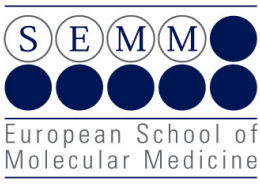

EUROPEAN SCHOOL OF MOLECULAR MEDICINE

NAPLES SITE – *Scientific Coordinator Prof. Francesco Salvatore*

UNIVERSITÀ DEGLI STUDI DI NAPOLI "FEDERICO II"

PhD in Molecular Medicine

***Curriculum* Human Genetics**

XX Ciclo

A yeast synthetic network for In-vivo Reverse-engineering and Modelling Assessment (IRMA)

Supervisor:

Dr Maria Pia Cosma

Internal Supervisor:

Dr Diego di Bernardo

External Supervisor:

Dr Andrea Califano

Ph.D. Student:

Irene Cantone

Acknowledgments

Dedicato a mio padre che mi ha insegnato a sognare, a mia madre che ha cercato di tenermi coi piedi per terra, e ad Antonio che ha colmato il senso di vuoto che mi ha accompagnato in questi anni. Grazie a mia sorella che è stata la mia "messa a terra", alle mie zie, ai miei zii e ai miei nonni perchè non mi hanno mai lasciato da sola.

I would like to thank at first my two supervisors, Dr. Maria Pia Cosma and Dr. Diego di Bernardo, who gave me the great possibility to join such an enthusiast and lively environment as TIGEM is. I was honored of working with them who are both brilliant scientists and good persons. I will always remember the great experience I had here.

I thank also to my first mentor, Prof. Roberto Di Lauro, who fed my passion for science and initiated me to research. His teaching helped me to succeed in this work.

Thanks to Lucia Marucci, Francesco Iorio, Maria Aurelia Ricci, Mukesh Bansal, Vincenzo Belcastro, Stefania Santini and Mario di Bernardo for their constant discussion and feedback, which helped me a lot. Without their contributions this work would not be like it is. Many thanks to Ciro Talotti for his help with yeast media and plates and for all the chats and the jokes we had while working. Thanks to the people of the informatics core at TIGEM and especially to Giampiero Lago, who was always available for helping me with any kind of trouble.

I would like to give a special thank to the friends who supported me even with a simple smile and whom I shared my daily life at TIGEM: Serena Abbondante, Francesco Iorio, Mukesh Bansal, Maria Aurelia Ricci, Alessandro Gennarino, Mario

Buono, Sara Salvia, Giusy della Gatta, Alberto Ambesi-Imbiombato. The greatest thanks go to Vinicia Polito whom I had the fortune to live all my ‘adventure’. We shared happiness and sorrows, goals and defeats, love and pains and we were Friends. She always believed in me and motivated me to stay in lab when I was going to give up. I will never thank her enough. Thank to Antonio Abate for lending me his enthusiasm when mine was not enough and for patiently listening at all my complaints about work, experiments and colleagues. Thanks also to Silvana Libertini, who many times listened at my scientific and personal problems and gave me her support. Without their constant moral support I would have never been able to finish my project successfully.

Table of Contents

Figure Index

Table Index

Abstract

Systems Biology approaches aim to reconstruct gene regulatory networks from experimental data. Conversely, Synthetic Biology aims at using mathematical models to design novel biological 'circuits' (synthetic networks) in order to seed new functions inside the cell. These disciplines require quantitative mathematical models and reverse-engineering techniques.

A plethora of modelling strategies and reverse-engineering approaches has being proposed during the last years. Even if successful applications have being demonstrated, at present their usefulness and predictive ability cannot still be assessed and compared rigorously. There is the pressing and yet unsatisfied need for a 'benchmark': a perfectly known biological circuit that can be used to evaluate pro and cons of such techniques when applied at *in vivo* networks.

In order to address this aim, we constructed in the simplest eukaryotic organism, the yeast *Saccharomyces cerevisiae*, a novel synthetic network for In-vivo Reverse-engineering and Modelling Assessment (IRMA). IRMA is composed of five well-studied genes that have been assembled to regulate each other in such a way to include a variety of regulatory interactions, thus capturing the behaviour of larger eukaryotic gene networks on a smaller scale. It was designed to be isolated from the cellular environment, and to respond to galactose by triggering transcription of its genes.

To demonstrate that IRMA is a unique resource to validate the System and Synthetic biology approaches, we analysed the transcriptional response of IRMA

genes following two different perturbation strategies: by performing a single perturbation and measuring mRNA changes at different time points, or by performing multiple perturbations and collecting mRNA measurements at steady state. We used these data as a 'gold standard' to assess either the predictive ability of mathematical modelling based on differential equations and, to compare four well-established reverse engineering algorithms, NIR, TSNI, BANJO and ARACNE.

We thus showed the usefulness of IRMA as the first simplified model of eukaryotic gene networks built "ad hoc" to test the power of network modelling and reverse-engineering strategies.

Chapter 1 – Introduction

"*The reductionist approach has successfully identified most of the components and many of the interactions but, unfortunately, offers no convincing concepts or methods to understand how system properties emerge...the pluralism of causes and effects in biological networks is better addressed by observing, through quantitative measures, multiple components simultaneously and by rigorous data integration with mathematical models*" (Sauer et al., 2007).

For over a century, biological research has been focused on the identification and the study of individual cellular components and their specific functions. This practice, sometimes called *reductionism*, purports to understand biological systems by dividing them into their smallest possible or discernible elements and understanding their elemental properties alone. Despite its enormous success, it is increasingly clear that a discrete biological function can only rarely be attributed to an individual molecule. Instead, most biological functions stems from the interactions among thousands of different molecular species orchestrating the biological processes needed to sustain life. The different cellular components, such as DNA, RNA, proteins and metabolites, almost never work alone, but interact with each other and with other molecules in highly structured complex ways. From these observations, it is clear that the central dogma of molecular biology where genetic material is transcribed into RNA and then translated into protein is only an oversimplified picture. Identifying regulatory, signalling and metabolic pathways, and understanding their coordinated action is a key challenge for biology in the twenty-first century.

Indeed, modern biology can be considered in a *holistic* sense. This term may not have a precise definition. Aristotele in the Metaphysics concisely summarized the principle of *holism* as follows: "The whole is more than the sum of its parts". Holistic science is an approach to research that emphasizes the study of complex systems. This practice is in contrast with *reductionism* since it recognizes feedback within systems as a crucial element for understanding their behavior and it is irreducible. Systems that have emergent properties are said to be irreducible meaning that it cannot be reduced to its individual parts or studied one part at a time, with the expectation of understanding the emergent properties of the system. This approach aims to study the cell at the systems level by unravelling the regulatory, signalling and metabolic interactions, and understanding their coordinated action.

Biotechnological advances in quantitative high-throughput technology in combination with the growing inter-disciplinarity between biology with engineering and natural sciences, have made this challenge achievable thanks to the emerging fields of Systems and Synthetic Biology (Hasty et al., 2002; Hayete et al., 2007; Kaern et al., 2003; Sprinzak and Elowitz, 2005).

Systems biology aims at developing a formal understanding of biological processes via the development of quantitative mathematical models. A model is a mathematical formalism to describe changes in concentration of each gene transcript and protein in a network, as a function of their regulatory interactions (gene regulatory network). When these interactions are unknown, reverse engineering can be used to infer them from experimental observations.

Synthetic biology aims at using such models to design novel biological ‘circuits’ (synthetic networks) in the cell able to perform specific tasks (e.g. periodic expression of a gene of interest), or to change a biological process in a desired way

(e.g. modify metabolism to produce a specific compound of interest) (Elowitz and Leibler, 2000; Gardner et al., 2000; Khosla and Keasling, 2003; Ro et al., 2006; Tigges et al., 2009).

The usefulness of a model in both Systems and Synthetic Biology lies in its ability to formalise the knowledge about the biological process at hand, to identify inconsistencies between hypotheses and observations, and to predict the behaviour of the biological process in yet untested conditions. There is a variety of mathematical formalisms proposed in literature (Di Ventura et al., 2006; Szallasi et al., 2006) to model biological circuits, with ordinary differential equations being the most common. We gave background of the different modelling strategies and their usefulness in simulating both natural and unnatural systems (the 'synthetic network approach') in Chapter 2.

Reverse engineering methods are used in Systems Biology to uncover unknown molecular interactions from gene expression data that are informative of the network dynamics. Typically, the data consist of measurements at steady state following multiple perturbations (i.e. gene overexpression, knockdown, or drug treatment), or at multiple time points, following one perturbation (i.e. time-series data). Successful applications of these approaches have been demonstrated in bacteria, yeast and, recently, in mammalian systems (Basso et al., 2005; Della Gatta et al., 2008; di Bernardo et al., 2005; Faith et al., 2007; Gardner et al., 2003). A plethora of novel reverse engineering approaches is being proposed, and their assessment and evaluation is of critical importance (Stolovitzky et al., 2007). In Chapter 3, we detailed the three well-established reverse-engineering approaches: ordinary differential equations (ODE), Bayesian Networks, and Information-theory.

In this scenario, the **goal of our work** was to provide the System Biology community with an *in vivo* benchmark, which can be used as "ground of truth" to test and compare different modelling approaches and reverse-engineering inference strategies.

To this aim we constructed, in the yeast *Saccharomyces cerevisiae*, a synthetic network of five genes regulating each other for In-vivo Reverse-engineering and Modelling Assessment (IRMA). We detailed experimental produres and computational methods used in this work in Chapter 4. In Chapter 5 we gave a background of the design principles at the basis of our synthetic network, such as the choice of model organism and of network topology. In Chapter 6 we described the construction of the synthetic network and we detailed the characterization of its behaviour by analysing the transcriptional response of network genes following two different perturbation strategies: by performing a single perturbation, and measuring mRNA changes at different time points, or by performing multiple perturbations, and collecting mRNA measurements at steady state.

We tested the usefulness of IRMA as a simplified biological model to benchmark both modelling and reverse-engineering approaches (chapter 7 and 8, respectively). In Chapter 9, we concluded the thesis and discussed the application of the IRMA network as a unique tool for System and Synthetic Biology.

Part of the work presented in this thesis resulted in a scientific publication (Cantone et al., 2009).

Chapter 2 – Modelling biological systems: from *In vivo* to *In silico* biology and back

In biology the term 'model' is commonly used for graphical descriptions of a mechanism underlying a cellular process, the intrusion of computational biology in the 'wet' lab has being modifying its use to refer to a set of equations expressing in a formal and exact manner the relations among the variables that characterize the state of a biological system. The approach of biologists towards knowledge building has been mostly empirical but experimental facts remain 'blind' without laws or principles derived from them. Conversely, theoretical approaches used by modellers have often failed to relate to real systems, such that theoretical concepts encapsulated in these studies are equally 'empty'. Instead, theory and experiments need to be viewed in close interplay. *In silico* predictions of the behaviour of a biological system can be used to complement *in vivo* experimental observations and accelerate the hypothesis generation-validation cycle of research (Locke et al., 2005). Modelling a cellular process can highlight which experiments are likely to be the most informative in testing model hypotheses, and allow testing for the effect of drugs (di Bernardo et al., 2005) or mutant phenotypes (Segre et al., 2002) on cellular processes—thus paving the way for individualized medicine.

A mathematical model is a formalization of the biological knowledge about a certain system, where each component of the system is described by an equation, which represents its behaviour as a function of its regulators. *A priori* knowledge, which derives from experiments, is essential and needs to be formalized for the chosen framework. Ideally, all information relevant to a system (not only

concentrations and rates of events, but also spatial distribution, diffusion parameters, and so on) would be known to make a maximally accurate *in silico* replica of the system. Unfortunately, even for the best-studied systems, the mass of accumulated data still falls short of describing, even qualitatively, the variety of elementary processes that each molecular species engages in (post-translational modifications, degradation, complex formation, and so on); even less known are details of spatial information and the timing of events. Consequently, assumptions are necessary (for example, that all gene copies of a multi-copy plasmid are transcriptionally active, or that a certain molecule freely diffuses inside a cell or is always monomeric). On the other hand, it can be beneficial to exclude some known data to accommodate available computational power and to facilitate the analysis (even at the expense of accuracy). For example, irrelevant interactions of highly connected proteins could be omitted; details such as the cell-cycle regulation of a certain protein could be temporarily set aside; abundant species such as ATP or ribosomes might be represented as constant pools; or transcription and translation events might be lumped together.

In order to accurately describe the behaviour of a system, the second step in modelling, after the derivation of the equations that describe system components, should be to estimate from experiments those parameters, which numerically describe system dynamics (e.g. synthesis and degradation kinetics, equilibrium constants, basal and maximal concentration of a molecule). With the majority of current experimental techniques yielding only qualitative or semi-quantitative data, biologists have two different options:

- Using descriptive information about the system for **qualitative modelling**;

- Performing target experiments to estimate unknown reactions parameters for **quantitative modelling**.

2.1 Qualitative modelling

In qualitative modelling, for simulations to be applied and useful in drawing non-obvious conclusions, we need to retrieve from biological data at least the information required for the formulation of logical statements describing, for instance, causal relationships between events involving model components. As an example, computer science algorithms used to perform code checks can assess the logical consistency of a set of statements: that is, check that no subset of statements is in contradiction with any other (Batt et al., 2005). Automated tools such as these and others used in qualitative reasoning approaches become indispensable if logical inferences are to be made on very large sets of experimental observations. In qualitative modelling, kinetic processes are simulated by tracking over discrete time the state of the system, defined in terms of a coarse range for each variable. The weak specification of such models conserves computer resources needed to explore the space of possible behaviours; moreover, it provides high-level predictions applying to a whole family of systems—for instance, the number of feedback loops or the ranges of variables supporting oscillations or switches. Although simulation of qualitative models can be fast, even a rough exploration of parameter space can become intractable as the size of the system increases, highlighting the need for increasing computer resources and methods to accelerate the parameters search.

Anyway, since biological systems have evolved tolerance to random fluctuations and perturbations, coarse ranges may suffice to predict correctly a system's behaviour (Csete and Doyle, 2002). For genes that are naturally found in only two states, the trade-off in accuracy may not even be high. On the other hand, simple models can, in some cases, predict behaviours that are far away from reality (Fig. 2.1) (Di Ventura et al., 2006).

2.2 Quantitative modelling

Compared with qualitative models, quantitative ones have a natural appeal in that they offer greater detail in mimicking reality. Moreover, rich qualitative insights on the system are possible using theoretical tools such as bifurcation and stability analysis, which, for example, indicate the precise boundaries of parameter ranges to which steady states or sustained oscillations correspond, or reveal the stability of the solutions before actually solving the dynamical equations representing the system.

Quantitative models can be either deterministic or stochastic. The most popular formalism is the deterministic ordinary differential equations (ODEs). Each equation in a set typically represents the rate of change of a species' continuous concentration as a sum or product of, more or less, empirical terms (typically law of mass action terms, Michaelis–Menten functions, and so on), accounting for the effect of biological events on such concentrations. By definition, the initial state of the system in a deterministic model uniquely sets all future states. As analytical solutions seldom exist, numerical solutions need then to be computed (once for each set of parameter values and initial conditions explored). In general, ODEs are best suited for

capturing the behaviour of systems where species are abundant and reaction events frequent, because species concentrations are then acceptably approximated as varying continuously and predictably. Thus, the deterministic approach approximate the average response of a system within a population of genetically identical cells, or the average response within single cells measured over a long time period.

As the number of molecular species and consequently of reaction events decrease, the probabilistic nature of biological events becomes more evident. In this case, the response of individuals within a population of genetically identical cells may be significantly different from the average population response. Population heterogeneity arises from stochasticity in molecular events or from noise. For instance, occurrence of noise have been found to be exploited by cells to survive a variety of environmental changes (Thattai and van Oudenaarden, 2004) or to increase sensitivity in signal transduction processes (Hanggi, 2002). To model such stochastic systems, two main methods are used. The first comprises using stochastic differential equations (SDEs; derived from ODEs by adding noise terms to the equations), the solutions for which can be numerically obtained either by computing many trajectories (Monte Carlo methods) or approximating their probability distribution and then calculating statistical measures (such as mean and variance). Notably, with this method noise is imposed on the system and represented by mathematical terms chosen a priori, instead of arising from the underlying physical interactions. The second is a very successful and exact method introduced nearly 30 years ago, and recently enhanced to cope with different reaction timescales or space constraints. With this approach, molecules are modelled individually and reaction events are calculated by their probability. The price to pay for having a more physically realistic model is the

considerable increase in computational time and the need for specialized algorithms (Ander et al., 2004).

2.3 Choice of Modelling strategy and pitfalls

The choice of mathematical formalism in modelling depends on what we know and on what we want to know about the real system. Considering the functional phenomenon being modelled and defining a clear biological question to answer helps to chose which modelling strategy is best suited to capture the essential properties of the biological system. Thus, if the system includes gradients, the mathematical formalism used should handle space. If the system seems noisy (for example, not all cells respond in the same way to the same stimulus), then a stochastic approach might clarify this point.

Problems can arise from the mathematical formalism used to simulate a system. To illustrate the impact of modelling choices, we will use as example the simulations of a simple gene network with a negative feedback (protein B forms multimers and sequesters the activator protein AP responsible for its transcription) which were modelled using three different formalisms (Figure 2.1; reported in (Di Ventura et al., 2006)): a simple boolean model, a quantitative deterministic model with ODEs, and a quantitative stochastic model. With a qualitative model, the boolean one, the built-in delay produces oscillations (Figure 2.1B). The other two models require additional events to be modelled explicitly (for example, degradation to balance production), and in contrast to what is observed with the coarser boolean model, oscillations did not occur unless multimerization was allowed (Figure 2.1B-F).

This simple example is useful to underline the importance of choosing the most appropriate model and also to remind that mathematical models represent a simplification of the real system, so when drawing conclusions from simulation results it is essential to keep in mind the limitations of a given approach to represent reality.

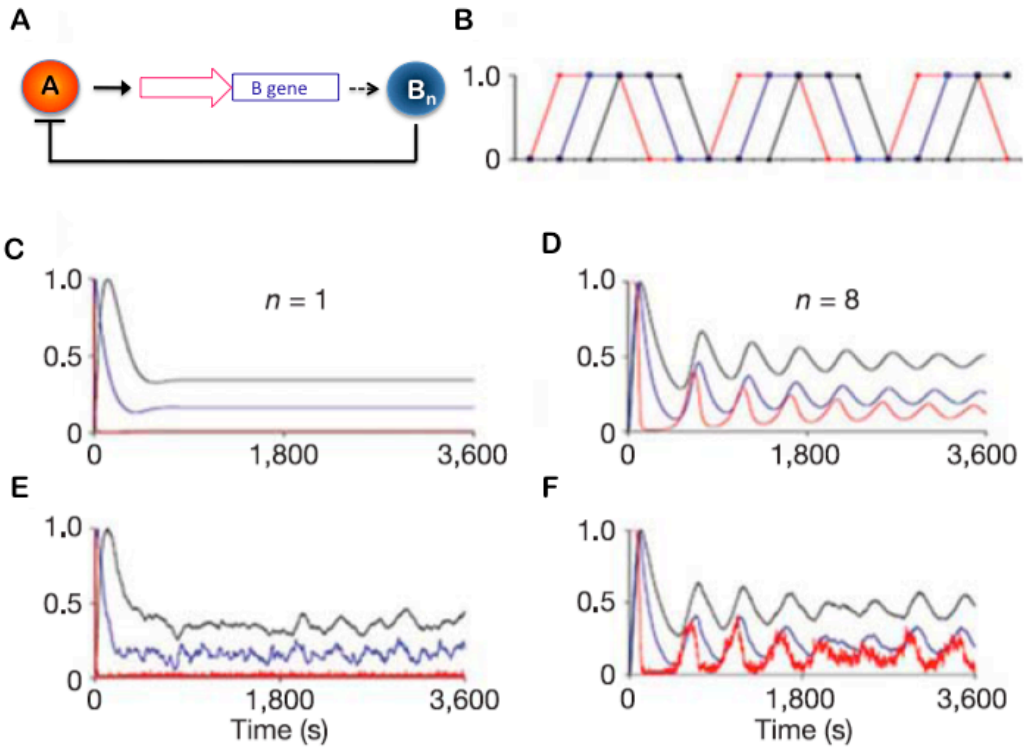

Figure 2.1 Simulation of a Feedback Loop using different mathematical formalisms. (A) Schematic representation of the negative feedback network used in the simulations. n indicates the number of B molecules in the active complex. **(B-F)** Time courses of activator protein A (red), B mRNA (blue) and B protein (black). The y axis represents the number of molecules, normalized for each species by the maximum value reached, except in (B), in which it represents presence or absence of the molecules. Simulation of discrete time boolean model (B) with synchronous update. Deterministic (C, D) and stochastic (E, F) simulations with B monomer (C, E) or octamer (D, F). Oscillations predicted by the boolean model are obtained in the deterministic/stochastic model only when B oligomerization is included. The figure has been modified from (Di Ventura et al., 2006).

2.4 Gaining understanding by mimicking nature

Given the complexity of natural systems, knowledge of all interactions happening at the molecular level does not generally provide the modeller with an intuitive and coherent comprehension about the process of interest. For instance, although we can understand how a signal is propagated in a linear cascade from cell membrane to nucleus, we find it difficult to make sense of a highly interconnected gene network. As a matter of fact, in many cases, although circuit components and their interactions have extensively been identified, this knowledge is not enough to explain the circuit mechanism as a whole. This could happen either because some components miss or at the opposite because not all the identified interactions are relevant. Another problem is ignorance of the effective rules by which proteins and genes interact. For example, *in vivo* values of kinetic parameters such as affinities, binding and degradation rates, and so on, are generally unknown. Finally, the intracellular environment is intrinsically 'noisy', and small copy numbers of molecular species limit the predictability of biochemical reactions. Taken together, these problems reduce our confidence in the combined understanding we get from perturbations, measurements and mathematical modelling.

To circumvent this problem one strategy is to decompose the network into more manageable modules with a defined function and to construct replicas of these small natural circuits, which can be used to find out the most adequate modelling strategy for describing a certain functional module. This approach is well-known in engineering, in which problems are often tackled via simplified empirical models of the process to be studied, where the complexity is reduced to facilitate its handling, but its key features are kept. For example, a jumbo-jet contains over six million parts

and is complex enough to be incomprehensible to the human mind without appropriate simplifications. Nevertheless a simplified toy model of a flying airplane retains some of the most complex and relevant features of the jumbo-jet (fluidodynamics and control) and it is routinely used to derive models and design principles for the full-scale plane (Csete and Doyle, 2002).

Similarly, the field of Synthetic biology has clearly started addressing these issues with small circuits of various designs guided by their own collection of models (Bennett and Hasty, 2008; Chin, 2006; Hasty et al., 2002; Sprinzak and Elowitz, 2005). Reconstructing simplified replicas of natural genetic circuits helps to understand the sufficient and essential biological features that drive a specific function. In this context, for example, the construction of synthetic circuits which are able to produce oscillations in bacteria (Atkinson et al., 2003; Elowitz and Leibler, 2000; Fung et al., 2005; Stricker et al., 2008), and in mammalian cells (Tigges et al., 2009) aims to a better understanding of natural circadian clocks (Panda et al., 2002). The natural circadian rhythms manifest themselves in the periodic variation of concentrations of particular proteins in the cell; for example, in *Drosophila* 'clock genes' (such as *PER*, *TIM* and *VRI*) oscillate with a 24 hours rhythm and self-synchronize to the day/night cycle. Using genetic and biochemical techniques, researchers have isolated genes and proteins involved in interlocked feedback loops of gene expression (Hardin, 2005) (Figure 2.2B) that are necessary for clock function. However, many fundamental questions remain difficult to answer: what sets the period of the oscillation, how does the clock operate reliably in diverse cellular conditions, and what features of its design are responsible for its reliable operation? To gain insight into such questions different simplified synthetic circuits were built to generate self-sustaining periodic oscillations. Even if they fail to operate as reliably,

their construction combined with modelling provide insights in reconstructing a specific function. For example a model of the ‘repressilator’ (Figure 2.2C) showed that the ring architecture is theoretically capable of sustaining oscillations but not all parameter choices give rise to oscillatory solutions. Modelling indicated that high protein synthesis and degradation rates, large cooperative binding effects, and efficient repression favoured oscillations.

In conclusion, this reconstructive approach offers several advantages:

- First, one can test the sufficiency of an arbitrary circuit for generating a particular function.
- Second, one may study the circuit mechanism without impairing cellular functions or inducing downstream consequences.
- Third, different circuit designs with similar functions can be directly compared to determine their relative advantages and disadvantages.

Pushing the engineering analogy even further, systems biology studies can help uncovering the 'organizational principles', or 'design principles', of biological systems. Although there are obvious radical differences between human-engineered and natural systems, natural systems do have solutions similar to human-engineered ones in terms of certain emergent properties (for example, modularity and noise attenuation), details of design (for example, feedback loops) and behaviour (for example, oscillations), as if conforming to a strict set of constraints. Actually, beyond their natural appeal, the use of systems-theoretical concepts is perhaps our only chance to logically formulate the way a complex biological process operates in a concise, synthetic, human-understandable manner.

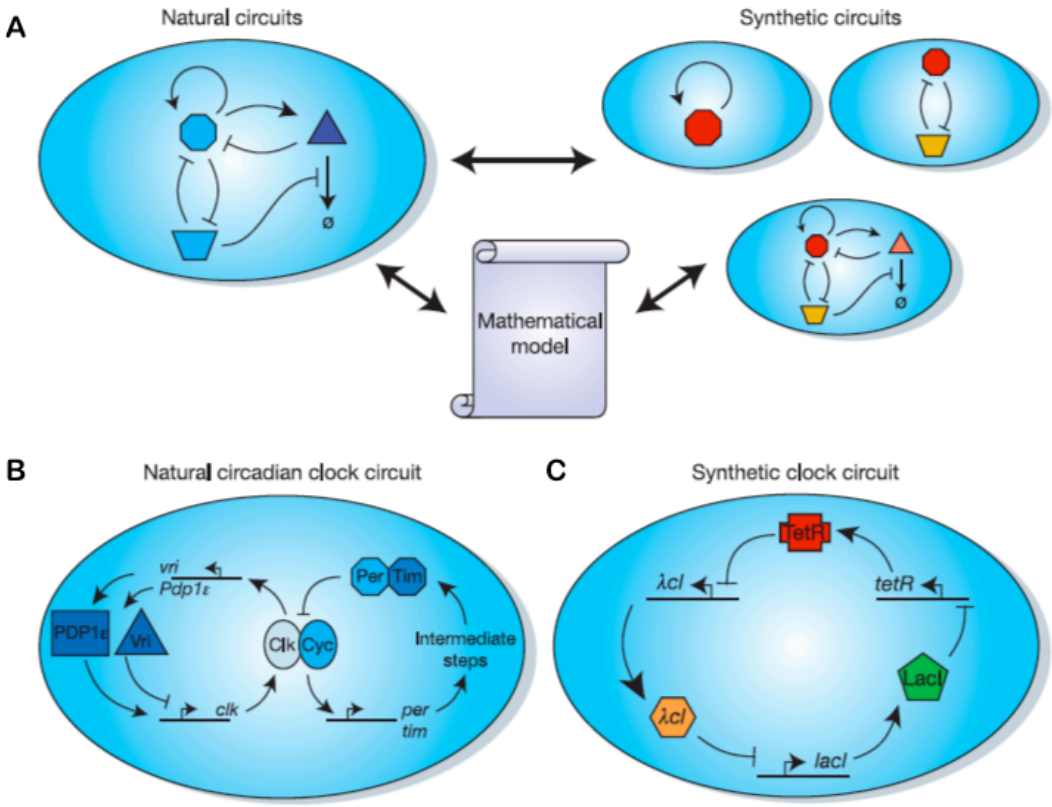

Figure 2.2 Natural and synthetic gene circuits. (A) The synthetic biology paradigm. Genetic circuits are composed of interacting genes and proteins (blue shapes, top left). The pointed and blunt arrows represent positive and negative regulation, respectively. Synthetic circuits (right) based on the natural circuit can be constructed from well-characterized components (red and orange shapes) with similar regulatory effects to form similar or simplified circuits. The dynamics of these synthetic replicas can be compared to the natural system as well as to mathematical models. Analysis of natural circuits, synthetic replicas and models together can help us understand mechanisms used by natural systems. **(B)** Schematic representation of *Drosophila* circadian clock. It contains a negative feedback loop in which Per and Tim, after a delay, repress their own production (via Clk/Cyc, right loop). Interlocked with this negative feedback loop there is another loop involving Vri and Pdp1ε (left loop). c, The 'repressilator' is a simple synthetic clock circuit consisting of a three-component negative feedback loop that operates in E. coli3. The three-element loop provides a delayed negative feedback on all components and permits oscillations. In this sense, it models the generation of oscillations by delayed negative feedback. However, as can be seen from the figure this simple synthetic circuit differs markedly from the natural circadian clock in both complexity and design. (Sprinzak and Elowitz, 2005).

2.5 Engineering new functions inside the cells

Although much remains to be learned in this field, simulation predictions of natural or engineered biological networks are helping us to identify the logical links between system design and system behaviour. As our knowledge will increase, synthetic biologist will be able to combine well-characterised modular genetic components form different organisms, in order to re-engineer cells and entire organisms for numerous applications (including medical, agricultural and ecological situations), or even to construct a synthetic cell, having the minimal and sufficient number of components to be considered alive (the so called 'minimal-cell'; (Luisi, 2007; Luisi et al., 2006)). Several examples of organisms, which were addressed to perform a specific task by seeding synthetic circuits inside cells, are present in literature: yeast cells producing an antimalaric drug (Ro et al., 2006), bacteria sensing environmental toxins and warfare agents (iGEM 2005 and (Looger et al., 2003)) or engineered to act like blood cells (iGEM 2007), or even to make a picture (Levskaya et al., 2005).

Thus, although a global and perfect understanding of a living system is not expected in the near future, the combination of modelling and experimentation offers the possibility of making inroads towards that goal, as well as developing new exciting, useful applications.

Chapter 3 – Reverse Engineering Approaches

Today, biotechnological advances in the development of high-throughput technology platforms, such as microarrays and protein chips, allow measuring simultaneously the different cellular components on genome-scale. Molecular biology is therefore rapidly evolving into a quantitative science, and as such, it is increasingly relying on engineering, applied physics and mathematics to make sense of quantitative high-throughput data.

The challenge for Systems Biology and in particular reverse-engineering is to infer gene networks (i.e. the regulatory interactions among genes), transforming high-throughput heterogeneous data sets into biological insights about the underlying mechanisms. To this aim computational algorithms are typically applied on gene expression data obtained after the perturbation of a gene of interest (i.e. after overexpression or silencing). Gene regulatory networks can be inferred from their inputs (perturbations) and outputs (gene expressions).

There are two broad classes of reverse-engineering algorithms: those based on the "physical interaction" approach that aim at identifying interactions among transcription factors (TFs) and their target genes (gene-to-sequence interaction) and those based on the "influence interaction" approach that try to relate the expression of a gene to the expression of the other genes in the cell (gene-to-gene interaction), rather than relating it to sequence motifs found in its promoter (gene-to-sequence).

3.1 The physical strategy: identifying TF interactions

The physical approach seeks to identify the protein factors that regulate transcription, and the DNA motifs to which the factors bind. In other words, it seeks to identify true physical interactions between regulatory proteins and their promoters.

One of the first methods to accomplish this task was introduced by Tavazoie and colleagues (Tavazoie et al., 1999). The method assumes that transcripts controlled by the same TFs will exhibit similar expression changes under a variety of experimental conditions. With a sufficient number of RNA expression experiments, the method clusters transcripts based on the similarity of their changes across all the experiments. Then the method applies a motif-finding algorithm to identify homologous DNA sequences in the promoter regions of the clustered transcripts. The approach assumes that homologous DNA sequences are probable TF binding motifs.

An advantage of this strategy is that it can reduce the dimensionality of the reverse-engineering problem by restricting possible regulators to TFs. It also enables the use of genome sequence data, in combination with RNA expression data, to enhance the sensitivity and specificity of predicted interactions. The limitation of this approach is that it cannot describe regulatory control by mechanisms other than transcription factors.

3.2 The influence strategy: inferring gene networks

The influence strategy for reverse-engineering seeks to identify regulatory influences between RNA transcripts. In other words, it aims to describe the transcription rate of a set of "output" mRNAs as a function of other "input" transcripts. This type of model is sometimes called a *gene regulatory network* or a *gene network*. In such a model the interaction between two genes does not necessarily imply a physical interaction, but can also refer to an indirect regulation via proteins, metabolites and ncRNA that have not been measured directly (Figure 3.1). Influence interactions include physical interactions, if the two interacting partners are a transcription factor, and its target, or two proteins in the same complex. So, even if gene network algorithms do not use or model protein and metabolite data they can provide a global view of gene regulation that is not restricted to TF/promoter interactions. However, the meaning of influence interactions is not well defined and depends on the mathematical formalism used to model the network. Nonetheless, influence networks have practical utility for:

1) Identifying functional modules, that is, identify the subset of genes that regulate each other with multiple (indirect) interactions, but have few regulations to other genes outside the subset;

2) Predicting the behavior of the system following perturbations, that is, gene network models can be used to predict the response of a network to an external perturbation and to identify the genes directly 'hit' by the perturbation (di Bernardo et al., 2005), a situation often encountered in the drug discovery process, where one needs to identify the genes that are directly interacting with a compound of interest;

3) Identifying real physical interactions by integrating the gene network with additional information from sequence data and other experimental data (i.e. ChIP, yeast two-hybrid assay, etc.).

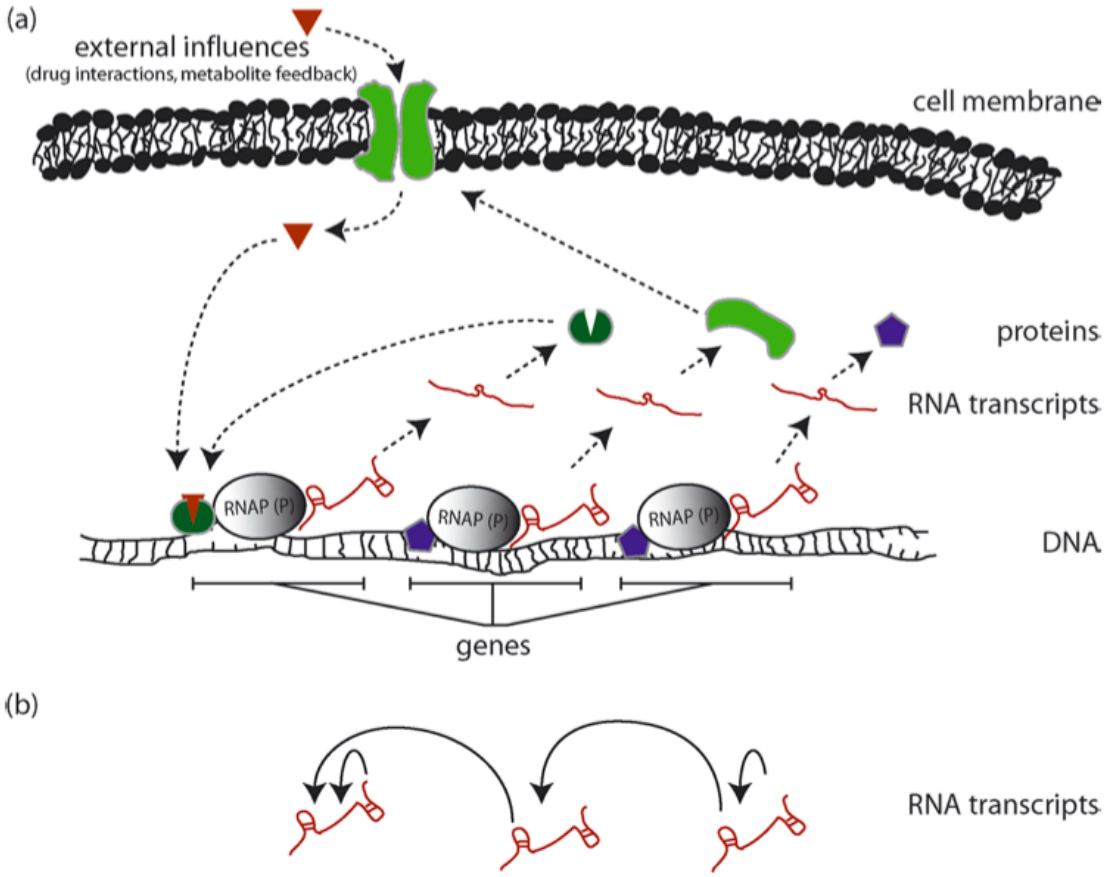

Figure 3.1 Biological networks are regulated at many levels. (a) Shows an example network where protein transcription factors (blue and green shapes) influence the expression of different transcripts (brown lines). One protein is a membrane-bound metabolite transporter. The metabolite it imports (brown triangle) binds one of the transcription factors enabling it to bind DNA and initiate transcription. **(b)** A gene network model of the real network in (a). Since the model is inferred from measurements of transcripts only, it describes transcripts as directly influencing the level of each other, even though they do not physically interact.

In general, one can represent a gene network model as a directed graph (Figure 3.2) Depending on the reverse-engineering approach used, one can describe this graph mathematically as a system of ordinary differential equations (ODEs), as a Bayesian network, or as an association network (Bansal et al., 2007; de Jong, 2002; Faith and Gardner, 2005). The representations provide different degrees of simplification of cell regulation; lend themselves to different computational strategies described below.

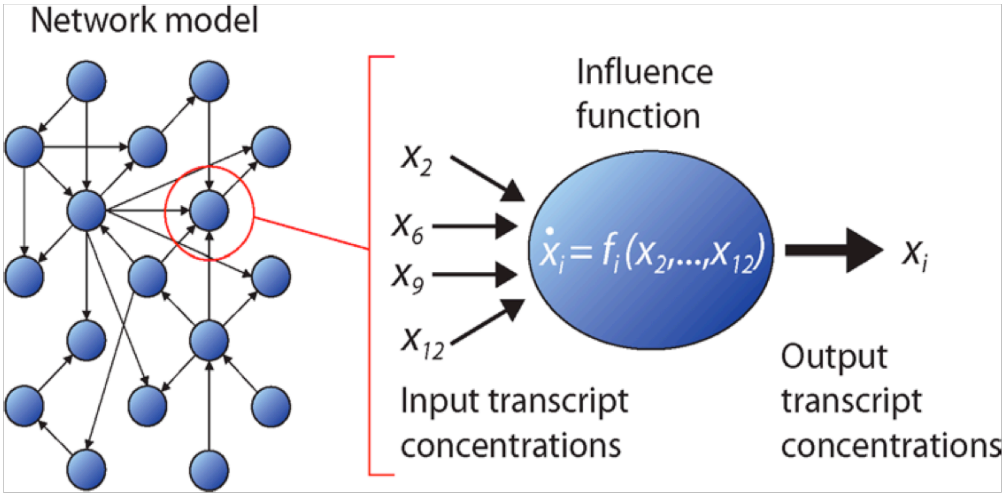

Figure 3.2 Gene network models representation. Gene network models are represented as directed graphs describing the influence of the levels of one set of transcripts (the inputs) on the level of another transcript (the output). One usually assumes that networks are sparse, i.e., only a small subset of transcripts act as inputs to each transcript. The relation between inputs and outputs is specified by an interaction function (f_i).

3.2.1 Ordinary Differential Equations

Reverse-engineering algorithms based on ordinary differential equations (ODEs) relate changes in gene transcripts concentration to each other and to an external perturbation. By external perturbation, we mean an experimental treatment that can alter the transcription rate of the genes in the cell. An example of perturbation is the treatment with a chemical compound (i.e. a drug), or a genetic perturbation involving overexpression or downregulation of particular genes. As ODEs are deterministic, the interactions among genes represent causal interactions, and not statistical dependencies as the other methods. The model consists of a differential equation for each of the N genes in the network, describing the transcription rate of the gene as a function of the other genes and of the perturbation. The parameters of the equations have to be inferred from the expression data.

Linear functions have proven to be the most versatile in the analysis of experimental data sets (Della Gatta et al., 2008; Gardner et al., 2003). 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. This advantage is crucial in light of the high cost of experimental data and the high dimensionality of biological systems. On the other hand, the linear model places strong constraint on the nature of regulatory interactions in the cell. Therefore, oscillations or multistationarity, which are both important properties of real 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.

In summary, the ODE-based approaches yield signed directed graphs and can be applied to both steady-state and time-series expression profiles. Another advantage of ODE approaches is that once the parameters describing interactions among genes are known, the model can be used to predict the behaviour of the network in different conditions (i.e. gene knockout, treatment with an external agent, etc.).

3.2.2 Bayesian Networks

A Bayesian network is a graphical model of probabilistic relationships among a set of random variables, with each variable representing one of the *N* genes in the network. The state of each gene in the network is specified by a probability distribution function, which is dependent on (i.e. *conditioned on*) a set of regulator genes that are called its *parents*. The conditional distribution, which describes relationships (i.e. the gene-gene interactions) in the network, has one important restriction: a gene may be a regulator of any other gene, provided that the network contains no cycles (i.e. a gene cannot directly, or indirectly, regulate itself). In order to reverse-engineer gene networks using a Bayesian approach, we must find the directed acyclic graph that best describes the gene expression data (in the case of time-series data, the directed graph can also contain cycles). These means to find two sets of parameters: the model topology (i.e. the regulators of each gene), and the conditional probability functions, which relate the state of the regulators to the state of the state of their targets. Advantages of using Bayesian networks are:

- It can handle incomplete data sets.
- It allows one to learn about casual relationships.
- It can facilitate the combination of domain knowledge and data.

- It offers an efficient and principled approach for avoiding the overfitting of data.
- Owing to its probabilistic nature, it can also handle noisy data as found in biological experiments.

Owing to its advantages, researchers have devoted considerable attention in recent years to the use of Bayesian network approaches for reverse-engineering gene network (Dojer et al., 2006; Friedman et al., 2000; Segal et al., 2003; Yu et al., 2004; Zhou et al., 2005).

The main limitation of Bayesian networks is that they disregard dynamical aspects completely and require the network structure to be acyclic (i.e. no feedback loops). To overcome these limitations one may use *Dynamic Bayesian networks* (Dojer et al., 2006; Yu et al., 2004). It is an extension of Bayesian networks, which can be used to infer cyclic phenomena such as feedback loops and are also able to infer interactions from time-series data in order to capture dynamic behaviours.

In summary, these approach yields signed directed graphs indicating regulation among genes and, like ODEs ones, can analyse both steady-state and time-series data.

A word of caution: Bayesian networks model probabilistic dependencies among variables (the genes of the network) and not causality, that is, the regulators of a gene are not necessarily the direct causes of its behaviour. However, we can interpret the edge as a causal link if we assume that a variable X is independent of all the other variables (except the target of X), which depends on (i.e. *conditional on*) all its direct causes. It is not known whether this assumption is a good approximation of what happens in real biological networks.

3.2.3 Information-theoretic approaches

In Information-theoretic approaches, the network among *N* genes is reconstructed by considering one pair of genes at the time, and checking whether the two genes are co-expressed across the experimental dataset. In other words, interactions between pairs of genes are assigned when they are expressed with a high statistical similarity in different experiments. Algorithms begin by adding connections between all gene pairs with expression profiles that exceed a threshold of similarity. Ideally, connections in this graph will describe true input–output relationships. However, many connections in this initial graph may associate genes that, for instance, are regulated by the same transcription factor, or that have a common regulator few nodes upstream in the network. In other words, the first step of the algorithm does not distinguish similar and causal relations, nor between direct and indirect relations. To address this problem, a pruning process is undertaken to remove connections that are better explained by a more direct path through the graph. What remains are the connections that are more likely to be causal interactions.

Co-expression can be measured either by correlation, which assumes linear dependence between variables, or by a more robust measure called Mutual Information, which makes no assumption about the form of dependence between variables (Bansal et al., 2007).

Because of its nature, information theoretic approaches yield undirected graphs thus differing from Bayesian networks and Ordinary Differential Equation approaches. Furthermore, Mutual Information based approaches (which are mostly

used) by definition require each experiments to be statistically independent from the others. Thus, they can only deal with steady-state experiments.

Another important aspect, which needs to be considered is that, since this class of algorithms is based on statistical dependence between the genes of the network, it requires a big training dataset of gene expression levels measured over many different experimental conditions.

Chapter 4 – Materials and Methods

4.1 Yeast Strains and Plasmids

All S. cerevisiae strains used to construct IRMA were YM4271 background (*MATa ura3-52 his3-Δ200 ade2-101 lys2-801 leu2-3 trp1-901 gal4-Δ542 gal80-Δ538 ade5::hisG*) kindly provided by M. Johnston (Liu et al., 1993). PCR generated cassettes were used for both integration of the new transcriptional units and contemporary gene deletion as detailed in the paragraphs below (Brown and Tuite, 1998). All the integrations were confirmed by PCR. Genotypes of strains and plasmids generated in this study are listed respectively in Table 1 and 2. Primers that were used to integrate each described cassette in the genome are reported in Table 3.

4.1.1 IRMA Network Construction

To construct the IRMA containing strain, sequential PCR-based genomic integrations were made with the cassettes described in the text below.

At first two HA epitopes were cloned in pAG32 (Goldstein and McCusker, 1999) among Hind III and Bgl II sites. The *2xHA-hphMX4* cassette was amplified by PCR and inserted in front of the stop codon of *ASH1* gene in YM4271 strain resulting in P278 strain.

To generate P280 strain *MET16* promoter (-446 to -1, ATG = +1) was amplified from W303 and cloned in YIplac128 between Hind III and Sac I; *GAL4* ORF was then

cloned between Sac I and Nde I thus resulting in plasmid p*MET16prGAL4*. The *MET16prGAL4-LEU2* cassette was integrated in *SHE2* locus (-11 to +751).

CBF1 ORF was amplified from W303 and cloned among Bam HI and Pac I of pFA6a-GFP(S65T)-*kanMX6* (Wach et al., 1997). Then, the *CBF1-GFP-kanMX6* cassette was integrated downstream of the *HO* promoter (between -1 to +1758) of P280 strain, obtaining P324.

ASH1 promoter (-591 to -1) was cloned in Pst I and Bam HI of YIplac211 where the *GAL80-3xFLAG* was then inserted between Bam HI and Sac I. The *ASH1prGAL80-3xFLAG-URA3* was integrated in *SWI5* locus (-50 to +2299) thus yielding P326. In this strain, *ACE2* gene was then also deleted (from -345 to +2314) by integrating *natMX4* cassette from pAG25 (Goldstein and McCusker, 1999).

Finally, *GAL10prSWI5$_{AAA}$-MYC9- KlTRP1* was integrated in *CBF1* locus (-1 to +1464) resulting in IRMA containing strain (P340). To build *GAL10prSWI5 $_{AAA}$ - MYC9- KlTRP1*, the *SWI5$_{AAA}$* locus was tagged at the C-terminus with nine Myc epitopes in K2072 strain, which was gently provided by K. Nasmyth (Moll et al., 1991). *SWI5$_{AAA}$ -MYC9- KlTRP1* was then amplified by PCR from the resulting strain and cloned in YIplac204 between Eco RI and Aat II. The *GAL10* promoter (-523 to -1) was cloned in YIplac204 between Hind III and Eco RI yielding the vector containing the entire integrated cassette.

4.1.2 Promoter Strength Strains

Strains used for promoter strength measurements were constructed by integrating the promoters containing cassettes in the genome of strains reported in Table 1. The *kanMX4-MET25pr* cassette was amplified by PCR from plasmid pYM-

N34 (Janke et al., 2004) and integrated upstream of the starting codon of *GAL4* (in P265, a wild type strain), and upstream of the ATG of *ASH1* and *SWI5* (in P358, a *she2Δace2Δ* strain) to obtain respectively P549, P362 and P364 strains.

In order to obtain strains which express the *CBF1* TF at different levels, we integrated at the 5' of this gene constitutive promoters of variable strength (*CYC1*, *ADH1*, *TEF*, *GPD* promoters) and the *CUP1* inducible promoter, which were amplified (together with the *kanMX4* resistance cassette) from plasmids pYM-N10, pYM-N6, pYM-N18, pYM-N14, pYM-N1 (Janke et al., 2004).

4.2 Time-series and Steady-state Experiments

For time-series experiments, yeast cells of IRMA-containing strain (P340) were grown at 30°C in YEP containing 2% glucose (YEPD) or 2% galactose and 2% raffinose (YEPGR) until mid-log phase. Cells were then collected by filtration, washed twice with YEP, shifted respectively in YEPGR (for switch-on experiments) or YEPD (for switch-off experiments) and grown at 28°C. Cells were harvested at different time points for RNA extraction.

For steady-state perturbation experiments, centromeric plasmids were constructed as follow. *CBF1*, *GAL4*, *SWI5*, *ASH1* and *GAL80* ORFs were amplified from W303 genome and cloned in pENTR/D-TOPO vector (Invitrogen). Each of these 'entry clones' was then recombined with pAG413*GPD*-ccdB (Addgene 14142) destination vectors by LR Clonase II enzyme, as previously described by (Alberti et al., 2007). IRMA containing strain was then transformed with the obtained plasmids

as described by (Gietz and Woods, 1994). Transformed cells were grown at 30°C in SC (Synthetic complete) medium lacking histidine with 2% glucose or 2% galactose plus 2% raffinose to 0.6-0.8 OD_{600} and then harvested for RNA extraction.

4.3 Promoter Strength Experiments

Strains P349, P362 and P364 were grown in SC medium lacking methionine with 2% glucose in presence of different amounts of methionine (from 1mM to 5 μM) in order to express variable TF levels; only to express *GAL4* (P349) the experiment was also performed by adding 2% galactose and 2% raffinose to the SC medium. In addition, P349 was transformed with pGPD-GAL80, P362 with pGPD-SWI5aaa and P364 with pGPD-ASH1. Different transformed clones were grown in SC medium lacking methionine with 2% glucose in presence of 1mM or 50 μM of methionine; pGPD-GAL80 transformed clones were also grown with 2% galactose and 2% raffinose. The cultured cells described above were harvested for RNA extraction at 0.6-0.8 OD_{600}.

W303 strain (with endogenous *CBF1* gene), P351, P353, P354 and P360 were grown in YEPD up to 0.6-0.8 OD_{600} and harvested for RNA extraction.

P365 cells were grown in SC medium containing 16nM CuSO4 until mid-log phase, then induced for 2h with different amounts of CuSO4 and harvested to collect RNA.

4.4 Semi-quantitative and quantitative RT-PCR

Total RNA was prepared as previously described by Cross et al. (Cross and Tinkelenberg, 1991), treated with 2.5 units/RNA(µg) of DNAse I (Roche) and cleaned up with RNeasy MiniElute Cleaneup Kit (Quiagen) to effectively remove traces of genomic DNA. Lack of genomic DNA contamination was checked by PCR amplification of total RNA samples without prior cDNA synthesis using primers annealing on *ACT1* intron. Cleaned RNA (1.5 µg) was reverse transcribed using SuperScript III First-Strand Synthesis System (Invitrogen).

Semi-quantitative PCR reactions were performed with AmpliTaq Gold (Applied Biosystems) using an amount of cDNA normalized on *ACT1* and *PDA1* gene expression.

Quantitative real-time PCR reactions were set up in duplicates using Platinum SYBR Green qPCR SuperMix-UDG with ROX (Invitrogen), and amplification was performed using a 7000 or 7900 ABI Real-Time PCR machine. Primers were designed using PrimerExpress software (Applied Biosystems). Data analyses were performed using the Applied Biosystems' SDS software version 1.2.3. *ACT1* values were used to normalize the amount of cDNA and ΔCts were calculated as the difference between the average *ACT1* Ct and the average gene*N* Ct. List of primers is given in Table 4.

4.5 Processing of expression data from quantitative RT-PCR

Real-time PCR for the "Glucose steady-state" and "Galactose steady-state" datasets were processed as follows: for each gene in each of perturbation experiments, expression levels were obtained with the ΔCt method yielding fold-changes in perturbed over non-perturbed conditions. Values were averaged across technical and biological replicates. Standard errors were computed using biological replicates.

Real-time data for the switch-on time-series consisted of five independent experiments with a sampling time of 20 min up to 5 hrs; for the switch-off time-series the dataset consists of four independent experiments with sampling time of 10 min up to 3hrs. The data were processed as follows: for each of the time-series, we computed a baseline, by taking the mean value of the time-series, and subtracted it from each of its points. An averaged time-series was then computed by taking the mean of each time-point across the different experiments. We then summed back the mean of the different baselines to the averaged time-series thus obtaining the switch-on and switch-off time-series data. The error bars in Figure 6.6 refer to standard errors.

For promoter strength analysis we had the five datasets described below. In each dataset we measured mRNA expression levels both of the promoter gene to be characterized and of its regulating TF/s (Figure 7.3 and 7.4). Expression levels were obtained with the ΔCt method and values were averaged across technical replicates.

MET16 promoter dataset is composed of 18 data points which were collected from various strains in which *CBF1* is expressed at different levels (P351, P353, P354, P360, P365). For *ASH1* promoter we collected 29 data points in strain P364 after induction of *SWI5*. *HO* promoter dataset includes 38 data points in strain P364

after induction of *SWI5* at different levels, plus 34 in strain P362 where *ASH1* was induced. Expression levels of *HO*, *SWI5* and *ASH1* were measured in all 72 data points. For *GAL10* promoter characterization two datasets were collected: one in glucose (composed of 32 data points) and another in galactose containing medium (composed of 14 data points). In both datasets, *GAL10*, *GAL80* and *GAL4* expression levels were measured.

4.6 Mathematical Model of the IRMA network

The mathematical model consists of five nonlinear Delay Differential Equations describing the rate of change in mRNA levels of the five genes (Figure 7.1). It was derived using Hill kinetics for the gene interactions and a phenomenological law to describe the interactions between the galactose pathway and the genes in the network. The problem of estimating parameter values was defined as a nonlinear programming problem (NLP) and handled using a Hybrid Genetic Algorithm to the purpose of merging the global-search properties of GAs with the fast local convergence of Least Square (LS) methods (Cantone et al., 2009). The *in silico* experiments, mirroring the Glucose steady-state and Galactose steady-state *in vivo* experiments, were carried out by numerically solving the mathematical model: we used as initial conditions the steady state predicted by the model in unperturbed conditions (either in glucose or in galactose), and we added a constant input, corresponding to the gene overexpression, to each of the five equations.

4.6.1 Parametrization of the IRMA network

To fit the 33 unknown parameters of the mathematical model, we relied both on promoters' strength data (steady-state data) and on non-logarithmic averaged time-series data set of the switch-on experiments (dynamic data). In both the processes, we used a Hybrid Genetic Algorithm (HGA) with two distinct cost functions (Cantone et al., 2009).

Regarding promoter strength data, for each promoter, we fitted to the data the equation at steady state of the gene whose expression is driven by the promoter itself. For example, in the case of *HO* promoter, the function fitted was the right-hand side of the *CBF1* equation (Figure 7.1), thus obtaining:

$$x_1 = \frac{\alpha_1}{d_1} + \frac{v_1}{d_1}\left(\frac{x_3^{h_1}}{(k_1^{h_1} + x_3^{h_1}) \cdot \left(1 + \frac{x_5{}^{h_2}}{k_2^{h_2}}\right)}\right)$$

where x_3 is [*SWI5*] and x_5 is [*ASH1*]. For the fitting, the HGA was used and the objective function was defined as:

$$J = \sum_{i=1}^{n}\left(\frac{y_{calc}^i - y_{\exp}^i}{y_{\exp}^i}\right)^2$$

where *n* is the number of experimental data points, y_{calc} are the predicted values of the mathematical model and y_{exp} are the experimental data points.

Results are shown in Figures 7.3 and 7.4. To quantify the strength of network connections, according to (Zhang et al., 2002), we defined the promoter strength parameter like $\frac{v_{max}}{k_m}$, where v_{max} is the maximal transcription rate and k_m is the Michaelis and Menten constant. In our case, since we are estimating the ratio between

the maximal transcription rate and the degradation rate *d* the promoter strength parameter becames $\frac{v_{max}}{k_m d}$. Table 5 lists the fitted kinetic parameters and the strength of each promoter.

The decreasing strength ranking of the network promoters is the following:

1) Gal4 → *GAL10* promoter in galactose
2) Gal4 → *GAL10* promoter in glucose
3) Ash1 → *HO* promoter
4) Cbf1 → *MET16* promoter
5) Swi5 → *ASH1* promoter
6) Swi5 → *HO* promoter

4.6.2 Fitting of switch-on data

The parameters, which were fitted from steady data (reusable in dynamic simulations of switch-on), are the Michaelis-Menten constants, the relative Hill coefficients, the values of $\widehat{\gamma}$ in glucose and galactose and the ratio between the values of the kinetic parameter $\widehat{v_3}$ in galactose and in glucose. From steady-state data we did not have estimation of the transcription rates and of the basal activities, but only of their ratio to the degradation rates. We could not fit also the degradation rates, and the starvation effect parameters (ψ_1 and ψ_2). The remaining 17 unknown parameters were estimated relying on the data set of nonlogarithmic averaged time-series of the switch-on experiments. The objective function of the HGA was defined as:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{m} \left(\frac{y_{calc}^{i}(j) - y_{\exp}^{i}(j)}{y_{\exp}^{i}(j)} \right)^2$$

where *n* and *m* are respectively the number of genes and experimental data points, y_{calc} are the predicted values of the mathematical model (using the inferred parameters) and y_{exp} are the experimental data points.

In simulations, the initial values of gene expression were set to the steady state values predicted by the model.

The overall estimated values of the 33 unknown parameters of our DDE model are shown in Table 6.

4.7 Reverse Engineering the IRMA network

4.7.1 The Ordinary Differential Equation Approach: the NIR and TSNI algorithms

As described in (Bansal et al., 2007), in the Ordinary Differential Equation approach, the gene network dynamics, describing the time evolution of the mRNA concentration transcribed by each gene, is modelled by a set of Ordinary Differential Equations, one equation for each gene *i*, in a network of N genes:

$$\frac{dx_{il}}{dt} = \sum_{j=1}^{N} a_{ij}x_{jl} + u_{il} = a_i^T x_l + u_{il}, \quad i = 1, \ldots, N, \quad l = 1, \ldots, M \tag{1}$$

where x_{il} is the mRNA concentration of gene *i* following the perturbation in the experiment l; a_{ij} represents the influence of gene *j* on gene *i*; u_{il} is an external perturbation to the expression of gene *i* in experiment l.

Identifying the gene interactions network means to retrieve the matrix A of the coefficients a_{ij} for each gene i in the model described below. This can be accomplished if we measure mRNA concentrations of all the N genes at steady state (i.e. $\dot{\mathbf{x}}_l = 0$) in M experiments and then solve the system of equations, as done in the Network Inference by Regression (NIR) algorithm (Gardner et al., 2003).

Alternatively, the same system of equations can be solved using a single perturbation experiment, and measuring multiple time-points following the perturbation, as done in the Time-Series Network Identification algorithm (TSNI) (Bansal et al., 2007; Bansal et al., 2006).

In order to infer the IRMA network from the "Glucose steady-state" and the "Galactose steady-state" datasets, we applied the NIR algorithm (Gardner et al., 2003). NIR solves equation (1) to obtain the network matrix A from gene expression data. We considered a fixed number of regressors for each predicted gene (k = 2), i.e. we assume that each gene can be regulated by a maximum of 2 other genes. The regressor set was chosen according to the residual sum of square error (RSS) minimization criterion. Since we have only 5 genes in the network we exhaustively searched the best regressors in the space of all the possible couples of genes, In each of the experiments, only one gene i was perturbed. NIR requires only the knowledge of which gene was perturbed in each experiment. The perturbation value was set equal to 1.

In order to reverse-engineer IRMA from the switch-on and switch-off datasets, we applied the TSNI (Time Series Network Identification) algorithm (Bansal et al., 2007). To estimate the coefficients of the gene interaction in IRMA, the TSNI algorithm solves the integral version of equation (1) and identifies the network of genes (matrix A). As in the case of NIR, TSNI requires knowledge of which gene has

been perturbed in the experiment. In our case, since the perturbation is obtained by shifting cells from glucose to galactose or vice versa, we assumed a constant input to the SWI5 gene, since this is the first gene that is affected by galactose treatment at the transcriptional level. We also assumed, as in the case of NIR, that each gene can be regulated by a maximum of 2 other genes (k = 2).

4.7.2 The Bayesian Network Approach: Banjo algorithm

A Bayesian network is a graphical model for probabilistic relationships among a set of random variables X_i, with $i = 1 \ldots n$. These relationships are encoded in the structure of a directed acyclic graph G whose vertices (or nodes) are the random variables Xi. The relationships between variables are described by a joint probability distribution $P(X_1, \ldots, X_n)$ that is consistent with the independence assertions embedded in the graph G and has the form:

$$P(X_1, \ldots, X_N) = \prod_{i=1}^{N} P(X_i = x_i \| X_j = x_j, \ldots, X_{j+p} = x_{j+p}) \quad (2)$$

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 Bayes theorem: $P(A, B) = P(B \| A) * P(A) = P(A \| B) * P(B)$. We observe that the JPD (joint probability distribution) can be decomposed as the product of conditional probabilities as in equation (2) only if the *Markov assumption* holds, that is each variable X_i is independent of its non-descendants, given its parent in the directed acyclic graph G. In order to reverse-engineer a Bayesian network model of a gene network we must find the directed

acyclic graph *G* (i.e. the regulators of each transcript) that best describes the gene expression data *D*. This is done by choosing a scoring function, which evaluates each graph *G* (i.e. a possible network topology) with respect to the gene expression data *D*, and then by searching for the graph *G* that maximizes the score.

Banjo (Bayesian Network Inference with Java Objects) is a gene network inference software that has been developed by (Yu et al., 2004). Banjo is based on the Bayesian networks formalism and implements both Bayesian and Dynamic Bayesian networks; therefore it can infer gene networks from steady-state gene expression data, or from time-series gene expression data. In Banjo, heuristic approaches are used to search the 'network space' in order to find the network graph *G* (that is Proposer/Searcher module in Banjo). For each network structure explored, the parameters of the conditional probability density distribution are inferred and an overall network's score is computed using the BDe metric in Banjo's Evaluator module. The output network will be the one with the best score (Banjo's Decider module). Banjo outputs a signed directed graph indicating regulation among genes. Banjo can analyse both steady-state and time-series data.

In the case of steady-state data, Banjo, as well as the other Bayesian networks algorithms, is not able to infer networks involving cycles (e.g. feedback or feed-forward loops).

In order to reverse-engineer IRMA, we applied Banjo on both time-series and steady-state datasets. Banjo recovers the Dynamic Bayesian Network that better describes the observed data. In order to estimate the JPD of all variables in the network, Banjo first discretizes the data using a *quantile discretization procedure* and then constructs a Bayesian Network that summarize the observations. Moreover, assuming independence above a certain level among the nodes of the network (length

of the chain of parents to be considered for a given node of the network), Banjo repeatedly applies the chain rule. The *minMarkovLag* and *maxMarkovLag* parameters that specify the depth of the parent chain were set to 1. The *random local move* and *simulated annealing*, were chosen as *Proposer/Searcher strategies*, respectively. The amount of time Banjo uses to explore the Bayesian Network space was set to one minute. All the other parameters such as *reannealingTemperature*, *coolingFactor*, and so on, were left with their default values. Of course the parameter values were not arbitrary chosen; those values were selected as best values (in terms of network inference accuracy), as described in (Bansal et al., 2007).

4.7.3 The Information-theoretic Approach: the ARACNE algorithm

Information-theoretic approaches use a pseudo-distance between probability distribution called Mutual Information (MI), to compare expression profiles from a set of microarrays. For each pair of genes (*i*, *j*), their MI_{ij} is computed and the edge $a_{ij} = a_{ji}$ is set to 0 or 1 depending on a significance threshold to which MI_{ij} is compared. The MI can be used to measure the degree of independence between two genes.

Mutual information MI_{ij} between gene *i* and gene *j* is computed as:

$$MI_{ij} = H_i + H_j - H_{ij} \tag{3}$$

where *H*, the entropy, is defined as:

$$H_k = -\sum_{k=1}^{n} p(x_k)\log(p(x_k)). \tag{4}$$

The entropy H_k has many interesting properties, specifically it reaches a maximum for uniformly distributed variables, i.e. the higher the entropy, the more

randomly distributed are gene expression levels across the experiments. From the definition, it follows that MI becomes zero if the two variable x_i and x_j are statistically independent ($P(x_i x_j) = P(x_i)P(x_j)$), since their joint entropy is $H_{ij} = H_i + H_j$. A higher MI indicates that the two genes are non-randomly associated to each other. It can be easily shown that MI is symmetric, $M_{ij} = M_{ji}$, therefore the network is described by an undirected graph G, thus differing from Bayesian networks and Ordinary Differential Equation approaches (directed acyclic graph). The definition of MI in equation (3) requires that each data point, i.e. each experiment, is statistically independent from the others, thus information-theoretic approaches, as described here, can deal with steady-state gene expression data set, or with time-series data as long as the sampling time is long enough to assume that each point is independent from the previous ones. Edges in networks derived by information-theoretic approaches represent statistical dependences among gene expression profiles.

ARACNE (Basso et al., 2005; Margolin et al., 2006) belongs to the family of information-theoretic approaches to gene network inference with their *relevance network* algorithm. ARACNE computes M_{ij} for all pairs of genes i and j in the data set. M_{ij} is estimated using the method of Gaussian kernel density (Steuer et al., 2002). Once M_{ij} for all gene pairs has been computed, ARACNE excludes all the pairs for which the null hypothesis of mutually independent genes cannot be ruled out ($H_0 : M_{ij} = 0$). A p-value for the null hypothesis, computed using Montecarlo simulations, is associated to each value of the mutual information. The final step of this algorithm is a pruning step that tries to reduce the number of false positives (i.e. inferred interactions among two genes that are not direct causal interaction in the real biological pathway). They use Data Processing Inequality (DPI) principle that asserts that if both (i, j) and (j, k) are directly interacting, and (i, k) are indirectly interacting

through j, then $M_{i,k} \leq min(M_{ij}, M_{jk})$. This condition is necessary but not sufficient, i.e. the inequality can be satisfied even if (i, k) are directly interacting, therefore the authors acknowledge that by applying this pruning step using DPI they may be discarding some direct interactions as well.

In order to reverse-engineer IRMA we applied ARACNE on the steady-state datasets, "Glucose steady-state" and "Galactose steady-state", and concatenating them to obtain a larger dataset. The lack of any statistical independence assumption for time-series data does not allow running ARACNE on them. All the parameters were left with their default values. For instance, the software automatically detects the *Kernel width* and *Number of bins*; no *threshold* and *p-value* between both MI values and MI P-value were used, respectively; *DPI tolerance*, which removes false positive "mirrored" connections, was left to its default value, 0.15.

4.7.4 Estimation of the Performance of the Algorithms

In order to assess the inference performances we computed the Positive Predicted Value (PPV) and the Sensitivity scores as described in (Bansal et al., 2007). We considered the following definitions:

TP = **Number of True Positives** = number of edges in the real network that are correctly inferred;

FP = **Number of False Positives** = number of inferred edges that are not in the real network;

FN = **Number of False Negatives** = number of edges in the real network that are not inferred.

Then we computed:

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

and

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

In order to compute the random PPV we considered the expected value of an hyper geometrically distributed random variable whose distribution function and expected value are, respectively:

$$P_x = \frac{\begin{pmatrix} M \\ x \end{pmatrix}\begin{pmatrix} N-M \\ n-x \end{pmatrix}}{\begin{pmatrix} N \\ x \end{pmatrix}}, \quad E[x] = M\frac{\begin{pmatrix} N-1 \\ n-1 \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}} = M\frac{n}{N}.$$

In our case, *N* =number of possible edges in the network; *M* = number of true edges, *n* = number of predicted edges. Then we computed as random PPV:

$$PPV_{rand} = \frac{TP_{rand}}{TP + FP} = \frac{E[x]}{n} = \frac{M}{N}.$$

Reverse-engineering algorithms can infer interactions with direction of regulation (A regulates B and not vice versa − directed graph), or just an undirected interaction (A regulates B, or, B regulates A − undirected graph). For unsigned directed networks the value of the random PPV is equal to 8/20=0.4, for unsigned directed networks, it is equal to 7/10 = 0.70.

4.7 Fluorescence Microscopy

For microscopy analysis, yeast cells were grown over-night in 5ml of YEPD or YEPG at 28°C. Ten µl of cell suspension were applied on a microscope slide, sealed with a coverslip and immediately inspected on Zeiss microscope (Axioplan 2 imaging) with a 63x oil immersion objective lens (Zeiss). Pictures were taken with Axiocam camera controlled by AxioVision software.

Strain Name	Genotype	Source
P277	*MATa*, *ura3-52 his3-Δ200 ade2-101 lys2-801 leu2-3 trp1-901 gal4-Δ542 gal80-Δ538 ade5::hisG*	YM4271[1]
W303 (P265)	*MATa*, *ura3-52 his3-11,15 ade2-1 leu2-3,112 trp1-1 can1-100*	
P15	*MATa*, *swi5Δ::URA3*	W303
P302	*MATa*, *ace2Δ::natMX4*	W303
P304	*MATa*, *ace2Δ::natMX4*	P15
P366	*MATa*, *cbf1Δ::hphMX4*	W303
P274	*MATa*, *gal4Δ::kanMX4*	W303
K2072	*MATa*, *ho SWI5(AAA)*	W303[2]
P323	*MATa*, *ho SWI5(AAA)-myc9-KlTRP1*	K2072
P278	*MATa*, *ASH1-2xHA-hphMX4*	P277
P280	*MATa*, *she2Δ::MET16*p-*GAL4-LEU2*	P278
P324	*MATa*, *hoΔ::CBF1-GFP-kanMX6*	P280
P326	*MATa*, *swi5Δ::ASH1*p-*GAL80-3xFlag-URA3*	P324
P331	*MATa*, *ace2Δ::natMX4*	P326
P340	*MATa*, *cbf1Δ::GAL10*p*SWI5(AAA)-myc9-KlTRP1*	P331
P351	*MATa*, *kanMX4-ADH*p*::CBF1*	P265
P353	*MATa*, *kanMX4-CYC1*p*::CBF1*	P265
P354	*MATa*, *kanMX4-GPD*p*::CBF1*	P265
P360	*MATa*, *kanMX4-TEF*p*::CBF1*	P265
P365	*MATa*, *kanMX4-CUP1*p*::CBF1*	P265
P349	*MATa*, *kanMX4-MET25*p*::GAL4*	P265
P355	*MATa*, *ace2Δ::natMX4*	K2072
P358	*MATa*, *she2Δ::hphMX4*	P355
P362	*MATa*, *kanMX4-MET25*p*::ASH1*	P358
P364	*MATa*, *kanMX4-MET25*p*::SWI5(AAA)*	P358

Information under "Source" enables the origins of the various strains to be traced.

[1] The strain was kindly provided by Johnston M.

[2] The strain was kindly provided by Nasmyth K.

Table 1. List of Yeast Strains and Their Genotype

Plasmid Name	**Cloned sequence (restriction sites)**	**Backbone**
pGal10pYIp204	*GAL10* -523 to -1 (Hind III - Eco RI)	YIplac204
pMet16pYIp128	*MET16* -446 to -1 (Hind III - Sac I)	YIplac128
pAsh1pYIp211	*ASH1* -591 to -1 (Pst I - Bam HI)	YIplac211
pGal10pSwi5aaaMyc9	$SWI5_{AAA}$ (ORF)-*MYC9*- *KlTRP1* (Eco RI - Aat II)	pGal10pYIp204
pMet16pGal4	*GAL4* (ORF) (Sac I – Nde I)	pMet16pYIp128
pAsh1pGal80-3xFlag	*GAL80* (ORF)-Nar I-3xFlag (Bam HI – Sac I)	pAsh1pYIp211
p2xHA	5'- AAC ATC TTT TAC CCA TAC GAT GTT CCT GAC TAT GCG GGA GGA TCC TAT CCA TAT GAC GTT CCA GAT TAC GCT GCT CAG TGC TGA -3' synthetic sequence (Hind III – Bgl II)	pAG32
pCbf1GFP(S65T)	*CBF1* (ORF) (Bam HI - Pac I)	pFA6a-GFP(S65T)-kanMX6
pENTRCbf1	*CBF1* (ORF)	pENTR/D-TOPO
pENTRCbf1-S	*CBF1* (ORF without stop codon)	pENTR/D-TOPO
pENTRGal4	*GAL4* (ORF)	pENTR/D-TOPO
pENTRGal4-S	*GAL4* (ORF without stop codon)	pENTR/D-TOPO
pENTRSwi5aaa	$SWI5_{AAA}$ (ORF)	pENTR/D-TOPO
pENTRSwi5aaa-S	$SWI5_{AAA}$ (ORF without stop codon)	pENTR/D-TOPO
pENTRAsh1	*ASH1* (ORF)	pENTR/D-TOPO
pENTRAsh1-S	*ASH1* (ORF without stop codon)	pENTR/D-TOPO
pENTRGal80	*GAL80* (ORF)	pENTR/D-TOPO

pENTRGal80-S	*GAL80* (ORF without stop codon)	pENTR/D-TOPO
pGPDCbf1	*CBF1* from pENTRCbf1	pAG413GPD-ccdB
pGPDGal4	*GAL4* from pENTRGal4	pAG413GPD-ccdB
pGPDSwi5aaa	$SWI5_{AAA}$ from pENTRSwi5aaa	pAG413GPD-ccdB
pGPDAsh1	*ASH1* from pENTRAsh1	pAG413GPD-ccdB
pGPDGal80	*GAL80* from pENTRGal80	pAG413GPD-ccdB

Table 2. List of plasmids

Primer	**Sequence 5' to 3'**	**Comments**
HAF	TACCGTTGCTTATTTTGTAATTACATAACTGAGACAGTAGAGAATAACATCTTTTACCCATACGAT	*ASH1* tagging (P278)
HA2	CGTGATAATGTCTCTTATTAGTTGAAAGAGATTCAGTTATCCATGTATCAATCGATGAATTCGAGCTCG	*ASH1* tagging (P278)
INTGal4F	AGAGAAAGCACAGTAAACCCTCCTTAATTTTCCTTTTGCATAATACCCACCATGATTACGCCAAGCTT	*MET16*p-*GAL4* in *SHE2* locus (P280)
INTGal4R	TATATGTTCTATTAACTAGTGGTACTTATTTGCTCTTTTTGAGCTAAGGCGTATCACGAGGCCA	*MET16*p-*GAL4* in *SHE2* locus (P280)
CBF1ATG	ATCCATATCCTCATAAGCAGCAATCAATTCTATCTATACTTTAAAATGAACTCTCTGGCAAATAAT	*CBF1-GFP* in *HO* promoter (P324)
CBF1C	AATTTTACTTTTATTACATACAACTTTTTAAACTAATATACACATTTATCGATGAATTCGAGCTCG	*CBF1-GFP* in *HO* promoter (P324)
INTGal80F	GAGCTAGGTAAATAGATCCTGAGAACGTGTTTAACATCTGCGATATACCATGATTACGCCAAGCTT	*ASH1*p-*GAL80-3xFlag* in *SWI5* locus (P326)
INTGal80R	ATTCCTAAAGTTATAGTTCACATTGTTATATATGTATCTATAAAGCGAGGCGTATCACGAGGCCA	*ASH1*p-*GAL80-3xFlag* in *SWI5* locus (P326)
Ace2NatFv	TCATAATATACGATATATATCTCAAAACGGCAAAATGTAAACATTCGTACGCTGCAGGTCGAC	*ACE2* deletion (P331, P355)
Ace2NatRv	TGTTACTATTATTTATTATGTTAATATCATGCATAGATAAATGTTATCGATGAATTCGAGCTCG	*ACE2* deletion (P331, P355)
SWITagF2	AATGGAACGGGGATTATGGTTTCGCCAATGAAAACTAATCAAAGGTCCGGTTCTGCCGCTAG	*SWI5* tagging (P323)

SWITagR2	TTTATTATTAAATATTAAAAAAAGTGTCCATAACATCAATGTTTTTTTCCTCGAGGCCAGAAGAC	*SWI5* tagging (P323)
SWI1	CAACATCAAGTGCTTAAAATATAATACGGTTTTCTACACTTTTATTAACGGACCATGATTACGCCAAGCT	*GAL10*p-*SWI5(AAA)-myc9* in *CBF1* locus (P340)
INTSwiKL	AAAGTAGAAATAGGCCCGTGATTGTCGCGGACCTTCAAGGATGTGACGTTCCTCGAGGCCAGAAGACTA	*GAL10*p-*SWI5(AAA)-myc9* in *CBF1* locus (P340)
MetGAL4Fv	TGCACGCCATCATTTTAAGAGAGGACAGAGAAGCAAGCCTCCTGAAAGATGCGTACGCTGCAGGTCGAC	*MET25*p in *GAL4* locus (P349)
MetGAL4Rv	CTTTTTAAGTCGGCAAATATCGCATGCTTGTTCGATAGAAGACAGTAGCTTCATCGATGAATTCTCTGTCG	*MET25*p in *GAL4* locus (P349)
PrCBF1Fv	CATCAAGTGCTTAAAATATAATACGGTTTTCTACACTTTTATTAACGATGCGTACGCTGCAGGTCGAC	Promoters in *CBF1* locus (P351-3-4,360, 365)
PrCBF1Fv	TGGATTTCCTCATCCTCAGTAGAAAGCTTATTATTATTTGCCAGAGAGTTCATCGATGAATTCTCTGTCG	Promoters in *CBF1* locus (P351-3-4,360, 365)
DShe2Fv	AGAGAAAGCACAGTAAACCCTCCTTAATTTTCCTTTTGCATAATACCCGTACGCTGCAGGTCGAC	*SHE2* deletion (P358)
DShe2Rv	TATATGTTCTATTAACTAGTGGTACTTATTTGCTCTTTTTGAGCTAATCGATGAATTCGAGCTCG	*SHE2* deletion (P358)
MetSwi5Fv	ATTGGATTCTAGGGCCAATGTTATTTCTGTCTTAAAGGAGAGCGAATCAACGTACGCTGCAGGTCGAC	*MET25*p in *SWI5* locus (P364)
MetSwi5Rv	AAAATTTAGGCTTTGTACTTTTGAGGCATCAAACCAAGAGTTTGATGTATCCATCGATGAATTCTCTGTCG	*MET25*p in *SWI5* locus (P364)
MetAsh1Fv	ATGTGGAACAGAAAAGAAATCGGGGCGCTTCCTCTTCTGTATTCCTTTAATTCGTACGCTGCAGGTCGAC	*MET25*p in *ASH1* locus (P362)
MetAsh1Fv	TCCGGACCAGCAGATAATGCATGCAGTGGTGTTTTGATGTATAAGCTTGACATCGATGAATTCTCTGTCG	*MET25*p in *ASH1* locus (P362)

All primers listed above were used for amplifying by PCR the cloned cassettes to be integrated in the specified locus. They consist of 45-50 nucleotides that are homologous to the targeted locus followed by 18-20 (underlined) nucleotides that anneal on the cassette.

Table 3. Oligonucleotides used for PCR-based integrations

Primers	Sequence 5’ to 3’	Comments
RTCbf1	GAGGATATGCACACTCACA	semi-quantitative
RTGFP	AGATTGTGTGGACAGGTAAT	semi-quantitative
Swi5_1242	AGACCAATATACACCAAGAGG	semi-quantitative
RTMyc9	CGTTCAAGTCTTCTTCTGAGA	semi-quantitative
RTAsh1	CTAGTTACAGTTCTGTCTCT	semi-quantitative
HARv	TCAGCACTGAGCAGCGTA	semi-quantitative
FLAG1	CCTTGCATGTTCACTAGAT	semi-quantitative
FLAG2	CGTCATCCTTGTAGTCGAT	semi-quantitative
Gal10Fv	TCATGCATTCTGCAAAGCTTC	semi-quantitative
Gal10Rv	CCCGTAAGTTTCACCGTTTTT	semi-quantitative
Met16Fv	TAATCAAGCTGGAAACGCCAC	semi-quantitative
Met16Rv	ATCGGCTGGCTTCATGAATT	semi-quantitative
HOFv	TCCAGGGTGAGAGTACTGT	semi-quantitative
HORv	CGGACAGCATCAAACTGTAA	semi-quantitative
Ash1Fv	CGCTTCCCTGATACATCAAA	semi-quantitative
Ash1Rv	TCAATTTCGCAGTTGCGTTC	semi-quantitative
CDC6Fv	TAGAATCCGTGGCGGTAACC	Both
CDC6Rv	TGGGCCATTCAGATCTTGGA	Both
PCL2Fv	TTAACAAACACTGGGCCGAAT	Both
PCL2Rv	TGGTGACGTCCCAATCAAAAT	Both
SIC1Fv	TATTGTTTCCCACGCAGCAA	Both
SIC1Rv	CTGCCTGGCAGATGTAGGTCT	Both
RME1Fv	AATTTCCGAAGGGCAAACAA	Both
RME1Rv	TGAATTCGTCTAAGTGCGCG	Both
PCL9Fv	TCGGTTCCTTCACTGACATCC	Both
PCL9Rv	TCAGATTCCACCAACGGTAGG	Both
PIR3Fv	TGCCTATGCTCCAAAGGACC	Both

PIR3Rv	CGGCTTCAATAGCAATACCGA	Both
Act1F	TTCTGAGGTTGCTGCTTTGGT	Both
Act1R	TGGTGTCTTGGTCTACCGACG	Both
Cbf1_826F	GCAAACATCGAAAAGTGGACG	quantitative
Cbf1_926R	GCATTTCCCAGTTCTTCCTGC	quantitative
Gal4_105F	TTCTCCTGGCTCAGTAGGGC	Both
Gal4_205R	AGTTACGAGAGGGTGGACGGT	Both
Swi5_1515F	TCCTCAATTCGGCACACACA	quantitative
Swi5_1615R	CGATTGAACCTCTGGGCAGT	quantitative
Ash1_1267F	TCATCTCCATCTCCCTCCACA	quantitative
Ash1_1367R	GGTGACCTTGGGCTTGGAGT	quantitative
Gal80_368F	TGAAACTTGAAGGCGATGCC	quantitative
Gal80_468R	TTGTTGTCCATTGGCTAGCG	quantitative
Met16_618F	CAACGAACTTTTGGACCTTGG	quantitative
Met16_718R	TGCCCTTCCATCTTCCTGC	quantitative
Gal10_1356F	CGGCGTTAATGCGAATCAT	quantitative
Gal10_1456R	ACTCGGCGGTAAAAACATCCT	quantitative
HO_1619F	AAGGCGAAAAATTGGGCATT	quantitative
HO_1715R	CCGCGGACAGCATCAAACT	quantitative

Table 4. Oligonucleotides used for semi-quantitative and quantitative RT-PCR

Promoter	k_m	v_{max}/d	h	$\frac{v_{max}}{k_m \cdot d}$ (Prom. Strength)
SWI5→*HO* pr	1	0.1	1	0.1
ASH1→*HO* pr	0.0356	0.1	1	2.8
CBF1→*MET16* pr	0.0372	0.0427	1	1.14
GAL4→$GAL10_{glu}$ pr	0.09	0.5	4	5.5
GAL4→$GAL10_{gal}$ pr	0.01	4.5	4	450
SWI5→*ASH1* pr	1.814	0.604	1	0.33

Table 5. Fitted promoter strength parameters. Numbers refer to absolute values.

Par.	Description	Estimated value
k_1	M. M. const. ($SWI5 \rightarrow HO$ pr)	1 $[a.u.]$
k_2	M. M. const. ($ASH1 \rightarrow HO$ pr)	0.0356 $[a.u.]$
k_3	M. M. const. ($CBF1 \rightarrow MET16$ pr)	0.0372 $[a.u.]$
k_4	M. M. const. ($GAL4 \rightarrow GAL10$ pr)	(Glu) 0.0938 $[a.u.]$ (Gal) 0.01 $[a.u.]$
k_5	M. M. const.($SWI5 \rightarrow ASH1$ pr ($GAL80$))	1.814 $[a.u.]$
k_6	M. M. const. ($SWI5 \rightarrow ASH1$ pr ($ASH1$))	1.814 $[a.u.]$
α_1	Basal act. of the HO pr	0 $[a.u.\, min^{-1}]$
α_2	Basal act. of the $MET16$ pr	$1.49 \cdot 10^{-4} [a.u.\, min^{-1}]$
α_3	Basal act. of the $GAL10$ pr	$3 \cdot 10^{-3} [a.u.\, min^{-1}]$
α_4	Basal act. of the $ASH1$ pr ($GAL80$)	$7.4 \cdot 10^{-4} [a.u.\, min^{-1}]$
α_5	Basal act. of the $ASH1$ pr ($ASH1$)	$6.1 \cdot 10^{-4} [a.u.\, min^{-1}]$
v_1	Max. HO pr trans.	0.04 $[a.u.\, min^{-1}]$
v_2	Max. $MET16$ pr trans.	$8.82 \cdot 10^{-4}$ $[a.u.\, min^{-1}]$
v_3	Max. $GAL10$ pr trans.	(Glu) 0.0022 $[a.u.\, min^{-1}]$ (Gal) 0.0201 $[a.u.\, min^{-1}]$
v_4	Max. $ASH1$ pr ($GAL80$) trans.	0.0147 $[a.u.\, min^{-1}]$
v_5	Max. $ASH1$ pr ($ASH1$) trans.	0.0182 $[a.u.\, min^{-1}]$
d_1	Deg. rate of $CBF1$	0.0222 $[min^{-1}]$
d_2	Deg. rate of $GAL4$	0.0478 $[min^{-1}]$
d_3	Deg. rate of $SWI5$	0.4217 $[min^{-1}]$
d_4	Deg. rate of $GAL80$	0.0980 $[min^{-1}]$
d_5	Deg. rate of $ASH1$	0.05 $[min^{-1}]$
γ	Affinity in the law of $SWI5$	(Glu) 0.2 [a.u.] (Gal) 0.6 [a.u.]
ψ_1	Starv. effect on $GAL4$ deg.	0.2014 $[min^{-1}]$
ψ_2	Starv. effect on $GAL80$ deg.	0.1676 $[min^{-1}]$
h_1	Hill coeff. ($SWI5 \rightarrow HO$ pr)	1
h_2	Hill coeff. ($ASH1 \rightarrow HO$ pr)	1
h_3	Hill coeff. ($CBF1 \rightarrow MET16$ pr)	1
h_4	Hill coeff. ($GAL4 \rightarrow GAL10$ pr)	4
h_5	Hill coeff. ($SWI5 \rightarrow ASH1$ pr ($GAL80$))	1
h_6	Hill coeff. ($SWI5 \rightarrow ASH1$ pr ($ASH1$))	1
τ	Time delay	100 [min]

Table 6. Estimated Parameters for DE model

Chapter 5 - Design Principles for the Construction of an *in vivo* Benchmark

"Influential ideas are always simple. Since natural phenomena need not be simple, we master them, if at all, by formulating simple ideas and exploring their limitations."

Al Hershey

Our final goal was to build an *in vivo* gene network as ground of truth against which system and synthetic biology approaches can be assessed.

5.1 Choice of model organism

We chose as model organism yeast *Sacchamyces cerevisiae* because it is the simplest eukaryote and it shares both transcriptional machinery structure and gene transcription mechanisms with higher eukaryotes.

Considering basic biological concepts, the inside of a yeast cell looks more like that of a human cell than that of a bacterium. The DNA is wrapped around proteins called histones to form bead-like structures called nucleosomes, and the chromosomes are sequestered in a cellular compartment called the nucleus. For these and other reasons, yeast is classified as a eukaryote, as are humans, flies, worms and plants.

Most of what we know about eukaryotic gene regulation comes from studies of the yeast *Saccharomyces cerevisiae*. Expression of a typical eukaryotic gene is

more complex than is the one of a bacterial gene, because there can be different layers of control which involve the presence of nucleosomes and nuclei. As a matter of fact, nucleosomes modifications can affect protein binding to DNA, and the sequestration of genes in the nucleus implies that regulators often must move from one compartment to another in order to perform their task. We reasoned that, even if these biological processes are not explicitly formalized when we build the model of a biological system, it is essential to consider an organism that has these features. This will help to evaluate and compare the modelling assumptions, which are at the basis of different system biology approaches, and to understand their limits.

Among eukaryotes, the yeast has got other convenient features, which lead us to choose it as model. This organism grows rapidly, about 20-fold faster than mammalian cells and is only 3-fold slower than *Escherichia coli*. It is unicellular and can be easily cultured and manipulated. Mutants can be selected or recognised by simple assays, and sequences in and around genes can be altered at will. The genome is completely sequenced and comprises about 6000 genes, only about 2000 more than *E. coli*. Here, I will use the term yeast to refer to *S. cerevisiae*.

5.2 Choice of Network Motifs

In order to obtain a good benchmark, we aimed to construct a synthetic network that captures the behaviour of larger eukaryotic gene networks on smaller scale. Indeed, we decided to include in our network a variety of regulatory interactions, which are peculiar of transcriptional regulatory networks.

5.2.1 Network Motifs

Living cells are the product of gene expression programs involving regulated transcription of thousands of genes. Gene expression programs depend on recognition of specific promoter sequences by transcriptional regulatory proteins, including transcription factors. How a collection of regulatory proteins associates with genes across a genome can be described as a **transcriptional regulatory network**. In the network, the nodes are genes and the edges represent transcriptional regulation of one gene by the protein product of another gene. Thus, a directed edge X → Y means that the product of gene X regulates the transcription rate of gene Y. Since the cell is not an isolated system, but it continuously responds to external stimuli in order to follow a specific developmental program or to adapt to changing environmental conditions, the transcriptional network is a dynamic system: after an input signal arrives transcription factor activities change, leading to changes in the production rate of proteins.

In order to study the complex dynamics of cellular networks, during the last years several studies aimed to identify the basic building-blocks of transcriptional networks and to study the functional relevance of these modular components (Lee et al., 2002; Milo et al., 2004; Milo et al., 2002; Shen-Orr et al., 2002; Yeger-Lotem and Margalit, 2003; Yeger-Lotem et al., 2004). The approach is based on the identification of meaningful patterns on the basis of statistical significance. To define statistical significance, the real network is compared to an ensemble of randomized networks, which have the same number of nodes and edges as the real one, but where the connections are made at random. If a pattern occurs in the real network significantly more often than in the randomized networks, it is defined as a **network motif**. The basic idea is that network motifs that occur in the real network more often than in randomized networks must have been preserved over evolutionary timescales against

mutations that randomly change edges. As a matter of fact, point mutations, which occur in a promoter sequence, can alter the binding of a specific transcription factor to the promoter thus resulting in the loss of an edge of the transcriptional network. Similarly, new edges can be added to the network by either point mutations or by duplication events in a promoter region, thus generating a new binding site for a transcription factor. Hence, conserved network motifs must have been selected in order to survive during evolution because they provide some advantage to the organism. If a motif did not offer a selective advantage, it would be washed out and occur about as often as in randomized networks.

The first studies, which aimed to systematically identify network motifs, were done in simple organisms such as *Escherichia coli* and *Saccharomyces cerevisiae* since their regulatory networks has been extensively studied and a large amount of information about direct transcriptional interactions can be found in databases such as RegulonDB and SGD respectively (Alon, 2006; Harbison et al., 2004; Lee et al., 2002; Shen-Orr et al., 2002). They found that much of the transcriptional network is made of repeating occurrence of some network motifs. The fact that the network motifs appear at frequencies much higher than expected at random suggests that they may have specific functions in the information processing performed by the network. Mathematical analysis of the dynamics of these simple motifs helped to associate at each one a specific function, as exemplified by the two examples reported in the following paragraph (Shen-Orr et al., 2002).

5.2.2 Network Motifs are associated with specific functions

One of the motifs, which are found highly represented in both *E. coli* and *S. cerevisiae*, is the 'feed-forward loop' (Figure 5.1A). This motif is defined by a transcription factor (X) that regulates a second transcription factor (Y), such that both X and Y jointly bind a common target (Z).

Mathematical analysis suggest that this motif may act as a switch that rejects transient activation signals and responds only to persistent ones, while allowing a rapid system shut-down. This can occur when both transcription factors (X and Y) are required to activate Z (that is they act as an 'AND-gate'). When X is activated, the signal is transmitted to the output Z by two pathways, a direct one from X and a delayed one through Y. If the activation of X is transient, Y cannot reach the threshold level to activate Z, and the input signal is not transduced through the circuit. Only when X signals for a long enough time so that Y can accumulate and reach the appropriate levels, Z will be activated (Figure 5.1B). Since expression of the ultimate target gene (Z) may depend on the accumulation of adequate threshold levels of the first (X) and the second regulators (Y), the feed-forward loop can also provide temporal control of a process.

Another motif, which can be used in the cell for both ordering events in a temporal sequence and responding to sustained signals rather than transient ones, is the 'regulator chain' motif. It consists of chains of three or more transcription factor in which one regulator binds the promoter for a second regulator, and the second binds the promoter for a third regulator and so forth. Compared to the feed-forward loop, the regulator chain motif has a slower system shut-down (thin red line in the Z dynamics panel of figure 5.1B).

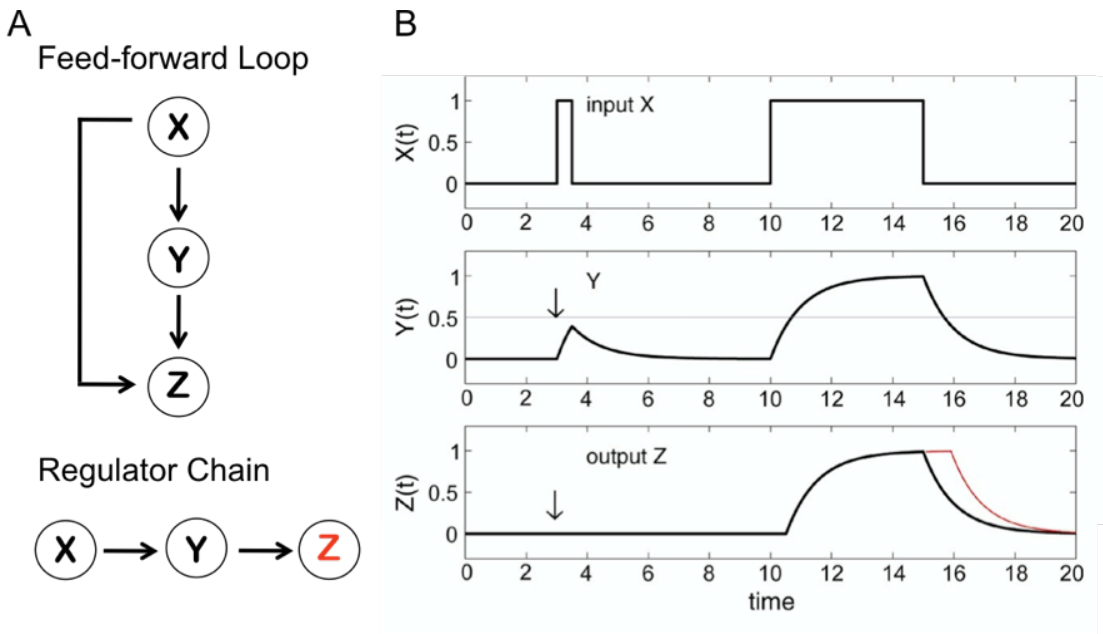

Figure 5.1 Dynamic features of the Feed-forward Loop and the Regulator Chain Motifs. (A) Schematic representation of a coherent Feed-forward Loop and of a Regulator Chain. **(B)** These circuits can reject transient variations in the activity of the input X (left side of the graph, indicated by an arrow), and respond only to persistent activation profiles (right side of the graph). This is because Y needs to integrate the input X over time to pass the activation threshold for Z (thin line). The case of regulator chain motif has a slower shut-down than the feed-forward loop (thin red line in the Z dynamics panel).

5.2.3 Network Motifs in yeast and higher eukaryotes

Even if various network motifs are common both in bacteria and yeast, there are some motifs that are more represented in eukaryotes than in prokaryotes and could therefore reflect differences in the regulatory mechanisms adopted by these organisms. For this reason we focused our attention on network motifs which occur more often in yeast than in bacteria (Lee et al., 2002; Yeger-Lotem et al., 2004).

Motifs identified in a study of genome-wide location analysis of all 141 transcription factors listed in the Yeast Proteome Database (Lee et al., 2002) are

depicted in Figure 5.2. They include the feed-forward loop and the regulator chain motifs (described above) among the most frequent ones. In particular, 49 feed-forward loops, and 188 chain motifs, which varied in size from 3 to 10 regulators, were identified. However, Lee et al identified the other frequent motifs, which we discuss below: the Auto-regulation, the Multi-Component Loop and the Single Input Motif.

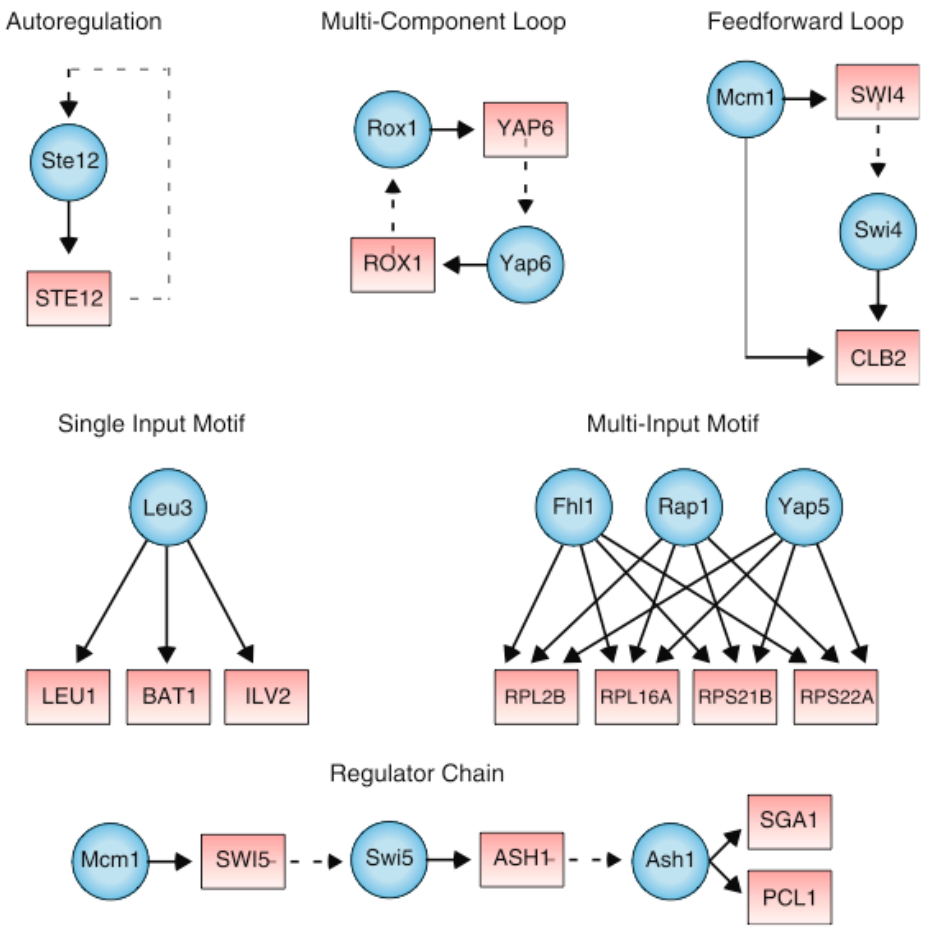

Figure 5.2 Examples of network motifs in the yeast regulatory network. Schematic representation of network motifs identified by Lee et al (Lee et al., 2002). Regulatory proteins are represented as blue circles, while their target promoters as red rectangles. A solid arrow indicates binding of a regulator to a promoter. The dashed arrow links the gene to its protein product, representing transcription and translation processes.

An ‘Auto-regulation’ motif consists of a transcription factor that binds its own promoter. This motif is though to reduce response time to environmental stimuli and increase stability of gene expression. For example, upon exposure to mating pheromone, the concentration of the pheromone-responsive TF, Ste12, rapidly increases because it auto-sustains its own gene expression. The consequent progressive increase of Ste12 protein leads to the binding of other genes required for the mating process. Only 10% of yeast regulators are autoregulated. In contrast, studies of *E. coli* networks indicate that most (52% to 74%) prokaryotic genes encoding transcriptional factors are autoregulated; indicating that eukaryotes probably evolved different and therefore more frequent mechanisms for stabilizing gene expression, such as feed-back loops.

A ‘Multi-component Loop’ motif consists of a regulatory circuit whose closure involves two or more factors. The closed loop structure provides the capacity for the feedback control and offers the potential to produce bi-stable systems, which switch between two alternative states. This motif is peculiar of yeast and of developmental networks of higher eukaryotes since, except the auto-regulation, feedback loops composed only by direct transcriptional interactions have not been identified in bacteria. This observation contributes to sustain the hypothesis that eukaryotes and bacteria often use different peculiar motifs for the same purpose.

Another motif, which can give rise to bi-stability and is often used by cells to respond to environmental stimuli, is the ‘Mixed Negative Feedback Loop’ is. This feedback motif is composed of one transcriptional and one protein-protein interaction and is a common network motif in many organisms (Figure 5.3B) (Lahav et al., 2004). In this motif, protein X is a transcription factor that activates the transcription of gene Y. The protein product of gene Y in turn interacts with X at the protein level, often in

a negative fashion (Figure 5.3A). This negative regulation can take several forms. In some case, Y enhances the rate of degradation of protein X, such as in mammals in the case of p53 transcriptionally activating Mdm2, which in turn targets p53 for degradation by protein-protein interaction, or as Ime1-Ime2 in *S. cerevisiae* and σ_H-dnaK/J in *E. coli*. In other cases, Y binds X and inhibits its activity as transcription factor by preventing its access to DNA, such as in the case of NFκB activating the transcription of its inhibitor IκB, which sequesters it in the cytoplasm thus preventing its entry in the nucleus where it works as transcription factor.

The 'Single Input' motif contains a single regulator that binds a set of target genes. In this way the expression of the target genes is coordinated under a specific condition. This is useful for coordinating the components of the same cell structure or the enzymes of a specific metabolic pathway, so that their proportion at the steady state can be fixed. For example, several enzymes required for the galactose catabolic pathway are controlled by Gal4 transcriptional factor in response to the presence of carbon source in the medium. In addition, mathematical analysis suggests that this motif can also result in a temporal program of expression resulting from differences in the activation threshold of the different target genes.

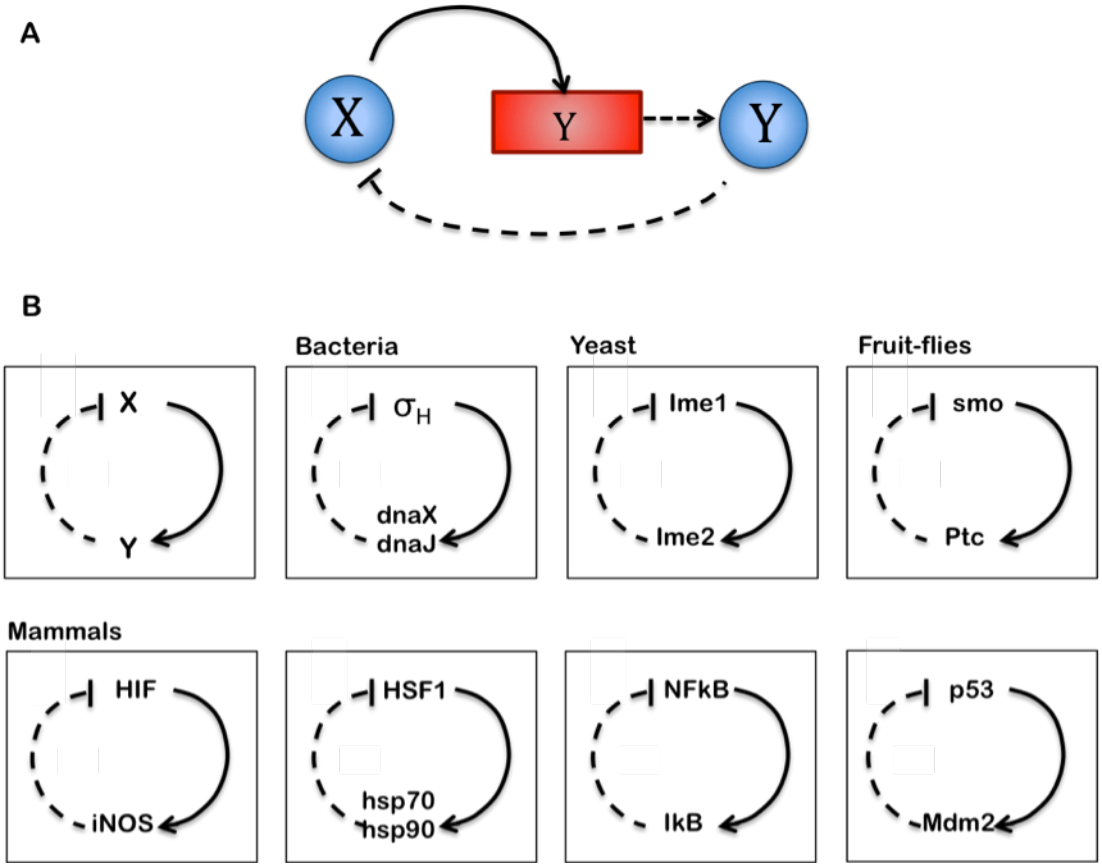

Figure 5.3 Negative Feedback Loop Motif. (A) A generic scheme of the negative feedback loop motif is shown. The regulatory protein X activates the transcription of the Y gene. Then Y inhibits X by a protein-protein interaction. Blue circles represent proteins and the rectangle represents gene promoter. **(B)** Examples of negative feedback loops with one transcription arm and one protein-protein arm are present in diverse systems of mammals (p53-Mdm2 and NFκB-IκB), fruit fly (HSF1-hsp70), yeast (smo-Ptc and Ime1-2) and bacteria (σ_H-dnaK/J).

5.2.4 Choice of Network Motifs in IRMA construction

In order to construct our synthetic network, we combined some of the described motifs in a way to obtain a circuit in which some components (Gal80 and Cbf1) can respond to different signals (galactose and methionine concentrations) from the environment and propagate the signal to the rest of the network. We decided to use the Regulator Chain (Cbf1-Gal4-Swi5 regulators) and the Single Input (Swi5 which activates three promoters that have different threshold) motifs in order to have a sequence of transcriptional events, which can be separately analyzed in time. We added to them both a positive (Swi5 activates *HO* transcription thus closing the circuit) and a negative transcriptional feedback loop (Ash1 represses *HO* transcription) thus obtaining a Multi Component Loop, with the aim of enriching the dynamic behavior of the network. Finally, in order to provide the circuit of a switch, we also used a negative feedback loop composed of a protein-protein interaction (Gal80-Gal4) that can turn off the system in response to an external stimulus (depletion of galactose from the culturing medium).

Indeed, our network, apparently simple, is articulated in its interconnections, which include a variety of regulatory interactions, thus capturing some features of larger eukaryotic gene networks on a smaller scale.

Chapter 6 – Results.

Construction and Characterization of a Gene Synthetic Network in Yeast

6.1 Construction of a Gene Synthetic Network in yeast

We designed a synthetic gene network of five genes for In vivo Reverse-engineering and Modelling Assessment (IRMA). The network, depicted in Figure 6.1A, is organized in such a way that each gene controls transcription of at least another gene in the network. In addition, it can be 'switched' on or off by culturing cells in galactose or in glucose, respectively, and it can be modulated by using different methionine concentrations in the growing medium.

6.1.1 Choice of Network Genes

Particular care was taken in the choice of genes in order to isolate the network from cellular environment. We searched in literature and in the SGD (*Saccharomyces cerevisiae* Genome Database; www.yeastgenome.org) for genes, which show some essential features.

- We chose non-essential and non-redundant TF-genes, which do not show synthetic lethality and, therefore, can be knocked out without affecting yeast viability.

- We selected well-characterised promoter/TF-encoding-gene pairs, belonging to distinct and non-redundant pathways, to further minimize external feedbacks on the network due to pathway crosstalk.

- We chose promoters for which a single transcription factor (TF) is sufficient and essential to activate transcription. Thus, by removing the endogenous TF, we maximally reduced influences from the cellular environment on each promoter.

Specifically, we selected as activators and repressors encoding genes: *SWI5, ASH1, CBF1, GAL4 a*nd *GAL80;* as promoter genes: *HO, ASH1, MET16* and *GAL10* (Figure 6.1A).

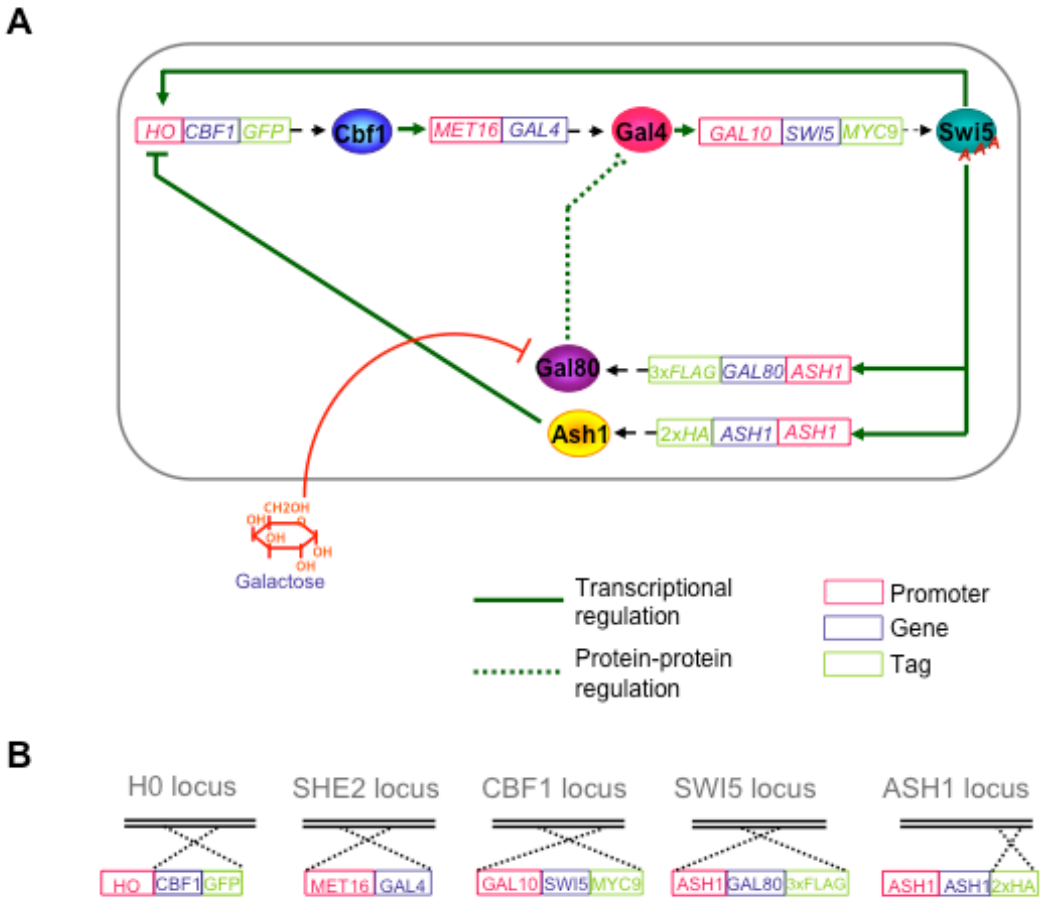

Figure 6.1. Construction of IRMA, a synthetic network in yeast. (A) Schematic diagram of the synthetic gene network is represented. New transcriptional units (rectangles) were built by assembling promoters (red) with non-self coding sequences (blue). Genes were tagged at the 3' end with the specified sequences (green). Each cassette encodes for a protein (represented as a circle) regulating the transcription of another gene in the network (solid green lines). The resulting network, IRMA, is fully active when cells are grown in presence of galactose, while it is inhibited by the Gal80-Gal4 interaction in presence of glucose. **(B)** Schematic diagram of genomic integrations of IRMA genes. Each cloned cassette was integrated by homologous recombination in a specified genomic locus of a Δgal4 Δgal80 Saccharomyces cerevisiae strain to contemporarily delete (*CBF1, SWI5, SHE2*) or to modify (*ASH1* tagging, *CBF1* integration under *HO* promoter) endogenous genes. *ACE2* gene deletion was achieved by integrating a drug resistance cassette, *natMX4* (not shown).

6.1.2 Selected promoter/TF-gene pairs

The first selected promoter/TF-gene pair in the network is the *HO* promoter controlled by two TFs: a cell-cycle independent Swi5 mutant ($swi5_{AAA}$), and Ash1 (Moll et al., 1991; Nasmyth et al., 1987). Since *ASH1* transcription is also controlled by Swi5, we chose as the second promoter/TF-gene pair the *ASH1* promoter controlled by $swi5_{AAA}$.

Swi5 mediates specific *HO* expression in the late G1 phase (Nasmyth et al., 1990). It is retained in the cytoplasm by Cdk8 phosphorylation and enters the nucleus to regulate transcription only in late anaphase, when Cdc14 dephosphorylates it (Visintin et al., 1998).

In order to overcome Swi5-mediated cell cycle control of the *HO* promoter in the network, we used the $swi5_{AAA}$ mutant in which the three phosphorylated serine residues (Ser-522, Ser-646, and Ser-664) are substituted by alanines. These mutations lead to constant Swi5 accumulation into the nucleus throughout the cell cycle (Moll et al., 1991).

Specific expression of *HO* in mother cells is achieved via Ash1-mediated repression of *HO* in daughter cells only (Bobola et al., 1996; Cosma, 2004; Jansen et al., 1996). In order to obtain a symmetrical Ash1 distribution in both mother and daughter cells, we also planned to delete the *SHE2* gene whose mRNA localizes Ash1 in daughters (Gonsalvez et al., 2003; Long et al., 1997). We thus obtained a homogeneous population of cells, where *HO* transcription is not developmentally regulated. In addition, we deleted Ace2 that cooperates with Swi5 in regulating the *ASH1* promoter (Voth et al., 2007).

The third selected promoter/TF-gene pair was the *MET16* promoter/*CBF1*. Cbf1 is a DNA binding protein that controls chromosome segregation and sulphur amino acids metabolism (Mellor et al., 1990). Specific Cbf1 binding upstream of the *MET* genes is required for its function during transcriptional activation even if it is not sufficient to activate transcription alone. Cbf1 tethers the activator Met4 to the *MET* promoters and forms the Cbf1-Met4-Met28 complex, which triggers expression of *MET* genes (Kuras et al., 1997). Among its transcriptional targets, we chose *MET16* since it is the only *MET* gene that strictly depends on the binding of Cbf1 (Ferreiro et al., 2004; O'Connell et al., 1995), while the others can still be expressed at a lower level in its absence (Kuras and Thomas, 1995).

In order to add a signalling molecule able to activate expression of network genes, we chose as the fourth and last promoter/TF-gene pair the *GAL1-10* promoter, which is tightly regulated by the carbon source via the Gal4 transcription factor. In the presence of galactose, Gal4 activator binds to the multiple UAS_{GAL} elements in the promoter and leads to activation of transcription. In absence of galactose, Gal4 is inactive because of the binding of Gal80 repressor to its activation domain, which prevents the interaction of the transcription machinery (Traven et al., 2006).

6.1.3 Network Transcription Factors are essential and sufficient for transcription of their target promoter

We experimentally verified that the selected TFs are essential and sufficient for transcription of their target promoter. To this aim we constructed yeast strains in which one of the selected TFs was deleted and, therefore, analysed the transcription of

the target promoter in the absence and in the presence of each TF by semi-quantitative RT-PCR (Figure 6.2).

We analysed transcription of both *HO* and *ASH1* genes in absence of Swi5, of its homologous Ace2 and of both (Figure 6.2A). In Δ*ace2* strain, both *HO* and *ASH1* transcription is decreased but not abolished, while in Δ*swi5* strain *HO* is not detectable thus confirming that Swi5, but not Ace2, is essential for activating *HO* transcription. Conversely, *ASH1* transcription is abolished only in the Δ*ace2*Δ*swi5* strain indicating that both Swi5 and Ace2 are essential for activating this gene.

In order to confirm that Cbf1 is essential for *MET16* transcription we analysed the wild type and the Δ*cbf1* strains. Figure 6.2B shows that *MET16* transcript is not detectable in absence of Cbf1.

Finally, we analysed Gal4 effect on *GAL10* expression (Figure 6.2C). Since Gal4 is required for the activation of the GAL genes in response to galactose, in the wild type strain expression of *GAL10* is detectable only when yeast cells are cultured in presence of this sugar, while it is not transcribed in glucose due to Gal80 inhibition. Yeast cells, which are null for Gal4, are not viable in presence of galactose as the only carbon source, for this reason we grew Δ*gal4* and Δ*gal4*Δ*gal80* strains in the presence of both galactose and raffinose. We show that, even in absence of Gal80 repressor, *GAL10* is not expressed in the absence of Gal4. We thus confirmed that Gal4 is essential for the transcription of *GAL10* gene.

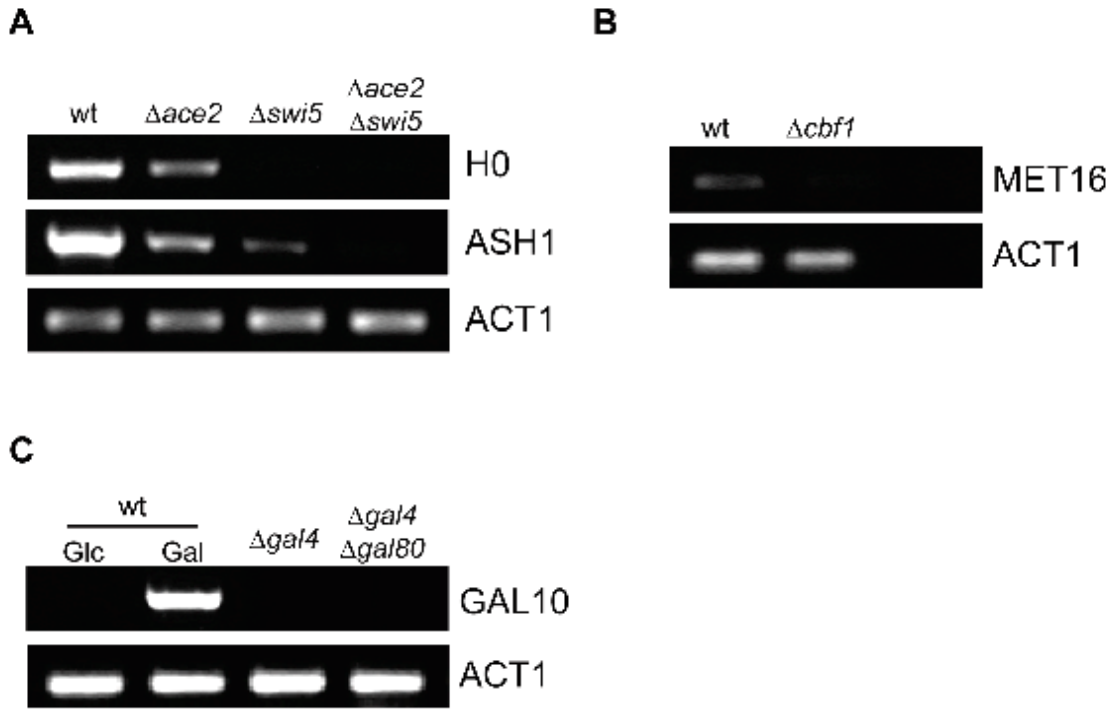

Figure 6.2. *HO*, *MET16* and *GAL10* promoters are not transcribed in absence of their specific activators. Semi-quantitative RT-PCRs were carried out to amplify the indicated genes (oligonucleotides are listed in Table 4). Cells were grown in YEPD or YEPGR at 30 °C up to mid-log phase. Strains used: W303 (wt); Δ*ace2* (P302); Δ*swi5* (P15); Δ*swi5ace2* (P304); Δ*cbf1* (P366); Δ*gal4* (P274); Δ*gal4gal80* (YM4271). **(A)** Transcription of *HO* and *ASH1* are dependent on Swi5 and Ace2 TFs. *HO* was not transcribed in *swi5* deletion strain; *ASH1* transcription was inactivated in the *ace2swi5* double deletion mutant. **(B)** Cbf1 is essential for *MET16* transcription. **(C)** *GAL10* is transcribed only in the presence of Galactose in wt cells. *GAL10* transcription is abolished in cells lacking Gal4 TF or both Gal4 and Gal80. Δ*gal4* and Δ*gal4gal80* were grown in YEPGR.

6.1.4 Synthetic Network Construction by contemporary gene knock-in and knock-out

We assembled the chosen promoters upstream of non-self gene coding sequences to obtain the IRMA network. The network (Figure 6.1A) includes positive and negative feedback loops, and other network motifs discussed in the previous chapter. These interactions were selected because they coexist normally in many sensory and developmental networks in higher eukaryotes as discussed in Chapter 5.

We combined minimal regions of the chosen promoters upstream of the chosen TF encoding genes, in vectors containing different yeast selectable markers. Thus, we built the following new transcriptional units: *HO* promoter/*CBF1-GFP*, *MET16* promoter/*GAL4*, *GAL1-10* promoter/*SWI5-MYC9*, *ASH1* promoter/*GAL80-3XFLAG* and *ASH1* promoter/*ASH1-2XHA* (Figure 6.1). A fluorescent tag was cloned at the 3' end of the *CBF1* ORF to easily monitor its protein product.

We integrated these cassettes in the genome of a *gal4*Δ*542 gal80*Δ*538* yeast strain (YM4271 strain) whose *GAL4* and *GAL80* loci were deleted (Liu et al., 1993). We targeted the cassettes in specific genomic loci to simultaneously integrate the newly built transcriptional units, and delete all the endogenous counterparts of our network genes, thus minimising influences from endogenous genes (Figure 6.1B).

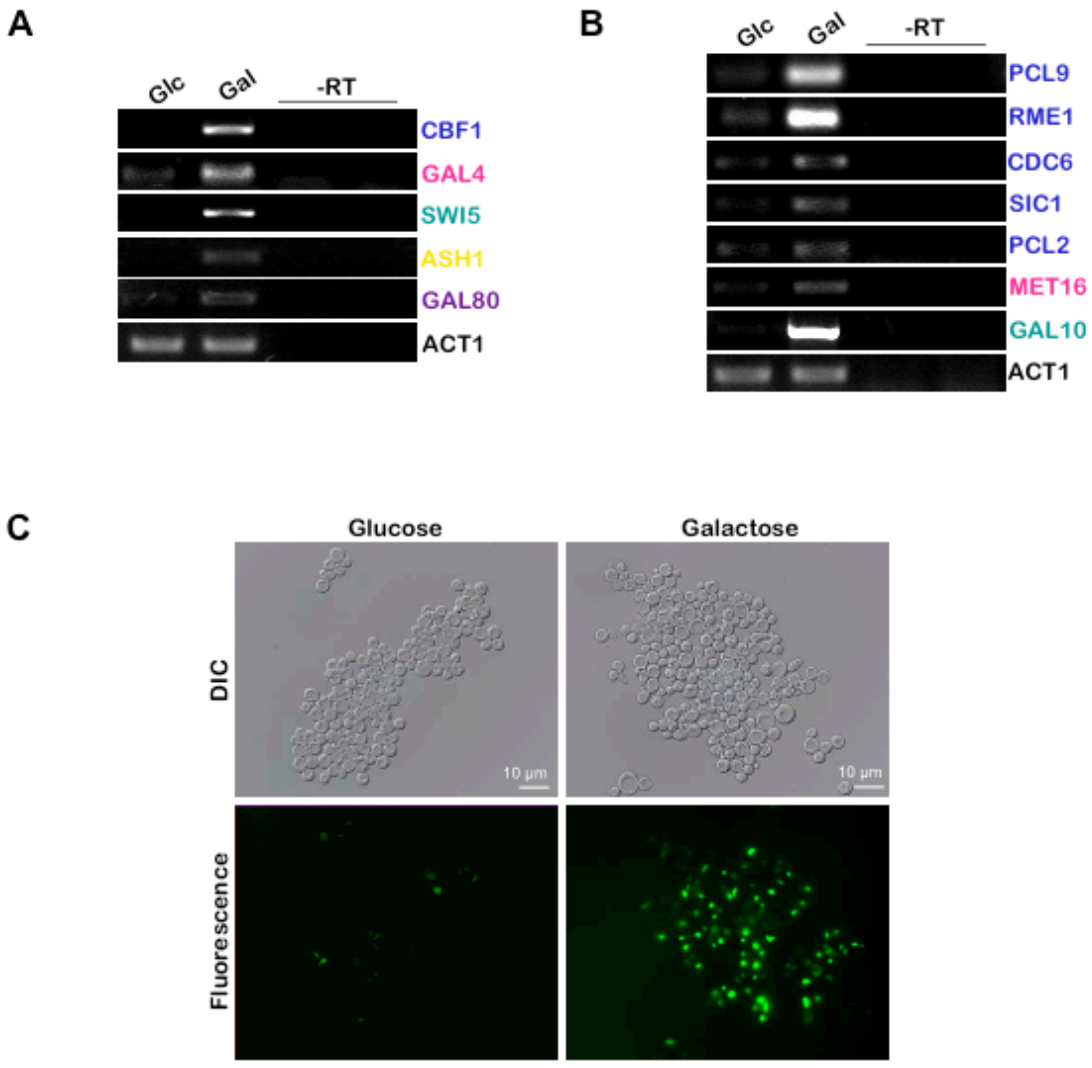

Figure 6.3. Galactose triggers activation of IRMA synthetic network. (A)-(B) Network genes, and cell genes that are network targets, are expressed only in the presence of galactose. Semi-quantitative PCR to amplify IRMA and IRMA-dependent genes was carried out using total RNA extracted from cells grown in Glucose (Glc) or Galactose-Raffinose (Gal) containing medium. **(C)** Live imaging of IRMA cells grown in glucose and galactose containing medium. Scale bar, 10µm; 63x magnification.

6.2 Network genes, and their endogenous targets, are activated by galactose

We tested transcription of network genes upon culturing cells in presence of galactose or glucose. Galactose activates the *GAL1-10* promoter, cloned upstream of $swi5_{AAA}$ in the network, and it is able to activate transcription of all the five network genes (Figure 6.3A).

We also checked for protein expression of Cbf1-GFP. Living yeast cells grown with different carbon sources (galactose or glucose) were analyzed by fluorescent microscopy. As shown in Figure 6.3C, positive green cells were visualized only when IRMA was cultured in galactose-containing medium.

Endogenous yeast genes, not included in the synthetic network, but under transcriptional control of IRMA genes, such as *PCL9*, *RME1*, *CDC6*, *SIC1* and *PCL2*, targets of Swi5, and *MET16*, target of Cbf1, which are not controlled by galactose in wild type yeast, became galactose dependent; furthermore *GAL10*, which is not expressed in the YM4271 background, became network and galactose dependent (Figure 6.3B). These genes should not influence the network behaviour by means of direct or indirect feedback loops, since their functions are unrelated to any known regulation of the chosen promoters. In conclusion, the synthetic network can regulate external genes, but is very robust against regulatory inputs from the rest of the genome.

6.3 Network Genes are modulated by Methionine

In wild type yeast cells, Cbf1, together with Met4 and Met28 proteins, activates the transcription of genes involved in the methionine biosynthesis pathway. Methionine modulates the expression of the *MET* genes by affecting the formation of the Cbf1-Met4-Met28 transcriptional complex (Kuras et al., 1997; Thomas and Surdin-Kerjan, 1997). High levels of methionine increase the ubiquitination and the subsequent degradation of the activator Met4, indeed inhibiting the transcription (Chandrasekaran et al., 2006; Chandrasekaran and Skowyra, 2008; Menant et al., 2006). Conversely, low levels of methionine lead to an increase in transcription levels. In figure 6.4, we analysed the expression levels of various *MET* genes (*MET16*, *MET10*, *MET14* and *MET25*) when yeast cells were grown in the presence of "low" (10 μM) or "high" (1 mM) methionine concentration by both semi-quantitative and quantitative real-time RT-PCR (q-PCR). The levels were compared with the standard yeast growing condition ('control' lane in 6.4A and normalization condition in 6.4B), in which we performed all the other experiments. In the control, yeast cells were grown in the standard complete medium (YPD), which contains an intermediate concentration of methionine (about 140 μM), and thus show an intermediate level of *MET* genes expression. Some *MET* genes, such as *MET16* and *MET10*, are quickly down regulated as methionine concentration increases, thus showing 'control' levels which are more similar to the low methionine ('off' state) than to the high methionine ('on' state) condition (Figure 6.4 A and B). Conversely, *MET14* and *MET25* show 'control' expression levels nearer to the 'on' state than to the 'off' one thus reflecting a slower kinetic in response to methionine; furthermore are not fully turned off at high methionine concentration (Figure 6.4 A and B).

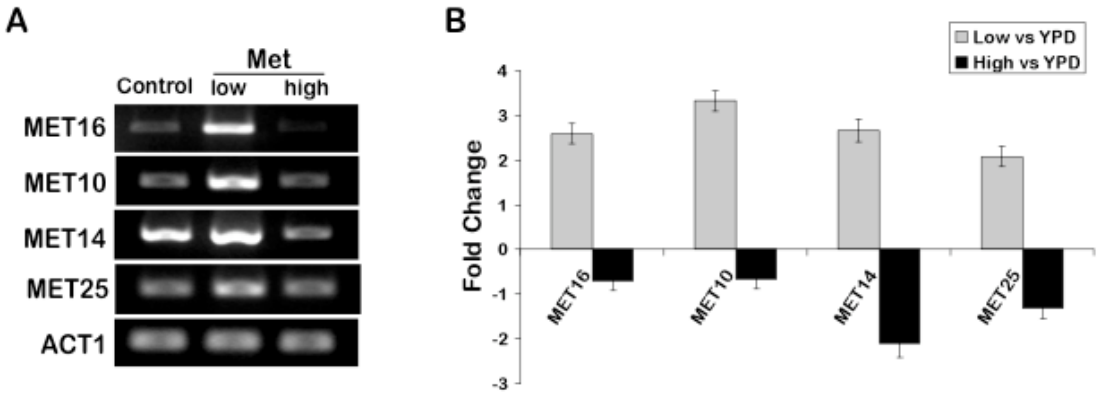

Figure 6.4. Expression of *MET* genes in wild type yeast cells. *MET* genes regulated by Cbf1 are transcriptionally activated in the presence of low levels of methionine ("low" lane in panel A and grey bars in panel B), while they are repressed at high methionine concentrations ("high" lane in panel A and black bars in panel B). Semi-quantitative **(A)** and quantitative **(B)** RT-PCR of *MET* genes were performed on total RNA extracted from yeast cells grown in the standard complete medium (YPD; "control" lane in panel A and normalization condition in panel B) and at two different methionine concentrations ("low" and "high" correspond to 10μM and 1mM, respectively). Data in panel B represent the ΔΔCt (mean ± SEM; n=2), which were calculated as the difference between the test condition ("low" or "high" methionine levels) and the standard condition (YPD medium).

In IRMA network we used *MET16* promoter to regulate *GAL4* expression. As shown by data in figure 6.4, *MET16* expression is tightly regulated by methionine concentrations since it is completely turned off in the presence of high methionine levels and, even at intermediate methionine levels (the 'control' condition), its transcription appears to be strongly decreased. We thus also tested transcription of network genes at steady state upon culturing cells in the presence of different concentrations of methionine ("low", "high" and "control" conditions are defoned as above explained) both in glucose and in galactose containing medium (Figure 6.5).

Even in the presence of glucose, when normally the network is turned off in the control standard growing condition (look at figure 6.3A and 'control' bars of figure 6.5), network genes are activated in low methionine containing medium, and reach the same expression levels that they have in the cells grown in the control in the presence of galactose, that is when the network is turned on. These data show that the increased *GAL4* expression, due to *MET16* activation after the removal of methionine, turns on all the network genes, while addition of methionine inhibits them, independently from galactose.

In the presence of galactose, when the network is turned on in the control condition, the increase of methionine concentration ("high" methionine) leads to a down regulation of network genes at levels, which are comparable to the ones of the control in glucose, and is therefore turned off. Conversely, at low methionine concentration, expression levels of network genes are even higher than the control ones in galactose.

Curiously, in the presence of galactose, *GAL4* steady-state levels, which are directly triggered by *MET16* promoter, do not show significant variations in response to methionine. This effect can have different explanations. Considering that low

methionine triggers the stabilization of Met4 protein, which forms a complex with Cbf1 and Met28, if we assume that Met28, and not Met4 or Cbf1, is the limiting factor for the formation of the transcriptional complex, neither Met4 stabilization or Cbf1 increase can lead to a further *MET16* activation when at the steady state all the Met28 protein has been recruited in the complex. A second possibility is that *MET16* promoter has already reached its maximum transcription rate or it is in proximity of its saturation level in the control condition in the presence of galactose. As a consequence, the *CBF1* increase, which results at the steady state in low methionine, has only a small effect on *GAL4* expression.

Anyway, when low methionine is compared to the control both in the presence of galactose, there is a strong increase in *SWI5* and consequently in its target genes (*CBF1*, *GAL80* and *ASH1*) expression. This effect is due to Gal4 protein stabilization in the presence of galactose (Muratani et al., 2005; Nalley et al., 2006), which at the steady state leads to the amplification of the small difference, which is seen at the transcriptional level.

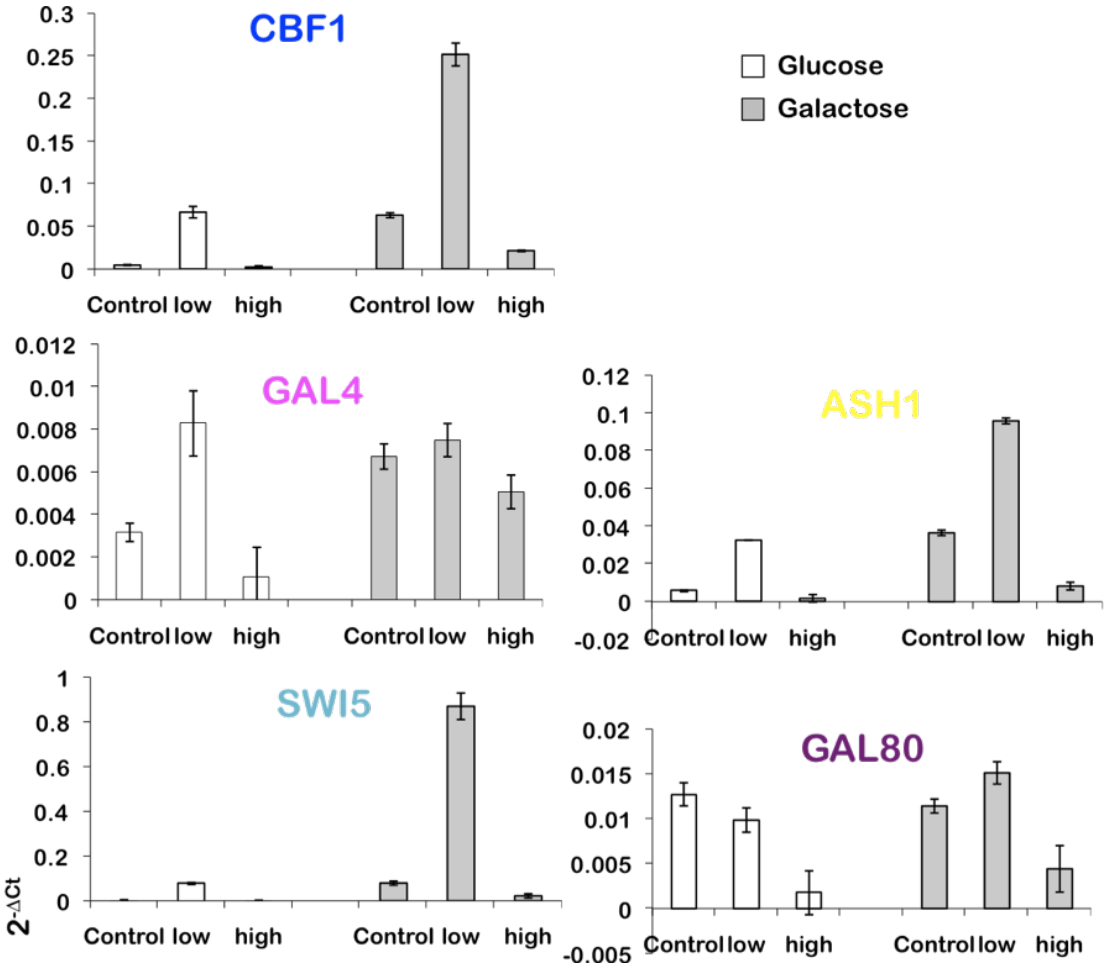

Figure 6.5. Methionine modulates IRMA genes expression. Expression levels of network genes at different methionine concentrations ("low" and "high" correspond to 10μM and 1mM, respectively; "control" is the standard complete medium, YEP, which contains 140μM)) in glucose (white bars) or in galactose/raffinose (grey bars) are shown. Data represent the $2^{-\Delta Ct}$ (mean ± SEM; n = 2).

6.4 Gene expression profiling of IRMA to study its dynamic behavior

In order to analyse the dynamic behaviour of the IRMA network, we performed perturbation experiments by shifting cells from glucose to galactose ("switch-on" experiments) and from galactose to glucose ("switch-off" experiments). We collected samples every 20 minutes up to 5 hrs in five independent experiments, for the switch-on dataset, and every 10 minutes up to 3 hrs, in four independent experiments, for the switch-off dataset. We analysed expression profiles of network genes by q-PCR.

In the switch-on experiment in Figure 6.6, the activation of *GAL4* by galactose led to transcription of all the other network genes. Their dynamic behaviour is evident; a seemingly oscillatory behaviour is present in *SWI5* with two peaks at 40 min and 180 min. The Swi5 targets, *CBF1*, *GAL80* and *ASH1*, are activated with different types of kinetics: *CBF1* is delayed with respect to the other two genes. This delay is due to the sequential recruitment of chromatin modifying complexes to the *HO* promoter, which follow the binding of Swi5 and other transcription factors. These events occur with a precise timing before *HO* transcription is finally triggered (Bhoite et al., 2001; Cosma et al., 1999). Of note, dynamics of *GAL80* and *ASH1* mRNAs are different. This is due both to differences in their degradation rates, and to the effect of cell manipulation on *GAL80* and *GAL4*. Specifically, the first point of the switch-on time-series, in Figure 6.6, was measured in glucose, right before shifting cells from glucose to galactose. During the standard washing steps, when the glucose medium is removed and the fresh new galactose-containing medium is added to the cells, we

observed a transient increase in mRNA levels of *GAL4* and *GAL80* (Figure 6.6, gray bar).

In order to check whether this effect was independent from galactose administration, we performed an *ad hoc* glucose-to-glucose shift experiment (Figure 6.7). *GAL4* and *GAL80* showed the same increase, once the cells were transferred back in the glucose medium, after the washing steps (Figure 6.7). We believe this increase is due to the transient deprivation of carbon source during the washing steps, which attenuates the degradation levels of *GAL4* and *GAL80* mRNAs (Jona et al., 2000). This effect is unrelated to their transcriptional regulation, being these two genes controlled by different promoters. Moreover, the expression levels of the *MET16* endogenous gene, whose promoter, in our network, is the same promoter as *GAL4*, do not show any increase in the glucose-to-glucose shift (Figure 6.7) and in both the switch-on and switch-off experiments (Figure 6.8), further excluding dependence on transcriptional regulation.

In the switch-off experiment (Figure 6.6), as expected, the transcription of the whole network is rapidly turned off with a delay in the silencing of *CBF1* expression.

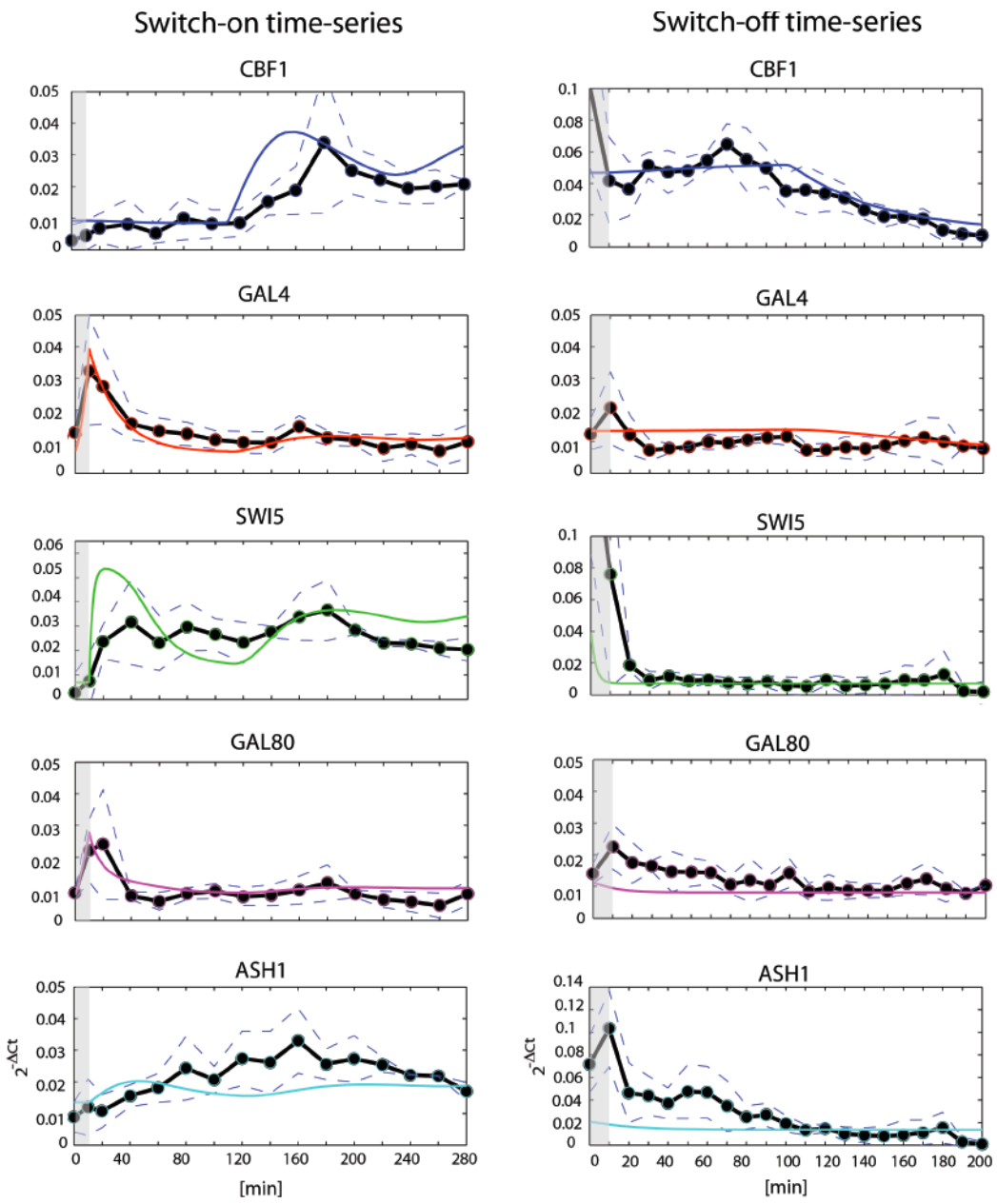

Figure 6.6. IRMA dynamic behavior in response to medium shift perturbations. Time-series expression profiles of network genes after a shift from glucose to galactose-raffinose containing medium – switch-on - (left) and from galactose-raffinose to glucose containing medium – switch-off - (right) are shown. Circles represent average expression data for each of the IRMA genes at different time points. Dashed lines represent standard errors. Continuous colored lines represent *in silico* data obtained from the DE–based model and show how the model fits experimental

data. Gray bar indicates the 10 min interval during which the washing steps and subsequent medium shift are performed (see main text). The first point in the switch-on time-series (left) is measured in glucose right before shifting the cells to galactose; the second point at 10 min is the first one in galactose just after the shift has occurred. Similarly (right) the first point in the switch-off time-series is measured in galactose before shifting the cells to glucose. In the switch-off experimental data, the first point of SWI5 at time 0 is off scale, with a value of 0.18. This was done to better show its behavior in the figure. Represented data are the $2^{-\Delta Ct}$ (mean ± SEM; n = 5 for switch-on and n = 4 for switch-off).

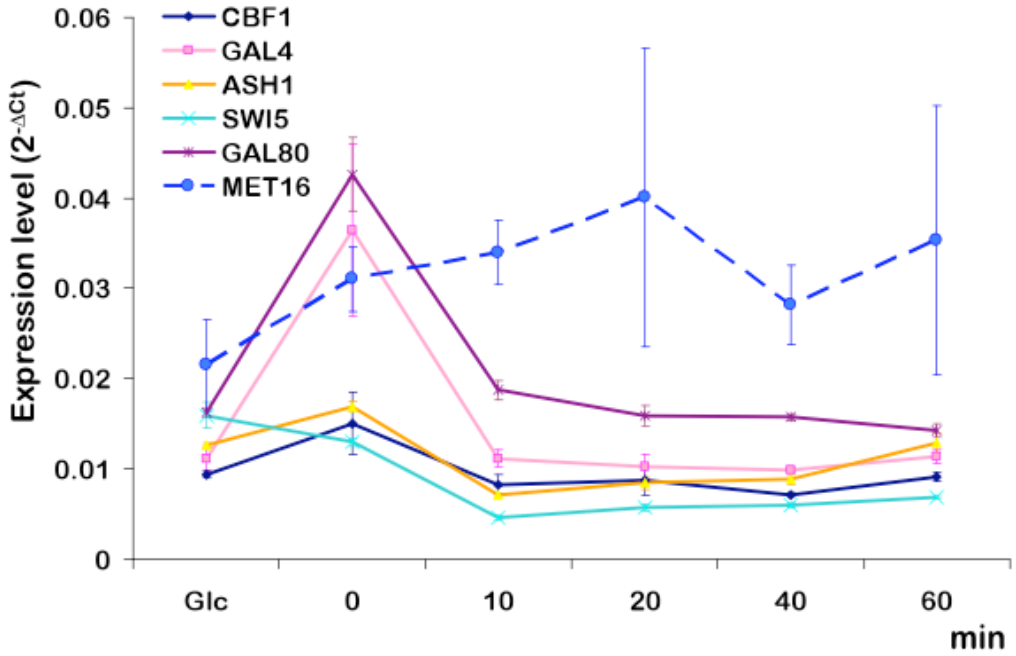

Figure 6.7. *GAL4* and *GAL80* increase after sugar shift is an IRMA independent effect. Expression profiles of the five IRMA genes and the *MET16* endogenous gene in glucose-to-glucose time-series are shown. Yeast cells were grown in glucose up to mid-log phase and then shifted back to glucose containing medium (0 time point), after filtering and washing steps in absence of any carbon source. Represented data are the $2^{-\Delta Ct}$ (mean ± SEM; n = 2). Solid lines represent IRMA genes; the dotted line is the *MET16* gene.

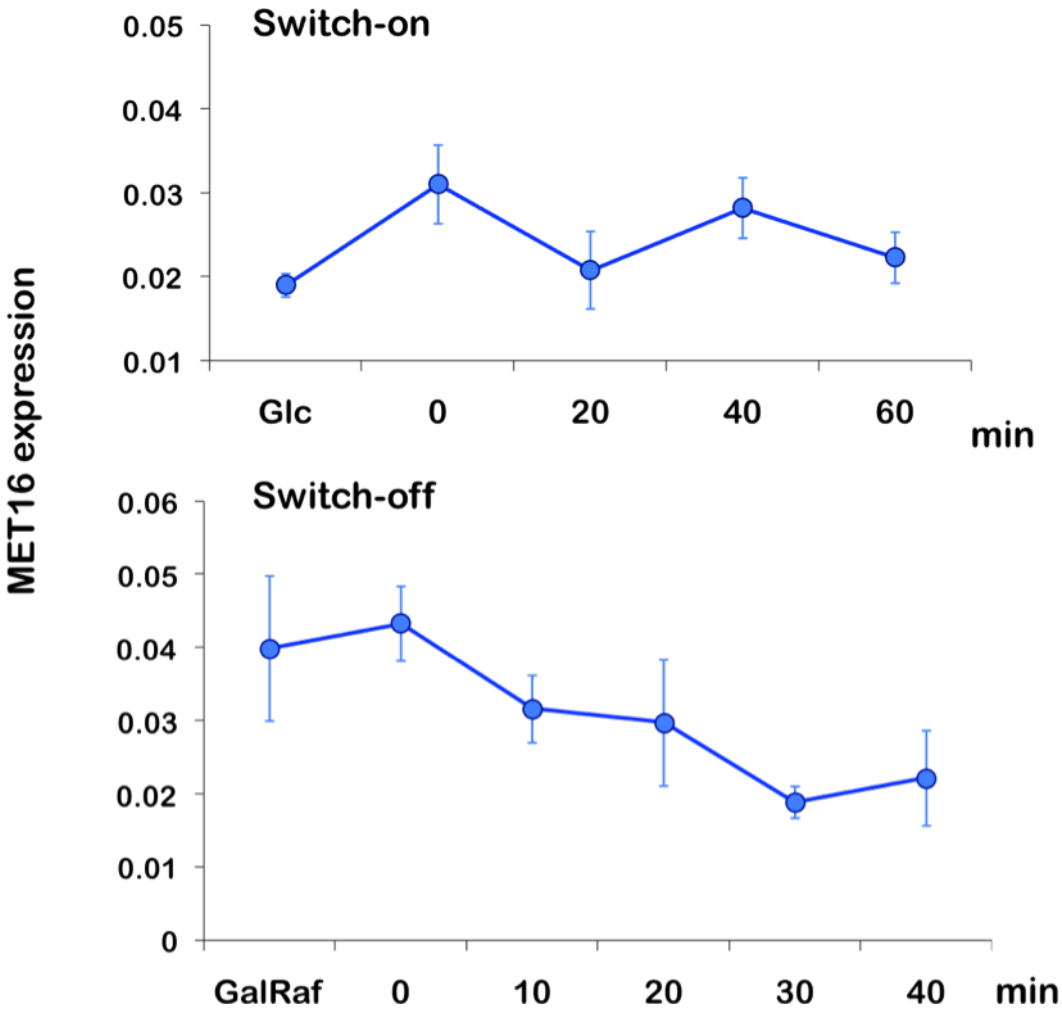

Figure 6.8 Transcription of *MET16* endogenous gene is not affected by sugar shift in switch-on and switch-off time-series. Expression levels of the *MET16* endogenous gene in IRMA cells during the glucose-to-galactose (switch-on) and galactose-to-glucose (switch-off) time-series, described in figure 6.6, are shown. The first point in the switch-on and switch-off time-series is measured in yeast cells which had been grown up to the steady state in glucose (Glc) and galactose-raffinose (Gal), respectively, before washing and shifting in a different medium. The 0 point is the one collected just after the medium shift had been performed. After the shift, samples were collected every 20 and every 10 min for the switch-on and the switch-off ,respectively (see main text). *MET16* does not show any transient increase after shifting cells from a medium to another one containing a different sugar, as conversely happens to *GAL4*, which is regulated by the *MET16* promoter. Represented data are the $2^{-\Delta Ct}$ (mean ± SEM; n = 5 for switch-on and n = 4 for switch-off).

6.5 Gene expression profiling of IRMA to study its static behaviour

In addition, we analysed the response of the network to genetic perturbations by overexpressing each of the five network genes under the control of the strong constitutive *GPD* promoter, in cells that were grown either in glucose, or in galactose. We then measured steady-state expression levels of IRMA genes by q-PCR. We thus obtained two datasets, one in glucose, and one in galactose, consisting of the response of the five network genes to each of the five perturbations. We will refer to these two experimental datasets as the "Glucose steady-state" and "Galactose steady-state" (Figure 6.9A-C and Figure 6.10A-C).

In vivo, upon overexpression of each of the five network genes, the other genes were either upregulated, or downregulated, with respect to their basal level (transformation with an empty vector) both in galactose and in glucose (Figure 6.9A-C and Figure 6.10A-C). Following overexpression of the three activators (*CBF1*, *GAL4* and *SWI5*), network genes' transcription increased in both growing conditions, reaching higher levels in galactose, when Gal80 repressor is inactive. In the *CBF1* overexpression experiment, *SWI5* responded with a significant increase, whereas *GAL4*, a direct target of *CBF1*, and the regulator of *SWI5* in the network, responded weakly. Gal4 protein is stable (Muratani et al., 2005; Nalley et al., 2006), and therefore even a small, or transient, increase in its mRNA level in galactose is able to induce the *GAL10* promoter, which in our network regulates *SWI5*.

Overexpression of *ASH1* induced smaller transcriptional variations, although a slight downregulation of the network genes is evident in galactose containing

medium, when the network is on. Remarkably, in the inducing medium, overexpression of *GAL80* resulted in a downregulation of the other genes, implying that the excess of Gal80 binds and represses the Gal4 protein, even in the presence of galactose.

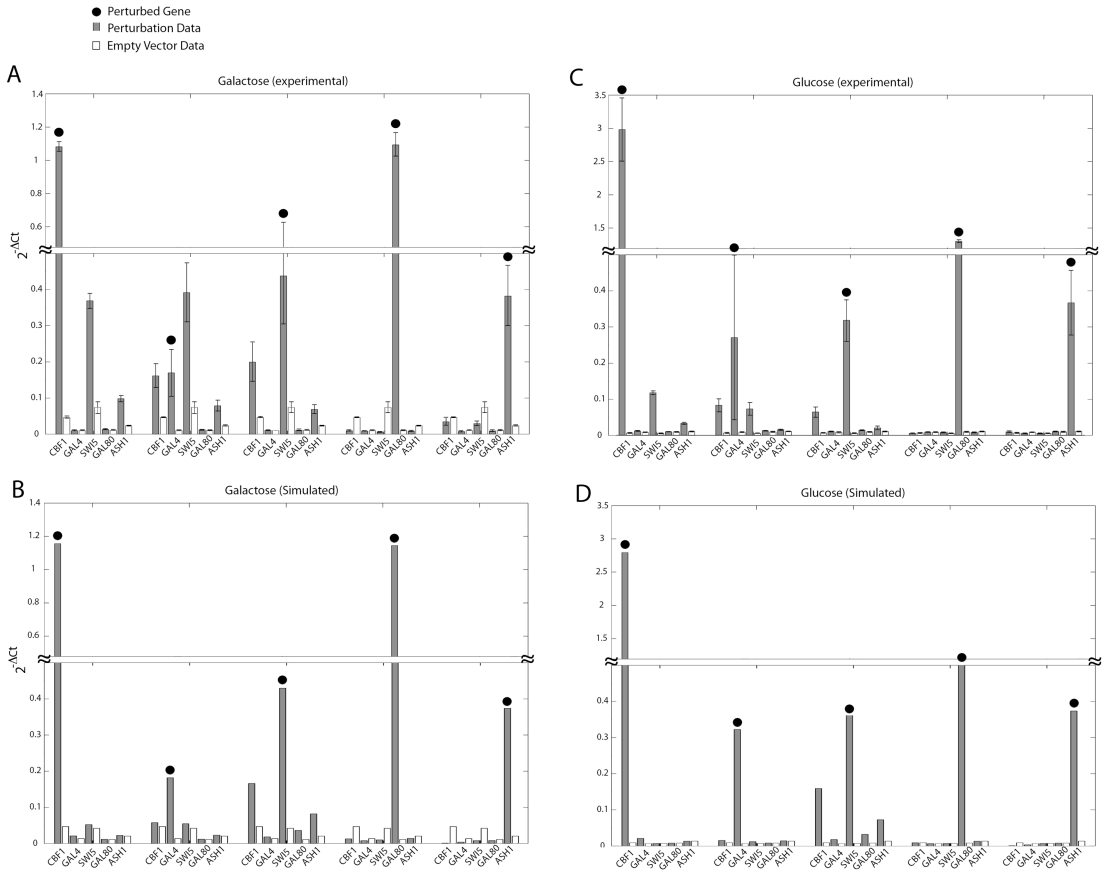

Figure 6.9. Experimental and simulated gene expression profiles show the static behaviour of IRMA in response to overexpression perturbation experiments. (A)-(C) *In vivo* expression levels of IRMA genes after overexpression of each gene (perturbed gene; indicated by the black dots on the bars) from the constitutive GPD promoter (grey bars) and after transformation of the empty vector (white bars). IRMA cells were transformed with each of the constructs containing one of the five genes or with the empty vector. At least three different colonies were grown in glucose (C) and in galactose-raffinose (A) up to the steady-state levels of gene expression. Quantitative PCR data are represented as $2^{-\Delta Ct}$ (mean ± SEM; n ≥ 3). **(B)-(D)** *In silico* expression levels of IRMA genes obtained by simulating the overexpression of each gene with the DE-based model.

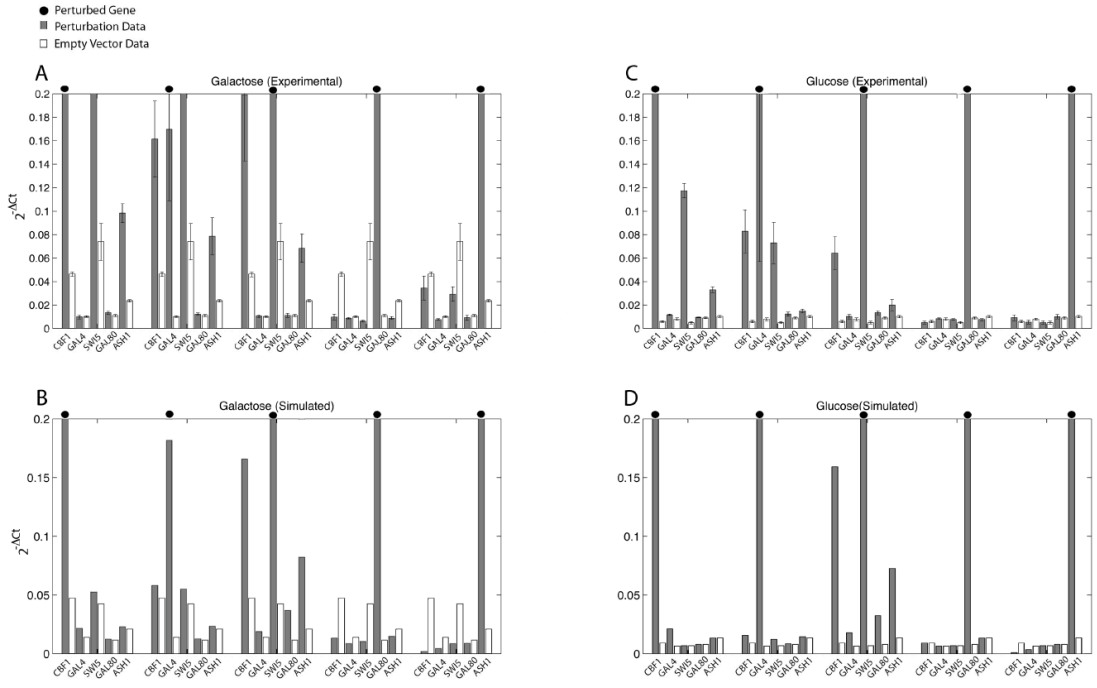

Figure 6.10. Experimental and simulated gene expression profiles show the static behaviour of IRMA in response to overexpression perturbation experiments (Magnification of figure 6.9). Panels (A-D) here correspond to the ones reported in figure 6.9. The y axis scale was lowered and set at 0.2 in order to better show the network genes behavior in the figure even when they are expressed at very low levels (e.g. in the presence of glucose).

Chapter 7 – Results.

IRMA as a Benchmark for Modelling

7.1 Mathematical model of the IRMA network

The most common strategy to model gene networks is the one based on nonlinear differential equations (DE) obtained from standard mass-balance kinetic laws (Alon, 2006; Szallasi et al., 2006). Therefore, the first step in modelling was to derive an appropriate deterministic model of IRMA where each gene of the network is described by an equation, which contains a synthesis term and a degradation term. For the sake of simplicity, we ignored protein levels (assuming proportionality between proteins and their corresponding mRNAs), and considered transcription and translation processes as a single synthesis step. Thus the variables in the mathematical model represent the mRNA abundance of each gene and the derived DE model consists of five equations describing the transcription rate of the five mRNAs - *CBF1*, *GAL4*, *SWI5*, *GAL80*, *ASH1* (Figure 7.1).

In the following subparagraphs, I will describe the assumptions, which lead to the formulation of model.

$$
\begin{aligned}
\frac{d[CBF1]}{dt} &= \alpha_1 + v_1 \left(\frac{[SWI5]^{h_1}(t-\tau)}{k_1^{h_1} + [SWI5]^{h_1}(t-\tau)} \right) \cdot \left(\frac{k_2^{h_2}}{k_2^{h_2} + [ASH1]^{h_2}} \right) - d_1[CBF1], \\
\frac{d[GAL4]}{dt} &= \alpha_2 + v_2 \left(\frac{[CBF1]^{h_3}}{k_3^{h_3} + [CBF1]^{h_3}} \right) - (d_2 - \Delta(\psi_1))[GAL4], \\
\frac{d[SWI5]}{dt} &= \alpha_3 + \hat{v_3} \left(\frac{[GAL4]^{h_4}}{\hat{k_4}^{h_4} + [GAL4]^{h_4}(1 + \frac{[GAL80]^4}{\hat{\gamma^4}})} \right) - d_3[SWI5], \\
\frac{d[GAL80]}{dt} &= \alpha_4 + v_4 \left(\frac{[SWI5]^{h_5}}{k_5^{h_5} + [SWI5]^{h_5}} \right) - (d_4 - \Delta(\psi_2))[GAL80], \\
\frac{d[ASH1]}{dt} &= \alpha_5 + v_5 \left(\frac{[SWI5]^{h_6}}{k_6^{h_6} + [SWI5]^{h_6}} \right) - d_5[ASH1],
\end{aligned}
$$

Figure 7.1 IRMA DE model. Each IRMA gene is represented by a differential equation (DE), which describes its synthesis and degradation rates as a function of its regulator inside the network. In the model, we assumed that protein translation is fast enough so that a quasi – steady state approximation can be made and proteins levels are considered proportional to their corresponding mRNA concentrations. So, the equations in the model describe the transcriptional rate (d[mRNA]/dt) of the IRMA genes; where d_i, i = 1, . . . , 5 are the degradation rates, k_j, j = 1, . . . , 6 are the Michaelis-Menten constants, α_i are the basal transcription activities of each promoter, υ_i represent the maximal transcription rates. The hat symbol (^) indicates that the value of the relative parameter depends on the medium in which the yeast is grown (glucose or galactose). Moreover, as explained below, γ is an affinity constant, τ models the delay in the *HO* promoter activation, and the terms $\Delta(\cdot)$ model the transient increase lasting 10 minutes (the time required to perform the washing steps during medium shift) in mRNA stability due to the cell manipulation, as described in the previous chapter, with magnitude ψ_1 and ψ_2.

7.1.1 Modelling the Binding of Transcription Factors to Promoters

Transcription of a gene results from the binding of a TF (X) to specific DNA sites (D) in the promoter region of the gene and can be thought as a two steps chemical reaction:

$$X + D \underset{k_{off}}{\overset{k_{on}}{\rightleftarrows}} [XD] \overset{k_{mRNA}}{\rightleftarrows} mRNA + D \quad (1)$$

Considering mRNA formation as the rate-limiting step of transcription ($k_{mRNA} << k_{on}$), we can focus our attention on the first step, which is transcriptional complex formation. If the TF is a repressor, transcription occurs only when the repressor is not bound to the promoter, that is, when D is free. Conversely, if the TF is an activator, transcription occurs when it binds to D forming the [XD] complex. The DNA site can thus be either free, D, or bound, [XD], resulting in a **conservation equation**:

$$D + [XD] = D_T \quad (2)$$

where D_T is the total concentration of the site.

Formation of the [XD] complex happens because the transcription factor, X, and its target DNA, D, diffuse in the cell and occasionally collide. This process can be described by **mass-action kinetics**: X and D collide and bind each other at a rate k_{on}. The rate of complex formation is thus proportional to the collision rate, given by the product of concentrations of X and D:

$$\text{rate of complex formation} = k_{on} \, X \, D$$

The complex [XD] dissociates at a rate k_{off}. The rate of change of [XD] based collision and dissociation processes is described by:

$$d[XD]/dt = k_{on} XD - k_{off} [XD] \quad (3)$$

At the **steady state**, when the concentrations of molecular species involved in the reaction do not change in time, the above equation is equal to zero. Solving it at the steady state, the balance between the association of X and D and the dissociation of [XD] leads to the chemical equilibrium equation:

$$K_d [XD] = XD \quad (4)$$

where K_d is the dissociation constant ($K_d = k_{off} / k_{on}$), which as the units of a concentration. The higher is the dissociation constant, the higher is the rate of dissociation of the complex, that is, the weaker is the binding of the TF to DNA.

Considering that [XD] transcriptional complexes dissociate within less than 1 sec, we can average over times much longer than 1 sec and consider the ratio of D over D_T as the probability that the D site is free averaged over many binding and unbinding events. We can derive the probability that D is free by substituting (D_T – D) with [XD] and solving Equation 3, which yields:

$$\frac{D}{D_T} = \frac{1}{1 + X / K_d} \quad (5)$$

If we consider that, in the case of a repressor, the promoter is transcribed when the DNA binding sites are free, we can describe the rate of transcription (the concentration of the mRNA over time) as the product between the probability that the D site is free with a constant υ:

$$\text{Promoter activity} = \frac{\upsilon}{1 + X / K_d} \quad (6)$$

Conversely, an activator protein increases the rate of transcription when it binds to its DNA site in the promoter. Using the same reasoning as above, we can derive the probability that X is bound to D substituting (D_T – [XD]) with D and solving Equation 3:

$$\frac{[XD]}{D_T} = \frac{X}{X + K_d} \tag{7}$$

In the case of an activator, the rate of transcription is proportional to the probability that it is bound to D and thus we have:

$$\text{Promoter activity} = \frac{\upsilon X}{X + K_d} \tag{8}$$

Even if we derived transcription rates assuming that the binding of the TF to the promoter is faster than the mRNA transcription (which corresponds to say that $k_{mRNA} << k_{on}$), this is not always true in biology. Thus, considering also the second step of the reaction in (1), in which from the [XD] complex we have mRNA transcription, and considering it as an enzymatic reaction, the rate of complex dissociation also depends on k_{mRNA} and Equation 3 becomes:

$$d[XD]/dt = k_{on} XD - k_{off} [XD] - k_{mRNA} [XD] \tag{9}$$

Reasoning as above, if we consider that the concentration of [XD] changes much more slowly than the one of X or D (quasi – steady state assumption in Equation 4) and that the total concentration of DNA binding sites do not change over time (in Equation 2), we obtain the same Equations as in (4), (5) and (7), in which now the dissociation constant is equal to the Michaelis and Menten constant ($K_M = k_{off} + k_{mRNA} / k_{on}$). Substituting the concentration of free D or of [XD] complex in the following:

$$d[mRNA] = k_{mRNA}\,[XD] \qquad (10)$$

we obtain the same as Equation (6) and (8), respectively (where $\upsilon = k_{mRNA}\, D_T$).

However, in order to obtain a more realistic description of promoter binding in our model we needed to take into account that many TFs inside the cell are composed of several repeated protein subunits, and they often has got more than one binding site inside the target promoter. These events can lead to cooperativity in the binding reaction of the TF to the DNA. In order to take into account the cooperative binding we added at Equation 5 and 7 the Hill coefficient, which provides a way to quantify this effect. We thus obtained a classical **Hill function**, a phenomenological equation that describes the fraction of a macromolecule saturated by ligand as a function of the ligand concentration. For the same reasons we explained above, the Hill function can be used to model the transcriptional activation or repression. For an activator X the equation is the following:

$$H^{+}(X\,;K) = \frac{X^{h}}{X^{h} + K^{h}} \qquad (10)$$

where *h* is the hill coefficient (a pure number that refers to the cooperativity of the activation binding reaction) and *K* is the Michaelis and Menten constant. In the case of inhibition the function becomes:

$$H^{-}(X\,;K) = \frac{1}{1 + (X/K)^{h}} \qquad (11)$$

Indeed, in our model, in order to describe mRNA synthesis we assumed that each promoter has got a basal level of transcription and can be activated or repressed by its regulators following a Hill kinetic (Kaern et al., 2003). The activation-repression rates are assumed to be proportional to the Hill function via some constants

υ_I, which represent the maximal transcription rates; while mRNAs degradation rates are supposed to be well described by using first order degradation kinetics.

7.1.2 Modelling Galactose Regulation

In deriving the model, particular care was taken in modelling the galactose regulation on the *GAL10* promoter, in order to capture its main features, but without increasing model complexity. In the yeast cells, regulatory genes of the galactose pathway are *GAL4*, *GAL80* and *GAL3*. Gal4 is the transcription factor, which binds to the four UAS_{GAL} in the *GAL10* promoter and consequently triggers its transcription in the presence of galactose. Gal80 is a repressor, which interacts with Gal4 activation domain thus blocking the recruitment of the transcriptional machinery in the absence of galactose. As a matter of fact, Gal4 can bind to its target promoter only when Gal80 is inhibited. The galactose sensor is Gal3 protein, which determines Gal80 dissociation from Gal4. In summary, when galactose is present, Gal3 binds Gal80 and this interaction triggers the relief of Gal80 inhibition and Gal4 activates *GAL10* promoter (Figure 7.2 A and B) (Traven et al., 2006).

In literature different models of the galactose pathway in yeast have been proposed but they are extremely detailed and they would have increased the number of equations to be used in our model (Acar et al., 2005; Bennett et al., 2008; Verma et al., 2004). In order to keep our model as simplest as possible, we assumed that galactose directly binds to Gal80 and influences its affinity for the *GAL10* promoter (we considered the Gal80 inhibition directly on *GAL10* and not on Gal4), that is we did not considered Gal3-Gal80 and Gal80-Gal4 complexes formation (Figure 7.2 C and D). Thus, to describe the interactions between galactose and the GAL genes of the

network, we used a phenomenological rate law for the activation of the promoter of *SWI5* (Figure 7.1). We assumed that the activation of *SWI5* by Gal4 is also inhibited by a Michaelis-Menten like term proportional to the amount of *GAL80* mRNA and inversely proportional to an affinity constant $\widehat{\gamma}$, which depends on the presence or the absence of galactose and was experimentally measured (Figure 7.4 and Table 6).

Furthermore, we modelled the effect of the transient deprivation of carbon source, which lead to an increase in *GAL4* and *GAL80* mRNA levels during medium shift in the switch-on time-series, as an additional transient perturbation to the degradation rates of *GAL4* and *GAL80* mRNAs lasting 10 min (the time estimated to perform the washing steps). This increase in mRNA stability is represented by the terms $\Delta(\psi_1)$ and $\Delta(\psi_2)$ in the degradation of *GAL4* and *GAL80*, respectively (Figure 7.1).

7.1.3 Modelling *HO* promoter regulation

In our network, the expression of *CBF1* driven by the *HO* promoter is both activated by Swi5 and repressed by Ash1. *HO* promoter transcription is triggered by an ordered recruitment of transcription and chromatin remodelling factors. The first factor that arrives to *HO* is Swi5, which activates the sequential chain of events that result in promoter transcription after a certain lag time (Bhoite et al., 2001; Cosma, 2002; Cosma et al., 1999; Mitra et al., 2006). This delay is also evident in our network in the activation of *CBF1* during switch-on and switch-off experiment (Figure 6.6). We thus included an explicit delay in the activation of *CBF1* by Swi5 (represented by τ in Figure 7.1).

Furthermore, in order, to model the multiple regulations by Swi5 and Ash1, there are two alternative strategies (Alon, 2006): modelling this effect as the non-competitive interaction between Ash1 and Swi5 (i.e. as a logical AND gate) or considering the multiple regulations as the combined action of the activation and repression terms (i.e. as an OR gate). From literature we know that Swi5 and Ash1 have got distinct binding sites on the *HO* promoter so, they do not compete for the binding neither interacts with each other (Maxon and Herskowitz, 2001). It has been reported that Ash1 inhibits *HO* transcription preventing the recruitment of the Swi/Snf complex by Swi5 probably through it interaction with the large Sin3-Rpd3 histone deacetylase complex (Carrozza et al., 2005; Mitra et al., 2006). Taking into account these evidences, we modelled Swi5 and Ash1 regulation on *CBF1* expression as a non-competitive interaction. This is a type of inhibition that reduces the maximum rate of the chemical reaction without changing the apparent binding affinity of the promoter.

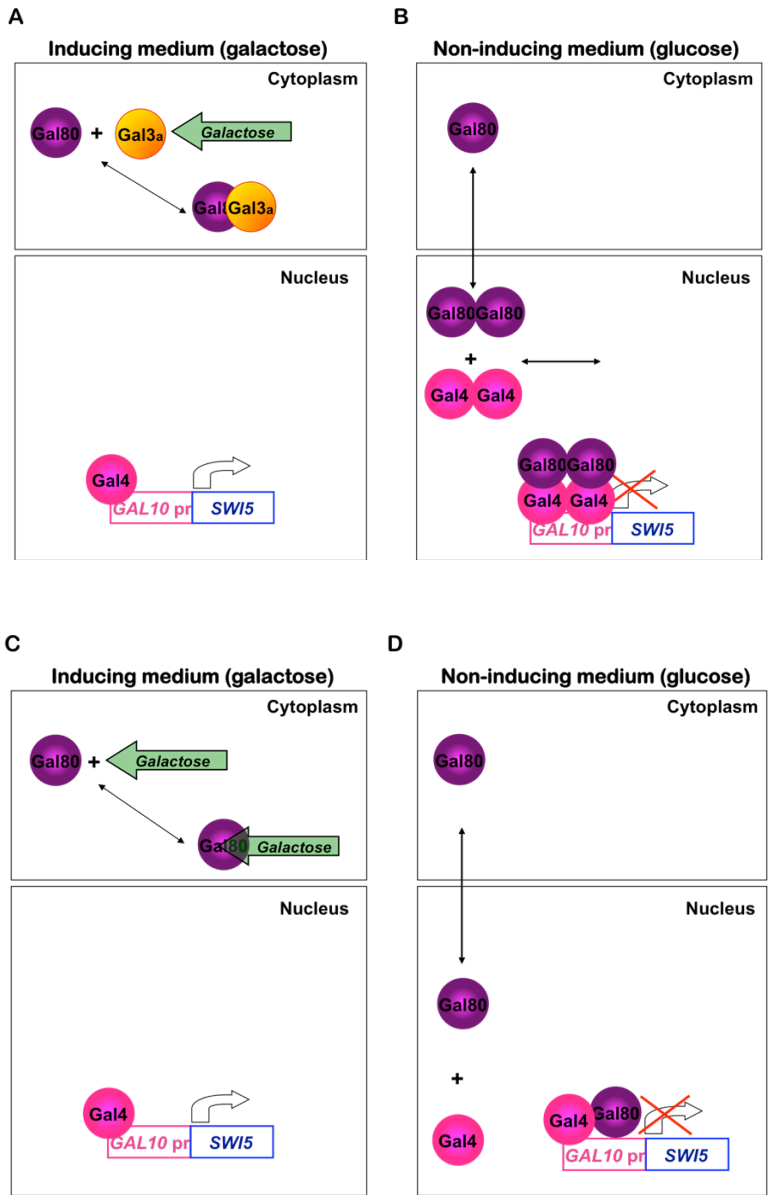

Figure 7.2. Galactose regulatory pathway. (A)-(B) Schematic model of galactose pathway is shown. (A) In the presence of galactose, Gal80 repressor is inhibited by Gal3 binding and Gal4 triggers transcrption of *GAL* genes. How exactly Gal3–Gal80 complex formation relieves Gal80 inhibition of Gal4 is not yet known. Gal3 interacts with Gal80 in the cytoplasm and then elicit a conformational change in the Gal80–Gal4 complex. Gal80 dissociates from Gal4 on binding to Gal3 and 'shuttles' between the cytoplasm and the nucleus. (B) Under non-inducing conditions, Gal80 binds to Gal4 dimers and blocks its interaction with the transcriptional machinery. **(C)-(D)** Simplified model of galactose pathway as assumed in IRMA mathematical model.

7.2 Estimating Model Parameters

In order to estimate the unknown parameters, we experimentally measured promoters' strength of the promoter, which we used in IRMA network, namely *GAL10*, *MET16*, *ASH1* and *HO*. We constructed different strains in which an inducible promoter replaced the endogenous one of each TF gene. In this way, we were able to induce the expression of the TF at different levels, and then we measured, by q-PCR, the transcription of the corresponding promoter gene, at steady state, for a total of 165 data points (Figure 7.3 and 7.4).

To characterize *MET16* promoter, we constructed a copper-inducible strain carrying *CBF1* under the control of *CUP1* promoter and we measured *CBF1* and *MET16* using different concentration of copper. Since *CBF1* and *MET16* showed small variations, even in a wide range of copper concentrations (black dots in the upper graph of Figure 7.3), in order to extend the dataset we constructed four more strains in which *CBF1* was under the control of constitutive promoters of different strength (the *CYC1*, *ADH1*, *TEF* and *GPD* promoters).

Similarly, in order to evaluate the strength of the *ASH1* promoter we constructed an inducible strain carrying *SWI5* under the control of the *MET25* promoter and we measured *SWI5* and *ASH1* using different concentration of methionine (Figure 7.3 lower panel).

For the *HO* promoter, since both Swi5 and Ash1 regulate it, we measured the level of expression of *ASH1*, *SWI5* and *HO* in:

1) *MET25-SWI5* inducible strain after inducing Swi5 by different methionine concentrations, both in the presence of an endogenous Ash1 and in colonies, which over express Ash1 at different levels;

2) *MET25-ASH1* inducible strain after inducing Ash1 by different methionine concentrations, both in the presence of an endogenous Swi5 and in colonies, which over express Swi5 at different levels.

In order to characterize *GAL10* promoter, which is regulated by both Gal4 and Gal80, we measured *GAL10*, *GAL4* and *GAL80* in an inducible strain carrying *GAL4* under the control of the *MET25* promoter, both in the presence of an endogenous Gal80 and in colonies, which over express Gal80 at different levels. The experiment was performed both in glucose and in galactose containing medium (Figure 7.4).

For both *HO* and *GAL10*, we plotted surfaces in order to show their expression as a function of the two regulators (Figure 7.4).

We thus obtained four datasets, one for each promoter, and we fitted to these data the equation at the steady state (Figure 7.1) of the gene whose expression in our network is driven by the promoter itself (Figure 7.3 and 7.4). In this way, we estimated 16 (out of the 33) parameters, which consist of the Michaelis-Menten and the relative Hill coefficients.

The remaining 17 unknown parameters, which could not be computed from promoters' data, were estimated from the switch-on time-series by using a novel stochastic optimization algorithm (Cantone et al., 2009). In order to simulate the switch-on data, we chose as initial conditions the steady state equilibrium of the model in glucose, thus recapitulating the experimental conditions. Simulated data fitted semi-quantitatively *in vivo* data, despite the simplifying assumptions, being on average within the experimental standard errors (Figure 6.6, left panel).

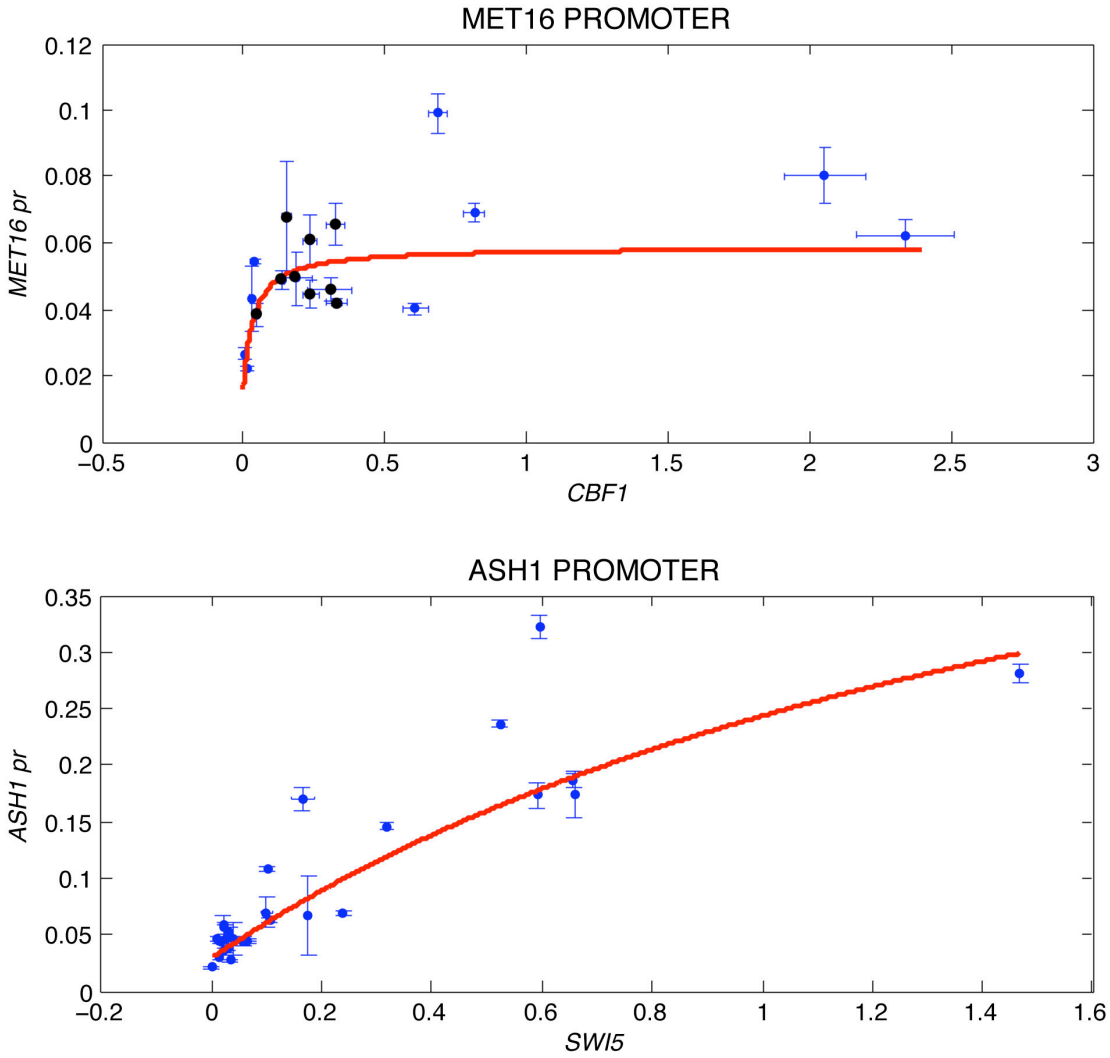

Figure 7.3. Fitting of *MET16* and *ASH1* promoter strength data to Hill function. Upper panel: *MET16* expression levels (y axis) after induction of *CBF1* expression (x axis). Lower panel: *ASH1* expression levels (y axis) after induction of *SWI5* expression (x axis). Dots represent experimental data ($2^{-\Delta Ct}$ ± SEM; n = 2). Error bars represent technical errors. Red lines represent the fitting of the Hill function for the target gene.

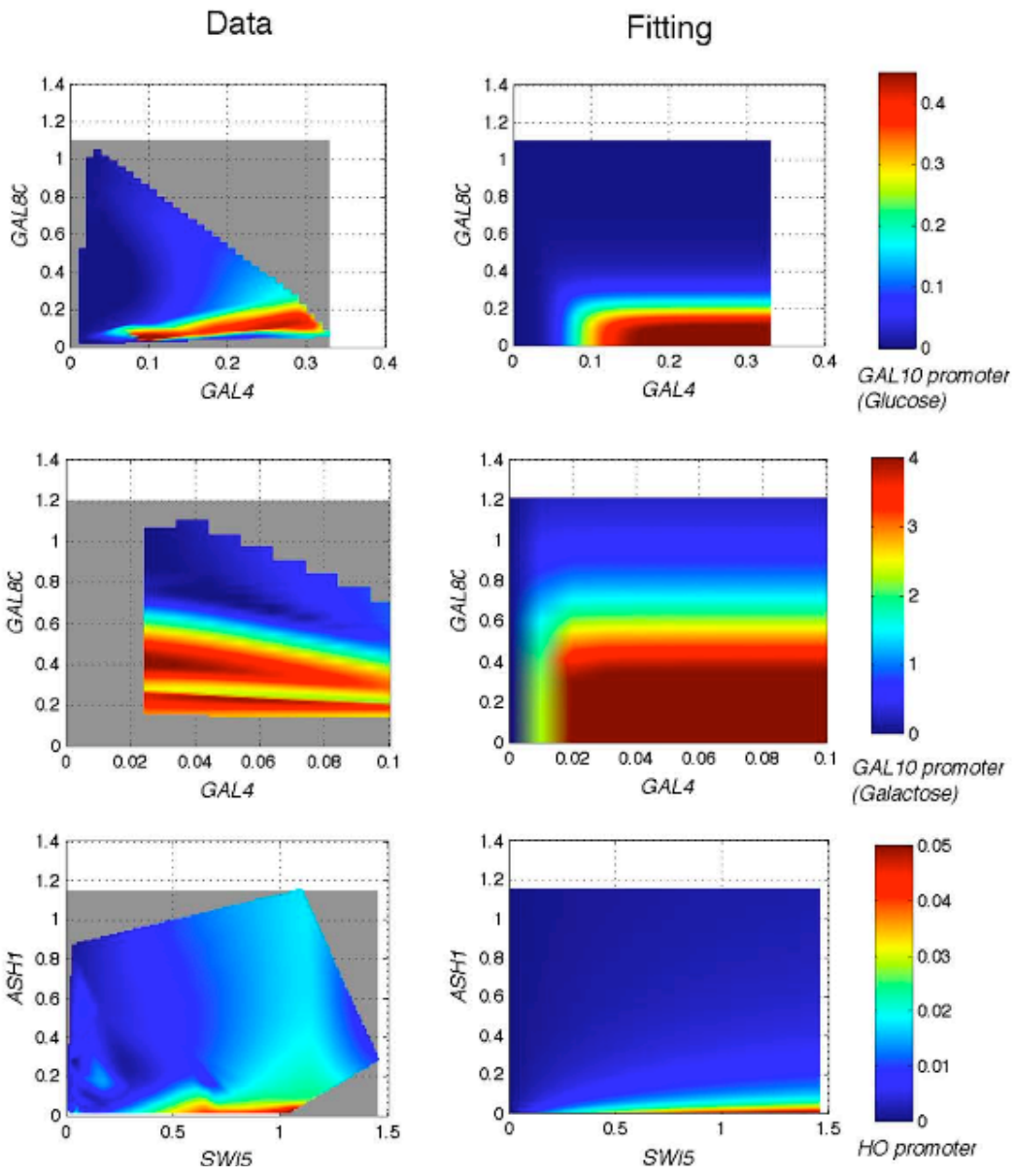

Figure 7.4. Fitting of *GAL10* and *HO* promoter strength data to Hill function. *GAL10* expression levels (z axis represented by colour) after induction of *GAL4* expression (x axis) and *GAL80* (y axis) at different levels in the presence of glucose (upper panel) and galactose (middle panel). Lower panel shows *HO* expression levels (z axis represented by colour) after induction of *SWI5* expression (x axis) and *ASH1* (y axis) at different levels. Left panels show experimental data (expression levels) as $2^{-\Delta Ct}$ (grey area represent regions in which data are not present). Right panels show fitting results.

7.3 Model Predictive Power

In order to test the model predictive performance we used both the switch-off time-series and the Glucose steady-state and Galactose steady-state datasets.

The model was able to predict, semi-quantitatively, the behaviour of the network during the switch-off experiment (Figure 6.6, right panel). Specifically, the model correctly predicted the delay in *CBF1* silencing, in contrast to the fast switch-off dynamics of SWI5. Furthermore, the small variations of *GAL4* and *GAL80*, which are due to the low expression level of these two genes in glucose containing medium, were captured by the model. Differences in the starting amount of *CBF1*, *SWI5* and *ASH1* during the switch-off may be due to the unmodelled effect of protein accumulation of network genes. Indeed, the switch-off experiment is performed after having grown cells overnight in galactose, prior to galactose removal.

In order to further validate the predictive power of the model, we performed the previously described Glucose steady-state and Galactose steady-state overexpression experiments *in silico*, by simulating an overexpression of each of the five genes using the model. In Figure 6.9 and 6.10, we compared *in vivo* and *in silico* experiments. There is a semi-quantitative agreement, both in the Galactose and Glucose steady-state experiments. The model, despite some discrepancies in the predicted transcription levels, correctly captured the overall trend among each perturbed set of genes. We observed that *SWI5* predicted expression levels are smaller than their experimental counterparts, and this effect propagates in turn to its targets.

To explain this behaviour, we noticed that the Gal4 protein is stable (Muratani et al., 2005; Nalley et al., 2006), and therefore, even a small, or transient, increase in

its mRNA level is able to induce the *GAL10* promoter, regulating *SWI5* in our network. Since we did not model explicitly protein dynamics, a small increase in *GAL4* mRNA, cannot fully activate *GAL10* in the model, and neither cause the consequent large increase in *SWI5* mRNA seen *in vivo*.

The model was able to recapitulate some of the expected biological features, such as the higher expression levels in the galactose containing medium, and the Gal80 repression activity when *GAL80* is over-expressed in the presence of galactose.

The model can also be used to link the observed dynamics to the topology of the network; we show by simulation that both the positive feedback loop (Swi5-Cbf1-Gal4) and the delay in the activation of the *CBF1* promoter are essential for the non-monotonic behaviour characterised by damped oscillations in the levels of *SWI5* and *CBF1*. Removing any of the interactions in the positive loop, or the delay, makes the oscillations smaller (Figure 7.5), or totally disappear (Figure 7.6).

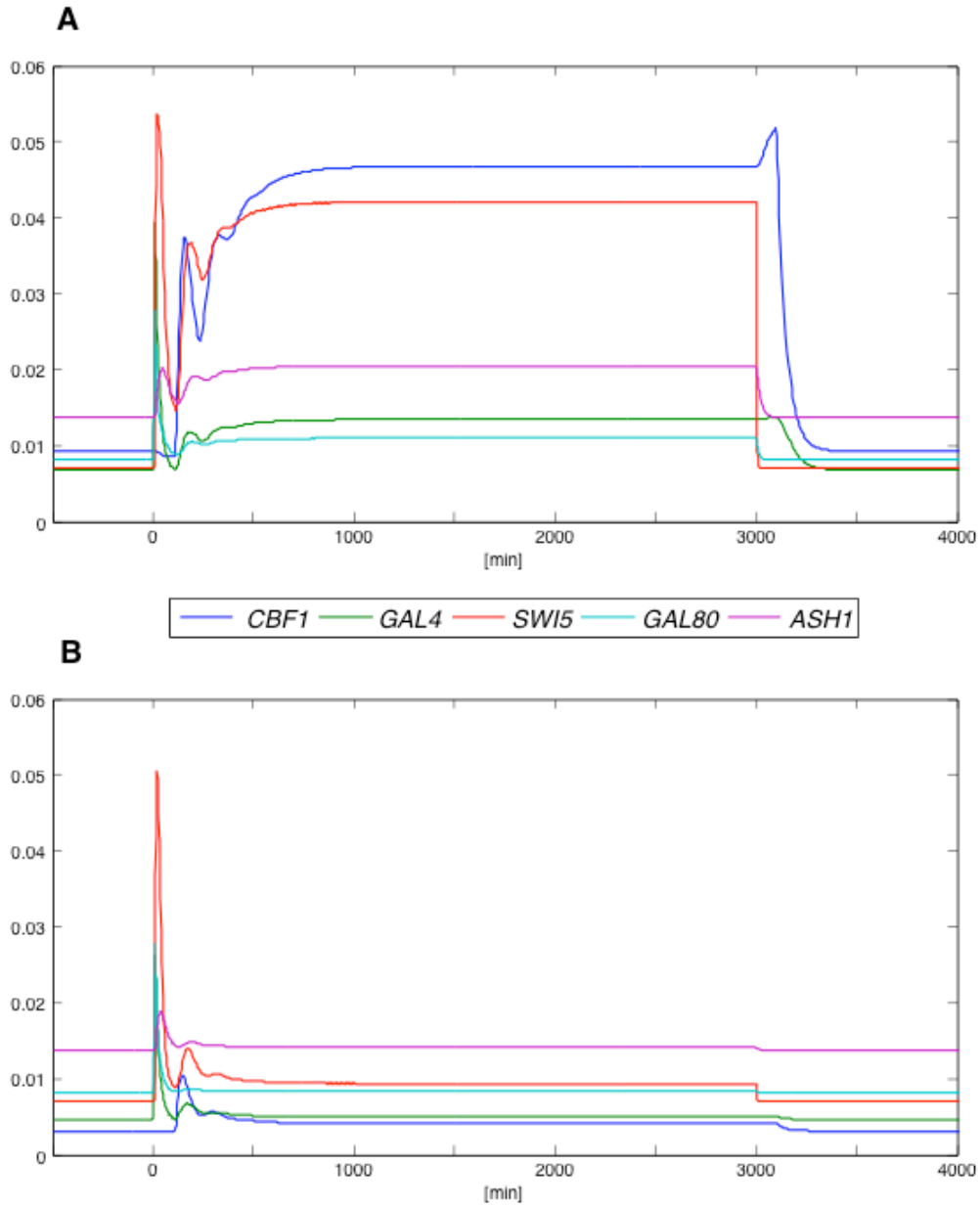

Figure 7.5. Simulations of the switch-on and switch-off time-series. (A) Simulations of network genes expression obtained by the DE model with parameters as in Table 6. **(B)** Simulations of network genes obtained by the model without the positive feedback loop (no activation of *CBF1* by Swi5 by setting k_1=3 [a.u.] in Table 6). Gene expression is in absolute values ($2^{-\Delta Ct}$).

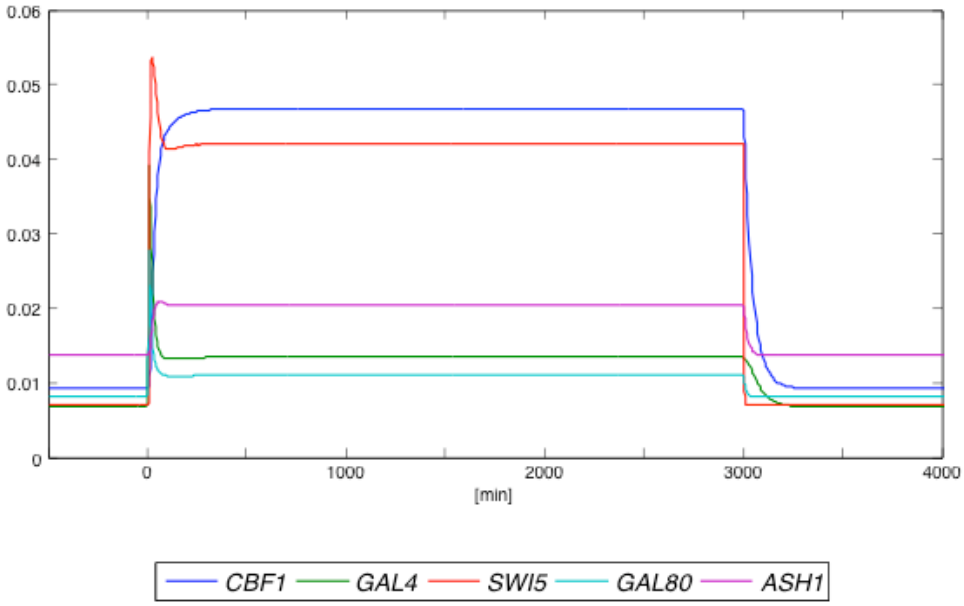

Figure 7.6. Influence of *CBF1* activation delay on the dynamics of the network. Simulations of network genes expression in the switch-on (starting at the 0 time) and switch-off time-series (after 3000 minutes) without delay in *CBF1* activation are shown. Gene expression is in absolute values ($2^{-\Delta Ct}$).

Chapter 8 – Results.

IRMA as a Benchmark for Reverse-engineering

8.1 Reconstructing the network: a reverse engineering approach

The IRMA synthetic network can be used to assess the ability of experimental and computational approaches to infer regulatory interactions from gene expression data. We used the switch-on and switch-off time-series, and the steady-state perturbations in galactose and glucose, in conjunction with four published algorithms as representatives of reverse-engineering approaches, BANJO (Bayesian network) (Yu et al., 2004), NIR and TSNI (Ordinary differential equations) (Della Gatta et al., 2008; Gardner et al., 2003), and ARACNE (Information theoretic) (Basso et al., 2005). ARACNE was not applied to the time-series data, since it is not appropriate in this case.

Figure 8.1, Figure 8.2 and Figure 8.3 show the results of the ODE, Bayesian and Information-theoretic reverse engineering approaches, respectively.

Reverse engineering performance was quantified in terms of percentage of correctly predicted interactions out of the total number of predicted interactions (i.e. Positive Predictive Value – PPV), and, in terms of percentage of all the true interactions that have been correctly identified by the algorithm (i.e. Sensitivity - Se) (Bansal et al., 2007).

In order to test the significance of the algorithms, we computed the "random" performance, which refers to the expected performance of an algorithm that randomly assigns edges between pair of genes. For example, for a fully connected network, the random algorithm would have a 100% accuracy (PPV=1) for all the levels of sensitivity (as any pair of genes is connected in the real network). In our network, the expected PPV for a random guess of directed interactions among genes is PPV=0.40 (40%), so any value higher than 0.4 will be significant. In the case of undirected interactions, the random PPV=0.70 (70%).

8.1.1 Reverse-engineering Time-series Data

On time-series data, the best performance both in terms of PPV and of Se was achieved by the ODE approach (TSNI) on the switch-on data with a PPV=0.80 and a Se=0.50 (Figure 8.1A). ODE performed better than random (PPV=0.60, Se=0.38) also on the switch-off data, in Figure 8.1B, albeit with a lower precision.

Dynamic Bayesian Networks (BANJO) performed better than random (PPV=0.60, Se=0.38) only on the switch-off experiment, with the same performance as TSNI for this data set (Figure 8.2B). Bayesian Networks failed to perform better than random on the switch-on data (Figure 8.2A) probably because of the lower number of points (15) as compared to the switch-off time-series (21 points).

By comparing the inferred networks from BANJO and TSNI in the switch-on and switch-off experiments, it is clear that both methods are extracting similar information, albeit with less precision in the case of BANJO. If we consider only the interactions inferred by both methods on the same dataset (compare Figure 8.1A with Figure 8.2A, and Figure 8.1B and 8.2B), we obtained only two interactions, both

correct (PPV=1). This result hints to the possibility that meta-algorithms, combining results from multiple reverse-engineering algorithms, may improve reverse-engineering performance.

We could not apply the ARACNE algorithm to our time-series datasets because application of information-theoretic approaches on time-series data requires that each data point is statistically independent from the previous one, and in our case this assumption cannot be made.

8.1.2 Reverse-engineering Steady-state Data

When reverse-engineering from steady-state data, NIR was able to recover the network with a PPV=0.60 and a Se=0.38 in the galactose dataset (Figure 8.1C), but it did not perform better than random (PPV=0.40 and Se=0.25) in the glucose dataset (Figure 8.1D). NIR and TSNI correctly recovered the same three regulatory interactions of Swi5, in galactose steady-state and switch-on time-series, respectively. BANJO was better than random both in the galactose dataset (PPV=0.60, Se=0.38) and in the glucose one (PPV=0.50, Se=0.38); albeit with a lower precision in the last one (Figure 8.2C and 8.2D). BANJO extracted very similar information from both steady-state and switch-off time-series, inferring on all of them the same two interactions, among the three correct (Figure 8.2B-C-D). These results imply that both dynamic time-series data, and static steady-state data, are informative for reverse-engineering.

By considering only interactions inferred by both methods on the same dataset, in the case of galactose, we selected only one interaction, albeit correctly (PPV=1); whereas in the glucose experiment, no interactions were in common. This is a further hint that combing results from multiple reverse-engineering algorithms may be

beneficial. ARACNE did not perform better than random, which in the case of undirected graph is very high (PPV=0.70) (Figure 8.3). ARACNE was designed for inference of large networks (of the order of thousands of genes), and it is not directly comparable to the other two approaches (Basso et al., 2005).

From these data, we can conclude that ODE-based algorithms and Bayesian Networks (BANJO) performed similarly for the steady-state data, but ODE-based algorithms require more information, that is, the genes that have been directly perturbed in the experiment (Bansal et al., 2007). Information-theoretic approaches should not be applied to small networks, due to their inability of inferring the direction of regulation. However, they are superior to other methods in the case of large networks due to their ability to require a minimal amount of data to infer gene-gene undirected interactions (Faith et al., 2007).

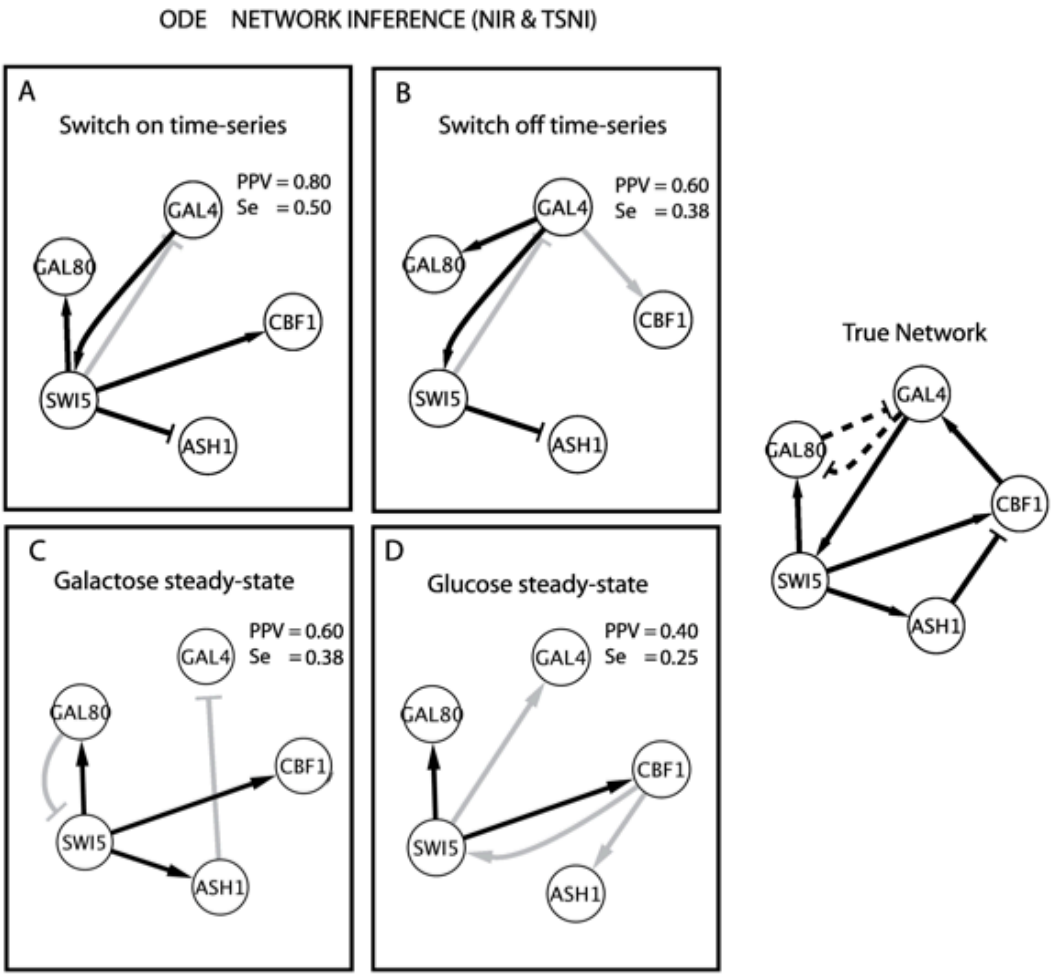

Figure 8.1. Reverse-engineering the IRMA gene network from steady-state and time-series experimental data using the ODE-based approach. The true network shows the regulatory interactions among genes in IRMA. Dashed lines represent protein-protein interactions. Directed edges with an arrow-end represent activation, whereas dash-end represents inhibition; **(A) and (B)** Inferred network using the TSNI reverse-engineering algorithm and the switch-on and switch-off time-series experiments. Solid grey lines represent inferred interactions that are not present in the real network, or that have the wrong direction (False Positives- FP). PPV (Positive Predictive Value = TP/(TP+FP)) and Se (Sensitivity = TP/(TP+FN)) values show the performance of the algorithm for an unsigned directed graph. TP=True Positive, FN=False negative. The random PPV for the unsigned directed graph is equal to 0.40. **(C) and (D)** Inferred network using the NIR reverse-engineering algorithm and the steady-state experimental data from network genes overexpression in cells grown in galactose or glucose medium, respectively.

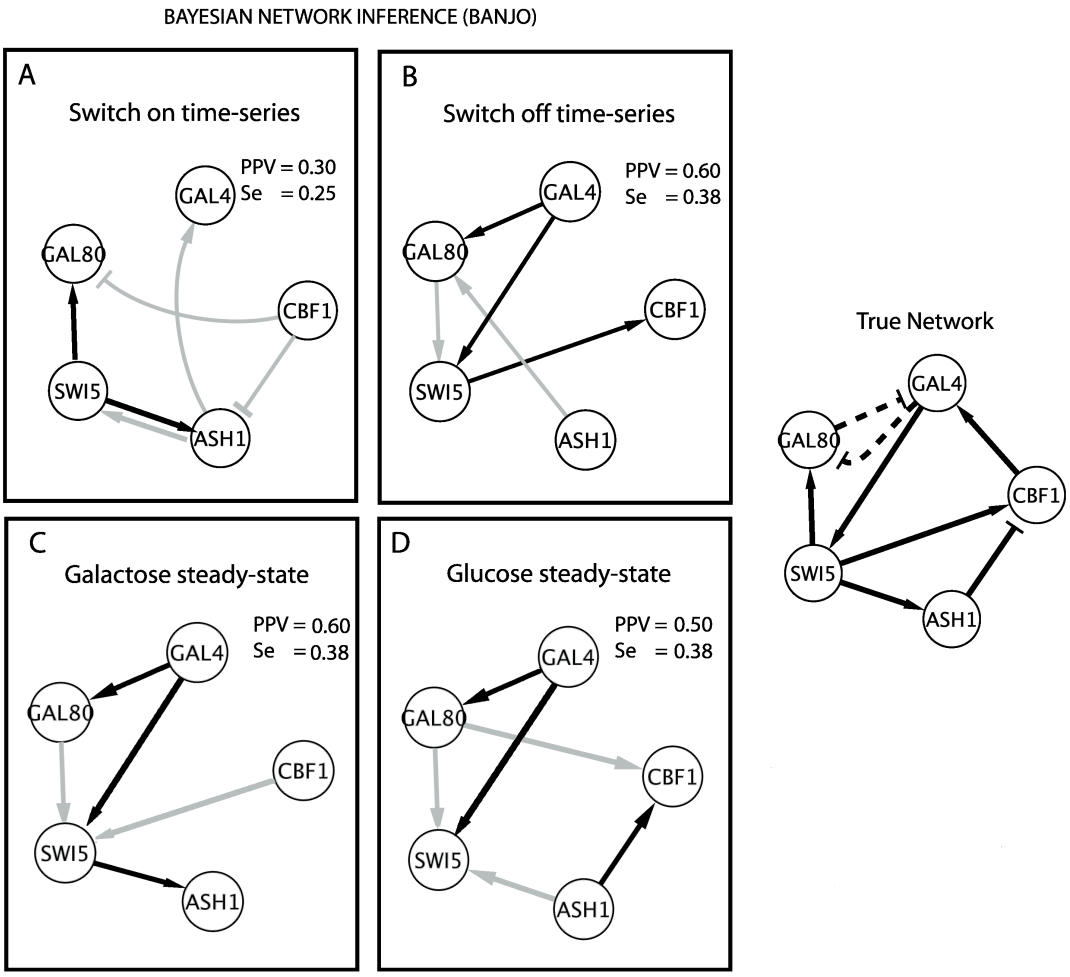

Figure 8.2. Reverse-engineering the IRMA gene network from steady-state and time-series experimental data using the Bayesian Network approach. (A) and (B) Inferred network using the BANJO algorithm and the switch-on and switch-off time-series experiments. Solid grey lines represent inferred interactions that are not present in the real network, or that have the wrong direction (False Positives- FP). PPV (Positive Predictive Value = TP/(TP+FP)) and Se (Sensitivity = TP/(TP+FN)) values show the performance of the algorithm for an unsigned directed graph. TP=True Positive, FN=False negative. The random PPV for the unsigned directed graph is equal to 0.40. **(C) and (D)** Inferred network using the BANJO algorithm and the steady-state experimental data from network genes overexpression in cells grown in galactose or glucose medium, respectively.

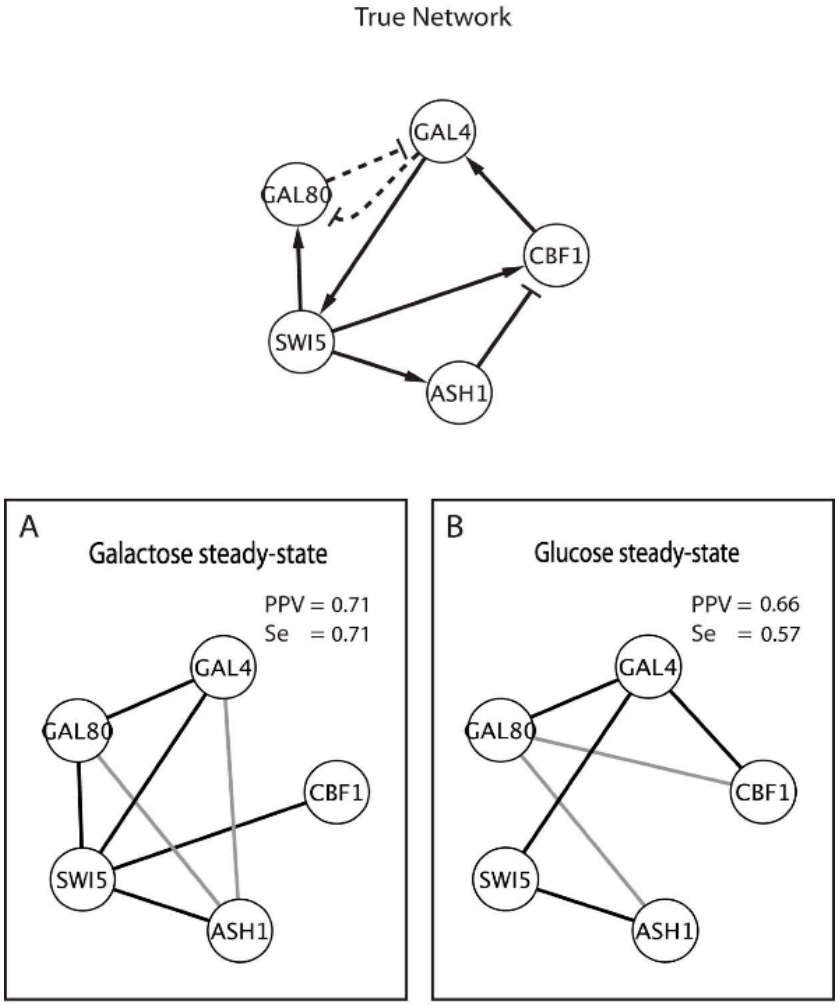

Figure 8.3. Reverse-engineering the IRMA gene network from steady-state experimental data using the Information Theoretic approach. (A) and **(B)** Inferred network using the ARACNE algorithm and the steady-state data from network genes overexpression in cells grown in galactose or glucose medium, respectively. Solid grey lines represent inferred interactions that are not present in the real network, or that have the wrong direction (False Positives- FP). PPV (Positive Predictive Value = TP/(TP+FP)) and Se (Sensitivity = TP/(TP+FN)) values show the performance of the algorithm for an unsigned directed graph. TP=True Positive, FN=False negative. The random PPV for the undirected graph is equal to 0.70.

8.2 Reverse-engineering protein-protein interaction

The networks inferred from the *in vivo* datasets (Figure 8.1) contain correctly identified interactions, but also false positive interactions. We observed that most of these false interactions involved the Gal4 and Gal80 proteins. By taking into account that these proteins form a complex, we can consider *GAL4* and *GAL80* as a single component, rather than as two different ones, and simplify the true network accordingly, as shown in Figure 8.4 (True Network – Simplified). This simplification is justified by considering that reverse-engineering is performed on mRNA concentration measurements, but not on protein levels, and therefore a complete recovery of the protein-protein interaction is unlikely.

The number of correctly inferred interactions for the ODE approach increased when checked against this simplified true network. All of the inferred interaction are correct in switch-on dataset (PPV=1 and Sensitivity=0.67), as shown in Figure 8.4A. The same correct interactions are inferred from Galactose steady-state dataset (Figure 8.4C) even if with a lower precision (PPV=0.80 and Sensitivity=0.67). Results of Glucose steady-state are still not better than random (in this case random PPV=0.50) (Figure 8.4D). In the case of the switch-off time-series the performance remained the same (the ratio between the obtained PPV and the random PPV is 1.5 both in the simplified and in the original network inference). This happens because the wrongly inferred interactions do not involve the Gal4-Gal80 complex (Figure 8.4B).

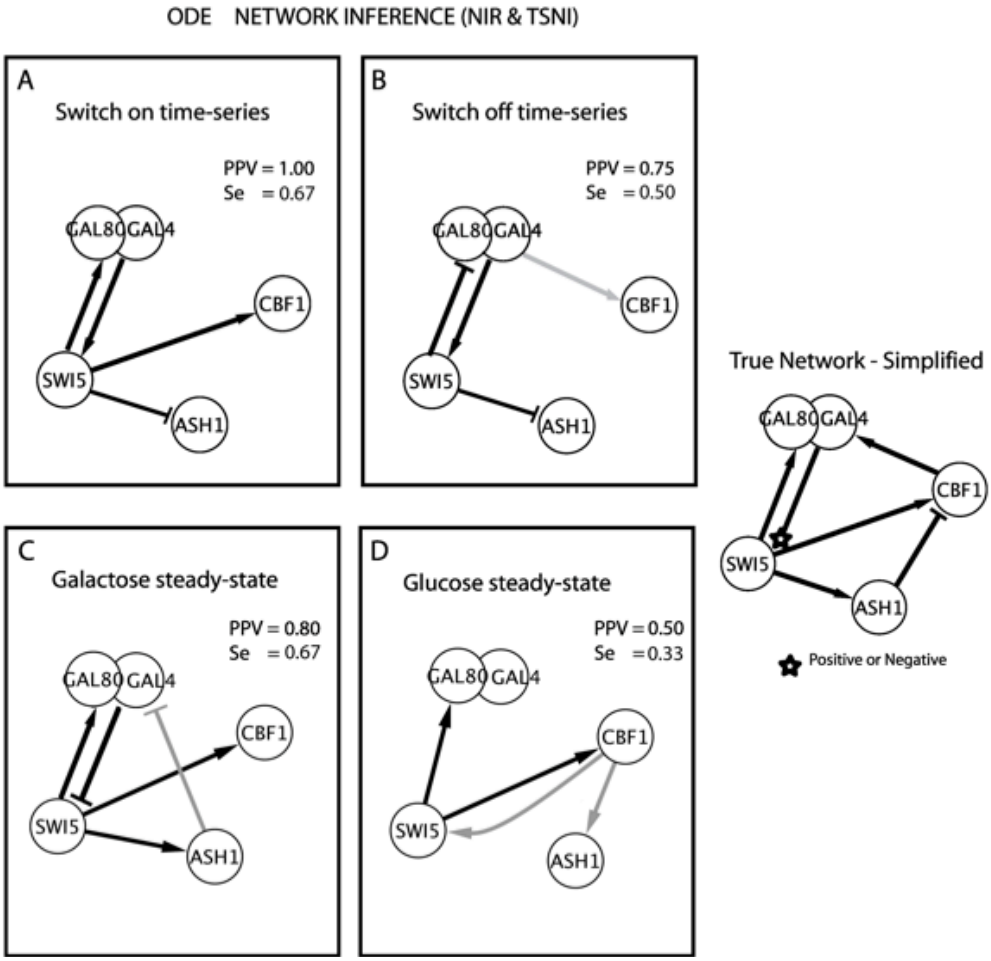

Figure 8.4. Reverse-engineering the IRMA gene network from steady-state and time-series experimental data using the ODE-based approach – Comparison with the *simplified* true network. The *Simplified* true network depicted on the right shows only the regulatory transcriptional interactions among genes in IRMA. We grouped the Gal4 and Gal80 proteins as a single component, so that all the interactions represent only transcriptional regulation. Directed edges with an arrow-end represent activation, whereas dash-end represents inhibition; **(A)-(B)** Inferred network using the TSNI reverse-engineering algorithm and the switch-on and switch-off time-series experiments. Solid grey lines represent inferred interactions that are not present in the real network, or that have the wrong direction (False Positives- FP). PPV and Se values summarize the performance of the algorithm for an unsigned directed graph. The random PPV for the unsigned directed graph is equal to 0.50. **(C)-(D)** Inferred network using the NIR reverse-engineering algorithm and the steady-state experimental data after gene overexpression in cells grown in galactose or glucose medium, respectively.

Chapter 9 – Discussion

One of the key challenges of Systems Biology is to reverse-engineer gene networks from gene expression data. To date, however, there is not a common benchmark *in vivo* that can be used to assess and compare the different strategies. *In silico* benchmarks, i.e. gene expression data simulated with a computational model, have been extensively used to this end (Camacho et al., 2007). However they are biased by the modelling strategies used to generate the data and therefore they are not 'objective'.

Together with Systems biology, the field of Synthetic biology is rapidly emerging with the aim of building 'de novo' biological circuits, or networks, to perform specific functions. Modelling the behaviour of biological circuits *in silico* before their construction is a defining feature of Synthetic biology. Ideally, different biological networks are modelled and their behaviour simulated *in silico* to check which network better performs the desired task. The selected network is then physically constructed. Gene synthetic networks formed of 2-3 genes such as the genetic toggle switch in *E. coli* (Gardner et al., 2000) and in mammalian cells (Kramer et al., 2004), the bacterial 'repressilator' (Elowitz and Leibler, 2000) or the mammalian oscillator (Tigges et al., 2009) have been built to add specific new functions to living organisms. However, none of these systems have been constructed with the aim to develop a benchmark to implement modelling strategies or reverse-engineering algorithms but have been thought to resemble a specific function.

In this work, we developed a synthetic network to assess and benchmark modelling and reverse-engineering strategies.

We showed that the semi-quantitative prediction of cell behaviour is possible, even with a simplified phenomenological differential equation model. One of the difficulties in obtaining a predictive and quantitative model in biology is the choice of the unknown kinetic parameters, especially for complex networks like the one in this work (33 parameters). Different set of parameters may yield similar results. Ideally, the kinetic parameters should be identified by appropriate experiments, and this is not always possible, particularly if one wants to obtain quantitative values (Rosenfeld et al., 2005). In this work, we were able to measure, semi-quantitatively, the strength of the promoters, and we estimated 16, out of 33 parameters from these data. Remarkably, despite all of the simplifications made, the model showed predictive power, albeit semi-quantitative. In order to have more quantitative predictions, the predictive 'scope' of the model has to be considered. In our case, the model was learned from a dynamic time-series of 5 hours after galactose addition, but then used to predict the behaviour of the system at long time-scales (i.e. steady state, or switch-off after having grown cells overnight in galactose). Since proteins were not modelled explicitly, their accumulation has larger effects in this case. The model relates only transcriptional levels of genes to each other assuming that protein and mRNA concentrations are proportional. We believe that the discrepancies between model predictions and experimental data are mainly due to this simplification. Hence, our model is semi-quantitative.

We observed that the model correctly predicts the steady-state transcriptional level of *GAL4* in galactose (t=0 in Figure 6.6 solid red line - right panel), which corresponds to the first point of the switch-off time-series (t=0 in Figure 6.6, black line with dots - right panel). However, *SWI5* (Gal4 direct target in the network), *CBF1* and *ASH1* (directly regulated by Swi5) have much higher

transcriptional levels at steady-state (t=0 in Fig. 6.6 black lines with dots – right panel), than the last point of the switch-on time-series (t=280 in Fig. 6.6 black lines with dots – left panel), despite *GAL4* transcriptional level being similar (~0.01).

Gal4 protein is more stable in galactose than in glucose and it accumulates after glucose-to-galactose shift, despite its low mRNA levels (Muratani et al., 2005; Nalley et al., 2006). The 17 unknown parameters, learned during the switch-on experiments, were estimated when Gal4 protein has not yet reached its maximal level. The model was then used to predict the switch-off time-series, where the level of Gal4 protein is initially very high. As a consequence, there is an increase in *SWI5* levels, which in turn drives *CBF1* and *ASH1* expression, which are not well captured by the model. We believe that this is a major determinant of the discrepancy between model predictions and experimental data.

The same explanation applies for the differences between model prediction and experiments in the steady-state overexpression data. In this case, cells stably over-expressed each of the five genes of the network. For the *CBF1* overexpression experiment (Figure 6.9 A and B), the model accurately predicts *GAL4* levels at the steady state but, for the same reason we explained above, *SWI5* and its target genes have higher levels in experiments as compared to the predictions.

More accurate models, including, for example, a detailed description of the galactose system, or those based on different formalisms, can be developed, depending on the biological question to be investigated, and assessed against the same ground-truth provided by our synthetic network.

We also confirmed the usefulness of the network as a benchmark for assessing reverse engineering. Our results enabled us to draw some definite conclusions: (1) when the dataset are informative, reverse-engineering algorithms are able to correctly

identify direct regulatory interactions, but some precautions must be taken when using Bayesian networks on dynamic time-series regarding the number of time-points. It is likely that the larger number of experimental time points (21 points) in the switch-off experiment as compared to the switch-on experiment (16 points) improved the performance of Dynamic Bayesian Networks, since this method needs to estimate joint probabilities, whereas the ODE approach is not greatly affected by the number of points, as long as the dynamics are well captured by the sampling time; (2) by comparing the results of different reverse-engineering algorithms on the same dataset it is possible to increase the accuracy of the predictions; (3) time-series and steady-state data are both useful for reverse engineering, but they can convey different information; (4) if knowledge of the perturbation effect is available (i.e. which gene has been over-expressed) and data points are limited, ODE are superior to Bayesian Networks. These conclusions were drawn from our small-scale network consisting of five genes only, yet they should hold also for large-scale networks. Comparison of reverse engineering methods using *in silico* expression data has shown that performances on small networks (in the order of 10 genes) are in line with those on larger networks (in the order of 100 or 1000 genes) (Bansal et al., 2006; Stolovitzky et al., 2007). Namely, if an algorithm works better than another on a small network, it will do so also on larger networks, as long as the number of experimental data points scales with the size of the network. IRMA, therefore can be used to test algorithms designed for large-scale networks, with some exceptions. Association-based algorithms (such as ARACNE) cannot be properly assessed, since the random precision for a small undirected network is too high. We observe, however, that transcription factor genes in the network regulate additional endogenous 'non-network' genes (i.e. their well characterised transcriptional targets). Thus, if a

sufficient number of genome-wide expression data is collected, then our network could be a useful benchmark, also in the case of large-scale networks.

Concomitantly with our synthetic network, we also generated both dynamic and static data sets after perturbing the system and we showed their usefulness for testing and comparing some of the available computational tools. These data are now available to the community and can be used as gold standard to test published or novel developed algorithms. The data sets produced in this work were obtained perturbing the system by changing the carbon source or by overexpressing single genes of the network, but we also showed that methionine modulates the expression of network genes (Figure 6.5) and can therefore be used to perturb system dynamics and collect new data sets.

Furthermore, being an *in vivo* system, IRMA allows also testing of different experimental strategies, which can support computational tools. As new experimental techniques, measuring for instance protein levels, will be developed it will be possible to test how reverse-engineering algorithms and refined models, which take into account also protein dynamics, work in combination with this type of data, instead than with gene expression ones.

In addition, IRMA transcriptional cassettes can be swapped, or substituted with different ones, to yield different topologies. It is also possible to extend the network, thus increasing both the number of genes and the number of interactions, by adding new cassettes. In our strain, one resistance gene (*HIS*), is available for integration of additional cassettes; furthermore new dominant resistance markers such as *ble(r)* and *pat*, which confer resistance to the antibiotic phleomycin and biaphalos respectively, have been flanked by LoxP sites (Gueldener et al., 2002). Thus they can

be Cre-excised and re-integrated in association with different transcriptional cassettes, multiple times.

High-throughput approaches often generate lists of target genes or proteins that need a heroic effort to be validated. On the other hand, computational approaches can help in inferring the regulatory interactions within a complex biological process; in reality, however, it is difficult to identify the appropriate computational approach to solve a specific biological problem, without an experimental validation of the computational predictions.

IRMA will help reducing the *in vivo* validation steps and represents the first comprehensive resource, providing both a yeast strain, and gold-standard data, to benchmark network-reconstruction and modelling strategies using an “a priori” known network.

Bibliography

Acar, M., Becskei, A., and van Oudenaarden, A. (2005). Enhancement of cellular memory by reducing stochastic transitions. Nature 435, 228-232.

Alberti, S., Gitler, A.D., and Lindquist, S. (2007). A suite of Gateway cloning vectors for high-throughput genetic analysis in Saccharomyces cerevisiae. Yeast 24, 913-919.

Alon, U. (2006). An Introduction to Systems Biology. Design principles of biological circuits, Vol 10 (London, CRC Press, Taylor & Francis Group).

Ander, M., Beltrao, P., Di Ventura, B., Ferkinghoff-Borg, J., Foglierini, M., Kaplan, A., Lemerle, C., Tomas-Oliveira, I., and Serrano, L. (2004). SmartCell, a framework to simulate cellular processes that combines stochastic approximation with diffusion and localisation: analysis of simple networks. Systems biology 1, 129-138.

Atkinson, M.R., Savageau, M.A., Myers, J.T., and Ninfa, A.J. (2003). Development of genetic circuitry exhibiting toggle switch or oscillatory behavior in Escherichia coli. Cell 113, 597-607.

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

Bansal, M., Gatta, G.D., and di Bernardo, D. (2006). Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. Bioinformatics 22, 815-822.

Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U., Dalla-Favera, R., and Califano, A. (2005). Reverse engineering of regulatory networks in human B cells. Nature genetics 37, 382-390.

Batt, G., Ropers, D., de Jong, H., Geiselmann, J., Mateescu, R., Page, M., and Schneider, D. (2005). Validation of qualitative models of genetic regulatory networks by model checking: analysis of the nutritional stress response in Escherichia coli. Bioinformatics 21 Suppl 1, i19-28.

Bennett, M.R., and Hasty, J. (2008). Systems biology: genome rewired. Nature 452, 824-825.

Bennett, M.R., Pang, W.L., Ostroff, N.A., Baumgartner, B.L., Nayak, S., Tsimring, L.S., and Hasty, J. (2008). Metabolic gene regulation in a dynamically changing environment. Nature 454, 1119-1122.

Bhoite, L.T., Yu, Y., and Stillman, D.J. (2001). The Swi5 activator recruits the Mediator complex to the HO promoter without RNA polymerase II. Genes & development 15, 2457-2469.

Bobola, N., Jansen, R.P., Shin, T.H., and Nasmyth, K. (1996). Asymmetric accumulation of Ash1p in postanaphase nuclei depends on a myosin and restricts yeast mating-type switching to mother cells. Cell 84, 699-709.

Brown, A.J.P., and Tuite, M., eds. (1998). Yeast gene Analysis (Academic press).

Camacho, D., Vera Licona, P., Mendes, P., and Laubenbacher, R. (2007). Comparison of reverse-engineering methods using an in silico network. Ann N Y Acad Sci 1115, 73-89.

Cantone, I., Marucci, L., Iorio, F., Ricci, M.A., Belcastro, V., Bansal, M., Santini, S., di Bernardo, M., di Bernardo, D., and Cosma, M.P. (2009). A Yeast Synthetic Network for In Vivo Assessment of Reverse-Engineering and Modeling Approaches. Cell 137.

Carrozza, M.J., Florens, L., Swanson, S.K., Shia, W.J., Anderson, S., Yates, J., Washburn, M.P., and Workman, J.L. (2005). Stable incorporation of sequence specific repressors Ash1 and Ume6 into the Rpd3L complex. Biochimica et biophysica acta 1731, 77-87; discussion 75-76.

Chandrasekaran, S., Deffenbaugh, A.E., Ford, D.A., Bailly, E., Mathias, N., and Skowyra, D. (2006). Destabilization of binding to cofactors and SCFMet30 is the rate-limiting regulatory step in degradation of polyubiquitinated Met4. Molecular cell 24, 689-699.

Chandrasekaran, S., and Skowyra, D. (2008). The emerging regulatory potential of SCFMet30 -mediated polyubiquitination and proteolysis of the Met4 transcriptional activator. Cell division 3, 11.

Chin, J.W. (2006). Programming and engineering biological networks. Current opinion in structural biology 16, 551-556.

Cosma, M.P. (2002). Ordered recruitment: gene-specific mechanism of transcription activation. Molecular cell 10, 227-236.

Cosma, M.P. (2004). Daughter-specific repression of Saccharomyces cerevisiae HO: Ash1 is the commander. EMBO Rep 5, 953-957.

Cosma, M.P., Tanaka, T., and Nasmyth, K. (1999). Ordered recruitment of transcription and chromatin remodeling factors to a cell cycle- and developmentally regulated promoter. Cell 97, 299-311.

Cross, F.R., and Tinkelenberg, A.H. (1991). A potential positive feedback loop controlling CLN1 and CLN2 gene expression at the start of the yeast cell cycle. Cell 65, 875-883.

Csete, M.E., and Doyle, J.C. (2002). Reverse engineering of biological complexity. Science 295, 1664-1669.

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

Della Gatta, G., Bansal, M., Ambesi-Impiombato, A., Antonini, D., Missero, C., and di Bernardo, D. (2008). Direct targets of the TRP63 transcription factor revealed by a combination of gene expression profiling and reverse engineering. Genome Res 18, 939-948.

di Bernardo, D., Thompson, M.J., Gardner, T.S., Chobot, S.E., Eastwood, E.L., Wojtovich, A.P., Elliott, S.J., Schaus, S.E., and Collins, J.J. (2005). Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. Nature biotechnology 23, 377-383.

Di Ventura, B., Lemerle, C., Michalodimitrakis, K., and Serrano, L. (2006). From in vivo to in silico biology and back. Nature 443, 527-533.

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

Elowitz, M.B., and Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators. Nature 403, 335-338.

Faith, J., and Gardner, T.S. (2005). Reverse-engineering transcription control networks. Physics of Life Reviews 2, 65-88.

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

Ferreiro, J.A., Powell, N.G., Karabetsou, N., Kent, N.A., Mellor, J., and Waters, R. (2004). Cbf1p modulates chromatin structure, transcription and repair at the Saccharomyces cerevisiae MET16 locus. Nucleic acids research 32, 1617-1626.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. J Comput Biol 7, 601-620.

Fung, E., Wong, W.W., Suen, J.K., Bulter, T., Lee, S.G., and Liao, J.C. (2005). A synthetic gene-metabolic oscillator. Nature 435, 118-122.

Gardner, T.S., Cantor, C.R., and Collins, J.J. (2000). Construction of a genetic toggle switch in Escherichia coli. Nature 403, 339-342.

Gardner, T.S., di Bernardo, D., Lorenz, D., and Collins, J.J. (2003). Inferring genetic networks and identifying compound mode of action via expression profiling. Science 301, 102-105.

Gietz, R.D., and Woods, R.A. (1994). High efficiency transformation with litium acetate. In Molecular Genetics of Yeast, A Practical Approach, J.R. Johnston, ed. (Oxford, IRL Press), pp. 121-134.

Goldstein, A.L., and McCusker, J.H. (1999). Three new dominant drug resistance cassettes for gene disruption in Saccharomyces cerevisiae. Yeast 15, 1541-1553.

Gonsalvez, G.B., Lehmann, K.A., Ho, D.K., Stanitsa, E.S., Williamson, J.R., and Long, R.M. (2003). RNA-protein interactions promote asymmetric sorting of the ASH1 mRNA ribonucleoprotein complex. Rna 9, 1383-1399.

Gueldener, U., Heinisch, J., Koehler, G.J., Voss, D., and Hegemann, J.H. (2002). A second set of loxP marker cassettes for Cre-mediated multiple gene knockouts in budding yeast. Nucleic acids research 30, e23.

Hanggi, P. (2002). Stochastic resonance in biology. How noise can enhance detection of weak signals and help improve biological information processing. Chemphyschem 3, 285-290.

Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J., et al. (2004). Transcriptional regulatory code of a eukaryotic genome. Nature 431, 99-104.

Hardin, P.E. (2005). The circadian timekeeping system of Drosophila. Curr Biol 15, R714-722.

Hasty, J., McMillen, D., and Collins, J.J. (2002). Engineered gene circuits. Nature 420, 224-230.

Hayete, B., Gardner, T.S., and Collins, J.J. (2007). Size matters: network inference tackles the genome scale. Molecular systems biology 3, 77.

Janke, C., Magiera, M.M., Rathfelder, N., Taxis, C., Reber, S., Maekawa, H., Moreno-Borchart, A., Doenges, G., Schwob, E., Schiebel, E., et al. (2004). A versatile toolbox for PCR-based tagging of yeast genes: new fluorescent proteins, more markers and promoter substitution cassettes. Yeast 21, 947-962.

Jansen, R.P., Dowzer, C., Michaelis, C., Galova, M., and Nasmyth, K. (1996). Mother cell-specific HO expression in budding yeast depends on the unconventional myosin myo4p and other cytoplasmic proteins. Cell 84, 687-697.

Jona, G., Choder, M., and Gileadi, O. (2000). Glucose starvation induces a drastic reduction in the rates of both transcription and degradation of mRNA in yeast. Biochimica et biophysica acta 1491, 37-48.

Kaern, M., Blake, W.J., and Collins, J.J. (2003). The engineering of gene regulatory networks. Annu Rev Biomed Eng 5, 179-206.

Khosla, C., and Keasling, J.D. (2003). Metabolic engineering for drug discovery and development. Nat Rev Drug Discov 2, 1019-1025.

Kramer, B.P., Viretta, A.U., Daoud-El-Baba, M., Aubel, D., Weber, W., and Fussenegger, M. (2004). An engineered epigenetic transgene switch in mammalian cells. Nature biotechnology 22, 867-870.

Kuras, L., Barbey, R., and Thomas, D. (1997). Assembly of a bZIP-bHLH transcription activation complex: formation of the yeast Cbf1-Met4-Met28 complex is regulated through Met28 stimulation of Cbf1 DNA binding. The EMBO journal 16, 2441-2451.

Kuras, L., and Thomas, D. (1995). Identification of the yeast methionine biosynthetic genes that require the centromere binding factor 1 for their transcriptional activation. FEBS Lett 367, 15-18.

Lahav, G., Rosenfeld, N., Sigal, A., Geva-Zatorsky, N., Levine, A.J., Elowitz, M.B., and Alon, U. (2004). Dynamics of the p53-Mdm2 feedback loop in individual cells. Nature genetics 36, 147-150.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., et al. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. Science 98, 799-804.

Levskaya, A., Chevalier, A.A., Tabor, J.J., Simpson, Z.B., Lavery, L.A., Levy, M., Davidson, E.A., Scouras, A., Ellington, A.D., Marcotte, E.M., et al. (2005). Synthetic biology: engineering Escherichia coli to see light. Nature 438, 441-442.

Liu, J., Wilson, T.E., Milbrandt, J., and Johnston, M. (1993). Identifying DNA-Binding Domains Using a Yeast Selection System. Methods: A Companion to Methods in Enzymology 5, 125-137.

Locke, J.C., Southern, M.M., Kozma-Bognar, L., Hibberd, V., Brown, P.E., Turner, M.S., and Millar, A.J. (2005). Extension of a genetic network model by iterative experimentation and mathematical analysis. Molecular systems biology 1, 2005 0013.

Long, R.M., Singer, R.H., Meng, X., Gonzalez, I., Nasmyth, K., and Jansen, R.P. (1997). Mating type switching in yeast controlled by asymmetric localization of ASH1 mRNA. Science 277, 383-387.

Looger, L.L., Dwyer, M.A., Smith, J.J., and Hellinga, H.W. (2003). Computational design of receptor and sensor proteins with novel functions. Nature 423, 185-190.

Luisi, P.L. (2007). Chemical aspects of synthetic biology. Chemistry & biodiversity 4, 603-621.

Luisi, P.L., Ferri, F., and Stano, P. (2006). Approaches to semi-synthetic minimal cells: a review. Die Naturwissenschaften 93, 1-13.

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

Maxon, M.E., and Herskowitz, I. (2001). Ash1p is a site-specific DNA-binding protein that actively represses transcription. Proceedings of the National Academy of Sciences of the United States of America 98, 1495-1500.

Mellor, J., Jiang, W., Funk, M., Rathjen, J., Barnes, C.A., Hinz, T., Hegemann, J.H., and Philippsen, P. (1990). CPF1, a yeast protein which functions in centromeres and promoters. The EMBO journal 9, 4017-4026.

Menant, A., Baudouin-Cornu, P., Peyraud, C., Tyers, M., and Thomas, D. (2006). Determinants of the ubiquitin-mediated degradation of the Met4 transcription factor. The Journal of biological chemistry 281, 11744-11754.

Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., and Alon, U. (2004). Superfamilies of evolved and designed networks. Science 303, 1538-1542.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: simple building blocks of complex networks. Science 298, 824-827.

Mitra, D., Parnell, E.J., Landon, J.W., Yu, Y., and Stillman, D.J. (2006). SWI/SNF binding to the HO promoter requires histone acetylation and stimulates TATA-binding protein recruitment. Molecular and cellular biology 26, 4095-4110.

Moll, T., Tebb, G., Surana, U., Robitsch, H., and Nasmyth, K. (1991). The role of phosphorylation and the CDC28 protein kinase in cell cycle-regulated nuclear import of the S. cerevisiae transcription factor SWI5. Cell 66, 743-758.

Muratani, M., Kung, C., Shokat, K.M., and Tansey, W.P. (2005). The F box protein Dsg1/Mdm30 is a transcriptional coactivator that stimulates Gal4 turnover and cotranscriptional mRNA processing. Cell 120, 887-899.

Nalley, K., Johnston, S.A., and Kodadek, T. (2006). Proteolytic turnover of the Gal4 transcription factor is not required for function in vivo. Nature 442, 1054-1057.

Nasmyth, K., Adolf, G., Lydall, D., and Seddon, A. (1990). The identification of a second cell cycle control on the HO promoter in yeast: cell cycle regulation of SW15 nuclear entry. Cell 62, 631-647.

Nasmyth, K., Stillman, D., and Kipling, D. (1987). Both positive and negative regulators of HO transcription are required for mother-cell-specific mating-type switching in yeast. Cell 48, 579-587.

O'Connell, K.F., Surdin-Kerjan, Y., and Baker, R.E. (1995). Role of the Saccharomyces cerevisiae general regulatory factor CP1 in methionine biosynthetic gene transcription. Molecular and cellular biology 15, 1879-1888.

Panda, S., Hogenesch, J.B., and Kay, S.A. (2002). Circadian rhythms from flies to human. Nature 417, 329-335.

Ro, D.K., Paradise, E.M., Ouellet, M., Fisher, K.J., Newman, K.L., Ndungu, J.M., Ho, K.A., Eachus, R.A., Ham, T.S., Kirby, J., et al. (2006). Production of the antimalarial drug precursor artemisinic acid in engineered yeast. Nature 440, 940-943.

Rosenfeld, N., Young, J.W., Alon, U., Swain, P.S., and Elowitz, M.B. (2005). Gene regulation at the single-cell level. Science (New York, NY 307, 1962-1965.

Sauer, U., Heinemann, M., and Zamboni, N. (2007). Genetics. Getting closer to the whole picture. Science 316, 550-551.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nature genetics 34, 166-176.

Segre, D., Vitkup, D., and Church, G.M. (2002). Analysis of optimality in natural and perturbed metabolic networks. Proceedings of the National Academy of Sciences of the United States of America 99, 15112-15117.

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

Sprinzak, D., and Elowitz, M.B. (2005). Reconstruction of genetic circuits. Nature 438, 443-448.

Steuer, R., Kurths, J., Daub, C.O., Weise, J., and Selbig, J. (2002). The mutual information: detecting and evaluating dependencies between variables. Bioinformatics 18 Suppl 2, S231-240.

Stolovitzky, G., Monroe, D., and Califano, A. (2007). Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference. Ann N Y Acad Sci 1115, 1-22.

Stricker, J., Cookson, S., Bennett, M.R., Mather, W.H., Tsimring, L.S., and Hasty, J. (2008). A fast, robust and tunable synthetic gene oscillator. Nature 456, 516-519.

Szallasi, Z., Stelling, J., and Periwal, V., eds. (2006). System Modeling in Cellular Biology: From Concepts to Nuts and Bolts (Boston, The MIT Press).

Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., and Church, G.M. (1999). Systematic determination of genetic network architecture. Nature genetics 22, 281-285.

Thattai, M., and van Oudenaarden, A. (2004). Stochastic gene expression in fluctuating environments. Genetics 167, 523-530.

Thomas, D., and Surdin-Kerjan, Y. (1997). Metabolism of sulfur amino acids in Saccharomyces cerevisiae. Microbiol Mol Biol Rev 61, 503-532.

Tigges, M., Marquez-Lago, T.T., Stelling, J., and Fussenegger, M. (2009). A tunable synthetic mammalian oscillator. Nature 457, 309-312.

Traven, A., Jelicic, B., and Sopta, M. (2006). Yeast Gal4: a transcriptional paradigm revisited. EMBO Rep 7, 496-499.

Verma, M., Bhat, P.J., and Venkatesh, K.V. (2004). Expression of GAL genes in a mutant strain of Saccharomyces cerevisiae lacking GAL80: quantitative model and experimental verification. Biotechnology and applied biochemistry 39, 89-97.

Visintin, R., Craig, K., Hwang, E.S., Prinz, S., Tyers, M., and Amon, A. (1998). The phosphatase Cdc14 triggers mitotic exit by reversal of Cdk-dependent phosphorylation. Molecular cell 2, 709-718.

Voth, W.P., Yu, Y., Takahata, S., Kretschmann, K.L., Lieb, J.D., Parker, R.L., Milash, B., and Stillman, D.J. (2007). Forkhead proteins control the outcome of transcription factor binding by antiactivation. The EMBO journal 26, 4324-4334.

Wach, A., Brachat, A., Alberti-Segui, C., Rebischung, C., and Philippsen, P. (1997). Heterologous HIS3 marker and GFP reporter modules for PCR-targeting in Saccharomyces cerevisiae. Yeast 13, 1065-1075.

Yeger-Lotem, E., and Margalit, H. (2003). Detection of regulatory circuits by integrating the cellular networks of protein-protein interactions and transcription regulation. Nucleic acids research 31, 6053-6061.

Yeger-Lotem, E., Sattath, S., Kashtan, N., Itzkovitz, S., Milo, R., Pinter, R.Y., Alon, U., and Margalit, H. (2004). Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. Proceedings of the National Academy of Sciences of the United States of America 101, 5934-5939.

Yu, J., Smith, V.A., Wang, P.P., Hartemink, A.J., and Jarvis, E.D. (2004). Advances to Bayesian network inference for generating causal networks from observational biological data. Bioinformatics 20, 3594-3603.

Zhang, X., Dennis, P., Ehrenberg, M., and Bremer, H. (2002). Kinetic properties of rrn promoters in Escherichia coli. Biochimie 84, 981-996.

Zhou, X.J., Kao, M.C., Huang, H., Wong, A., Nunez-Iglesias, J., Primig, M., Aparicio, O.M., Finch, C.E., Morgan, T.E., and Wong, W.H. (2005). Functional annotation and network reconstruction through cross-platform integration of microarray data. Nature biotechnology 23, 238-243.