J. Math. Biol. (2011) 62:685–706
DOI 10.1007/s00285-010-0350-z

Mathematical Biology

Derivation, identification and validation of a computational model of a novel synthetic regulatory network in yeast

Lucia Marucci · Stefania Santini · Mario di Bernardo · Diego di Bernardo

Received: 4 November 2009 / Revised: 19 May 2010 / Published online: 12 June 2010

Abstract Systems biology aims at building computational models of biological pathways in order to study in silico their behaviour and to verify biological hypotheses. Modelling can become a new powerful method in molecular biology, if correctly used. Here we present step-by-step the derivation and identification of the dynamical model of a biological pathway using a novel synthetic network recently constructed in the yeast *Saccharomyces cerevisiae* for In-vivo Reverse-Engineering and Modelling Assessment. This network consists of five genes regulating each other transcription. Moreover, it includes one protein–protein interaction, and its genes can be switched on by addition of galactose to the medium. In order to describe the network dynamics, we adopted a deterministic modelling approach based on non-linear differential equations. We show how, through iteration between experiments and modelling, it is possible to derive a semi-quantitative prediction of network behaviour and to better understand the biology of the pathway of interest.

Electronic supplementary material The online version of this article (doi:10.1007/s00285-010-0350-z) contains supplementary material, which is available to authorized users.

L. Marucci (✉) · D. di Bernardo
Telethon Institute of Genetics and Medicine (TIGEM), 80131 Naples, Italy
e-mail: marucci@tigem.it

D. di Bernardo
e-mail: dibernardo@tigem.it

L. Marucci · S. Santini · M. di Bernardo · D. di Bernardo
Department of Computer and Systems Engineering,
Federico II University, 80125 Naples, Italy

M. di Bernardo
e-mail: mario.dibernardo@unina.it

Keywords Mathematical modelling · Synthetic biology · Hill functions · Parameters identification

Mathematics Subject Classification (2000) 92-08 · 93A30 · 93B30

1 Introduction

The emerging discipline of Synthetic Biology can be defined as the engineering of biology. Up to now, two major goals have been actively investigated: the building of new biological networks in the cell that perform a specific task [e.g. periodic expression of a gene (Elowitz and Leibler 2000) or genetic switching (Gardner et al. 2000)] and the modification of networks that occur in nature in order to achieve some desired functionalities (e.g. production of a specific compound useful for medical applications Ro et al. 2006). Synthetic Biology is an interdisciplinary area requiring a deep synergy between biology, biotechnology and nanotechnology on one side and mathematical modelling, information technology and control theory on the other. Such combination of disciplines is needed to construct robust and predictable synthetic networks. In particular, quantitative models are needed for a precise and unambiguous description of synthetic circuits (Kaznessis 2007). Mathematical models allow to rigorously compare hypotheses and observations, thus providing additional insight into the biological mechanisms. Model derivation from experimental data can be carried out following three major approaches: white-box, black-box and gray-box. In white-box modelling, the model and parameter values are entirely derived from first principles, while in black-box modelling the model is completely derived from input–output data. The third alternative, the so-called gray-box approach (Nelles 2000), combines the two above approaches. Specifically, first principles are used to partially derive the model structure, while parameters or terms in the model are determined by measurement data. The approach we use in this paper is a gray-box one. In this case, modelling entails three main steps to be executed iteratively: (i) derivation of the model equations; (ii) identification of the model parameters from experimental data and/or literature; (iii) validation [or invalidation (Anderson and Papachristodoulou 2009)] of the model.

Step (i) requires introducing simplifying hypothesis and choosing a proper formal framework. A huge variety of mathematical formalisms have been proposed in the literature, such as directed graphs, Bayesian networks, Boolean networks and their generalizations, ordinary and partial differential equations, qualitative differential equations, stochastic equations, and rule-based formalisms (see, for example, De Jong 2002; Ventura et al. 2006; Szallasi et al. 2006 and references therein). Deterministic formalisms are commonly used to describe the average behaviour of a population of cells (De Jong 2002). They have been shown to be viable for the analysis of synthetic networks in a number of works (e. g. Elowitz and Leibler 2000; Gardner et al. 2000; Kramer et al. 2004; Tigges et al. 2009; Stricker et al. 2008). The reaction mechanism is described by applying the law of mass action: the rate of any given elementary reaction is proportional to the product of the concentrations of the species reacting in the elementary process (reactants) (Alon 2006). The DEs modelling approach is based on the following biological assumptions: the quantified concentrations do not vary with respect to space and they are continuous functions of time. These assumptions hold for processes evolving on long time scales in which the number of molecules of the species in the reaction volume is sufficiently large. In different experimental cases, approaches based on partial differential equations or stochastic models would be more appropriate (Szallasi et al. 2006).

Step (ii) is required to estimate unknown model parameters from the available experimental data. A crucial issue that arises when estimating model parameters, is the structural identifiability (Walter and Pronzato 1997). The notion of identifiability addresses feasibility of estimating unknown parameters from data collected in well-defined stimulus-response experiments (Cobelli and Distefano 1980). Structural non-identifiability is related to the model structure independently from experimental data. In contrast, practical non-identifiability also takes into account the amount and the quality of measured data used for parameters calibration. Of note, a parameter that is structurally identifiable may still be practically non-identifiable, due to the unavoidable presence of noise in biological experimental data (Raue et al. 2009). Unfortunately, while being well assessed in the case of linear dynamical systems, the identifiability analysis of highly non-linear systems remains an open problem (Boubaker and Fourati 2004).

The parameter estimation problem can be formulated from the mathematical viewpoint as a constrained optimization problem where the goal is to minimize the objective function, defined as the error between model predictions and real data. In biological applications, the objective function usually displays a large number of local optima as measurements are strongly affected by noise. For this kind of problems, classical optimization methods, based on gradient descent from an arbitrary initial guess of the solution, can be unfeasible and show convergence difficulties. The above considerations suggest to look at stochastic optimization algorithms, like evolutionary strategies, which rely on random explorations of the whole space of solutions, are not sensitive to initial conditions and avoid trapping in local optimal points. In Moles et al. (2003), the performance of both local and global-search optimization methods is compared in the identification of the 36 unknown parameters of a non-linear biochemical network. The authors show that only evolutionary strategies are able to successfully solve the parameters estimation problem, while gradient based methods tend to converge to local minima. Among the stochastic techniques, genetic algorithms (GA) (Mitchell 1998) provide a very flexible approach to non-linear optimization. Their application showed good results in the parametrisation of synthetic networks (Weber et al. 2007; Tigges et al. 2009).

Finally, step (iii) is required to check the validity and usefulness of the model, that is to evaluate its ability in predicting the behaviour of the actual physical system. Theoretically, the modeller should be confident that the formalism is able to describe *all* input–output behaviours of the system (Smith and Doyle 1992). This condition can be never guaranteed, since it would require an infinite number of experiments. However, it is possible to test a necessary condition: the model is able to describe *all observed* input–output behaviours of the system (Smith and Doyle 1992). To this aim, one possible approach is to use a cross-validation like procedure (Arlot and Celisse 2010) by splitting the experimental data in two sets: one of them is used for the parameter identification, while the other one is used to validate the predictive power of the model. If the predictive performance of the model is not satisfactory, it is invalidated (Anderson and Papachristodoulou 2009). Thus, it is necessary to refine the model (for example, by increasing the level of detail) and/or to perform new experiments, going back to step (i) of the modelling procedure.

In what follows, we describe the gray-box modelling of IRMA network (Fig. 1) as a representative example of the modelling problem for a small biological pathway, and present the detailed derivation of the model whose equations were given in Cantone et al. (2009). Note that usually, when a genetic circuit is presented to the Synthetic Biology community, only the best performing mathematical model is showed without providing any detail of how the model was obtained. Here, instead, we provide the "history" of the derivation of the final model and experimental data-set, highlighting the major modelling choices and challenges faced during the process. The aim is to build a model able to correctly predict the dynamical changes in the mRNA concentrations of the five network genes following both internal and external perturbations (i.e. gene over-expression, galactose addition, etc.). We choose differential equations (DEs) to model the dynamics of the genes. The task is challenging since, to our knowledge, up to now quantitative DEs mathematical models have been developed for synthetic networks composed of a smaller number of genes than IRMA (e.g. Gardner et al. 2000; Elowitz and Leibler 2000; Tigges et al. 2009; Kramer et al. 2004; Stricker et al.

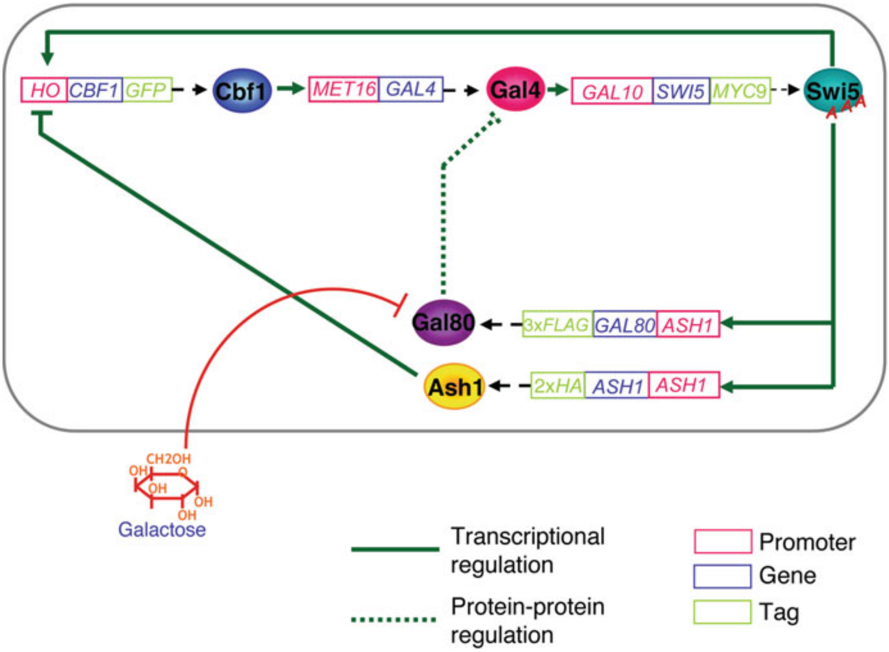

Fig. 1 Diagram of the network. Schematic diagram of the synthetic gene network. New transcriptional units (*rectangles*) were built by assembling promoters with non-self coding sequences. Genes were tagged at the 3′ end with the specified sequences. Each cassette encodes for a protein (*circle*) regulating the transcription of another gene in the network (*solid lines*)

2008). Regarding the identifiability issue, we adopted the novel approach proposed by Raue and colleagues (see Raue et al. 2009), able to deal with non-linear models with an high number of parameters. This approach exploits the profile likelihood and is able to detect both structural and practical non-identifiable parameters. For the parameters identification, in order to cope with the high number of unknown quantities, the noise of experimental data and the presence of non-linear aspects in the optimisation procedure, we used an ad hoc designed hybrid genetic algorithm (see Sect. 3 for further details). Finally, for the model validation, we tested the predictions of the model against data not used for the parameters identification.

2 Results and discussion

2.1 Derivation of model equations: step (i)

For each species in the network, i.e. each mRNA (italic capital letters) and correspondent protein concentration (roman small letters), we wrote one equation which expresses its change in time as the result of production and degradation:

$$\frac{d[\mathit{CBF1}]}{dt} = \alpha_1 + v_1 H^{+-}([\mathrm{Swi5}], [\mathrm{Ash1}]; k_1, k_2, h_1, h_2) - d_1[\mathit{CBF1}], \tag{1}$$

$$\frac{d[\mathrm{Cbf1}]}{dt} = \beta_1[\mathit{CBF1}] - d_2[\mathrm{Cbf1}], \tag{2}$$

$$\frac{d[\mathit{GAL4}]}{dt} = \alpha_2 + v_2 H^{+}([\mathrm{Cbf1}]; k_3, h_3) - d_3[\mathit{GAL4}], \tag{3}$$

$$\frac{d[\mathrm{Gal4}]}{dt} = \beta_2[\mathit{GAL4}] - d_4[\mathrm{Gal4}], \tag{4}$$

$$\frac{d[\mathit{SWI5}]}{dt} = \alpha_3 + v_3 H^{+}([\mathrm{Gal4}^{free}]; k_4, h_4) - d_5[\mathit{SWI5}], \tag{5}$$

$$\frac{d[\mathrm{Swi5}]}{dt} = \beta_3[\mathit{SWI5}] - d_6[\mathrm{Swi5}], \tag{6}$$

$$\frac{d[\mathit{GAL80}]}{dt} = \alpha_4 + v_4 H^{+}([\mathrm{Swi5}]; k_5, h_5) - d_7[\mathit{GAL80}], \tag{7}$$

$$\frac{d[\mathrm{Gal80}]}{dt} = \beta_4[\mathit{GAL80}] - d_8[\mathrm{Gal80}], \tag{8}$$

$$\frac{d[\mathit{ASH1}]}{dt} = \alpha_5 + v_5 H^{+}([\mathrm{Swi5}]; k_6, h_6) - d_9[\mathit{ASH1}], \tag{9}$$

$$\frac{d[\mathrm{Ash1}]}{dt} = \beta_5[\mathit{ASH1}] - d_{10}[\mathrm{Ash1}]. \tag{10}$$

The first two terms, on the right-hand side of the mRNA equations, represent the production, where α are the basal transcription rates; v are the maximal transcription rates modulated by the Hill functions, $H^{+}(y; k, h) = \frac{y^h}{y^h+k^h}$, $H^{-}(z; k, h) = \frac{k^h}{y^h+k^h}$

and $H^{+-} = H^{+}(y; k, h)(\cdot, +)H^{-}(z; k_1, h_1)$, modelling transcriptional activation, repression or a combination of the two, respectively; $(\cdot, +)$ indicates that we can either sum or multiply the Hill functions in the case of multiple regulation; y and z represent transcription factor levels, h are the Hill coefficients (pure numbers that refer to the cooperativity of the activation binding reaction) and k are the Michaelis-Menten constants, equal to the amount of transcription factor needed to reach half maximal activation (or repression). For protein equations, the production rates are β, i.e. the maximal translation rates. Degradations of mRNAs and proteins are represented by d, i.e. the degradation constants. Gal4^{free} in Eq. (5) depends on the interactions of the galactose pathway with the network genes.

In the model, the concentrations and the Michelis–Menten parameters k are reported in arbitrary units (a.u.), the basal activities α in (a.u. min^{-1}), the maximal transcription rates v in (a.u. min^{-1}), the translation rates β in (min^{-1}), the degradation constants d in (min^{-1}).

When writing the above model, we made the following assumptions: **[A1]** the *transcriptional activity* of each promoter is leaky (α); **[A2]** the degradation kinetics of both mRNAs and proteins are first-order; **[A3]** the protein production terms are proportional to the corresponding mRNA concentrations; **[A4]** the transcriptional *activation–repression* of each promoter by a transcription factor can be modelled as a Hill function (Kaern et al. 2003) and the *HO* promoter driving the expression of *CBF1* can be modelled either by adding the H^{+} and H^{-} functions (i.e.the promoter is activated by *SWI5 OR* repressed by *ASH1*), or by multiplying them (i.e. the promoter is activated by *SWI5 AND* repressed by *ASH1*) (Alon 2006; Mangan and Alon 2003). We chose between these two forms only during step (iii) of the modelling process, as described later.

In order to define the Gal4^{free} term in Eq. (5), we needed to describe the effect of the galactose pathway on the network dynamics. The biological mechanism is shown in Supplementary Fig. 1 (A). The concentration of Gal4^{free} is the amount of Gal4 protein that is not involved in the formation of the protein–protein complex with Gal80 and hence activates the *GAL10* promoter driving *SWI5* expression. In the literature, very detailed models of the galactose pathway have been presented (Bennett et al. 2008; Verma et al. 2004). We decided to simplify such paradigms and assumed (**[A5]**) that Gal80 directly binds to galactose ([GAL], the input of our model) in galactose growing condition, while Gal4 and Gal80 form the complex Gal4Gal80, when the yeast is grown in glucose [Supplementary Fig. 1 (B)]. Under this assumption, the simplified physical mechanism can be described by the mass balance laws:

$$[\text{Gal4}] = [\text{Gal4}^{free}] + [\text{Gal4Gal80}], \tag{11}$$

and

$$[\text{Gal80}] = [\text{Gal80}^{free}] + [\text{Gal4Gal80}] + [\text{GALGal80}], \tag{12}$$

where [Gal4Gal80] and [GALGal80] indicate the concentrations of the complexes. The rates of these complexes can be modelled assuming reversible reactions for them ($A + B \rightleftarrows AB$), i.e.

$$\frac{d[\mathrm{Gal4Gal80}]}{dt} = K_1[\mathrm{Gal4}^{free}][\mathrm{Gal80}^{free}] - K_2[\mathrm{Gal4Gal80}], \tag{13}$$

$$\frac{d[\mathrm{GALGal80}]}{dt} = K_3[\mathrm{GAL}][\mathrm{Gal80}^{free}] - K_4[\mathrm{GALGal80}], \tag{14}$$

with K being rate constants (K_1 is measured in [a.u.$^{-1}$min^{-1}], K_2 and K_4 in [min^{-1}], K_3 in [nM^{-1}min^{-1}] if the concentration of galactose, [GAL], is measured in [nM]).

The full model is described by Eqs. (1)–(10) together with Eqs. (13), (14) and consists of 12 equations and 41 parameters (Model A). If we assume that the time scale for the protein synthesis rate (including translocation and post-translational modifications) is much smaller than the time scale for the mRNA synthesis rate (Hatzimanikatis and Lee 1999), the protein concentrations are monotonically increasing functions of their corresponding mRNA concentrations at any time. Thus, by considering mRNA transcription and translation as a single step of synthesis for the five genes of the network (**[A6]**), Eqs. (2), (4), (6), (8) and (10) can be removed together with their associated 10 unknown parameters, leading to a *simplified non-linear model* (Model B) of IRMA (degradation constants renumbered and Hill functions in explicit form):

$$\begin{aligned}\frac{d[\mathit{CBF1}]}{dt} &= \alpha_1 + v_1\left(\frac{[\mathit{SWI5}]^{h_1}}{k_1^{h_1} + [\mathit{SWI5}]^{h_1}}\right)(\cdot, +)\left(\frac{k_2^{h_2}}{k_2^{h_2} + [\mathit{ASH1}]^{h_2}}\right)\\ &\quad - d_1[\mathit{CBF1}],\end{aligned} \tag{15}$$

$$\frac{d[[\mathit{GAL4}]}{dt} = \alpha_2 + v_2\left(\frac{[\mathit{CBF1}]^{h_3}}{k_3^{h_3} + [\mathit{CBF1}]^{h_3}}\right) - d_2[\mathit{GAL4}], \tag{16}$$

$$\frac{d[\mathit{SWI5}]}{dt} = \alpha_3 + v_3\left(\frac{([\mathit{GAL4}] - [\mathrm{Gal4Gal80}])^{h_4}}{k_4^{h_4} + ([\mathit{GAL4}] - [\mathrm{Gal4Gal80}])^{h_4}}\right) - d_3[\mathit{SWI5}], \tag{17}$$

$$\frac{d[\mathit{GAL80}]}{dt} = \alpha_4 + v_4\left(\frac{[\mathit{SWI5}]^{h_5}}{k_5^{h_5} + [\mathit{SWI5}]^{h_5}}\right) - d_4[\mathit{GAL80}], \tag{18}$$

$$\frac{d[\mathit{ASH1}]}{dt} = \alpha_5 + v_5\left(\frac{[\mathit{SWI5}]^{h_6}}{k_6^{h_6} + [\mathit{SWI5}]^{h_6}}\right) - d_5[\mathit{ASH1}], \tag{19}$$

$$\begin{aligned}\frac{d[\mathrm{Gal4Gal80}]}{dt} &= K_1([\mathit{GAL4}] - [\mathrm{Gal4Gal80}])([\mathit{GAL80}] - [\mathrm{Gal4Gal80}]\\ &\quad - [\mathrm{GALGal80}]) - K_2[\mathrm{Gal4Gal80}]\end{aligned} \tag{20}$$

$$\begin{aligned}\frac{d[\mathrm{GALGal80}]}{dt} &= K_3[\mathrm{GAL}]([\mathit{GAL80}] - [\mathrm{Gal4Gal80}] - [\mathrm{GALGal80}])\\ &\quad - K_4[\mathrm{GALGal80}],\end{aligned} \tag{21}$$

where $(\cdot, +)$ in (15) indicates that, according to assumption [A4], the multiple regulation of *CBF1* can be modelled either as an AND or an OR logic gate. Note the complexes equations (20), (21) were derived from the rate equations (13), (14) by substituting the expressions of $[Gal4^{free}]$ and $[Gal80^{free}]$ derived from the mass balance laws (11) and (12) under the assumption [A6]. In what follows, we will explore both possibilities, showing how comparison of the model predictions with the experimental data motivated the final choice. Equations (20) and (21) were obtained by substituting Eqs. (11) and (12) in Eqs. (13) and (14), under the assumption of proportionality between the protein levels of Gal4 and Gal80 and the corresponding mRNAs. Model B consists of seven differential equations [(15)–(21)] and contains 31 unknown parameters.

2.2 Identification of model parameters: step (ii)

In order to reduce the number of unknown parameters in Model B, we assumed that: **[A7]** all the promoters have null basal activity and unitary transcription rate, so that $\alpha = 0$ and $v = 1$; **[A8]** for the cooperativity coefficients h in the Hill functions we can consider only two options: set them all to 1 (*monomers approach*), or set all to 1 with the exception of h_3 and h_4, which are equal to 2 (*dimers approach*) in order to model the higher cooperativity of Cbf1 and Gal4, respectively, on the *MET16* promoter and the *GAL10* promoter (Hemmerich et al. 2000; Giniger and Ptashne 1988). Parameters K_1, K_2, K_3 and K_4 in Eqs. (20), (21) were fixed a priori from literature (Anders et al. 2006). Their values are reported in Table 1. The remaining 11 parameters were unknown and needed to be estimated from experimental data. To this end, we collected data of mRNAs expression levels during a time course experiment, by shifting cells from glucose to galactose ("switch-on" experiment) as described in Cantone et al. (2009) and in Sect. 3.

There are four versions of Model B due to assumptions [A4] (AND/OR regulation of the *HO* promoter driving *CBF1* expression) and [A8] (dimers versus monomers). We labeled the four different versions of Model B as B1 (AND/Monomers), B2 (OR/monomers), B3 (AND/dimers) and B4 (OR/dimers).

Identifiability analysis showed that all the 11 unknown parameters of Models B1, B2, B3 and B4 are structurally, but not practically, identifiable in the sense of Raue et al. (2009). Thus, the non-identifiability does not arise from incomplete observation of the internal model states or redundant parametrisation, but from the noisy nature and/or from the insufficient amount of experimental data. This makes the qualitative identification procedure we used the only viable option. We proceeded with the identification in order to evaluate the descriptive performance of the models and to start discriminating between different modelling possibilities (see Sect. 3 for detail about the identification procedure). This was done by comparing in silico and in vivo data for each of the four B models, using direct inspection and comparison of the corresponding values of the cost function J (Eq. (31) in Sect. 3).

The identified parameters are listed in Table 1. Results for Models B1 and B2 are shown in Fig. 2a and Supplementary Fig. 2 (A), respectively. Model B1 has a lower cost function ($J = 4.37$) than the value obtained with Model B2 ($J = 7.951$). Hence, modelling the multiple regulation of *CBF1* as a product (AND) seems to capture more accurately the dynamics of the *HO* promoter. Results for the Models B3 and B4 are shown in Supplementary Fig. 2 (B) and (C), respectively. Model B3 (AND/dimers) has a cost function $J = 2.83$, much lower than the other three models, and, thus, it was selected for the next step.

Table 1 Parameters of the mathematical models

Parameter	Model B1	Model B2	Model B3	Model B4	Model C	Model D	Model D ref.	Experim. id.
k_1 (a.u.)	0.329	9.637	1.757	10	1.884	1	1	1
k_2 (a.u.)	8.027	0.002	0.071	0.001	30	0.035	0.035	0.035
k_3 (a.u.)	3.387	0.711	0.886	0.240	0.229	0.037	0.037	0.037
k_4 (a.u.)	4.003	0.853	1.011	0.133	0.216	0.09 Glu 0.01 Gal	0.9 Glu 0.1 Gal	0.09 Glu 0.01 Gal
k_5 (a.u.)	1.750	1.972	7.375	1.313	0.16	1.884	1.884	1.884
k_6 (a.u.)	0.951	0.107	7.191	0.116	0.160	1.884	1.884	1.884
α_1 (a.u. min^{-1})	0	0	0	0	0	0	0	–
α_2 (a.u. min^{-1})	0	0	0	0	1.10×10^{-4}	1.49×10^{-4}	1.49×10^{-4}	–
α_3 (a.u. min^{-1})	0	0	0	0	3.2×10^{-4}	3×10^{-3}	3×10^{-3}	–
α_4 (a.u. min^{-1})	0	0	0	0	0	7.4×10^{-4}	7.4×10^{-4}	–
α_5 (a.u. min^{-1})	0	0	0	0	7.37×10^{-5}	6.1×10^{-4}	6.1×10^{-4}	–
v_1 (a.u. min^{-1})	1	1	1	1	0.065	0.04	0.04	–
v_2 (a.u. min^{-1})	1	1	1	1	0.002	8.82×10^{-4}	8.82×10^{-4}	–
v_3 (a.u. min^{-1})	1	1	1	1	0.025	0.002 Glu 0.020 Gal	0.017 Glu 0.155 Gal	v_3 Glu/v_3 Gal 9
v_4 (a.u. min^{-1})	1	1	1	1	0.007	0.014	0.014	–
v_5 (a.u. min^{-1})	1	1	1	1	0.002	0.018	0.018	–
v_{tr} (a.u. min^{1})	–	–	–	–	–	–	0.080	–
d_1 (min^{-1})	6.632	9.946	0.964	10	0.033	0.022	0.022	–
d_2 (min^{-1})	0.273	1.268	0.013	0.124	0.042	0.047	0.047	–
d_3 (min^{-1})	0.109	0.640	0.001	0.297	0.047	0.421	0.590	–
d_4 (min^{-1})	1.712	1.335	0.405	2.228	0.141	0.098	0.098	–
d_5 (min^{-1})	1.186	8.644	0.133	9.885	0.018	0.050	0.050	–
d_{pr} (min^{-1})	–	–	–	–	–	–	0.0144	–
h_1	1	1	1	1	1	1	1	1
h_2	1	1	1	1	1	1	1	1
h_3	1	1	2	2	1	1	1	1
h_4	1	1	2	2	1	4	4	4
h_5	1	1	1	1	1	1	1	1
h_6	1	1	1	1	1	1	1	1
h_7	–	–	–	–	1	4	4	4
K_1 (a.u.$^{-1}$ min^{-1})	100	100	100	100	–	–	–	–
K_2 (min^{-1})	1	1	1	1	–	–	–	–
K_3 (nM^{-1} min^{-1})	0.1	0.1	0.1	0.1	–	–	–	–
K_4 (min^{-1})	1	1	1	1	–	–	–	–
β_1 (min^{-1})	–	–	–	–	0.223	0.201	0.201	–
β_2 (min^{-1})	–	–	–	–	0.285	0.167	0.167	–
γ (a.u.)	–	–	–	–	10^{-4}Glu 5.55 Gal	0.2 Glu 0.6 Gal	0.2 Glu 0.6 Gal	0.2 Glu 0.6 Gal
τ (min)	–	–	–	–	100	100	100	–
GAL (nM)	5.55×10^7	5.55×10^7	5.55×10^7	5.55×10^7	–	–	–	–
J (cost function)	4.37	7.951	2.83	6.819	16.79	21.83	22	

2.3 Validation of model descriptive and predictive performance: step (iii)

In order to assess the predictive ability of Model B3, i.e. if the model is able to predict the behaviour of the network to new perturbations, we measured the gene expression

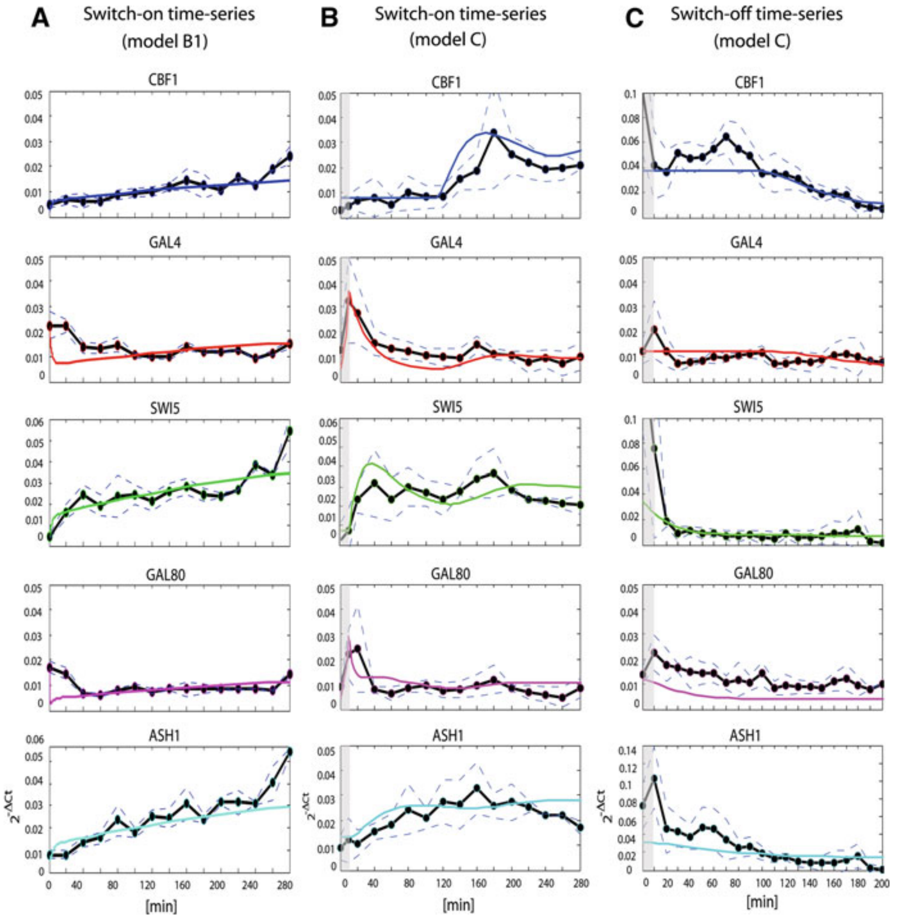

Fig. 2 Identification results on time-series data. *Circles* represent average expression data for each of the IRMA genes at different time points. *Dashed lines* represent standard errors. *Continuous lines* represent in silico data. **a** Identification results of Model B1 on the preliminary 5 h "switch-on" time-series (average of 4 time-series). **b** Identification results of Model C on the new 5 h "switch-on" data-set (average of 5 time-series). **c** Validation of Model C on the 3 h "switch-off" data-set (average of 4 time-series)

response of the five network genes following exogenous over-expression of each of the five genes under the control of a strong constitutive promoter, as described in Cantone et al. (2009) and in Sect. 3. Such over-expression experiments were performed both in glucose and in galactose. We will refer to these two experimental data-sets as the "Galactose steady-state" and "Glucose steady-state" (Fig. 3a, b).

Performing in silico the over-expression experiments (see Sect. 3), it is clear that Model B3, despite its good descriptive performance, has a very poor predictive power (Supplementary Fig. 3).

Therefore, we tested the predictive performance also for Model B1 (the second best as regards descriptive performance). Results are shown in Fig. 3c, d and in the Supplementary Excel File. Model B1 is able to partly describe and predict the network

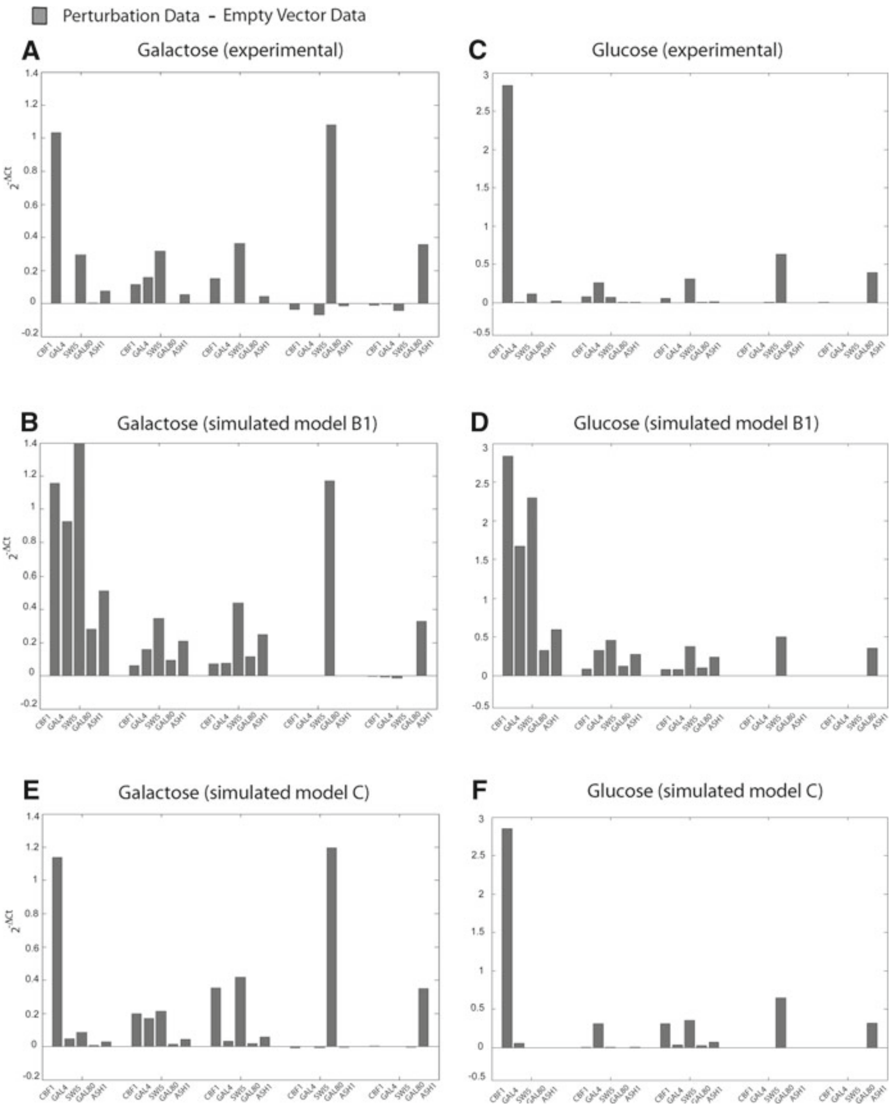

Fig. 3 Experimental and simulated over-expression experiments. **a**, **b** Difference between in vivo expression levels of IRMA genes after over-expression of each gene from the constitutive *GPD* promoter and levels after transformation of the empty vector. IRMA cells were transformed with each of the constructs containing one of the five genes or with the empty vector. At least three different colonies were grown in glucose (**b**) and in galactose–raffinose (**a**) up to the steady-state levels of gene expression. Quantitative PCR data are represented as $2^{-\Delta Ct}$ (average data from different colonies). **c**, **d** In silico data obtained by simulating the over-expression of each gene with Model B1. **e**, **f** In silico data obtained by simulating the over-expression of each gene with Model C

behaviour. There are still some major pitfalls: (a) two quantities (the concentrations of the two complexes) are present in the model, but cannot be measured experimentally. They were introduced by assuming a simplified mechanism for the interactions between the medium, Gal4 and Gal80 (assumption [A5], Supplementary Fig 1), but they are not physically consistent, and thus not measurable. (b) The "switch-on" data-set shows almost monotonic dynamics for the genes of IRMA, regardless of its complex topology. This data-set is an average of four independent experiments, three lasting 3 h, and just one lasting 5 h. Moreover, in such data, the early dynamic behaviour of the genes *GAL4* and *GAL80* is highly unexpected. We should observe an increase of all the mRNA concentrations following addition of galactose (switch-on), whereas *GAL4* and *GAL80* show a decrease during the initial 40 min, which Model B is unable to reproduce.

This modelling stage indicates that Model B has to be refined, and that new experiments are needed in order both to obtain a better characterisation of the dynamics of the synthetic network and to try to cope with the practical non-identifiability of the parameters.

2.4 Additional experimental investigation

We performed one additional 5 h "switch-on" time-series (see Sect. 3), this time including as the first point of the time-series the expression level of the network genes after growing cells overnight in glucose, just before shifting them from glucose to galactose (Cantone et al. 2009). The second point, taken after 10 min, is measured just after the shift has occurred and is equivalent to the first point of the previous time-series. The addition of this point to the data is fundamental to clarify the inconsistency in the early dynamics of *GAL4* and *GAL80*. The new averaged data-set (Fig. 2b) shows that the standard washing steps, needed to shift cells from glucose medium to the fresh new galactose-containing medium, induce a transient increase in mRNA levels of *GAL4* and *GAL80* (Fig. 2b, grey bars). This effect is not dependent on galactose addition, but uniquely on the washing steps (Cantone et al. 2009), and it is probably due to the transient deprivation of carbon source during washing, which attenuates the degradation levels of *GAL4* and *GAL80* mRNAs (Jona et al. 2000).

Also, in the new averaged data-set, the activation of *CBF1* appears to be delayed with respect to the other Swi5 targets, respectively, *GAL80* and *ASH1*. Such delay, not evident from the preliminary data-set, is physically due to the sequential recruitment of chromatin modifying complexes to the *HO* promoter, which follows binding of Swi5 (Bhoite et al. 2001; Cosma et al. 1999).

We performed four additional experiments, shifting cells from galactose to glucose, thus switching off gene expression in the network, as described in Cantone et al. (2009) and in Sect. 3. The averaged time-series data-set (Fig. 2c) was used for a further validation of the model predictive performance. We will refer to this data-set as the "switch-off" data-set.

2.5 Model refinement: Model C [step (i)]

At this stage, we had to properly refine the model both to be able to capture the new features highlighted by the new data-set and to remove unsuitable model complexity. First of all, we made the following extra modelling assumptions: **[A9]** a fix time delay,

τ, equal to 100 min, is added in the activation of the HO promoter by Swi5; **[A10]** a transient decrease in the mRNA degradation of *GAL4* and *GAL80* of value $\Delta\beta_1$ and $\Delta\beta_2$ ([min^{-1}]) is added for an interval of 10 min to describe the effect of the washing steps (Δ represents the transient duration of the washing effect).

Secondly, in order to remove from the model the unmeasured complexes concentrations describing the effects of galactose on the network, we considered two possible approaches: (1) to take the quasi steady-state approximation of the protein complexes dynamics (i.e. by setting the left-hand sides of (13) and (14) to 0); (2) to consider a new phenomenological non-linear function describing the effect of galactose. In the first case, steady-state approximation leads to the presence of an algebraic constraint (Supplementary information) thus turning the problem into a differential algebraic model with delays (DDAEs). This kind of problems are particularly cumbersome to solve and analyse from a mathematical viewpoint (see Kumar and Daoutidis 1999 for further details). To avoid this, we proceeded by finding a simple but effective phenomenological non-linear function to model the effect of the galactose pathway on the dynamics of *SWI5*, which is regulated by the *GAL10* promoter.

We assumed **[A11]** that the protein-protein interaction between Gal80 and Gal4 can be modelled as a direct inhibition of *GAL80* on the promoter of *SWI5*, and that the strength of such inhibition depends on the medium (strong inhibition in glucose, weak inhibition in galactose). Actually, we assumed that the *GAL10* promoter is activated by *GAL4* and non-competitively inhibited by *GAL80* (Copeland 2000).

The resulting phenomenological DDEs model (Model C), derived from Model B1, is:

$$\frac{d[CBF1]}{dt} = \alpha_1 + v_1\left(\frac{[SWI5(t-\tau)]^{h_1}}{k_1^{h_1} + [SWI5(t-\tau)]^{h_1}}\right) \cdot \left(\frac{k_2^{h_2}}{k_2^{h_2} + [ASH1]^{h_2}}\right) - d_1[CBF1], \tag{22}$$

$$\frac{d[GAL4]}{dt} = \alpha_2 + v_2\left(\frac{[CBF1]^{h_3}}{k_3^{h_3} + [CBF1]^{h_3}}\right) - (d_2 - \Delta\beta_1)[GAL4], \tag{23}$$

$$\frac{d[SWI5]}{dt} = \alpha_3 + v_3\left(\frac{[GAL4]^{h_4}}{(k_4^{h_4} + ([GAL4]^{h_4})(1 + \frac{[GAL80]^{h_7}}{\varphi^{h_7}})}\right) - d_3[SWI5], \tag{24}$$

$$\frac{d[GAL80]}{dt} = \alpha_4 + v_4\left(\frac{[SWI5]^{h_5}}{k_5^{h_5} + [SWI5]^{h_5}}\right) - (d_4 - \Delta\beta_2)[GAL80], \tag{25}$$

$$\frac{d[ASH1]}{dt} = \alpha_5 + v_5\left(\frac{[SWI5]^{h_6}}{k_6^{h_6} + [SWI5]^{h_6}}\right) - d_5[ASH1], \tag{26}$$

which consists of only five equations without any additional constraint.

The constant $\hat{\gamma}$ in (24) is the Michaelis–Menten coefficient of the phenomenological description of the inhibition of *GAL80*, which is assumed to be dependent on the medium (we use the symbol $\widehat{}$ to indicate medium-dependent quantities). This phenomenological DDEs model consists of five differential equations [(22)–(26)] and 31 unknown parameters.

2.6 Identification of the model parameters and validation of its descriptive and predictive performance: steps (ii) and (iii)

We set all of the Hill coefficients to 1 (monomers). For the identification of the remaining parameters, we used again the "switch-on" data-set, but this time using as initial values the simulated steady-state mRNA levels in glucose. The identifiability analysis showed that all the unknown parameters of Model C are again structurally identifiable, but not practically. Identification results are shown in Fig. 2b and the inferred parameters in Table 1. The model captures the delay in *CBF1* activation and the small variations of *GAL4* and *GAL80*.

In order to validate the model predictive performance, we used again the "Glucose steady-state" and "Galactose steady-state" over-expression experiments, and compared them with their in silico counterparts by simulating the over-expression of each of the five genes using Model C (Fig. 3e, f, Supplementary Excel File).

We further validated the predictive performance of the Model C against the "switch-off" time-series by simulating in silico the "switch-off" experiment (i.e. setting the medium-dependent parameters to their values in glucose and starting the simulation from the steady-state equilibrium in galactose) (Fig. 2c).

Model C has good descriptive and predictive performance. At this stage, it represents the best compromise between model complexity and performance given the experimental data-set. The model is indeed able to qualitatively predict network behaviour to new perturbations, thus achieving the aim we set for the modelling task. However, the 24 identified parameter values are likely to be different from their physical values. For example, model parameters (Table 1) indicate that the inhibition of Ash1 on *CBF1* is so weak that can be neglected, even if in the literature it has been reported otherwise (Cosma et al. 1999).

2.7 Experimental identification of the Hill function parameters

At this point, we needed to clarify the biological properties of the *HO* promoter by taking direct measurements of the promoters parameters. We thus measured the transcriptional response of the promoters of *GAL10*, *MET16*, *ASH1* and *HO*; the latter when regulated by both Swi5 and Ash1. For details see Cantone et al. (2009) and Sect. 3. Actually, we could have performed these experiments from the beginning, since the Hill functions were almost unchanged during the model refinement, with the exception of the *GAL10* and *HO* promoters modelling. However, since each experiment is costly and time consuming, we tried at each step to only perform those experiments that the mathematical modelling deemed indispensable. The need of performing promoter strength experiments arose after the identification of Model C since we did not trust the identified Hill functions parameters.

The model is now significantly improved, and the number of parameters that are practically not identifiable from the "switch-on" data-set can be significantly reduced.

For all of the promoters, we fitted the Hill function used in Model C. For each promoter, we fitted to 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 Eq. (22), thus obtaining:

$$[CBF1] = \frac{\alpha_1}{d_1} + \frac{v_1}{d_1}\left(\frac{[SWI5]^{h_1}}{(k_1^{h_1} + [SWI5]^{h_1}) \cdot \left(1 + \frac{[ASH1]^{h_2}}{k_2^{h_2}}\right)}\right). \tag{27}$$

For the fitting, the hybrid genetic algorithm was used (Sect. 3). In order to identify the phenomenological law for the *GAL10* promoter in Eq. (24), we fitted all the possible forms of the inhibition law (non-competitive,uncompetitive and competitive (Copeland 2000). Uncompetitive inhibition was found to give the best fitting (data not shown). Finally, it became apparent from the new experimental data and the results of the fitting, that galactose not only weakens the inhibition of Gal80 on the *GAL10* promoter (assumption [A11] in Model C), but also allows a faster activation of the *GAL10* promoter. Moreover, in galactose such activation is possible for values of *GAL4* lower than in glucose.

The kinetic parameters that were physically estimated are given in Table 1, while the data and the relative fitting in Fig. S4 and S5 in the Supplemental Data of Cantone et al. (2009).

2.8 Further model refinement: Model D [back to step (i)]

To model the effect of galactose and, in particular, the behaviour of the *GAL10* promoter, Model C needed to be further refined. In particular, since galactose was found to affect all of the parameters describing the *GAL10* promoter activity, we considered two additional parameters in the model to be explicitly dependent on the medium.

Thus, we derived a new model (Model D) consisting of the Eqs. (22), (23), (25), (26) of Model C and of the following equation for *SWI5*:

$$\frac{d[SWI5]}{dt} = \alpha_3 + \hat{v}_3\left(\frac{[\mathrm{GAL4}]^{h_4}}{(\hat{k}_4{}^{h_4} + ([\mathrm{GAL4}]^{h_4})(1 + \frac{[GAL80]^{h_7}}{\hat{\gamma}^{h_7}})}\right) - d_3[SWI5]. \tag{28}$$

where the symbol $\widehat{\ }$ indicates parameters dependent on the medium. From the analysis of data, we found that the value assumed by $\widehat{v_3}$ in galactose is 9 times bigger than the one in glucose. Analogously, the value of $\widehat{k_4}$ is 9 times bigger in glucose than in galactose (see Table 1).

2.9 Identification of parameters and validation of model D: step (ii) and (iii)

The refined DDEs model (Eqs. (22), (23), (25), (26), (28)) contains 33 unknown parameters. From the promoter data-set, we estimated 16 parameters, including the medium-dependent ones (Table 1). From such data, we could not fit degradation constants, nor the washing effect parameters ($\Delta\beta_1$ and $\Delta\beta_2$). Thus, the remaining 17 parameters were evaluated from the "switch-on" experiment (Table 1). In simulations, the initial values of mRNA concentrations were set to the steady state values predicted by the model in glucose. The in silico "switch-on" time-series is shown in Fig. 2 of Cantone et al. (2009) and in Supplementary Fig. 4. Also in this case, we tested the predictive ability of the model performing in silico the previously described "Glucose steady-state" and "Galactose steady-state" over-expression experiments and the "switch-off" time-series (Fig. 4c, d, Fig. S3 in Cantone et al. (2009), Supplementary Excel File). By comparing data and simulations, it appears that Model D is quite similar to Model C, the only difference being that, this time, some of its physical parameter values have been directly measured. Now, Model D parameters confirm that the Ash1 inhibition of the *HO* promoter is indeed strong, as reported in the literature (Cosma et al. 1999).

There are still discrepancies between the in vivo and in silico initial values of *CBF1*, *SWI5* and *ASH1* in the "switch-off" data-set, and in the predicted steady state of mRNA levels in galactose. We attribute them to the unmodelled effect of protein dynamics, which have been removed from the original model due to the lack of experimental measurements. In particular, we noticed that the Gal4 protein is stable (Muratani et al. 2005), 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 do 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. In order to verify this thesis, we modified Model D by additionally modelling the protein level of Gal4. Thus, in the model we added the following equation for Gal4 protein (which is assumed to be linearly dependent on *GAL4* mRNA):

$$\frac{d[\mathrm{Gal4}]}{dt} = v_{tr}[GAL4] - d_{pr}[\mathrm{Gal4}]. \tag{29}$$

As a consequence, a new variable in the activation law of Swi5 has been inserted:

$$\frac{d[SWI5]}{dt} = \alpha_3 + \hat{v_3}\left(\frac{[\mathrm{Gal4}]^{h_4}}{(\hat{k_4}^{h_4} + ([\mathrm{Gal4}]^{h_4})(1 + \frac{[GAL80]^{h_7}}{\hat{\gamma}^{h_7}})}\right) - d_3[SWI5]. \tag{30}$$

We fitted the parameters in Eq. (29) from the "switch-on" data-set (Table 1). In particular, the estimated degradation rate of Gal4 protein is lower than all the other degradation rates, in accordance with the experimental results in Muratani et al. (2005). Consequently, we slightly modified two parameters of the *GAL10* promoter (see Table 1). Note that such parameters were previously estimated from the promoter data-set, but in such experiments we measured the levels of the *GAL10* promoter depending on the mRNA and not on the protein level of Gal4. The in silico "switch-on" and "switch-off" time-series look almost identical to the simulations of Model D (data not shown), but the quality of the predictions of the "Glucose steady-state" and "Galactose steady-state" over-expressions is significantly improved (see Fig. 4e, f, Supplementary Excel File). In particular, the increase in *SWI5* expression, due to the accumulation of Gal4 protein, is captured (e.g. Fig. 4e, over-expression of *CBF1* and *GAL4*).

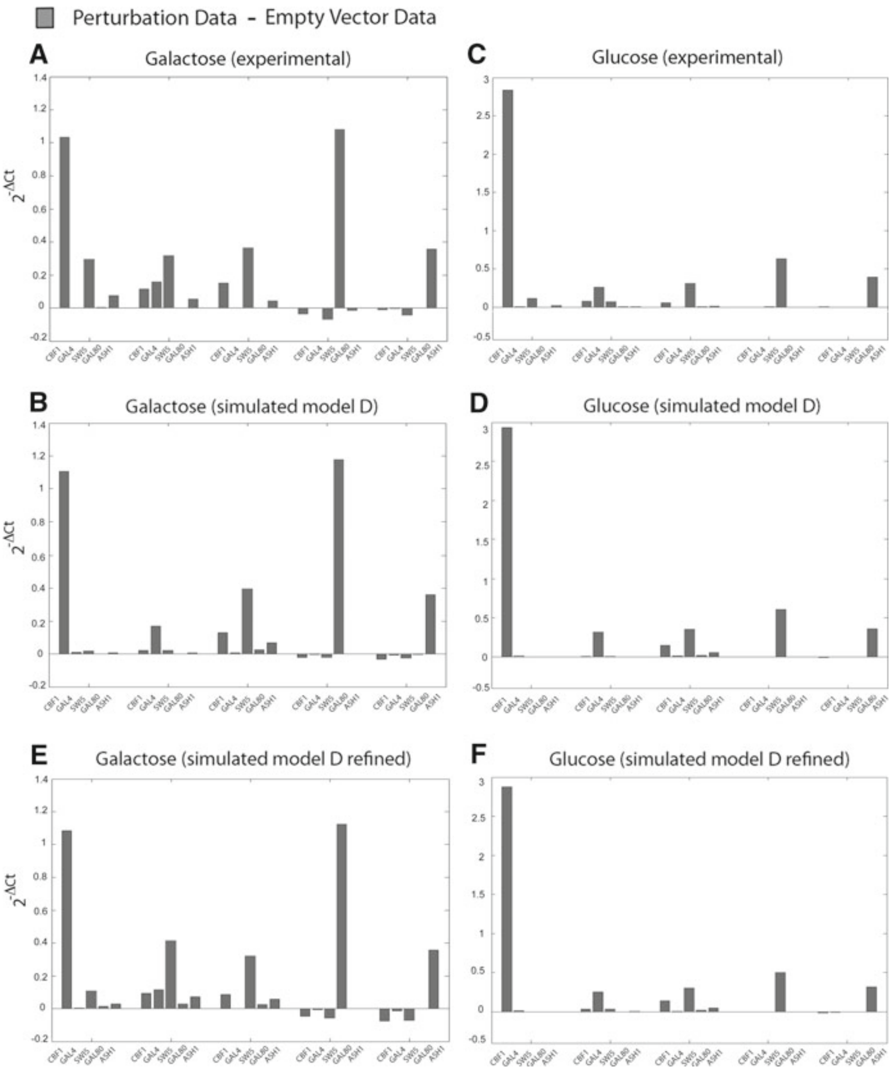

Fig. 4 Experimental and simulated over-expression experiments. **a**, **b** Difference between in vivo expression levels of IRMA genes after over-expression of each gene from the constitutive *GPD* promoter and levels after transformation of the empty vector. IRMA cells were transformed with each of the constructs containing one of the five genes or with the empty vector. At least three different colonies were grown in glucose (**b**) and in galactose–raffinose (**a**) up to the steady-state levels of gene expression. Quantitative PCR data are represented as $2^{-\Delta Ct}$ (average data from different colonies). **c**, **d** In silico data obtained by simulating the over-expression of each gene with Model D. **e**, **f** In silico data obtained by simulating the over-expression of each gene with Model D refined

2.10 Discussion

We described in detail the steps required to build a mathematical model of a synthetic biological pathway. This framework can be applied equally well to naturally occurring networks in the cell, thus transforming the drawing of a biological pathway into a computational model. Such a model can then be easily probed in silico and its predictions checked against experimental data in order to validate the correctness of biological hypotheses. When inconsistencies between modelling and experiments arise, this is a clue that something important is missing in our drawing of the biological pathway. We can identify this missing link by appropriately modifying the computational model using our biological knowledge, until a better agreement between simulated and experimental data is achieved.

In our example, modelling pointed to an inconsistency between the in silico and in vivo behaviour of *GAL4* and *GAL80* during the glucose-to-galactose shift ("switch-on"); their decrease in concentration could not be captured by the model, which was simply based on the drawing in Fig. 1, i.e. on the known biological function of the promoters and proteins in the network. This hinted to the possibility of an unmodelled effect and prompted further experimental investigation of what this could be. We discovered that cell manipulation during the washing steps (needed to perform the medium shift) induced a transient increase in *GAL4* and *GAL80*.

Modelling can also suggest that additional experimental investigation is needed. In particular, we had to face the issue of practical identifiability for the model parameters. Biological systems, as well as economical ones, often suffer from this problem due to the intrinsic experimental noise (Ljung 1998). However, when it is possible, extra-experiments can be performed in order to reduce the number of practicalnon-identifiable parameters. In our case, we enlarged the available data-set by performing the promoter strength experiments. To further address the issue of obtaining better experimental measurements, we are currently setting up a novel experimental platform based microfluidics (Bennett et al. 2008).

The whole modelling procedure is schematically described in Fig. 5. During the modelling process, the modeller needs to simplify some aspects of the model and to increase the level of details of others, always taking into account the amount and quality of experimental data. For example, we showed that adding an equation for Gal4 protein improves the predictive power of the model. The quality of the fitting and the predictions could be further improved by modelling the proteins levels of all the genes in the network. However, in the actual version of the network, it is not possible to measure protein levels with the exception of only one gene (Cbf1). Thus, the assumption of steady state for protein dynamics is required, not only in order to simplify the model, but mainly to do not introduce the problem of over-fitting and non-uniqueness of parameters for proteins. In order to decide what can be simplified, and what needs to be modelled in more details, it is necessary to go through iterative refinement steps both in the model and in the experimental data-set.

3 Materials and methods

3.1 Construction of IRMA and experimental data-set

The promoters of the network, chosen in such a way that for each of them a single transcription factor (TF) is sufficient and essential to activate transcription, were assembled upstream of non-self gene coding sequences. Further details can be found in Cantone et al. (2009). All data presented refer to mRNA levels of the five IRMA genes and were measured by quantitative real-time RT-PCR (q-PCR).

For the preliminary "switch-on" data-set (used for the identification of Models B1, B2, B3 and B4), we collected samples every 20 min up to 5 h in four independent experiments and we measured mRNA levels of the five IRMA genes by quantitative real-time RT-PCR (q-PCR). Out of the four time-series, three were 3 h long, and one lasted 5 h. The averaged data-set is shown in Fig. 2a. We then performed one additional 5 h "switch-on" time-series. The experimental set up is identical, but we included as the first point of the time-series the expression level of the network genes after growing cells overnight in glucose. The new averaged "switch-on" data-set was used for the identification of Models C and D and is shown in Fig. 2b.

The "switch-off" data-set (Fig. 2c) is the average of four experiments performed by shifting cells from galactose to glucose and collecting samples every 10 min up to 3 h (Cantone et al. 2009).

For the "Galactose steady-state" and "Glucose steady-state" data-set (Fig. 3a, b), the over-expression of each gene was performed in cells grown either in glucose, or in galactose. The steady-state expression levels of IRMA genes were measured by q-PCR.

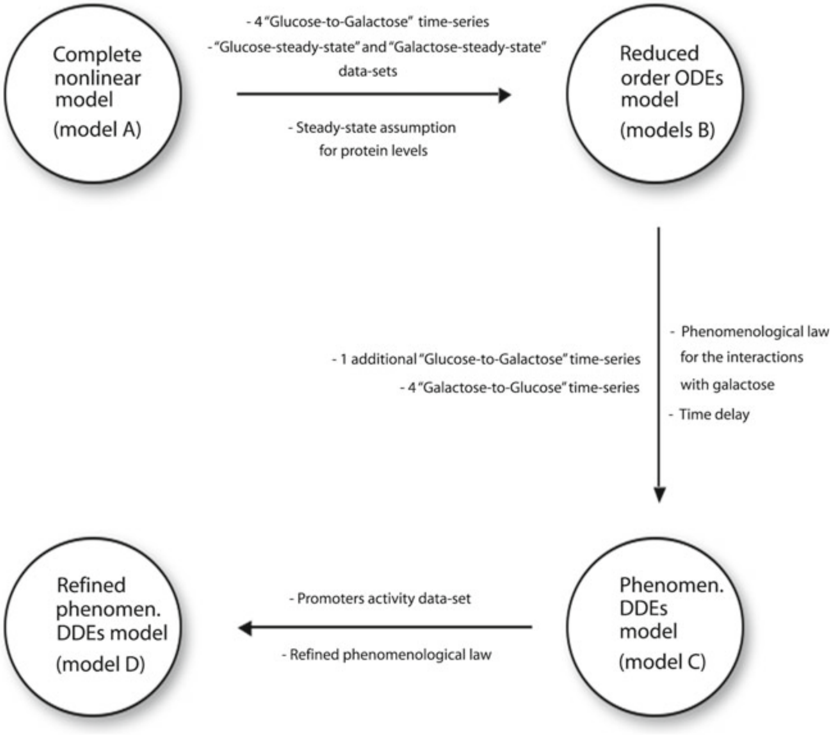
Complete nonlinear model (model A)
- 4 "Glucose-to-Galactose" time-series
- "Glucose-steady-state" and "Galactose-steady-state" data-sets
- Steady-state assumption for protein levels
Reduced order ODEs model (models B)
- Phenomenological law for the interactions with galactose
- Time delay
- 1 additional "Glucose-to-Galactose" time-series
- 4 "Galactose-to-Glucose" time-series
Phenomen. DDEs model (model C)
- Promoters activity data-set
- Refined phenomenological law
Refined phenomen. DDEs model (model D)

Fig. 5 Scheme of the whole modelling and experimental procedure. Schematic representation of the steps performed in the refinement of the mathematical model of IRMA and in the set-up of the experiments

For the promoters data-set, each of the transcription factor genes was stably expressed at different levels in a wild-type strain, and the transcription of the corresponding promoter genes were measured by q-PCR at steady-state, for a total of 165 data points (Cantone et al. 2009).

3.2 Identification of model parameters

The problem of estimating the unknown parameters can be formulated as a minimization problem, where the in silico predictions of the model are compared to the experimental data and the parameter values are adjusted in order to minimise the disagreement between the two. To quantify the disagreement between in silico and in vivo data, we computed the cost function J_θ as

$$J_\theta = \sum_{i=1}^{n} \sum_{j=1}^{m} \left(\frac{y^i_{calc}(j;\theta) - y^i_{\exp}(j)}{y^i_{\exp}(j)} \right)^2, \tag{31}$$

where θ are the parameters to be identified, n is the number of genes, m is the number of experimental data points, $y^i_{calc}(j)$ is the in silico mRNA concentration of gene i at time j, $y^i_{exp}(j)$ is the experimental mRNA concentration for the same gene. A variety of approaches can be used to find parameters that minimise J. Genetic Algorithms (GA) were indeed found to provide a flexible approach to non-linear optimization (Mitchell 1998) and described to be particularly effective in the parametrisation of synthetic networks (Weber et al. 2007; Tigges et al. 2009). Here we use the hybrid genetic algorithm (HGA) described in Cantone et al. (2009), using the above cost function to fit time-series data.

For the estimation of the parameters of the Hill functions from the promoter strength data-set, the objective function to be minimized by the HGA was defined as:

$$J = \sum_{i=1}^{n} \left(\frac{y^i_{calc} - y^i_{\exp}}{y^i_{\exp}} \right)^2. \tag{32}$$

3.3 Simulation settings

Simulated data of time-series were obtained by numerically solving the model equations. In the simulations of Models B1, B2, B3 and B4 we set as initial values for all mRNA concentrations the first available experimental data point. The initial value of the complex [Gal4Gal80] was derived using Eq. (20). The initial condition for the complex [GALGal80] was instead set to 0 (i.e. no galactose when the experiment starts). In the simulations of Models C and D we used as initial values the simulated steady-state mRNAs levels (respectively in glucose for the switch-on time-series and in galactose for the "switch-off").

Simulated data of over-expression experiments were performed by adding a constant production term to the equations describing the mRNA dynamics of the gene being perturbed, thus simulating constitutive expression from the strong promoter.

Numerical simulations were run using Matlab 2008b (The MathWorks). In the case of Models B1, B2, B3 and B4 (ODE models), we used *ode23* solver [a detailed discussion of the numerical methods used by *ode23* can be found in Bogacki and Shampine (1989)]. For Models C, D and D refined (DDEs models), we adopted the *dde23* solver, which solves delay differential equations with constant delays [a detailed discussion of the numerical methods used can be found in Shampine and Thompson (2001)]. For the identifiability analysis, we followed the approach proposed in Raue et al. (2009) and we used the PottersWheel fitting toolbox (Maiwald and Timmer 2008).

Acknowledgments This work was supported by EU-STREP FP6 grant (COBIOS) to D.d.B. and M.d.B. We wish to thank Andreas Raue, Thomas Maiwald and Filippo Menolascina for helpful discussion about the identifiability issue. We thank Oscar Giovanni Sassner, Giulia Cuccato and Girolamo Valentinetti for general support.

Conflict of interest statement The authors declare that they have no competing financial interests.

References

Alon U (2006) An Introduction to systems biology: design principles of biological circuits. Chapman & Hall, London

Anders A, Lilie H, Franke K, Kapp L, Stelling J, Gilles ED, Breunig KD (2006) The galactose switch in kluyveromyces lactis depends on nuclear competition between gal4 and gal1 for gal80 binding. J Biol Chem 281(39):29337–29348

Anderson J, Papachristodoulou A (2009) On validation and invalidation of biological models. BMC Bioinform 10:132

Arlot S, Celisse A (2010) A survey of cross-validation procedures for model selection. Stat Surv 4:40–79

Bennett MR, Pang WL, Ostroff NA, Baumgartner BL, Nayak S, Tsimring LS, Hasty J (2008) Metabolic gene regulation in a dynamically changing environment. Nature 454:1119–1122

Bhoite LT, Yu Y, Stillman DJ (2001) The swi5 activator recruits the mediator complex to the ho promoter without rna polymerase II. Genes Dev 15:2457–2469

Bogacki P, Shampine LF (1989) A 3(2) pair of Runge–Kutta formulas. Appl Numer Math 2:1–9

Boubaker O, Fourati A (2004) Structural identifiability of non linear systems: an overview. Ind Technol 3:1224–1248

Cantone I, Marucci L, Iorio F, Ricci MA, Belcastro V, Bansal M, di Bernardo M, Santini S, di Bernardo D, Cosma MP (2009) A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. Cell 137:171–181

Cobelli C, Distefano JJ III (1980) Parameter and structural identifiability concepts and ambiguities: a critical review and analysis. Am J Physiol 239:R7–R24

Copeland R (2000) Enzymes: a practical introduction to structure, mechanism, and data analysis, 2nd edn. Wiley, New York

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

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

Elowitz MB, Leibler S (2000) A synthetic oscillatory network of transcriptional regulators. Nature 403(6767):335–338

Gardner TS, Cantor CR, Collins JJ (2000) Construction of a genetic toggle switch in Escherichia coli. Nature 403(6767):339–342

Giniger E, Ptashne M (1988) Cooperative dna binding of the yeast transcriptional activator gal4. Proc Natl Acad Sci 85:382–386

Hatzimanikatis V, Lee KH (1999) Dynamical analysis of gene networks requires both mRNA and protein expression information. Metabol Eng 1(4):275–281

Hemmerich P, Stoyan T, Wieland G, Koch M, Lechner J, Diekmann S (2000) Interaction of yeast kinetochore proteins with centromere-proteinytranscription factor cbf1. Proc Natl Acad Sci USA 97(23):12583–12588

Jona G, Choder M, Gileadi O (2000) Glucose starvation induces a drastic reduction in the rates of both transcription and degradation of mrna in yeast. Biochim Biophy Acta 1491:37–48

Kaern M, Blake WJ, Collins JJ (2003) The engineering of gene regulatory networks. Annu Rev Biomed Eng 5:179–206

Kaznessis Y (2007) Models for synthetic biology. BMC Syst Biol 1(1):47

Kramer BP, Viretta AU, Daoud-El-Baba M, Aubel D, Weber W, Fussenegger M (2004) An engineered epigenetic transgene switch in mammalian cells. Nat Biotechnol 22(7):867–870

Kumar A, Daoutidis P (1999) Control of nonlinear differential algebraic equation systems: with application to chemical processes. CRC Press, West Palm Beach

Ljung L (1998) System identification: theory for the user, 2nd edn. Prentice-Hall, Englewood Cliffs

Maiwald T, Timmer J (2008) Dynamical modeling and multi-experiment fitting with potterswheel. Bioinformatics 24(18):2037–2043

Mangan S, Alon U (2003) Structure and function of the feed-forward loop network motif. Proc Natl Acad Sci USA 100(21):11980–11985

Mitchell M (1998) An introduction to genetic algorithms (complex adaptive systems). MIT Press, Cambridge

Moles CG, Mendes P, Banga JR (2003) Parameter estimation in biochemical pathways: a comparison of global optimization methods. Genome Res 13(11):2467–2474

Muratani M, Kung C, Shokat KM, Tansey WP (2005) The f box protein dsg1/mdm30 is a transcriptional coactivator that stimulates gal4 turnover and cotranscriptional mrna processing. Cell 120:887–899

Nelles O (2000) Nonlinear system identification: from classical approaches to neural networks and fuzzy models, 1st edn. Springer, Berlin

Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, Klingmuller U, Timmer J (2009) Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics 25(15):1923–1929

Ro D-K, Paradise EM, Ouellet M, Fisher KJ, Newman KL, Ndungu JM, Ho KA, Eachus RA, Ham TS, Kirby J, Chang MCY, Withers ST, Shiba Y, Sarpong R, Keasling JD (2006) Production of the antimalarial drug precursor artemisinic acid in engineered yeast. Nature 440(7086):940–943

Shampine LF, Thompson S (2001) Solving ddes in matlab. Appl Numer Math 37:441–458

Smith RS, Doyle JC (1992) Model validation: a connection between robust control and identification. IEEE Trans Automat Contr 37:942–952

Stricker J, Cookson S, Bennett MRR, Mather WHH, Tsimring LSS, Hasty J (2008) A fast, robust and tunable synthetic gene oscillator. Nature 456(7221):516–519

Szallasi Z, Stelling J, Periwal V (2006) System modeling in cellular biology: from concepts to nuts and bolts. MIT Press, Cambridge

Tigges M, Marquez-Lago TT, Stelling J, Fussenegger M (2009) A tunable synthetic mammalian oscillator. Nature 457(7227):309–312

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

Verma M, Bhat JP, Venkatesh KV (2004) Quantitative analysis of gal genetic switch of *Saccharomyces cerevisiae* reveals that nucleocytoplasmic shuttling of gal80p results in a highly sensitive response to galactose. J Biol Chem 278:48764–48769

Walter E, Pronzato L (1997) Identification of parametric models from experimental data. Springer, Berlin

Weber W, Stelling J, Rimann M, Keller B, Baba MDE, Weber CC, Aubel D, Fussenegger M (2007) A synthetic time-delay circuit in mammalian cells and mice. Proc Natl Acad Sci USA 104(8):2643–2648