Cell

Resource

A Yeast Synthetic Network for In Vivo Assessment of Reverse-Engineering and Modeling Approaches

Irene Cantone,[1] Lucia Marucci,[1,2] Francesco Iorio,[1] Maria Aurelia Ricci,[1] Vincenzo Belcastro,[1] Mukesh Bansal,[1] Stefania Santini,[2] Mario di Bernardo,[2] Diego di Bernardo,[1,2,3,*] and Maria Pia Cosma[1,3,*]

[1]Telethon Institute of Genetics and Medicine (TIGEM), Naples 80131, Italy

[2]Department of Computer and Systems Engineering, University of Naples "Federico II," Naples 80125, Italy

[3]These authors contributed equally to this work

*Correspondence: dibernardo@tigem.it (D.d.B.), cosma@tigem.it (M.P.C.)

DOI 10.1016/j.cell.2009.01.055

SUMMARY

Systems biology approaches are extensively used to model and reverse engineer gene regulatory networks from experimental data. Conversely, synthetic biology allows "de novo" construction of a regulatory network to seed new functions in the cell. At present, the usefulness and predictive ability of modeling and reverse engineering cannot be assessed and compared rigorously. We built in the yeast *Saccharomyces cerevisiae* a synthetic network, IRMA, for in vivo "benchmarking" of reverse-engineering and modeling approaches. The network is composed of five genes regulating each other through a variety of regulatory interactions; it is negligibly affected by endogenous genes, and it is responsive to small molecules. We measured time series and steady-state expression data after multiple perturbations. These data were used to assess state-of-the-art modeling and reverse-engineering techniques. A semiquantitative model was able to capture and predict the behavior of the network. Reverse engineering based on differential equations and Bayesian networks correctly inferred regulatory interactions from the experimental data.

INTRODUCTION

Cellular complexity stems from the interactions among thousands of different molecular species. 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), scientists are beginning to unravel these regulatory, signaling, and metabolic interactions and to understand their coordinated action.

Systems biology aims to develop 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).

The usefulness of a model lies in its ability to formalize the knowledge about the biological process at hand, to identify inconsistencies between hypotheses and observations, and to predict the behavior of the biological process in yet untested conditions. There are 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.

Synthetic biology aims to use such models to design unique 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) (Gardner et al., 2000; Khosla and Keasling, 2003; Ro et al., 2006).

Interactions among genes, when unknown, can be identified from gene expression data using reverse-engineering methods. Typically, the data consist of measurements at steady state after multiple perturbations (i.e., gene overexpression, knockdown, or drug treatment) or at multiple time points after 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 reverse-engineering approaches is being proposed, and their assessment and evaluation is of critical importance (Stolovitzky et al., 2007). There are three well-established reverse-engineering approaches: ordinary differential equations (ODEs), Bayesian networks, and information theory.

ODEs relate changes in gene transcripts concentration to each other and to an external perturbation. The model consists of a differential equation for each of the 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.

A Bayesian network is a graphical model of probabilistic relationships among a set of random variables, with each

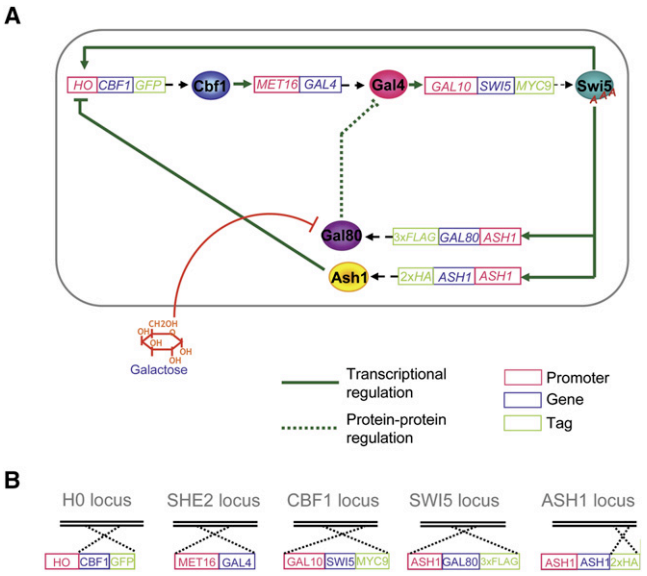


Figure 1. Construction of IRMA, a Synthetic Network in Yeast

(A) Schematic diagram of the synthetic gene network is represented. New transcriptional units (rectangles) were built by assembling promoters (red) with nonself 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 Δ*gal4* Δ*gal80 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 integration of a drug resistance cassette, *natMX4* (not shown).

variable representing one of the genes in the network. These relationships (i.e., the gene-gene interactions) are encoded in a directed graph without 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).

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 coexpressed across the experimental data set. Coexpression can be measured either by correlation or by a more robust measure called mutual information (Bansal et al., 2007).

Here, we constructed, in the yeast *Saccharomyces cerevisiae*, a synthetic network of five genes regulating each other for in vivo reverse-engineering and modeling assessment (IRMA). We chose the simplest eukaryotic organism, *Saccharomyces cerevisiae*, because it can easily be grown and manipulated. The synthetic network includes a variety of regulatory interactions, thus capturing the behavior of larger eukaryotic gene networks on a smaller scale. The network was designed to be negligibly affected by endogenous genes, and to respond to galactose, which triggers transcription of its genes. Our network, apparently simple, is in fact very articulated in its interconnections, which include regulator chains, single-input motifs, and multiple feedback loops, generated by the combination of transcriptional activators and repressors.

We analyzed the transcriptional response of network genes after two different perturbation strategies: performing a single perturbation and measuring mRNA changes at different time points, or performing multiple perturbations and collecting mRNA measurements at steady state.

We tested the usefulness of IRMA as a simplified biological model to benchmark both modeling and reverse-engineering approaches.

RESULTS

Construction of a Gene Synthetic Network in Yeast

The network, shown in Figure 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.

We chose promoters for which a single transcription factor (TF) is sufficient and essential to activate transcription (Figure S1 available online). Thus, by removing the endogenous TF, we maximally reduced influences from the cellular environment on each promoter. We selected well-characterized promoter/TF-encoding gene pairs, belonging to distinct and nonredundant pathways, to further minimize external feedbacks on the network due to pathway crosstalk. We chose nonessential and nonredundant TF genes, which can be knocked out without affecting yeast viability—specifically, as activators and repressors encoding genes *SWI5*, *ASH1*, *CBF1*, *GAL4*, and *GAL80*, and as promoter genes *HO*, *ASH1*, *MET16*, and *GAL10* (Figure 1A).

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 deleted 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

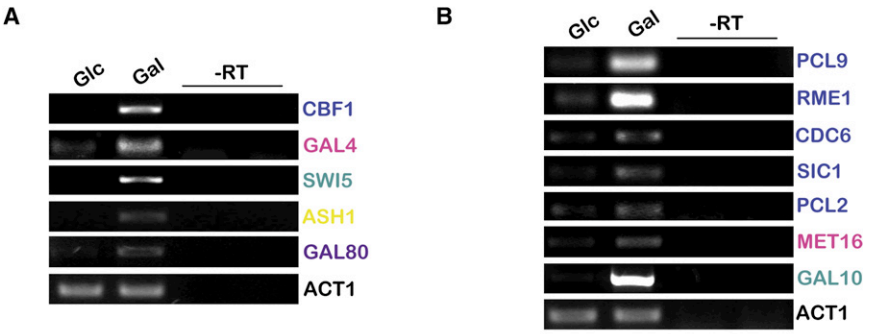
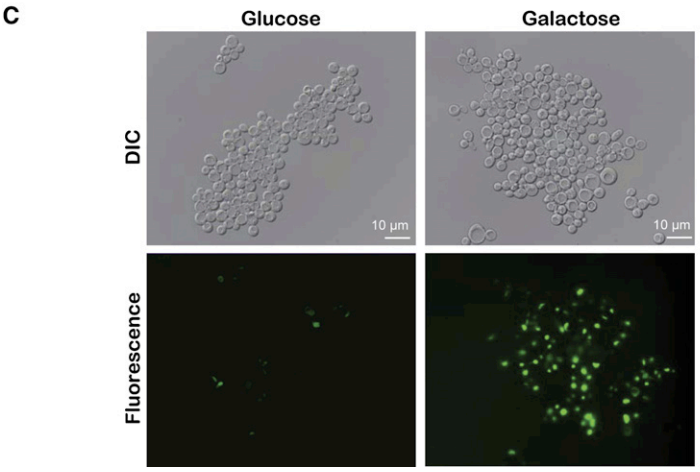


Figure 2. Galactose Triggers Activation of IRMA Synthetic Network

(A and B) Network genes, and cell genes that are network targets, are expressed only in presence of galactose. Semiquantitative PCR to amplify IRMA and IRMA-dependent genes was carried out with 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 bars represent 10 μm; 63× magnification.



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 sulfur amino acids metabolism (Mellor et al., 1990). 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 signaling 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, preventing interaction of the transcription machinery (Traven et al., 2006).

We assembled the chosen promoters upstream of nonself gene coding sequences to obtain the IRMA network. The network (Figure 1A) includes positive and negative feedback loops and one protein-protein interaction. These interactions coexist normally in many transcriptional pathways in higher eukaryotes (Lee et al., 2002).

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 1A). A fluorescence tag was cloned at the 3′ end of the *CBF1* open reading frame (ORF) to easily monitor its protein product.

We integrated these cassettes in the genome of *gal4Δ542 gal80Δ538* yeast strain YM4271, 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 minimizing influences from endogenous genes (Figure 1B).

Network Genes, and Their Endogenous Targets, Are Transcriptionally 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 2A).

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 2C, 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 2B). These genes should not influence the network behavior by means of direct or indirect feedback loops, since their functions are

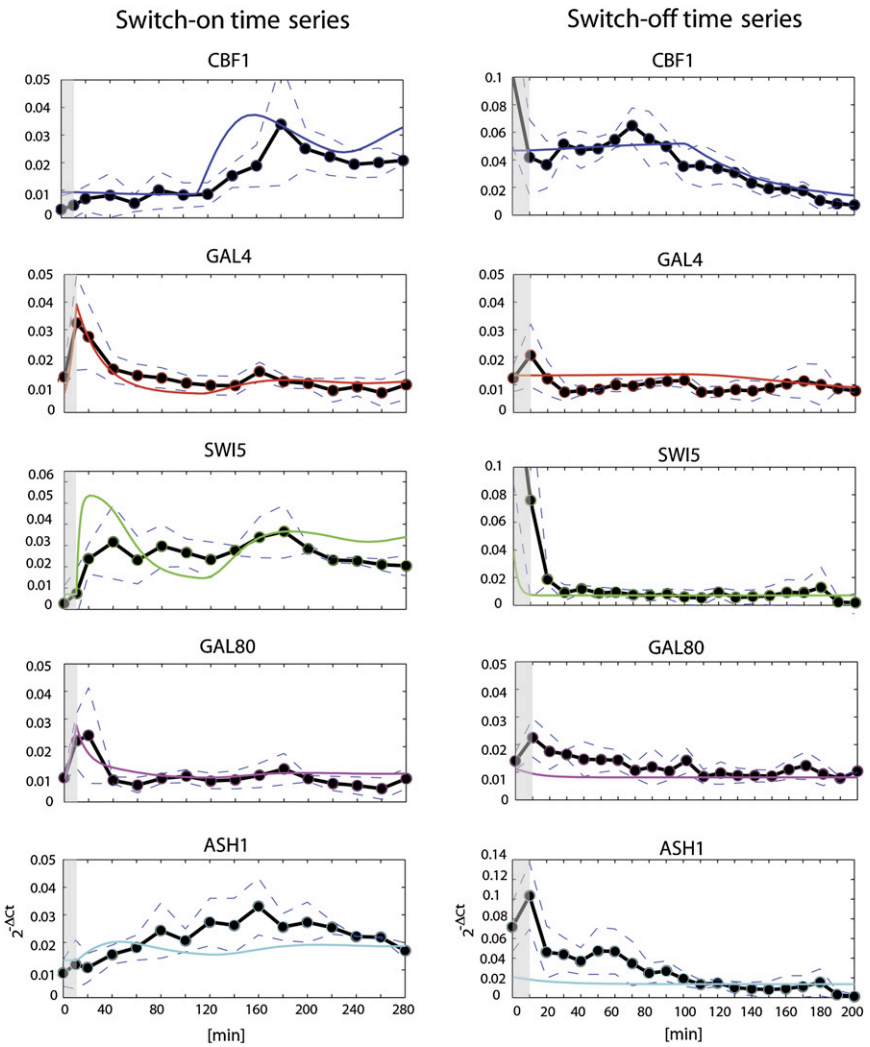


Figure 3. In Vivo and In Silico Gene Expression Profiles Show the Dynamic Behavior of the Network in Response to Medium Shift Perturbations

Expression profiles of network genes after a shift from glucose- to galactose-raffinose-containing medium—switch on—(left) and after a shift 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 ODE-based model and show how the model fits the experimental data. Gray bars indicate 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 of 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 of 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), which were processed as explained in Supplemental Data.

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.

Gene Expression Profiling of IRMA to Study Its Static and Dynamic Behavior

In order to analyze the dynamic behavior of the 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 min up to 5 hr in five independent experiments for the switch-on data set, and every 10 min up to 3 hr in four independent experiments for the switch-off data set. We analyzed expression profiles of network genes by quantitative real-time RT-PCR (q-PCR). In the switch-on experiment in Figure 3, the activation of *GAL4* by galactose led to transcription of all the other network genes. Their dynamic behavior is evident; a seemingly oscillatory behavior 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 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 3, was measured in glucose, right before shifting of 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 3, 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 S2). *GAL4* and *GAL80* showed the same increase, once the cells were transferred back in the glucose medium, after the washing steps. We believe that 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 because these two genes are controlled by different promoters. Moreover, the expression levels of the *MET16* endogenous gene, whose promoter, in our

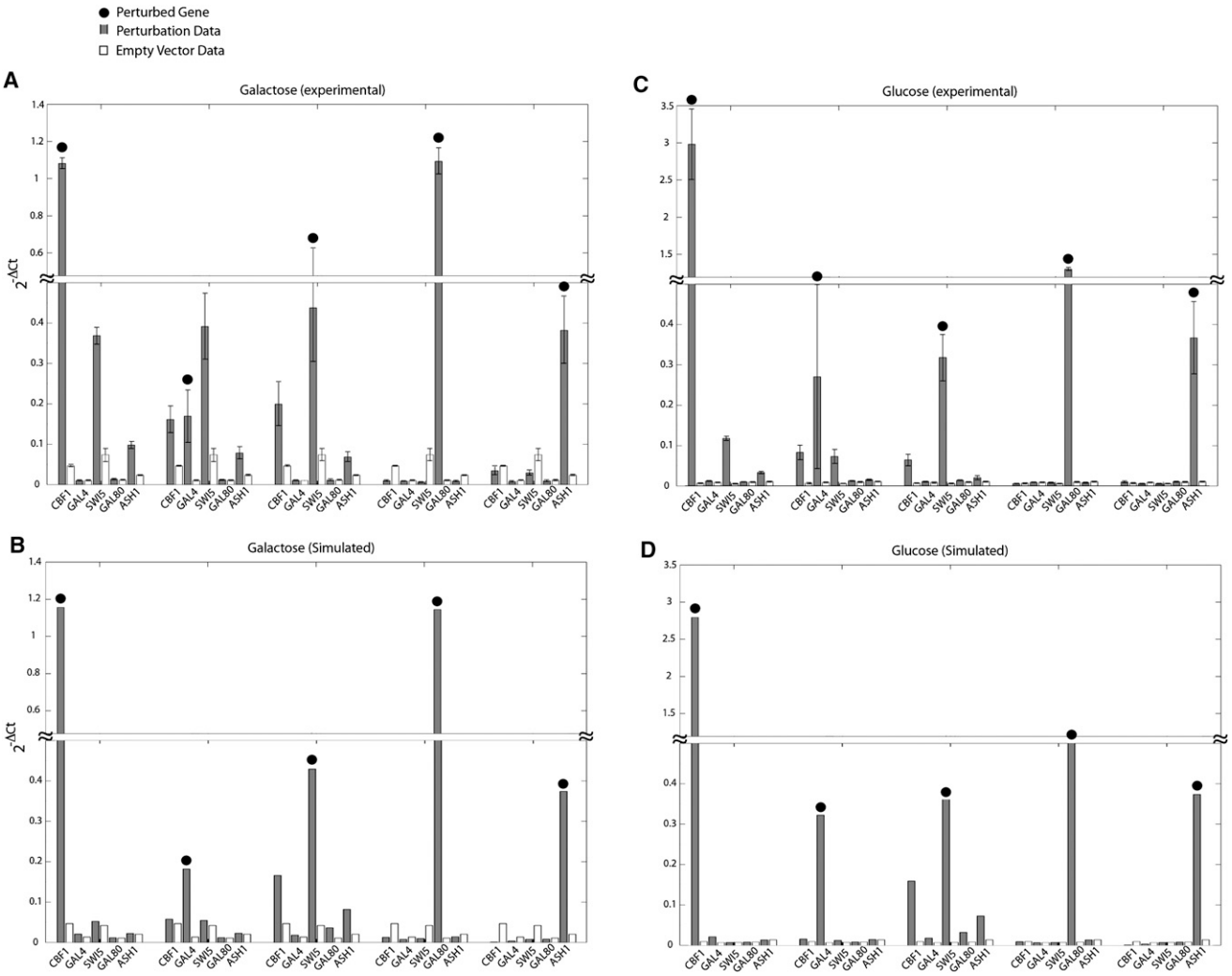


Figure 4. Experimental and Simulated Gene Expression Profiles Show the Static Behavior of IRMA in Response to Overexpression Perturbation Experiments

(A and 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 (gray 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; $n \geq 3$).

(B and D) In silico expression levels of IRMA genes obtained by simulating the overexpression of each gene with the ODE-based model.

network, is the same promoter as *GAL4*, do not show any increase in the glucose-to-glucose shift experiment, further excluding dependence on transcriptional regulation (Figure S2).

In the switch-off experiment (Figure 3), as expected, the transcription of the whole network is rapidly turned off with a delay in the silencing of *CBF1* expression.

In addition, we analyzed 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 galactose. We then measured steady-state expression levels of IRMA genes by q-PCR. We thus obtained two data sets, 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 data sets as the ''glucose steady state'' and ''galactose steady state'' (Figures 4A–4C and S3A–S3C).

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 (Figures 4A–4C and S3A–S3C). After 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 was 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 regulating *SWI5* in our network.

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.

Mathematical Model of the Network

The most common strategy for modeling gene networks is the one based on nonlinear differential equations (DEs) obtained from standard mass-balance kinetic laws (Alon, 2006; Szallasi et al., 2006). We derived a DE model of the network (Supplemental Results). 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. The DE model consists of five equations describing the transcription of the five mRNAs—*CBF1*, *GAL4*, *SWI5*, *GAL80*, *ASH1*—with 33 unknown parameters. We used Hill kinetics functions to describe transcription and considered first-order degradation terms.

In deriving the model, we took particular care in modeling the galactose pathway (Gal4 and Gal80 proteins in the network), in order to capture its main features, but without increasing model complexity. We used a phenomenological rate law to describe the activation of the *GAL10* promoter driving the transcription of *SWI5*. This promoter is activated by the amount of Gal4 that is not involved in the formation of a protein-protein complex with Gal80. In the inducing medium, the inhibition of Gal80 is relieved by the activated form of the Gal3 protein. Thus, we assumed that *SWI5* is also inhibited by a Michaelis-Menten-like term proportional to the concentration of the *GAL80* mRNA.

We included an explicit delay in the activation of *CBF1* by Swi5. This delay is apparent in the switch-on and switch-off data in Figure 3 and has been well described in the literature regarding the *HO* promoter activation by Swi5 (Bhoite et al., 2001; Cosma, 2002; Cosma et al., 1999).

We modeled the effect of cell manipulation 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).

In order to estimate the unknown parameters, we experimentally measured promoters' strength of *GAL10*, *MET16*, *ASH1*, and *HO*. We stably expressed each TF genes at different levels and measured, by q-PCR, the transcription of the corresponding promoter gene, at steady state, for a total of 165 data points (Figures S4 and S5). We then estimated 16 (out of the 33) parameters (Michaelis-Menten and the relative Hill coefficients) from these data using a stochastic optimization algorithm (described in the Supplemental Results). In addition, *GAL10* promoter was assessed in both galactose and glucose growing conditions (Figure S5, Supplemental Results, and Supplemental Experimental Procedures).

The remaining 17 unknown parameters, which could not be computed from promoters' data, were estimated from the switch-on time series (described in the Supplemental Results). The switch-off data were used to test the model predictive performance.

Figure 3 shows the experimental data and the model-simulated data for the switch-on and switch-off time series experiments. In order to simulate the switch-on data, we chose as initial conditions the steady-state equilibrium of the model in glucose, recapitulating the experimental conditions.

Simulated data fitted semiquantitatively in vivo data, despite the simplifying assumptions, being on average within the experimental standard errors (Figure 3).

The model was able to predict, semiquantitatively, the behavior of the network during the switch-off experiment (Figure 3). 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 unmodeled effect of protein accumulation of network genes. Indeed, the switch-off experiment is performed after cells have been grown 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 Figures 4 and S3, we compared in vivo and in silico experiments. There is a semiquantitative 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 behavior, 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 explicitly model protein dynamics, a small increase in *GAL4* mRNA cannot fully activate *GAL10* in the model and does not cause the 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 overexpressed 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 nonmonotonic behavior characterized 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 or totally disappear (Figures S6 and S7).

Reconstructing the Network: A Reverse-Engineering Approach

The synthetic network can be used to assess the ability of experimental and computational approaches to infer regulatory

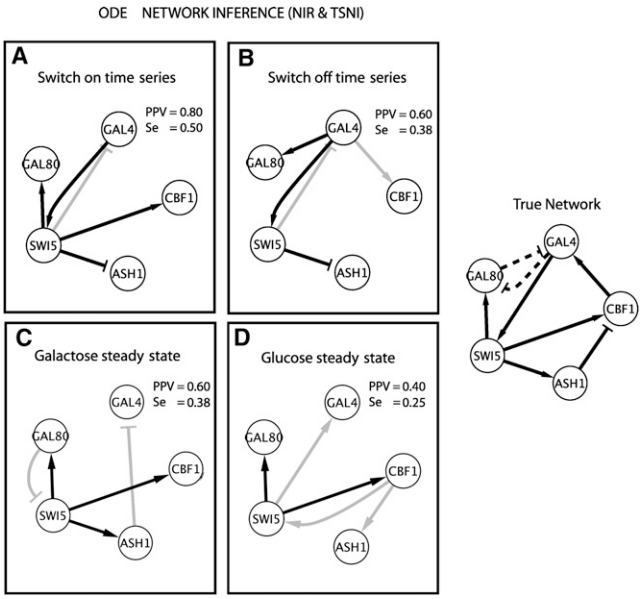


Figure 5. Reverse Engineering of the IRMA Gene Network from Steady-State and Time Series Experimental Data Using the ODE-Based Approach

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 a 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 gray lines represent inferred interactions that are not present in the real network, or that have the wrong direction (FP, false positive). 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 gene overexpression in cells grown in galactose or glucose medium, respectively.

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.

Figures 5, S8, and S9 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%).

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 5A). ODE performed better than random (PPV = 0.60, Se = 0.38) also on the switch-off data, in Figure 5B, 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 S8B). Bayesian networks failed to perform better than random on the switch-on data (Figure S8A) probably because of the lower number of points (16) as compared to the switch-off time series (21 points).

By comparison of 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 data set (compare Figure 5A with Figure S8A, and Figures 5B and S8B), 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.

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 data set (Figure 5C), but it did not perform better than random (PPV = 0.40 and Se = 0.25) in the glucose data set (Figure 5D). 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 data set (PPV = 0.60, Se = 0.38) and the glucose one (PPV = 0.50, Se = 0.38), albeit with a lower precision in the latter (Figures S8C and S8D). 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 ones (Figures S8B–S8D). These results imply that both dynamic time series data and static steady state are informative for reverse engineering.

By considering only interactions inferred by both methods on the same data set, 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 S9). 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 BANJO performed similarly for the steady-state data, but

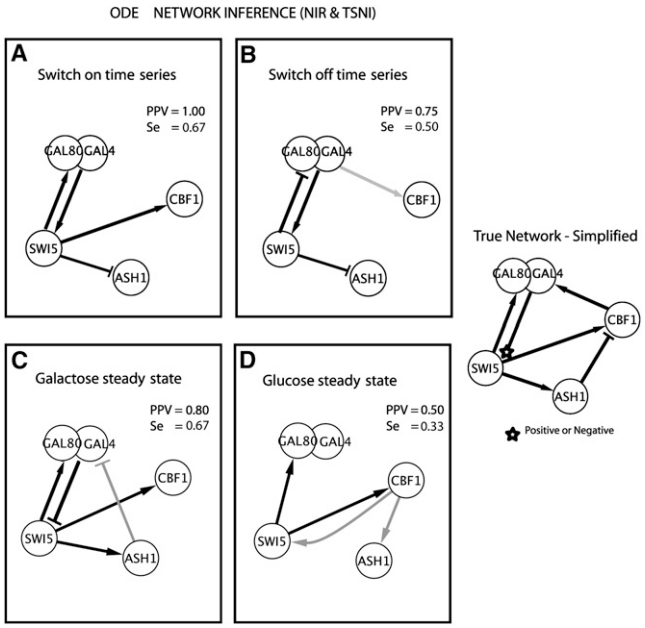


Figure 6. Reverse Engineering the IRMA Gene Network from Steady-State and Time Series Experimental Data Using the ODE-Based Approach—Comparison with the *Simplified* True Network

Simplified true network 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 a 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 gray lines represent inferred interactions that are not present in the real network, or that have the wrong direction (FP, false positive). 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 and 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.

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, because of their inability of inferring the direction of regulation. However, they are superior to other methods in the case of large networks because of their ability to require a minimal amount of data to infer gene-gene undirected interactions (Faith et al., 2007).

The networks inferred from the in vivo data sets (Figure 5) 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 6 ("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 interactions are correct in switch-on data set (PPV = 1 and Se = 0.67), as shown in Figure 6A. The same correct interactions are inferred from galactose steady-state data set (Figure 6C) even if with a lower precision (PPV = 0.80 and Se = 0.67). Results of glucose steady state are still not better than random (in this case random PPV = 0.50) (Figure 6D). 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 6B).

DISCUSSION

In this paper, we developed a synthetic network to assess and benchmark modeling and reverse-engineering strategies. We showed that the semiquantitative prediction of cell behavior 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). A 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, semiquantitatively, 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 semiquantitative. In order for there to be 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 hr after galactose addition, but then used to predict the behavior of the system at long time scales (i.e., steady state, or switch off after cells were grown overnight in galactose). Since proteins were not modeled explicitly, their accumulation will have larger effects in this case.

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 data sets are informative, reverse-engineering algorithms are able to correctly identify direct regulatory interactions, but some precautions must be taken when Bayesian networks are used on dynamic time series regarding the number of time points. It is likely that the larger number of experimental time points in the switch-off experiment (21 points) 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 data set, 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 overexpressed) and data points are limited, ODEs 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 with in silico expression data has shown that performances on small networks (in the order of ten genes) are in line with those on larger networks (in the order of 100 or 1000 genes) (Bansal et al., 2007; 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-characterized transcriptional targets). Thus, if a sufficient amount of genome-wide expression data is collected, then our network could be a useful benchmark, also in the case of large-scale networks.

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 reintegrated 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 a comprehensive resource, providing both a yeast strain and gold-standard data to benchmark network reconstruction and modeling strategies with an "a priori" known network.

EXPERIMENTAL PROCEDURES

Yeast Culture, 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. Genotypes of strains and plasmids generated in this study are listed respectively in Tables S3 and S4. Details of strain construction are given in the Supplemental Experimental Procedures.

For time series experiments, yeast cells of an 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 mRNA extraction.

For steady-state perturbation experiments, centromeric plasmids were constructed as follows. *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 pAG413GPD-ccdB (Addgene 14142) destination vectors by LR Clonase II enzyme, as previously described by Alberti et al. (2007). The IRMA-containing strain was then transformed with the obtained plasmids. Transformed cells were grown at 30°C in synthetic complete (SC) medium lacking histidine with 2% glucose or 2% galactose plus 2% raffinose to 0.6–0.8 OD_{600} and then harvested for mRNA extraction.

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. 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. The in silico experiments, mirroring the glucose steady-state and galactose steady-state in vivo experiments, were carried out by numerical solving of the mathematical model. As initial conditions, we used the steady states predicted by the model in unperturbed conditions (either in glucose or in galactose), and in addition we applied a constant input, corresponding to the gene overexpression, to each of the five equations. Details for modeling and parameter identification can be found in the Supplemental Results.

SUPPLEMENTAL DATA

Supplemental Data include Supplemental Results, Supplemental Experimental Procedures, nine figures, six tables, and IRMA in vivo data sets and can be found with this article online at http://www.cell.com/supplemental/S0092-8674(09)00156-1.

ACKNOWLEDGMENTS

This work was supported by EC Framework 6 (COBIOS) and by a "Fondazione Telethon" Institutional grant to TIGEM; I.C. and M.B. were supported by SEMM (European School of Molecular Medicine). F.I. is part of the University of Salerno Ph.D. program. M.A.R. is part of University of Naples "Federico II" undergraduate program. V.B. is part of the Open University Ph.D. program and supported by Italian MIUR-FIRB ITALBIONET to D.d.B. We thank Graciana Diez-Roux, Andrea Ballabio, and Joerg Stelling for helping with the manuscript; Kim Nasmyth for providing the K2072 yeast strain; Mark Johnston for providing the YM4271 strain; and Susan Lindquist for providing Gateway vectors.

Received: April 21, 2008
Revised: August 6, 2008
Accepted: January 29, 2009
Published online: March 26, 2009

REFERENCES

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, Volume 10 (London: CRC Press, Taylor & Francis Group).

Bansal, M., Belcastro, V., Ambesi-Impiombato, A., and di Bernardo, D. (2007). How to infer gene networks from expression profiles. Mol. Syst. Biol. *3*, 78.

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. Nat. Genet. *37*, 382–390.

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 Dev. *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.

Cosma, M.P. (2002). Ordered recruitment: gene-specific mechanism of transcription activation. Mol. 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.

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. Nat. Biotechnol. *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.

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 Res. *32*, 1617–1626.

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.

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 Res. *30*, e23.

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. Mol. Syst. Biol. *3*, 77.

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. Biochim. Biophys. 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.

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.

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 *298*, 799–804.

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.

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.

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. EMBO J. *9*, 4017–4026.

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., 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.

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.

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. Mol. Cell. Biol. *15*, 1879–1888.

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 *307*, 1962–1965.

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

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.

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

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

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. Mol. 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. EMBO J. *26*, 4324–4334.

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.