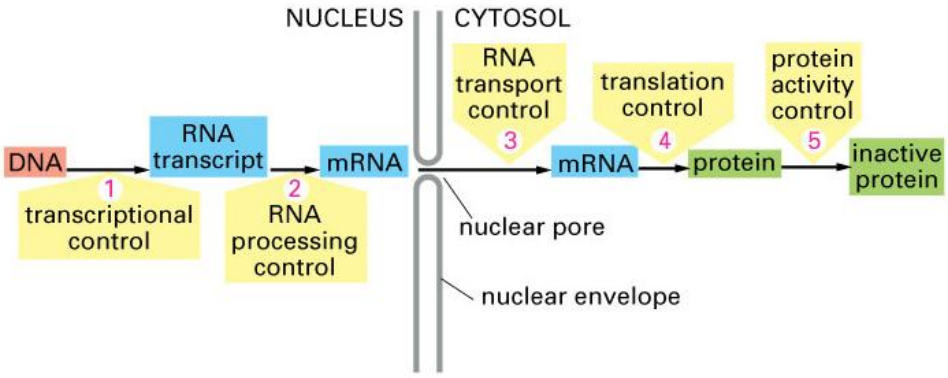

Figure 1: Eucaryotic gene expression may be controlled at several different steps. Examples of regulation at each step are known, although for most genes the main site of control is step 1: transcription from DNA to RNA. ©Alberts [2002].

0.3 Biological networks

The behavior of a living cell is regulated by complex networks of interactions between DNA, RNA, proteins and small molecules. In this thesis we concentrate mainly on *gene regulatory networks* and *signal transduction pathways*.

Gene regulatory networks

Gene (or *genetic* or *transcriptional*) *regulatory networks* describe the regulatory interactions at the RNA level. We say that gene 1 regulates gene 2 if the protein coded from gene 1 is a TF for gene 2, and we represent this interaction with an arrow from 1 to 2. In Figure 2, an example of the different levels of gene regulation and the corresponding regulatory network is illustrated. Inferring the

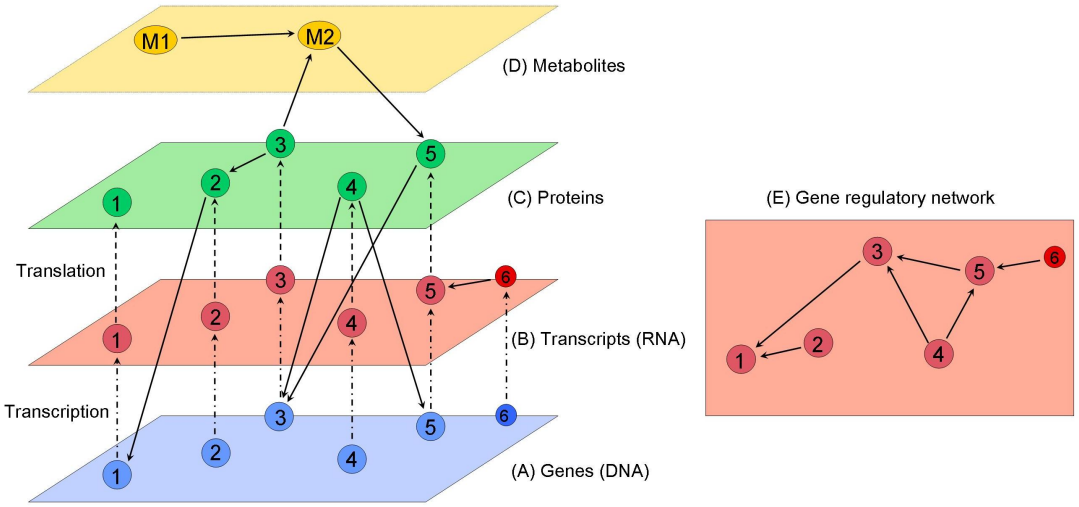

Figure 2: Different levels of gene regulation: example of a gene regulatory network. In the four plane figure (left), full arrows correspond to gene regulation at different levels; dashed arrows correspond to transcription (from genes to transcripts), translation (from transcripts to proteins). (A) Genes are transcribed into mRNA. Transcription process may be regulated by protein that are TFs (incoming arrows). (B) mRNA transcripts are translated into proteins. Transcript 6 is not translated, it corresponds to a miRNA that regulates transcript 5 by silencing it. (C) TFs bind to DNA, activating or repressing the expression of other genes. Proteins may interact also in their own plane, protein 3 regulates protein 2. Protein 3 regulates also levels of metabolite M2. (D) Active metabolites may regulate other metabolites or proteins. (E) Representation of the simplified gene regulatory network (GRN), objective of network inference starting from mRNA levels. In the GRN, protein-protein interaction may not be observed, but we see their indirect effects: protein 3 regulates transcription of gene 1 (mediated by protein 2). The regulatory module in which are involved genes 3, 4 and 5 is a feed-forward loop.

structure of such networks is one of the principal goals of functional genomics,

Hecker et al. [2009].

Signal transduction pathways
When an external signal reaches the cell membrane, it is recognized by a transmembrane receptor. Binding of a ligand to its receptor triggers an intracellular cascade, leading from the receptor to the activation of transcription. This cascade connecting external signals to a transcriptional response corresponds to a *signalling* or *signal transduction pathway*. As an example of signalling pathway, we report the type I IFN pathway that is central to the investigations performed in second part of this thesis, Figure 3.

0.3.1 Topological properties of biological networks

Biological networks are commonly represented as graphs in which nodes correspond to molecular species: genes, protein, metabolites; whereas edges are regulatory interactions. In particular, in a transcriptional network each node corresponds to a gene and departing/incoming edges correspond to genes that are regulated by/regulate that gene. The number of connections (edges) per node is called connectivity or degree of the node, whereas the number of departing and incoming edges are called out-degree and in-degree, respectively. Number, directionality and strength of connections of a given node reflect the importance and the centrality of that node. A useful measure of network topology is thus the connectivity distribution.

Sparsity.
Gene regulatory networks are sparse. In a network with N nodes there are at most $E_{max} = N^2$ possible edges (allowing also self-regulations), but the actual number of edges E is much smaller: $E/E_{max} << 1$. The sparse nature reflects the fact that only few very important interactions build up the network.

Scale-free topology.
Gene regulatory networks show many nodes with few connections and few nodes with many connections (hubs). Thus, the out-degree distribution of gene networks

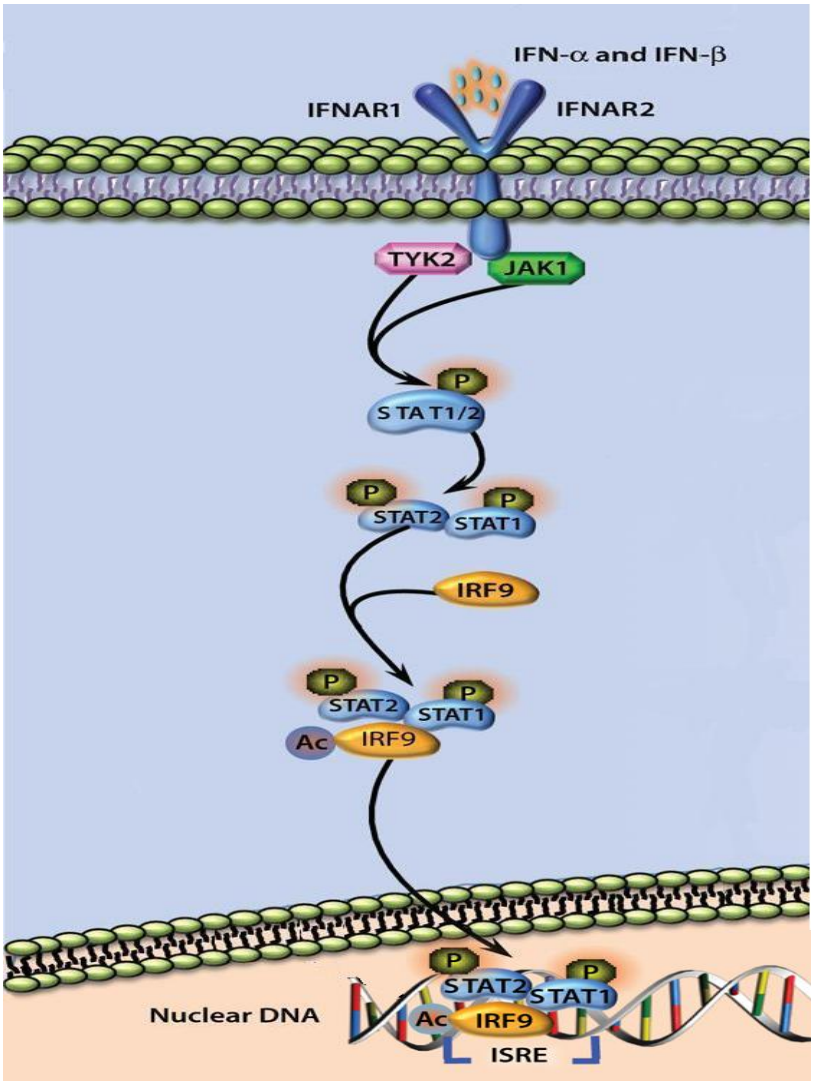

Figure 3: Type I IFN signalling pathway. IFN-α binds to a specific cell surface receptor, composed by two subunits: IFNAR1 and IFNAR2. The signal is transduced via the JAK-STAT pathway and induces transcription of hundreds of genes containing IFN stimulatory response elements (ISREs). This illustration is a detail of Fig 1 from reference Munir [2010].

has a heavy tail and can be well approximated by a power-law, Barabasi and Oltvai [2004]. The heavy-tailed distribution is often called *scale free*, and is written as

$$P(k) \propto k^{-\gamma},$$

where k is the outdegree and γ is the scaling parameter. In real world networks the scaling parameter typically lies in the range $1 < \gamma < 3$.

Small-world and high clustering coefficient.
Gene regulatory networks show a small world property, namely a small path length between any two nodes of the network. Another topological property of graphs is the clustering coefficient, which looks at the interconnections among the neighbors of a given node. These two properties are deeply connected, as the clustering coefficient quantitatively characterizes the small world property. The average clustering coefficient of gene regulatory networks is several orders of magnitude higher than that of random networks with the same average connectivity.

Network motifs.
Recurring interaction motifs, i.e. small subgraphs with well-defined topologies, have been showed to characterize gene regulatory networks, and more generally cellular networks, Milo et al. [2002], Shen-Orr et al. [2002]. These interaction motifs, such as auto-regulation and feed-forward loop (FFL), see Figure 4, have a higher frequency in biological networks compared with random networks, Erdos and Renyi [1959]. In particular, in the second part of the thesis, we focus on FFL network motifs, see Alon [2006] for a comprehensive treatment with quantitative models. Feed-forward loop is a widespread regulatory module in biology, and among the 13 possible three-node patterns, the FFL was the only one found significant (5). Introducing also the sign of regulation, arrows stand for activations while $\dashv$ for repressions, eight coherent FFL may be defined, see Figure 6

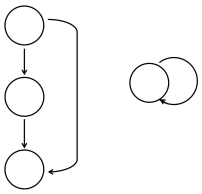

Figure 4: Example of FFL and auto-regulation motifs.

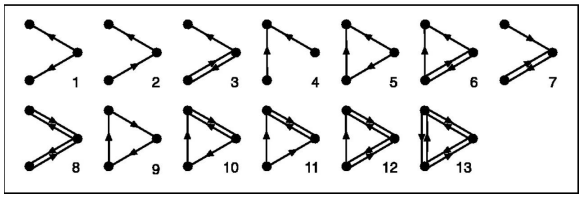

Figure 5: The 13 connected three-node directed subgraphs. This figure is taken from Milo et al. [2002]

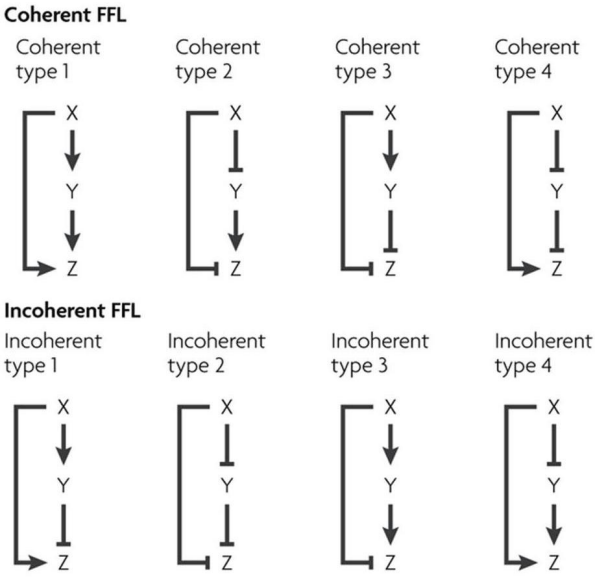

Figure 6: Feed-forward loops. The eight types of FFLs are shown. In coherent FFLs, the sign of the directed path from transcription factor X to output is the same as the overall sign of the indirect path through transcription factor Y. Incoherently FFLs have opposite signs for the two paths. This figure is taken from Alon [2007].

0.4 Inference of gene regulatory networks

In the last fifteen years, many different approaches have been proposed for the identification of gene regulatory networks based on high-throughput microarray data. Some methods reconstruct undirected networks, based on pairwise profile comparisons between genes. They use different measures of similarity, e.g. Pearson correlation, partial Pearson correlation Schäfer and Strimmer [2005], mutual information Margolin et al. [2006]. Other methods reconstruct directed graphs. Among these, a possible approach is to model the rates of change in the expression levels of different genes either with systems of differential equations (continuous case), Chen et al. [1999], or with Boolean relationships (discrete case), Shmulevich et al. [2002]. The most popular models are Bayesian networks Barenco et al. [2006], Friedman et al. [2000], I. Nachman and Friedman [2004]. They allow to represent the interactions among genes as relationships of causal independence and their stochastic nature is well suited to describe the noisy microarray data. Recent advances led to the development of dynamic Bayesian networks which are able to model the changes in the regulatory interactions over time Ferrazzi et al. [2007]. The increasing interest in this area can be judged by the sheer number of review papers published in the last ten years, e.g. Bansal et al. [2007], Gardner and Faith [2005], Markowetz and Spang [2007], Smolem et al. [2000], Soranzo et al. [2007]. Driven by this interest in the spring of 2006 a group of systems biologists at the NYAS started the Dialogue for Reverse Engineering Assessments and Methods (DREAM) initiative, Stolovitzky et al. [2007]. The main objective is to catalyze the interaction between experiment and theory in the area of cellular network inference and find fair methods for the assessment of new algorithms. The problem of reverse engineering is still an open challenge, due to its intrinsic ill-posedness, Wang et al. [2006]. The number of genes in the network is often much larger than the number of observations ($n >> p$), and that of course does not allow for unique solutions. Moreover, the gene expression process, involving transcription, translation and post-translational modifications, is too complex to be represented by the over-simplified models of gene-regulatory networks, see Figure 2. However, many lessons have been learned from the DREAM efforts and, among them, two are notable: 'topology is very important in the ability of a

method to reconstruct the network' and 'an algorithm should be designed using all the available biological knowledge of the system', Stolovitzky et al. [2009]. In order to infer the gene regulation, it becomes of central importance to plan experimental designs able to reconstruct causal relationships among genes. This kind of approach will be developed in the second part of the thesis. In the next subsection we introduce the notion of perturbation data that will be instrumental for our study.

0.4.1 Perturbation data: RNA interference gene silencing

High-throughput assays such as microarrays allow the simultaneous monitoring of thousand of transcripts. A major use of microarrays has been to measure expression changes in response to a stimulus or to specific perturbations. Examples of very informative perturbations include silencing and over-expression of genes of interest, Reimand et al. [2010], Sopko et al. [2006]. As changes in the transcriptional response are likely to be triggered by the perturbed gene, such experiments have been used to study the functions of perturbed genes and to test if they are essential to cell survival. RNA interference (RNAi) is a cellular mechanism of post-transcriptional gene silencing, Fire et al. [1998]. The discovery that small RNAs act as functional regulators is one of the most important novelties of the last decades, because it gives an active role to RNA, which was previously intended only as a mere intermediary in the gene expression process. RNAi screens are widely used in functional genomics because it takes only a few days to observe an almost complete protein depletion and, compared with the long time required for gene knockout at the DNA level, it makes RNAi silencing more suitable for genome-wide applications.

In the first part of the thesis we consider large gene regulatory networks and we develop a method of inference based on a Bayesian model that accounts for the biological knowledge about the network topology.
In the second part of the thesis, we develop a new experimental design, based on a RNAi silencing of selected genes, to infer causal relationships and reconstruct gene regulatory modules from basic building FFL motifs.

Part I

Bayesian models for inference of gene regulatory networks with topological constraints

Chapter 1

Bayesian Modelling and Inference

In this part of the thesis we propose a Bayesian model for the reconstruction of gene regulatory networks starting from time-course gene-expression profiles. Our aim is to infer the regulatory interactions between genes, accounting also for the directionality of the regulation.

1.1 Modelling gene transcription

One of the principal means of control of the behavior of the cell is the control of the gene expression process. Transcription is the process by which the DNA sequence of a gene is expressed into mRNA molecules that are then translated into proteins. The regulation of transcription is due to special proteins called Transcription Factors (TFs) that can act as activators or inhibitors of the process. We say that a gene regulates the transcription process of another gene if the protein it encodes is a transcription factor for that gene, and is present in its active form (for instance phosphorilated). Modelling the dynamics of transcription and accounting for its regulatory mechanism, requires the knowledge of a number of biological quantities: the mRNA abundance levels, the active levels of TF proteins, and a set of gene-specific constants such as the basal expression level, the rate of decay of its mRNA, and the affinity that a specific transcription factor has for the given substrate Khanin et al. [2007]. Unfortunately some of these biological quantities, such as protein activity, are not yet available on a

genome-wide scale. A common approach in modelling transcription assumes that the mRNA abundance level of a regulator approximates reasonably well the active level of the TF protein it produces. In the following subsection a simplified linear model for gene transcription is introduced.

1.1.1 Linear gene transcription model

We consider a linear gene transcription model with Gaussian error on the log-transcription scale. We assume that $y_j(t)$, the mRNA abundance level of gene j at time t, results from a multiplicative effect of the mRNA abundance levels of a collection of other genes. By considering the log-transformed data, the relationship between the log-transcription level of gene j at time t and all the others is assumed to be

$$\log y_j(t) = \sum_{i \neq j} x_{ij} \alpha_{ij} \log y_i(t) + \alpha_{0j} + \epsilon_j(t), \tag{1.1}$$

where x_{ij} is the (i, j) element of the connectivity matrix X defined as

$$x_{ij} = \begin{cases} 0 & \text{if gene } i \text{ does not regulate gene } j; \\ 1 & \text{if gene } i \text{ regulates gene } j. \end{cases}$$

The parameters α_{ij} represent the strength of interaction associated with x_{ij}, the α_{0j} is a sort of background mean expression level, and $\epsilon_j(t)$ is an i.i.d. Gaussian error with unknown variance σ^2. Models of gene transcription closer to biological reality describe the process with first order differential equations, see e.g. Appendix A.

1.2 A Bayesian hierarchical model

Starting from the available time-course gene-transcription data collected in the observation matrix, y, we aim at the reconstruction of the directed graph which represents the regulatory influences at the gene level. We assume the simplified gene transcription model (1.1) to hold and its unknown parameters being part

of our Bayesian model. The connectivity matrix X is the structural parameter, representing the algebraic counterpart of the gene interaction graph.

1.2.1 Scale-free topological constraint

The matrix X is a parameter that can be estimated from the available data. In this work we follow the approach presented in Wit and Thomson [2005], where a scale-free topological constraint was introduced in agreement with the main features exhibited by biological networks Aloy and Russell [2004]: a relatively short path length between any two nodes (*small world property)*, the presence of many genes with few connections and few highly connected genes (*hubs)*, the lethal impact for the overall architecture of the network of the deletion of a hub (*centrality and lethality principle)* Jeong et al. [2001].

As the departing connectivity of each gene, *outdegree*, has been found to follow approximately a power law Guelzim et al. [2002], in our model we impose a scale-free structure on the data via a power law prior on the outdegree:

$$P(x_{i\cdot} = k) \propto k^{-\gamma}, \tag{1.2}$$

where $x_{i\cdot} = \sum_j x_{ij}$ is the outdegree of gene i, and γ is the nonnegative scaling parameter. The way in which we actually incorporate this constraint in our model will be explained in subsection 1.3.2.

1.2.2 Model description

The hierarchical model structure is represented in the directed acyclic graph (DAG) Figure 1.1. The observed data y depend from the structural parameter X, the parameters α, and the variance σ^2. The scaling parameter of the power-law constraint appears in the hierarchical model as a hyper-parameter on X.

1.2.3 Interpretation of parameteres

The power law exponent γ can only take positive values, it represents a tuning parameter for our model; the higher gamma, the stronger the penalty on large

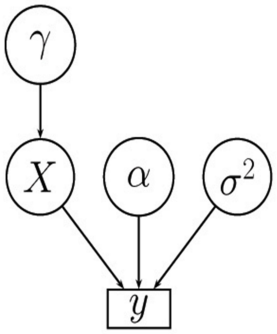

Figure 1.1: Directed acyclic graph showing the dependencies among the parameters of the model.

numbers of gene-interactions. On γ we impose a uniform prior over the range $(1, 3)$, which are the most common values quoted in the recent literature. The noise variance σ^2 is assumed to follow an Inverse Gamma prior distribution.

1.3 Markov chain Monte Carlo inference

The DAG structure of the hierarchical model (Figure 1.1) in principle makes an MCMC implementation a standard option for performing Bayesian inference. We implemented a hybrid Gibbs and Metropolis-Hastings sampler in R with a few modifications that are described below. At every sweep parameters γ and σ^2 are updated in a standard manner. Due to the dimensionality of the connectivity matrix X and linear model parameters α, efficiency considerations force us to adopt different update rules for those parameters. Given that the matrix α is not our main interest, we estimate it within every sweep via an empirical Bayes approach. This means that X and α are updated jointly, whereby a proposal for X is supplemented with a value $\hat{\alpha}(X, y)$, to wit the maximum likelihood estimate for α given the connectivities and the data. The update for X itself involves a novel idea, using a fast ratio of p-values of a relevant test-statistic rather than the more involved ratio of conditional probabilities. This approach has strong similarities with Plagnol and Tavaré [2004].

1.3.1 Algorithm

Given an arbitrary set of starting values, the basic algorithm proceeds as follows:

(A1) X update: return updated values for X, α_{0j} and α_{ij}, $\forall i, j$.

(A2) σ^2 update: return updated value for σ^2.

(A3) γ update: return updated value for γ.

At each iteration of the algorithm we do a sweep of these three steps. Informal convergence criteria suggest that the sampler converges after 5,000 iterations. We obtain 300,000 iterations and consider the output after a 5,000 sweep burn-in, with a thinning value of 25.

1.3.2 Gene interaction matrix update

As we indicated above, we use a novel MCMC procedure for the update X' of X. The reason for this is that calculating the ratio $\frac{P(X'|\gamma)}{P(X|\gamma)}$ is time-consuming. We thus introduce the scalar summary statistic

$$T_\gamma(X) = \sum_{k=1}^{N-1} \frac{(E_k - O_k)^2}{E_k},$$

where E_k are the expected counts under a power law distribution with parameter γ and O_k are the observed counts in the connectivity matrix X. This goodness-of-fit test-statistic is distributed as a $\chi^2_{(N-3)}$ and checks to what extent the data is consistent with the scale-free topology with scaling parameter the current γ. We replace the ratio $\frac{P(X'|\gamma)}{P(X|\gamma)}$ with the faster density ratio

$$\frac{p_{\chi^2_{(N-3)}}(T_\gamma(X'))}{p_{\chi^2_{(N-3)}}(T_\gamma(X))}. \tag{1.3}$$

Combining the information coming from the data, coded in the likelihood with the density ratio, we obtain the alternative acceptance probability

$$\text{AP} = \min\left\{1, \frac{p_{\chi^2_{(N-3)}}(T_\gamma(X'))}{p_{\chi^2_{(N-3)}}(T_\gamma(X))} \frac{p(y|X', \gamma, \sigma)}{p(y|X, \gamma, \sigma)}\right\}. \tag{1.4}$$

This acceptance probability guarantees that without data one would end up sampling from a network indistinguishable from one with a scale-free distribution. With data, the solution will converge to the most scale-free distribution that is consistent with the data. This approach, although novel, has direct links with Approximate Bayesian Computation Plagnol and Tavaré [2004], where for large probability calculations summary statistics are used instead.

In summary, for the X update we use the following procedure:

1. Propose a new value X' by flipping m random elements of X from 0 to 1 or viceversa, and its associated $\hat{\alpha}(y, X')$.
2. Compute
   - the likelihood with both the current value of $(X, \alpha(y, X))$ and the proposal $(X', \alpha(y, X'))$.
   - the density ratio as in equation (1.3).
3. Combine the information from Step 2 and compute the acceptance probability AP as in 1.4.
4. Accept the proposal $(X', \alpha(y, X'))$ with probability AP in step 3.
5. Proceed in the algorithm with σ^2 and γ updates.

1.4 Evaluating performance of inference

In this section we introduce the score metrics that are used to assess the performances of our algorithm in the reconstruction of genetic networks. Performances are evaluated by comparing the original (simulated or real) network structure with the inferred one, in terms of the elements of the adjacency matrices. Network prediction can be thought of as a binary classification task in which every potential network edge is classified as either present (positive, P) or absent (negative, N). A *true positive* (TP) or a *true negative* (TN) corresponds to an edge, respectively, present or absent in both the predicted and the underlying true network. A *false positive* (FP) corresponds to an edge present in the inferred network

but absent in the true one; while a *false negative* (FN) corresponds to an edge absent in the predicted network but present in the true one. The decision made by the algorithm may be summarized by a confusion matrix, see Table 1.1.

Table 1.1: Confusion matrix.

Edge	Actual Positive	Actual Negative
Inferred Positive	TP	FP
Inferred Negative	FN	TN

Binary classification is a standard paradigm in the field of machine learning with well established evaluation metrics. Usually the accuracy of the inferred networks is assessed using the following metrics: *recall* (R) (also known as *sensitivity* or *true positive rate*, TPR), *precision* (P) (also known as *positive predictive value*), *specificity* (SPC) (also known as *true negative rate*), and *false positive rate* (FPR). The *F1-measure*, defined as the harmonic mean of precision and recall with the two measures equally weighted, is a commonly used summary metric. All the performance metrics range between 0 and 1, with 1 representing the perfect matching.

$$\begin{aligned} R = TPR &= \frac{TP}{TP+FN} \\ P &= \frac{TP}{TP+FP} \\ SPC &= \frac{TN}{TN+FP} \\ FPR &= 1-SPC \\ F1-measure &= \frac{2PR}{P+R} \end{aligned}$$

As there are two types of error, it is common to simultaneously explore complementary metrics over a range of parameter values. One possible choice is precision-recall. Precision is a measure of fidelity, whereas recall is a measure of completeness. Another common choice is true positive rate (TPR) and false positive rate (FPR), the axes of the receiver operator characteristic (ROC) curve. With retrieval results ranked in decreasing order of reliability, the performance metrics at each depth can be used to construct a point in either precision-recall space or ROC space. Both representations may be useful to evaluate the quality

of network reconstruction.

Chapter 2

Data

2.1 *In silico* data

In silico data were simulated using the Netsim gene network simulator, Di Camillo et al. [2009]. By using a new hierarchical modular topology model, Netsim is able to mimic some important structural features of biological networks: scale-free topology, small-world property, high clustering coefficient (independent of the number of nodes in the network). Moreover, it integrates fuzzy logic, to describe the different types of interactions among regulators of each gene, with differential equations that allow to obtain a wide range of gene dynamics.
Using this flexible simulator, we generated 10 networks with 300 genes to assess the average behaviour of our algorithm in reconstructing scale-free networks and in hub detection. The simulated networks were generated according to a scale-free topology, with scaling parameter fixed at 2.2, they were connected graphs and auto-regulation was not allowed.
Starting from the network structures, two different types of data were generated: natural responses dynamics and knock-out (KO) steady-state data.
Natural responses correspond to time series of 101 time points in which gene profiles of each network are initialized at random and then evolve towards a steady state. From this time series 20 time points were sub-sampled in order to obtain significant dynamics, see Figure 2.1 for an example.
Knock-out data correspond to steady state after the silencing of selected highly

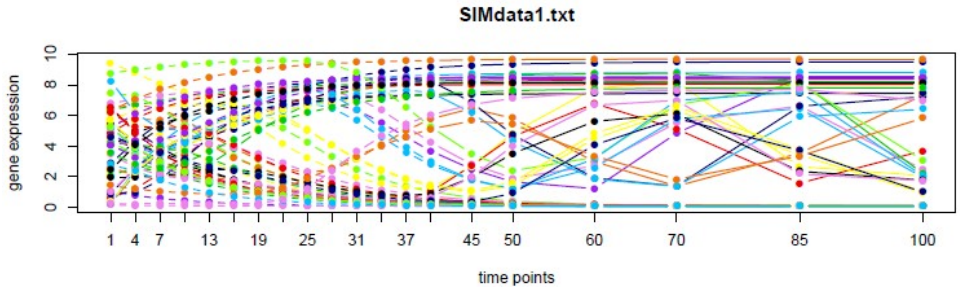

Figure 2.1: Dynamics of simulated data.

connected genes of the networks. For each network, both the wild type experiment (steady state reached after natural response) and a sample at steady state for each silenced gene were provided. The KO of a gene was realized by setting to zero its expression level and evolving the system until a steady state, while the silenced gene was kept at zero.

2.2 Real data: *Saccharomyces cerevisiae*

The algorithm was first applied to a well known dataset of yeast microarray gene expression provided by Spellman et al. [1998], which can be downloaded at http://genome-www.stanford.edu/cellcycle/data/rawdata/combined.txt. The dataset consists of four synchronization experiments (alpha factor arrest, elutriation, and arrest of CDC15 and CDC28 temperature sensitive mutants) which were performed for a total of 77 microarrays during cell-cycle. This dataset represents $\log_2$-ratios that have been normalized so that for each synchronization method, the average $\log_2$-ratio over the course of the experiments is equal to 0. We considered a specific 25-gene subnetwork of the yeast gene interaction network, the SGS1 neighbor subnetwork, described by Tong et al. [2004].

Chapter 3

Results

3.1 *In silico* data

The algorithm, detailed in section 1.3.1, was applied to the *in silico* data described in section 2.1. In our procedure we can identify two key tuning factors for the reconstruction of the network: the number m of elements of the X matrix that are changed in the proposal of a new X' at each iteration, and the final threshold on the posterior mode of each possible interaction, used to define the links of the reconstructed network. Preliminary analysis on simulated data were directed to the understanding of how these two factors affect the reconstruction performances in terms of precision, recall, and F1-measure. In the proposal of a new X' only a very low percentage of the links of X should be switched to avoid acceptance probabilities almost always zero. We compared different possible values for the m parameter, corresponding respectively to updating a percentage of 0.3%, 0.5%, and 0.8% of the full X matrix. The comparisons showed that the best choice was to update 0.3% of the links at each iteration, i.e. $m = 270$.
We explored the behavior of the algorithm, by considering the links in a ranked list according to their posterior probability and plotting them in the precision-recall space, and in the ROC space. We report in Figure 3.1 and Figure 3.2, at the end of the chapter, the performances obtained for the first six networks. The remaining four networks led to perfectly comparable plots. The results obtained by our algorithm are in line with those of other reverse-engineering methods,

Soranzo et al. [2007]. For the reconstruction of the network, a critical point of our procedure is the choice of the cut-off on the posterior probability assigned to each link. The algorithm is very sensitive to variations in the cut-off parameter, a property which can be used to set a natural threshold. These considerations are well showed by Figure 3.3 and Figure 3.4, where we plot the outdegree distribution of network 1 and network 4 at different cut-off levels. The figures also show that the out-degree distribution in consistent with the imposed scale-free topological constraint as soon as the cut-off is fixed above a minimum level.

3.2 *Saccharomyces cerevisiae* public data

We applied our algorithm to the Spellman dataset, described in section 2.2 to reconstruct the SGS1 neighbor subnetwork. These data were analyzed using a preliminary version of the algorithm and published in Grassi and Wit [2008a], to which we refer the reader for further details. The influence network inferred by our procedure is displayed in Figure 3.5, where the different intensities of the directed edges represent the magnitude of the corresponding α parameters. To establish the presence of an edge between two genes, a threshold of 0.8 on the posterior probability of X was set. In Figure 3.6, the histogram of the outdegrees for the reconstructed network is plotted. The distribution is quite skewed, in agreement with the imposed power law behavior. Figure 3.7 shows the outdegrees obtained reconstructing the network with the same model but without topological constraint, using a stepwise AIC procedure. The decaying outdegree distribution is not preserved by such method, and it typically overestimates the number of links, leading to a non-sparse network.

3.3 Discussion

Despite the simplicity of the gene interaction model, the novelty of our method resides in the introduction of a topological constraint on the structure of the connectivity matrix directly within the MCMC inference procedure. In the update of the gene interaction matrix we replace the $P(X|\gamma)$ computation based on the full interaction matrix, with the density of the chi-square checking how much the

outdegree distribution of the X matrix differs from a power-law. This update has a direct connection with Approximate Bayesian Computation methods in which, in the computation of untractable or time expensive likelihoods, the data are replaced with a lower order statistics. This novel approach needs a theoretical analysis of the convergence properties of the algorithm and in collaboration with Prof Wit, who proposed the first version of the method, we are working also on these aspects. Future work will be directed to the theoretical characterization of the procedure and to its generalization and transfer to different contexts. The gene transcription model was a simplification, but our main aim was to propose a method able to include constraints in an, in principle, easy way. Actually the presented method is very flexible and it allows also the introduction of a priori information on known links . The algorithm was tested on simulated data and a small subnetwork of the yeast interaction network and results showed that the introduction of the topological constraint was effective.

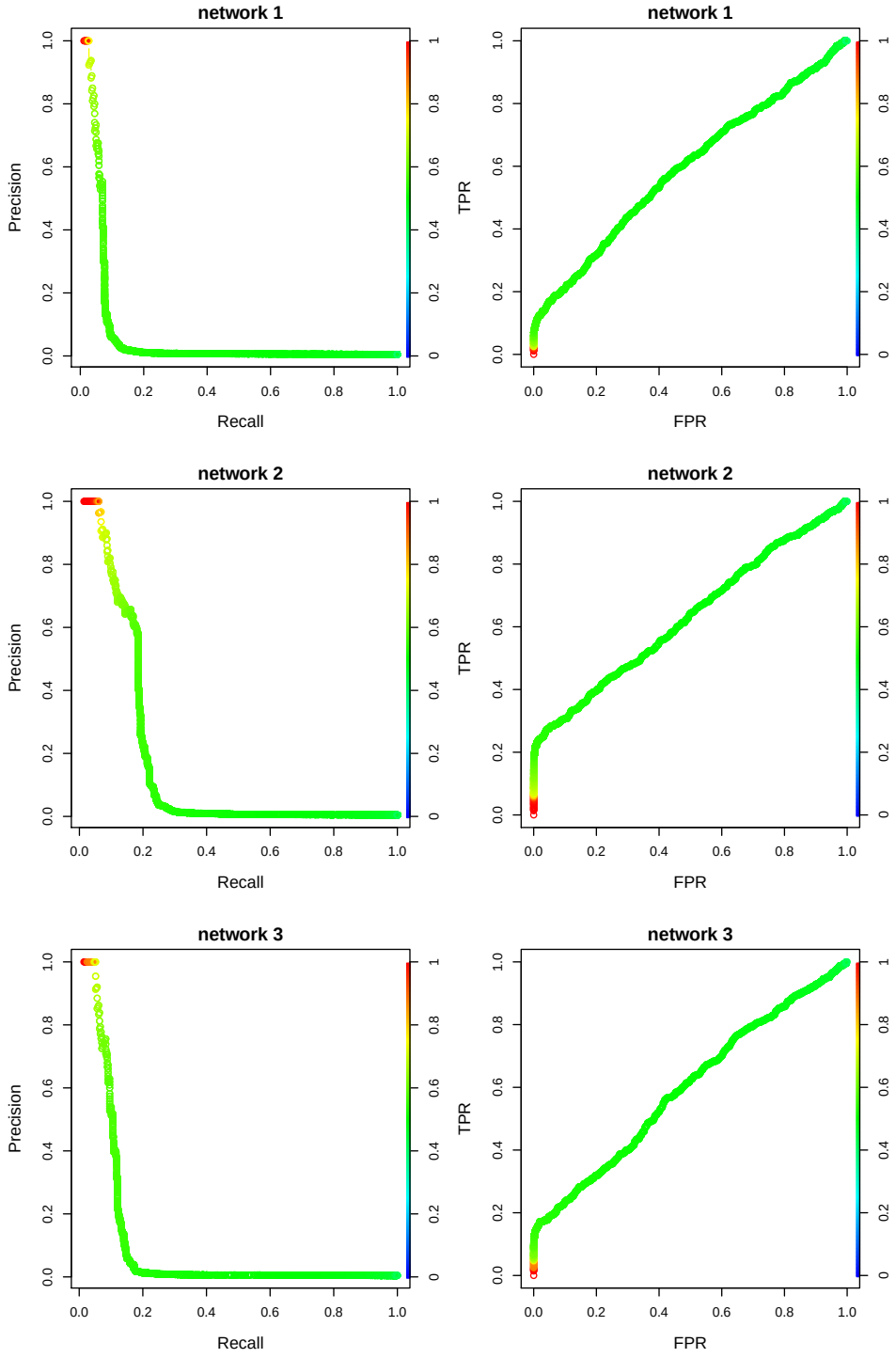

Figure 3.1: Simulated networks 1, 2, 3. Performance of the algorithm in the PR-plane and ROC curve.

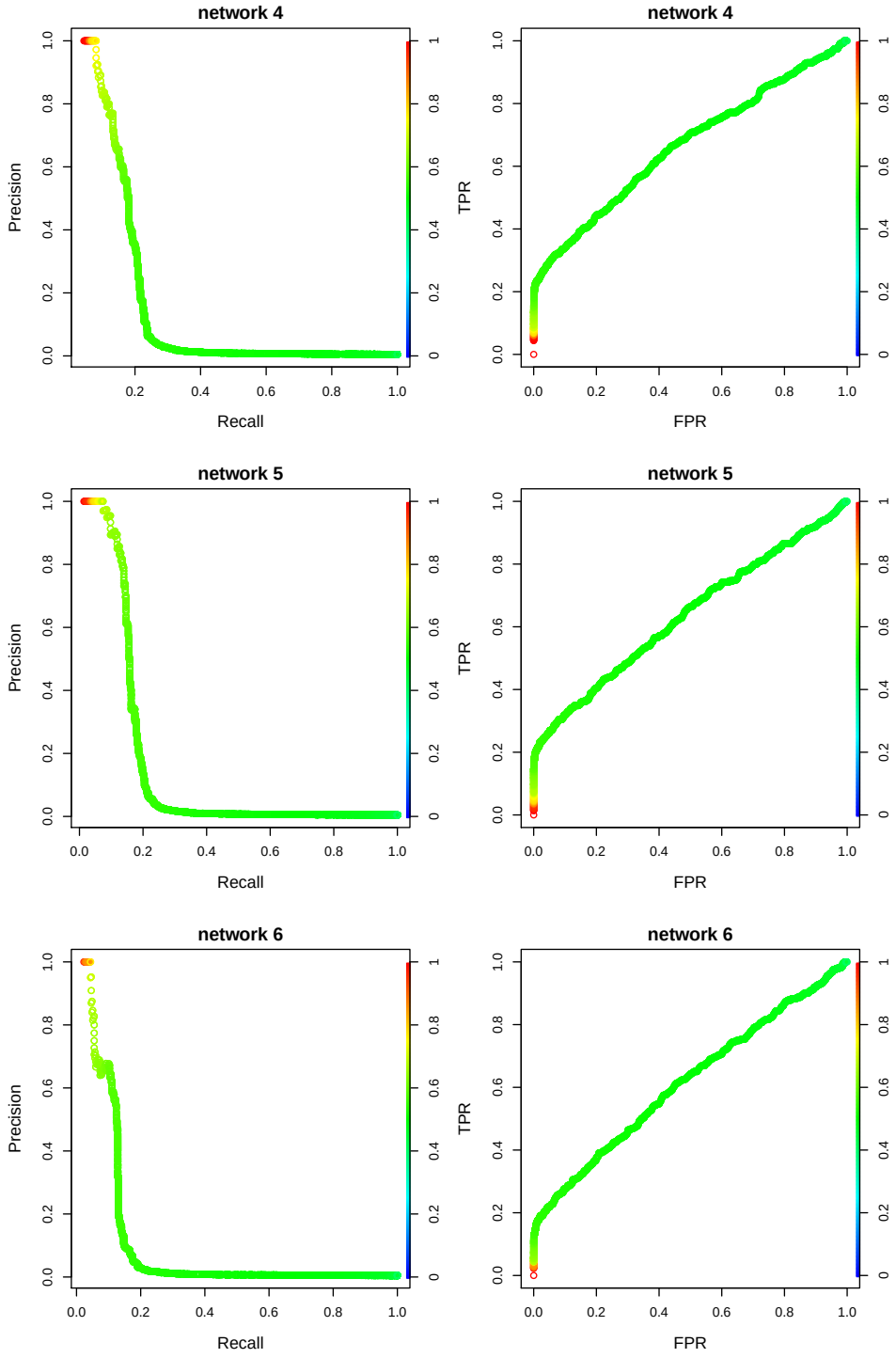

Figure 3.2: Simulated networks 4, 5, 6. Performance of the algorithm in the PR-plane and ROC curve.

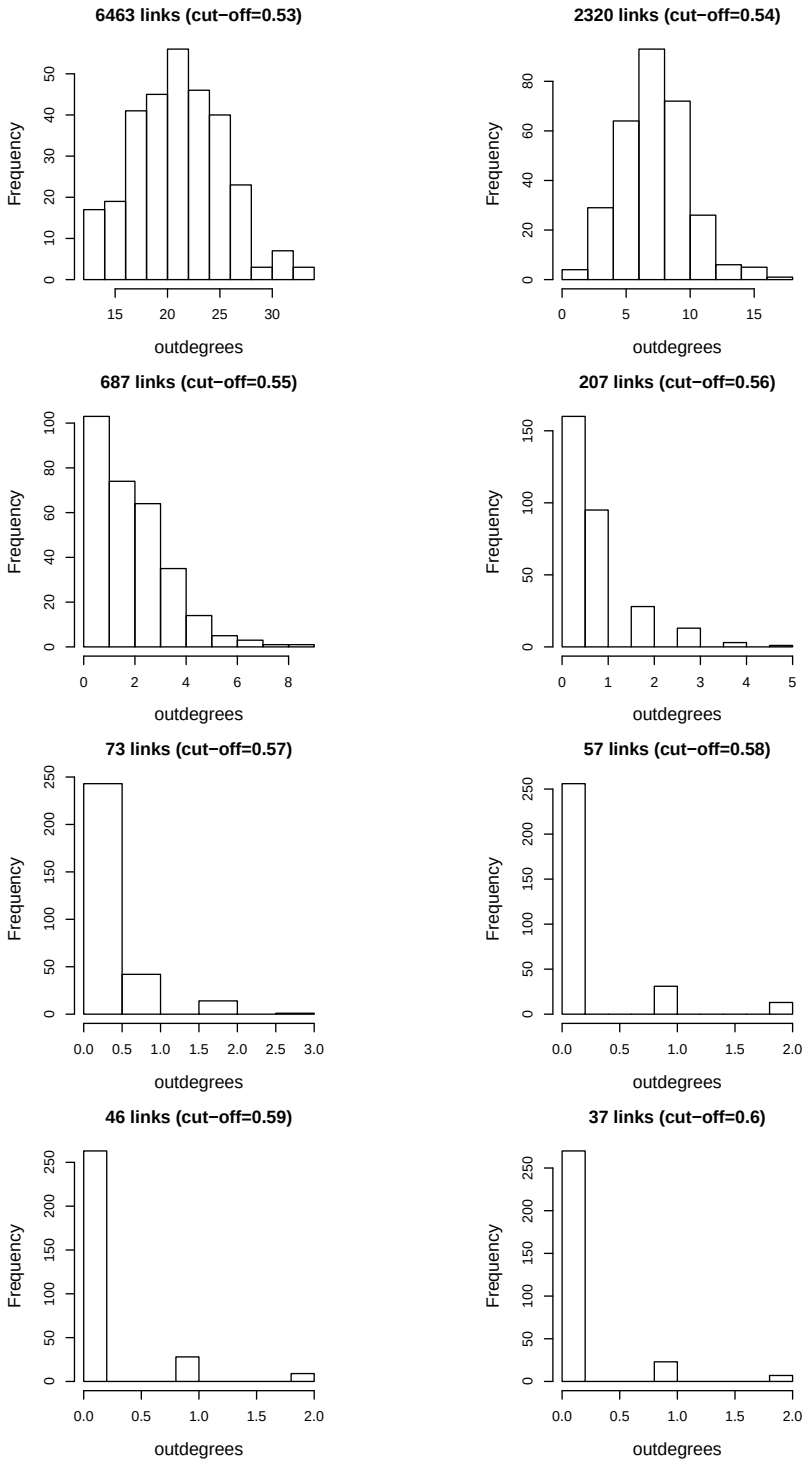

Figure 3.3: Network 1. Sensitivity to the cut-off on the posterior probability.

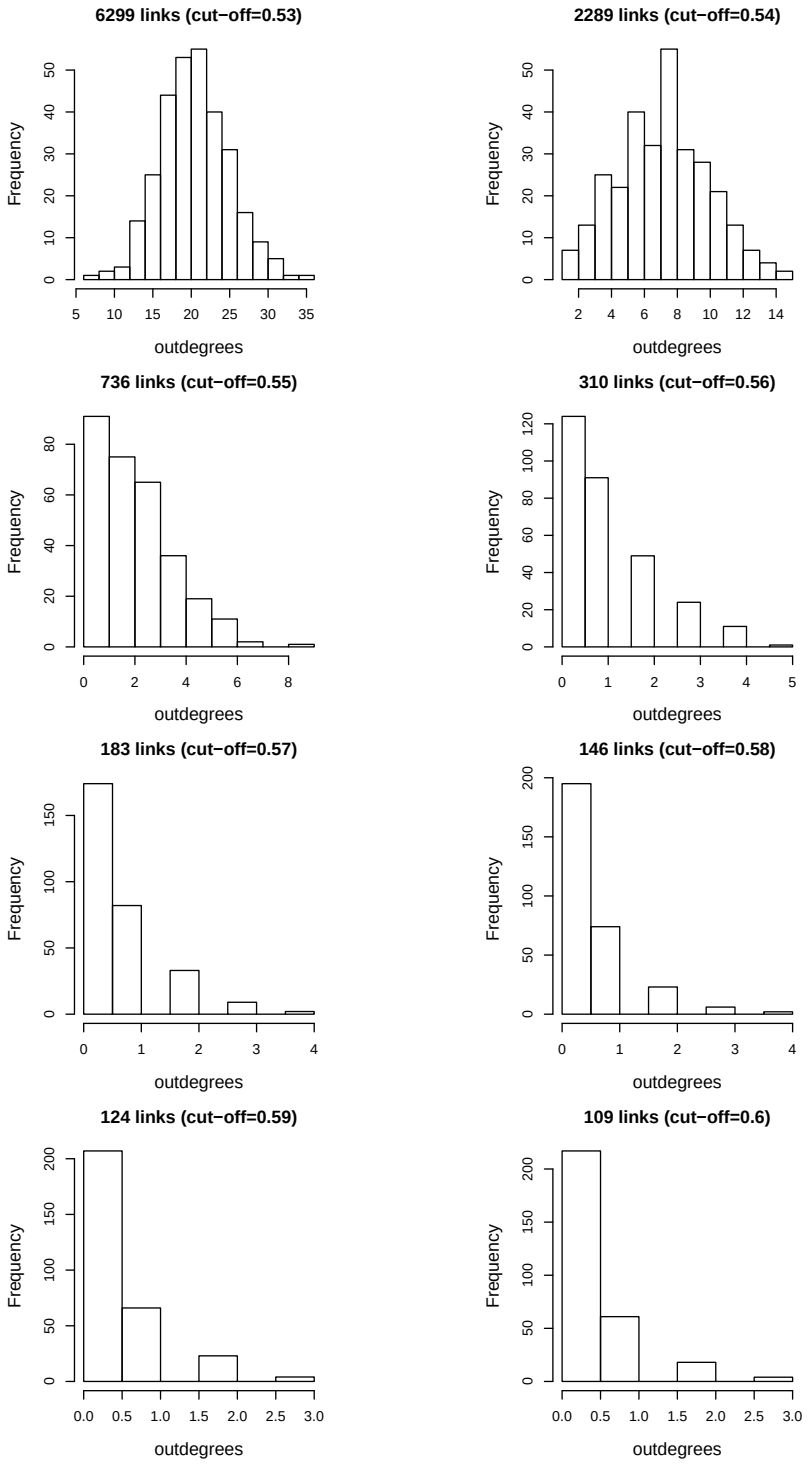

Figure 3.4: Network 4. Sensitivity to the cut-off on the posterior probability.

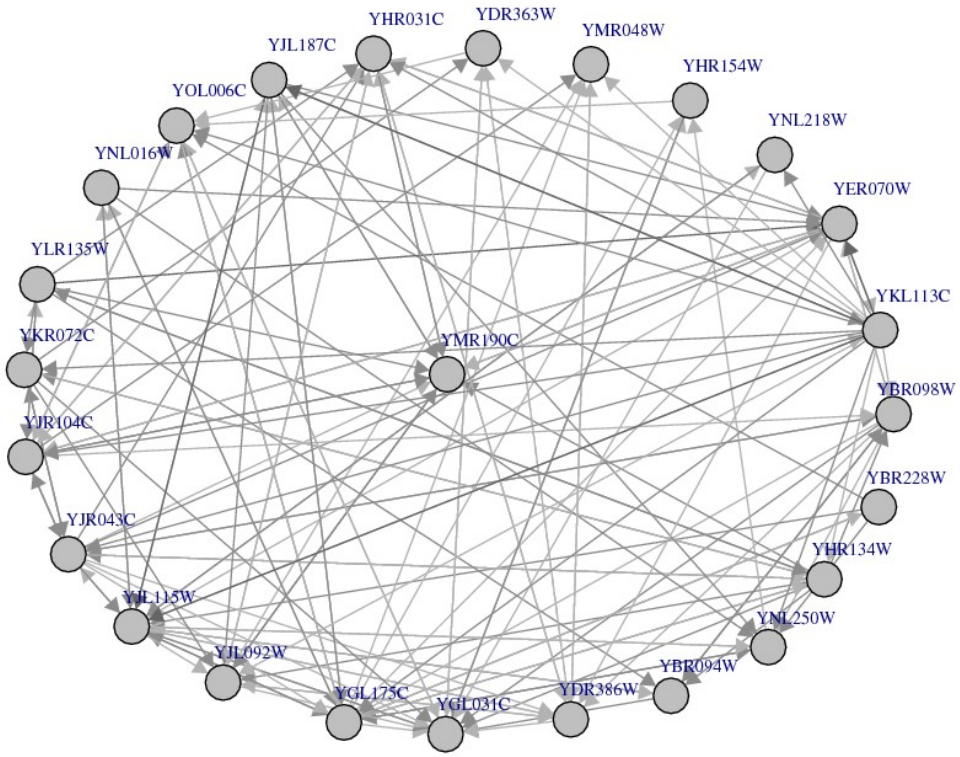

Figure 3.5: Reconstructed influence network, with YMR190C being the systematic name of SGS1.

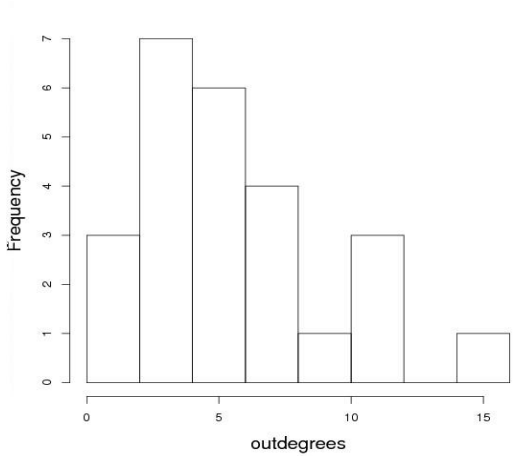

Figure 3.6: Histogram of outdegree for the reconstructed network.

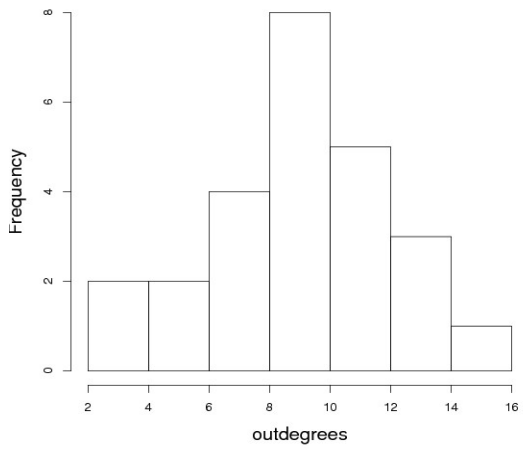

Figure 3.7: Histogram of outdegrees obtained via stepwise regression method.

Part II

Inference of regulatory modules from RNAi silencing of IFN-α transcriptional response modulators

Chapter 4

Protocol, design and realization

4.1 Biological motivation

IFN-α is a potent cytokine endowed with remarkable anti-angiogenic activity, firstly noticed many years ago Sidky and Borden [1987], and subsequently confirmed in different tumor models, see review Minuzzo et al. [2007]. Its anti-angiogenic activity has mainly been attributed to indirect effects, including the suppression of pro-angiogenic factor synthesis, such as inhibition of basic fibroblast growth factor (bFGF) overproduction by tumor cells, down-regulation of IL-8 and vascular endothelial growth factor (VEGF) gene expression. On the other hand, IFN-α also produces direct effects on endothelial cells (EC), including impairment in EC proliferation and migration. To date, the gene expression profile induced by this cytokine has been predominantly studied in tumor cell lines and some primary cells, including peripheral blood mononuclear cells (PBMC), T lymphocytes, and dendritic cells. Recent experiments have showed that treatment of EC with IFN-α induced a marked transcriptional up-regulation, and the set of genes regulated by this cytokine was characterized at a fixed time point (5 h) from stimulation, Indraccolo et al. [2007]. A dynamic study of the transcriptional profiles of IFN-α regulated genes and the identification of functional interactions between these transcripts is likely central to achieve a full understanding of the biology of interferons and their anti-tumor activity and may also have clinical relevance, given the broad use of this cytokine in patients.

At Istituto Oncologico Veneto-IRCSS, Padova, the group leaded by Dr Stefano Indraccolo has a long term experience on interferons, as documented by Indraccolo et al. [2007], Minuzzo et al. [2007], Moserle et al. [2008]. Aiming to a better characterization of the transcriptional response to IFN-α in endothelial cells, a new genomic experiment was designed with the final goal of reconstructing the functional network induced by this cytokine and identifying key modulators of IFN-α activity. The project was funded by the Azienda Ospedaliera di Padova (PRIL 2008-2010). The experiment consisted of a stimulation *in vitro* of human EC with human recombinant IFN-α. Gene expression profiles of a panel of selected IFN-α-regulated transcripts, measured by using quantitative PCR methods, were analyzed. The goal is to **identify possible IFN-α modulators in EC and infer some functional regulations of the underlying transcriptional network**. The Project is articulated into three main parts. The first part includes the design and the realization of the biological experiments. The second part is focused on the analysis of the transcriptional data at different levels: preprocessing and normalization, identification of possible modulators of IFN-α activity, application of methodologies of reverse-engineering. The third part consists in the experimental validation of the new findings.
In Figure 4.1 all the steps conducted to realize the study are sketched.

4.1.1 Aims

We designed the study in order to:

- gain a better understanding of IFN-α transcriptional response in EC;
- investigate possible modulators of IFN-α activity;
- infer causal relationship among IFN-α-regulated genes;
- study gene regulation in the IFN-α signaling pathway.

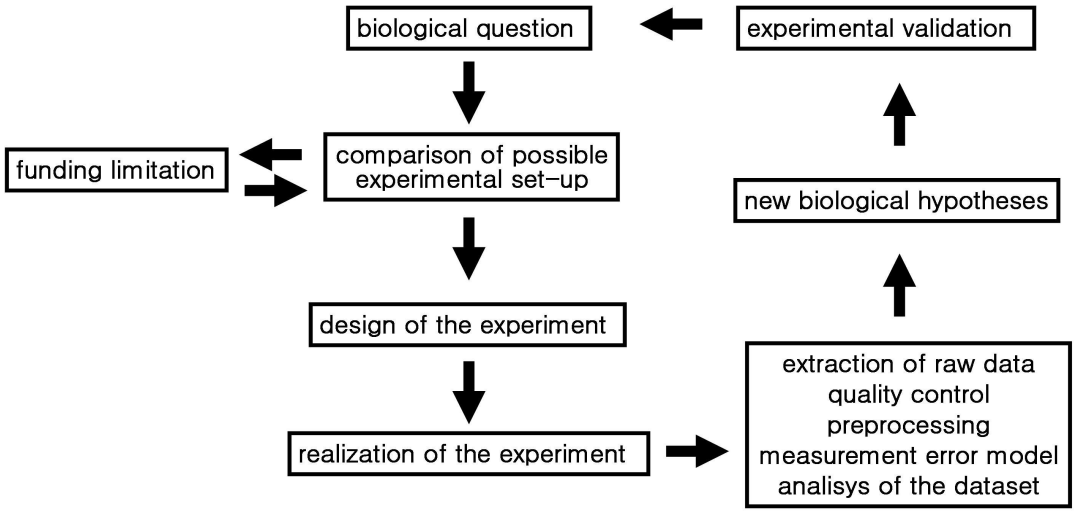

Figure 4.1: Several steps involved in the realization of the study.

4.2 Experimental set-up

Platform

Quantitative RT-PCR Cards[1], 96 gene format.

Cell lines and stimulation

Cell lines: primary human umbilical vein endothelial cells (HUVEC).

Pooling: each sample is obtained by pooling cells from 4 donors.

IFN-α dose: 1,000 IU/ml

Dynamic data:

IFN-α stimulation (7 time points)

IFN-α removal (3 time points)

Perturbation data:

Knock-down (KD) of 6 genes at 4 time points (0h, 2h, 8h, 12h)

Knock-down experiments include both the transfection with the siRNA silencing

[1]TaqMan Custom Array, 7900HT Fast Real-Time PCR System; Applied Biosystems

a specific gene and the corresponding control transfection.

Technical and biological replicates

2 technical replicates

2 biological replicates of each pool

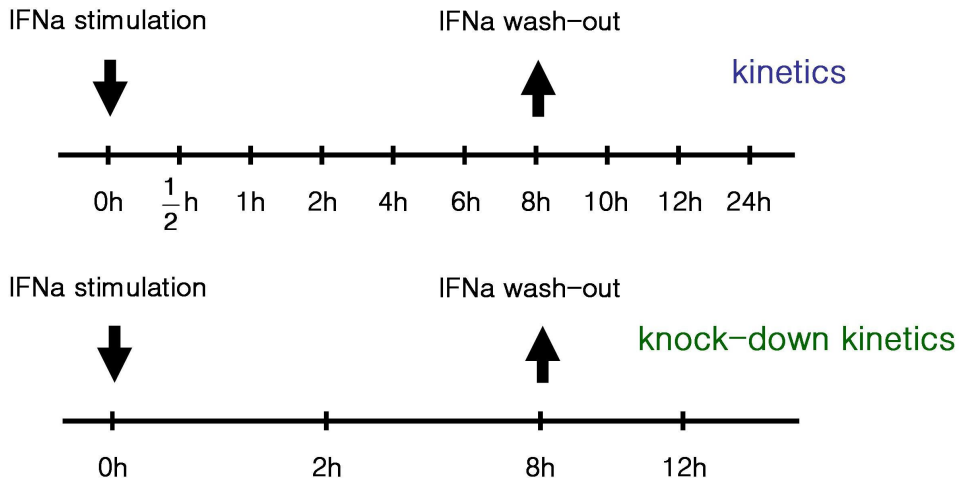

Figure 4.2: Experimental set-up: dynamics and perturbation data. Dynamic data consists of a time series of 10 time points with double stimulation: IFN-α stimulation at 0h and wash-out at 8h. Perturbations are obtained by knocking-down 6 candidate modulators of IFN-α. Perturbation data consists of a short time series of 4 time points with double stimulation as before.

4.2.1 Monitored transcripts

A panel of 96 genes, including the top of the transcriptional signature induced by IFN-α at 5h and the genes belonging to IFN-α signal transduction pathway, are screened using custom TaqMan Low-Density Arrays. Transcripts to be monitored were chosen based mainly on a recent genome-wide study of IFN-α effects on EC at a fixed time point (5 hours from stimulation) Indraccolo et al. [2007]. Preliminary analyses of this dataset compared with other datasets relative to

different cellular types stimulated with IFN-α, were carried out and served to define the probes to be spotted in the custom card, see Appendix B for a summary of these analyses. The 96 genes were selected as follows:

- **9 genes from IFN-α pathway**:
  IFNA1, IFNAR1, IFNAR2, IFNB1, IRF9, JAK1, TYK2, STAT1, STAT2.
- **75 genes up-regulated in EC** (at 5h from stimulation):
  top 77 in the FC ranking (including STAT1, STAT2), see Table B.1 in Appendix B.
- **9 genes of biological interest**:
  BNIP3, VEGFA, TP53, IRF3, BLZF1, EFL1, ATF3, IFI16, IFI27.
- 1 mandatory control: 18S.
- **2 genes as supplementary control (HK)**: LMNA, HMBS.

4.3 Realization of biological experiments

4.3.1 Materials

Cell lines

HUVEC are responsive to cytokine stimulation and are commonly used for physiological and pharmacological investigations. Endothelial cells play a pivotal role in a variety of physiological processes, such as angiogenesis and as the selective blood barrier; and pathophysiological processes, including arterial disease and cancer development.

TaqMan low density arrays (LDA)

Simultaneous real-time PCR reactions of 96 selected genes were performed by using TaqMan Low-Density Arrays in an ABI PRISM 7900 HT Sequence Detection System (Applied Biosystems). Each card contains two biological samples, each in technical duplicates (see Figure 4.3).

Stealth RNAi™

RNA interference (RNAi) is widely used for knockdown of gene expression in

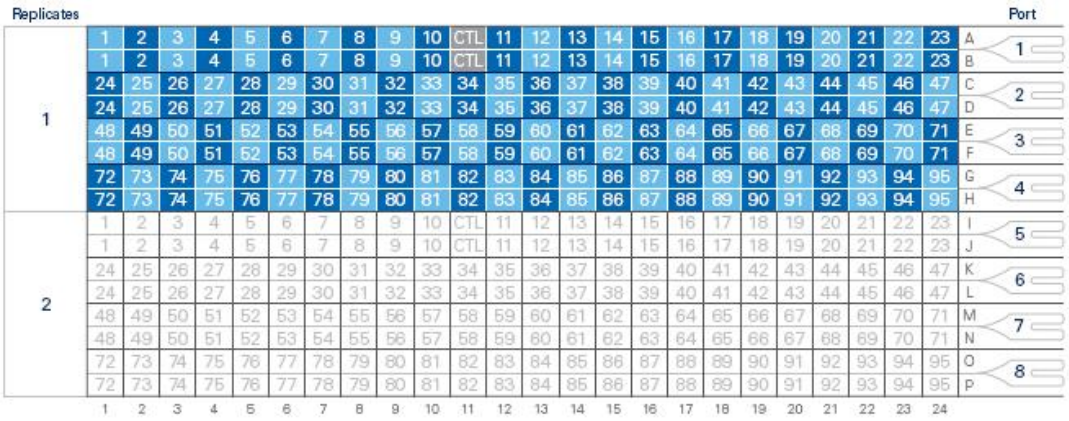

Figure 4.3: TaqMan card design, format 96. Two unique samples per card, each one with 95 unique assays and 1 mandatory control in technical duplicates.

eukaryotic cells. Indeed, transfection of small/short interfering RNA (siRNA) is an immediate, specific, powerful, easy to use tool for silencing specific sequences of mRNA. A siRNA works by turning off target genes by annealing to the mRNA strand and leading to its degradation.
Stealth® siRNAs designed by Invitrogen were used for RNAi knockdown of 6 selected genes: STAT1, IRF7, IRF1, IFIH1, OAS2, GBP1. These genes were picked up based on both unpublished data (GBP1) and consolidated biological information on their effects of modulators of IFN-α signalling Borden et al. [2007]. For each target gene a set of three siRNAs was tested for RNAi activity and the Stealth® siRNA showing the highest level of knockdown was selected to be used for RNAi experiments. Negative control with low content of GC was used as the calibrator siRNA.

4.3.2 Biological Methods

Cell culture, RNA extraction, reverse-transcription

HUVEC were provided by the Dipartimento di Scienze Biomediche e Biotecnologiche, Università degli Studi di Brescia and by the Dipartimento di Pediatria, Università degli Studi di Padova. Only passages from 2 to 6 were used. They were grown in M200 medium supplemented with 10% LSGS (low serum growth

supplement), and 1% antibiotic-antimycotic liquid (all reagents from GIBCO™, Invitrogen). Culture plates were coated with a 0.001% collagen solution (Sigma Aldrich). Cells were cultured at 37°C and 5% CO_2.
Total RNA was isolated from cells using TRIzol Reagent (Invitrogen) and was quantified with NanoDrop (NanoDrop Technologies). Reverse-transcription reaction of 2 μg of total RNA was performed using the High Capacity RNA-to-cDNA Reverse Transcription Kit (Applied Biosystems).
Real time quantitative PCR reactions were performed with TaqMan Universal PCR Master Mix.

siRNA Transfection
Reverse transfection of each stealth siRNA was performed using the transfection reagent Lipofectamine RNAi MAX (Invitrogen). RNAs was extracted 48h after the transfection as described in the above paragraph.

4.3.3 Temporal realization of the experiments

From January 2010, several pilots were conducted to choose the biological parameters for the full experiments: culture conditions, time points, siRNA. Following this preliminary phase, the experiments started and were performed at IOV. After more than one year of intense work we are still performing experiments. The main constraint on the possibility of speeding up the experiment schedule was due to the difficulty of obtaining adequate numbers of primary HUVEC samples and collecting enough cells from four donors at a passage lesser than 6.

4.4 Normalization and quantification

4.4.1 Real time RT-PCR data

Real-time reverse transcription polymerase chain reaction (RT-PCR) is the most widely used technique for mRNA quantification of a limited number of genes. Indeed this method is perfectly suited for validation of high-throughput microarray data, and for studies of a panel of selected candidate genes or pathway com-

ponents Vandesompele et al. [2009]. This depends on several factors: it is a homogeneous assay, which eliminates the requirement for post-PCR processing; with a wide dynamic range of linear quantification, high speed, reliability, high sensitivity (low template input required) and resolution (small differences can be measured).
Quantitative 'RT-PCR' has two main steps: cDNA synthesis by reverse transcription of mRNA (RT step) and subsequent quantification of cDNA by amplification to a detectable level (PCR step). 'Real-time' PCR is based on the sequential monitoring of product formation during the polymerase chain reaction. PCR amplification curves are sigmoidal (S shaped) and can be split into three phases: baseline (background signal or lag phase), log-linear (exponential phase), and plateau.

Amplicon (PCR product) concentration is measured by fluorescence detection, facilitated by the binding of fluorescent dyes or fluorescently labeled sequence-specific probes to the amplicon. Real-time RT-PCR quantifies gene expression levels by measuring the threshold cycle (CT), which represents the number of cycles needed to produce a certain defined threshold fluorescence. The threshold fluorescence may be determined by different computational strategies or set manually but must be identical for all samples to be compared within a run and must be placed in order to guarantee the PCR in the exponential phase of amplification. The CT is inversely related to the amount of target molecules in the reaction. Real-time PCR exploits the fact that, under ideal conditions, Yuan et al. [2006], the quantity of PCR products in exponential phase is proportional to the quantity of initial template. During the exponential phase PCR product will ideally double during each cycle if efficiency is perfect, i.e. 1. It is possible to obtain a PCR amplification efficiency close to 1 in the exponential phase if the PCR conditions, primer characteristics, template purity, and amplicon lengths are optimized.

4.4.2 Normalization by reference genes

The reference gene concept is currently the preferred way of normalizing real-time PCR data, Vandesompele et al. [2009]. Reference genes are internal controls

that are affected by the same sources of variation in the experiment workflow as the genes of interest. Choosing the correct reference genes to normalize gene expression in RT-PCR is essential. Typically, a stably expressed endogenous reference, i.e. an housekeeping gene, is used.

4.4.3 A mathematical model for relative quantification: the comparative CT method

The classic comparative CT method can be used to calculate the expression level of the gene of interest relative to a calibrator or reference sample using the CT data, Schmittgen and Livak [2008]. The equation describing the exponential amplification of PCR is:

$$X_n = X_0 \cdot (1 + E_X)^n \tag{4.1}$$

where:

$$\begin{aligned} X_n &= \text{ number of target molecules at cycle } n \\ X_0 &= \text{ initial number of target molecules} \\ E_X &= \text{ efficiency of target amplification} \\ n &= \text{ number of cycles} \end{aligned}$$

In the following we indicate with the subscripts X and R, the quantities related to the target and the reference gene, respectively. More precisely, the threshold cycle (CT) indicates the fractional cycle number at which the amount of amplified target reaches the fixed threshold (K_X or K_R). Thus, for the target gene

$$X_T = X_0 \cdot (1 + E_X)^{CT_X} = K_X,$$

and for the reference gene

$$R_T = R_0 \cdot (1 + E_R)^{CT_R} = K_R.$$

Dividing X_T by R_T we get

$$\frac{X_T}{R_T} = \frac{X_0 \cdot (1+E_X)^{CT_X}}{R_0 \cdot (1+E_R)^{CT_R}} = \frac{K_X}{K_R} = K.$$

Assuming that the efficiencies of the target and the reference are equal: $E_X = E_R = E$,

$$\frac{X_0}{R_0} \cdot (1+E)^{(CT_X - CT_R)} = K.$$

Defining

$X_N = \frac{X_0}{R_0}$, normalized amount of target

$\Delta CT = CT_X - CT_R$, difference in threshold cycles between target and reference

$$X_N = K \cdot (1+E)^{-\Delta CT}.$$

Dividing X_N of treatment T (silencing in our case) by the X_N of the calibrator CB (corresponding control silencing),

$$\frac{X_N^T}{X_N^{CB}} = \frac{K \cdot (1+E)^{-\Delta CT^T}}{K \cdot (1+E)^{-\Delta CT^{CB}}} = (1+E)^{-\Delta\Delta CT}$$

where $\Delta\Delta CT = \Delta CT^T - \Delta CT^{CB}$.

For amplicons designed and optimized according to Applied Biosystems guidelines (amplicon size < 150bp), the efficiency E is close to 1. We can thus express the relative expression level R in the target sample relative to the calibrator sample as

$$R = \frac{X_N^T}{X_N^{CB}} \approx 2^{-\Delta\Delta CT}. \tag{4.2}$$

Chapter 5

Experimental data

5.1 Data

5.1.1 Data extraction

CT values were extracted using the Sequence Detection System 2.4 software (Applied Biosystems) and the following analyses were done in R. Normalization was performed using the classic comparative CT method described in section 4.4.3.

5.1.2 Housekeeping choice

The choice of the housekeeping was based on the assessment of variation of CT values. The lesser the variance, the more stably the gene is expressed across the experimental samples. In Figure 5.1, boxplots with the CT distribution of candidate housekeeping genes (LMNA, HMBS, and JAK1) are shown, the silencing of OAS2 was taken as an example.

LMNA turns out to be the best housekeeping in IFN-α dynamics, while JAK1 was the most stable gene in all the perturbation experiments. Normalization was performed with respect to the best housekeeping in the two kinds of experiments.

5.1.3 Dynamic data

Availability of the full activation and deactivation kinetics of the monitored genes allows a more complete description of the transcriptional effects of IFN-α , the

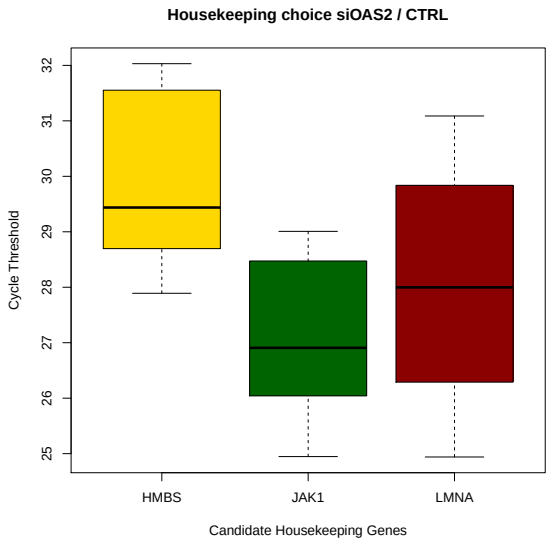

Figure 5.1: Variation of candidate HKs in OAS2 silencing experiment and corresponding control silencing. JAK1 was the gene with lower variance.

identification of very early induced genes and, among them, of possible IFN-α modulators. As an example of the full kinetics, we report here the gene expression profiles of the two biological replicates for gene OAS2, see Figure 5.2.

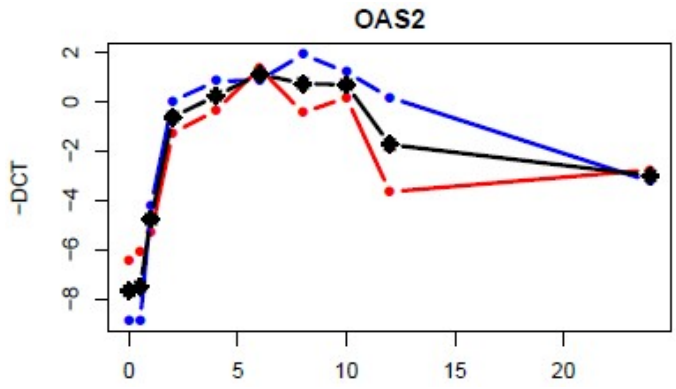

Figure 5.2: OAS2 kinetics of activation/deactivation. In red and blue the two biological replicates; in black the mean kinetics. Time from IFN-α stimulation are on the x-axis (hours), $-\Delta CT$ on the y-axis. Wash-out was performed at 8h.

5.1.4 Selection of perturbation data sampling times

Dynamic data were collected first and the first biological replicate of the full IFN-α kinetics was used to properly define the four sampling times for the short kinetics of perturbation data. The number of time points was set to 4, basically due to funding limitation. We decided to keep the double stimulation (IFN-α at 0h and wash-out at 8h), thinking that the final dataset would be enriched by this choice. The remainig two time points were to be chosen, the first among 1/2h, 1h, 2h, 4h, 6h and the second among 10h, 12h, 24h. Two criteria were used for this choice: number of undetermined genes and percentage of variation with respect to the stimulation with IFN-α or its removal. The two time points in the activation and in the deactivation phases showing the best balance between a low number of undetermined genes and a high percentage of variation with respect to the corresponding stimulation (IFN-α and wash-out) were 2h and 12h.

5.1.5 Perturbation data

As an example of the knock-down kinetics, we report here the plot of $-\Delta CT$ for the siRNA silencing OAS2 and the corresponding control siRNA, in biological duplicates. The kinetics of SAMD9 was taken as an example, see Figure 5.3.

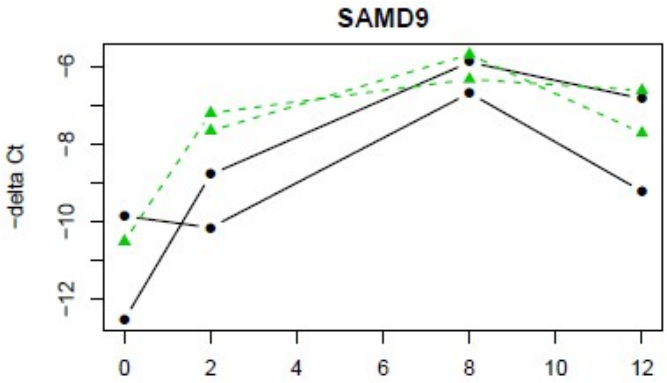

Figure 5.3: SAMD9 kinetics after silencing of OAS2 and stimulation with IFN-α. Black circles correspond the control siRNA; green triangles to the siRNA targeting OAS2. On the x-axis time (hours), on the y-axis $-\Delta CT$. The regulation at 2h is found significant by our selection procedure.

5.2 Measurement error model for $\Delta\Delta CT$

In our analysis, we estimate the biological variance of $\Delta\Delta CT$ directly starting from the observed $\Delta\Delta CT$ values. Different procedures, based on error propagation, were also considered and are reported in Appendix C for completeness. To estimate the biological variance of $\Delta\Delta CT$, a flexible model for error variance was used:

$$\sigma^2_{\Delta\Delta CT}(t_i) = \alpha + \beta(\Delta\Delta CT_{OBS}(t_i))^\gamma,$$

where α, β, and γ are parameters linking the variance to the intensity of the observations, $\Delta\Delta CT_{OBS}$. The range of $\Delta\Delta CT_{OBS}$ intensities was subdivided in intervals corresponding to subsets with the same number of samples and for each subset mean $\pm$ SD of variance estimates were calculated and the mean of the variance estimates was plotted against the mean intensity. The general error model described above was fitted on these data. Parameters and their precision (expressed in terms of coefficient of variation, CV) were estimated in Matlab using the `lsqnonlin` function to solve the weighted least square problem. For the fitting, the complete perturbation dataset was used with the only exception of IRF7 silencing data. This information was discarded because the two biological replicates of siIRF7 experiment were not comparable and a new biological replicate is currently planned to be performed, see 7.2.6 for further details. The best fitting, in terms of Weighted Residual Sum of Squares (WRSS) and precision of the parameter estimate, was obtained by the model with constant variance, $\hat{\sigma}^2 = 0.395$, see Figure 5.4.

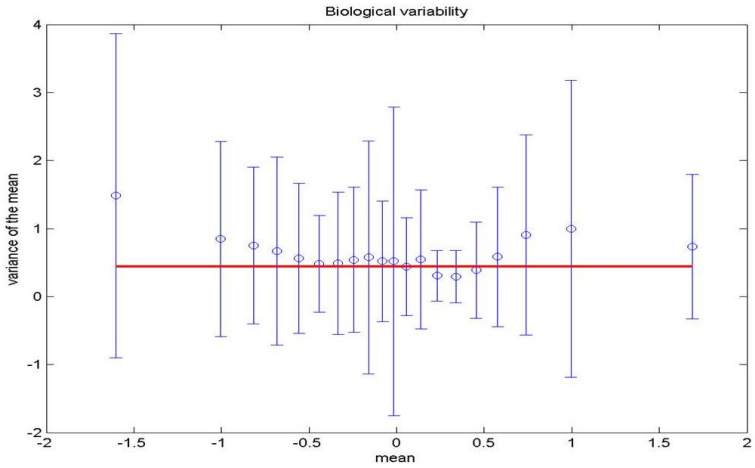

Figure 5.4: DDCT measurement error model.

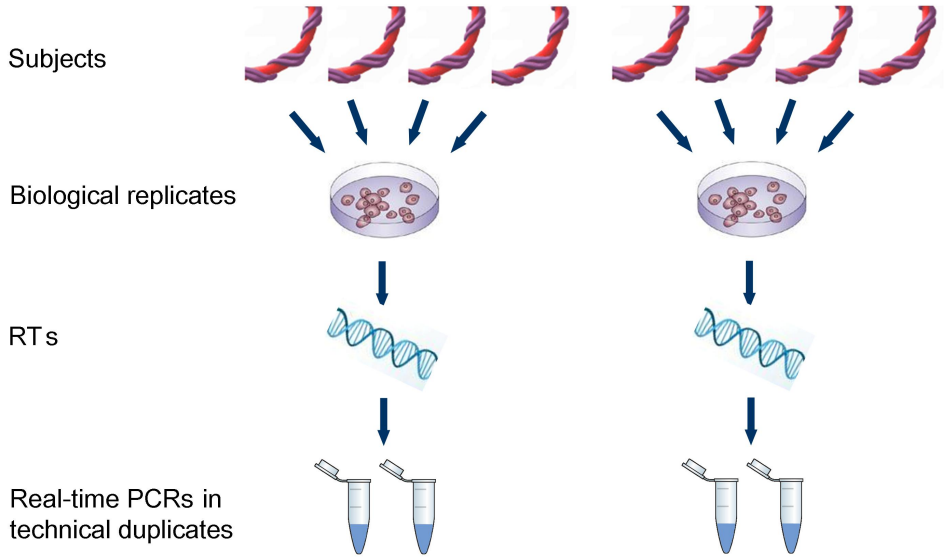

Figure 5.5: Biological and technical variability in the experiment design for each time point. Cells are extracted from 8 different donors. Two biological samples are obtained, each one by pooling cells isolated from 4 donors. For each RT, the real time PCR assay is performed in technical duplicates, run on the same card.

Chapter 6

Data analyses

6.1 Dynamic data analysis: K-means clustering

In order to identify the main temporal expression patterns characterizing IFN-α transcriptional response, $-\Delta CT$ temporal profiles of the selected genes were clustered using the K-means algorithm. The number of clusters K was set to 5 and the measure of similarity adopted was the Pearson correlation.

6.2 Silencing data analysis: selection procedure

Silencing data analysis is aimed at extracting information from the knockdown data about significant modulations that may represent direct or indirect effects due to inactivation of target genes. To obtain a **significance analysis of IFN-α induced transcriptional modulations** we used two different ingredients: the measurement error model of $\Delta\Delta CT$ biological variability, discussed in Subsection 5.2, and a selection procedure that uses the variance estimated from the error model to assign a p-value to each modulation. For each silencing experiment, the selection of significant modulations was performed, starting from the mean $\Delta\Delta CT$ values (across the two biological replicates).
We proposed a two-stage approach that first filters observations by a variance-based criterion and then performs a variable-by-variable statistical test procedure, only on the observations that pass the filter. A final multiple test correction is ap-

plied to control the number of modulations erroneously identified as differentially expressed (false positives).

6.2.1 Filtering based on $\Delta\Delta CT$ variance

In our experimental design, only two biological replicates for each silencing were performed. This prevented us from the application of statistical test procedures that account for variability. So, biological variance was not used to select significant modulations but only to filter out unreliable observations. We considered the distribution of $\Delta\Delta CT$ variance, based on all the perturbation experiments, and defined the filtering cut-off (threshold) as the 95-th percentile.
A second filter was applied to exclude from the analysis non-informative genes that were inserted in the study as candidate housekeeping genes, JAK1 (best HK, used for normalization), LMNA, HMBS and 18S (mandatory control in the card).

6.2.2 Null hypothesis distribution

The null hypothesis H_0, namely no difference in the effects of a siRNA targeting a specific gene and of the calibrator siRNA, was tested. Recalling that $\Delta\Delta CT(t_i) = \Delta CT_T(t_i) - \Delta CT_{CB}(t_i)$, where $i = 1, \ldots, 4$; a gene at time point t_i is considered not differentially expressed between the two conditions if the $\Delta\Delta CT(t_i)$ is close to 0.
H_0 was generated by assuming that the $\Delta\Delta CT$ averaged on the two biological replicates is normally distributed, according to a $N(0, \frac{\hat{\sigma}^2}{2})$, where $\hat{\sigma}^2$ is the biological variance of $\Delta\Delta CT$, estimated with the measurement error model in Figure 5.4. A variable-by-variable test was performed on each observation (gene at a given time point) and the corresponding p-value was calculated.

6.2.3 Multiple testing correction

A Bonferroni multiple test correction to control the false positive rate (FPR) in the gene callings was applied. The corrected significance level for each test was $\tilde{\alpha} = \frac{\alpha \cdot K}{92}$, where 92 is the number of genes tested (excluding non-informative genes) and K represents the number of time points on which the gene is tested.

6.3 Inference of regulatory modules

Starting from the results of the selection procedure we developed a novel method to reconstruct regulatory subnetworks, composed of FFLs, the basic building blocks of many biological networks. Among the regulations selected by the significance analysis of IFN-α induced transcriptional modulations, we pick up those involving silenced genes. For each regulation between two perturbed genes, we check for possible significant modulated genes shared by the two and construct the corresponding FFLs. Merging the FFLs together we obtain a completely connected regulatory module (clique, in graph theory) of significant modulations in which the two silenced genes and their common regulated genes are present.
The rationale behind this approach is that in biological networks FFLs are overrepresented, therefore to distill information worth of biological validation we elicit among the significant modulations those most trusted, namely those that belong to FFLs.

Chapter 7

Results

7.1 Dynamic data analysis

Dynamic data were used to

- fix the time points at which perturbation data were collected (see Subsection 5.1.4);
- gain a better understanding of the kinetics of activation and deactivation of IFN-α regulated transcripts.
- cluster genes sharing the same pattern of IFN-α modulation.

7.1.1 K-means clustering

Five main temporal patterns of modulations were identified in response to IFN-α stimulation (0h) and subsequent wash-out (8h). In 7.1 the resulting clusters for the $-\Delta CT$ profiles are plotted. Profiles of the single genes are represented in gray, while colored profiles are the means of each cluster. Clusters 1, 2, and 3 show different patterns of up-regulation; in these clusters most genes belong to the top signature of IFN-α at 5h. Cluster 4 shows a fluctuating profile characterized by several peaks. Interestingly IFNAR1, IFNAR2, and TYK2 belong to this cluster. Cluster 5 shows a profile of activation/deactivation characterized by a plateau phase from 2 to 10 hours. STAT1 and JAK1 belong to it.

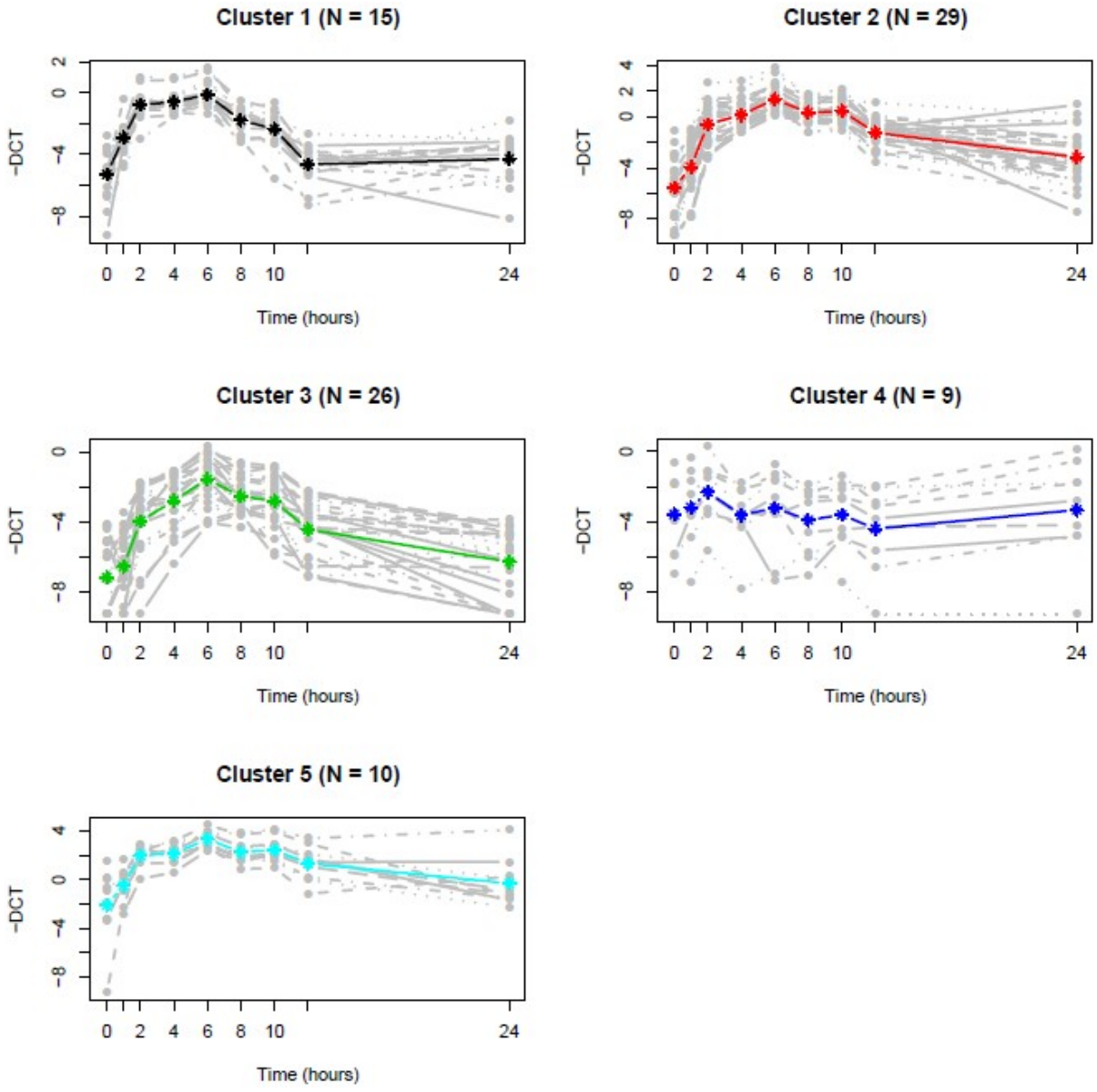

Figure 7.1: K-means clustering. Patterns of up-regulation.

Detailed list of the genes in each cluster.

Cluster 1

APOL2; BLZF1; IFI16; IFIT2; IFIT3; IFIT5; IRF1; IRF9; SAMD9; SLC25A28; TDRD7; TMEM140; TNFSF10; TRIM21; ZC3HAV1.

Cluster 2
APOL1; APOL6; BST2; C19orf66; DDX58; DDX60; GMPR; HERC6; IFI35; IFI44L; IFI6; IFIH1; IFIT1; IFITM1; LGALS9; MX2; OAS1; OAS2; OAS3; PLSCR1; PML; PSMB9; RSAD2; RTP4; SAMHD1; SLC15A3; SP110; STAT2; USP18.

Cluster 3
APOBEC3G; APOL3; C1orf38; CASP1; CFB,C2; CXCL10; DHX58; DSP; FAM46A; GCH1; HERC5; IDO1; IFI30; IFI44; IL15RA; IRF7; OASL; PLEKHA4; PSMB8; RARRES3; SECTM1; TAP2; TLR3; TRANK1; TRIM14; XAF1.

Cluster 4
ATF3; BNIP3; CX3CL1; IFNAR1; IFNAR2; IRF3; TP53; TYK2; VEGFA.

Cluster 5
CXCL11; ELF1; GBP1; IFI27; ISG20; JAK1; MX1; STAT1; TAP1; UBE2L6.

7.2 Silencing data analysis: significance analysis of IFN-α-induced regulations

According to the selection procedure detailed in section 6.2, significant regulations, due to the silencing of each candidate modulator of IFN-α activity, were selected from perturbation data by using the measurement error model of Figure 5.4.

7.2.1 STAT1 knock-down

The silencing of STAT1 produced the most evident effect among the perturbations performed (18 down-regulated genes and 5 up-regulated genes). This result was expected as STAT1 is a transcription factor central to the IFN-α signaling pathway. In Figure 7.2 the genes down-regulated by the silencing of STAT1 are reported. All significant down-regulations occur in the stimulation phase of the

kinetics (2h and 8h), only CXCL11 expression is affected also in the IFN-α removal phase (12h). Notably, among the down-regulated genes, there is also IFIH1,

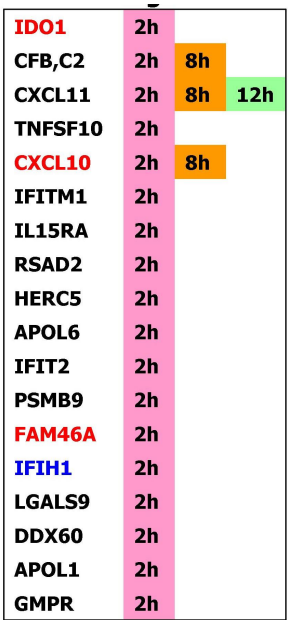

Figure 7.2: Genes significantly down-regulated by STAT1 silencing.

another candidate modulator, target of a specific siRNA experiment. Figure 7.3 shows the genes up-regulated by STAT1 silencing; at 8h from stimulation we find four genes significantly modulated, including the subunit 1 of IFN-α receptor (IFNAR1) and the ligand itself (IFNA1).

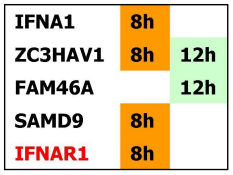

Figure 7.3: Genes significantly up-regulated by STAT1 silencing.

7.2.2 IFIH1 knock-down

Silencing of IFIH1 caused mainly down-regulation (10 down-regulated genes vs. 3 up-regulated), partially overlapping with STAT1 silencing. In Figure 7.4 and in Figure 7.5 the genes, respectively, down and up-regulated by the silencing of IFIH1 are reported.

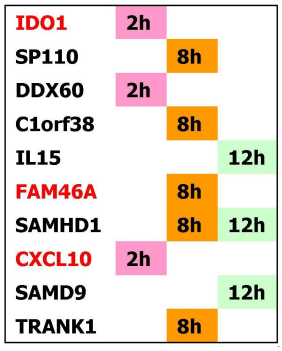

Figure 7.4: Genes significantly down-regulated by IFIH1 silencing.

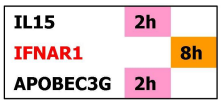

Figure 7.5: Genes significantly up-regulated by IFIH1 silencing.

7.2.3 OAS2 knock-down

The silencing of OAS2 induced down-regulation of 7 genes and up-regulation of 6 genes, listed in Figures 7.6 and 7.7. Almost all the significant modulations occur in the stimulation phase of the kinetics. Interestingly, among the up-regulated genes there are IFNA1 itself, the receptor TLR3 and IL15.

7.2.4 GBP1 knock-down

GBP1 seems to be a negative modulator of IFN-α transcriptional response. Indeed, the silencing of GBP1 produced mainly up-regulation effects (18 up-regulated

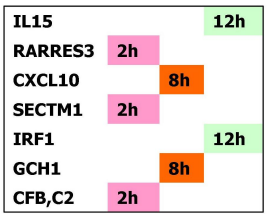

Figure 7.6: Genes significantly down-regulated by OAS2 silencing.

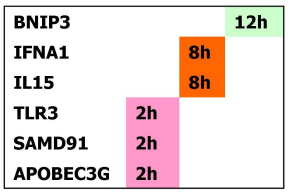

Figure 7.7: Genes significantly up-regulated by OAS2 silencing.

genes vs. 3 down-regulated genes). Several modulations occur in the IFN-α removal phase of the kinetics. The significant up-regulations induced by GBP1 silencing in the stimulation phase of the kinetics include some members of the IFN-induced protein family ('IFI' prefix), IFNB1, and IFNAR2. Full lists of significant modulations are presented in Figures 7.8 and 7.9.

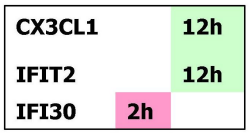

Figure 7.8: Genes significantly down-regulated by GBP1 silencing.

7.2.5 IRF1 knock-down

The silencing of IRF1 induced up-regulation of 10 genes and down-regulation of 2 genes. Significant up-regulations occurs both in the IFN-α stimulation and in the removal phase; among the modulations there is also IFNB1 at 2h (but a critical evaluation of the two biological replicates separately shows that the significance of

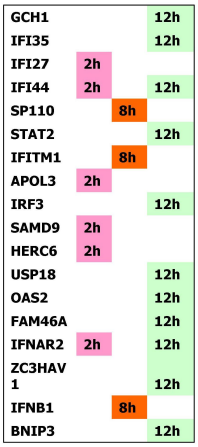

Figure 7.9: Genes significantly up-regulated by GBP1 silencing.

this modulation is due only to one biological replicate). Significant modulations are presented in Figures 7.10 and 7.11.

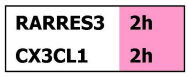

Figure 7.10: Genes significantly down-regulated by IRF1 silencing.

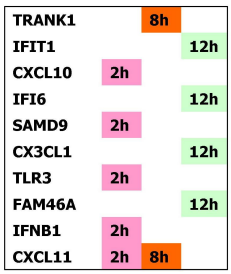

Figure 7.11: Genes significantly up-regulated by IRF1 silencing.

7.2.6 IRF7 knock-down

One of the two biological replicates relative to IRF7 knock-down showed some experimental problems: the silencing induced by the RNAi procedure was effec-

tive only at 0h, and it was not maintained in the following time points (2h, 8h, 12h). That replicate was discarded and we are presently performing the wet lab experiments for a new one. In order to have a preliminary indication about the effects of IRF7 knock-down, our selection procedure was adapted to the analysis of a single biological replicate. Significant modulations due to the silencing of IRF7, based only on the good biological replicate, are showed in Figures 7.12 and 7.13.

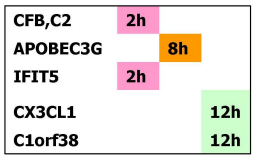

Figure 7.12: Genes significantly down-regulated by IRF7 silencing (based on a single biological replicate).

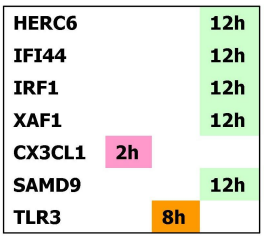

Figure 7.13: Genes significantly up-regulated by IRF7 silencing (based on a single biological replicate).

7.3 Inference of regulatory modules

7.3.1 IFN-α -sentinel genes

Although in the custom cards about ninety of IFN-α-regulated-genes were present, only a small number of them was affected by silencing of selected candidate modulators. The genes significantly modulated by at least two different RNAi perturbations are listed in Table 7.1. These 23 transcripts, particularly sensitive

Table 7.1: Genes significantly modulated by at least two different siRNAs. The capital letters U, D and D&U indicate, respectively, up-regulation, down-regulation or both the regulations.

Transcripts	siSTAT1	siIFIH1	siOAS2	siGBP1	siIRF1
APOBEC3G		U	U		
BNIP3			U	U	
CFB,C2	D		D		
CX3CL1				D	D&U
CXCL10	D	D	D		U
CXCL11	D				U
DDX60	D	D			
FAM46A	D&U	D		U	U
GCH1			D	U	
IDO1	D	D			
IFI35				U	U
IFIT2	D			D	
IFITM1	D			U	
IFNA1	U		U		
IFNAR1	U	U			
IFNB1				U	U
IL15		D&U	D&U		
RARRES3			D		D
SAMD9	U	D	U	U	U
SP110		D		U	
TLR3			U		U
TRANK1		D			U
ZC3HAV1	U			U	

to changes in the IFN-α pathway, define a sort of IFN-α mini-signature, that could be used to test the integrity of the pathway. They could represent either terminal genes in the IFN-α pathway or, in turn, modulators of other IFN-α-stimulated genes. In this list there is also an unexpected gene not directly related to IFN-α, BNIP3, selected by our biologists because it is a marker of hypoxia and autophagy. Interestingly, three genes of the mini-signature seem to have the role of IFN-α-sentinels: SAMD9 is modulated by all the selected perturbations, while CXCL10 and FAM46 are modulated by four out of five perturbations.

7.3.2 STAT1-IFIH1 regulatory modules

As stated in Paragraph 7.2.1, among the genes significantly down-regulated by STAT1 there is also IFIH1. Comparing the significant regulations obtained from STAT1 and IFIH1 knock-downs in the stimulation phase, we can hypothesize five feed-forward loops in which STAT1 activates IFIH1 and both STAT1 and IFIH1 coherently modulate a third gene. Four genes are activated by STAT1 and IFIH1: IDO1, DDX60, CXCL10, and FAM46A, see Figure 7.14.

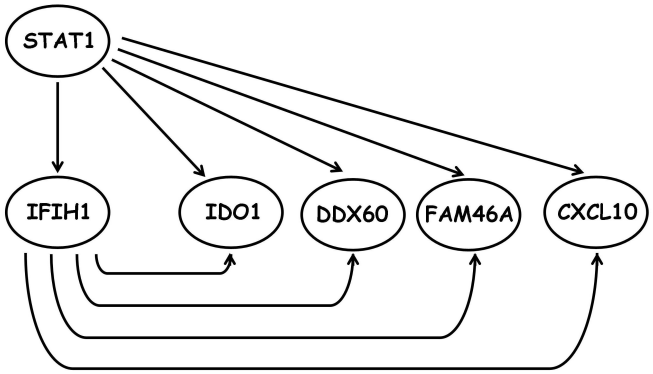

Figure 7.14: Hypothesis to be validated: 4 possible coherent feed-forward loop involving STAT1, IFIH1, IDO1, DDX60, CXCL10, and FAM46A.

Probably the most interesting finding concerns repression of IFNAR1 from both STAT1 and IFIH1. This results in a new feed-forward loop represented in Figure 7.15. Further analysis of the single biological replicates, has shown that regulation of IFNAR1 by STAT1 and IFIH1 is significant, independently in each

biological replicate. This information seems very promising for the future functional validation that will support the results of our study.
Regulations in the post-IFN-α removal phase require more caution in the interpretation and will be treated in Paragraph 7.3.3.

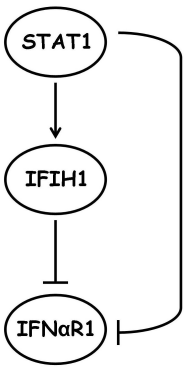

Figure 7.15: Hypothesis to be validated: possible coherent feed-forward loop involving STAT1, IFIH1, IFNAR1.

7.3.3 Interpretation of regulations in the post-IFN-α removal phase

The double stimulation (IFN-α at 0h and wash-out at 8h) has proved useful to enrich the observation dataset. The removal of exogenous IFN-α allowed to see many modulations at 12h that otherwise would have been missed, but forcing the interpretation of regulations in this phase of the kinetics may be misleading. We thus decide to consider these regulations as influence effects, without specifying whether they represent repressions or activations. Reconsidering the results presented in subsection 7.3.2, including also the regulations at 12h, we obtain the subnetwork presented in Figure 7.16. Following this approach we can hypothesize two other transcriptional subnetworks, sketched in Figures 7.17 and 7.18. Notice that all the inferred regulations are consistent also in terms of directionality, when present. The only exception is the regulation of CXCL10 from OAS2 and IRF1: in this case OAS2 activates while IRF1 represses CXCL10. All the observed transcriptional regulations are not necessarily direct regulations but may

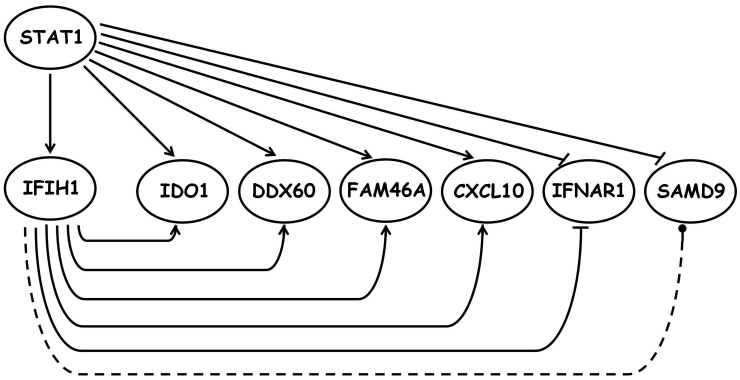

Figure 7.16: Subnetwork of regulations between STAT1, IFIH1 and their common targets. Solid lines show regulations in the stimulations phase; dashed lines show regulations occurred in the post-IFN-α removal phase. Arrow styles represent: arrow, activation; $\dashv$, repression; circle, unknown.

be mediated by other genes, so this is not a contradiction.

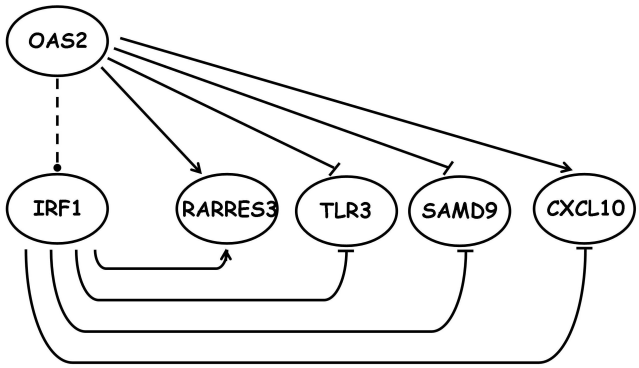

Figure 7.17: Subnetwork of regulations between OAS2, IRF1 and their common targets. Solid lines show regulations in the stimulations phase; dashed lines show regulations occurred in the post-IFN-α removal phase. Arrow styles represent: arrow, activation; $\dashv$, repression; circle, unknown.

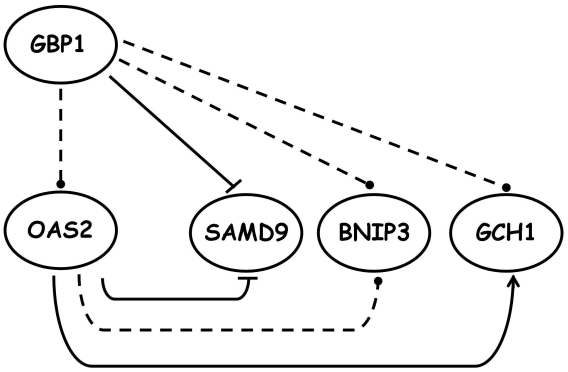

Figure 7.18: Subnetwork of regulations between GBP1, OAS2 and their common targets. Solid lines show regulations in the stimulations phase; dashed lines show regulations occurred in the post-IFN-α removal phase. Arrow styles represent: arrow, activation; $\dashv$, repression; circle, unknown.

Conclusions

The main object of this thesis is the inference of gene regulation from quantitative gene expression data. This goal is of central importance for health care as complex genetic diseases such as cancer are caused by deregulation or aberrant regulation of genes. The thesis is structured into two main parts corresponding to a theoretical and a practical approach to the inference of gene regulation.

In the first part of the thesis we presented a Bayesian hierarchical model for the reconstruction of regulatory networks from gene expression data. We assume that the transcriptional interactions can be described via a simple linear model on the logarithmic scale. Due to the typically high dimensionality of the data, a scale-free topological constraint on the outdegree distribution has been imposed. We develop an MCMC algorithm for inferring the structure of the network. A key element of our procedure resides in the introduction of scale-free topological constraint on the network structure directly within the MCMC update of the connectivity matrix. This new update involves the introduction of a χ^2-statistic and it has a direct connection with Approximate Bayesian Computation methods in which, in the computation of untractable or time expensive likelihoods, the data are replaced with a lower order summary statistic. We tested the performance of the algorithm on *in silico* data and on real data from a public dataset reconstructing a small regulatory network in yeast. Results show that our method is effective in introducing the scale-free topological constraint but does not improve the state of the art in reconstructing the network. Actually the present method, despite the simplicity of the assumed gene interaction model, is very flexible and it allows the easy introduction of a priori information. Future work will be directed to exploit this possibility, including information on known hubs, biological knowledge, or information from independent analyses. Independent work will be

done, in collaboration with Prof Wit, to obtain a theoretical characterization of the novel MCMC update proposed for X in order to generalize and transfer it to different contexts.

In the second part of the thesis we described the design, the realization and the analysis of a new genomic experiment, in collaboration with Dr Indraccolo, to characterize the biological activity of IFN-α, an extremely important cytokine also used in cancer patients. The specific goal of the project is the identification of regulations from the IFN-α transcriptional network and the discovery of new modulators of this cytokine in human endothelial cells. This is a very innovative and ambitious goal as previously published studies have so far been limited to the description of the IFN-α response in different cell types in a 'static' manner, and only very recently a dynamic study on IFN-α effects was published Pappas et al. [2009] in the context of CD4$^+$ T cells. The experience obtained from the theoretical work presented in the first part of the thesis and the difficulties encountered in the reconstruction of true interactions have been very useful for the design of the new experiment. We planned an experimental setup with a dataset as informative as possible, including a double stimulation and perturbation of the system by RNAi silencing of candidate modulators. Instead of studying a large scale regulatory network we decided to consider relevant regulatory modules. We focused on the smallest building modules, which are recurring throughout large networks, and are widely studied, the feed-forward loop motifs, described by Alon [2007]. We thus designed customized TaqMan arrays which allowed us to measure the gene expression of about 90 validated transcripts regulated by IFN-α, by state-of-the-art quantitative PCR techniques. A perturbation-based strategy was applied: we focused on few candidate modulators and perturbed the system by siRNA inactivation of these targets followed by stimulation with IFN-α. Silencing effects were evaluated with respect to a calibrator siRNA using the $\Delta\Delta CT$ method and a significance analysis of IFN-α-induced transcriptional changes was developed. The significative modulations were elicited through a two-stage selection procedure, that first filters observations by a variance based criterion and then applies a variable-by-variable statistical test procedure, with a Bonferroni multiple testing correction. The analysis performed provided useful information about the impact of each candidate modulator and about the genes

regulated by it. We were able to reconstruct some regulations from the IFN-α-induced transcriptional network and identify modulators of IFN-α signalling pathway. Among the inferred regulations, of specific interest is the feed-forward loop in which IFIH1 is activated from STAT1 and they both coherently repress IFNAR1. A transcriptional regulation of either STAT1 or IFIH1 on an element upstream in the pathway has never been reported yet and, if it will be confirmed by functional validation, this could be a self-standing innovative result. STAT1 was confirmed as a positive modulator, able to activate several ISGs. GBP1 was firstly noticed as a negative modulator of IFN-α pathway. Three regulatory subnetworks, involving three couples of genes (STAT1-IFIH1, OAS2-IRF1, and GBP1-OAS2), were identified. From a systems biology perspective, the inferred regulatory modules may be merged to reconstruct a transcriptional subnetwork in which the candidate modulators are the only regulators. Such information is completely new and could help the IFN-α community in developing new biological hypotheses and design new biological studies. Another very interesting result was the identification of a mini-signature that can be used to test the responsiveness of IFN-α signalling and the identification of three sentinel-genes, i.e. SAMD9, FAM46A, and CXCL10. The discovery of modulators of IFN-α transcriptional response will be extremely useful to identify future targets of intervention in knock-out experiments, which will allow to check the relevance of these genes in vivo in relationship to endogenous IFN-α signalling. Moreover, some of these genes could possibly mediate anti-angiogenic effects of other angiogenesis inhibitors, and their discovery could open new avenues of intervention. Finally, the information generated by this work will provide a basis to investigate possible genetic alterations of target genes (hitting both modulators and sentinels) in tumor cells, which could explain their often reduced responses to IFNs compared to normal cells.

The reductionist approach used in this thesis, passing from a first experience of inference in large scale problem, to a smaller one focused on a ninetieth of transcripts, including genes from the IFN-α signaling pathway central to our problem, proved successful. The results obtained in this second part of the thesis are of great biological interest were useful to generate a reasonable number of new biological

hypotheses that will be validated by Dr Indraccolo and his team. The novel experimental design developed and the novel method used to reconstruct regulatory modules are transportable and widely applicable to various biological problems.

Appendix A

A.1 Michaelis-Menten model for gene trascription

As a possible extension of the linear model on log-scale presented in subsection 1.1.1, we considered, Grassi and Wit [2008b], a non-linear kinetic model of gene transcription which extends to multiple input motifs the Michaelis-Menten kinetics, thoroughly investigated in Khanin et al. [2007] for a single input motif (a regulatory structure in which a set of genes is regulated by a single TF). Let $y_j(t)$ represent the true mRNA abundance level of gene j at time t, then a simple transcription model consists of a production term with a saturating behavior, $p_j(t)$, and a degradation term $\delta_j y_j(t)$,

$$\dot{y}_j(t) = p_j(t) - \delta_j y_j(t). \tag{A.1}$$

Let $\eta_i(t)$ be the active levels of TF proteins (both activators and repressors) of gene j, and K_j be the number of its regulators. Supposing that the effect on the transcription of gene j is due to the cooperative action of its TFs, then we can model the production term as

$$p_j(t) = \beta_j \frac{\sum_{i=1}^{K_j} z^{ij} \eta_i(t)}{\gamma_j + \sum_{i=1}^{K_j} \eta_i(t)} + \alpha_j, \tag{A.2}$$

where α_j is the basal level of mRNA production for gene j, β_j is its rate of

production, γ_j is its half-saturation constant, and z^{ij} are 0/1 indicators that are not null only when i is an activator of gene j, so that in the numerator we sum over all activators and in the denominator over all regulators.

Appendix B

B.1 Preliminary analyses to select the genes to be monitored

To select the genes to be inserted in the custom cards, we analyzed some recent results obtained by the IOV research group, Indraccolo et al. [2007] and Moserle et al. [2008]. Both works bear the signature of IFN-α, but in different cell types: the first studies EC (specifically HUVEC) and FH (human fibroblasts), while the latter two sub-populations of EOC (epithelial ovarian cancer): the side-population (SP) and non-side population (non-SP). Gene expression analyses were carried out in all four cases at 5 hours from IFN-α stimulation. The main effect of IFN-α stimulation was of up-regulation for all the considered cell types, while no genes were significantly down-regulated.

The quantification of differential expression between cells treated with IFN-α and untreated was performed using the following definition of Fold-Change

$$FC = \text{sign}(\bar{x}_T - \bar{x}_C) \cdot 2^{|\bar{x}_T - \bar{x}_C|},$$

where $\bar{x}_T$ and $\bar{x}_C$ indicate, respectively, the gene expression value on a log-scale averaged across treated and control replicates (RMA output).
The methods applied to extract information from the four datasets are:

1. Analysis based on Fold Change: intersection of the lists of genes up-regulated

in the 4 cell types with FC cut-off =5.

2. Gene Sets Enrichment Analysis (GSEA) method, see Subramanian et al. [2005] .

B.1.1 Analysis based on Fold Change: up-regulation with FC cut-off 5

Fixing the cut-off on FC to 5, 95 genes are selected as upregulated in endothelial cells, see Table B.1.

Genes up-regulated in EC with FC cut-off 5							
1	RSAD2	26	GBP1	51	LBA1	76	GMPR
2	IFIT1	27	SLC15A3	52	CASP1	77	APOL2
3	CXCL11	28	IFI35	53	DDX60	78	UBA7
4	CXCL10	29	IRF7	54	TDRD7	79	LAP3
5	IFI44L	30	TRIM14	55	UBE2L6	80	PARP12
6	MX2	31	SAMD9	56	DSP	81	APOBEC3F
7	MX1	32	TNFSF10	57	C19orf66	82	CCL8
8	OASL	33	GCH1	58	TMEM140	83	CXCL9
9	IFIT3	34	IFIT5	59	PLSCR1	84	KCTD14
10	OAS2	35	FAM46A	60	TRIM21	85	TNFSF18
11	IFIT2	36	SP110	61	LGALS9	86	ETV7
12	OAS1	37	ISG15	62	IL15RA	87	CCRL1
13	HERC5	38	APOL6	63	DHX58	88	MSX1
14	RTP4	39	C1orf38	64	IL15	89	MYD88
15	IDO1	40	USP18	65	APOBEC3G	90	PDZD2
16	HERC6	41	STAT2	66	CX3CL1	91	NMI
17	PSMB9	42	STAT1	67	TAP2	92	EIF2AK2
18	IFI6	43	TAP1	68	IRF1	93	GBP2
19	DDX58	44	XAF1	69	PML	94	PHF11
20	IFI44	45	ISG20	70	APOL3	95	SP140L
21	IFIH1	46	ZC3HAV1	71	PLEKHA4		
22	SECTM1	47	APOL1	72	RARRES3		
23	TLR3	48	SLC25A28	73	PSMB8		
24	IFITM1	49	BST2	74	CFB		
25	OAS3	50	SAMHD1	75	IFI30		

Table B.1: List of the 95 genes up-regulated in EC, using a FC cut-off = 5 as criterion of selection. Gene list is in decreasing order with respect to FC. The first 52 genes correspond to the genes selected with a more restrictive FC cut-off=10. **The gene selected to be monitored in the custom card are the first 77, two of which, STAT1 and STAT2, were already included because part of the IFN-α signaling pathway.**

Venn diagram, in Figure B.1, quantitatively illustrates the intersections between

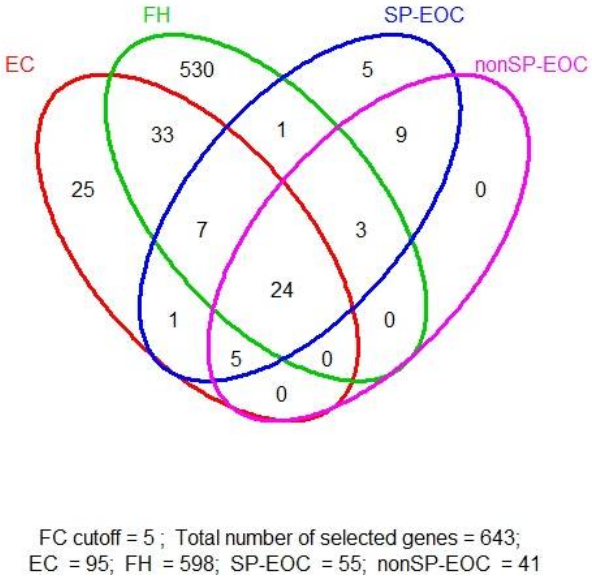

Figure B.1: Venn diagram, FC cut-off 5

the lists of genes showing an induction of at least 5 times (treated vs. control) in the 4 cell types. In Table B.2 and Table B.3 are shown the IDs of genes up-regulated in EC and belonging to each of the intersections highlighted by the Venn diagram.

B.1.2 Functional analysis with GSEA

In our analysis, we tested the enrichment of all the 'functional' gene sets corresponding to metabolic or signaling pathways from Biocarta, GenMAPP,or Kegg databases. In Table B.4, the intersection of results obtained for EC and for the other cell types are shown. Notice that the IFN-α pathway is the most enriched in all four cellular types.

Up-regulated in EC, FH, SP and nonSP-EOC (24)
STAT1
MX1
GBP1
TAP1
PLSCR1
OAS1
IFIT1
IFIT5
IFI44L
CXCL10
IFIT3
OAS2
MX2
ISG15
OASL
SP110
RSAD2
IFI44
IFIT2
OAS3
DDX58
USP18
RTP4
ZC3HAV1

Up-regulated only in EC (25)
CFB
GBP2
LGALS9
CXCL9
RARRES3
APOBEC3G
SAMHD1
MSX1
IL15
PDZD2
APOL1
IDO1
EIF2AK2
CCL8
LAP3
PLEKHA4
DHX58
CCRL1
APOL3
TNFSF18
APOL2
ETV7
FAM46A
KCTD14
CX3CL1

Table B.2: *FC* cut-off 5. Genes up-regulated in all four cellular types (left) and only in EC (right).

Up-regulated in EC and FH (33)
DSP
IFI30
BST2
UBE2L6
IRF1
TRIM14
UBA7
NMI
GMPR
GCH1
IFI6
ISG20
TAP2
STAT2
CASP1
XAF1
TLR3
PML
IL15RA
C1orf38
IRF7
PSMB8
MYD88
IFI35
LBA1
TDRD7
SECTM1
SP140L
APOBEC3F
PARP12
SLC15A3
SLC25A28
PHF11

Up-regulated in EC and SP-EOC (1)
CXCL11

Up-regulated in EC, SP and nonSP-EOC (5)
TNFSF10
DDX60
IFIH1
SAMD9
HERC5

Up-regulated in EC, FH, SP-EOC (7)
IFITM1
PSMB9
TRIM21
C19orf66
TMEM140
HERC6
APOL6

Table B.3: *FC* cut-off 5. Genes up-regulated in EC and at least another cellular type.

Gene sets enriched in EC, FH, SP-EOC, and nonSP-EOC
IFNAPATHWAY
IFNGPATHWAY
IL22BPPATHWAY

Gene sets enriched only in EC
HSA04620_TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY
HSA04060_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION
FOLATE_BIOSYNTHESIS
HSA00791_ATRAZINE_DEGRADATION
HSA00790_FOLATE_BIOSYNTHESIS

Gene sets enriched in EC and FH
PROTEASOME

Gene sets enriched in EC, FH, SP-EOC
LDLPATHWAY

Table B.4: GSEA (up-regulation) results: intersection of the 10 gene sets with highest ES in each of the 4 cellular types.

Appendix C

C.1 Biological variance estimation

In order to estimate the biological variance of $\Delta\Delta CT$, three different strategies may be adopted: 1) an error propagation model starting from the raw CT values; 2) an error propagation model that takes the ΔCT as basic unit of information, due to the inherent relative nature of real RT-PCR data; 3) direct estimate of $\Delta\Delta CT$ variance.

C.2 Error propagation

Let us suppose to be interested in estimating the variance of a function f of n observable variables $x_1, x_2, \ldots, x_n$. In the general case in which the quantity of interest f is a non-linear function, it is usually linearized by a first-order Taylor expansion

$$f \approx f^0 + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} x_i,$$

where $\frac{\partial f}{\partial x_i}$ is the partial derivative of f with respect to the i-th variable. Since f^0 is a constant it does not contribute to the variance of f. Therefore σ_f^2, the

variance of f, is given by

$$\sigma_f^2 = \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i} \right|^2 \cdot \sigma_{x_i}^2 + 2 \cdot \sum_{j=2}^{n} \sum_{k=1}^{j-1} \left(\frac{\partial f}{\partial x_j} \cdot \frac{\partial f}{\partial x_k} \cdot \mathrm{cov}(x_j, x_k) \right),$$

where $\sigma_{x_i}^2$ is the variance of x_i, and $\mathrm{cov}(x_j, x_k)$ is the covariance of x_j and x_k.

C.3 Estimation of propagation error in real-time RT-PCR

The random error observed in real-time RT-PCR originates from both the RT step and the real time PCR step Ståhlberg et al. [2004]. In Nordgård et al. [2006], the propagation error associated with the sole real time PCR step was analyzed. As in our experimental set-up we have biological replicates, we need an expression for the propagation error associated with the biological variability. This quantity is actually confounded with the variability of the whole procedure as at each biological replicate corresponds a proper RT followed by a proper real-time PCR, see Figure 5.5.

Case 1: error propagation starting from CT values
We want the biological variance of $\Delta\Delta CT$ to be expressed in terms of the variance of the measured raw CT data. Recall that $\Delta\Delta CT = \Delta CT^T - \Delta CT^{CB} = CT_X^T - CT_R^T - CT_X^{CB} + CT_R^{CB}$. Using the error propagation formula with $f = \Delta\Delta CT$ and $\mathbf{x} = \left\{CT_X^T, CT_R^T, CT_X^T, CT_X^{CB}, CT_R^{CB}\right\}$ we get:

$$\begin{aligned}\sigma_{\Delta\Delta CT}^2 &= \left|\frac{\partial f}{\partial CT_X^T}\right|^2 \sigma_{CT_X^T}^2 + \left|\frac{\partial f}{\partial CT_R^T}\right|^2 \sigma_{CT_R^T}^2 + \left|\frac{\partial f}{\partial CT_X^{CB}}\right|^2 \sigma_{CT_X^{CB}}^2 + \left|\frac{\partial f}{\partial CT_R^{CB}}\right|^2 \sigma_{CT_R^{CB}}^2 + \mathrm{COV1} \\ &= \left(\sigma_{CT_X^T}^2 + \sigma_{CT_R^T}^2 + \sigma_{CT_X^{CB}}^2 + \sigma_{CT_R^{CB}}^2\right) + \mathrm{COV1}, \end{aligned} \tag{C.1}$$

where $\mathrm{COV1} = 2\left[-\mathrm{cov}(CT_X^T, CT_R^T) - \mathrm{cov}(CT_X^T, CT_X^{CB}) + \mathrm{cov}(CT_X^T, CT_R^{CB}) + \right.$

$$+\mathrm{cov}(CT_R^T, CT_X^{CB}) - \mathrm{cov}(CT_R^T, CT_R^{CB}) - \mathrm{cov}(CT_X^{CB}, CT_R^{CB})\big] .$$

Case 2: error propagation starting from ΔCT values

An alternative way of propagating the error may be developed by considering the basic unit of information given by the ΔCT according to the inherent relative nature of RT-PCR data. In this case $\Delta\Delta CT = \Delta CT^T - \Delta CT^{CB}$ and:

$$\begin{aligned} \sigma^2_{\Delta\Delta CT} &= \left|\frac{\partial f}{\partial \Delta CT^T}\right|^2 \sigma^2_{\Delta CT^T} + \left|\frac{\partial f}{\partial \Delta CT^{CB}}\right|^2 \sigma^2_{\Delta CT^{CB}} + \mathrm{COV2} \\ &= \left(\sigma^2_{\Delta CT^T} + \sigma^2_{\Delta CT^{CB}}\right) + \mathrm{COV2}, \end{aligned} \tag{C.2}$$

where $\mathrm{COV2} = -2\,\mathrm{cov}(\Delta CT^T, \Delta CT^{CB})$.

Case 3: error propagation starting from $\Delta\Delta CT$

The last possibility is to directly estimate $\sigma^2_{\Delta\Delta CT}$. This is the approach followed in this thesis, whose detailed model and fitted result are presented in section 5.2.

References

Bruce Alberts. *Molecular biology of the cell*. Garland Science, 2002. 2, 5

Uri Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman & Hall/CRC, 2006. 9

Uri Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461, 2007. 10, 72

P. Aloy and R. B. Russell. Taking the mystery out of biological networks. *EMBO Reports*, 5(4):349–350, Apr. 2004. 17

Victor Ambros and Xuemei Chen. The regulation of genes and genomes by small RNAs. *Development*, 2007. 5

Mukesh Bansal, Vincenzo Belcastro, Alberto Ambesi-Impiombato, and Diego di Bernardo. How to infer gene networks from expression profiles. *Molecular Systems Biology*, 3, 2007. 11

Albert-Laszlo Barabasi and Zoltan N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004. 9

M. Barenco, D. Tomescu, D. Brewer, R. Callard, J. Stark, and M. Hubank. Ranked prediction of p53 targets using hidden variable dynamic modeling. *Genome Biology*, 7(3):R25, 2006. 11

Ernest C Borden, Ganes C Sen, Gilles Uze, Robert H Silverman, Richard M Ransohoff, Graham R Foster, and George R Stark. Interferons at age 50: past,

current and future impact on biomedicine. *Nature Reviews Drug Discovery*, 6 (12):975–990, 2007. 42

T. Chen, H. L. He, and G. M. Church. Modeling gene expression with differential equations. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 29–40, 1999. ISSN 1793-5091. 11

Barbara Di Camillo, Gianna Toffolo, and Claudio Cobelli. A gene network simulator to assess reverse engineering algorithms. *Annals Of The New York Academy Of Sciences*, 1158:125–142, 2009. 23

P. Erdos and A. Renyi. On random graphs. *Publicationes Mathematicae*, 6:290, 1959. 9

Fulvia Ferrazzi, Paola Sebastiani, Marco F Ramoni, and Riccardo Bellazzi. Bayesian approaches to reverse engineer cellular systems: a simulation study on nonlinear gaussian networks. *BMC Bioinformatics*, 8(Suppl 5):S2, 2007. 11

Andrew Fire, SiQun Xu, Mary K. Montgomery, Steven A. Kostas, Samuel E. Driver, and Craig C. Mello. Potent and specific genetic interference by double-stranded rna in caenorhabditis elegans. *Nature*, 391(6669):806–811, 1998. xviii, 12

Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using Bayesian Networks to Analyze Expression Data. *Journal of Computational Biology*, 7(3-4):601–620, August 2000. ISSN 1066-5277. doi: 10.1089/106652700750050961. 11

Timothy S. Gardner and Jeremiah J. Faith. Reverse-engineering transcription control networks. *Physics of Life Reviews*, 2(1):65–88, March 2005. 11

Angela Grassi and Ernst Wit. Bayesian modelling for genetic networks with topological constraints. In *In Proceedings of the Fifth International WCSB Workshop on Computational Systems Biology, June 11-13, Leipzig, Germany*, pages 45–48, 2008a. URL `http://www.cs.tut.fi/wcsb08/wcsb08.pdf`. 26

Angela Grassi and Ernst Wit. Genetic networks with topological constraints: a bayesian approach. In *In Atti del Primo Congresso Nazionale di Bioingegneria, July 3-5, Pisa, Italy*, pages 67–68, 2008b. 75

N. Guelzim, S. Bottani, P. Bourgine, and F. Képès. Topological and causal structure of the yeast transcriptional regulatory network. *Nature Genetics*, 31: 60 – 63, May. 2002. 17

Michael Hecker, Sandro Lambeck, Susanne Toepfer, Eugene van Someren, and Reinhard Guthke. Gene regulatory network inference: Data integration in dynamic modelsA review. *Biosystems*, 96(1):86–103, 2009. 7

A. Regev I. Nachman and N. Friedman. Inferring quantitative models of regulatory networks from expression data. *Bioinformatics*, 20:248 – 256, Aug. 2004. 11

Stefano Indraccolo, Ulrich Pfeffer, Sonia Minuzzo, Giovanni Esposito, Valeria Roni, Susanna Mandruzzato, Nicoletta Ferrari, Luca Anfosso, Raffaella Dell'Eva, Douglas M Noonan, Luigi Chieco-Bianchi, Adriana Albini, and Alberto Amadori. Identification of genes selectively regulated by IFNs in endothelial cells. *J Immunol*, 178(2):1122–35, 2007. 37, 38, 40, 77

H. Jeong, S. P. Mason, A.-L. Barabasi, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411:41 – 42, May. 2001. 17

R. Khanin, V. Vinciotti, V. Mersinias, C. P. Smith, and E. Wit. Statistical reconstruction of transcription factor activity using michaelis-menten kinetics. *Bioinformatics*, 63:816 – 823, Sep. 2007. 15, 75

Hiroaki Kitano. Systems Biology: A Brief Overview. *Science*, 295(5560):1662–1664, 2002. 1

Adam Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Favera, and Andrea Califano. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics*, 7(Suppl 1):S7+, 2006. 11

Florian Markowetz and Rainer Spang. Inferring cellular networks - a review. *BMC Bioinformatics*, 8(Suppl 6):S5+, 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-S6-S5. 11

R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network Motifs: Simple Building Blocks of Complex Networks. *Science*, 298 (5594):824–827, 2002. 9, 10

Sonia Minuzzo, Lidia Moserle, Stefano Indraccolo, and Alberto Amadori. Angiogenesis meets immunology: Cytokine gene therapy of cancer. *Molecular Aspects of Medicine*, 28(1):59 – 86, 2007. 37, 38

Lidia Moserle, Stefano Indraccolo, Margherita Ghisi, Chiara Frasson, Elena Fortunato, Silvana Canevari, Silvia Miotti, Valeria Tosello, Rita Zamarchi, Alberto Corradin, Sonia Minuzzo, Elisabetta Rossi, Giuseppe Basso, and Alberto Amadori. The side population of ovarian cancer cells is a primary target of IFN-α antitumor effects. *Cancer Research*, 68(14):5658–5668, 2008. 38, 77

Muhammad Munir. Trim proteins: Another class of viral victims. *Science Signaling*, 3(118):jc2, 2010. 8

Oddmund Nordgård, Jan Terje Kvaly, Ragne Kristin Farmen, and Reino Heikkil. Error propagation in relative real-time reverse transcription polymerase chain reaction quantification models: the balance between accuracy and precision. *Analytical Biochemistry*, 356(2):182–193, 2006. 83

D J Pappas, G Coppola, P A Gabatto, F Gao, D H Geschwind, J R Oksenberg, and S E Baranzini. Longitudinal system-based analysis of transcriptional responses to type i interferons. *Physiol Genomics*, 38(3):362–71, 2009. 72

E. Pennisi. Systems biology. Tracing life's circuitry. *Science*, 302(5651):1646–1649, 2003. 1

V. Plagnol and S. Tavaré. Approximate bayesian computation and mcmc. In H. Niederreiter, editor, *In Monte Carlo and Quasi-Monte Carlo Methods 2002*, pages 99–114. Springer-Verlag, 2004. 18, 20

Jüri Reimand, Juan M. Vaquerizas, Annabel E. Todd, Jaak Vilo, and Nicholas M. Luscombe. Comprehensive reanalysis of transcription factor knockout expression data in Saccharomyces cerevisiae reveals many new targets. *Nucleic Acids Research*, 38(14):4768–4777, 2010. 12

J. Schäfer and K. Strimmer. An empirical bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21:754–764, Mar. 2005. 11

Thomas D. Schmittgen and Kenneth J. Livak. Analyzing real-time PCR data by the comparative CT method. *Nature Protocols*, 3(6):1101–1108, 2008. 45

Shai S. Shen-Orr, Ron Milo, Shmoolik Mangan, and Uri Alon. Network motifs in the transcriptional regulation network of Escherichia coli. *Nature Genetics*, 31(1):64–68, 2002. 9

Ilya Shmulevich, Edward R. Dougherty, Seungchan Kim, and Wei Zhang. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18(2):261–274, February 2002. ISSN 1367-4803. doi: 10.1093/bioinformatics/18.2.261. 11

Younan A. Sidky and Ernest C. Borden. Inhibition of angiogenesis by interferons: Effects on tumor-and lymphocyte-induced vascular responses. *Cancer Research*, 47(19):5155–5161, 1987. 37

P. Smolem, D.A. Baxter, and J.H. Byrne. Modeling transcriptional control in gene networks. *Bulletin of mathematical biology*, 62(2):247–292, 2000. 11

Richelle Sopko, Dongqing Huang, Nicolle Preston, Gordon Chua, Balázs Papp, Kimberly Kafadar, Mike Snyder, Stephen G. Oliver, Martha Cyert, Timothy R. Hughes, Charles Boone, and Brenda Andrews. Mapping pathways and phenotypes by systematic gene overexpression. *Molecular cell*, 21(3):319–330, 2006. 12

Nicola Soranzo, Ginestra Bianconi, and Claudio Altafini. Comparing association network algorithms for reverse engineering of large-scale gene regulatory networks: synthetic versus real data. *Bioinformatics*, 23(13):1640–1647, 2007. 11, 26

P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9:3273 – 3297, 1998. 24

Anders Ståhlberg, Joakim Hkansson, Xiaojie Xian, Henrik Semb, and Mikael Kubista. Properties of the reverse transcription reaction in mrna quantification. *Clinical Chemistry*, 50(3):509–515, 2004. 83

Gustavo Stolovitzky, Don Monroe, and Andrea Califano. Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference. *Annals of the New York Academy of Sciences*, 1115(1):1–22, December 2007. 11

Gustavo Stolovitzky, Robert J. Prill, and Andrea Califano. Lessons from the DREAM2 Challenges: A Community Effort to Assess Biological Network Inference. *Annals of the New York Academy of Sciences*, 1158(1):159–195, March 2009. ISSN 0077-8923. doi: 10.1111/j.1749-6632.2009.04497.x. 12

Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. 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, 2005. 78

A. H. Y. Tong, G. Lesage, G. D. Bader, H. Ding, H. Xu, X. Xin, J. Young, G. F. Berriz, R. L. Brost, M. Chang, Y. Chen, X. Cheng, G. Chua, H. Friesen, D. S. Goldberg, J. Haynes, C. Humphries, G. He, S. Hussein, L. Ke, N. Krogan, Z. Li, J. N. Levinson, H. Lu, P. Menard, C. Munyana, A. B. Parsons, O. Ryan, R. Tonikian, T. Roberts, A-M. Sdicu, J. Shapiro, B. Sheikh, B. Suter, S. L. Wong, L. V. Zhang, H. Zhu, C. G. Burd, S. Munro, C. Sander, J. Rine, J. Greenblatt, M. Peter, A. Bretscher, G. Bell, F. P. Roth, G.W. Brown, B. Andrews, H. Bussey, and C. Boone. Global mapping of the yeast genetic interaction network. *Science*, 303:808 – 813, Feb. 2004. 24

J. Vandesompele, M. Kubista, and M. W. Pfaffl. Reference gene validation software for improved normalization. In J. Logan, K. Edwards, and N. Saunders, editors, *Real-Time PCR: Current Technology and Applications*, pages 47–64. Caister Academic Press, 2009. 44

Yong Wang, Trupti Joshi, Xiang-Sun Zhang, Dong Xu, and Luonan Chen. Inferring gene regulatory networks from multiple microarray datasets. *Bioinformatics*, 22(19):2413–2420, October 2006. doi: 10.1093/bioinformatics/btl396. 11

E. Wit and N. Thomson. Bayesian genetic networks with topological constraints. In *Proc. International Workshop on Statistical Modelling*, Sydney, Australia, Jul. 2005. 17

Joshua Yuan, Ann Reed, Feng Chen, and C. Neal Stewart. Statistical analysis of real-time PCR data. *BMC Bioinformatics*, 7(1):85+, 2006. 44