Telethon Institute of Genetics and Medicine (TIGEM)

University of Naples "Federico II"

RESEARCH DOCTORATE (PhD) in COMPUTATIONAL BIOLOGY and BIOINFORMATICS

"Identification of transcriptional and post-translational regulatory networks from gene expression profile: an information-theoretic approach."

Supervisor:
Dr. Diego di Bernardo

Candidate:
Gennaro Gambardella

Co-Supervisor
Dr. Adriano Peron

Coordinator:
Dr. Sergio Cocozza

YEAR 2010/2013

Out of the night that covers me,
Black as the pit from pole to pole,
I thank whatever gods may be
For my unconquerable soul.

In the fell clutch of circumstance
I have not winced nor cried aloud.
Under the bludgeonings of chance
My head is bloody, but unbowed.

Beyond this place of wrath and tears
Looms but the Horror of the shade,
And yet the menace of the years
Finds and shall find me unafraid.

It matters not how strait the gate,
How charged with punishments the scroll.
I am the master of my fate:
I am the captain of my soul.

(Traduzione)

Dal profondo della notte che mi avvolge,
Buia come un pozzo che va da un polo all'altro,
Ringrazio qualunque dio esista
Per l'indomabile anima mia.

Nella feroce stretta delle circostanze
Non mi sono tirato indietro né ho pianto forte.
Sotto i colpi d'ascia della sorte
Il mio capo è sanguinante, ma indomito.

Oltre questo luogo d'ira e di lacrime
Si profila il solo Orrore delle ombre,
E ancora la minaccia degli anni
Mi trova e mi troverà senza paura.

Non importa quanto stretto sia il passaggio,
Quanto piena di castighi la vita,
Io sono il padrone del mio destino:
Io sono il capitano della mia anima.

Invictus *(W. E. Henley 1849-1903)*

Table of Contents

List of Tables

List of Figures

Abstract

The reconstruction of the regulatory interactions among DNA, RNAs and proteins in a cell is probably the most important and key challenge in molecular biology. In the last decade, the introductions of new high-throughput technologies, such as microarrays and, more recently, next generation sequencing (NGS) have facilitated this task. Different Systems Biology approaches have been proposed to reconstruct the transcriptional, post-transcriptional and the post-translational regulatory networks of a cell starting from genomics data. The two aims of the research here described are: (1) the development and the application of a computational method for the identification of tissue-specific, or more broadly, condition-specific pathways; (2) the development and the application of a computational approach for the identification of post-translational modulators of transcription factor activity from gene expression profiles.

In **Chapter 1**, I provide a brief overview of the different molecular networks known to exist in a living cell. **Chapter 2** illustrates a comparative study of the different approaches to reverse-engineering gene networks from gene expression profiles (GEPs) and their limitations. Current state-of-the-art reverse-engineering approaches model gene networks as static processes, i.e. regulatory interactions among genes in the network (such as direct physical interactions or indirect functional interactions) do not change across different conditions or tissue types. However, different cell-types, or the same cell-type but in different conditions, may carry out very different functions, thus it is expected that their regulatory networks may reflect these differences.

In **Chapter 3 and 4**, I describe the development of a novel approach named DINA (Differential Network Analysis) for the identification of differentially co-regulated pathways. DINA is based on the hypothesis that genes belonging to a condition-specific pathway are actively co-regulated only when the pathway is active, independently of their absolute level of expression. I first reverse-engineered 30 tissue-specific networks from a collection of about 3000 GEPs. I then applied DINA to these networks in order to identify tissue-specific pathways starting from a list of 110 KEGG-annotated pathways. As expected, DINA predicted many metabolic pathways to be tissue-specific and prevalently active in liver and kidney. I then built a simplified model of hepatocellular carcinoma (HCC) to mimic the HCC progression using three condition-specific regulatory networks obtained from three different cell-lines: (i) primary hepatocyte, (ii) HepG2 and (iii) Huh7. Using these three cell-type specific networks, I demonstrated that DINA can be used to make hypotheses on dysregulated pathways during disease

progression. DINA is also able to predict which Transcription Factors (TFs) may be responsible for the pathway condition-specific co-regulation. I tested this approach to identify regulators of tissue-specific metabolic pathways, and I correctly identified Nuclear Receptors as their main regulators. With this method, I was also able to identify a new putative tissue-specific negative regulator of hepatocyte metabolism: Yeats2.

In **Chapter 5, 6 and 7**, I propose a generalized method that I called *Differential Multi-Information* (DMI) to identify post-translational modulators M of a transcription factor TF by observing the changes in co-regulation (measured by Multi-Information) among a set of *n* target genes $G_1 \cdots G_n$ in the presence or absence of the modulator M. My working hypothesis is that the set of target genes will be strongly co-regulated only when the modulator M is present, since the modulator will active the TF. The DMI algorithm requires in input a set of known target genes regulated by a common TF, and it returns in output a ranked list of predicted post-translational modulators of the TF. I first validated the approach using an "in-silico" datasets consisting of 100 GEPs and 760 genes. Next, I tested DMI performance on a real gene expression profile dataset, by identifying the post-translational modulators of 7 transcription factors for which I was able to collect a list of high-confident targets. This set of transcription factors included transcription factors such as P53, MYC and STAT3. Finally, as a case of study, I tested the DMI method on a transcription factor TFEB recently identified as a master regulator of lysosomal biogenesis and autophagy. By comparing the results of DMI with a High Content Screening (HCS) using siRNA oligo libraries against all the known phosphatases, I was able to show that DMI can achieve a very high precision. All these results confirm that DMI could be instrumental in identifying post-translational regulatory interactions in an efficient and cost-effective manner.

Chapter 1

Introduction to regulatory networks

Interactions among molecules in a cell lead to a complex regulatory network, which is only partially understood. Biological networks are regulated at many levels by different kinds of mechanisms. For the sake of simplicity, here I will consider only three types of regulations that can occur in a cell and thus three types of regulatory networks: (i) transcriptional regulatory networks describing transcriptional regulations such as protein-DNA interactions; (ii) post-transcriptional regulatory networks where post-transcriptional regulations occurs at the RNA level before its eventual translation; (iii) post-translational regulatory networks where post post-translational regulations occurs at the protein level, i.e. a protein covalently modified "on the fly". In this Chapter, I will give a brief introduction on each one of these three types of regulations and their regulatory networks.

1.1 Transcriptional Regulation

In eukaryotic organisms protein-coding gene expression is promoted by RNA polymerase II and it can be regulated at several steps, including transcription initiation and elongation, and mRNA processing [1]. Genes that are transcribed by the RNA polymerase II usually contain two distinct families of *cis*-acting transcriptional regulatory DNA elements: (*a*) a promoter region located near the gene, which structure can be quite complex containing many specific DNA sequences and response elements that provide a secure initial binding site for RNA polymerase and the other proteins required for the transcription, and (*b*) distal regulatory elements, which can be enhancers, silencers, insulators, or locus control regions (LCR). In particular, these *cis*-acting transcriptional regulatory elements contain recognition sites for trans-acting DNA-binding transcription factors, which function either to enhance or repress transcription [1].

The factors that are involved in the transcription of eukaryotic protein-coding genes by RNA polymerase II can be categorized into three distinct groups: (*i*) general (or basic) transcription factors (GTFs), (*ii*) promoter-specific activator proteins (activators or transcription factors), and (*iii*) coactivators (Figure 1). Moreover, the existence of multiple regulatory elements on the promoter regions of a

protein-coding gene lead to a combinatorial control of regulation, with a consequential exponentially increases for the potential number of unique expression patterns.

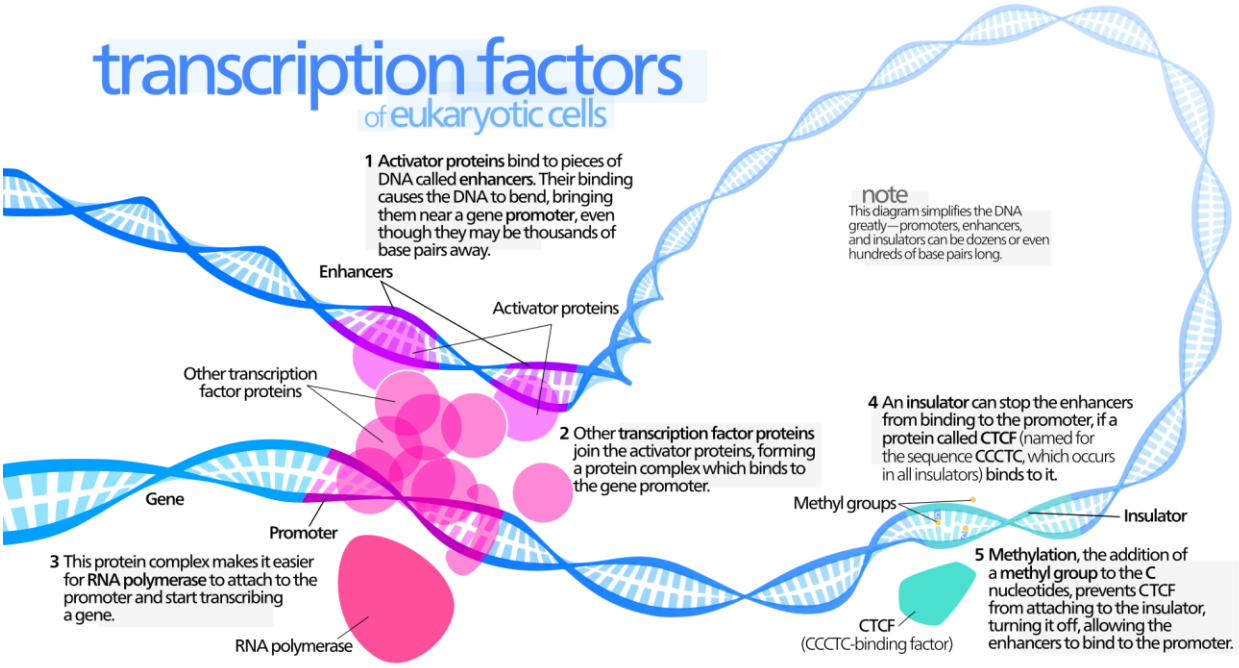


Figure 1 – Simplified diagram of transcriptional regulations (source Wikipedia).

As shown in Figure 2, the transcriptional machinery start assembling GTFs to form a transcription pre-initiation complex (PIC), this complex is useful to direct RNA polymerase II to the transcription start site (TSS). In particular, the first step in PIC assembly is binding of TFIID, a multi-subunit complex consisting of TATA-box-binding protein (TBP) and a set of tightly bound TBP- associated factors (TAFs). Then, many other steps are required before the transcription of a gene starts and, in particular, before that a fully functional RNA polymerase II elongation complex is formed [1].

The only assembly of a PIC on the core promoter is not sufficient to have a functional transcriptional machinery, but it is sufficient to direct only low levels of accurately initiated transcription from DNA templates in vitro. This process is generally referred to as basal transcription. Transcriptional activity is greatly stimulated by a second class of factors, termed activators. These activators, in general, are sequence-specific DNA-binding proteins able to recognize specific sequence to bind on the core promoter. In the last decade, different classes of activators, discriminated by different DNA-binding domains, have been discovered, each one associated with their own class of specific DNA sequences [1].

The DNA-binding sites for activators are also called transcription factor-binding sites (TFBSs). These sites are generally very small, consisting of 6–12 bp, although the binding specificity of an activator is usually stated by no more than 4–6 nucleotides (i.e. consensus sequence) within the site. For this reason the TFBSs for a specific activator is typically described by this consensus sequence for which an activator is relatively constrained, while the other nucleotides can vary. The particular sequence of a TFBS is fundamental, and it can also affect the structure of a bound activator in a way that alters its activity [1].

Finally, activators have also been proposed to function by recruiting activities that modify chromatin structure [2, 3]. Chromatin barrier can prevent the transcriptional machinery from interacting directly with promoter DNA and thus preventing activator binding and PIC assembly.

Although the phenomenon of transcriptional synergy has long been recognized, the mechanism underlying it has remained sill elusive and needs to be still studied.

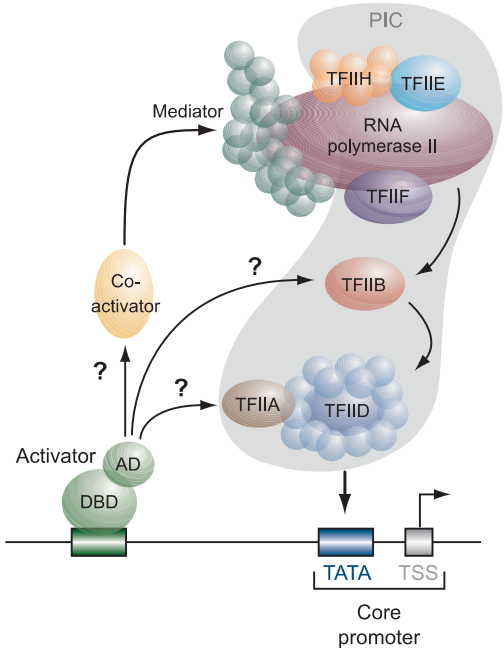


Figure 2 – Cartoon schema of the eukaryotic transcriptional machinery. Factors involved in eukaryotic transcription by RNA polymerase II can be classified into three groups: general transcription factors (GTFs), activators, and coactivators. In addition to the RNA polymerase II, the class of GTFs includes TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH. These proteins are assembled on the core promoter in order to form a pre-initiation complex (PIC), able to direct the RNA polymerase II on the transcription start site (TSS). Transcriptional activity is then promoted by other activators proteins (i.e. transcription factors), which bind specific areas of the promoter region and work stimulating PIC formation. Two parts principally compose activators: (*i*) DNA-binding domain (DBD) and (*ii*) a separable activation domain (AD) that is required for the activator to stimulate transcription. (Figure taken from the review work of Glenn et al. "*Transcriptional Regulatory Elements in the Human Genome*" [1])

1.2 Post-transcriptional regulation

Post-transcriptional regulation of gene expression performs an important role in many cellular processes including cell development, metabolism and cancer progression. In the recent years the discovery of the role of microRNAs (miRNAs) as keys molecules of post-transcriptional regulators of gene expression has emerged [4]. MicroRNAs are a large family of small, approximately 21-nucleotide-long, non-coding RNAs (Figure 3). In particular, miRNAs are able to modulate gene expression at post-transcriptional level regulating mRNA translation or stability in the cytoplasm [5-8]. It has been estimated that miRNAs may regulate about 30% of all protein-coding genes [4].

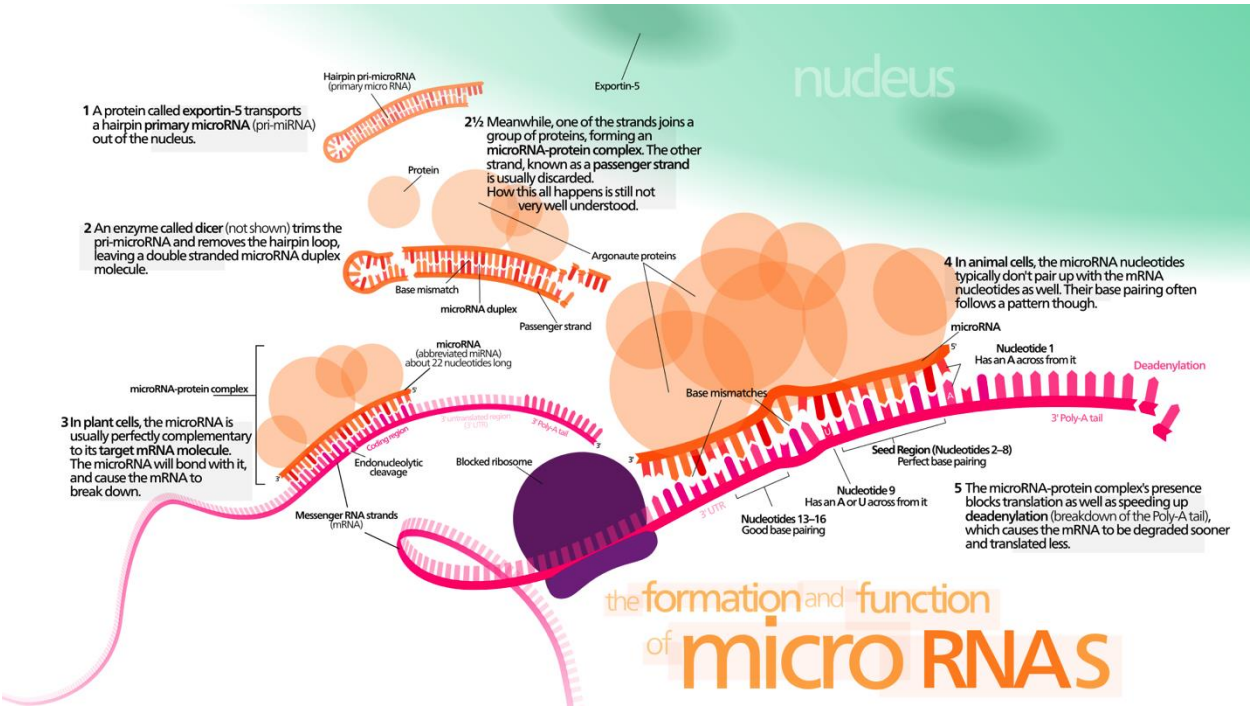


Figure 3 – Simplified diagram of microRNA (source Wikipedia).

As shown in Figure 4 microRNAs are processed from precursor molecules called pre-miRNAs, which can be either transcribed by RNA polymerase II from transcripts belonging portions of non-coding genes (exons) or portions of introns. A single pre-miRNA usually contains sequences for many different miRNAs. After its transcription, pre-miRNAs fold into hairpin structures containing imperfectly base-

paired stems. Then they are processed in two steps, catalysed by the RNase III type endonucleases Drosha (also known as RN3) and Dicer [4].

In animals, pre-miRNAs are transported to the cytoplasm by exportin5 protein, where they are thus cleaved by Dicer to yield about 20-bp miRNA duplexes. After this step, usually only one strand is then selected to function as a mature miRNA, while the other strand is degraded. Only occasionally, both arms of the pre-miRNA hairpin give rise to mature miRNAs [4].

During their processing, miRNAs are assembled into ribonucleoprotein (RNP) complexes called micro-RNPs (miRNPs) or miRNA-induced silencing complexes (miRISCs). Anyway, the assembly process of miRNA is a dynamic process, for which many details are still not well understood. The keys components of miRNPs are proteins of the Argonaute (AGO) family. In mammals, four AGO proteins (AGO1 to AGO4) are involved in miRNA pathway, but only AGO2 functions in RNAi, while the other three are involved in the miRNA repression. In particular, AGO2 seems to be involved first in the cleavage of the passenger (or sense) strand of the double-stranded siRNA, thus forming the single-stranded RNA that is used by the RISC complex as the guide strand to bind the target mRNA; then RISC can undergo multiple rounds of mRNA cleavage, mediating a robust silencing effect on the target gene. Apart from AGOs, miRNPs can contain further proteins that function as regulatory factors or effectors mediating inhibitory function of miRNPs (for more details on the production of miRNAs, refer to Box 1 in [4]).

With few exceptions, miRNA-binding sites in metazoan mRNAs lie in the 3’ UTR and usually multiple copies are required for effective repression of target genes [9-13]. However, in many cases, miRNAs pair imperfectly with their RNA targets, following a set of rules determined by experimental and bioinformatics analyses [9-13].

In addition to miRNA many other classes of non-coding RNAs have been, and are being, discovered but since these have no direct relevance to this thesis, I will not mention them in this Chpater.

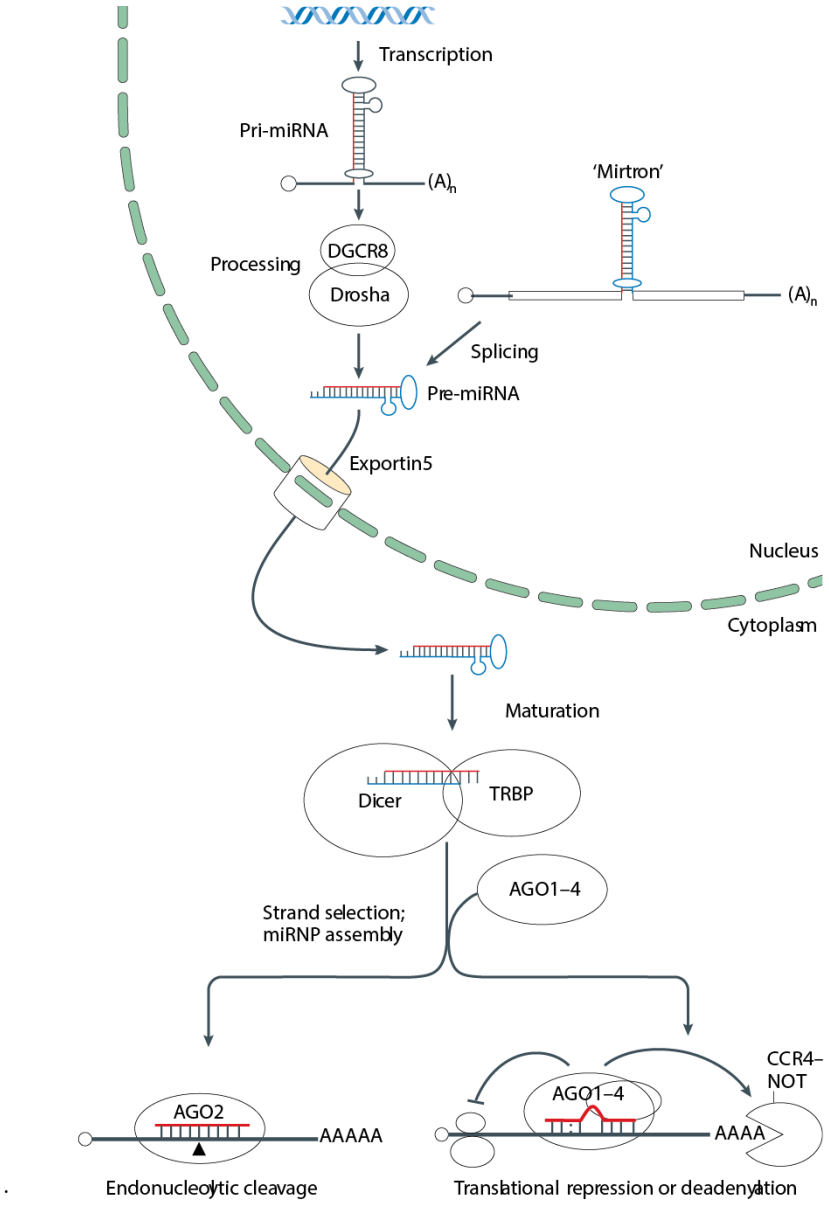


Figure 4 – The biogenesis of miroRNAs and their assembly into microribonucleoproteins (Figure taken from the review work of Filipowicz et al. "*Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight?*" [4]).

1.3 Post-translation regulation

The discovery that the human genome is composed only by about 20,000 genes [14] has stressed the complexity and the importance of regulation of gene expression and protein activity. Whereas Transcription Factors and noncoding RNA have a predominant role in regulation of gene expression, other proteins, such as kinases and phosphatases, control protein activation via post-translational modifications (PTMs). A post-translational modification is a chemical mechanism in which amino-acid residues in a protein are covalently modified "on the fly". Through this mechanism a cell is able to tightly regulate its functions regulating the activity, localization and interaction of many molecules including proteins, nucleic acids, lipids, and cofactors [15].

A simple example of PTM-mediated information processing can be found in reversible phosphorylation on a single site (Figure 5), where a single residue on a substrate can be dynamically regulated through phosphorylation or dephosphorylation. Hence, phosphorylation is a binary modification, where a given serine, threonine or tyrosine residue is either phosphorylated or not. Since a protein has in general n sites of phosphorylation then the total number of possible states is combinatorial and in particular equal to 2^n. This, obviously considering the simplest case in which each site can have only 1 modification, but in general the situation is more complex since a single site can also has k associated modifications.

Post-translational modification can occur at any step in the life of a protein. Many proteins are modified shortly after translation to regulate the folding, the stability or to direct sign a protein that must be translocated in another compartment of the cell as membrane, cytoplasm and so on. Other modifications instead can occur after the folding and the localization in order to influence the biological activity of the protein as for example occur for many transcription factors.

In addition to phosphorylation events many other kinds of post-translational modifications can occur, including glycosylation, ubiquitination, nitrosylation, methylation, acetylation and lipidation [15]. Protein glycosylation is considered one of the major post-translational modifications, with significant effects on protein folding, conformation, distribution, stability and activity. Ubiquitination is a PTM useful to mark a protein, which must be degraded by the proteasome machinery. Ubiquitin is an 8-kDa polypeptide consisting of 76 amino acids that is appended to the protein. Nitrosylation is a reversible reaction and a critical PTM used by cells to stabilize proteins and regulate gene expression, by providing nitric oxide (NO). Methylation consists in the addition of a methyl group to a substrate or the substitution of an atom or group by a methyl group. This kind of PTM contributes to epigenetic

inheritance by, for example, modifying histone tails, but it can occur also at the DNA level (DNA methylation). Amino acid residues can be conjugated to a single methyl group or multiple methyl groups to increase the effects of modification. Acetylation is the transfer of an acetyl group to nitrogen and it occurs in almost all eukaryotic proteins through both irreversible and reversible mechanisms. It is a common method of regulating gene transcription. For example histone acetylation is a reversible event used by the cell to reduce the chromosomal condensation and thus to promote the transcription. Usually, the acetylation of these lysine residues is regulated by transcription factors containing histone acetyletransferase (HAT) activity. While transcription factors with HAT activity act as transcription co-activators, histone deacetylase (HDAC) enzymes are co-repressors that reverse the effects of acetylation by reducing the level of lysine acetylation and increasing chromosomal condensation. Lipidation is PTM used to target proteins to membranes in organelles including endoplasmic reticulum (ER), Golgi apparatus, mitochondria, vesicles (endosomes, lysosomes) and the plasma membrane.

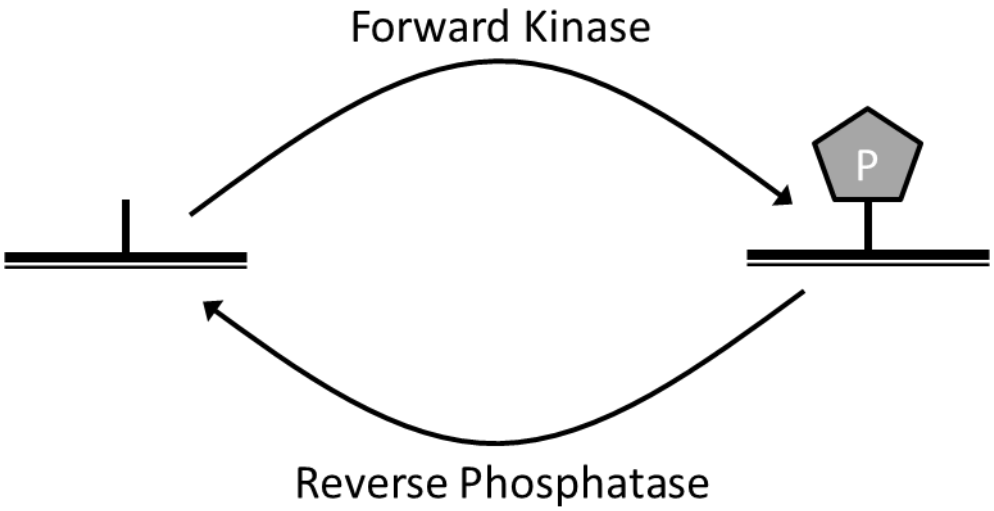


Figure 5 – Example of reversible phosphorylation. A single site is dynamically regulated adding a phosphate group by a forward kinase and removing a phosphate group by a reverse phosphatase.

1.4 Conclusion and open challenges

Following the completion of the human genome sequence draft in 2000, more than 20,000 protein-coding genes were identified and annotated [14]. Once this first milestone had been achieved, the nest and key challenge was elucidate the regulatory networks controlling expression of these genes in a condition-specific, cell-specific and tissue-specific manner. To explain the molecular mechanisms behind these specific cellular expression patterns, it is fundamental to identify the specific transcriptional regulatory sequences, such as transcription factor binding sites, associated with each one of the predicted genes, and PTM regulating transcription factor activity. Indeed in a cell thousands of molecules interact with each other in a complex regulatory network (Figure 6). Moreover, this network changes its behaviour dynamically since it is subjected also to external stimuli.

The advent of high-throughput technologies, such as microarrays and Next Generation Sequencing (NGS) based approaches, has facilitated this task. These high-throughput data are an extremely useful resource for the study of transcription regulation. High-throughput techniques provide genome-wide information regarding the establishment of spatial and temporal gene expression patterns and the mechanisms required for their establishment. In order to extract biological useful information from these data and to make evidence-based hypotheses on regulatory mechanisms active in the cell, it is essential to develop new computational “reverse-engineering” methods (refer to Chapter 2). The challenge for Systems Biology and in particular for reverse engineering is to infer gene networks, transforming high-throughput heterogeneous data sets into biological insights about the underlying mechanisms.

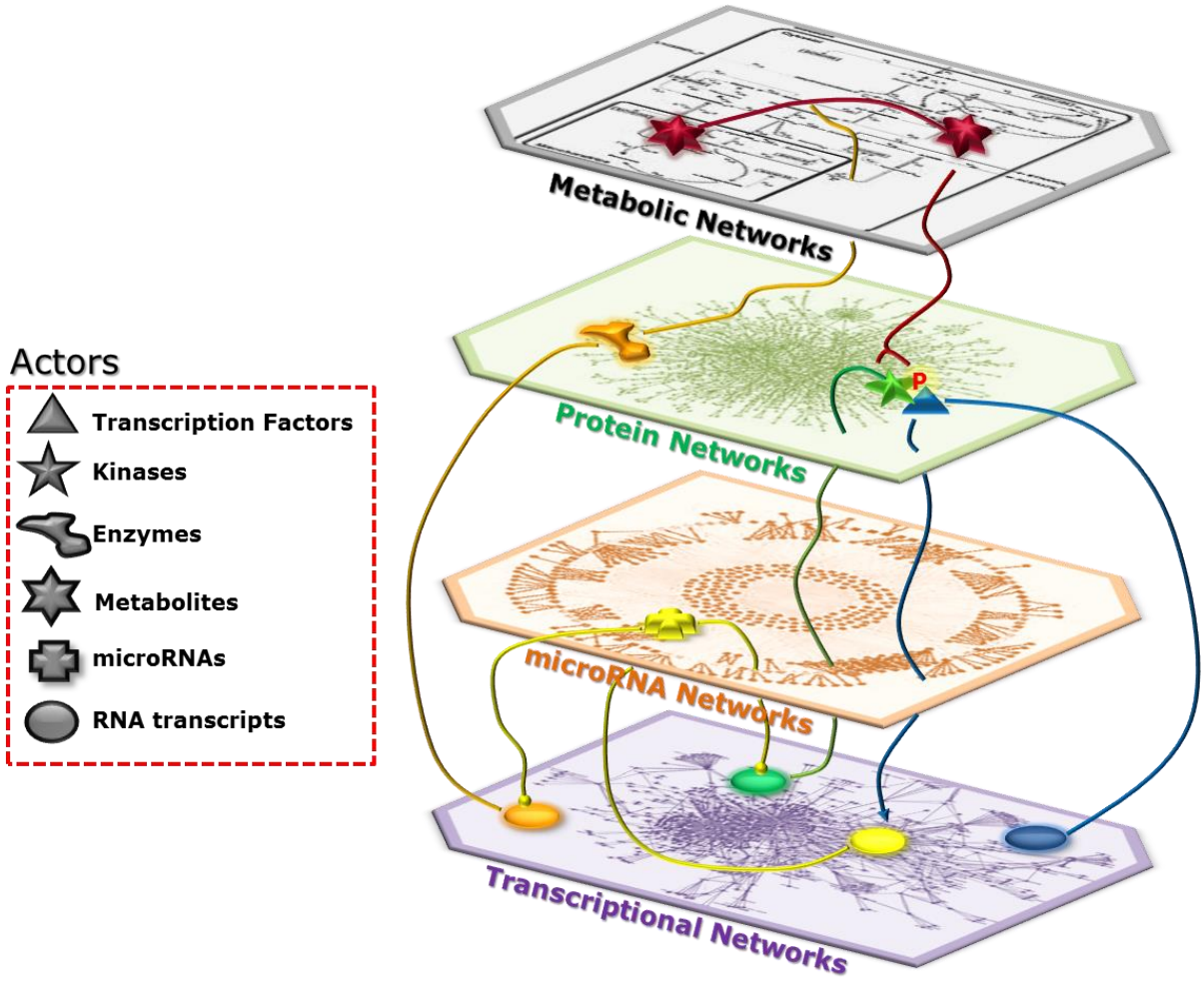


Figure 6 – A simplified example of regulatory networks in a cell. For simplicity we considered only four regulatory networks: (*i*) the transcriptional regulatory network where interaction among transcript are described; (*ii*) the microRNA regulatory network where microRNA interactions are reported; (*iii*) the protein regulatory network containing interactions among proteins, including post-translational interactions; (*iv*) the metabolic regulatory network where relationship among metabolites are reported. Obviously, all these regulatory networks are part of a single regulatory network in a cell, since the elements of these networks are interconnected. Consider, for example, the *orange* gene encoding for an *orange* enzyme able to catalyse a metabolic reaction, whose en product acts as a signal for a *green* kinase, which in turn activates a *blue* transcription factor. Finally, one of the downstream targets the *blue* transcription factor contain a *yellow* microRNA used to interfere the production of the *orange* protein.

Chapter 2

Introduction to reverse-engineering

With the diffusion of high-throughput technologies such as microarrays and more recently Next Generation Sequencing, massive genome-wide datasets measuring gene expression, and other relevant biological information, have been produced. For this reason in the last decade, new computational techniques for the analysis of these high-dimensional data have been developed giving rise to the interdisciplinary field of computational Systems Biology. In this chapter, I will introduce the main state-of-the-art reverse-engineering approaches for the reconstruction of gene regulatory network from gene expression profiles.

2.1 Introduction

As anticipated in the conclusion of the previous chapter, in the last decade, reverse-engineering techniques have been mainly focused on inferring transcriptional regulatory networks. The reason for this can be found in the diffusion of DNA microarray technology. This technology has enabled researchers to efficiently measure the concentration of all RNA transcripts in a cell with a relative low cost. On the contrary, measuring the concentration of others molecules in the cell, including proteins and metabolites, is generally more difficult, and hence such data are not abundant in the literature.

Reverse-engineering techniques can be principally divided in two classes: “physical” and “influence" approaches. A physical approach seeks to identify the protein factors that regulate transcription, and the DNA motifs on the promoter regions where these factors bind. With this technique, it is possible reduce the dimensionality of the reverse-engineering problem, since the “actors” to identify are reduced only to the TFs present in the genome. The second class, which we call the “influence” approach, seeks to reconstruct the relationship among RNA transcripts. Such model does not generally describe physical interactions between molecules since transcription is rarely controlled directly by RNAs, but with this approach it is possible to implicitly capture regulatory mechanisms at the protein and metabolite level that are not directly measured. Hence “influence” approaches are not restricted to describing only transcription factor/DNA interactions, but also indirect interactions occurring among genes via proteins, metabolites and non-coding RNAs that have not been

measured directly. The concept of influence interaction is not well defined and it is strongly dependent on the mathematical formalism used by the reverse-engineering method. Therefore, to avoid confusion, here I will refer to functional interactions by the term “connection”, whereas the term “interaction” will be used only when a physical interaction between the DNA, RNA or protein products of the genes is occurring [16].

Methods to infer a gene regulatory network rely on the analysis of the transcriptional response of a population of cells to multiple experimental conditions. High-throughput technologies such as microarray and more recently next generation sequencing allow measuring genome-wide expression under specific experimental conditions. The capability to “obtain” the fingerprint of a cell at a specific time and condition, together with the large number of expression data now available allow to use methods from engineering mathematics and statistics to explore and analyse gene expression data

An inferred gene network is therefore a collection of gene-gene connections (or TF-gene interactions) captured from expression data. A gene network is not only able to store information regarding the relationship among the transcripts but it also contains much more information that can be used to study the behaviour of a cell such as for example the topological organization of its nodes (genes). A community in a network of genes, for example, identifies a group of genes that are highly connected among them and sparsely connected with genes outside the group. These communities can be used to detect the functional modules in the cell, that is, groups of genes cooperate to accomplish specific functions.

In this chapter, I will present different approaches to infer or “reverse-engineer” influence gene regulatory networks from gene expression profiles measured by microarray technology. A description of other methods based on the physical approach and more details on computational aspects can be found in [17-22]. From an engineering point of view, knowledge of how gene expressions change following the perturbation experiment allows to identify the network of regulatory interactions occurring among them. This identification process can take different names depending on the field of application, such as: *system identification* and *reverse-engineering*.

2.2 Microarray technology and microarray data repositories

A DNA microarray (DNA chip) is a collection of small DNA oligomers laying on a solid surface of approximately 1 or 2 cm (chip). DNA oligomers on the chip are organized in approximately 250,000 "spots" (depending on the chip model), and each spot contains millions of copy of the same DNA sequence, called probe. Microarrays allow to simultaneously measuring the expression of thousands of genes starting from total RNA extracted from a cell population.

When samples are prepared, in the first step the total RNA is converted into cDNA through reverse transcriptase and then it is tagged with a fluorescent marker. Finally, cDNA is placed on the microarray chip and the complementarity between two fragments allows the hybridization of a cDNA sequence on the corresponding DNA "spot". The number of hybridized probes in a spot and the expression level of the gene are directly related. Hence, with microarray technology the expression level of a gene is quantified through fluorescence analysis.

There exist different types of microarrays, but we can divide they in two main categories:: (i) two channels microarrays and (ii) one channel microarrays. The main difference is that with two channels microarrays concurrently measure on the same chip the gene expression levels of treated cells and control cells.

Moreover there exist other types of microarray that can be used to measure single nucleotide polymorphism, fusion genes, alternative splicing, and so on. Here, we concentrate on DNA single channel microarrays and we use the term hybridization and gene expression profile (GEP) to refer to a set of gene expression levels collected on a single microarray chip. Moreover, we will refer to a set of hybridizations with the term experiment.

The fluorescence levels collected from microarray hybridization are called raw data. The hybridizations in the same experiment are usually "normalized" where noise is removed, the background fluorescence is subtracted and the average fluorescence level among the spots associated to the same probe is computed. The output of this normalization step yields a set of comparable gene expression profiles. We refer to a set of normalized GEPs (or experiment) with the term processed GEPs.

Two major GEPs public repositories exists: the Gene Expression Omibus (GEO [23]) and Array Express [24]. In both repositories, GEPs are logically divided into experiments (i.e. a collection of GEPs usually performed in a single laboratory). Still in both repositories GEPs are stored with their METADATA useful to trace different kinds of information including the experimental protocol used, the

type of samples (i.e. cell types or tissues), and many other information that the MIAME standard [25] requires. Obviously they periodically mirror their experiments.

2.3 Reverse-engineering transcriptional networks: methods and applications

2.2.1 Bayesian networks

Definition: A Bayesian Network (BN) is a Directed Acyclic Graph (DAG) $G = (V, A)$ together with a set of local probability distributions P. The vertices V corresponds to variables, and the arcs or edges A represent probabilistic dependency between the variables. An arc from variable X to variable Y states a probabilistic dependence between the two variables, i.e. the state of Y depends on the state of X. In this case, X is called a parent of Y. A node with no parents is unconditional. P contains the local probability distributions of each node X conditioned on its parents.

Bayesian Networks give a graphical representation of probabilistic relationships among a set of random discrete or continuous variables X_i, with $i = 1 \ldots n$. An example of Bayesian network is provided in (Figure 7). One of the main properties of BNs is their ability to handle incomplete data sets and their robustness to noise that is typical of biological experiments. This last property is a consequence of the robustness of the statistical model that they adopt. Since these advantages, many researchers, in the last years have devoted considerable attention in the use of Bayesian network approaches for reverse-engineering gene networks [26-33].

Bayesian networks are able to describe the relationship between variables at both qualitative and quantitative level. At a qualitative level, the relationships between variables X_i are relations of dependence and conditional independence represented as directed graph G. More precisely vertices of the graph G correspond to variables and a direct edge between a pair of vertices represents dependencies between the two variables.

At a quantitative level instead, relations between variables are described by a family of joint probability distributions $P\ (X_1, \ldots, X_n)$ that are consistent with the independence assertions embedded in the graph G and it has the form:

$$P\ (X_1, \ldots, X_n) = \prod_{1=1}^{N} P(X_i = x_i | X_j = x_j, \ldots, X_{j+p} = x_{j+p}) \qquad (2.1)$$

where the $p\ +\ 1$ genes on which the probability is conditioned are called the parents of gene i and represent its regulators, and the joint probability density is expressed as a product of conditional probabilities by applying the chain rule of probabilities and independence. This rule is based on the Bayes theorem:

$$P(A, B) = P(B \mid A)P(A) = P(A \mid B)P(B) \qquad (2.2)$$

The JPD (joint probabilty distribution) can be decomposed as the product of conditional probabilities as in Eq. 2.1 only if the Markov assumption holds, i.e. each variable X_i is independent of its non-descendants, given its parents in the directed acyclic graph G. A schematic overview of the theory underlying Bayesian networks is given in Figure 7.

The joint probability distribution, $P(A, \ldots, H)$, for the Bayesian network in Figure 7 is given by

$$P(D)P(E)P(H)P(B \mid D)P(C \mid E)P(A \mid B, C)P(F \mid A, H)P(G \mid A) \qquad (2.3)$$

In order to reverse engineer a gene network using a Bayesian network model, two sets of parameters have to be estimated: (*i*) the conditional probability functions relating the state of the regulators to the state of the transcripts and (*ii*) the directed acyclic graph G (i.e. the regulators of each transcript) that "best" describes the gene expression data D, where D is assumed to be a steady-state data set. The model-learning algorithm usually assumes a specific form of the conditional probability function.

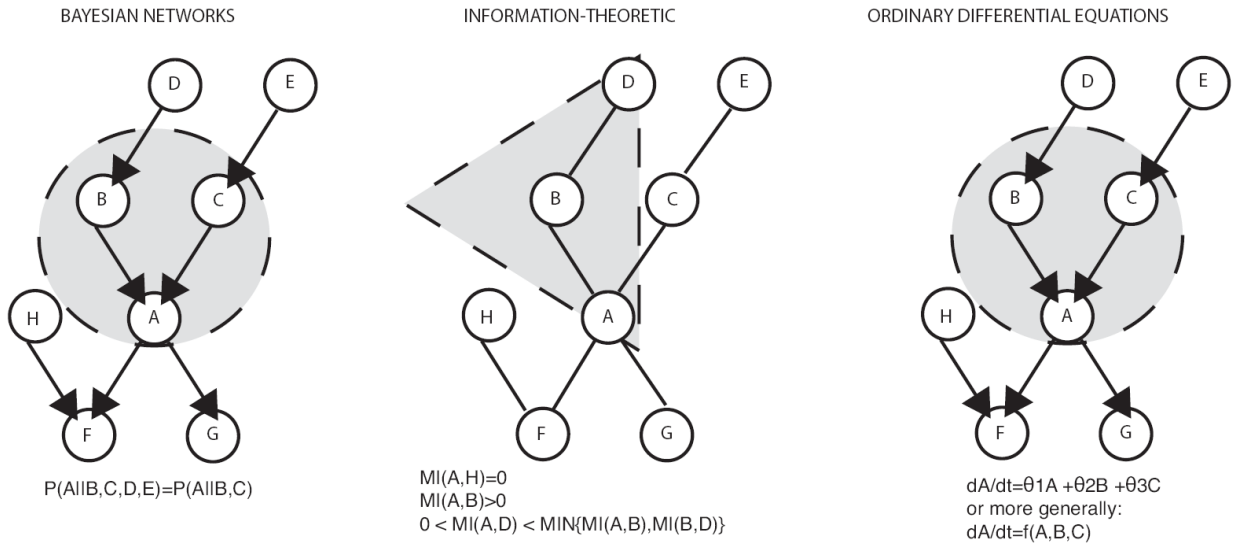


Figure 7 - Systematic overview of the theory underlying different models for inferring gene regulatory networks. Bayesian networks: A is conditionally independent from D and E given B and C; Information-Theoretic networks: Mutual information is 0 for statistically independent variables, and Data Processing Inequality helps pruning the network; Ordinary Differential Equations: Deterministic approach where the rate of transcription of gene A is a function (f) of the level of its direct causal regulators. (Figure taken from Bansal et al. "*How to infer gene networks from expression profiles*" [16])

The network structure is usually determined using a heuristic search due the NP-hard complexity of the problem. Many heuristics can be used, including greedy-hill climbing approach, Markov Chain Monte Carlo method or simulated annealing. For each network structure G visited in the search, the algorithm learns the maximum likelihood parameters for the conditional probability functions. It then computes a score that evaluates each graph G (i.e. a possible network topology) with respect to the gene expression data D. The score can be defined using Bayes rule:

$$P(G|D) = \frac{P(D|G)P(G)}{P(D)} \qquad (2.4)$$

where $P(G)$ can either contain some a priori knowledge on network structure, if available, or can be a constant non-informative prior, and $P(D \mid G)$ is a function, to be chosen by the algorithm, that evaluates the probability that the data D has been generated by the graph G. The most popular scores are the Bayesian Information Criteria (BIC) and the Bayesian Dirichlet equivalence (BDe). Both scores incorporate a penalty for complexity to guard against over fitting of data.

In Bayesian networks, the learning problem is usually underdetermined and several high scoring networks are found. A possible solution consists to use model averaging or bootstrapping to

select the most probable regulatory connection and to obtain confidence estimates for the connection.

But the main limitation of Bayesian networks remains that they assume the acyclic structure of the network (i.e. no feedback loops). Dynamic Bayesian networks [34-36] overcome this limitation and can be used to infer cyclic phenomena such as feedback loops that are prevalent in biological systems and they are also able to infer interactions from a time-series dataset.

2.2.2 Associative Networks

Associative networks are used to represent pairs of transcripts that coherently change their expression levels across a set of different conditions (i.e. co-expressed genes). Association networks connect transcripts that exhibit high statistical dependence by observing changes in their responses across all the experiments in the dataset. As a measure of similarity the pair-wise correlation (Pearson or Spearman) is often used. The Person correlation coefficient is computed as:

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{k=1}^{n}(y_i - \bar{y})^2}} \qquad (2.5)$$

The Spearman correlation coefficient between pair of transcript X and Y is computed instead as:

$$r_{XY} = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \qquad (2.6)$$

where $d_i = x_i - y_i$ is the difference between the ranks of each observation on the two variables are calculated.

Since many false positive connections are usually identified using these techniques, a pruning process can be undertaken to remove connections that are better explained by a more direct path through the graph. In [35] de la Fuente et al. proposed the application of partial correlation to prune the network and to remove redundant connections. Briefly, partial correlation measures the correlation between two variables after subtracting the correlation between each variable and one or more additional variables.

Correlation based methods are able to identify only linear dependencies between two variables. Mutual information instead makes no assumptions about the form of the dependence and in particular it is able to discover also nonlinear relationships among variables. Regarding association networks based on *Mutual Information*, let us consider a variable X with C possible states, $x_1 \dots x_C$ each with its corresponding probability $p(x_i)$. The average amount of information gained from a measurement that specifies one particular value x_i is given by entropy $H(X)$ and is computed as:

$$H(X) = -\sum_{i=1}^{C} p(x_i) \log\big(p(x_i)\big) \qquad (2.7)$$

The entropy $H(X)$ has the following four properties:

1. If an outcome of the measurement is completely determined by x_i i.e. the probability $p(x_i)$ is one and all other probabilities $p(x_j)$ with $i \neq j$ are zero, then $H(X) = 0$.
2. For equiprobable events the entropy $H(X)$ is maximum and is given by:

$$p(x_i) = \frac{1}{C} \rightarrow H(X) = \log(C) \qquad (2.8)$$

3. Entropy remains unchanged when impossible events are added.
4. If the logarithm to base C is used, the entropy is normalized (i.e. $0 \leq H(X) \leq 1$)

The joint entropy $H(X,Y)$ of two discrete variables X and , with Y assuming values in the set $\{y_1 \dots y_C\}$, is given by:

$$H(X,Y) = -\sum_{i=1}^{C}\sum_{j=1}^{C} p\big(x_i, y_j\big) \log p\big(x_i, y_j\big) \qquad (2.9)$$

$p\big(x_i, y_j\big)$ denotes the joint probability that X is in state x_i and Y in state y_j. If the systems X and Y are statistically independent the joint probability factorize and the joint entropy $H(X,Y)$ becomes:

$$H(X,Y) = H(X) + H(Y) \qquad (2.10)$$

The mutual information $MI(X,Y)$ or MI_{XY} between variables X and Y is defined as:

$$MI(X,Y) = H(X) + H(Y) - H(X,Y) \geq 0 \qquad (2.11)$$

There exist two main strategies to estimate mutual information: (*i*) histogram techniques and (*ii*) kernel density approach. There are many algorithms that have successfully applied the association network based on MI [37, 38] and shown its application in biological systems.

The definition of MI requires each data point (i.e. each experiment) to be statistically independent from the others. Thus information-theoretic approaches, as described here, can deal with steady-state gene expression dataset, or with time-series data as long as the sampling time is long enough to assume that each point is independent.

Edges in networks derived by information-theoretic approaches represent statistical dependences among gene expression profiles. As in the case of Bayesian network, the edge does not represent a direct causal interaction between two genes, but only a statistical dependency. Theoretically, the main difference between MI and Pearson correlation coefficient is that MI can quantify also nonlinear dependencies between variables meaning that zero value of correlation cannot imply that two variables are statistically independent.

2.2.3 Ordinary differential equations (ODEs)

Reverse-engineering algorithms based on *Ordinary Differential Equations* (ODEs) [22, 39-41] relate changes in gene transcript concentration in a system subject to an external perturbation. Where by external perturbation I mean an experimental treatment (i.e. small molecules inducing overexpression or down regulation of other genes) that can alter the transcription rate of some of the genes in the cell. ODEs are a deterministic approach unlike Bayesian networks and association network approaches. A set of ordinary differential equations, one for each gene, describes the gene regulation as a function of other genes:

$$\dot{x}_i(t) = f_i(x_1, \ldots, x_N, u, \boldsymbol{\theta_i}) \qquad (2.12)$$

where $\boldsymbol{\theta_i}$ is a set of parameters describing interactions among genes (the edges of the graph), $i = 1, \ldots, N$ and $x_i(t)$ is the concentration of transcript i measured at time t, $\dot{x}_i(t)$ is the rate of transcription of transcript i, N is the number of genes, and u is an external perturbation to the system. Since ODEs are deterministic, the interactions among genes ($\boldsymbol{\theta_i}$) represent causal interactions, and not statistical dependencies as the other methods.

The functional form of the influence functions f_i can be either linear of nonlinear, nonlinear functions can lead to an exponential rise in the unknown parameters to be estimated. Researchers have studied various functions, including sigmoidal functions [42], linear [22, 43, 44] and non-linear [45] functions.

Reverse-engineer a network using ODEs means to choose a functional form for f and then to estimate the unknown parameters $\boldsymbol{\theta_i}$ for each i from the gene expression data D using some optimization technique.

The easiest form that this function can assume is the linear form where Equation 2.12 becomes:

$$\dot{x}_i(t) = \sum_j \omega_{ij} x_j + p_i \qquad (2.13)$$

where ω_{ij} represents the influence of transcript j on transcript i and p_i is an externally applied perturbation to the level of transcript i. Linear functions have proven to be the most versatile in the analysis of experimental data sets [22, 43]. In part, this is due to the simplifying power of linear functions; they dramatically reduce the number of parameters needed to describe the influence function and avoid problems with overfitting. Thus, the amount of data required to solve a linear model is much less than that required by more complex nonlinear models a crucial advantage since the high cost of experimental data and the high dimensionality of the systems. On the other hand, linear functions do not show a rich variety of dynamic behaviour. They only have one isolated stationary state in which the temporal change of transcript vanishes, once reaching this state the concentrations of the net-work components remain constant. Furthermore, a linear model places strong constraint on the nature of regulatory interactions in the cell. Therefore, oscillations or multistationarity, which are both important properties of true biological networks, and are nonlinear phenomena, cannot be captured with linear models. Also, higher noise in the microarray data limits their application to make only qualitative statement and not quantitative statement about the underlying network.

ODE-based approaches can be applied to both steady state and time-series expression profiles. Advantage of ODE approaches is that once the parameters, $\boldsymbol{\theta_i}$ for all i are known, Equations 2.12 and 2.13 can be used to make predictions on the behavior of the network to different conditions (i.e. gene knock-out, treatment with an external agent, etc.) [46].

2.2.4 Examples of reverse-engineering application

There are many examples of successful application of reverse-engineering methods applied to mammalian cells. In this paragraph I decided to show two examples using association networks based on MI [37, 38]. In the 2005 Basso et al. [37] described the application of *ARACNe* (Algorithm for the Reconstruction of Accurate Cellular Networks) on 7907 genes in Human B cells. In ARACNe the MI between each pair of genes was estimated using a kernel density approach. The algorithm builds an initial graph by connecting all the transcript pairs with a mutual information value associated to a p-value higher than a chosen significance threshold computed using Monte Carlo simulation. Final pruning of the network is achieved by application of the *Data Processing Inequality* (DPI) principle. DPI asserts that if both $(x; y)$ and $(y; z)$ are directly interacting while $(x; z)$ is indirectly interacting through y, then $MI(x; z) \leq MI(x, y)$ and $MI(x; z) \leq MI(y; z)$. Using this method they were able to reconstruct the entire regulatory network of human B cells. The topology of this network suggested a hierarchical and scale-free network, with few highly interconnected genes (hubs) accounting for most of the connections. They found that MYC was one of the major hubs in this network and they experimentally verified some of the new inferred targets, showing that this approach can be generally useful for the analysis of normal and pathologic networks in mammalian cells.

In a more recent work, Belcastro et al. [38] collected a massive and heterogeneous dataset of 20,255 GEPs from ArrayExpress containing a large variety of human samples and experimental conditions. They also collected 8,895 GEPs from mouse samples. In this work, they developed a novel mutual information reverse-engineering approach [47]. Using this novel method that takes into account this large amount of data, they were able to reconstruct a gene network for human and mouse organisms. The inferred connections were compared against known interactions to assess their biological significance and successfully experimentally validated a subset of not previously described protein–protein interactions. They also showed the existence of co-expressed modules within the networks, consisting of genes strongly connected to each other, which carry out specific biological

functions, and tend to be in physical proximity at the chromatin level in the nucleus. Finally they showed that the network can be used to predict the biological function and subcellular localization of a protein, and to elucidate the function of a disease gene. As a case of study they experimentally verified that granulin precursor (GRN) gene, whose mutations cause frontotemporal lobar degeneration, is involved in lysosome function.

These two studies represent only a small example of the utility reverse-engineering methods and many other works in literature demonstrate that this approach can successfully be used in biology to elucidate the complexity of regulatory networks existing in a cell.

2.3 Differential networks: methods and applications

State-of-the-art reverse-engineering methods model gene networks as static processes, i.e. regulatory interactions among genes in the network (such as direct physical interactions or indirect functional interactions) do not change across different conditions or tissue types. However, different cell-types, or the same cell-type but in different conditions, may carry out very different functions, and it is expected that their regulatory networks reflect these differences. Several methods have been proposed to identify active sub-networks across different conditions from changes in gene expression. Looking genes at their functional level comparing the topological structure of gene regulatory networks across different conditions provides a more informative level to study genes alteration and their role.

One of the first attempts was a general method to search for “active sub-networks” connecting genes with unexpectedly high levels of Differential Expression (DE) [48]. This method requires in input a single network and it identifies a set of genes (i.e. sub-network) whose expression changes across two conditions. However, changes in expression may be very mild or absent, even when the sub-network is active. Hence, looking only at the differential expression levels of genes could be not sufficient.

Therefore, more recent approaches attempted to identify Differential Co-regulation (DC) of genes in the sub-network [49-57]. By differentially co-regulated genes we mean set of genes which are co-expressed only in a specific condition but not in others [58, 59]. Some of these methods works gene-to-gene and try to identify single genes that exhibit a differential co-expressed pattern across conditions. Many studies have shown that observing the strongly altered genes in connectivity could be sufficient to identify genes that play an important role in a disease phenotype [49-51].

Some of the most advanced methods instead, go beyond pair-wise co-regulation, and aim at automatically identifying denovo sub-network(s) containing genes whose co-regulation changes the most across two or more conditions [55-57]. For example Watson et al. [53] proposed a method called CoXpress in which clusters across two gene regulatory networks that have significant average correlation in the first condition and not in the other are considered differentially co-expressed. This approach has been recently extended by Choi et al. [54] using a new pair-wise measure alble to take into account also change in the sign of the correlation.

Other proposed method are more complex and use advanced optimization techniques such as genetic algorithms, which, however, are computationally intensive [52], since they require checking all of the possible sub-networks to identify the ones that are most dysregulated. Hence, these methods are limited in the number of different conditions that can be compared [53-55] and they may require fine-tuning of the algorithm parameters [57].

Recently Langfelder et al. [55] have published a very interesting work to study the network module preservation across multiple tissues or conditions. In this work the authors are able to measure the conservation of network modules across a set of condition-specific networks. Their method combine a permutation test and topological score that aggregate different indexes comprehending also the connectivity (or degree) of the genes.

Other studies have shown how is possible to find who are the genes that have the major contribution in the change of topology between two networks. In this area, recently, Odibat et al. [56] developed an algorithm able to optimizes an interactive objective function that is a linear combination of differential connectivity and differential betweenness centrality.

The main differences among all of these approaches are in how the genes to be tested are selected, how co-regulation is measured (i.e. Pearson Correlation Coefficient or Mutual Information), and how differential co-regulation across the conditions is quantified.

2.4 Reverse-engineering post-transcriptional regulatory interactions

Post-translational modification is a chemical mechanism in which amino-acid residues in a protein are covalently modified “on the fly” (Chapter 1). Through this mechanism, a cell is able to tightly regulate its activity regulating the localization and interaction of many molecules. Despite the number of human kinases that have already been identified, our understanding of phosphorylation-dependent signalling

networks is still very fragmentary. Considering that only about 500 human protein kinases are the actors of all the known signaling networks, only for a third of phosphorylation-sites identified so far, the kinases/phosphates is known. [60]

Due to the importance of the post-translational events, many researches were motivated to map phosphorylation networks. This has lead to the development of new computational methods to predict the substrate specificities of protein kinases. Some of these methods use a "sequence" approach. They are essentially based on experimental identification of the consensus sequence motifs recognized by the active site of kinase catalytic domains [61-63]. However, these motifs often lack sufficient information to uniquely identify the physiological substrates of specific kinases.

Other methods instead take into account also other kind of data [64-66] including gene expression and gene co-regulation. Here, in particular I will briefly discuss two of these methods. One has been recently proposes by Linding et al. [64], who developed a novel integrative computational approach, called NetworKIN. This method is designed to link experimentally identified phosphorylation sites to protein kinases. To predict protein kinases substrate, NetworKIN method combines consensus information from motifs with protein association networks. Basically, the method consists of two steps. In the first step, it uses neural networks and position-specific scoring matrices to assign each phosphorylation site to one or more kinase families, based on the intrinsic preference of kinases for consensus substrate motifs. In the second step, the context for each substrate is restricted using a probabilistic framework extracted from the STRING database [67], which integrates information from different kind of sources including: curated pathway databases, co-occurrence in abstracts, physical protein interaction assays, mRNA expression studies, and genomic context. They successfully applied their method to DNA damage signaling, showing that 53BP1 and Rad50 are phosphorylated by CDK1 and ATM, respectively.

In another recent work, Califano et al. [65] proposed a scheme where the ability of a transcription factor TF to regulate a target gene G is modulated by a third gene M (the modulator). Pairwise analysis of mRNA expression profiles will generally fail to reveal this complex picture because M and TF (e.g., a kinase and a transcription factor it activates/deactives) are generally statistically independent, and because the correlation between the expression of TF and G is averaged over an entire range of values of M and thus significantly reduced. The authors proposed a novel method called MINDy (Modulator Inference by Network Dynamics) that use conditional Mutual Information $I(TF, G|M)$ from gene expression profiles to detect such regulation, by conditioning on the expression level of the modulator. Hence, MINDy is a gene expression profile method based on a pair-wise

information-theoretic measure known as the conditional mutual information, for the identification of genes that modulate the transcriptional program of a transcription factor at the post-translational level. Since accurate estimation of the conditional mutual information requires large datasets MINDy needs a large gene expression profile dataset to work. The authors used MINDy to dissect the post-translational regulation of MYC activity in human B lymphocytes. Using this approach they were able to infer novel post-translational modulators of MYC [65] showing that also with gene expression it is possible infer post-transcriptional regulation.

Chapter 3

Reverse Engineering Tissue-Specific Transcriptional Networks

Due to the large diffusion of microarray technologies in the last decade, public repositories have been released to store and annotate these data for the scientific community. A standard called MIAME has also been proposed [25] to consistently annotate these gene expression data from microarray. In this chapter, I will present a software pipeline able to retrieve and classify in a semi-automatic way microarray experiments containing feature of interest. Using this pipeline, I was able to collect 2390 microarray experiments divided in 30 tissues. Finally the reverse-engineering technique used for the identification of the 30 tissue specific networks and their validation will be presented. Part of this Chapter has been published in the work of Gambardella et al. "Differential Network Analysis for the identification of condition-specific pathway activity and regulation" *in press* in the journal Bioinformatics [68].

3.1 Construction of a semantic database for tissue-specific gene expression profiles

One of the main problems in using Gene Expression Profiles (GEPs) from public repositories is the poor state of the experiments meta-data containing information on the biological samples and experimental protocols. In order to select tissue-specific GEPs, I built a semi-automatic tool to download and classify the MAGE-ML [25] annotation files present in the ArrayExpres [24]. This tool includes a *semantic database* [68] which structure is built round the eVoc human tissue ontology [69]. The database allows storage and retrieval of the classified experiments. I was thus able to assign to each GEP the correct tissue, according to the available meta-data, and selected only GEPs with a reliable annotation. I was thus able to collect 2930 high-quality GEPs (Affymetrix HG-U113A and HG-U133 Plus 2.0 platforms) for 30 different tissues (Table 1).

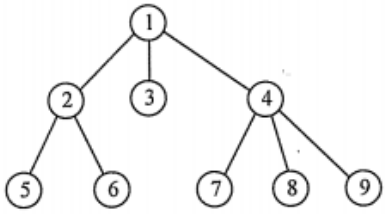
Table 1 – The Number of gene expression profiles collected for each tissue, using the semi-automatic tool to retrieve and classify Geps presents on ArrayExpress.

Tissues	# of Geps
ADIPOSE TISSUE	49
ADRENAL GLAND	88
BLOOD	268
BONE MARROW	180
BRAIN STEM	62
BRONCHUS	79
CARTILAGE	60
CEREBELLUM	49
CEREBRUM	109
COLON	66
DUODENUM	61
HEART	124
INTESTINE	108
KIDNEY	206
LIVER	86
LUNG	108
LYMPH NODE	43
MAMMARY GLAND	264
MIDBRAIN	55
MUCOSA	61
OVARY	56
PANCREAS	51
PLACENTA	47
PROSTATE	102
SKELETAL MUSCLE	94
SKIN	76
TESTIS	61
THYROID	43
UMBELLICAN CORD	84
UTERUS	190

The main feature of this database is its "*semantic*" property, consisting in the fact that it integrates a tissue-based ontology, which can be used to classify and retrieve all the hybridizations regarding a tissue of interest. Formally, to retrieve all the hybridization of a tissue means to retrieve all the hybridizations

that have an IS-A relationship with the tissue of interest. In a Relational Database Management System (RDBMS) to manage in an efficient way this kind of recursive query, ad-hoc structure are necessary.

Here, I decided to use, and thus to implement, the solution proposed by Kimball et al. in [70] where they employed a *bridge table* to model a 1 to N association with the table containing the ontology terms. This strategy allows flattening the hierarchies present in the ontology.



NAV_I

kPadre	kFiglio	livello	foglia
1	1	0	FALSE
1	2	1	FALSE
1	3	1	TRUE
1	4	1	FALSE
1	5	2	TRUE
....	....	....	....
2	2	0	FALSE
2	5	1	TRUE
2	6	1	TRUE
5	5	0	TRUE
....	....	....	....

Figure 8 –Example of *bridge table* necessary to store the tree shown on the in left. The *bridge table* contains one row for each pathway in the tree, as well as a row for the zero-length pathway from a node to itself. Each row of the *bridge table* contains the node key of the parent and of its descendant, the number of levels between the parent and the descendant and finally, a flag to indicate that there are no further nodes above the parent, which indicates if this descendant is a leaf or not.

As shown in Figure 8 the *bridge table* contains one row for each pathway in the tree, from a node (i.e. ontology term) to every other node in the tree, as well as a row for the zero-length pathway from a node to itself. Each row of the *bridge table* contains the node key of the parent and of its descendant, the number of levels between the parent and the descendant and finally, a flag to indicate that there are no further nodes above the parent, which indicates if this descendant is a leaf or not.

Obviously the dimension of the *bridge table* grows exponentially with the depth of the three, but here, the ontology I used has maximum depth equal to 6 and thus only 1200 rows are necessary to store it, which is an easily manageable number by any RDBMS.

3.2 Spearman Correlation Coefficient

The Spearman Correlation Coefficient (SCC) is a nonparametric measure of statistical dependence between two variables. It measures the strength of a monotonic relationship between paired data. Usually, SCC is denoted by the Greek letter ρ or r_S. The Spearman correlation coefficient is defined in the same way as the Pearson Correlation Coefficient (PCC) but it uses the ranks of the variables rather than their values. Given two variables X and Y of size n, we first have to convert the n raw scores X_i, Y_i into the ranks, x_i, y_i and then we can compute r_S as:

$$r_S = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

If there are identical values (i.e. rank ties or value duplicates), we assign a rank equal to the average of their positions in the ascending order of the values. In the case of absence of identical values r_S can be easly computed as:

$$r_S = \frac{6 \sum_i d_i^2}{n(n^2 - 1)}$$

where the difference $d_i = x_i - y_i$ is the difference between the ranks of the observations for the two variables.

The sign of the Spearman correlation indicates the direction of association between the variable X and the variable Y. The SSC will be positive if and only if Y tends to increase when X increases, otherwise if Y tends to decrease when X increases the SSC will be negative. A SCC equal to zero means that there is no tendency for Y to either increase or decrease when X increases. The SCC becomes equal to 1 when X and Y are perfectly monotonically related. By definition a perfect monotone relationship between two variable X and Y implies that for any two pairs of data values X_i, Y_i and X_j, Y_j, the differences $X_i - X_j$ and $Y_i - Y_j$ always have the same sign for each i and j. In particular if the sign is positive we are talking about an increasing monotone relationship and of an decreasing monotone relationship if the sign is negative.

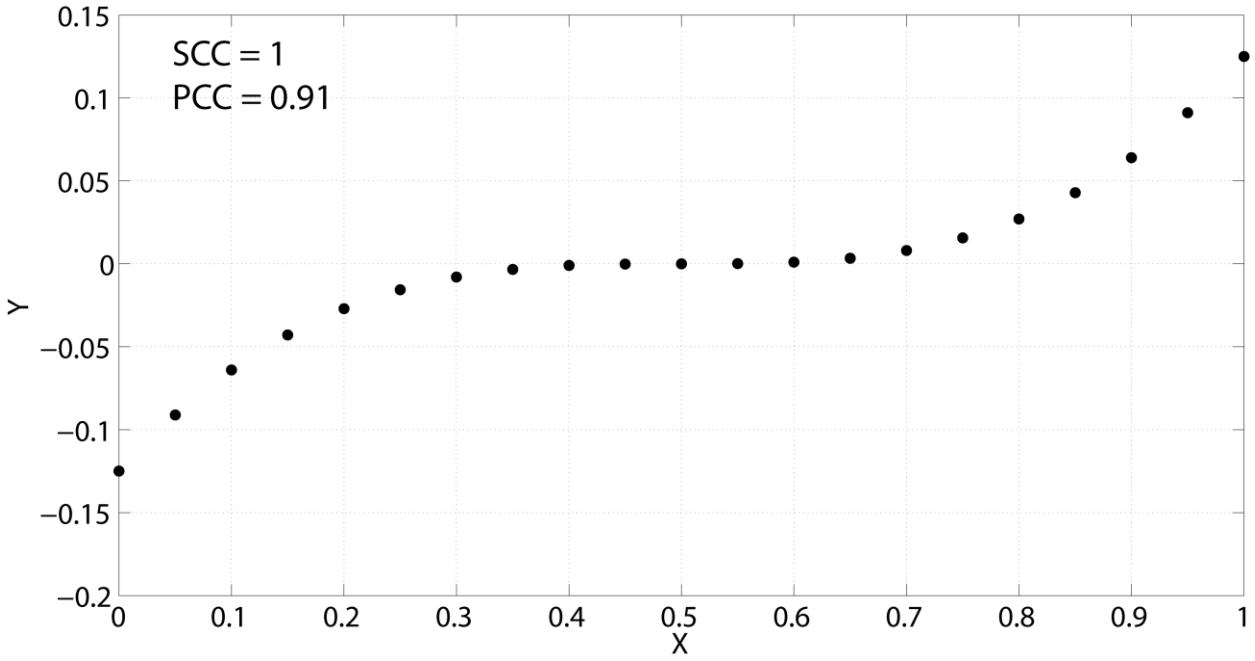


Figure 9 - Spearman correlation coefficient (SCC) between two variables is equal to 1 when they are monotonically related, even if their relationship is not linear, unlike the Person correlation coefficient (PCC).

The calculation of SCC and its significance test, unlike the Person Correlation Coefficient (PCC), do not require the normality distribution assumption of the data, but only the ordinal assumption. For this reason SCC is considered a nonparametric statistic.

The significance of SCC can be tested using the following formula [71]

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

which is distributed approximately as Student's t distribution with $n - 2$ degrees of freedom under the null hypothesis.

3.3 Reverse engineering of tissue-specific gene co-regulation networks

Using the database described in Section 3.1, I was able to classify 2930 microarrays (Affymetrix HG-U133A and HG-U133plus2) extracted from ArrayExpress in 30 different tissues [68]. To avoid batch effect, I decided to normalize microarrays independently for each tissue. The normalization algorithm used was RMA (Robust Multichip Average) as implemented in the R package Bioconductor [72]. Then I computed the Spearman Correlation Coefficient, as described in Section 3.2, for each pair of probes in each tissue, obtaining (excluding control probes) a final correlation matrix of dimension 22,215 x 22,215 for each tissue. The SCC significance for each pair of probes was computed fitting a Students' *t distribution* on the t-statistics obtained from all the SCC value.

Formally, in order to control the number of False Positives due to the multiple hypotheses test problem, I estimated the degrees of freedom of the *t distribution* from the data by fitting the parameters of a Student's t-location-scale distribution to the t statistics computed for all the probe pairs. I estimated the parameters by minimizing the squared error between the theoretical and the empirical distribution. In particular a Student's t-location-scale distribution of degree υ, mean μ and standard deviation σ has the density function as follow:

$$\frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\sigma\sqrt{\upsilon\pi}Gamma(\frac{\upsilon}{2})}\left[\frac{\upsilon+\left(\frac{x-\mu}{\sigma}\right)^2}{\upsilon}\right]^{-\left(\frac{\upsilon+1}{2}\right)}$$

with location parameter μ, scale parameter σ and shape parameter $\upsilon > 0$ and if x has a t-location-scale distribution with parameters μ, σ and υ, by definition $\frac{x-\mu}{\sigma}$ has a Student's t distribution with υ degrees of freedom.

Finally, to obtain the gene-wise SCC matrices, I used the following strategy: starting from the probe-wise SCC matrices, I first excluded probes that were associated to more than one gene using the tool proposed here [73], but still keeping genes associated to more than one probe. Specifically, using this tool to annotate probes of Affymetrix platforms [73], I was able to map 12161 genes from the probes in the HG-U133A platform. Out of these 12161 genes, 68% of the genes were associated to only one probe (Figure 10A) and only 11% of genes were associated to more than 2 probes (Figure 10B).

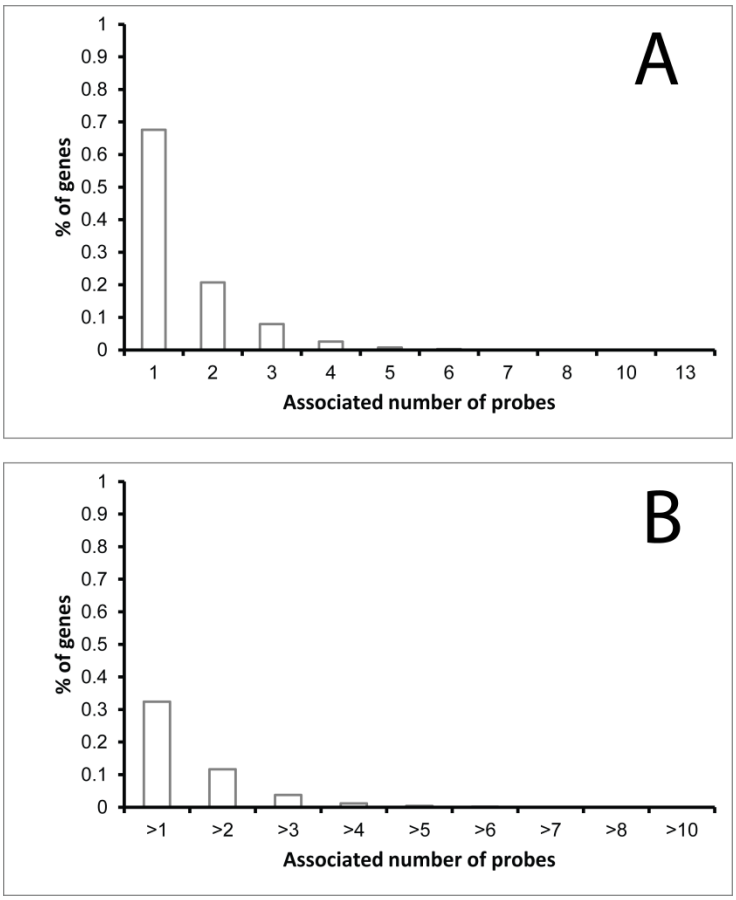


Figure 10 – Relationship between probes and genes on the HG-133A Affymentrix platform (**a**) Distribution of the number of probes associated to genes on the HG-U133A Affymetrix platform. x-axis: number of probes; y-axis: the pergentage of genes. (**b**) cumulative distribution of the genes versus associated number of probes. x-axis: the number of probes associated; y-axes: the pergentage of genes.

Hence, for the same pair of genes, there could be multiple values of the SCC, because the same gene might be associated to multiple probes in the microarray. In this case, I decided to assign to the gene pair, the "signed" maximal absolute value of SCC across all the different probe-pairs, deriving at the end of this procedure 30 gene-wise networks from the 30 probe-wise networks.

An alternative way to transform the probe-wise SCC matrices to gene-wise SCC matrices would have been to apply a "gene centered" normalization [74] of the microarrays using a custom CDF, prior to the SCC computation, thus eliminating the problem of multiple SCC values. We however decided to preserve information on possible alternative transcripts for future work and for experimental validation.

Finally, although Mutual Information (MI) has been shown to be a better alternative to correlation in identifying co-regulated genes [37, 38], because of the limited number of GEPs available for some tissues, I decided to use the Spearman Correlation Coefficient (SCC). I also decided not to use any network pruning techniques [35, 37] since I am not interested in distinguishing between direct and indirect interactions, but rather in how the co-regulation network among genes changes across the different tissues.

3.4 Validation and analysis of transcriptional networks

To verify the biological relevance of the tissue-specific gene networks, I computed the PPV-Sensitivity curve for each of the 30 co-regulation networks using as a "Golden Standard" the Ractome database [75]. Reactome is a database composed by manually curated interactions of genes and proteins participating to the same pathways.

For each network, I computed the percentage of co-regulated genes for which a regulatory interaction was confirmed by the Golden standard ($\text{Positive Predictive Value} = TP/(TP + FP)$). As shown in Figure 11, all of the 30 tissue specific co-regulatory networks have a Positive Predicted Value (PPV) significantly higher than what would be expected by chance [68].

Moreover, since each network was constructed using a different number of GEPs, I also verified that the different performance in PPV across the networks was not related to the number of GEPs used for the construction of each network. As shown in Figure 12 there is no correlation between the number of experiments used to build the network and its performance in PPV. PPV performance is measured in this case as Area Under the Curve (AUC) up to 1% of sensitivity.

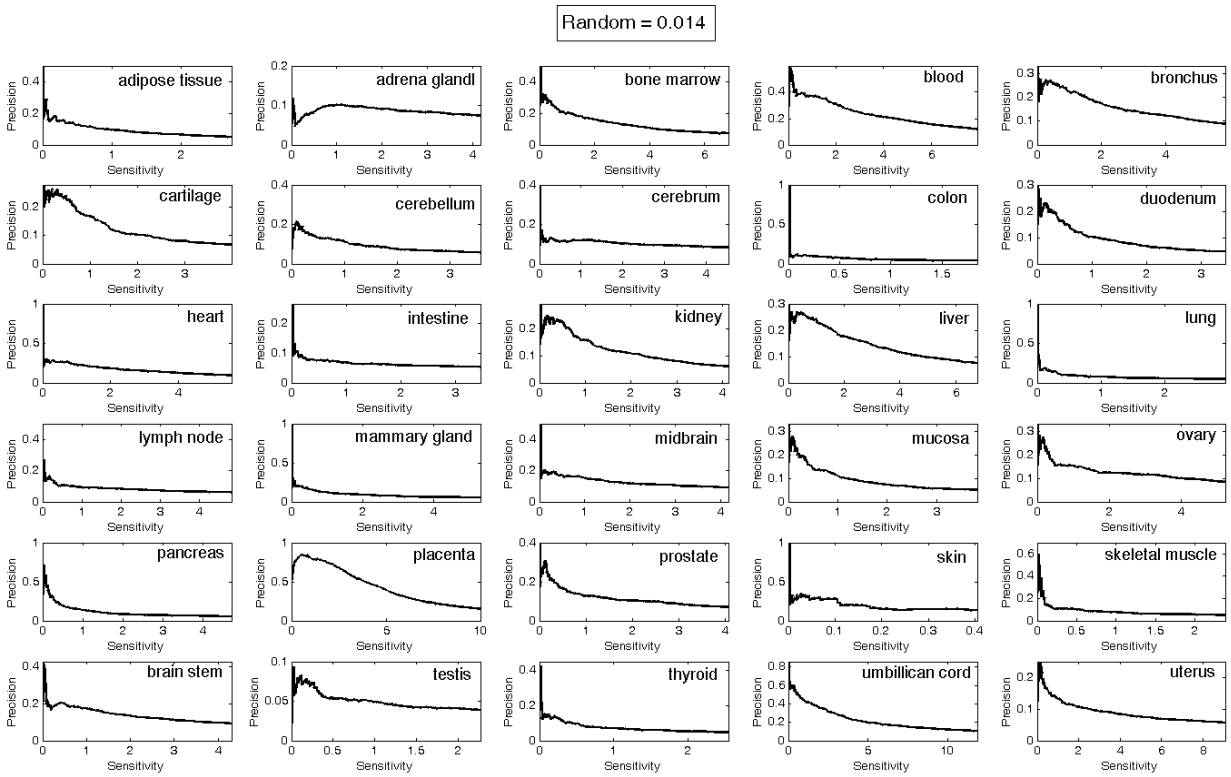


Figure 11 – Biological relevance of the 30 tissue specific co-regulation networks. I used as a Golden Standard, an interactome consisting of about 25,000 experimentally verified biological interactions from the Reactome database. The Positive Predictive Value ($\boldsymbol{PPV} = \boldsymbol{TP}/(\boldsymbol{TP} + \boldsymbol{FP})$) vs. Sensitivity ($\boldsymbol{TP}/(\boldsymbol{TP} + \boldsymbol{FN})$) curve for each of the 30 co-regulation networks is reported. The random performance is also shown for comparison (PPV Random = 0.014). (Figure taken from [68])

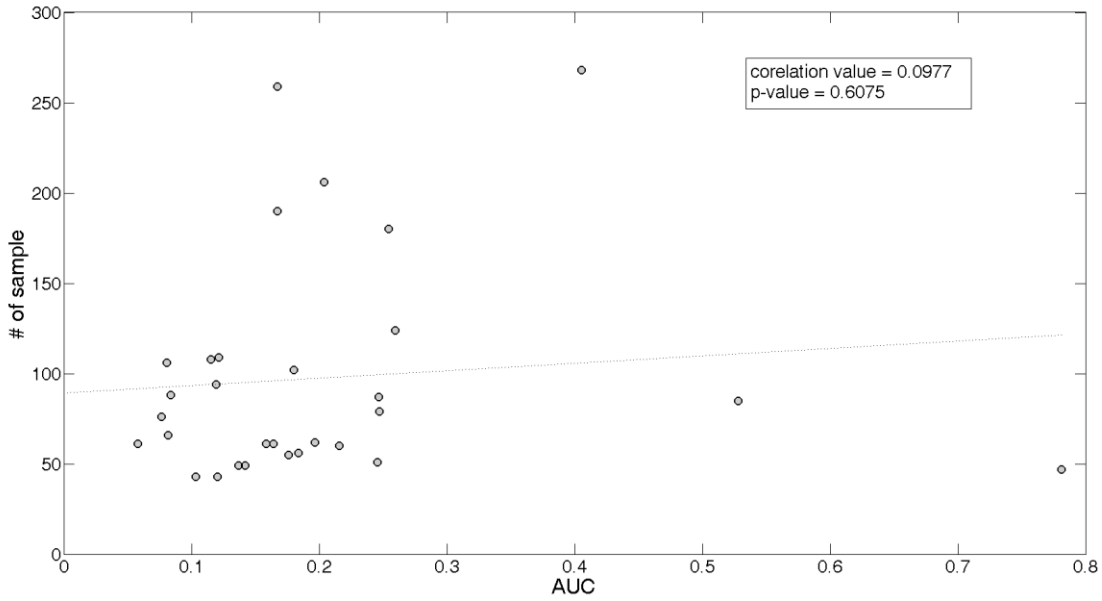


Figure 12 - Relationship between Area Under the Curve (AUC) and the number of GEPs in each tissue. Differences in performance (AUC) across the different networks is not due differences in the number of GEPs in each tissue. x-axis: AUC (up to a sensitivity of 1%) for each of the 30 networks. y-axis: number of GEPs in each tissue used to infer the network. The PCC (Pearson Correlation Coefficient) is 0.0977 with a p-value of 0.6075. (Figure taken from [68])

As a further proof of the biological relevance of the tissue specific-regulatory networks, I corroborated some recent results observed by analyzing tissue-specific Protein-Protein Interaction (PPI) networks by Lehner et al [76, 77]. In this work, the authors investigated human tissue-specific protein-protein interactions across 79 tissues, by assigning to each protein a tissue-specificity, if the corresponding coding genes was expressed in that tissue. They were able to show that tissue specific proteins (identified as the ones whose genes are expressed in one or few tissues) make few protein interactions, as compared to universally expressed proteins, and they tend to appear more recently during evolution.

Here, using a completely different approach based on tissue specific co-regulation networks inferred from tissue specific GEPs, I was able to draw similar conclusions: i.e. genes with tissue specific expression tend to be co-regulated with a smaller number of genes, as compared to ubiquitously expressed house-keeping genes (Figure 13A) and that evolutionary conservation is related with tissue-specific regulation (Figure 13B-C). In particular, in order to identify the specificity of a gene in a tissue, I used a set of GEPs measured across 79 human tissues [78] (i.e. the same set of GEPs used by Lehner). Specifically, in this dataset a gene is considered expressed in a tissue, if its normalized expression level is > 200 [78].

Finally, as a further proof of the biological relevance of the tissue-specific co-regulation networks, we identified which connections were conserved across the majority of the 30 networks. As shown in Figure 14, 3235 co-regulatory connections involving 993 distinct genes are conserved in at least half of the tissue-specific networks. In particular, these connections are enriched for “housekeeping” cellular functions such as ribosomal and cell cycle functions.

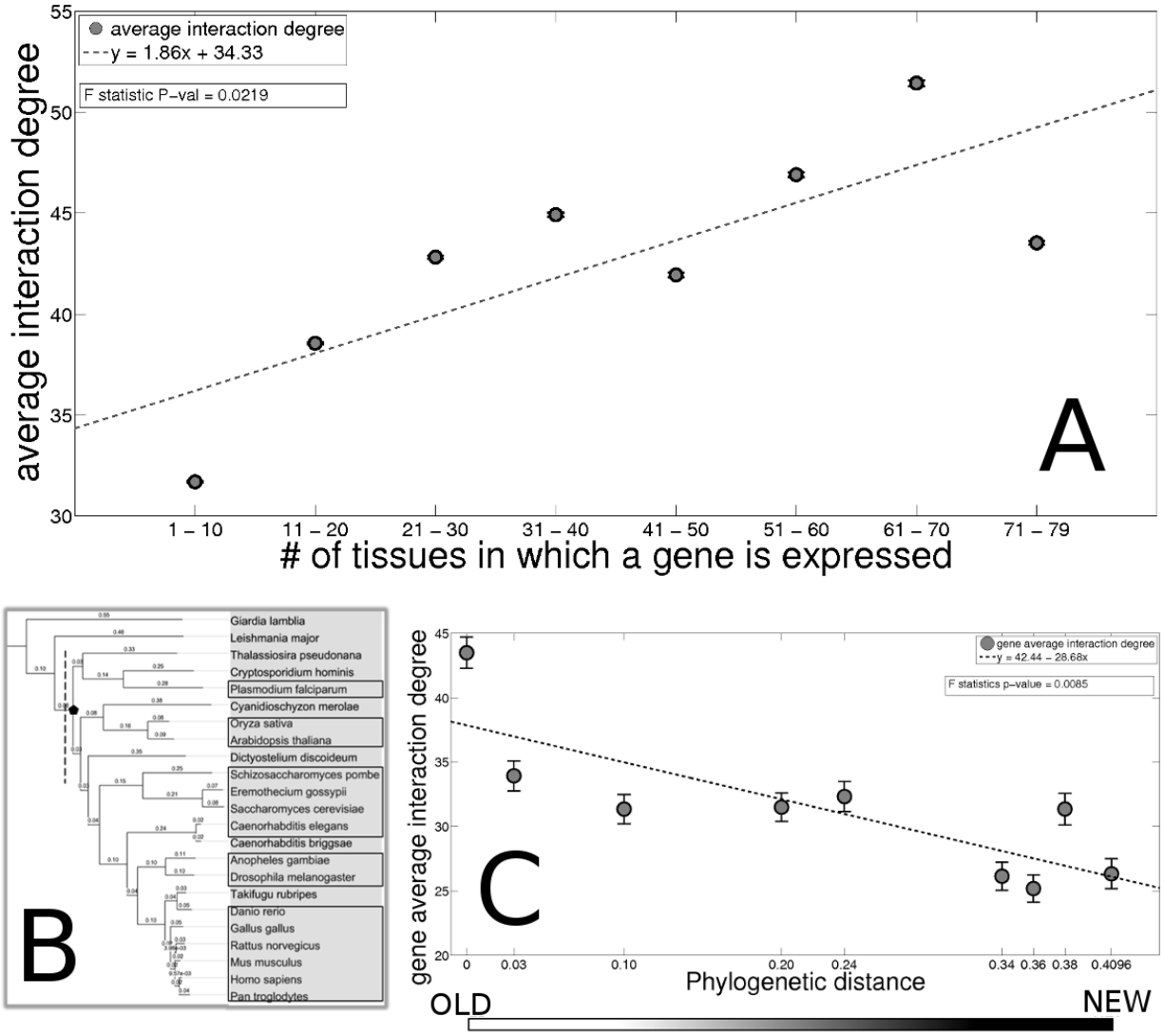


Figure 13 - Tissue specific genes and gene conservation as function of connection degree across the 30 co-expression networks. (**a**) Average connection degree of a gene across the 30 co-regulation networks as a function of the number of tissues in which it is expressed. x-axis: number of tissues in which a gene is expressed computed from the GENE ATLAS dataset [78]. y-axis: the average gene interaction degree across the 30 tissue specific co-regulation networks. The dashed line represents the linear regression with the corresponding p-value 0.037. (**b**) The phylogenetic tree, the pentagon marks the common ancestor of the 15 species (highlighted) used in the analysis. Numbers on each branch are the phylogenetic distances computed by Ciccarelli et al in [79]. (**c**) x-axis: phylogenetic distance of a gene computed as the distance between the root of the tree (pentagon in **a**) and the common ancestor of the species in which the gene is conserved. That is, the value 0 identifies genes conserved in all the 15 species, while the value 0.4096 identifies genes present only in human. y-axis: average gene interaction degree across the 30 co-regulation networks. The dashed line was obtained by linear regression shows the tendency of old genes to be more co-regulated compared with young genes (P-value = 0.0085).

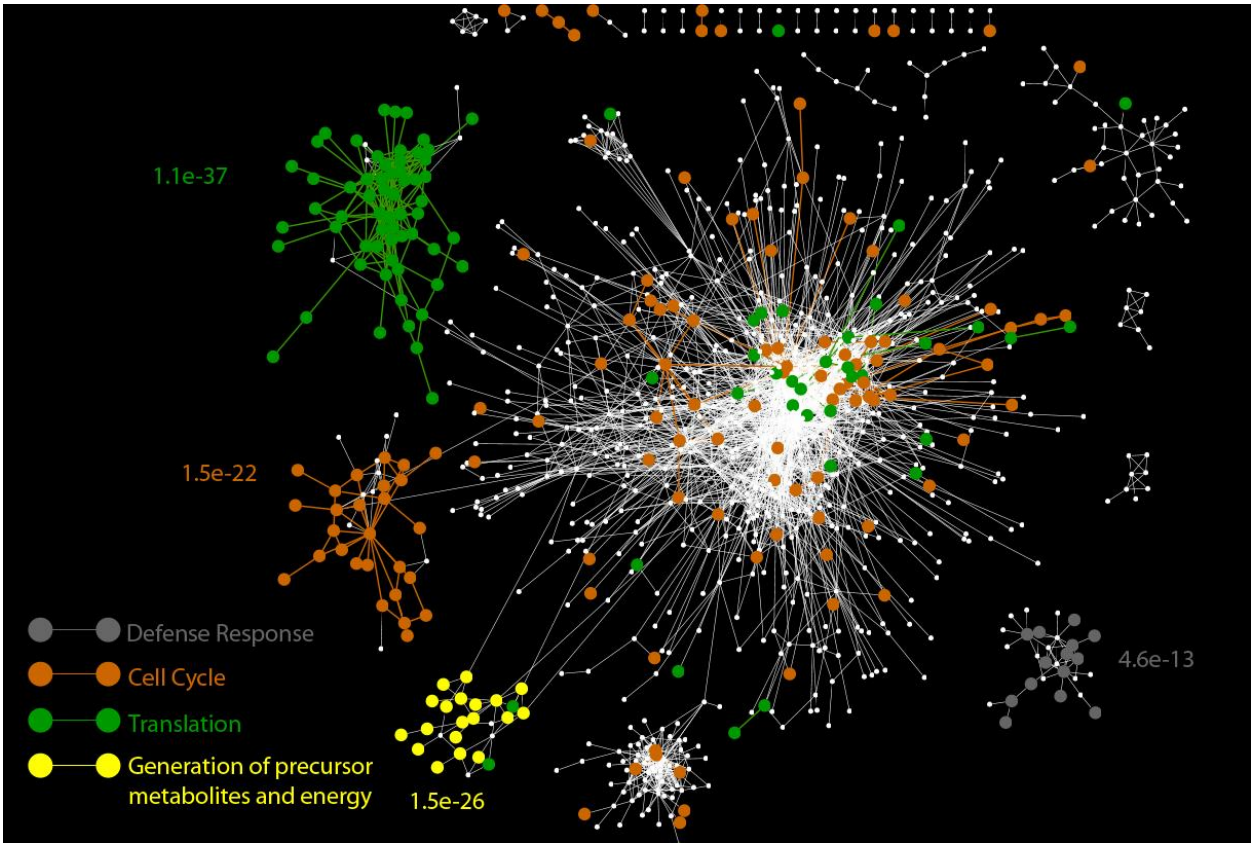


Figure 14 - Conserved connections across the majority of the 30 tissue specific co-regulatory networks. The graph represents the 3235 co-regulatory connections, involving 993 distinct genes, which are conserved in at least the half of the tissue specific co-regulation networks and their Gene Ontology Enrichment. (Figure taken from [68])

Chapter 4

A new approach to Differential Network Analysis

Identification of differential expressed genes has led to countless new discoveries. However, differentially expressed genes are only a proxy for finding condition-specific or dysregulated pathways. In this chapter, we will address the problem in which we want to identify how networks of regulatory and physical interactions rewire in different tissues or in during disease progression. I will first present a new method that I developed. I called this method DINA (DIfferential Network Analysis). This new procedure, starting from a collection of condition-specific gene expression profiles and a set of genes used as a query (i.e. a pathway), it is able to identify whether genes in the query set are co-regulated in a condition-specific manner. In this chapter, I will be also present an extension of DINA to predict which transcription factors may be responsible for the pathway condition-specific co-regulation. Part of the work here described is *in press* in the journal Bioinformatics [68].

4.1 A new approach to DIfferential Network Analysis (DINA)

The working hypothesis behind this new method I developed is that genes belonging to a condition-specific pathway are actively co-regulated only when the pathway is active, independently of their absolute level of expression. To this end, I developed a network-based algorithm, DINA (DIfferential Network Analysis), which is able to identify whether a set of genes (e.g. a pathway) is significantly co-regulated only in specific conditions, but not in others, as schematically described in Figure 15A.

The input to DINA is a set of M genes (i.e. genes belonging to a known pathway) and a set of N condition specific networks. DINA used the input and the networks to compute a "co-regulation probability" for the input M genes in each of the N networks; this probability is simply proportional to the number of edges among the genes in each network.

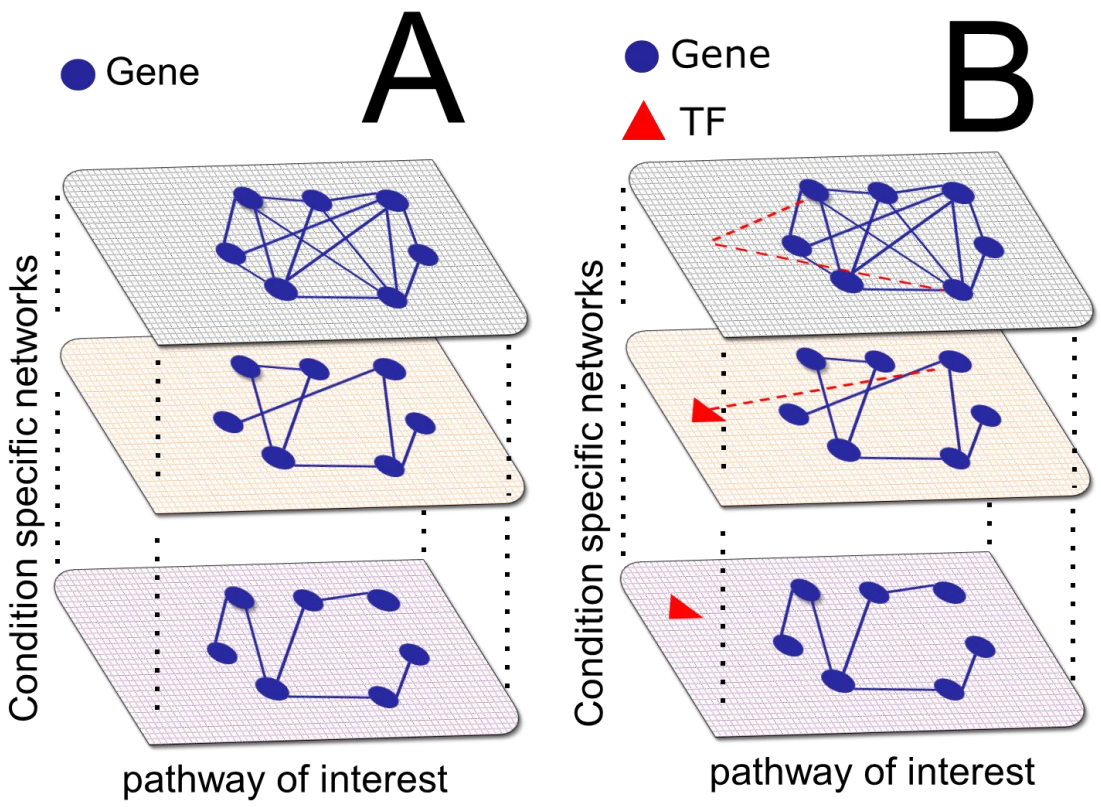


Figure 15 - Differential Network Analysis. (**a**) Graphical description of the DIfferential Network Analysis (DINA) method to quantify the variability of co-regulation among the genes in a pathway across multiple networks. (**b**) Graphical description of the method used to identify the transcriptional regulators of the genes in a pathway across multiple networks. (Figure taken from [68])

DINA [68] then quantifies how variable the co-regulation probability is across a set of N condition specific networks. The variability is quantified using an entropy-based measure (H) and its significance is estimated using a permutation test that I will describe in details in the following section.

In information theory, entropy is a measure of the uncertainty associated to a random variable. If V is a discrete random variable assuming N categorical values, its entropy $H(V)$ can be estimated as follows:

$$H(V) = \sum_{i}^{N} P(V = i) \log \frac{1}{P(V = i)}$$

From the property of entropy, H reaches its maximum value when each event is equi-probable and its minimum, i.e. $H(V) = 0$, when there is no uncertainty.

Specifically, in our settings, V assumes N categorical values, representing the N condition-specific networks. In order to compute $P(V = i)$ I first computed the number of edges n_i connecting the M genes in the i_{th} network (adding a pseudo-count of 1) with $i \in \{1, \dots, N\}$, and I then computed $P(V = i) = \frac{n_i}{\sum_{j=1}^{N} n_j}$.

$P(V = i)$ can be interpreted as a probability, because it is a number greater than 0 and it sums to 1 across all the N condition-specific networks by definition. Specifically, $P(V = i)$ will be equal to 1 if and only if the genes in the pathway are specifically co-regulated (i.e. connected) in network i and not co-regulated (connected) in any other network. In this case, H(V) will be equal to zero, indicating that the M genes are condition-specific. Thus $P(V = i)$ represents the probability that M genes in a pathway are co-regulated only in the i_{th} network and not in the other networks, and the value of $H(V)$ quantifies how condition specific the M genes are.

In order to assess the entropy significance (i.e. the condition specificity of a pathway) I used permutation test. The null distribution of $H(V)$ is approximated by selecting a set of N random networks, with the same density as the original networks, and a set of M random genes. Random networks are obtained from the original network by randomly shuffling the gene labels. This procedure is repeated 10,000 times in order to estimate the $H(V)$ p-value for the M genes in the query set. The p-value is estimated by counting how many times the $H(V)$ value computed on the query set is greater than the H(V) value computed on the random network. The p-value can then be corrected using the Benjamini-Hochberg method if multiple query set are given in input to DINA [80].

Summarizing, the key concept of the proposed method is this: if the M genes of interest have the same co-regulation probability across the N networks, then the entropy H will be high; on the other hand, if the M genes have a high co-regulation probability only in one (or few) networks (i.e. the pathway activity is condition-specific), than the entropy H will be low (hence we are interested in pathways associated to a low H).

4.2 Identification of transcriptional regulators of tissue-specific pathways

I then developed also a simple method for the identification of Transcription Factors regulating a condition-specific pathway [68] identified by DINA. The input to this algorithm is a pathway of interest (i.e. a set of M genes) and the output is a ranked list of Transcription Factors according to a p-value, estimated as described below. The algorithm iteratively tests all the transcription factors by computing for each one the number of edges connecting it to the genes in the pathway of interest in each of the condition specific networks, as depicted in Figure 15B. The method assigns a p-value to each tested TF using *the non-parametric Fisher's exact test*, by comparing, in each tissue, the number of edges between the TF and the genes in the pathway to the number of all the possible edges between the TF and the genes minus the real number of edges. The p-value was then corrected using the Benjamini-Hochberg method [80].

4.3 A case of study: Identification of tissue-specific pathways

To test whether DINA was able to identify tissue-specific pathways, i.e. pathways which are actively regulated only in specific tissues, I used the full manually curated list of 186 KEGG pathways from MsigDb [81, 82]. This list includes many different types of pathways including signaling, metabolic and regulatory pathways. By definition a pathway in KEGG is a set of genes known to work as a functional module in a regulatory network. From this list, I removed from the set of 186 KEGG pathways those pathways not well represented in our gene networks, described in Chapter 3, i.e. those pathways for which at least 80% of the genes had to be present obtaining a final list of 110 KEGG pathways to test.

By applying DINA to the tissue-specific networks and the list of 110 KEGG pathways, I obtained 22 significant tissue specific pathways (with corrected p-value<0.01) [68]. One of these is for example the Glycine, Serine and Threonine metabolic pathway (KEGG hsa00260). This pathway was correctly identified by DINA to be mainly regulated in liver and kidney, where most of the glycine to serine metabolism occurs [83]. Interestingly, among the 22 significant pathways, 9 are indeed metabolic pathways enriched in liver and kidney (Table 2 pathways in red).

Table 2 – The list of 22 significant tissue specific pathways identified by DINA with a p-value threshold of 0.01. Coloumn H contains the entropy value coputed by DINA.

KEGG PATHWAY	H	P-val(corrected)
KEGG_COMPLEMENT_AND_COAGULATION_CASCADES	3.72	0
KEGG_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION	3.85	0
KEGG_VALINE_LEUCINE_AND_ISOLEUCINE_DEGRADATION	3.90	0
KEGG_DNA_REPLICATION	3.91	0
KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION	4.09	0
KEGG_HEMATOPOIETIC_CELL_LINEAGE	4.24	0
KEGG_CALCIUM_SIGNALING_PATHWAY	4.25	0
KEGG_CELL_ADHESION_MOLECULES_CAMS	4.39	0
KEGG_JAK_STAT_SIGNALING_PATHWAY	4.41	0
KEGG_CHEMOKINE_SIGNALING_PATHWAY	4.46	0
KEGG_PROPANOATE_METABOLISM	4.11	0.00095
KEGG_GLYCINE_SERINE_AND_THREONINE_METABOLISM	4.12	0.00095
KEGG_BUTANOATE_METABOLISM	4.14	0.001477
KEGG_PEROXISOME	4.31	0.001477
KEGG_MISMATCH_REPAIR	4.15	0.001995
KEGG_ARGININE_AND_PROLINE_METABOLISM	4.18	0.001995
KEGG_FATTY_ACID_METABOLISM	4.16	0.002891
KEGG_LYSINE_DEGRADATION	4.23	0.002891
KEGG_PURINE_METABOLISM	4.53	0.002891
KEGG_OTHER_GLYCAN_DEGRADATION	4.41	0.00665
KEGG_FOCAL_ADHESION	4.60	0.006916
KEGG_TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY	4.47	0.008184

Figure 16A shows the *co-regulation probability* of the 32 genes in the Glycine, Serine and Threonine metabolic pathway in each of the thirty tissues, as previously defined and for comparison, Figure 16B shows the average expression level of the genes in the pathway in each of the 30 tissues. Notable is that expression levels do not change significantly across the tissues, whereas the co-regulation probabilities

(Figure 16A) are strikingly different, showing that to look only at the expression level, could be not sufficient to obtain the right answer.

To underline the fail of an approach only based on gene expression, we also checked in the Gene Atlas Dataset for the expression level of the genes encoding for the enzymes involved in this pathway [78], Using the canonical expression level threshold of 200 [78], we found that only 13 out of 32 genes are expressed in liver, and only 2 out of 32 are expressed in Kidney (Figure 17).

As shown in Figure 18 and Figure 19, similar considerations can be applied to the other significant metabolic pathways identified by DINA (Table 2 pathways in red). Hence, an approach based on expression levels (and not co-regulation) would not have been able to identify these tissue-specific metabolic pathways.

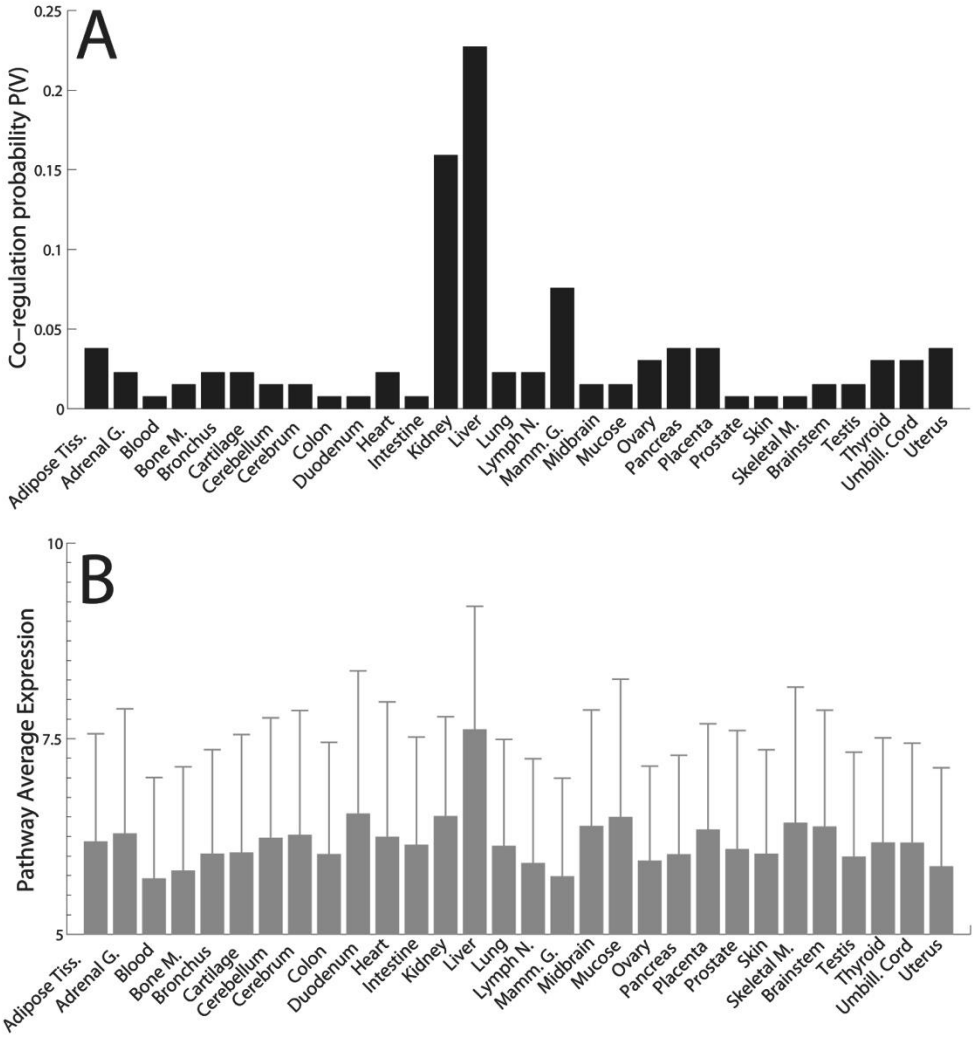


Figure 16 - Differential Network Analysis of the Glycine pathway (KEGG hsa00260). (**a**) Co-regulation probability of the 32 genes in the Glycine path-way (hsa00260) across the thirty tissues. (**b**) Average expression level of the 32 genes in the Glycine pathway (hsa00260) across the thirty tissues (error bars represent one standard deviation). (Figure taken from [68])

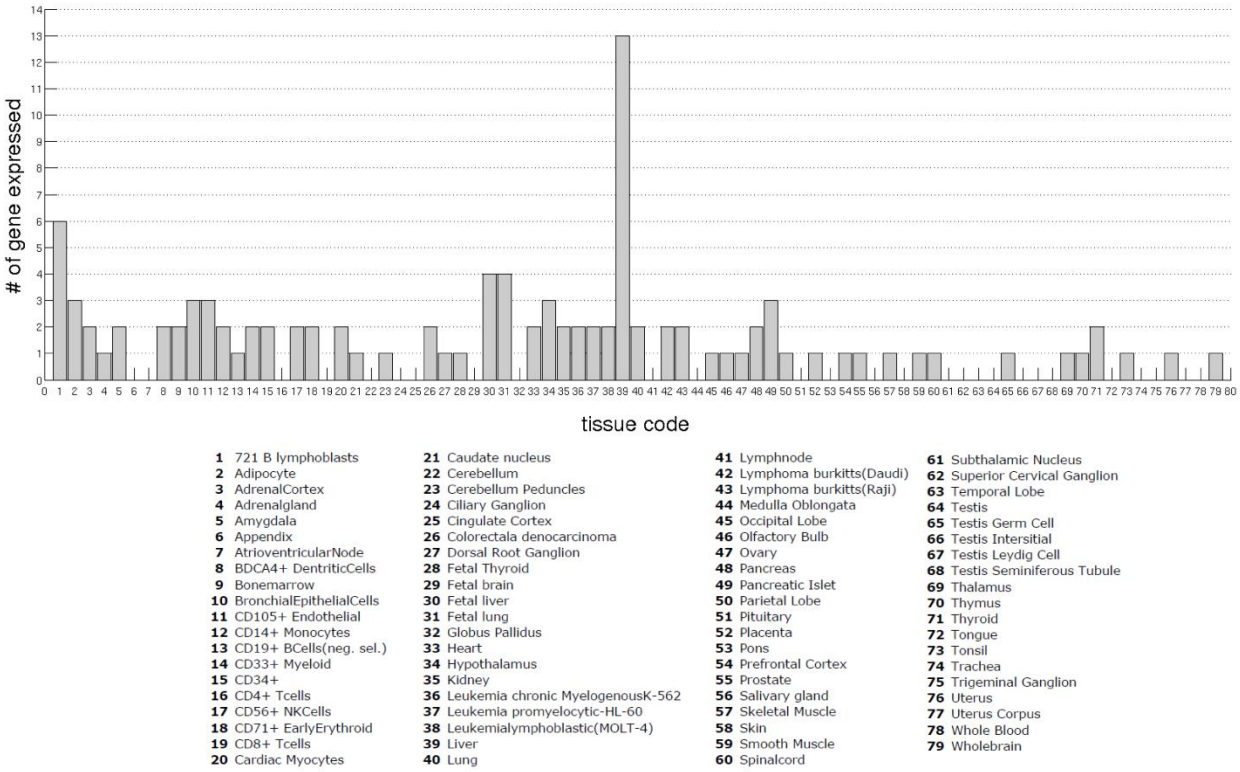


Figure 17 - The number of expressed genes that encode for the enzymes in the glycine pathway (KEGG hsa00260) across the 79 tissues of gene atlas dataset. Out of 32 genes only 13 are expressed in liver, 4 in fetal liver and only 2 in Kidney. (Figure taken from [68])

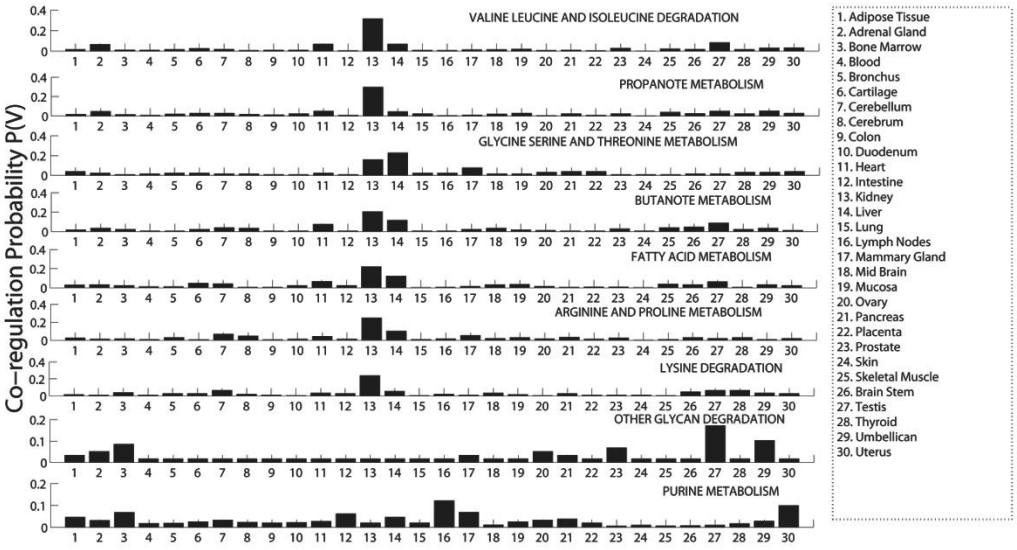


Figure 18 –Co-regulation probability among the genes that encode for the enzymes in the 9 significant metabolic pathways identified by DINA reported in Table 2. (Figure taken from [68])

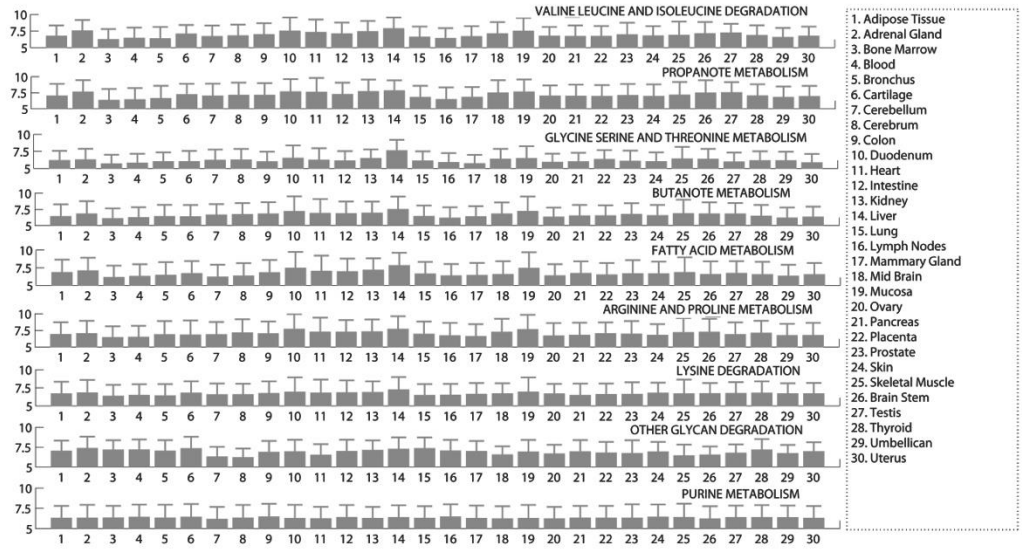


Figure 19 - Average expression among the genes that encode for the 9 significant metabolic pathways identified by DINA reported in Table 2. (Figure taken from [68])

4.4 A case of study: Identification of disease-specific pathways dysregulation

Gene expression alteration is a common molecular hallmark of cancer progression. In the last decade, the identification of cancer genetic signatures has been successfully exploited for understanding the mechanisms of cancer development [84], as well as, for anticancer therapies selection [85] and diseases prognosis [86]. Moreover, recently, with the help of reverse-engineering approaches also some specific cancer-regulated gene regulatory networks have been identified [87, 88]. Here, I wondered whether DINA could be successfully employed to identify selective alterations of co-regulated gene networks in cancer. Since several cell-lines modelling HCC progression are available, as well as, GEPs measured in these cell lines, as a study model, I focused my attention on Hepatocellular Carcinoma (HCC).

It is well established that HCC progression involves alterations in many fundamental signalling pathways, such as EGF-Ras-MAPK, AKT-mTOR, Jak-Stat, and NF-kB cascades [89]. In addition, inactivating mutations of the tumor suppressor p53, or p53 loss of expression, are among the most frequent genetic events associated with hepatocyte transformation [90, 91] and the dysregulation of p53-dependent genes have been observed in HCC [92, 93].

Here, in order to mimic the progression of HCC, I reverse engineering cell-type specific gene networks using the Spearman Correlation Coefficient by collecting a total of 161 GEPs for three human cell-lines: primary hepatocytes (40 GEPs), hepatoblastoma-derived cell line HepG2 (39 GEPs) and Hepatocarcinona-derived cell line (Huh7) (82 GEPs). From these GEPs I first reverse-engineered the three co-regulation networks for each one of the different cell-lines. Then, I tested the ability of DINA to identify differential co-regulation of p53-dependent genes across these three condition specific gene regulatory networks, since the Huh7 cell-line has an inactivating mutation for p53 [91].

To this purpose, I built a gene signature made up by 34 transcriptional targets of p53 [94] and I then applied DINA to this gene signature, as shown in Figure 20.

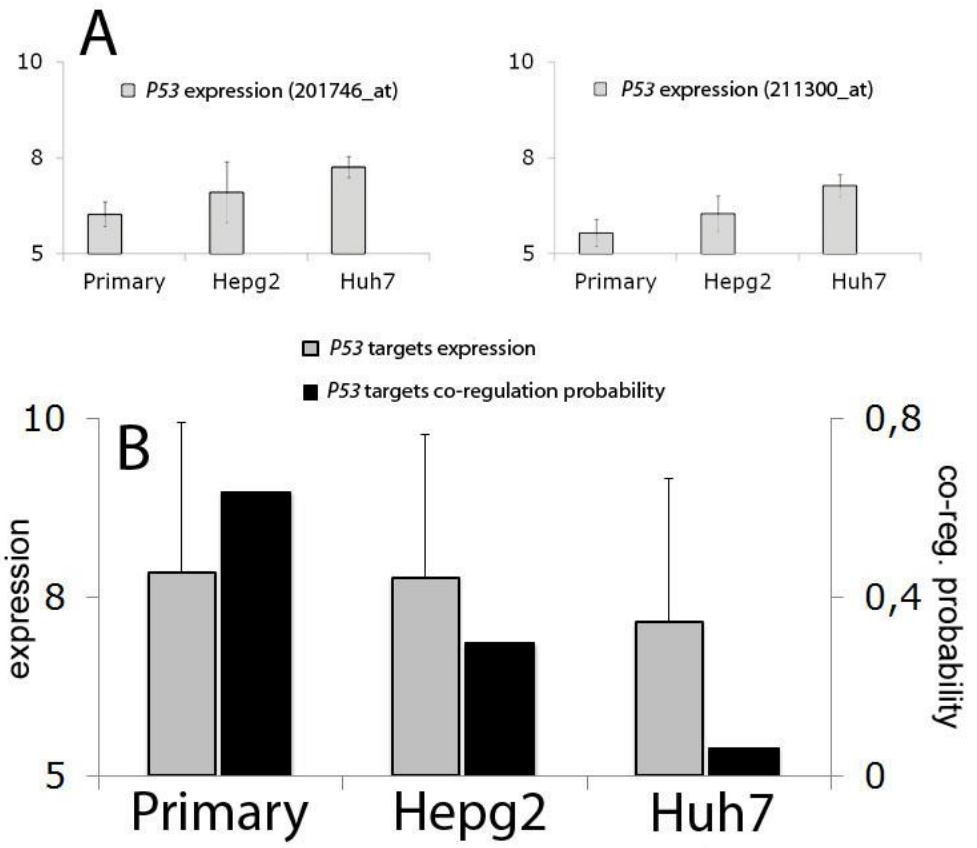


Figure 20 - Differential Network Analysis of the p53 gene signature in primary and transformed hepatocytes. The gene signature consists of 34 experimentally verified transcriptional targets of p53. (**a**) p53 expression level in the three cell-lines for the two probes present in Affy HG-U133A platform. (**b**) Comparison between the co-regulation probability of the genes in the signature (black) and their average expression level. (Figure taken from [68])

As shown in Figure 20, DINA successfully detected a differential co-regulation of the p53 target genes across the three cell lines: the co-regulation probability is high in normal hepatocytes and, to a lesser extent, in hepatocellular carcinoma HEPG2 cell line, carrying a wild type p53 protein, and decreases significantly in Huh7 cell line, carrying an inactive p53 protein [91] (Figure 20B). Interestingly, the expression level of the p53-target genes did not correlate with the functional status of the p53 protein in the different cell lines, thus supporting my previous observation (Figure 20A) that an expression-based method would be less powerful than the DINA in identifying dysregulated pathways.

I next applied DINA to identify dysregulated pathways during hepatocytes transformation. The DINA-based analysis of the 110 KEGG pathways identified at least five pathways whose co-regulation is significantly disrupted in the hepatocarcinoma cell lines compared to the normal hepatocytes (Table 3).

Similarly to the previous results, the average expression levels of the genes in these pathways did not change significantly between normal and transformed hepatocytes (Table 3).

Table 3 – List of pathways that DINA get dysregulated. The entropy value (H) are reported with their p-values (corrected and not). Red bold pathways are significantly disrupted pathway found by DINA. As a comparison also the average expression of the pathways is reported for each cell line.

PATHWAYS	H	P-val	P-val (Corected)	Avg. exp. primary hep.	Avg. exp. hepg2.	Avg. exp. huh7.
KEGG_PRIMARY_BILE_ACID_BIOSYNTHESIS	**0.5490**	**0.0000**	**0.0000**	7.36	6.50	6.37
KEGG_PEROXISOME	**0.5912**	**0.0004**	**0.0220**	8.21	7.60	7.48
KEGG_PHENYLALANINE_METABOLISM	**0.7229**	**0.0009**	**0.0330**	8.54	7.62	7.26
KEGG_GLYOXYLATE_AND_DICARBOXYLATE_METABOLISM	**0.8031**	**0.0014**	**0.0385**	8.58	8.20	8.16
KEGG_TYROSINE_METABOLISM	**0.6211**	**0.0019**	**0.0418**	7.57	7.03	6.59
KEGG_COMPLEMENT_AND_COAGULATION_CASCADES	0.8866	0.0064	0.1173	8.62	7.60	7.45
KEGG_VALINE_LEUCINE_AND_ISOLEUCINE_DEGRADATION	0.7811	0.0133	0.2063	8.40	7.84	8.04
KEGG_PPAR_SIGNALING_PATHWAY	0.9003	0.0150	0.2063	7.60	7.19	7.10
KEGG_TRYPTOPHAN_METABOLISM	0.8254	0.0172	0.2102	8.47	7.29	7.13
KEGG_FATTY_ACID_METABOLISM	0.8554	0.0201	0.2211	8.03	7.57	7.47
KEGG_ALANINE_ASPARTATE_AND_GLUTAMATE_METABOLISM	0.9041	0.0267	0.2325	8.06	7.69	7.47
KEGG_GLYCINE_SERINE_AND_THREONINE_METABOLISM	0.9193	0.0317	0.2325	8.45	7.54	7.70
KEGG_LINOLEIC_ACID_METABOLISM	0.9364	0.0300	0.2325	8.07	5.87	5.36
KEGG_PROPANOATE_METABOLISM	0.9154	0.0305	0.2325	8.30	7.80	7.73
KEGG_BUTANOATE_METABOLISM	0.8987	0.0277	0.2325	8.04	7.56	7.27
KEGG_BETA_ALANINE_METABOLISM	0.9997	0.0386	0.2654	7.73	7.15	7.17
KEGG_PROXIMAL_TUBULE_BICARBONATE_RECLAMATION	1.0211	0.0483	0.3125	7.39	7.38	7.17

Interestingly, the most significant loss of co-regulation observed in transformed hepatocytes involves the peroxisome metabolism (KEGG ko04146), the primary bile acid biosynthesis (map00120), and the glyoxylate and dicarboxylate metabolism (map00630): these pathways are responsible for fundamental functions in liver cells such as the synthesis of bile acids, cholesterol, the oxidation of fatty acid, the metabolism of phenyalanine, the glyoxylate and the tyrosine metabolism. Moreover, among the other dysregulated pathways identified by DINA, we found disruption of fundamental pathways regulating liver cancer progression such as the PPAR signaling pathway (Table 3).

In order to gain further insights into the dysregualtion of the peroxisome metabolism, I analyzed the changes in the gene co-regulation network among the corresponding genes across the three cell

lines. Figure 21A and Figure 21B demonstrate that there is a major loss of co-regulation among peroxisome related genes in both HepG2 and Huh7 hepatocarcinoma cell lines; moreover this loss mainly results from dysregulation of genes involved in peroxisomal fatty acid β-oxidation (e.g. ACOX, EHHADH, ACAA1) and genes involved in the control of the H202 metabolism (e.g. CAT and SOD). Notably, these genes are regulated by the peroxisome proliferator-activated receptor alpha (PPARalpha) [95] and the LXR family transcription factors [92].

Thus, the DINA results indicate that the dysregulation in the activity of these liver specific transcription regulators may represent a recurrent event associated with hepatocarcinoma. Consistent with these results, peroxisome and PPARalpha pathway alterations have been definitely associated with liver cell proliferation and with hepatocarcinoma development [96], confirming the efficacy and specificity of DINA algorithm in identifying condition-specific pathway regulation.

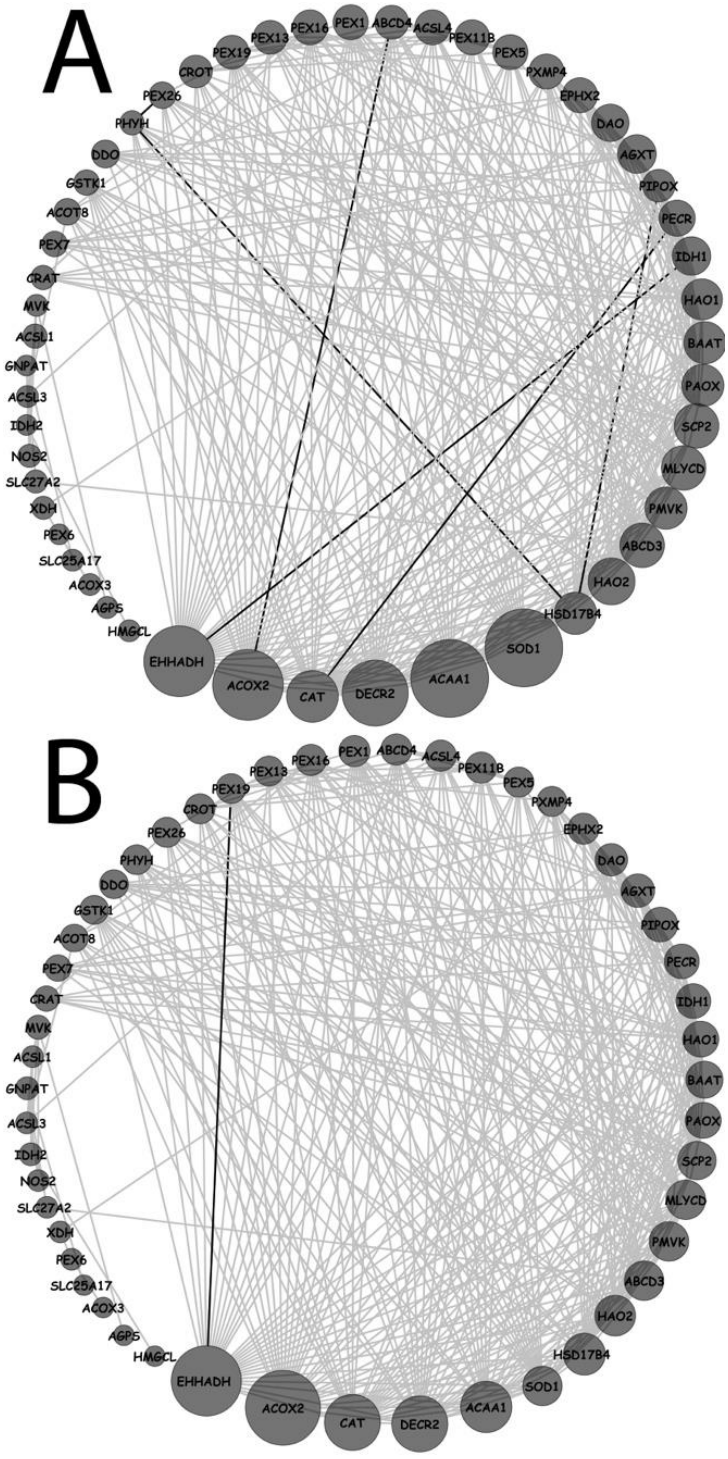


Figure 21 - Differential Network Analysis of the peroxisome KEGG pathway (M6391) in primary and transformed hepatocytes. Genes in the peroxisome pathway are represented as circles; a significant co-regulation between two genes as a line. The size of the circles is proportional to the difference in the number of edges between the networks in transformed hepatocytes versus primary hepatocytes. Gray lines represent edges lost in the network com-pared to primary cells. (**a**) HepG2 versus primary hepatocyte; (**b**) HepG2 versus primary hepatocyte. (Figure taken from [68])

4.5 A case of study: YEATS2: a negative transcriptional regulator of metabolic pathways

I wondered whether it was possible to identify transcription factors (TFs) regulating tissue-specific pathways identified by DINA [68]. The working hypothesis was that a TF, controlling a tissue-specific pathway, may be co-regulated with its target genes only in that tissue but not in others (Figure 15B). Since the regulation of metabolic pathways has been well studied in the past, we decided to identify TFs involved in the regulation of the 9 metabolic pathways (Table 2 in red) previously identified by DINA.

To this end, I used the list of 1358 human genes including both genes, whose protein product has a verified TF activity [97], as well as, genes encoding proteins with an indirect transcriptional activity, such as co-factors or scaffolding proteins. For each of the 9 metabolic pathways previously identified as tissue-specific, and for each TF in the list, I applied the method presented in the paragraph 4.2 to select TFs sharing a significant number of edges with the genes in the pathway only in the tissue(s) where the pathway is active, as shown in Figure 15B.

Table 4 lists the master TFs controlling the majority (i.e. 7 out of 9) of the metabolic pathways according to our analysis. Considering only genes encoding proteins with a known TF activity (Table 4 in bold), this method correctly identified many nuclear receptors as specific regulators of these pathways (NR1H4, NR1I3, ESRRG, HNF4A). The nuclear receptor super-family is one of the largest groups of TFs involved in the regulation of different metabolic processes [98], such as the regulation of liver metabolism [99].

For example, as shown in Table 4, one of the six nuclear receptors is HNF4A, probably the most famous nuclear receptor in liver, whose mutations are responsible for monogenic autosomal dominant non-insulin-dependent diabetes mellitus type I (MODY1) [100]. The protein encoded by this gene controls the expression of several genes, including hepatocyte nuclear factor 1 alpha (HNF1A), a transcription factor that regulates the expression of several hepatic genes.

When we considered also genes encoding proteins indirectly involved in transcription [97] (Table 4 not in bold), we identified, among others, SIRT4 (sirtuin 4), a member of the sirtuins' family that plays a key role in human metabolic regulation [101-103].

Among the list of protein indirectly involved in the regulation of transcription, the gene YEATS2 has attracted my attention because was predicted by DINA to be the most significant negative regulator shared by most of the metabolic pathways (Table 4 not in bold). I checked for the expression level of YEATS2 as described in [104], and I observed that YEATS2 gene is expressed at very low levels in both

liver and kidney and, at the time of writing of this thesis, very little is known about its function. Recently, it has been demonstrated that YEATS2 interacts with the ATAC complex (Ada-Two-A-Containing) [105], which, together with SAGA (Spt-Ada-Gcn5-Acetyl-Transferase), is able to modulate transcription, both by causing chromatin modification and by interacting with the TATA-binding proteins (TBPs) [105, 106]. But no association of this gene with metabolism is existent until now in literature.

In order to validate DINA's prediction about the involvement of YEATS2 in the transcriptional regulation of metabolism in liver, we decided to further investigate its function by perturbing hepatocytes homeostasis by starvation [107].

Table 4 - Transcription factors identification for the tissue-specific metabolic pathways identified by DINA. List of transcription factors regulating the majority (i.e. 7 out of 9) of the tissue-specific metabolic pathways. In bold genes with know TF activity, in normal text genes encoding protein indirectly involved in transcription. The column citations contain works reporting the association of the transcription factor and metabolism. The column *Role* contains the function (activator or inhibitor) of the TF predicted by DINA.

Gene Symbol	Name	Role	Citations
NR1H4	nuclear receptor subfamily 1, group H, member 4	activator	[108-110]
ESRRG	estrogen-related receptor gamma	activator	[109, 111]
TRPS1	trichorhinophalangeal syndrome I	inhibitor	
NR1I3	nuclear receptor subfamily 1, group I, member 3	activator	[109, 112, 113]
HNF4A	hepatocyte nuclear factor 4, alpha	activator	[109, 114]
ZNF394	zinc finger protein 394	inhibitor	
TBR1	T-box, brain, 1	activator	
DAB2	disabled homolog 2	activator	
DIP2C	disco-interacting protein 2	activator	
TRIM15	tripartite motif-containing 15	activator	
ASB9	ankyrin repeat and SOCS box-containing 9	activator	
YEATS2	YEATS domain containing 2	inhibitor	
SIRT4	sirtuin 4	activator	[101-103]

During starvation, a switch from anabolism to catabolism occurs: cells start to mobilize stored nutrients, such as glycogen and triglycerides, cell growth is arrested and autophagy is promoted [107, 115]. During starvation there are large changes in gene expression that affect specific metabolic pathways. For example, genes involved in fatty acid β-oxidation are up-regulated [112] whereas genes involved in biosynthesis are down-regulated [116].

In collaboration with Nicoletta Moretti at TIGEM (Naples, Italy), we performed a starvation time-course experiment for 8 hours in primary murine hepatocytes, by switching cells from a nutrient-rich medium to a starvation medium. Cells were collected at different time points during starvation (30 min, 1h, 2h, 4h, 6h, 8h). Cells grown in nutrient-rich medium were used as control. We measured by quantitative real-time PCR (qRT-PCR) the variation in the expression level of Yeats2 in response to starvation in primary hepatocytes (Figure 22). As shown in Figure 22, Yeats2 seems to be an early response gene, quickly down-regulated upon starvation during the first two hours in primary hepatocytes. We also analysed the expression profiles of a subset of genes whose expression levels increase following starvation [112, 117]: Pgc1a, Acaa1a, Acot2, Cyp4a10, Cyp4a14, ApoA4 and Plin4 (Figure 22).

These selected genes were up-regulated, as expected, during the first four hours of starvation, as shown in Figure 22: Pgc1a (Peroxisome proliferator-activated receptor gamma, co-activator 1 alpha) encoding for a transcriptional co-activators that plays a key role in the regulation of both carbohydrate and lipid metabolism [Leone2005]; Acaa1a (Acetyl-CoA acyltransferase 1A) encoding a peroxisomal thiolase operating in catabolism of fatty acid [112] together with ACOT2 (Acyl-CoA thioesterase 2) which is localized in peroxisomes [118]; Cyp4a10 (Cytochrome P450, family 4, subfamily a, polypeptide 10) and Cyp4a14 (Cytochrome P450, family 4, subfamily a, polypeptide 14) encoding two members of Cytochrome P family able to oxidize a variety of structural compounds, as well as fatty acids [112, 119]. Genes involved in lipid transport showed an up-regulation as well, such as ApoA4 (Apolipoprotein A4), which enhances lipid absorption by promoting the assembly and secretion of Chylomicrons [112, 120].

In order to probe further the role of Yeats2 and its involvement in regulation of metabolism in liver, I analysed an existing in vivo time-series microarray experiment (ArrayExpress ID E-MEXP-748) from liver, muscle and adipose tissue of ApoE3Leiden transgenic mice, exhibiting a humanized lipid metabolism, treated with high-fat diet (HFD) for 0, 1, 6, 9, or 12 weeks [121]. Upon HFD feeding, genes involved in metabolic pathways, such as lipid metabolic processes, were found to be up-regulated in liver [121]. Based on these observations, we decided to investigate the expression of Yeats2 in this mouse model considering only the liver tissue, and we found that Yeats2 expression is strongly down-regulated in HFD mouse liver (p-value of $3.38\ e10^{-8}$) [121, 122].

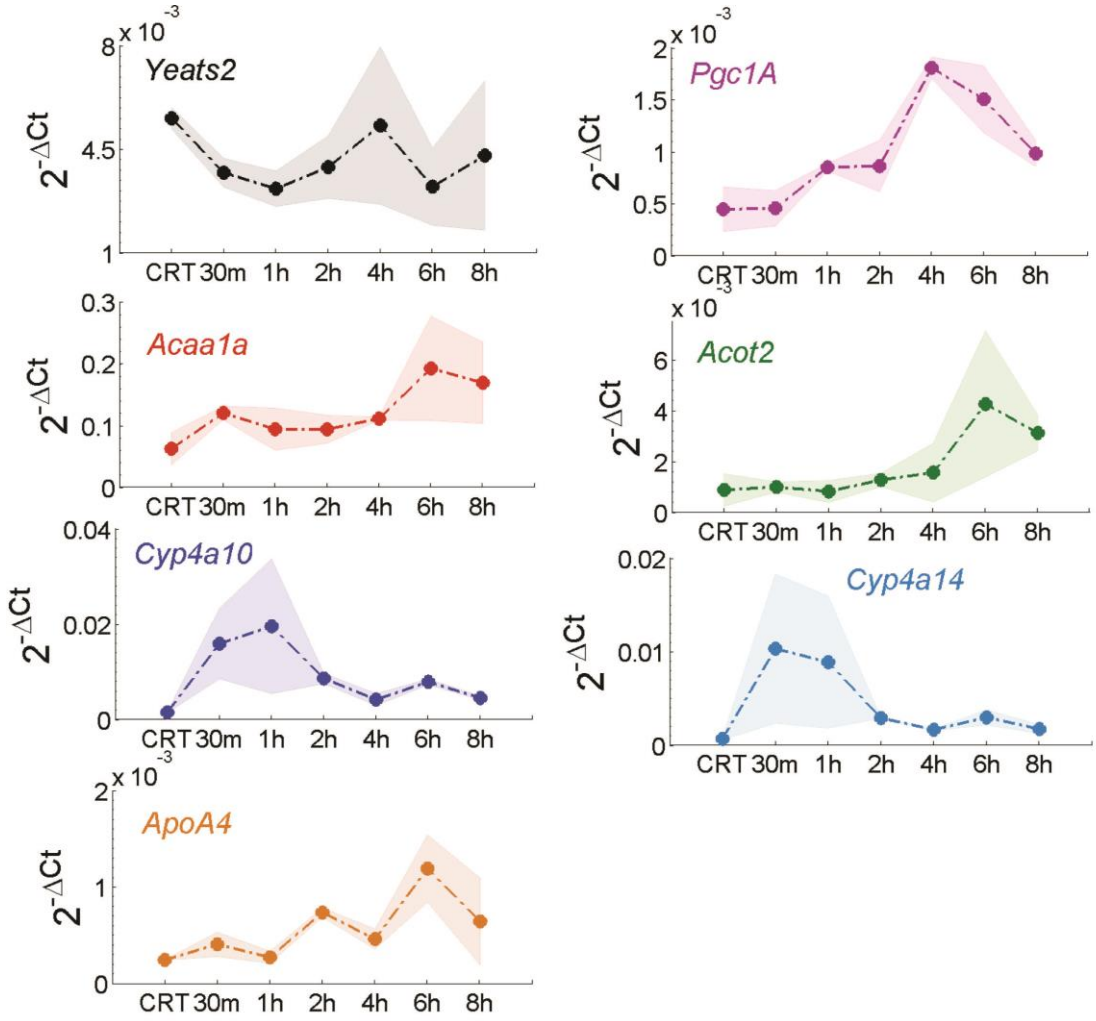


Figure 22 - Yeats2 expression in hepatocyte cells during starvation. Real-time quantitative PCR measurements of the expression of Yeats2 and a set of marker genes at the indicated time-points following starvation. CRT indicates cell in rich medium. BF indicated the Bayes Factor estimated using BATS algorithm [123]. The gray area represents the standard deviation across the two bio-logical replicates. Gene expression was quantified using the ΔCT method with Gapdh used as normalization gene. (Figure taken from [68])

4.6 A web-tool implementing DINA

I developed a web tool available at http://dina.tigem.it using Java Servlet Page (JSP) technology. The web-tool enables the user to query the 30 tissue specific networks with a set of genes gene in order to verify if the set is predicted to be tissue-specific or not in humans. As shown in Figure 23, the index page of the tool contains a text box in which the user can insert the gene list. As shown in Figure 24 the co-

regulation probabilities across the 30 co-regulatory networks for the user's gene list is reported in a resulting page using the radar-chart. Also the normalized entropy value (i.e. $0 \leq H \leq 1$) of the pathway is reported in the result's page.

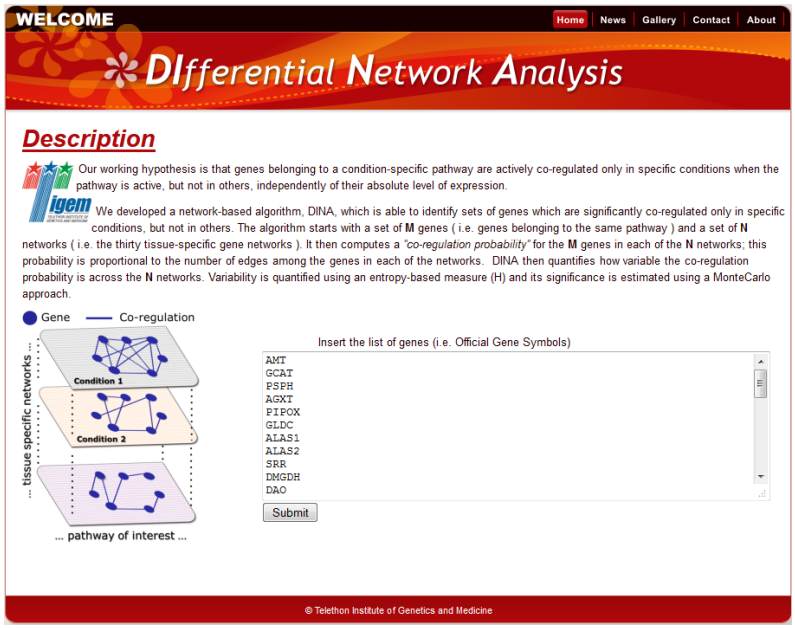


Figure 23 – Index page of DINA web tool. A User can insert his gene signature in order to evaluate if a gene signature is tissue specific or not.

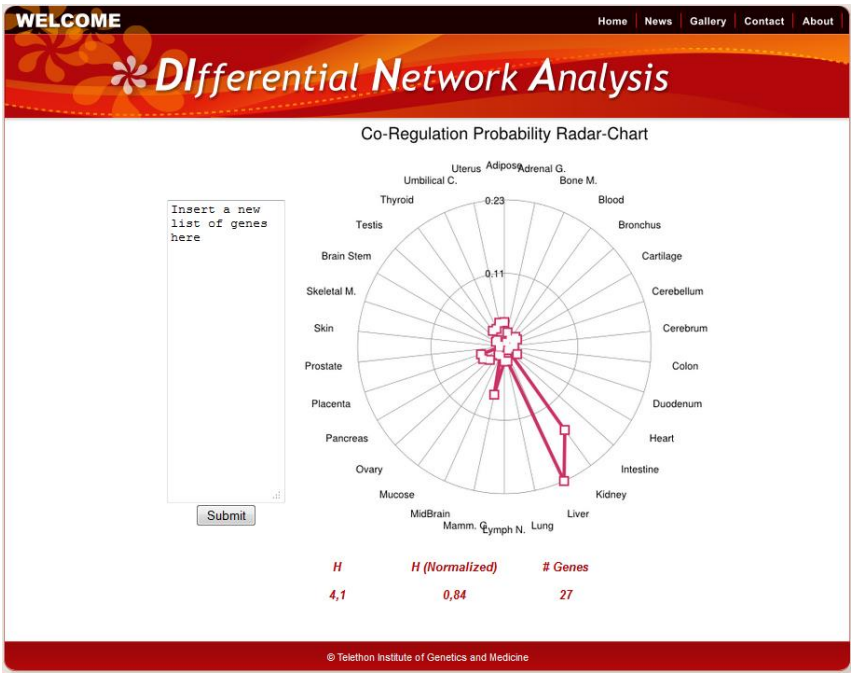


Figure 24 – Page of the results for DINA web tool. The co-regulation probability across the 30 tissue specific co-regulatory network of the input gene signature is showed in radar-chart.

4.7 Discussion and Conclusions

In this chapter, I proposed a network-based approach (DINA) for the identification of condition specific pathways (or gene signatures). DINA is based on the hypothesis that genes belonging to a condition-specific pathway are actively co-regulated only in specific conditions where the pathway is active, but not in others, independently of their absolute level of expression. DINA is based on an entropy-like measure and it is able to assess the specificity of a pathway by quantifying the variability in the co-regulation probability of genes in the pathway across a set of condition-specific networks. Since DINA is based on detecting differences in the number of edges among genes in a pathway across a set of networks, it can be applied to any kind of network, independently of how this is generated. Differently from other available methods, DINA does not aim at identifying *de-novo* sub-networks of genes, but rather at identifying whether a known pathway (or a user's gene signature) is differentially co-regulated across a set of conditions.

Using thirty tissue-specific gene co-regulatory networks, I was able to show that DINA can be successfully applied to identify tissue-specific pathways. Indeed, as expected, several metabolic

pathways were predicted by DINA to be the most differentially regulated across the tissues and thus tissue-specific pathways specifically active in liver and kidney. Usually, tissue-specificity of a gene, or of a pathway, is assessed by quantifying the expression level of the genes in the concerned tissue [124]. However, observing gene expression only could be not sufficient, as in the case of metabolic pathways [125]. Here, I show that an alternative possibility is to check if the genes involved in the same pathway are specifically co-regulated in the concerned tissue. Of note, a similar approach has been successfully applied in yeast [126].

In this Chapter, I also proposed a method based on *the Fisher's exact test* to identify tissue-specific Transcription Factors assuming that its tissue-specific targets tend to be co-regulated with it in a tissue-specific manner. Hence, I tested this approach to identify regulators of tissue-specific metabolic pathways, and correctly identified Nuclear Receptors as their main regulators. With this method I was also able to identify a new putative tissue-specific negative regulator of hepatocyte metabolism Yeats2.

Finally, I showed that DINA could be employed to analyse GEPs obtained during disease progression to make hypotheses on dysregulated pathways using as a study case a simplified model of hepatocellular carcinoma. It remains to be seen whether changes in signaling pathway activity can be detected using only a transcription based approach such as DINA.

Chapter 5

A new differential multi information approach for the identification of PTMs

5.1 A new method based on Differential Multi Information (DMI)

Transcriptional regulations in a cell may be modulated at many different levels, such as transcription factor activation/deactivation by phosphorylation, or formation of active complexes with one or more cofactors. These modulators exert their function at the post-transcriptional/post-translational level, and therefore capturing this kind of regulatory interactions using transcriptional data, such as gene expression profiles (GEPs), is considered not possible. Recently, it has been shown that this is not the case, since fluctuations in the transcriptional level of modulators across different conditions can be exploited to infer post-translational regulation [65].

Our working hypothesis is the scenario in which a modulator (i.e. kinase/phosphatase) is expressed and activates a Transcription Factor (TF) (Figure 25B). As a result, the TF direct targets become co-regulated among them since they are all controlled by the same TF (Figure 25B). On the contrary, when the modulator M is not expressed, the TF is inactive and hence its targets are not co-regulated. It is important to observe that our working hypothesis does not depend on changes in expression level of the target genes, but rather on changes in "co-regulation". Indeed, as described in the previous chapters in the case of metabolic pathways, changes in co-regulation of the metabolic pathway enzymes are predictive of an active pathway, but their expression level does not change considerably among different tissues. Hence, the idea behind the method we developed consists of quantifying changes in **co-regulation** among a set of downstream targets G^1 … G^n of the TF in presence (or absence) of a modulator M.

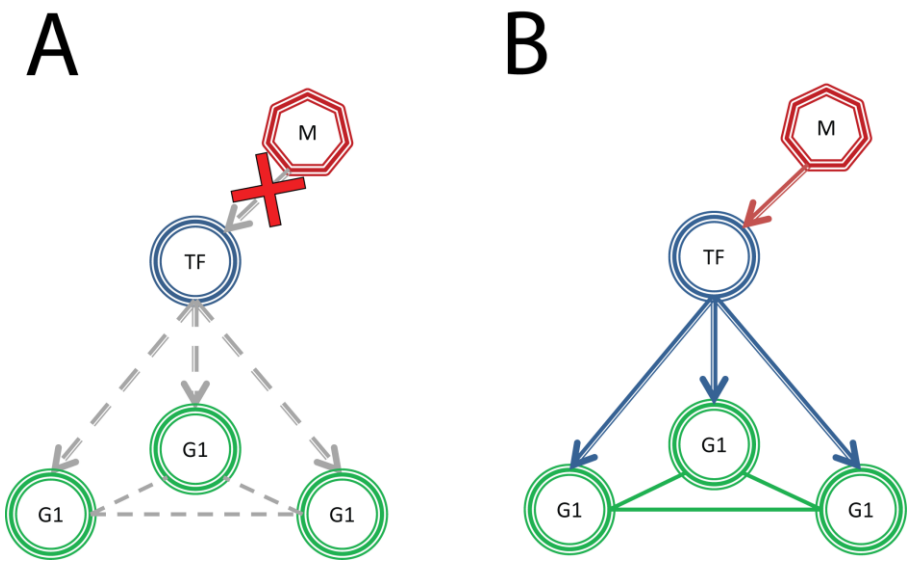


Figure 25 – Hypothetical scenario in which a hypothetical Transcription Factor (**TF**) is activated by phosphorylation or de-phosphorylation through a Modulator (**M**). **G1**, **G2** and **G3** are three downstream targets of the **TF**. (**a**) In absence of the Modulator (**M**) the downstream targets (**G1**, **G2** and **G3**) are not co-regulated since the Transcription Factor (**TF**) is not active. (**b**) In presence of the Modulator (**M**) the downstream targets (**G1**, **G2** and **G3**) become co-regulated through the active Transcription Factor (**TF**).

This method requires a way to estimate the co-regulation among n variables, to this end I chose to compute the distance of their Joint Probability Density from the product of their Marginal Probabilities. This measure is known as the **Multi-Information** $I(X^1 \cdots X^n)$ of n random variables and is computed as the KL-Divergence (D_{KL}) between their joint distribution and the product of their marginal.

$$I(X^1 \cdots X^n) = D_{KL}(\mathbf{G}||\mathbf{G}') = \int_{\mathbb{R}^n} p(x^1 \cdots x^n) \log \frac{p(x^1 \cdots x^n)}{p(x^1) \ldots p(x^n)} d(x^1 \cdots x^n)$$

Where $\mathbf{G}$ is the joint distribution of the targets and $\mathbf{G}'$ is the product of the marginal probabilities. $X^1 \cdots X^n$ are random variables representing the gene expression level of targets G^1 … G^n respectively. By this definition, the multi-information provides a measure of dependence among all the n variables and is non-negative because of the non-negativity of the D_{KL}.

Multi-Information can be also considered as a variant of the "Multivariate Mutual Information" as defined by McGill [127]. According to this definition, the Multivariate mutual information $I(X^1 \cdots X^n; Y)$ between a joint random variable $\mathbf{X}$ and some label Y is a Shannon Mutual Information but it measures only the interactions between the variables $\mathbf{X}$ that are dependent on Y, discarding all

the other non-relevant intearctions. Multi-Information, instead, measures all the interactions among all the variables giving a completely different answer. The properties of the Multi-information are further discussed in [128].

In order to apply the Multi-Information measure to detect the modulator M of the TF given its targets $G^1 \dots G^n$, I developed the approach depicted in Figure 2. Specifically, I first discretize the expression of the modulator M in n bins and then compute the difference in Multi-Information (ΔI) among the targets in the bin where M is "High" and "Low" (Figure 26). Hence, ΔI quantifies how much the modulator is able to influence the co-regulation of the targets of a given transcription factor. Since M is not known a-priori this approach is iterated for each of the modulator M to be tested and results are then ranked by ΔI and by significance as detailed in 5.1.2.

In order to discretize the expression of the modulator M, I decided to use the quantile discretization approach, in which each bin receives an equal number of data values and the data range of a bin varies according to the data values it contains. Specifically, each expression value of the modulator M is replaced by an integer value within the discrete interval $[1, nbin]$, where $nbin$ is the total number of bins in which we wish to discretize the expression of M. Each integer corresponds to the bin where the expression value falls into. Thus, samples whose expression values are replaced with the number 1 will be those where M is expressed at a low level, whereas samples whose expression values are replaced by the number $nbin$ will be those where M is highly expressed [129, 130].

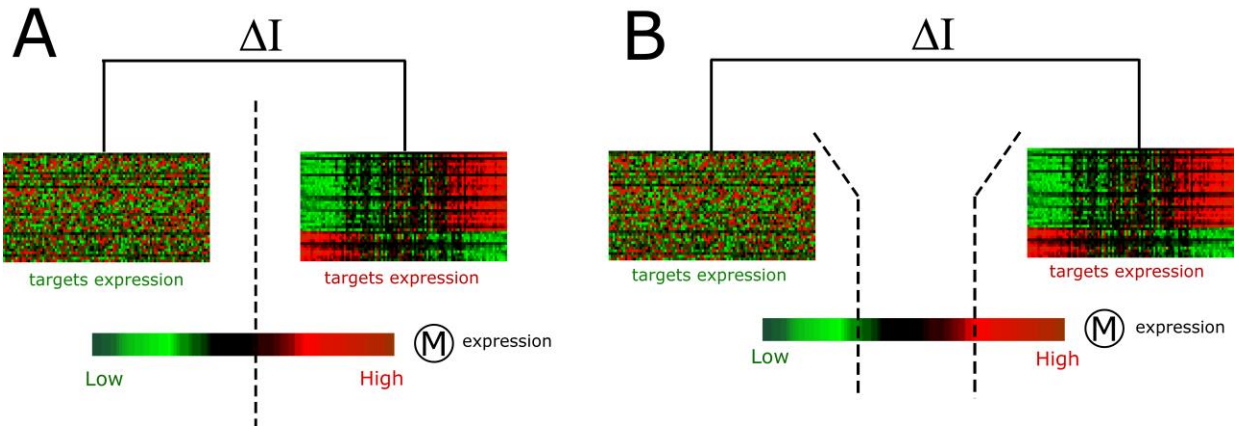


Figure 26 – For each step of the algorithm a candidate modulator **M** is tested. In the first step of the method the expression of the modulator **M** is discretized in $\boldsymbol{n}$ bins and the Differential Multi-Information ($\Delta \boldsymbol{I}$) of the targets is computed always between the two bins where **M** expression is either "'High"' or "Low''. In the samples of "High" bin, the targets are strongly co-regulated, since they are controlled by the same transcription factor (activated by **M**). In the samples of the "Low" bin, the targets are not co-regulated since **M** is not able to activate the transcription factor. **(a)** Example of 2-bins discretization for the expression of **M**. **(b)** Example of 3-bins discretization for the expression of **M**.

The problem can be seen from a different point of view as the one of finding the modulator M that best partition the data in order to maximize the Multi-Information of the targets in the partition where M is "High", and at the same time minimize the Multi-Information in the other partition where M is "Low".

5.1.2 Significance estimation of DMI using permutation tests

Since I cannot make any assumption regarding the probability distribution of the ΔI, the only approach that I can use here is a permutation test in which we estimate the empirical distribution of ΔI. Moreover, the value of multi-information is dependent on the number of variables among which we are estimating it, meaning that the same multi-information value computed among n and m variables (with $n \neq m$) is not directly comparable.

Hence, I proceeded as follows: first, I fixed a set of d target genes (i.e. variables) and a number of bins b to discretize the modulator M expression, I then computed the significance of a modulator M by randomly selecting d genes in L number of trials, and each time I re-computed the ΔI value thus generating its empirical distribution. The statistical power of the test is thus dependent of the parameter L whose upper bound is limited by the computational cost. Indeed the computational complexity of this permutation test is of the order $O(L \cdot MI(n, d))$. It is obviously a function of the computational complexity of Multi-Information (MI) estimation algorithm (details in paragraph 5.2) that is itself function of the parameters d (number of genes) and n (number of samples used for its estimation). Considering that the permutation test must be executed for the upper and lower bin of each one of the m modulators to test, then the total computational complexity becomes $O(2mL \cdot MI(n, s))$ In the following tests, I decided to use L=1000 as a compromise between statistical power and computational cost.

5.2 Rényi Multi-Information and its estimation $\hat{I}_{\alpha}$

In order to estimate the Multi-Information in each bin, I decided to use the Rényi Multi-Information (RMI). The RMI of d real-valued random variables $\mathbf{X} = (X^1 \dots X^d)$ with joint density $f: \mathbb{R}^d \to \mathbb{R}$ and

marginal densities $f_i: \mathbb{R} \to \mathbb{R}, 1 \leq i \leq d$ is defined for any real parameter α assuming the underlying integrals exist [131]. For $\alpha \neq 1$ Renyi multi-information is defined as:

$$I_\alpha(\boldsymbol{X}) = I_\alpha(f) = \frac{1}{\alpha - 1} \int_{\mathbb{R}^d} \frac{f^\alpha(x^1 \dots x^d)}{\left(\prod_{i=1}^d f_i(x^i)\right)^{\alpha-1}} d(x^1 \dots x^d)$$

For $\alpha = 1$ it is defined by the limit $I_1 = \log_{\alpha \to 1} I_\alpha$. In fact, the classical multi-information is just special case of Rényi multi-information with $\alpha = 1$.

Definition (generalized nearest-neighbor graph) [131]**:** Let V be a finite set of points in an Euclidean space $\mathbb{R}^d$ and let S be a finite non-empty set of positive integers. We define the generalized nearest-neighbor graph $NN_S(V)$ as a directed graph on V. The edge set of $NN_S(V)$ contains for each $i \in S$ an edge from each vertex $x \in V$ to its i-th nearest neighbor. That is, if we sort $V \setminus \{x\} = \{y_1, y_2, \dots, y_{|V|-1}\}$ according to the Euclidean distance to x: $\| x - y_1 \| \leq \| x - y_2 \| \leq \cdots \leq \| x - y_{|V|-1} \|$ then y_i is the i-th nearest-neighbor of x and for each $i \in S$ there is an edge from x to y_i in the graph. ■

Definition (sum of the p-th powers of Euclidean lengths of its edges) [131]**:** Let V be a finite set of points in an Euclidean space $\mathbb{R}^d$ and let S be a finite non-empty set of positive integers. For $p \geq 0$ let us denote by $L_p(V)$ the sum of the p-th powers of Euclidean lengths of its edges.

$$L_p(V) = \sum_{(x,y) \in E(NN_S(V))} \|x - y\|^p$$

where $E(NN_S(V))$ denotes the edge set of $NN_S(V)$■

We are now ready to present the estimator of Rényi Multi-Information based on the generalized nearest-neighbor graph and a copula transformation. For more details and the theorem demonstrations, please refer to [131].

Suppose we are given an i.i.d. sample $\mathbf{X}_{1:n} = (\mathbf{X}_1 \dots \mathbf{X}_n)$ where each $\mathbf{X}_j = (X_j^1, X_j^2, \dots, X_j^d)$ has density $f: \mathbb{R}^d \to \mathbb{R}$ and marginal densities $f_i: \mathbb{R} \to \mathbb{R}, 1 \leq i \leq d$. In [131] the authors showed that we can estimate the entropy $H_\alpha(f)$ for $\alpha \in (0,1)$ as:

$$\widehat{H}_\alpha(\mathbf{X}_{1:n}) = \frac{1}{1-\alpha}\log\frac{L_p(\mathbf{X}_{1:n})}{\gamma n^{1-p/d}} \quad \text{where} \quad p = d(1-\alpha)$$

with $L_p(\cdot)$ equal to the sum of the *p*-th powers of Euclidian lengths of edges of the nearest-neighbor graph $NN_S(\cdot)$ for some finite non-empty $S \subset \mathbb{N}^+$ and with γ representing a numeric constant dependent on d, p and S that can be estimated empirically from a large i.i.d. sample, as detailed in [131],.
Finally, as described in [131], we can estimate the Rényi Multi-Information I_α by

$$\hat{I}_\alpha(\mathbf{X}_{1:n}) = -\widehat{H}_\alpha\left(\hat{\mathbf{Z}}_1, \hat{\mathbf{Z}}_2, \dots, \hat{\mathbf{Z}}_n\right)$$

where $\widehat{H}_\alpha$ is defined as before and the sample $\left(\hat{\mathbf{Z}}_1, \hat{\mathbf{Z}}_2, \dots, \hat{\mathbf{Z}}_\mathrm{n}\right) = (\hat{\mathbf{F}}(\mathbf{X}_1), \hat{\mathbf{F}}(\mathbf{X}_1), \dots, \hat{\mathbf{F}}(\mathbf{X}_\mathrm{n}))$**.** $\hat{\mathbf{F}}(\cdot)$ is called *empirical copula transformation* [132] and basically consist into the following transformation where the *j*-th coordinate of $\widehat{\mathbf{Z}}_i$ equals

$$\hat{Z}_i^j = \frac{1}{n}\mathrm{rank}(X_i^j, (X_1^j, X_2^j, \dots, X_n^j))$$

where $\mathrm{rank}(x, A)$ is the number of elements of A less than or equal to x.

The computational complexity $T(n)$ for the estimation of Rényi Multi-Information depends of the complexity of the K nearest-neighbors algorithm, which is linear in the number of point and the number of features for each point, and the complexity of copula transformation, which is quadratic in the number of point. Hence, the computational complexity for the estimation of Rényi Multi-Information is:

$$T(n) = O(n^2 d + nd)$$

where n is the number of i.i.d. samples used for its estimation (i.e. number of experiments) and d is the number of features of each i.i.d sample (i.e. number of genes).

5.2.1 The rate of convergence of $\hat{I}_\alpha$

I tested the convergence of the estimated Rényi Multi-Information to the true value. I generated two dataset of 4000 i.i.d. samples from a multivariate Gaussian distribution of dimension 3 with 0 mean and an identity covariance matrix, corresponding to independent variables ($MI = 0$) or a randomly chosen symmetric covariance matrix, corresponding to a $MI > 0$. Figure 27 show the estimation of $\hat{I}_\alpha$ among the three variables in the case of dependent variables (i.e. $MI > 0$) and its error. Figure 28 instead shows the estimation of $\hat{I}_\alpha$ among the three variables in the case of independent variables (i.e. $MI = 0$).

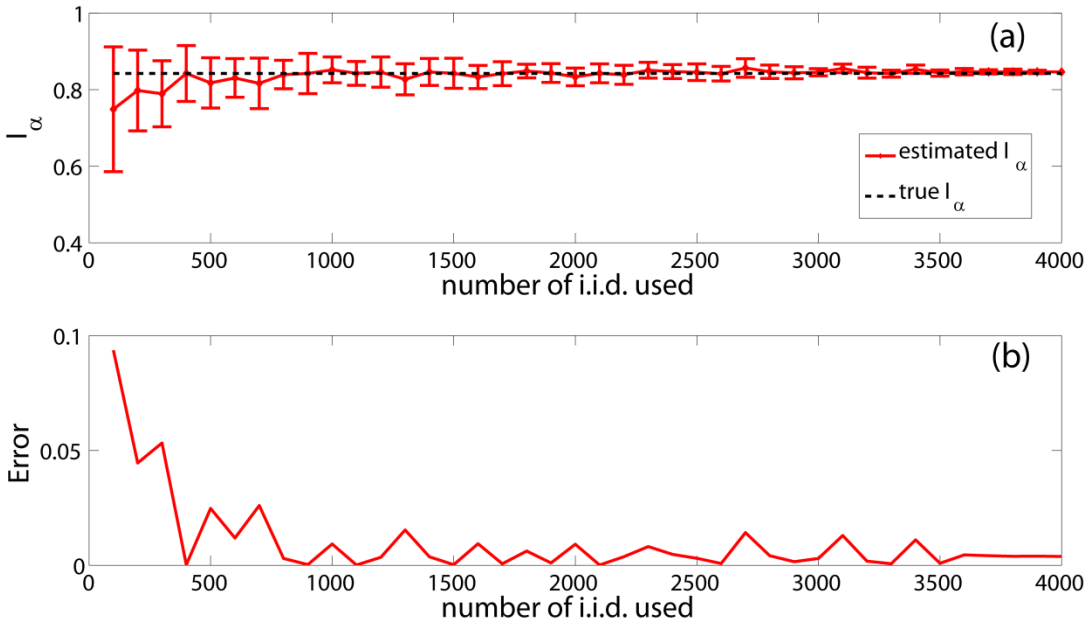


Figure 27 - $\hat{I}_\alpha$ for 3 variables as a function of the number of i.i.d used. The 3 variables are dependent variables. The estimation of $\hat{I}_\alpha$ is computed 20 times for each point and its standard deviation is reported. **(A)** The convergence of $\hat{I}_\alpha$ to the true value of I_α. **(B)** The error to estimate $\hat{I}_\alpha$ as a function of the number of i.i.d. used.

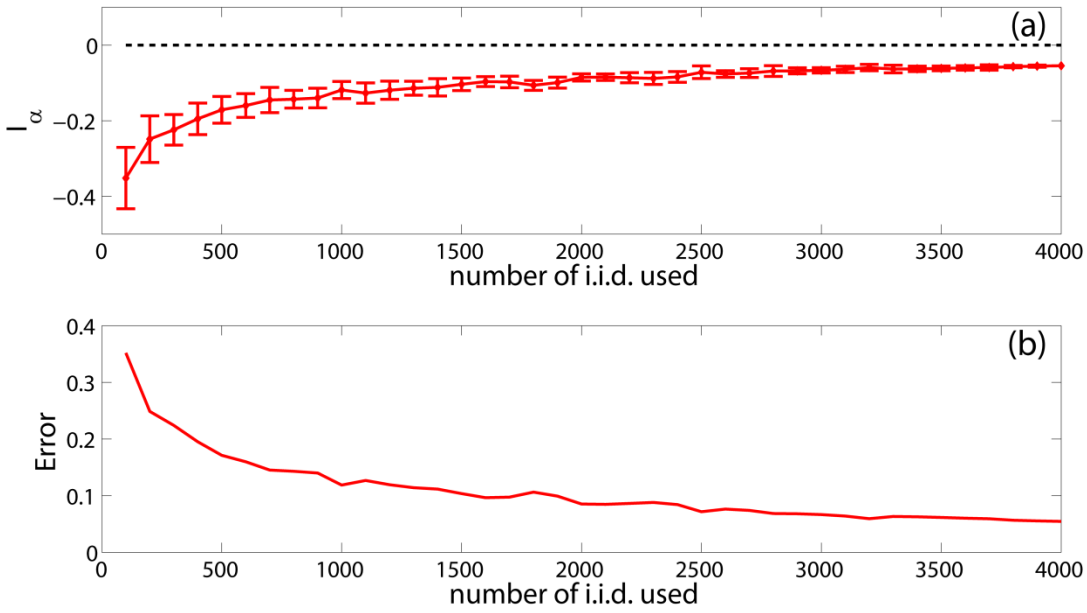


Figure 28 - $\hat{I}_\alpha$ among 3 variables as a function of the number of i.i.d used. The 3 variables are independent variables. The estimation of $\hat{I}_\alpha$ is computed 20 times for each point and its standard deviation is reported. (A) The convergence of $\hat{I}_\alpha$ to the true value of I_α. (B) The error to estimate $\hat{I}_\alpha$ as a function of the number of i.i.d. used.

5.3 Alternative approaches to identify post-translational modulators.

The difference in Rényi Multi-Information (DMI) I described above is certainly not the only possible way to measure changes in the dependence among variables and hence to identify post-translational modulators of TF activity. Therefore in addition to DMI I also explored two alternative approaches to as detailed below.

5.3.1 Multidimensional Independent Test (MIT)

Since I am interested in detecting changes in co-regulation among the targets $G^1 \cdots G^i$ of a TF due to a modulator M, an alternative strategy is to use an independence test between the variable $\mathbf{G} = (G^1 \cdots G^i)$ and the variable $\mathbf{M} = (M^1 \cdots M^j)$ representing the modulators (i.e. here$\mathbf{M} = (M^1)$).

Formally, this is equivalent to consider a sample of $\mathbb{R}^i \times \mathbb{R}^j$ valued random vectors $(G_1, M_1) \dots (G_n, M_n)$ with independent and identically distributed (i.i.d.) pairs defined on the same probability space and testing the null hypothesis (H_0) that $\mathbf{G}$ and $\mathbf{M}$ are independent:

$$H_0: P(\mathbf{G}, \mathbf{M}) = P(\mathbf{G}) \times P(\mathbf{M})$$

and possibly making minimal assumptions regarding the probability distributions. To test H_0 we can use one of the two methods presented in [133]. The authors proposed two different approaches to test the independence between two multidimensional variables. The first method consists in partitioning the underlying space, and in evaluating a test statistics on the resulting discrete empirical measures. I did not test this method because of the computational complexity, which is exponential in the number of bins used for the discretization step elevated to the sum of the dimensions of the two variables to test. The method that I tested is the other method proposed by Gretton in [133], it is a kernel-density estimation test that is based on a cross-covariance operator and on reproducing kernel Hilbert space (RKHS).

A possible limitation of this strategy (i.e. test only the independence between two multidimensional variables) consists in the fact that if the modulator M being tested is not a real modulator but instead it is a itself a target of the TF, then it will be strongly co-regulated with the targets G, and the method would detect a dependence between M and G, and hence M would be flagged as a modulator of the TF. This modulator however would obviously be a false positive.

5.3.2 Conditional Multidimensional Independent Test (CMIT)

Another possible approach could be of using the CMIT test described in [134]. Here, the authors present a new measure of conditional dependence among random variables, based on normalized cross-covariance operators and on reproducing kernel Hilbert spaces. The proposed criterion does not depend on the choice of kernel in the limit of infinite data, for a wide class of kernels. At the same time, it has a straightforward empirical estimate with good convergence behaviour.

The CMIT test applied to the problem of finding modulator M of a TF can be implemented as follows: given the target genes $\mathbf{G} = (G^1 \cdots G^i)$ of the transcription factors $\mathbf{F}$ and a modulators $\mathbf{M}$, we

can test the conditional independence of $\mathbf{G}$ and $\mathbf{F}$ given $\mathbf{M}$ using the CMIT test, thus solving the problem of finding the modulator for which the targets and the transcription factor(s) are co-regulated.

In order to apply the CMIT test, we have to know which is the TF that regulates our targets (this condition is usually satisfied) but we also have to assume that the TF and targets are co-regulated and so statistically dependent, introducing others constraints in our model.

5.4 Performance of DMI, MIT and CMIT on simulated datasets

In the following section, I will illustrate the test I performed to assess the efficacy of the Differential Multi-Information (DMI) procedure I introduced in the previous sections. I will start with a definition of the performance measure I decided to use followed by a description of the datasets used for the test and the results I obtained. I also compared the performance of the DMI approach with other possible measures of dependence among n variables.

5.4.1 PPV-Sensitivity Curve

Positive Predicted Value (PPV) is defined as the fraction of predicted modulators (M) that are correctly identified by the algorithm, that is:

$$PPV = \frac{TP}{TP + FP}$$

where TP are the true positive and FP are the false positives. Sensitivity, instead, is defined as the fraction of all the true modulators that are retrieved by the algorithm:

$$Sensitivity = \frac{TP}{TP + FN}$$

where FN are the false negatives. Hence, a PPV-Sensitivity curve represents the Precision against the Sensitivity with the predictions of the algorithm ranked according to the value of ΔI.

In order to assess the goodness of the performance, I also computed the expected performance of an algorithm that randomly choses modulators assuming a uniform probability distribution over the set modulators. This means that:

$$PPV_{rand} = \frac{\#\text{ of true modulators}}{\#\text{ of tested modulators}}$$

5.4.2 Generation of the "in silico" dataset

I generated two datasets consisting of 100 simulated GEPs each. The list of the two "in-silico" datasets I generated and the parameters I used are reported in Table 5. The first dataset D1 consists of 60 genes, among which only 10 genes are assumed to be the known targets of the TF (i.e. are $\mathbf{G} = (G^1 \cdots G^{10})$), 50 genes are assumed to be possible modulators M of the TF, but only 20 of them are the true modulators. In addition I assumed that 10 of the remaining 30 false modulators are indeed unknown targets of the TF and hence co-regulated with the TF itself, thus making it harder for the methods to distinguish them from the true modulators.

The second dataset D2 is constructed in a similar way but with a larger number of genes. Specifically D2 consists of 760 genes. As for dataset D1, only 10 genes are assumed to be the known targets of the TF (i.e. are $\mathbf{G} = (G^1 \cdots G^{10})$), whereas the remaining 750 genes are assumed to be possible modulators M of the TF, with only 50 of them being the true modulators.

In order to generate the 100 GEPs in each dataset I proceeded as follows. For clarity, I will describe only the generation of the D1 dataset, since the D2 dataset was generated in a similar manner. I divided the 100 GEPs in two subsets of 50 GEPs each: one subset in which the TF is active (hence the true modulators are expressed and the targets are co-regulated) and one in which the TF is inactive (hence the true modulators are not expressed and the targets are independent). In the first subset, I simulated the expression profiles of the 10 TF's *targets* (known and unknown) using a multivariate Gaussian distribution with a covariance matrix with off-diagonal elements equal to $\rho = 0.8$, mean equal to 0 and a random standard deviation in the interval $]0,0.5[$. In the second subset, the expression values of the TF's targets are sampled again from the same multivariate Gaussian distribution but with a diagonal covariance matrix (i.e. off-diagonal element are set to 0). The expression profiles of the 20 *false modulators* were generated using a normal distribution with mean 0 and random standard deviation in

the interval $]0,0.5[$ in both the subsets. Finally, in order to simulate the expression profiles of the *true modulators* I followed this strategy: in the first subset (i.e the 50 samples where the targets are co-regulated), I sampled from a normal distribution $N(1,0.1)$ (i.e. with average expression around 1), on the contrary in the second subset (i.e the 50 samples where the targets are independent), I used a normal distribution $N(0,0.1)$ (i.e. with average expression equal to 0).

Table 5 - Description of the parameters used in each "in-silico" dataset. The column *Dataset* indicate the name of the dataset. The column $\boldsymbol{\rho}$ the strength of the dependence of the targets $\boldsymbol{\rho} \in [\mathbf{0}, \mathbf{1}]$. The column *#Targets* represents the number of targets of the TF used as input for the tested methods. The column *#True Mod.* contains the number of modulators of the TF present in the dataset. The column *#False Mod*. contains the number of genes in the dataset that are not modulator of the TF. Finally, the column *#Uknw. Targ*. contains the number of unknown targets of the TF and hence co-regulated with the TF itself, thus making it harder for the methods to distinguish them from the true modulators (these are thus themselves false modulators).

Dataset	ρ	#Targets	#True Mod.	#False Mod.	#Uknw. Targ.
D1	0.8	10	20	30	10
D2	0.6	10	50	700	500

5.4.3 Comparison of the "In-silico" performance of the different methods.

I compared my method based on Differential Multi-Information with the other two algorithm presented in the previous sections. The DMI was applied as described in section 5.1 and 5.2 using a number of bin for discretization of the GEPs equal to 2 (unless specified).

The PPV-sensitivity curve for the D1 dataset (Table 5) is reported in Figure 29 where the Differential Multi-Information (ΔI_α) method archive the best performance, ranking the 20 modulators in the top 20 positions. The results were ranked in descending order according to the value of the estimated ΔI_α. We observe that when using the ΔI_α method the value of the ΔI_α for the false modulators is nearly 0 but it is very large for the true modulators.

The results for the dataset D2 are reported in Figure 30 where the ΔI_α method still achieves the best performance ranking the 50 true modulators in the top 50 positions. In this case as well the ΔI_α of the true modulators is very large compared with the $\Delta I_\alpha$of the false modulators that is close to 0.

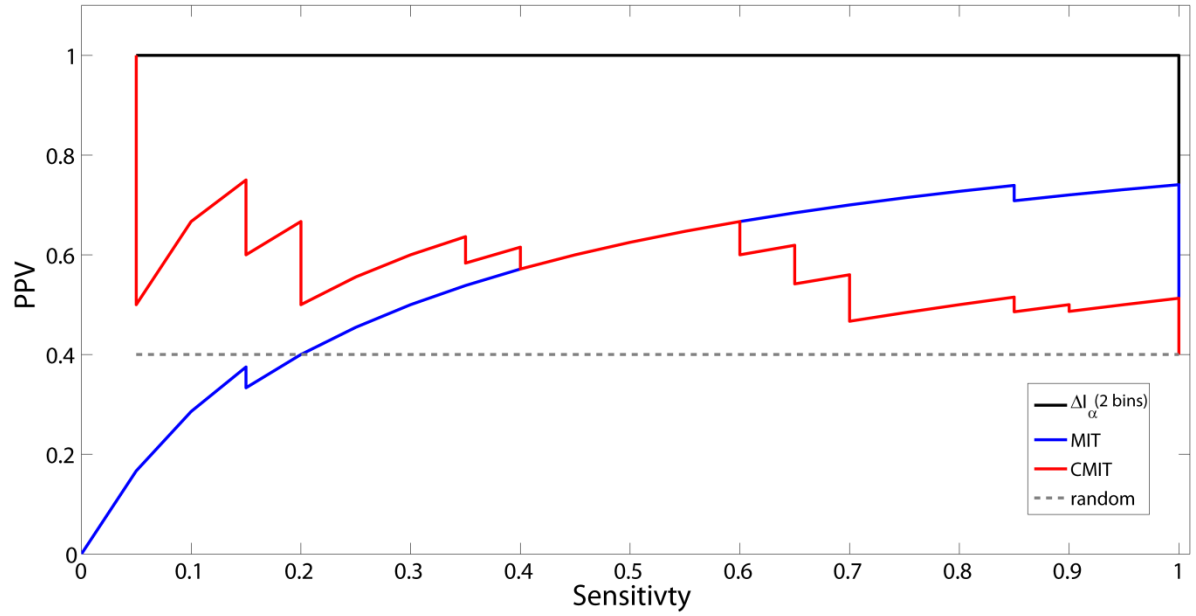


Figure 29 – PPV-sensitivity curve using the D1 dataset for 3 tested methods: $\boldsymbol{\Delta I_\alpha}$, MIT and CMIT. The $\boldsymbol{\Delta I_\alpha}$ method achieves the best performance ranking the 20 real regulators in the top 20 positions. The random PPV value is also shown for comparison (black dotted line).

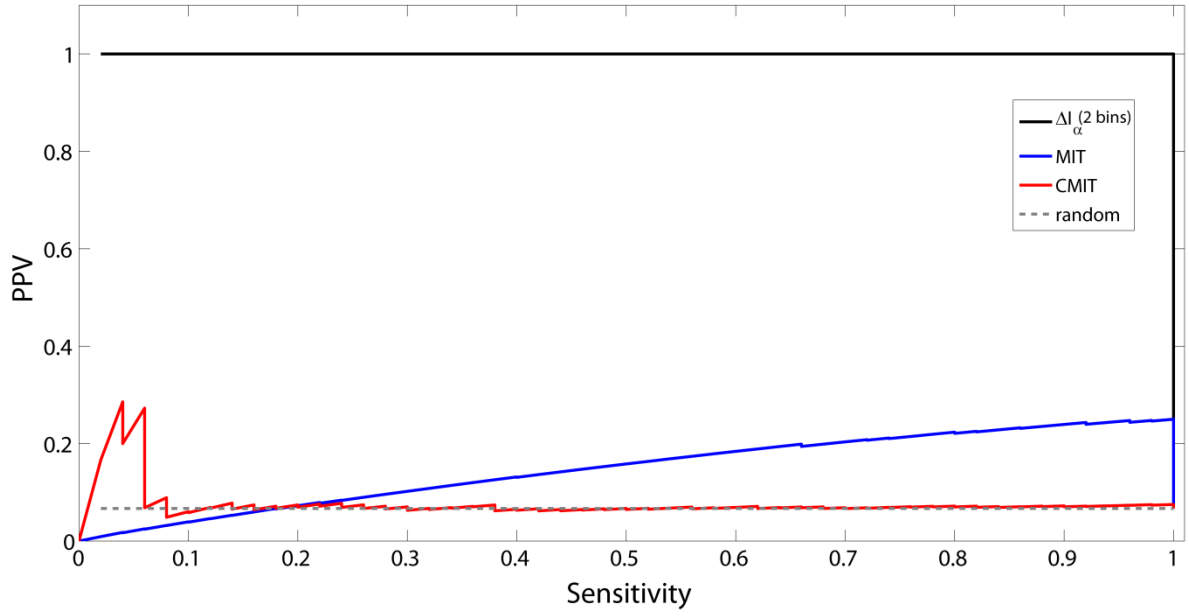


Figure 30 - PPV-sensitivity curve using the D2 dataset for 3 tested methods: $\boldsymbol{\Delta I_\alpha}$, MIT and CMIT. The $\boldsymbol{\Delta I_\alpha}$ method archives the maximal performance ranking the 50 real regulators in the top 50 positions. Also the random value has shown as comparison (black dotted line).

I also simulated 4 additional datasets with the same parameters of D2 (Table 5) but with different number of experiments in the two subsets (i.e. the one in which the targets are dependent and the one in which the targets are not). Specifically, in these 4 datasets the number of GEPs in the first subset, in which the TF's targets are dependent, is either 30, 40, 60 or 70 out of 100 GEPs. The PPV-sensitivity curves for all the three methods applied to these 4 dataset are reported in Figure 31. Again the DMI method still achieves the best performance.

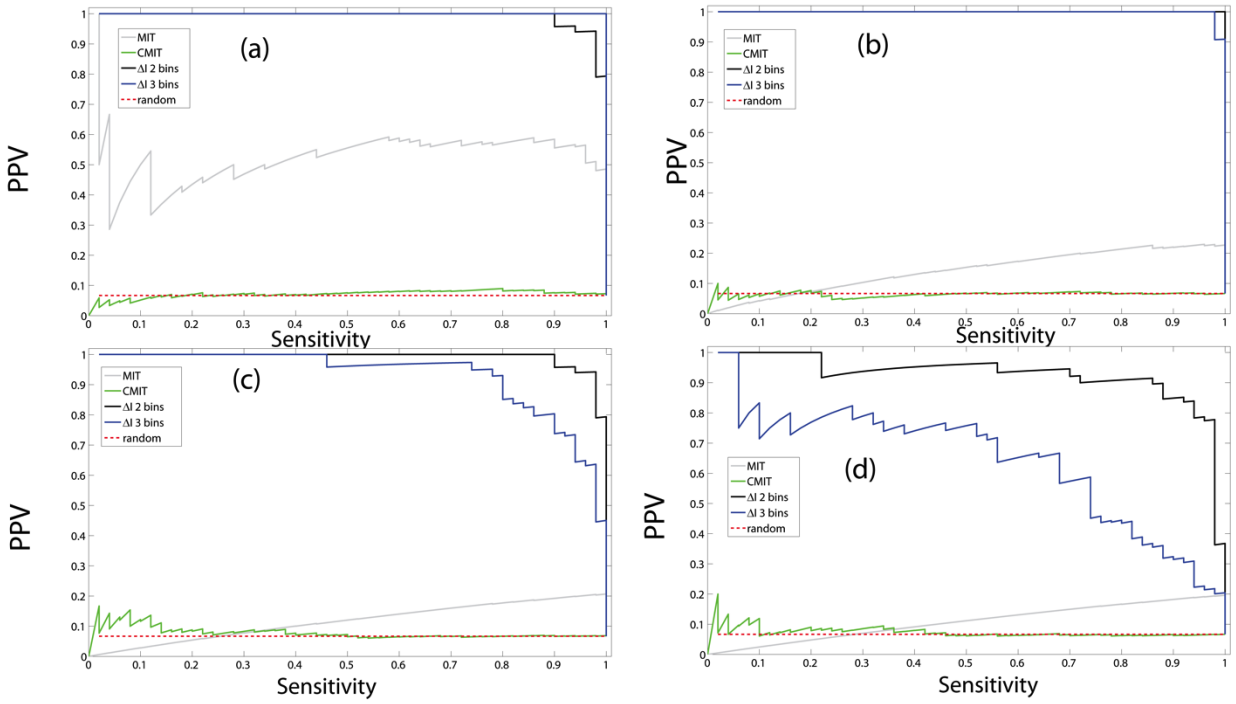


Figure 31 – PPV-sensitivity curve using "in-silico" dataset D2 where the targets are dependent in the 30 **(a)**, 40 **(b)**, 60 **(c)** and 70 **(d)** of the experiments.

Finally, I also computed the p-value using the permutation test as described in the subsection 5.1.2. In this case I first applied a threshold of p<0.05 to deem a prediction as significant and then I ranked only the significant predictions according to ΔI_α in descending order as before. The PPV-sensitivity curves are shown in Figure 32. In Table 6, I reported the Area Under the Curve computed from the PPV-sensitivity curves. For completeness I also repeated the DMI approach using three discretization bins rather then 2 and reported the results in Table 2.

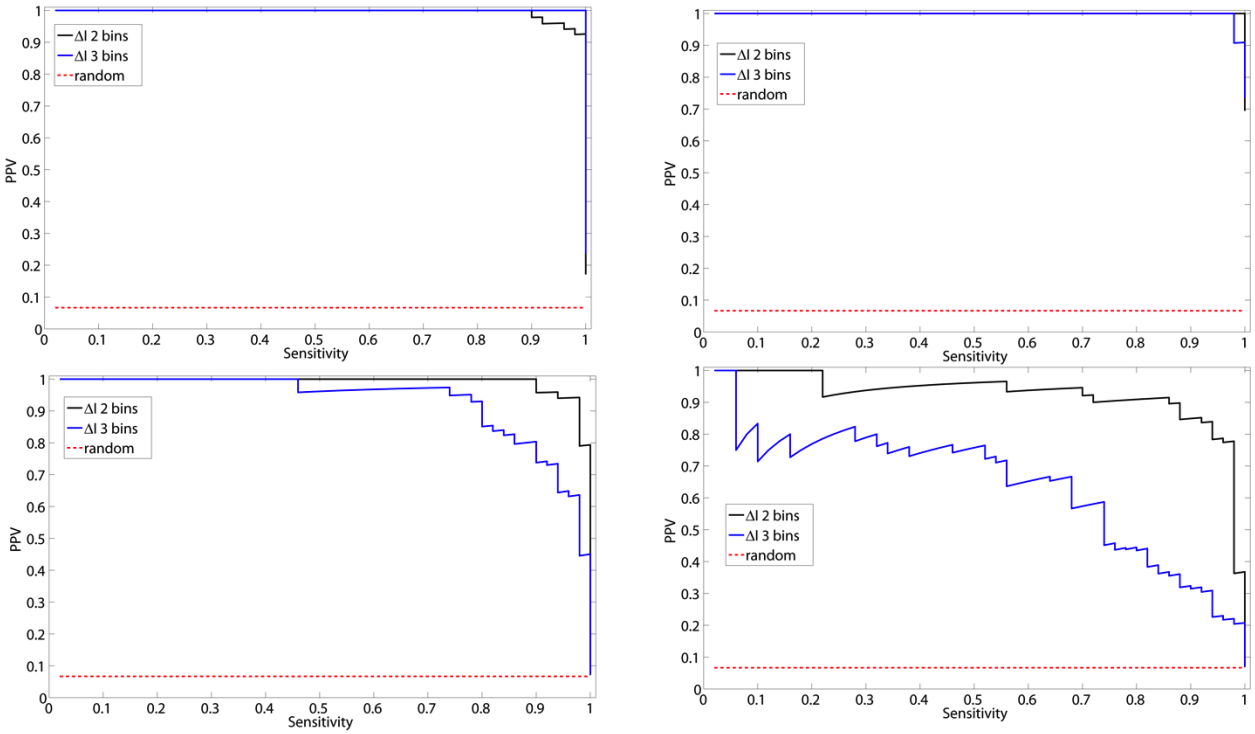


Figure 32 - PPV-sensitivity curve using "in-silico" dataset D2 where the targets are dependent in the 30 **(a)**, 40 **(b)**, 60 **(c)** and 70 **(d)** out of the 100 GEPs. Only modulators with p-value = 0 have been selected.

Table 6 – Area Under the Curve (AUC) of the PPV-sensitivity curves for the DMI method using the p-value to cut the results, with either 2 or 3 bins used for discretization of the GEPs according to the modulator expression level.

Targets dependence	AUC % 2 bins	AUC % 3 bins
30%	98.1%	**100.0%**
40%	**100.0%**	98.8%
60%	**98.2%**	92.4%
70%	**92.0%**	64.0%

5.5 Discussion and Conclusions

In this Chapter I presented and tested an original method based on Multi-Information to identify post-translational “modulators” of a transcription factor from gene expression profiles. A modulator could be for example a kinase or phosphatase able to activate/deactivate the transcription factor. This method is based on the estimation of multi-information among the targets of the transcription factor and it uses a permutation test to assess the significance of a possible modulator. Using an in-silico dataset, I evaluated and compared the performance of the proposed method.

Chapter 6

Evaluation of the DMI method

This chapter describes the evaluation of the DMI method using real experimental datasets. The chapter begins with a brief description of the experimental dataset, then continues with the presentation of the results obtained by testing the performance of DMI method on a set of transcription factors of interest, including a brief description of each of them.

6.1 Description of the experimental dataset

As presented in the Chapter 5, the DMI method requires in input a list of targets G of a TF, a set of Gene Expression Profiles (GEPs) and a list of possible modulators M to test. In order to evaluate the performance of DMI when applied to real experimental data, I chose 7 Transcription Factors, with known transcriptional targets, as listed in

Table 7, and whose activity is regulated by a set of well-characterised kinases. For each TF I first collected its known targets (as detailed in the following sections), I then retrieved the kinases modulating the TF activity from PhosphoPOINT [135] and BioGrid [136] databases. I thus obtained a "Golden Standard" for each TF consisting of experimentally verified kinases.

I then exploited this Golden Standard to assess the performance of the DMI method in correctly identifying the modulators each of the TF. To this end, I applied DMI to a compendium of 5,372 high quality human GEPs representing 369 different cell and tissue types, disease states and cell lines, described in [137]. GEPs were measured using the Affymetrix HG-U133A platform and normalized using the Robust Multi-array Average (RMA) normalization as implemented in the R package Bioconductor [72].

As list of possible modulators M, I collected the 491 kinases present on HG-U133A platform and associated the Gene Ontology (GO) molecular function term "protein kinase activity".

Table 7 – List of the 7 transcription factors tested including their official gene symbol and their full name.

Gene Symbol	Name
TP53	tumor protein p53
MYC	myelocytomatosis oncogene
STAT3	signal transducer and activator of transcription 3
SMAD3	SMAD family member 3
GATA2	GATA-binding protein 2
ELK1	ETS domain-containing protein Elk-1
ETS1	v-ets erythroblastosis virus E26 oncogene homolog 1 (avian)

6.2 Identification of kinases regulating P53

P53 protein is a tumor suppressor Transcription Factor protein encoded by TP53 gene located on the short arm of chromosome 17 (17p13.1). Since it was discovered 30 years ago [115] as a cellular partner of SV40 Large Tumor Antigen, the oncoprotein of this tumor virus, more than 50,000 PubMed-listed publications have been written on it. The notion of pivotal tumor suppressor and the fact that it constitutes a natural anti-cancer defense for our body have attracted a lot of interest in the study of this gene. Disrupting of TP53 functions leads to reduced tumor suppression, indeed more than 50% of human tumors contain a mutation or deletion of the TP53 gene [138].

The protective function carried out by the P53 protein is achieved through several different mechanisms including regulation of apoptosis, genomic stability, and inhibition of angiogenesis. When the DNA has sustained damage, P53 transcribes a set of genes whose protein products are responsible for DNA repair. To perform this task, P53 can also induce growth arrest by halting the cell cycle for long enough to allow DNA repair. If the DNA cannot be repaired, an apoptosis program is initiated to eliminate the DNA-damaged cell [115].

One of the main open challenges in p53 biology is to understand how p53 is able to discriminate which targets must be activated or repressed to obtain a specific cellular outcome (repair verus apoptosis). In the recent years different models have been proposed. For example, some researchers suggested that apoptosis is triggered only when the amount of active P53 protein present in a cell

reaches a threshold level [139]. Another model is one in which P53 target gene selection is determined mainly by co-factors, P53-binding factors and post-translational modifications [140].

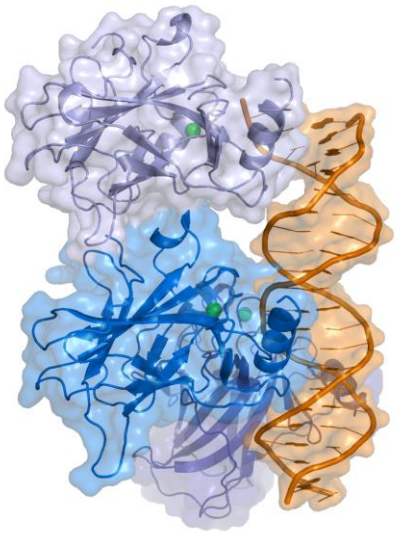


Figure 33 - 3D structure of P53 human protein (source Wikipedia)

A lot is known on how P53 activity is tightly regulated in the cell by co-factors [141] as well as post-translational modifications [142]. Several Serine and Threonine sites in the P53 protein are targeted for phosphorylation in response to a myriad of stress types, which include, but are not limited to, DNA damage, oxidative stress, osmotic shock and ribonucleotide depletion.

In order to test the performance of DMI in identifying kinases regulating P53 activity, I selected a set of 34 bona fide transcriptional (Table 8) targets reported in [94]. I also identified 69 kinases known to interact with P53 protein (Table 8) as described in section 6.1. I then applied DMI on the set of 5,372 GEPs and tested each of the 491 kinases using a different discretization bins and with or without permutation tests using 1000 permutations.

Table 8 – List of 34 bona fide targets used as input for our method and list of know kinases interact with P53 protein used as a golden standard.

P53	**OFFICIAL GENE SYMBOLS**
TARGETS	BDKRB2, BTG2, CCNG1, CD82, CDC25C, CDKN1A, CRYZ, CTSD, CX3CL1, DKK1, EGFR, FAS, GADD45A, GML, HGF, IER3, IGFBP3, MDM2, MET, MMP2, ODC1, PCBP4, PLK2, RB1, S100A2, SCARA3, SCD, SERPINE1, SFN, SLC38A2, TAP1, TGFA, THBS2, TP53I3
KINASES	ABL1,ALK,ATM,ATR,AURKA,AURKB,BCR,BMX,CCNH,CDK1,CDK2,CDK5,CDK8,CDK9,CDKN1A, CHEK1,CHEK2,CSNK1A1,CSNK1D,CSNK1E,CSNK1G1,CSNK1G2,CSNK1G3,CSNK2A1,CSNK2A2, CSNK2B,DAPK1,DAPK3,EIF2AK2,EPHA3,ERBB4,ERCC2,ERCC3,GSK3A,GSK3B,GTF2H1,HIPK1, HIPK2,HIPK3,IGF1R,IKBKB,LYN,MAP3K1,MAPK1,MAPK10,MAPK14,MAPK3,MAPK8,MAPK9, MNAT1,PLK1,PLK3,PPP4C,PRKCA,PRKCD,PRKD1,PRKDC,PTK2,RYK,SMG1,STK11,STK4,TAF1,T RIM24,TRIM27,TRIM28,TTK,VRK1,VRK2

The resulting PPV-Sensitivity curves are reported in Figure 34, with the performances of a random algorithm shown for comparison. As shown in from Figure 34 the best number of bins to use for the discretization step seems to be 3, where we achieve the best performance with an Area Under the Curve (AUC) of about 27.2% (Table 9) using the p-value.

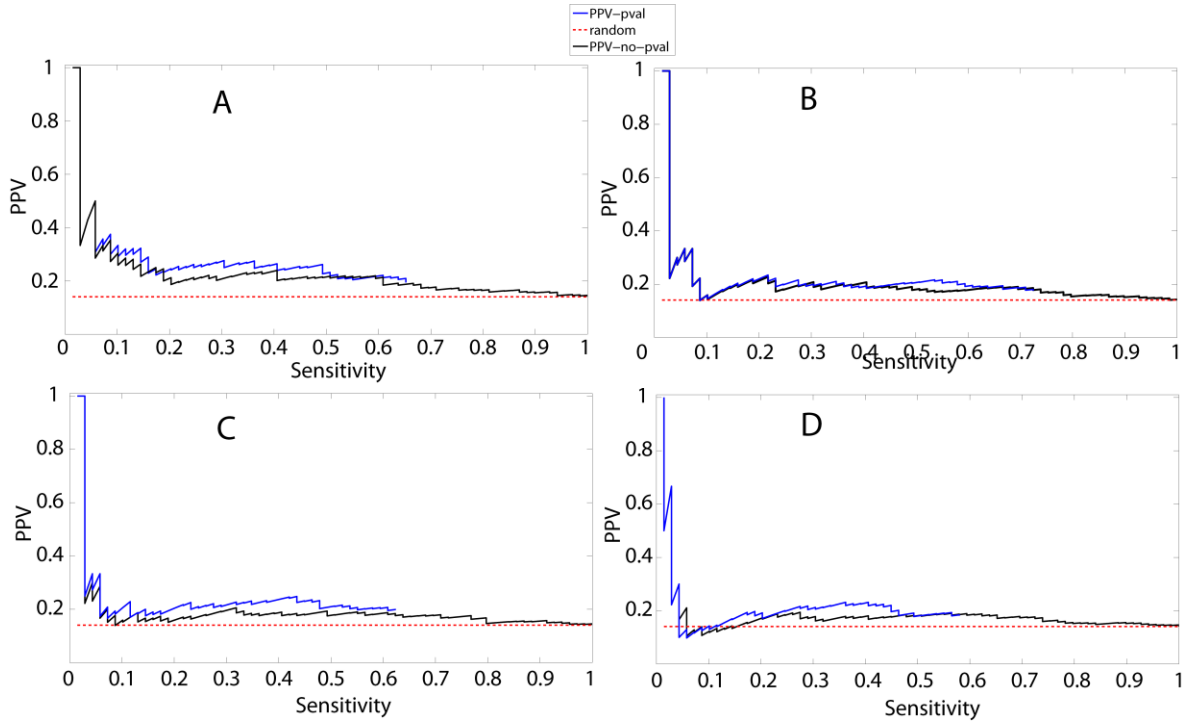


Figure 34 – PPV sensitivity curve and relative Area Under the Curve (AUC) for the identification of post-translational modulators of P53 using different number of bins for the expression discretization of the modulator. Red dotted line represents the performance of a random algorithm. (**a**) 3 bin discretization. (**b**) 5 bin discretization. (**c**) 7 bin discretization. (**d**) 10 bin discretization.

Table 9 - Area Under the Curves (AUC) of PPV-Sensitivity for the identification of post-translational modulators of the P53 transcription factor. *AUC (ΔI rank)* column is the AUC of the PPV-sensitivity curve using only the ΔI values to rank the modulators. *AUC (pval)* column is the AUC of the PPV-sensitivity curve removing the non-significant values of ΔI using the p-value. *AUC_norm (pval)* column is the normalized value of the AUC respect to the maximal value it can achieve: $\mathbf{1} \cdot \mathbf{max}(\boldsymbol{sensitivity})$, where $\mathbf{max}(\boldsymbol{sensitivity})$ is the maximal sensitivity can be achieved after removing non-significant values. In bold the best AUC for the different discretization strategies.

Disc. Type	**AUC (ΔI rank)**	**AUC (pval)**	**AUC_norm (pval)**
3 bins	21.7%	17.7%	**27.2%**
5 bins	19.2%	15.5 %	**21.4%**
7 bins	18.3%	14.2%	**22.8%**
10 bins	17.0%	11.3%	**19.5%**

6.3 Identification of kinases regulating MYC

MYC is probably the most important and studied oncogene with more than 19,000 PubMed publications associated to it. MYC protein belongs to the family of transcription factors containing bHLH/LZ (basic Helix-Loop-Helix Leucine Zipper) domain. MYC protein, through its bHLH domain can bind to DNA, while the leucine zipper domain allows the dimerization with its partner MAX, another bHLH transcription factor.

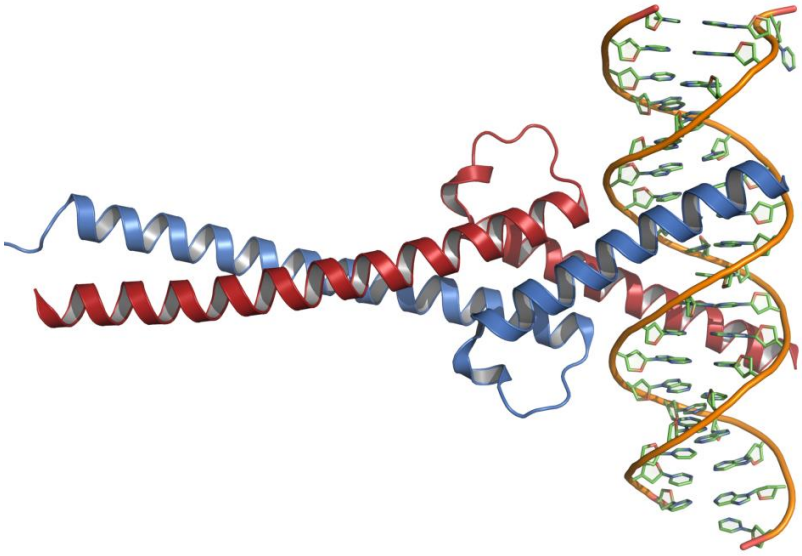


Figure 35 - 3D structure of MYC human protein (source Wikipedia).

When it was discovered about 25 years ago, it changed the definition of oncogene [143]. Before MYC, oncogenes gain-of-function was considered to be caused only by somatic mutations in their coding sequence. With MYC new mechanisms of oncogenes activation, including gene amplification, chromosomal activation and insertional mutagenesis have been discovered. With the discovery of MYC, probably, the most important breakthrough was the comprehension that MYC dysregulation is not due only to mutations and rearrangements of the MYC genomic locus, but also to dysregulation of the control mechanisms targeting its expression. Indeed MYC expression is tightly controlled in cell. Several efforts have been directed in understanding what is the normal expression pattern of MYC and how its

expression is regulated. MYC was identified as the first eukaryotic cellular gene to be regulated by transcription elongation control [144-147] and loss of this control is evident in cancer. Different studies in mouse models have revealed that abnormalities through direct or indirect mechanisms affecting MYC mRNA expression are able to drive cancer development [148, 149]. Regarding post-translational control, phosphopeptide analysis revealed that specific serine and threonine residues of MYC are phosphorylated in vivo [150]. Amati and colleagues showed that MYC–MAX heterodimerization is essential for MYC transformation [151].

Recent studies based on chromatin immunoprecipitation assays (ChIP) have helped researchers to better identify MYC transcriptipnal targets. In particular these studies have demonstrated that MYC can be considered as a "global transcriptional regulator", since it is able to bind approximately 10-15% both coding and no-coding regions of the genome [152, 153]. Multiple pathways are regulated by MYC in order to drive any one of a plethora of biological programmes, including cell proliferation [150] and cell differentiation [154-158].

I conclude this brief overview of MYC by quoting [143]: "MYC is downstream of many signal transduction pathways, functioning as a central hub that integrates multiple intracellular and extracellular cues. MYC then processes and interprets these instructions, much like the central processing unit of a computer".

In order to apply the DMI method to identify kinases regulating MYC activity, I retrieved 68 experimentally verified MYC transcriptipnal targets (Table 10) from the Myc target gene database [159]. The "Golden Standard" contains 59 know kinases (Table 10) interacting with MYC protein and generated as described in section 6.1.

Table 10 - List of 68 experimentally verified targets of MYC and know kinases interact with MYC protein used as a golden standard.

MYC	**OFFICIAL GENE SYMBOLS**
TARGETS	ACP5, AKAP1, APEX1, APP, ARPC4, BAX, BCAT1, CCKBR, CDC25C, CDKN1B, CKS2, CSTB, CTSC, DDX10, DDX18, DDX5, DKC1, EIF2S1, EIF4A1, ENO1, FASN, H2AFZ, HMOX1, HSPE1, ID3, IMPA2, LAMB2, LAMP1, MAT2A, MSH2, MSN, NAP1L1, NPM1, ODC1, PA2G4, PCNA, PHB, POLD2, PPAT, PPIA, PPID, PREP, PRPS2, PSMB1, PTPN1, PYCR1, RARA, RPL10, RPL13, RPL22, RPL27, RPL5, RPS16, RPS19, RPS20, RPS5, SERPINE1, SHMT1, SNRPD3, SRM, TERT, TFRC, TGFB1, TOP1, TP53, TXN, UCHL1, VHL

KINASES	ADRBK2, ALPK1, BCR, BMPR1A, BTK, CAMK1G, CAMK2G, CCNH, CDK12, CDK2, CDK4, CDK8, CDK9, CSNK1E, CSNK2A1, CSNK2A2, ERCC3, GRK1, GSK3A, GSK3B, HCK, IGF2R, IKBKAP, IRAK1, LATS1, LIMK2, MAP2K1, MAP2K3, MAP2K7, MAP3K13, MAPK1, MAPK3, MAPK7, MAPK8, MATK, MYLK, NEK2, NEK9, NTRK1, PAK6, PBK, PDK1, PIM1, PKN1, PLK1, POLR2B, POLR2E, POLR2I, PRKDC, RAF1, SPEG, SQSTM1, TIE1, TRIB1, TRIM28, TXK, WEE1, WNK1, YES1

As in the case of p53, I applied DMI to the set of 5,372 GEPs using the list of targets listed in Table 4 and tested each of the 491 kinases using a different number of discretization bins and with or without permutation tests using 1000 permutations. The resulting PPV-Sensitivity curves for the identification of the kinases regulating MYC are reported in Figure 36. Also the performances of a random algorithm is shown for comparison.

In this case the AUC is lower than in the case of p53, however also in this case the best number of bins to use for the discretisaiton seems to be three, even if the AUC is low in this case, nevertheless the PPV achieve its maximum with three bins (Figure 4A). The AUCs for all the PPV-sensitivity curves are reported in Table 11.

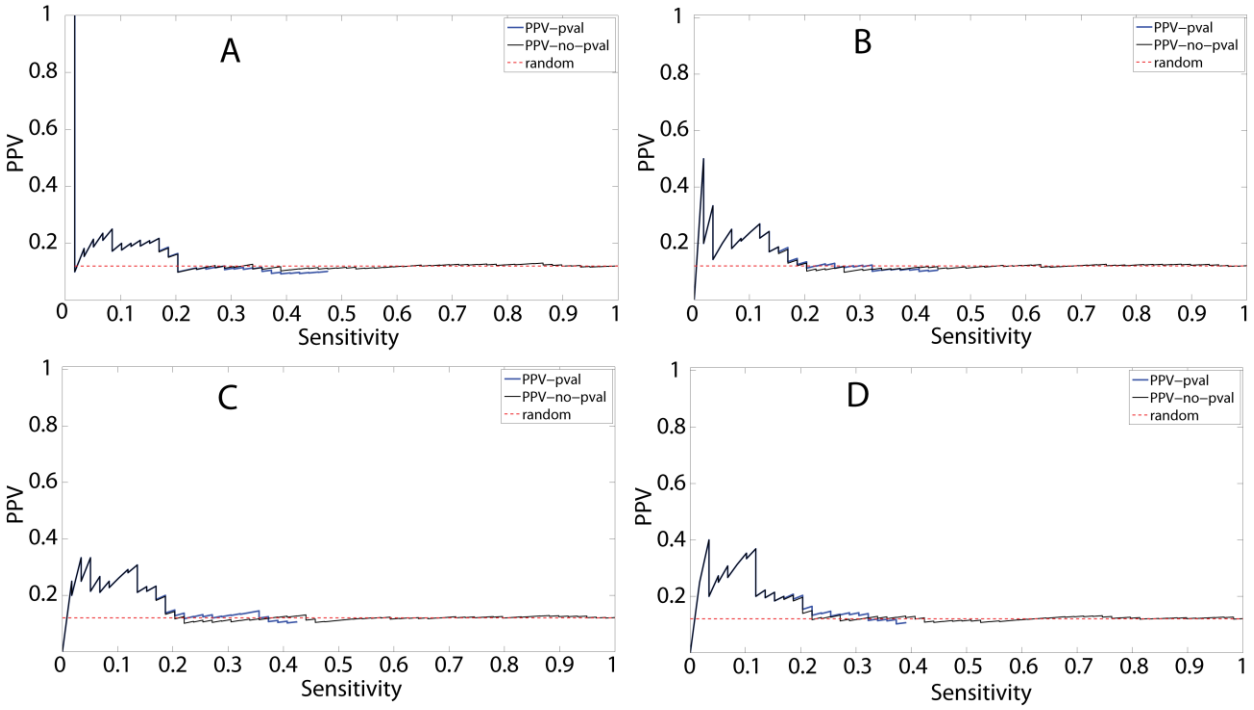


Figure 36 – PPV-sensitivity curve and relative Area Under the Curve (AUC) for the identification of post-translational modulators of MYC using different number of bins for the discretization of the modulator expression. Red dotted line represents the performance of a random algorithm. (**a**) 3 bin discretization. (**b**) 5 bin discretization. (**c**) 7 bin discretization. (**d**) 10 bin discretization.

Table 11 - Area Under the Curves (AUC) of PPV-Sensitivity for the identification of post-translational modulators of the P53 transcription factor. *AUC (*$\Delta \boldsymbol{I}$ *rank)* column is the AUC of the PPV-sensitivity curve using only the $\Delta \boldsymbol{I}$ values to rank the modulators. *AUC (pval)* column is the AUC of the PPV-sensitivity curve removing the non-significant values of $\Delta \boldsymbol{I}$ using the p-value. *AUC_norm (pval)* column is the normalized value of the AUC respect to the maximal value it can achieve: $\mathbf{1} \cdot \mathbf{max}(\boldsymbol{sensitivity})$, where $\mathbf{max}(\boldsymbol{sensitivity})$ is the maximal sensitivity can be achieved after removing non-significant values. In bold the best AUC for the different discretization strategies.

Disc. Type	**AUC (**$\Delta \boldsymbol{I}$ **rank)**	**AUC (pval)**	**AUC_norm (pval)**
3 bins	12.9%	6.4%	**13.5%**
5 bins	13.4%	6.8%	**15.5%**
7 bins	14.0%	7.3%	**17.4%**
10 bins	14.6%	7.4%	**19.0%**

6.4 Identification of kinases regulating STAT3

Signal transducer and activator of transcription 3 (STAT3) is a transcription factor which in humans is encoded by the STAT3 gene. STAT3 protein belongs to the family of STAT protein, a family of proteins usually phosphorylated by receptor-associated kinases. Characteristic of this family proteins is that when phosphorylated they form homo or heterodimers that translocate to the nucleus, where they act as transcription activators. STAT3 is the major member of STAT family consisting of STAT1, STAT2, STAT3, STAT4, STAT5α, STAT5β, and STAT6, plays important roles in cell differentiation and proliferation. In a variety of human cancers, constitutive activation of STAT3 is sufficient to induce tumor formation [160, 161].

I collected 10 bona fide targets of STAT3 combining expression data and ChIp data from two different studies [162, 163]. I selected the 10 genes identified in common in both studies. Known kinases regulating STAT3 activity were collected as described before for a total of 40 kinases.

Table 12 - List of 10 "bona fide" collected targets of STAT3 and the 40 known kinases interacting with STAT3 protein used as a golden standard

STAT3	**OFFICIAL GENE SYMBOLS**
TARGETS	ABCA1, ADM, BCL6, CEACAM1, CXCL2, OAS2, OASL, SERPINA3, SERPINB3, SERPINE2
KINASES	ALK, BMX, CCND1, CDK9, CDKN1A, EGFR, EIF2AK2, EPHA5, ERBB2, FER, FES, FGFR1, FGFR2, FGFR3, FGFR4, FLT1, HCK, IGF1R, IRAK1, JAK1, JAK2, JAK3, LCK, MAP3K7, MAPK1, MAPK3, MAPK8, MET, MTOR, NLK, PDGFRA, PDGFRB, PRKCD, PRKCZ, PTK2, PTK2B, RET, RPS6KA5, SRC, SYK

PPV-Sensitivity curves for the identification of the kinases regulating MYC are reported in Figure 37. Also the performances of a random algorithm are showed as a comparison.

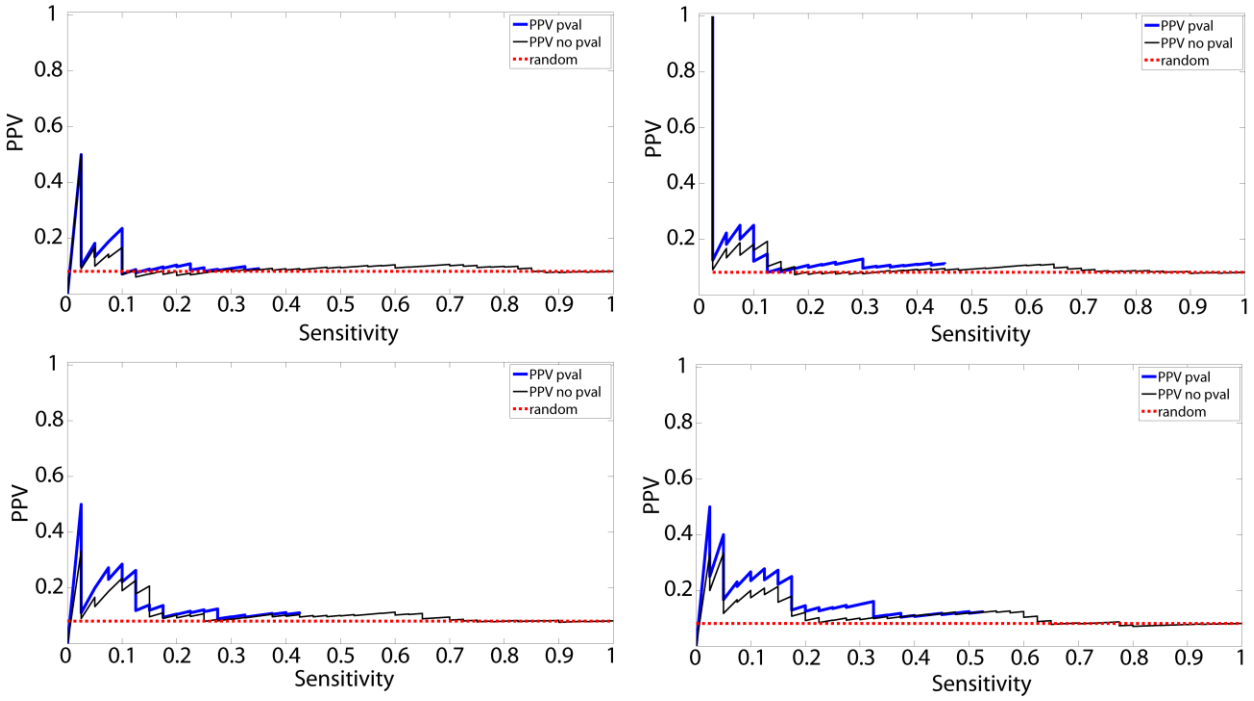


Figure 37 - PPV-sensitivity curve and relative Area Under the Curve (AUC) for the identification of post-translational modulators of STAT3 using different number of bins for the discretization of the modulator expression. Red dotted line represents the performance of a random algorithm. (a) 3 bin discretization. (b) 5 bin discretization. (c) 7 bin discretization. (d) 10 bin discretization**.**

Table 13 - Area Under the Curves (AUC) of PPV-Sensitivity for the identification of post-translational modulators of the P53 transcription factor. *AUC (*$\Delta \boldsymbol{I}$ *rank)* column is the AUC of the PPV-sensitivity curve using only the $\Delta \boldsymbol{I}$ values to rank the modulators. *AUC (pval)* column is the AUC of the PPV-sensitivity curve removing the non-significant values of $\Delta \boldsymbol{I}$ using the p-value. *AUC_norm (pval)* column is the normalized value of the AUC respect to the maximal value it can achieve: $\mathbf{1} \cdot \mathbf{max}(\boldsymbol{sensitivity})$, where $\mathbf{max}(\boldsymbol{sensitivity})$ is the maximal sensitivity can be achieved after removing non-significant values. In bold the best AUC for the different discretization strategies.

Disc. Type	**AUC (**$\Delta \boldsymbol{I}$ **rank)**	**AUC (pval)**	**AUC_norm (pval)**
3 bins	9.6%	4.2%	**11.9%**
5 bins	9.3%	5.3%	**11.8%**
7 bins	10.6%	6.2%	**14.5%**
10 bins	11.2%	8.9%	**16.8%**

6.5 DMI performance on additional transcription factors.

As discussed in the Chapter 5, the DMI method requires as input a list of target genes of a TF, a set of Gene Expression Profiles (GEPs) and a list of possible modulators to test.

In order to test the performance of DMI on additional transcription factors, I selected four transcription factors (Table 14) and their putative targets by combining data from MsigDb [164] and Chip [163] databases that are known to be regulated through post-translational mechanisms.

Table 14 – List of the four transcription factors selected to test the DMI method. The columns report the official gene symbol, the name, the number of gene targets used as input for DMI and the number of kinases known to modulate the TF.

Gene Symbol	Name	# of Targ.	# of Kin.
SMAD3	SMAD family member 3	43	26
GATA2	GATA-binding protein 2	14	2
ELK1	ETS domain-containing protein Elk-1	177	10
ETS1	v-ets erythroblastosis virus E26 oncogene homolog 1 (avian)	12	6

SMAD3 is a member of the SMAD family proteins belonging to the transforming growth factor beta (TGF-β) superfamily of modulators [165]. One of the first observations showing SMAD proteins at downstream of TGF-β pathway was the capability of SMAD proteins to accumulate in the nucleus in response to TGFβ or BMP [166, 167]. SMAD proteins undergo a constant process of nucleocytoplasmic shuttling mediated by phosphorylation [168, 169] and dephosphorylation [170] events.

GATA2 is a member of GATA family of transcription factors, a family of evolutionarily conserved proteins playing a crucial role in the development and differentiation of eukaryotic organisms expressed particularly in hematopoietic cell lineages [171]. There are several studies on the characteristics and functions of the principal member (GATA1) of this transcription factor family, instead little is known about GATA2 [172]. Variations in the transcriptional activity of GATA2 transcription factor are also modulated by post-translational modifications [173] affecting nuclear localization, DNA-binding, protein stability, and/or cofactor recruitment.

ELK1 is transcription factor member of the ETS family and of the ternary complex factor (TCF) subfamily. ELK1 seems play in many contexts [174], including long-term memory formation, drug

addiction, Alzheimer's disease, Down syndrome, breast cancer, and depression. The protein activity of ELK1 is strictly modulated through post-translational mechanisms. For example, it is activated by phosphorylation by three classes of MAP kinases, ERK, JNK, and p38 [175, 176] and it is repressed through dephosphorylation by the Protein phosphatase 2B (PP2B) [177, 178].

Also the ETS1 proto-oncoprotein [179] is a member of the ETS family of transcription factors. The DNA binding activity of ETS1 is tightly modulated by kinases and transcription factors [179]. ETS1 is expressed by many cell types and in hematopoietic cells it is involved in the regulation of cellular differentiation.

PPV-Sensitivity curves for the identification of the kinases regulating these four transcription factors using the 3 bins discretization strategy are reported in Figure 38, while their AUCs are reported in Table 15. In Figure 38, also the performances of a random algorithm are showed as a comparison.

Table 15 – Area Under the Curves (AUC) of PPV-Sensitivity for the identification of post-translational modulators of the P53 transcription factor. *AUC (*$\Delta \boldsymbol{I}$ *rank)* column is the AUC of the PPV-sensitivity curve using only the $\Delta \boldsymbol{I}$ values to rank the modulators. *AUC (pval)* column is the AUC of the PPV-sensitivity curve removing the non-significant values of $\Delta \boldsymbol{I}$ using the p-value. *AUC_norm (pval)* column is the normalized value of the AUC respect to the maximal value it can achieve: $\mathbf{1} \cdot \mathbf{max}(\boldsymbol{sensitivity})$, where $\mathbf{max}(\boldsymbol{sensitivity})$ is the maximal sensitivity can be achieved after removing non-significant values. In bold the best AUC for the different discretization strategies.

Name	AUC ($\Delta \boldsymbol{I}$ rank)	AUC (pval)	AUC_norm (pval)
SMAD3	12.2%	7.3%	**15.9%**
GATA2	**6.5**	**6.5**	**6.5**
ELK1	11.5%	11%	**13.6%**
ETS1	6.8%	5.4	**11%**

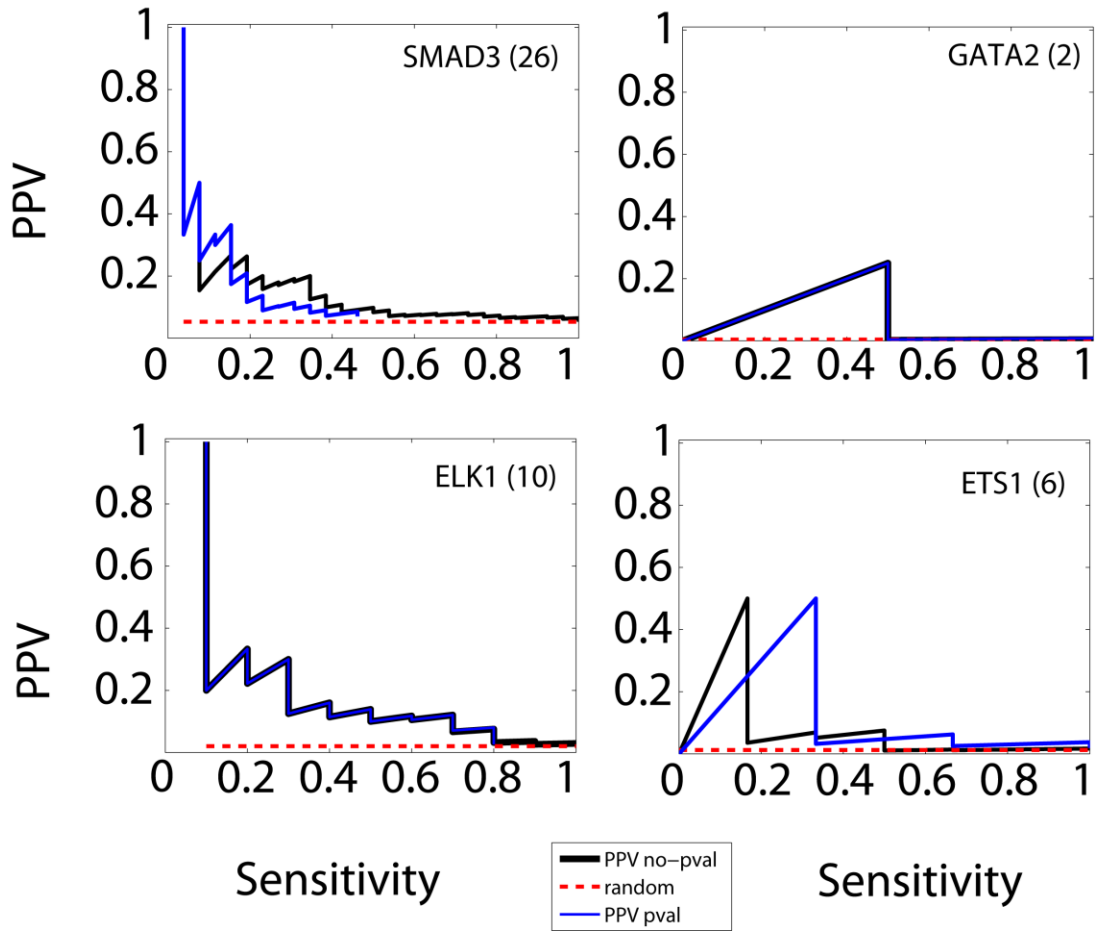


Figure 38 – PPV-sensitivity curve for the 4 transcription factors SMAD3, GATA2 ELK1 and ETS1 and in parentheses the number of know kinases interacting with them present in the "Golden Standard".

6.6 Discussion and Conclusions

In this Chapter, I tested the DMI method on a set of transcription factors for which I was able to collect a bona fide set of targets and a set of known kinases regulating their activity to be used as a "Golden Standard". I computed for each one of the collected transcription factors, the PPV-sensitivity curve using as the "Golden Standard" obtained by mining PhosphoPOINT and BioGrid databases containing experimentally verified protein interactions. It is important to remember that it is not possible have a complete "Golden Standard" containing all of the real modulators for a transcription factor of interest, because of the partial knowledge inherent in biology. Nevertheless, in these tests the method

performed very well achieving a high precision in all of the tests (Table 9, Table 11, Table 13 and Table 15).

Chapter 7

A case of study: Identification of TFEB modulators

In this Chapter, I tested the DMI algorithm presented in the Chapter 5 and 6 for the identification of phosphatases interacting with the TFEB transcription factor. The Chapter starts with a brief introduction to TFEB biology, followed by an overview of High Content Screening (HCS) for the identification of modulators of a transcription factor of interest. I then discuss the application of DMI to this case study and the comparison of predicted modulators using the DMI computational approach with the results obtained using HCS.

7.1 Introduction to TFEB

Lysosomes are membrane-delimited organelles present in all mammalian cells except red blood cells. They are engaged in the degradation of macromolecules delivered from the cells own cytoplasm (autophagy) as well as materials taken up from the extracellular space. Malfunctions in lysosomes lead to Lysosomal Storage Disorders (LSDs) a class of diseases characterized by the progressive accumulation of undigested macromolecules in the cell, resulting in cellular dysfunction that leads to diverse pathological manifestations [180].

Transcription factor EB is a protein that in humans is encoded by the TFEB gene. TFEB belong to members of the microphthalmia–transcription factor E (MiT/TFE). TFEB is latent cytoplasmic transcription factor. Its inactive form resides in the cytoplasm (Figure 39), and when activated, it is translocated into the nucleus where it is able to active its target genes [181]. In the 2009 Sardiello et al. [181] showed the association of TFEB with lysosomal biogenesis and an increment of the degradation of complex molecules when this transcription factor is overexpressed, introducing for the first time the biological and medical importance of TFEB for its potential therapeutic involvement in Lysosomal Storage Disorders.

In the 2011 Settembre et. al. [182] demonstrated a link between TFEB and autophagy pathway during starvation, also proving that TFEB nuclear translocation is induced by MAPK1 kinase phosphorylation. The authors proved that in order to be translocated into the nucleus, TFEB needs the addition of a phosphate group on one of its three serine sites Ser^{142}, Ser^{332} or Ser^{402}, indirectly showing

the existence of a phosphatase able to block TFEB into the cytoplasm removing the phosphate group from one of the three serine sites previously identified.

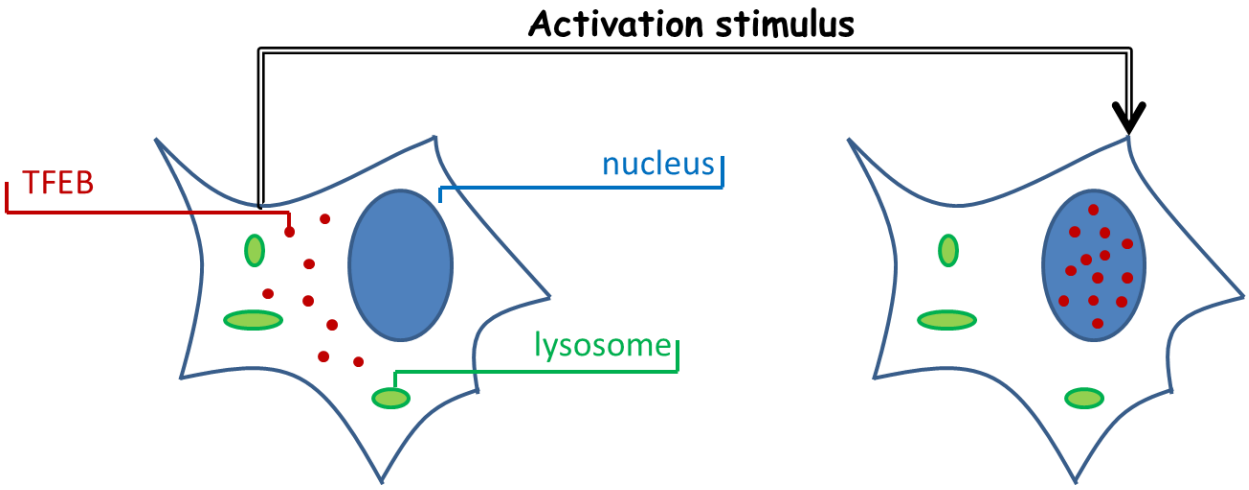


Figure 39 – TFEB is a latent cytoplasmic transcription factor, its inactive form resides into the nucleus. When it is activated it is translocated into the nucleus and it is able to activate its downstream lysosomal targets.

7.2 Introduction to High Content Screening

In the last two decades, different high-throughput technologies to investigate molecular pathways and drug mode of action have been developed. In the study of the transcriptome, for example, two of these technologies such as oligo-based microarrays and Next Generation Sequencing method for mRNAs/microRNAs (RNA-sequencing) had a huge impact. At the same time, high-throughput technologies for the direct study of cellular behavior, including change in morphology or macromolecules localization, have been an open challenge due to incompatibility of the required techniques with high-throughput strategies. Recently, hardware improvements in microscopy, as well as auto-focusing and automatic sampe handling with dedicated robots, have led to the development of automated microscopes. These microscopy improvements combined with quantitative measurements from acquired images of fluorescence, morphology or macromolecules localization, have given rise to the new concept of High Content Screening (HCS). HCS can be considered a high-throughput technology based on automated imaging approach to measure individual cell spatio-temporal events in biological

systems, such as, organelle morphology and complex phenotypes. HCS provides the opportunity to measure cell sub-populations and to combine multiple measurements per cell.

In high content screening cells are treated with drugs or via transfection of siRNA oligomers (RNAi) to study their effect on a phenotype of interest. Structures and molecular components of single cells are automatically analyzed (Figure 40). A common approach consists in using a florescent marker to measure changes in cell phenotype using automated image analysis. Through the use of fluorescent tags with different absorption and emission maxima, it is possible to measure several different cell components in parallel. Moreover, the image analysis techniques are able to detect changes at a subcellular level, including the translocation of a protein from an organelle to another, as well as, the translocation of a transcription factor from the cytoplasm to the nucleus. With this new high-throughput technology a large number of data points can be measured and animalized per cell.

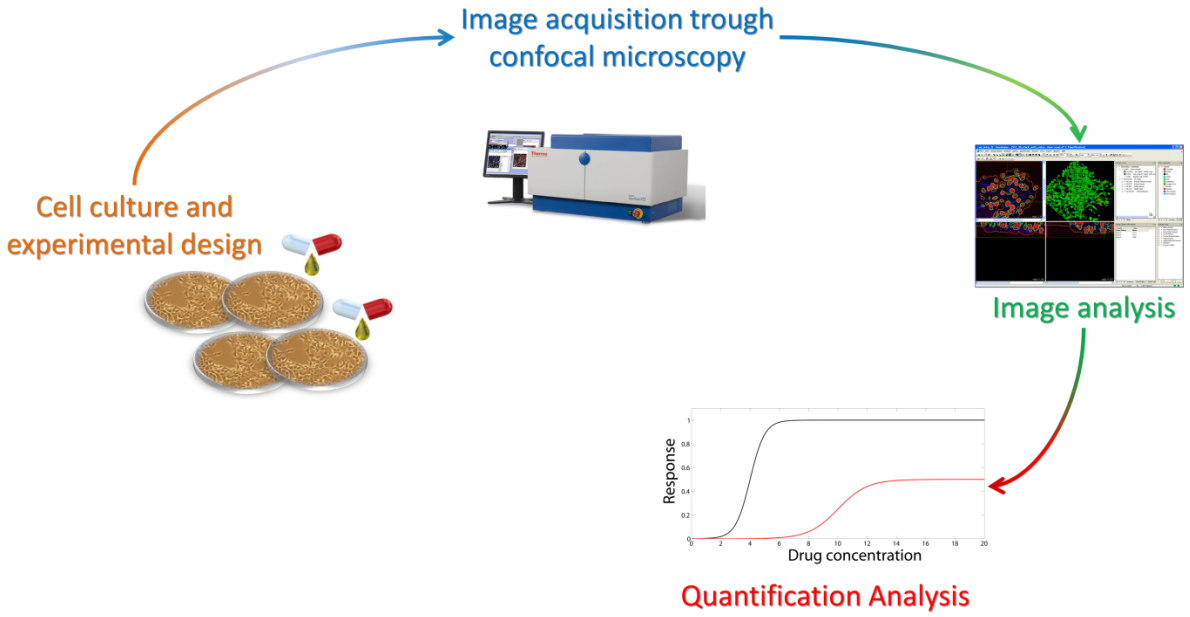


Figure 40 – Steps used in the high content screening experiments. First cells are treated, and then using confocal microscope connected to a pc images are automatically acquired. Finally quantitative measures are automatically extracted from image.

HCS has been successfully implemented in drug discovery [183] to identify small molecules, including peptides or siRNAs, able to modify the phenotype of a cell in the desired manner. A detailed guide for the use and applications of high contents screening can be found here [184].

7.3 Identification of kinases and phosphatases regulating TFEB

As introduced in the section 7.1, TFEB translocation into the nucleus is regulated at a post-translational level by MAPK1 kinase [182], while the opposite mechanism allowing the export from the nucleus into the cytoplasm still remains unknown. We used the DMI method to identify phosphatases able to modulate the activity of TFEB by blocking it into the cytoplasm (or exporting from the nucleus) by removing the phosphate group on one of its three serine sites Ser^{142}, Ser^{332} or Ser^{402} used by MAPK1. For this reason, I compiled a list of all the known 174 phosphatases present in the human genome to test them with DMA to identify the possible modulators of TFEB activity. I used as input for DMI 22 experimentally verified lysosomal target of TFEB and the previously describes set of 5372 GEPs [181] (Table 16).

Table 16 - List of 22 experimentally verified lysosomal targets of TFEB.

TFEB	**OFFICIAL GENE SYMBOLS**
TARGETS	ARSA, ARSB, ATP6V0E1, ATP6V1H, CLCN7, CTSA, CTSB, CTSD, CTSF, GALNS, GLA, GNS, GRN, HEXA, LAMP1, MCOLN1, NAGLU, NEU1, PSAP, SCPEP1, SGSH, TPP1

7.4 Comparison with High Content Screening results

HCS has been used in our institute by Dr Diego Medina in order to identify possible phosphatases able to block the translocation of TFEB into the nucleus. This was achieved by setting up a high-content-screening experiment using a library of siRNA oligos directed against 231 human phosphatases. The screening was performed in three nutrient conditions including, normal medium used as control, serum starvation and aminoacids starvation plus refeeding, to test TFEB translocation (starvation cause TFEB to translocate to the nucleus). The readout of these assays was the nuclear translocation of TFEB. We used the list of identified phosphatases from the first screening as a golden standard for our method. Specifically, we considered only the top six common phosphatases between the two conditions. The PPV-sensitivity curve is reported in Figure 41.

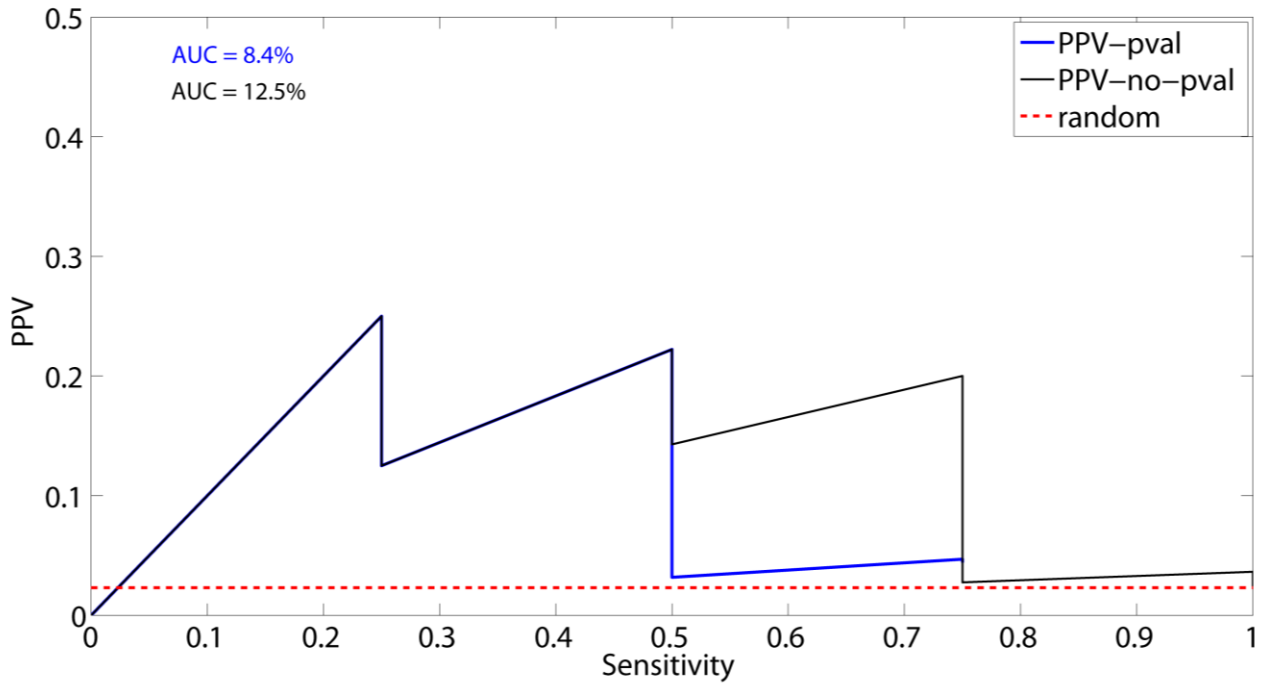


Figure 41 – PPV-sensitivity curve for the identification of TFEB phosphatase modulators, using as a golden standard six phosphatases identified using the High Content Screening approach. P-value has been computed performing 1000 permutation tests.

These results underline the possibility to use our method combined with high content screening to reduce false positives discovery rate of modulators for a transcription factor of interest.

7.5 Discussion and Conclusions

As a case of study, in collaboration with Diego Medina and Adrea Ballabio at Telethon Institute of Genetics and Medicine (TIGEM), I tested the DMI method I developed for the identification of post-translational modulators of a transcription factor on a TF of interest in my institute. TFEB is a latent cytoplasmic transcription factor and thus seems to be a perfect case of study. Comparing the results of DMI with a completely full biological approach such as High Content Screening, a very high PPV is achieved. These results confirm that DMI method could be instrumental in identifying post-translational regulatory interactions in an efficient and cost-effective manner. Moreover this method, combined with high content screening, could be useful to reduce false positives discovery rate in the identification of post-translation modulator for a TF of interest.

References

1. Maston, G.A., S.K. Evans, and M.R. Green, *Transcriptional regulatory elements in the human genome.* Annu Rev Genomics Hum Genet, 2006. **7**: p. 29-59.
2. de la Serna, I.L., et al., *MyoD targets chromatin remodeling complexes to the myogenin locus prior to forming a stable DNA-bound complex.* Mol Cell Biol, 2005. **25**(10): p. 3997-4009.
3. Lemon, B. and R. Tjian, *Orchestrated response: a symphony of transcription factors for gene control.* Genes Dev, 2000. **14**(20): p. 2551-69.
4. Filipowicz, W., S.N. Bhattacharyya, and N. Sonenberg, *Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight?* Nat Rev Genet, 2008. **9**(2): p. 102-14.
5. Valencia-Sanchez, M.A., et al., *Control of translation and mRNA degradation by miRNAs and siRNAs.* Genes Dev, 2006. **20**(5): p. 515-24.
6. Pillai, R.S., S.N. Bhattacharyya, and W. Filipowicz, *Repression of protein synthesis by miRNAs: how many mechanisms?* Trends Cell Biol, 2007. **17**(3): p. 118-26.
7. Standart, N. and R.J. Jackson, *MicroRNAs repress translation of m7Gppp-capped target mRNAs in vitro by inhibiting initiation and promoting deadenylation.* Genes Dev, 2007. **21**(16): p. 1975-82.
8. Nilsen, T.W., *Mechanisms of microRNA-mediated gene regulation in animal cells.* Trends Genet, 2007. **23**(5): p. 243-9.
9. Doench, J.G. and P.A. Sharp, *Specificity of microRNA target selection in translational repression.* Genes Dev, 2004. **18**(5): p. 504-11.
10. Brennecke, J., et al., *Principles of microRNA-target recognition.* PLoS Biol, 2005. **3**(3): p. e85.
11. Lewis, B.P., C.B. Burge, and D.P. Bartel, *Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.* Cell, 2005. **120**(1): p. 15-20.
12. Grimson, A., et al., *MicroRNA targeting specificity in mammals: determinants beyond seed pairing.* Mol Cell, 2007. **27**(1): p. 91-105.
13. Nielsen, C.B., et al., *Determinants of targeting by endogenous and exogenous microRNAs and siRNAs.* RNA, 2007. **13**(11): p. 1894-910.
14. *Finishing the euchromatic sequence of the human genome.* Nature, 2004. **431**(7011): p. 931-45.
15. Prabakaran, S., et al., *Post-translational modification: nature's escape from genetic imprisonment and the basis for dynamic information encoding.* Wiley Interdiscip Rev Syst Biol Med, 2012. **4**(6): p. 565-83.
16. Bansal, M., et al., *How to infer gene networks from expression profiles.* Mol Syst Biol, 2007. **3**: p. 78.
17. Honkela, A., et al., *Model-based method for transcription factor target identification with limited data.* Proc Natl Acad Sci U S A, 2010. **107**(17): p. 7793-8.
18. Foat, B.C., A.V. Morozov, and H.J. Bussemaker, *Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE.* Bioinformatics, 2006. **22**(14): p. e141-9.

19. Prakash, A. and M. Tompa, *Discovery of regulatory elements in vertebrates through comparative genomics.* Nat Biotechnol, 2005. **23**(10): p. 1249-56.

20. Gardner, T.S. and J.J. Faith, *Reverse-engineering transcription control networks.* Phys Life Rev, 2005. **2**(1): p. 65-88.

21. Beer, M.A. and S. Tavazoie, *Predicting gene expression from sequence.* Cell, 2004. **117**(2): p. 185-98.

22. Gardner, T.S., et al., *Inferring genetic networks and identifying compound mode of action via expression profiling.* Science, 2003. **301**(5629): p. 102-5.

23. Edgar, R., M. Domrachev, and A.E. Lash, *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.* Nucleic Acids Res, 2002. **30**(1): p. 207-10.

24. Parkinson, H., et al., *ArrayExpress--a public database of microarray experiments and gene expression profiles.* Nucleic Acids Res, 2007. **35**(Database issue): p. D747-50.

25. Brazma, A., et al., *Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.* Nat Genet, 2001. **29**(4): p. 365-71.

26. Nachman, I., A. Regev, and N. Friedman, *Inferring quantitative models of regulatory networks from expression data.* Bioinformatics, 2004. **20 Suppl 1**: p. i248-56.

27. Segal, E., et al., *Rich probabilistic models for gene expression.* Bioinformatics, 2001. **17 Suppl 1**: p. S243-52.

28. Dojer, N., et al., *Applying dynamic Bayesian networks to perturbed gene expression data.* BMC Bioinformatics, 2006. **7**: p. 249.

29. Yu, J., et al., *Advances to Bayesian network inference for generating causal networks from observational biological data.* Bioinformatics, 2004. **20**(18): p. 3594-603.

30. Segal, E., et al., *Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.* Nat Genet, 2003. **34**(2): p. 166-76.

31. Hartemink, A.J., et al., *Combining location and expression data for principled discovery of genetic regulatory network models.* Pac Symp Biocomput, 2002: p. 437-49.

32. Zou, M. and S.D. Conzen, *A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data.* Bioinformatics, 2005. **21**(1): p. 71-9.

33. Smith, V.A., E.D. Jarvis, and A.J. Hartemink, *Influence of network topology and data collection on network inference.* Pac Symp Biocomput, 2003: p. 164-75.

34. de Jong, H., *Modeling and simulation of genetic regulatory systems: a literature review.* J Comput Biol, 2002. **9**(1): p. 67-103.

35. de la Fuente, A., et al., *Discovery of meaningful associations in genomic data using partial correlation coefficients.* Bioinformatics, 2004. **20**(18): p. 3565-74.

36. Della Gatta, G., et al., *Direct targets of the TRP63 transcription factor revealed by a combination of gene expression profiling and reverse engineering.* Genome Res, 2008. **18**(6): p. 939-48.

37. Basso, K., et al., *Reverse engineering of regulatory networks in human B cells.* Nat Genet, 2005. **37**(4): p. 382-90.

38. Belcastro, V., et al., *Transcriptional gene network inference from a massive dataset elucidates transcriptome organization and gene function.* Nucleic Acids Res, 2011. **39**(20): p. 8677-88.

39. Tegner, J., et al., *Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling.* Proc Natl Acad Sci U S A, 2003. **100**(10): p. 5944-9.

40. van Someren, E.P., et al., *Least absolute regression network analysis of the murine osteoblast differentiation network.* Bioinformatics, 2006. **22**(4): p. 477-84.

41. van Someren, E.P., L.F. Wessels, and M.J. Reinders, *Linear modeling of genetic networks from experimental data.* Proc Int Conf Intell Syst Mol Biol, 2000. **8**: p. 355-66.

42. Weaver, D.C., C.T. Workman, and G.D. Stormo, *Modeling regulatory networks with weight matrices.* Pac Symp Biocomput, 1999: p. 112-23.

43. D'Haeseleer, P., et al., *Linear modeling of mRNA expression levels during CNS development and injury.* Pac Symp Biocomput, 1999: p. 41-52.

44. Chen, T., H.L. He, and G.M. Church, *Modeling gene expression with differential equations.* Pac Symp Biocomput, 1999: p. 29-40.

45. Gustafsson, M., et al., *Reverse engineering of gene networks with LASSO and nonlinear basis functions.* Ann N Y Acad Sci, 2009. **1158**: p. 265-75.

46. di Bernardo, D., et al., *Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks.* Nat Biotechnol, 2005. **23**(3): p. 377-83.

47. Vincenzo, B. *Parallel Computing Algorithms for Reverse-Engineering and Analysis of Genome-Wide Gene Regulatory Networks from Gene Expression Profiles*. 2010.

48. Ideker, T., et al., *Discovering regulatory and signalling circuits in molecular interaction networks.* Bioinformatics, 2002. **18 Suppl 1**: p. S233-40.

49. Reverter, A., et al., *Simultaneous identification of differential gene expression and connectivity in inflammation, adipogenesis and cancer.* Bioinformatics, 2006. **22**(19): p. 2396-404.

50. Leonardson, A.S., et al., *The effect of food intake on gene expression in human peripheral blood.* Hum Mol Genet, 2010. **19**(1): p. 159-69.

51. Lai, Y., et al., *A statistical method for identifying differential gene-gene co-expression patterns.* Bioinformatics, 2004. **20**(17): p. 3146-55.

52. Kostka, D. and R. Spang, *Finding disease specific alterations in the co-expression of genes.* Bioinformatics, 2004. **20 Suppl 1**: p. i194-9.

53. Watson, M., *CoXpress: differential co-expression in gene expression data.* BMC Bioinformatics, 2006. **7**: p. 509.

54. Choi, Y. and C. Kendziorski, *Statistical methods for gene set co-expression analysis.* Bioinformatics, 2009. **25**(21): p. 2780-6.

55. Langfelder, P., et al., *Is My Network Module Preserved and Reproducible?* PLoS Comput Biol, 2011. **7**(1): p. e1001057.

56. Odibat, O. and C.K. Reddy, *Ranking differential hubs in gene co-expression networks.* Journal of Bioinformatics and Computational Biology, 2012. **10**(01): p. 1240002.

57. Ma, H., et al., *COSINE: COndition-SpecIfic sub-NEtwork identification using a global optimization method.* Bioinformatics, 2011.

58. Ideker, T. and N.J. Krogan, *Differential network biology.* Mol Syst Biol, 2012. **8**: p. 565.

59. de la Fuente, A., *From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases.* Trends Genet, 2010. **26**(7): p. 326-33.

60. Manning, G., et al., *The protein kinase complement of the human genome.* Science, 2002. **298**(5600): p. 1912-34.

61. Hjerrild, M., et al., *Identification of phosphorylation sites in protein kinase A substrates using artificial neural networks and mass spectrometry.* J Proteome Res, 2004. **3**(3): p. 426-33.

62. Obenauer, J.C., L.C. Cantley, and M.B. Yaffe, *Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs.* Nucleic Acids Res, 2003. **31**(13): p. 3635-41.

63. Puntervoll, P., et al., *ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins.* Nucleic Acids Res, 2003. **31**(13): p. 3625-30.

64. Linding, R., et al., *Systematic discovery of in vivo phosphorylation networks.* Cell, 2007. **129**(7): p. 1415-26.

65. Wang, K., et al., *Genome-wide identification of post-translational modulators of transcription factor activity in human B cells.* Nat Biotechnol, 2009. **27**(9): p. 829-39.

66. Minguez, P., et al., *Deciphering a global network of functionally associated post-translational modifications.* Mol Syst Biol, 2012. **8**: p. 599.

67. von Mering, C., et al., *STRING: known and predicted protein-protein associations, integrated and transferred across organisms.* Nucleic Acids Res, 2005. **33**(Database issue): p. D433-7.

68. Gambardella, G., et al., *Differential Network Analysis for the identification of condition-specific pathway activity and regulation.* Bioinformatics (in press), 2013.

69. Hide, W., et al., *Application of eVOC: controlled vocabularies for unifying gene expression data.* C R Biol, 2003. **326**(10-11): p. 1089-96.

70. Kimball, R. and M. Ross, *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*2002: John Wiley \& Sons, Inc. 416.

71. Kendall, M.G., et al., *Kendall's Advanced Theory of Statistics, Classical Inference and the Linear Model*1994: John Wiley & Sons.

72. Irizarry, R.A., et al., *Summaries of Affymetrix GeneChip probe level data.* Nucleic Acids Res, 2003. **31**(4): p. e15.

73. Ballester, B., et al., *Consistent annotation of gene expression arrays.* BMC Genomics, 2010. **11**: p. 294.

74. Ferrari, F., et al., *Novel definition files for human GeneChips based on GeneAnnot.* BMC Bioinformatics, 2007. **8**: p. 446.

75. Croft, D., et al., *Reactome: a database of reactions, pathways and biological processes.* Nucleic Acids Res, 2011. **39**(Database issue): p. D691-7.

76. Bossi, A. and B. Lehner, *Tissue specificity and the human protein interaction network.* Mol Syst Biol, 2009. **5**: p. 260.

77. Lehner, B., et al., *Analysis of a high-throughput yeast two-hybrid system and its use to predict the function of intracellular proteins encoded within the human MHC class III region.* Genomics, 2004. **83**(1): p. 153-67.

78. Su, A.I., et al., *A gene atlas of the mouse and human protein-encoding transcriptomes.* Proc Natl Acad Sci U S A, 2004. **101**(16): p. 6062-7.

79. Ciccarelli, F.D., et al., *Toward automatic reconstruction of a highly resolved tree of life.* Science, 2006. **311**(5765): p. 1283-7.

80. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society. Series B (Methodological), 1995. **57**(1): p. 289-300.

81. Kanehisa, M., *The KEGG database.* Novartis Found Symp, 2002. **247**: p. 91-101; discussion 101-3, 119-28, 244-52.

82. Liberzon, A., et al., *Molecular signatures database (MSigDB) 3.0.* Bioinformatics, 2011. **27**(12): p. 1739-40.

83. van de Poll, M.C., et al., *Renal metabolism of amino acids: its role in interorgan amino acid exchange.* Am J Clin Nutr, 2004. **79**(2): p. 185-97.

84. Watters, J.W. and C.J. Roberts, *Developing gene expression signatures of pathway deregulation in tumors.* Mol Cancer Ther, 2006. **5**(10): p. 2444-9.

85. Rothenberg, M.L., D.P. Carbone, and D.H. Johnson, *Improving the evaluation of new cancer treatments: challenges and opportunities.* Nat Rev Cancer, 2003. **3**(4): p. 303-9.

86. Dracopoli, N.C., *Development of oncology drug response markers using transcription profiling.* Curr Mol Med, 2005. **5**(1): p. 103-10.

87. Mani, K.M., et al., *A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas.* Mol Syst Biol, 2008. **4**: p. 169.

88. Carro, M.S., et al., *The transcriptional network for mesenchymal transformation of brain tumours.* Nature, 2010. **463**(7279): p. 318-25.

89. Tanaka, S. and S. Arii, *Molecular targeted therapies in hepatocellular carcinoma.* Semin Oncol, 2012. **39**(4): p. 486-92.

90. Hinds, P.W., et al., *Immunological evidence for the association of p53 with a heat shock protein, hsc70, in p53-plus-ras-transformed cell lines.* Mol Cell Biol, 1987. **7**(8): p. 2863-9.

91. Bressac, B., et al., *Abnormal structure and expression of p53 gene in human hepatocellular carcinoma.* Proc Natl Acad Sci U S A, 1990. **87**(5): p. 1973-7.

92. Hu, T., et al., *Hepatic peroxisomal fatty acid beta-oxidation is regulated by liver X receptor alpha.* Endocrinology, 2005. **146**(12): p. 5380-7.

93. Hailfinger, S., et al., *Regulation of P53 stability in p53 mutated human and mouse hepatoma cells.* Int J Cancer, 2007. **120**(7): p. 1459-64.

94. Lim, Y.P., et al., *The p53 knowledgebase: an integrated information resource for p53 research.* Oncogene, 2007. **26**(11): p. 1517-21.

95. Reddy, J.K. and T. Hashimoto, *Peroxisomal beta-oxidation and peroxisome proliferator-activated receptor alpha: an adaptive metabolic system.* Annu Rev Nutr, 2001. **21**: p. 193-230.

96. Gonzalez, F.J. and Y.M. Shah, *PPARalpha: mechanism of species differences and hepatocarcinogenesis of peroxisome proliferators.* Toxicology, 2008. **246**(1): p. 2-8.

97. Ravasi, T., et al., *An atlas of combinatorial transcriptional regulation in mouse and man.* Cell, 2010. **140**(5): p. 744-52.

98. Francis, G.A., et al., *Nuclear receptors and the control of metabolism.* Annu Rev Physiol, 2003. **65**: p. 261-311.

99. Elfaki, D.A., E. Bjornsson, and K.D. Lindor, *Review article: nuclear receptors and liver disease--current understanding and new therapeutic implications.* Aliment Pharmacol Ther, 2009. **30**(8): p. 816-25.

100. Desvergne, B., L. Michalik, and W. Wahli, *Transcriptional regulation of metabolism.* Physiol Rev, 2006. **86**(2): p. 465-514.

101. Chalkiadaki, A. and L. Guarente, *Sirtuins mediate mammalian metabolic responses to nutrient availability.* Nat Rev Endocrinol, 2012. **8**(5): p. 287-96.

102. Nasrin, N., et al., *SIRT4 regulates fatty acid oxidation and mitochondrial gene expression in liver and muscle cells.* J Biol Chem, 2010. **285**(42): p. 31995-2002.

103. Ahuja, N., et al., *Regulation of insulin secretion by SIRT4, a mitochondrial ADP-ribosyltransferase.* J Biol Chem, 2007. **282**(46): p. 33583-92.

104. Wu, C., et al., *BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources.* Genome Biol, 2009. **10**(11): p. R130.

105. Wang, Y.L., et al., *Human ATAC Is a GCN5/PCAF-containing acetylase complex with a novel NC2-like histone fold module that interacts with the TATA-binding protein.* J Biol Chem, 2008. **283**(49): p. 33808-15.

106. Krebs, A.R., et al., *SAGA and ATAC histone acetyl transferase complexes regulate distinct sets of genes and ATAC defines a class of p300-independent enhancers.* Mol Cell, 2011. **44**(3): p. 410-23.

107. Ding, W.X. and X.M. Yin, *Analyzing macroautophagy in hepatocytes and the liver.* Methods Enzymol, 2009. **453**: p. 397-416.

108. Forman, B.M. and R.M. Evans, *Nuclear hormone receptors activate direct, inverted, and everted repeats.* Ann N Y Acad Sci, 1995. **761**: p. 29-37.

109. Makishima, M., *Nuclear receptors as targets for drug development: regulation of cholesterol and bile acid metabolism by nuclear receptors.* J Pharmacol Sci, 2005. **97**(2): p. 177-83.

110. Vazquez, M.C., A. Rigotti, and S. Zanlungo, *Molecular Mechanisms Underlying the Link between Nuclear Receptor Function and Cholesterol Gallstone Formation.* J Lipids, 2012. **2012**: p. 547643.

111. Sanoudou, D., et al., *Role of Esrrg in the fibrate-mediated regulation of lipid metabolism genes in human ApoA-I transgenic mice.* Pharmacogenomics J, 2010. **10**(3): p. 165-79.

112. Bauer, M., et al., *Starvation response in mouse liver shows strong correlation with life-span-prolonging processes.* Physiol Genomics, 2004. **17**(2): p. 230-44.

113. Miao, J., et al., *Functional inhibitory cross-talk between constitutive androstane receptor and hepatic nuclear factor-4 in hepatic lipid/glucose metabolism is mediated by competition for binding to the DR1 motif and to the common coactivators, GRIP-1 and PGC-1alpha.* J Biol Chem, 2006. **281**(21): p. 14537-46.

114. Tirona, R.G., et al., *The orphan nuclear receptor HNF4alpha determines PXR- and CAR-mediated xenobiotic induction of CYP3A4.* Nat Med, 2003. **9**(2): p. 220-4.

115. Levine, A.J. and M. Oren, *The first 30 years of p53: growing ever more complex.* Nat Rev Cancer, 2009. **9**(10): p. 749-58.

116. Sokolovic, M., et al., *The transcriptomic signature of fasting murine liver.* BMC Genomics, 2008. **9**: p. 528.

117. Yoon, J.C., et al., *Control of hepatic gluconeogenesis through the transcriptional coactivator PGC-1.* Nature, 2001. **413**(6852): p. 131-8.

118. Hunt, M.C. and S.E. Alexson, *The role Acyl-CoA thioesterases play in mediating intracellular lipid metabolism.* Prog Lipid Res, 2002. **41**(2): p. 99-130.

119. van den Bosch, H.M., et al., *Gene expression of transporters and phase I/II metabolic enzymes in murine small intestine during fasting.* BMC Genomics, 2007. **8**: p. 267.

120. Au, W.S., H.F. Kung, and M.C. Lin, *Regulation of microsomal triglyceride transfer protein gene by insulin in HepG2 cells: roles of MAPKerk and MAPKp38.* Diabetes, 2003. **52**(5): p. 1073-80.

121. Kleemann, R., et al., *Time-resolved and tissue-specific systems analysis of the pathogenesis of insulin resistance.* PLoS One, 2010. **5**(1): p. e8817.

122. Kapushesky, M., et al., *Gene expression atlas at the European bioinformatics institute*. Vol. 38. 2010. D690-8.

123. Angelini, C., et al., *BATS: a Bayesian user-friendly software for analyzing time series microarray experiments.* BMC Bioinformatics, 2008. **9**: p. 415.

124. Shlomi, T., et al., *Network-based prediction of human tissue-specific metabolism.* Nat Biotechnol, 2008. **26**(9): p. 1003-10.

125. Gille, C., et al., *HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology.* Mol Syst Biol, 2010. **6**: p. 411.

126. Kharchenko, P., G.M. Church, and D. Vitkup, *Expression dynamics of a cellular metabolic network.* Mol Syst Biol, 2005. **1**: p. 2005 0016.

127. McGill, W.J., *Multivariate information transmission.* Psychometrika, 1954. **19**(2): p. 97-116.

128. Studen, M. and J. Vejnarov, *The multiinformation function as a tool for measuring stochastic dependence*, in *Learning in Graphical Models*, M.I. Jordan, Editor 1998, Kluwer. p. 261-297.

129. Warnat, P., R. Eils, and B. Brors, *Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes.* BMC Bioinformatics, 2005. **6**: p. 265.

130. Liu, H., et al., *Discretization: An Enabling Technique.* Data Min. Knowl. Discov., 2002. **6**(4): p. 393-423.

131. Pál, D., B. Póczos, and C. Szepesvári, *Estimation of Renyi Entropy and Mutual Information Based on Generalized Nearest-Neighbor Graphs.* 2010.

132. Dedecker, J., et al., *Weak Dependence: With Examples and Applications*2007: Springer.

133. Gretton, A., *Consistent Nonparametric Tests of Independence.* J. Mach. Learn. Res., 2010. **99**: p. 1391-1423.

134. Fukumizu, K., et al. *Kernel measures of conditional dependence*. in *In Adv. NIPS*. 2008.

135. Yang, C.Y., et al., *PhosphoPOINT: a comprehensive human kinase interactome and phospho-protein database.* Bioinformatics, 2008. **24**(16): p. i14-20.

136. Chatr-Aryamontri, A., et al., *The BioGRID interaction database: 2013 update.* Nucleic Acids Res, 2013. **41**(Database issue): p. D816-23.

137. Lukk, M., et al., *A global map of human gene expression.* Nat Biotechnol, 2010. **28**(4): p. 322-4.

138. Hollstein, M., et al., *p53 mutations in human cancers.* Science, 1991. **253**(5015): p. 49-53.

139. Vousden, K.H. and C. Prives, *Blinded by the Light: The Growing Complexity of p53.* Cell, 2009. **137**(3): p. 413-31.

140. Espinosa, J.M., *Mechanisms of regulatory diversity within the p53 transcriptional network.* Oncogene, 2008. **27**(29): p. 4013-23.

141. Coutts, A.S. and N.B. La Thangue, *The p53 response: emerging levels of co-factor complexity.* Biochem Biophys Res Commun, 2005. **331**(3): p. 778-85.

142. Bode, A.M. and Z. Dong, *Post-translational modification of p53 in tumorigenesis.* Nat Rev Cancer, 2004. **4**(10): p. 793-805.

143. Meyer, N. and L.Z. Penn, *Reflecting on 25 years with MYC.* Nat Rev Cancer, 2008. **8**(12): p. 976-90.

144. Bentley, D.L. and M. Groudine, *A block to elongation is largely responsible for decreased transcription of c-myc in differentiated HL60 cells.* Nature, 1986. **321**(6071): p. 702-6.

145. Bentley, D.L. and M. Groudine, *Sequence requirements for premature termination of transcription in the human c-myc gene.* Cell, 1988. **53**(2): p. 245-56.

146. Eick, D. and G.W. Bornkamm, *Transcriptional arrest within the first exon is a fast control mechanism in c-myc gene expression.* Nucleic Acids Res, 1986. **14**(21): p. 8331-46.

147. Nepveu, A. and K.B. Marcu, *Intragenic pausing and anti-sense transcription within the murine c-myc locus.* EMBO J, 1986. **5**(11): p. 2859-65.

148. Adams, J.M., et al., *The c-myc oncogene driven by immunoglobulin enhancers induces lymphoid malignancy in transgenic mice.* Nature, 1985. **318**(6046): p. 533-8.

149. Leder, A., et al., *Consequences of widespread deregulation of the c-myc gene in transgenic mice: multiple neoplasms and normal development.* Cell, 1986. **45**(4): p. 485-95.

150. Facchini, L.M. and L.Z. Penn, *The molecular role of Myc in growth and transformation: recent discoveries lead to new insights.* FASEB J, 1998. **12**(9): p. 633-51.

151. Amati, B., et al., *Oncogenic activity of the c-Myc protein requires dimerization with Max.* Cell, 1993. **72**(2): p. 233-45.

152. Dang, C.V., et al., *The c-Myc target gene network.* Semin Cancer Biol, 2006. **16**(4): p. 253-64.

153. Patel, J.H., et al., *Analysis of genomic targets reveals complex functions of MYC.* Nat Rev Cancer, 2004. **4**(7): p. 562-8.

154. Coppola, J.A. and M.D. Cole, *Constitutive c-myc oncogene expression blocks mouse erythroleukaemia cell differentiation but not commitment.* Nature, 1986. **320**(6064): p. 760-3.

155. Dmitrovsky, E., et al., *Expression of a transfected human c-myc oncogene inhibits differentiation of a mouse erythroleukaemia cell line.* Nature, 1986. **322**(6081): p. 748-50.

156. Gandarillas, A. and F.M. Watt, *c-Myc promotes differentiation of human epidermal stem cells.* Genes Dev, 1997. **11**(21): p. 2869-82.

157. Langdon, W.Y., et al., *The c-myc oncogene perturbs B lymphocyte development in E-mu-myc transgenic mice.* Cell, 1986. **47**(1): p. 11-8.

158. Prochownik, E.V. and J. Kukowska, *Deregulated expression of c-myc by murine erythroleukaemia cells prevents differentiation.* Nature, 1986. **322**(6082): p. 848-50.

159. Zeller, K.I., et al., *An integrated database of genes responsive to the Myc oncogenic transcription factor: identification of direct genomic targets.* Genome Biol, 2003. **4**(10): p. R69.

160. Buettner, R., L.B. Mora, and R. Jove, *Activated STAT signaling in human tumors provides novel molecular targets for therapeutic intervention.* Clin Cancer Res, 2002. **8**(4): p. 945-54.

161. Bromberg, J.F., et al., *Stat3 as an oncogene.* Cell, 1999. **98**(3): p. 295-303.

162. Dauer, D.J., et al., *Stat3 regulates genes common to both wound healing and cancer.* Oncogene, 2005. **24**(21): p. 3397-408.

163. Lachmann, A., et al., *ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments.* Bioinformatics, 2010. **26**(19): p. 2438-44.

164. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.

165. Massague, J., J. Seoane, and D. Wotton, *Smad transcription factors.* Genes Dev, 2005. **19**(23): p. 2783-810.

166. Hoodless, P.A., et al., *MADR1, a MAD-related protein that functions in BMP2 signaling pathways.* Cell, 1996. **85**(4): p. 489-500.

167. Liu, F., et al., *A human Mad protein acting as a BMP-regulated transcriptional activator.* Nature, 1996. **381**(6583): p. 620-3.

168. Shi, Y. and J. Massague, *Mechanisms of TGF-beta signaling from cell membrane to the nucleus.* Cell, 2003. **113**(6): p. 685-700.

169. Xu, L. and J. Massague, *Nucleocytoplasmic shuttling of signal transducers.* Nat Rev Mol Cell Biol, 2004. **5**(3): p. 209-19.

170. Inman, G.J., F.J. Nicolas, and C.S. Hill, *Nucleocytoplasmic shuttling of Smads 2, 3, and 4 permits sensing of TGF-beta receptor activity.* Mol Cell, 2002. **10**(2): p. 283-94.

171. Weiss, M.J. and S.H. Orkin, *GATA transcription factors: key regulators of hematopoiesis.* Exp Hematol, 1995. **23**(2): p. 99-107.

172. Vicente, C., et al., *The role of the GATA2 transcription factor in normal and malignant hematopoiesis.* Crit Rev Oncol Hematol, 2012. **82**(1): p. 1-17.

173. Towatari, M., et al., *Regulation of GATA-2 phosphorylation by mitogen-activated protein kinase and interleukin-3.* J Biol Chem, 1995. **270**(8): p. 4101-7.

174. Besnard, A., et al., *Elk-1 a transcription factor with multiple facets in the brain.* Front Neurosci, 2011. **5**: p. 35.

175. Price, M.A., F.H. Cruzalegui, and R. Treisman, *The p38 and ERK MAP kinase pathways cooperate to activate Ternary Complex Factors and c-fos transcription in response to UV light.* EMBO J, 1996. **15**(23): p. 6552-63.

176. Cruzalegui, F.H., E. Cano, and R. Treisman, *ERK activation induces phosphorylation of Elk-1 at multiple S/T-P motifs to high stoichiometry.* Oncogene, 1999. **18**(56): p. 7948-57.

177. Sugimoto, T., S. Stewart, and K.L. Guan, *The calcium/calmodulin-dependent protein phosphatase calcineurin is the major Elk-1 phosphatase.* J Biol Chem, 1997. **272**(47): p. 29415-8.

178. Tian, J. and M. Karin, *Stimulation of Elk1 transcriptional activity by mitogen-activated protein kinases is negatively regulated by protein phosphatase 2B (calcineurin).* J Biol Chem, 1999. **274**(21): p. 15173-80.

179. Dittmer, J., *The Biology of the Ets1 Proto-Oncogene.* Molecular Cancer, 2003. **2**(1): p. 29.

180. Saftig, P., *Lysosomes (Medical Intelligence Unit)*2010: Springer.

181. Sardiello, M., et al., *A gene network regulating lysosomal biogenesis and function.* Science, 2009. **325**(5939): p. 473-7.

182. Settembre, C., et al., *TFEB links autophagy to lysosomal biogenesis.* Science, 2011. **332**(6036): p. 1429-33.

183. Zanella, F., J.B. Lorens, and W. Link, *High content screening: seeing is believing.* Trends Biotechnol, 2010. **28**(5): p. 237-45.

184. Haney, S.A., *High Content Screening: Science, Techniques and Applications*2008: Wiley.