Molecular Systems Biology 5; Article number 287; doi:10.1038/msb.2009.45
Citation: ***Molecular Systems Biology*** **5:287**

www.molecularsystemsbiology.com

molecular systems biology

Elucidating regulatory mechanisms downstream of a signaling pathway using informative experiments

Ewa Szczurek[1,2,3,*], Irit Gat-Viks[1,4], Jerzy Tiuryn[3] and Martin Vingron[1]

[1] Computational Molecular Biology Department, Max Planck Institute for Molecular Genetics, Berlin, Germany, [2] International Max Planck Research School for Computational Biology and Scientific Computing, Berlin, Germany and [3] Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland
[4] Present address: Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA 02142, USA
* Corresponding author. Computational Molecular Biology Department, Max Planck Institute for Molecular Genetics, Ihnestr. 73, 14195 Berlin, Germany.
Tel.: +49 30 8413 1261; Fax: +49 30 8413 1152; E-mail: szczurek@molgen.mpg.de

Received 15.8.08; accepted 26.5.09

Signaling cascades are triggered by environmental stimulation and propagate the signal to regulate transcription. Systematic reconstruction of the underlying regulatory mechanisms requires pathway-targeted, informative experimental data. However, practical experimental design approaches are still in their infancy. Here, we propose a framework that iterates design of experiments and identification of regulatory relationships downstream of a given pathway. The experimental design component, called MEED, aims to minimize the amount of laboratory effort required in this process. To avoid ambiguity in the identification of regulatory relationships, the choice of experiments maximizes diversity between expression profiles of genes regulated through different mechanisms. The framework takes advantage of expert knowledge about the pathways under study, formalized in a predictive logical model. By considering model-predicted dependencies between experiments, MEED is able to suggest a whole set of experiments that can be carried out simultaneously. Our framework was applied to investigate interconnected signaling pathways in yeast. In comparison with other approaches, MEED suggested the most informative experiments for unambiguous identification of transcriptional regulation in this system.
Molecular Systems Biology **5:** 287; published online 7 July 2009; doi:10.1038/msb.2009.45
Subject Categories: metabolic and regulatory networks; signal transduction
Keywords: experimental design; logical modeling; signal transduction; transcription regulation

Introduction

Revealing the mechanism of transcription regulation in the cell, the interplay of transcription factors and the way they influence their target genes, is a central problem in molecular biology. Diverse approaches have been proposed for the identification of transcriptional regulation based on high-throughput gene expression data (e.g. Akutsu *et al*, 1998; Bussemaker *et al*, 2001; Bolouri and Davidson, 2002; Gardner *et al*, 2003; Segal *et al*, 2003; Nachman *et al*, 2004; Hartemink, 2005). All these methods heavily depend on the available experiments and are prone to the problem of ambiguity in the identification of regulatory relationships. For example, it is possible that a transcription factor remains inactive in all experiments and therefore its targets cannot be revealed. Alternatively, consider two transcription factors located in distinct signaling pathways with a different role, different environmental stimulation and different target genes. In a given set of experiments, if the target genes have similar expression profiles, they will be falsely considered as co-regulated. Moreover, taking any of the two transcription factors as the common regulator of these targets will be equally supported by the experimental data, leading to ambiguous hypothesis about their transcriptional regulation. To avoid such problems, the experiments must generate enough information to draw clear conclusions about regulatory relationships.

In this study, we introduce an algorithm called MEED (model expansion experimental design). MEED is meant to guide experimentalists who focus their research on a chosen signaling pathway and are interested in the regulation of its downstream targets. We assume the researcher has initial qualitative knowledge about the studied pathway and wishes to systematically perturb the pathway components to characterize the gene expression response. Such experimental studies, in which a specific signaling system is perturbed to investigate its downstream regulation mechanisms (rather than global mapping of cellular transcription), became

common in the recent years (e.g., Roberts *et al*, 2000; Yoshimoto *et al*, 2002; e.g., O'Rourke and Herskowitz, 2004). In our framework, the researcher's expertise about the signaling network should be formalized in a predictive logical model and provided as input to MEED. The model represents biological components, such as signal transduction molecules, environmental stimulations and transcription factors, as well as the (possibly cyclic) logical relations among them (Gat-Viks *et al*, 2004). Given this input model, the algorithm aims to select the least number of experiments, which together allow for unambiguous identification of target genes and the way they are regulated by components in the model. To this end, MEED relies only on model predictions and does not use additional data. In particular, no initial high-throughput transcription factor–DNA binding data are required. Our algorithm instructs the researcher under which environmental conditions and with what eventual perturbations the experiments should be conducted. The suggested experiments are ordered such that the researcher can choose to carry out only the most informative ones from the list.

Experimental design (ED) in systems biology was previously applied for the reconstruction and refinement of *in-silico* models (Ideker *et al*, 2000; Tong and Koller, 2001; King *et al*, 2004; Yeang *et al*, 2005; Barrett and Palsson, 2006; Vatcheva *et al*, 2006). The extant methods evaluate and assign a score to each experiment independently. If the experimenter aims to carry out several experiments simultaneously, the independence assumption becomes critical: a set of highest scoring experiments might provide redundant information about the system under study. In other words, some of the highest scoring experiments might be dispensable given other highest scoring experiments. Therefore, all extant methods can design efficiently only one (the highest scoring) experiment. The next experiment can be designed only after the suggested experiment has been carried out in a lab and the measurements were processed. MEED, in contrast, scores a set of experiments together and considers potential dependencies between them. In this way, our algorithm is able to design a set of informative, non-redundant experiments that can be carried out in parallel.

We propose a general framework, in which the experiments designed using MEED are used in a *model expansion* procedure. Building on Gat-Viks and Shamir (2007), the procedure reconciles experimental data with model predictions to elucidate regulatory mechanisms downstream of a given pathway model. Model expansion identifies target genes together with their regulators in the model and the logic of regulation. We utilized our framework to investigate regulatory relationships downstream of the interconnected osmotic stress and pheromone pathways in *Saccharomyces cerevisiae*. Using experiments chosen by MEED from available experimental studies, we applied the expansion procedure and identified regulatory modules comprising groups of genes co-regulated by molecules in the pathway through a specific regulatory mechanism. We iterated experimental design to propose additional experiments for resolving the ambiguity remained after the expansion step. In comparison with other approaches, the experiments suggested by MEED make it possible to draw less ambiguous conclusions about transcriptional regulation. Moreover, our comparative analysis shows the importance of considering dependencies between experiments as part of the ED process.

Results

The proposed framework consists of three components: modeling of the studied signaling network, an experimental design algorithm MEED and an expansion procedure. The framework aims to discover transcriptional control downstream of the given signaling pathway using an optimal set of experiments. Software implementing our framework is available on our website: http://meed.molgen.mpg.de/.

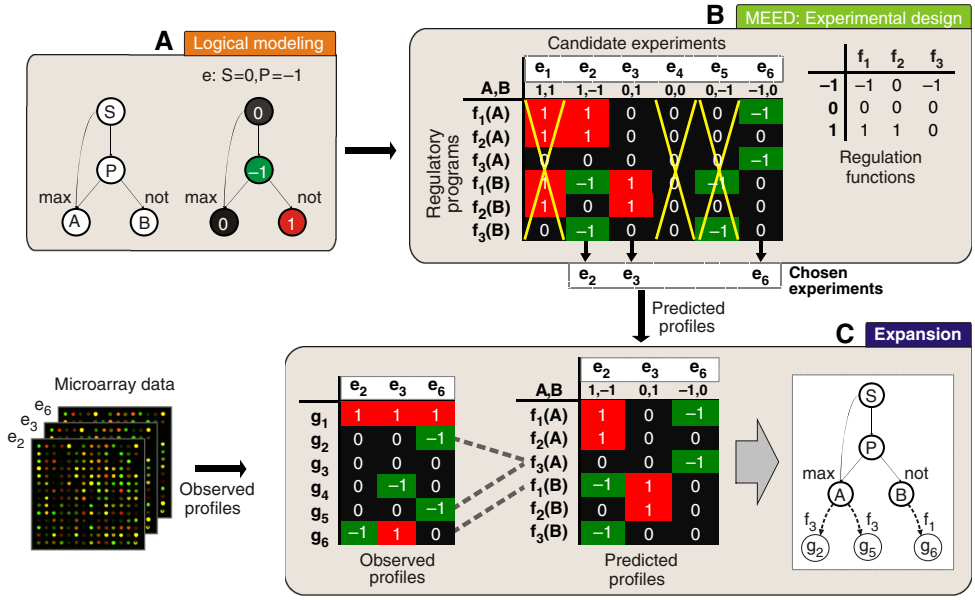
The first component, the model, formalizes prior knowledge about regulatory relations between signaling molecules. The model is predictive: For a given experiment (i.e., extracellular stimulation and genetic perturbation), the model predicts the activation states of the regulators in the network. In this study, we assume that regulatory relations are discrete logical functions and that the model describes the steady state of the system after exposure to the experiment. In addition, we predefine a repertoire of logical functions that formalize transcriptional regulation mechanisms, such as activation or repression (only regulation by a single molecule is considered). By applying all predefined logical functions to the model-predicted state of a given regulator under a given experiment, we obtain predictions about all possible readouts of the regulator's target genes. This is done for all regulators and all candidate experiments. In this way, we calculate predicted expression profiles for all putative targets of the regulators.

The MEED algorithm aims to select from the set of candidate experiments optimizing two objectives: (i) to minimize the number of selected experiments and (ii) to maximize diversity between the predicted expression profiles. The second condition aims to avoid an ambiguous situation in which two genes with distinct regulatory mechanisms attain the same expression profile under the suggested experiments. Only in the case in which the two genes have two distinct expression profiles, it is possible to distinguish their regulatory mechanisms. Next, the chosen experiments should be carried out in a lab and used to identify regulator–target relationships. To this end, the expansion procedure matches the model-predicted expression profiles of putative targets for the set of experiments selected by MEED with real expression measurements observed under the same experiments. The introduced framework is described in detail below.

Outline of our framework

We apply a model-based approach to design a set of experiments that can be used for unambiguous identification of regulatory relationships downstream of a given signaling pathway (Box 1). First, we formalize the available information on the pathway in a logical model with discrete variables (Gat-Viks *et al*, 2004). The environmental signals, which trigger the signaling cascade, are represented as variables called *stimulators*. Remaining variables correspond to the signaling molecules and can be in three possible *states*: 1 (*activated*), −1 (*deactivated*) or 0 (*neutral*). Variables representing proteins having transcriptional control over response of target

Box 1 Experimental design for model expansion



(**A**) Logical modeling. Left: a toy model. S—a stimulator variable representing the environmental signal, P—a variable representing a signaling molecule. A, B—regulators representing transcription factors. max, not: regulation functions. A, B, and P can be perturbed in the experiment. Right: prediction of regulator states. e—experiment, in which environmental signal is medium (stimulation S=0) and the signaling molecule is knocked out (perturbed variable P, perturbation state =−1). (**B**) Our MEED algorithm. Right: each of the regulators A and B influences its target genes through three possible regulation functions, f_1, f_2 and f_3. The regulation functions are represented by truth tables, in which the first column contains the states of a regulator, and each other column i contains the predicted responses of a target gene controlled by the regulator using f_i. For example, f_1 determines that activated (state=1) regulator upregulates (state=1) its target, and deactivated (state=−1) regulator downregulates (state=−1) its target. Left: matrix of predicted responses. Rows—regulatory programs, each represents a chosen regulator acting on a target gene through a chosen regulation function. Columns—the candidate experiments. MEED is restricted to choose experiments only from this set. For each candidate experiment, the predicted states of regulators A and B appear below its identifier e_1–e_6. For example, in experiment e_3, the predicted states of A and B are 0 and 1, respectively. A matrix entry—a predicted response of a potential target gene assuming it is regulated by its row's regulatory program in its column's experiment. Hence, a row of matrix entries is a predicted profile for a given regulatory program. If the predicted profiles are different, they are referred to as distinguished. MEED aims to find the smallest subset of candidate experiments, which distinguishes between the same pairs of regulatory programs as the full set of candidate experiments. Here, MEED chooses three out of the candidate experiments: e_2, e_3, and e_6, which distinguish all regulatory programs (the remaining ones are marked as deleted). (**C**) The expansion procedure. The experiments proposed by MEED are carried out and the measurements are used in the expansion procedure. Left: the measurements in the chosen experiments are referred to as observed profiles. Middle: a matrix as in B, including only experiments chosen by MEED. The expansion procedure identifies regulatory programs for the genes by matching of predicted and observed profiles (marked as dashed gray lines). Right: genes matching identical regulatory programs constitute regulatory modules. Here, two regulatory modules are found: the regulatory program f_3(A) controls the module of g_2 and g_5, and regulatory program f_1(B) controls g_6.

genes are called *regulators*. The state of each variable is determined by a discrete *regulation function* of its upstream effectors' state. The model can be visualized as a network in which nodes are variables and edges are direct regulatory influences (Box 1A, left). We refer to the topology of this network as to the model's *structure*.

The model can be used to predict the behavior of the regulators in experiments that manipulate the model components. An *experiment* is formalized in the model by defining: (i) *stimulation*—states of the stimulators fixed according to the levels of environmental signals applied in the experiment; (ii) *perturbed variables*—the model variables that are subject to perturbation (in this study we consider only experiments with a single perturbed variable); and (iii) *perturbation states*—fixed states of the perturbed variables, which represent the type of experimental manipulation, such as knockout (perturbation state is −1) or over-activation (perturbation state is 1). The model's regulatory functions can be utilized to predict the states of the regulators in the pathway in a given experiment. These calculated states are called *predicted states* (Box 1A, right; for calculation of predicted states in both cyclic and acyclic models, see Materials and methods).

With the predicted states of regulators in hand, we can also predict the response of potential target genes to a given list of experiments. To do this, we first predefine a set of regulation functions that describe biologically relevant logical relationships between regulators and their targets. Next, we define regulatory programs, which correspond to particular mechanisms of transcriptional regulation in the studied system. A *regulatory program* consists of a set of regulators from the model and a regulation function. The regulators tell 'who' regulates and the regulation function tells 'how'. In this paper, we consider only regulatory programs with a single regulator (e.g., Box 1B). Having a regulatory program and predicted states of its regulators in a given experiment, we may calculate the *predicted response* of this program's potential target genes. The predicted response specifies whether the potential target is in state 1 (*upregulated*), −1 (*downregulated*) or 0 (*neutral*). Finally, the vector of predicted responses in a given set of experiments defines a *predicted profile* of the regulatory

program. Assuming the model is correct, the predicted profile reflects the transcriptional response of a potential target gene controlled by the program in the given set of experiments (Box 1B, left).

The *expansion procedure* aims to find new target genes, which are regulated by the predefined regulatory programs. Given as input a list of experiments together with their measurements and the logical model, the procedure applies probabilistic matching between the observed expression profiles of the genes and the predicted profiles of the regulatory programs: for each gene, we find the predicted profile that matches its observed profile with the highest probability. If the probability exceeds a predefined cut-off threshold, we conclude that the gene is controlled by the regulatory program of this predicted profile. In such case, we say that the regulatory program *matches* the gene. A group of genes that match the same regulatory program constitutes a *regulatory module* (Box 1C). Hence, a regulatory module corresponds to a set of genes that are co-expressed and are predicted to be co-regulated by the same regulator in the model and through a common regulatory mechanism.

Of course, matching of profiles in the expansion procedure can be hampered. In a given set of experiments, some of the predicted profiles might be identical and therefore their regulatory programs cannot be *distinguished* by these experiments (for a full definition in both cyclic and acyclic models, see Materials and methods). In such a case, a single observed profile of genes in a regulatory module could match more than one predicted profile, making it impossible to identify a unique regulatory program for this module. Such regulatory module will be called an *ambiguous module*. As a practical remedy to this ambiguity problem, we propose an algorithm called MEED, which aims to minimize the number of experiments while still maintaining the maximal number of different predicted profiles. The algorithm is given as input a set of *candidate experiments* (i.e., a full set of experiments to choose from; for example, only experiments that can be conducted in a lab). MEED tries to select the smallest subset of the candidate experiments, which can distinguish all regulatory programs. In the case in which the candidate experiments themselves cannot distinguish all regulatory programs, the identified subset should distinguish between the same pairs of regulatory programs as the full candidate set (Box 1B). In our framework, experiments suggested by MEED are used by the expansion procedure to uniquely identify regulatory modules downstream of a given model (Box 1B and C). Both MEED and the expansion procedure utilize the same model and regulatory programs. Therefore, if the suggested experiments distinguish between all pairs of regulatory programs, all identified regulatory modules are unambiguous.

The experimental design problem defined above is computationally intractable (see proof in Supplementary information S1). Therefore, MEED implements an approximation algorithm, which selects the experiments greedily and returns an ordered *experiment list*. MEED calculates an entropy-based score for a list of experiments according to their joint ability to distinguish between regulatory programs. In each greedy step, the algorithm extends the current experiment list with one additional experiment that gives the highest improvement to the list's score (i.e., contributes the most number of new distinguished regulatory program pairs; see Materials and methods for details). The decision is made solely based on model predictions and the chosen experiments need not to be carried out in a lab before the next greedy step. Proposition 2 in Supplementary information S2 gives the approximation factor for our algorithm (the approximation holds for both cyclic and acyclic models).

Experimental design

To assess the performance of our algorithm, we first compare it with alternative ED methods in four tests on 1000 cyclic random models each. The experiments proposed by the analyzed methods were evaluated with respect to their efficiency in distinguishing between regulatory programs (using the *FUP* score; see Materials and methods). The random models were obtained by reshuffling of four human canonical signaling pathways and random assignment of regulation functions (see Supplementary information S3 for the randomization procedure and Supplementary Table S1 for details about the pathways). The regulatory programs were defined by taking all variables that are not stimulators as the regulators, and one regulation function representing activation (referred to as *activation both* in Supplementary Table S2).

Figure 1 presents a comparison of MEED to alternative ED methods in a test on 1000 randomizations of the human TNF pathway. The first compared method, called INDEP (Supplementary information S4), applies the same measure as MEED, but the score is assigned to each experiment independently, ignoring potential dependencies between experiments. In contrast to INDEP, each consecutive experiment designed by MEED radically increases the number of distinguished regulatory program pairs. With this ability, MEED significantly outperforms INDEP, showing the importance of scoring a set of experiments together rather than each experiment independently (Figure 1A and C; see Supplementary information S4 for a toy example). Next, MEED is compared with network-based ED methods (Supplementary information S5), which choose the perturbed variables according to key topological features of the model structure (by in- and out-degree, total number of connections, topological and reverse topological order, referred to as IN-DEGREE, OUT-DEGREE, CONNECTIONS, TOPOL and REV-TOPOL, respectively). These methods are divided into two types according to how they determine stimulation and perturbation states for a predefined perturbed variable: either at random, or with the use of reasoning and scoring of our MEED algorithm (referred to as *random* and *hybrid* methods, respectively). Figure 1B and C shows the advantage of our algorithm over all network-based methods in application to randomizations of the TNF pathway, indicating that MEED reduces the amount of experimental effort required to distinguish between regulatory programs. Notably, the hybrid methods perform better than the random methods, but worse than MEED. Hence, even having predetermined specific molecules to be perturbed, the experimenter can still gain from consulting MEED regarding the type of perturbation and the level of stimulation. Supplementary information S6 provides a detailed description of the analysis and Supplementary Figure S1 presents

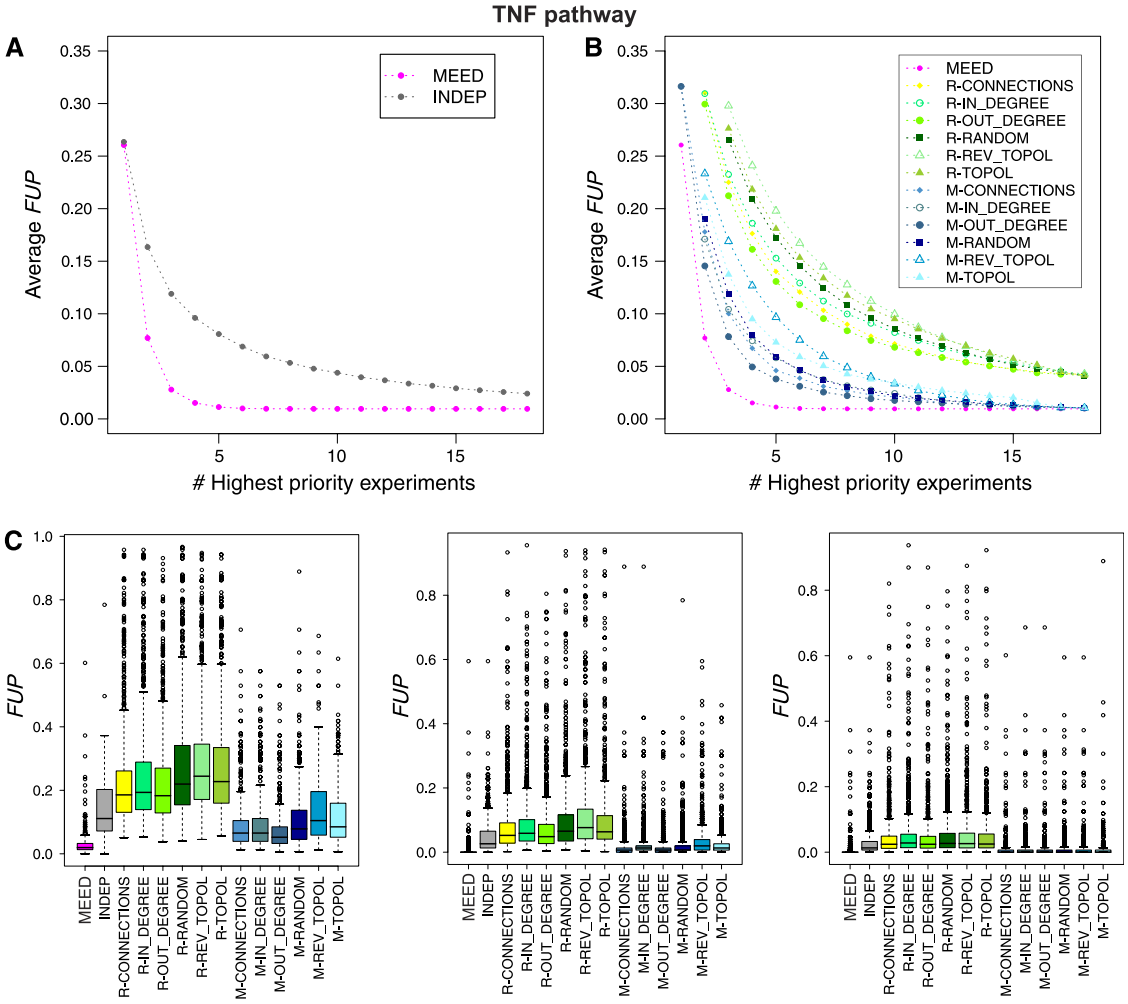
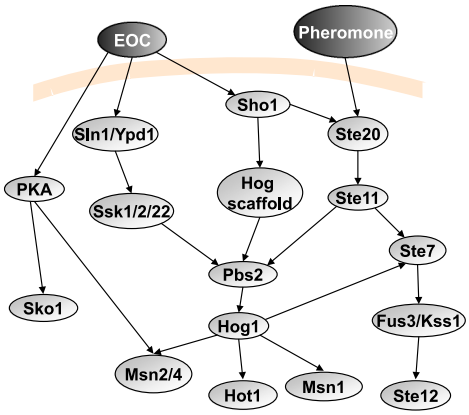


Figure 1 Comparative performance analysis on random models. The comparison is carried out on 1000 cyclic models generated by random reshuffling of the TNF canonical human signaling pathway. **(**A, B**)** *x*-axis: the number of highest priority experiments used from the compared experiment lists to distinguish between regulatory programs, *y*-axis: the *FUP* score averaged over the 1000 random models (only the results with average *FUP*<0.35 are reported). The lower the averaged cumulative *FUP*, the higher the performance of a given ED method. (**A**) Comparison with the INDEP method. Our MEED algorithm has significant advantage over independent experiment scoring. (**B**) Comparison with the network-based methods. The network-based methods choose the perturbed variables according to key features of the structure, whereas stimulations and perturbation states are chosen either at random (the random methods, R-prefixed, green shaded) or following our MEED algorithm (the hybrid methods, M-prefixed, blue shaded). (**C**) Box plots of the *FUP* scores (*y*-axis) for groups of 3, 9 and 15 highest priority experiments from the experiment lists proposed by all analyzed methods (*x*-axis). The results show that MEED consistently outperforms other methods on the tested random models. In general, the hybrid methods have a better performance than the random methods. This evident tendency implies that even allowing MEED to decide only on stimulations and perturbation states, regardless the way the perturbed variables were chosen, can still provide significant improvement.

similar results on randomizations of the remaining three human signaling networks, which are larger and have more stimulators.

Next, we apply MEED to select experiments for the investigation of the yeast cellular response to hyperosmotic and pheromone triggers. The response is mediated by signaling cascades that involve the PKA pathway, as well as the HOG and mating/pseudohyphal growth pathways. The model of the system (based on Gat-Viks and Shamir 2007; see Figure 2) is referred to as the *yeast signaling model* or, in short, the *yeast model*. The model contains two stimulators: environmental osmotic concentration (EOC) and pheromone. We assume that all variables (apart from the Hog-scaffold variable; altogether fifteen variables) are regulators and can be perturbed. A complete depiction of the model, including the regulation functions of all variables, is presented in Supplementary Figure S2. In this study, we focus on the regulation of the immediate response, exploring only the system state before the potential feedback mechanisms affect the signaling network. Therefore, the model does not contain several possible mechanisms of feedback control (e.g., Hog1 protein phosphatases whose production is stimulated after the osmotic shock, or glycerol production that leads to restoration of turgor pressure and stops further activation of the HOG pathway; see Hohmann, 2002) and we utilize only data consisting of measurements that were made shortly after stimulation (Supplementary Table S3). We consider only transcriptional control by single regulators. With this restriction, there are 27 (3^3) possible regulation functions reflecting different means of regulation. To avoid the problem of overfitting (Gat-Viks and Shamir, 2007), we limit ourselves to six biologically relevant regulation functions: *necessary*



#	Stimulation	Perturbation
1.	0.5 M KCl 40 min	—
2.	Control	*sko1*
3.	0.5 M KCl 40 min	*pbs2*
4.	50 nM α-factor 30 min	*fus3D*
5.	Control	*sln1*
6.	0.5 M KCl 40 min	*hog1*
7.	0.5 M KCl 40 min	*ssk1*
8.	0.5 M KCl 40 min	*sho1*
9.	Control	*STE12*
10.	0.5 M KCl 40 min	*ste11*
11.	Control	*MSN2*

Figure 2 Experiment list proposed using MEED for the yeast signaling model. The model is depicted on the left as a network with nodes (ovals) corresponding to environmental conditions (dark gray) and signaling components (light gray). Arrows represent regulatory influences. The list of the experiments designed using MEED is given in a table on the right, listing stimulation (control—YPD) and perturbation (green: knockout and red: overexpression).

activation/inhibition, sufficient activation/inhibition and both (necessary and sufficient) *activation/inhibition* (modified from Yeang and Jaakkola, 2006; detailed truth tables are provided in Supplementary Table S2). In total, we consider 90 regulatory programs (six for each of the fifteen regulators in the pathway).

To have access to experimental data for expansion, we restricted all analyzed ED methods to choose only from candidate experiments that are available in microarray databases. Our candidate set of experiments consists of 25 genome-wide profiles that are reported in five publications (Roberts *et al*, 2000; Hahn *et al*, 2004; Mnaimneh *et al*, 2004; O'Rourke and Herskowitz, 2004; Chua *et al*, 2006; see Supplementary Table S3 for details).

For the yeast model, MEED proposes a list of 11 out of 25 candidate experiments (Figure 2). Figure 3A and B shows that, similar to the results obtained for random networks, MEED distinguished regulatory programs more efficiently than INDEP and the network-based methods. For the yeast model, M-TOPOL performs best from the network-based approaches. The set of all 25 candidate experiments (therefore, also the experiments selected by MEED) cannot distinguish between pairs of regulatory programs within five groups (listed in Supplementary information S7). Accordingly, adding more experiments from this candidate set to the experiment list designed by MEED does not enable to distinguish between more regulatory programs.

Expansion of the yeast signaling model

To test our framework in practice, we performed expansion of the yeast model using the measurements from the 11 experiments chosen by MEED. In the expansion procedure, genes were assigned to regulatory modules by a probabilistic matching of the observed profiles of the genes to the predicted profiles of the regulatory programs (Supplementary information S8). For comparison, we repeated the expansion procedure using experiments selected by independent experiment scoring (INDEP), the best-performing network-based method (M-TOPOL; Figure 3B), as well as two extant ED methods, introduced by Ideker *et al* (2000) and by Barrett and Palsson (2006). Unlike MEED, the two extant methods take as input high-throughput measurements (gene expression or binding data) to build initial network models, and apply an 'on-line' procedure, that is, they use the data from each chosen experiment to propose the next one (see Supplementary Table S4 for a detailed comparative summary of the algorithms). INDEP, MEED and M-TOPOL were applied to choose from the same set of 25 candidate experiments. Expansion using the four highest priority experiments proposed by MEED is utilized to provide initial data for the methods of Ideker *et al* (2000) and Barrett and Palsson (2006), and thus these methods choose only from the remaining 21 candidate experiments. The information required by the methods in each 'on-line' step is also provided by applying the expansion procedure (Supplementary Information S9 reports how the two extant algorithms were implemented; Supplementary Figure S3 describes the main differences between our framework and the extant ED frameworks). For the yeast model, MEED achieves better performance than the extant methods in distinguishing regulatory programs (measured with *FUP* score, see Materials and methods; Figure 3A). The method of Ideker *et al* (2000) reaches its stop criterion already after choosing three experiments.

Using the 11 experiments proposed by MEED, the expansion procedure identifies 26 regulatory modules controlled by the yeast signaling network. More regulatory modules are identified using any number of highest priority experiments proposed by MEED than the same number of experiments proposed by the method of Barrett and Palsson (2006). Moreover, the eleven experiments chosen by MEED enable lower percentage (2 out 26) of ambiguous modules (modules that were matched to more than one regulatory program; Figure 3C). Supplementary Figure S5 reports a similar comparison with M-TOPOL and the method of Ideker *et al* (2000).

The quality of expansion is further evaluated by an *ambiguity* score, reporting the average number of regulatory programs that were identified for each gene. Intuitively, the more regulatory programs matching each ambiguous module, and the more genes it contains, the higher the overall ambiguity score. Unlike the *FUP* score, which evaluates a given ED method based only on model predictions, the ambiguity score evaluates results of expansion, which utilizes

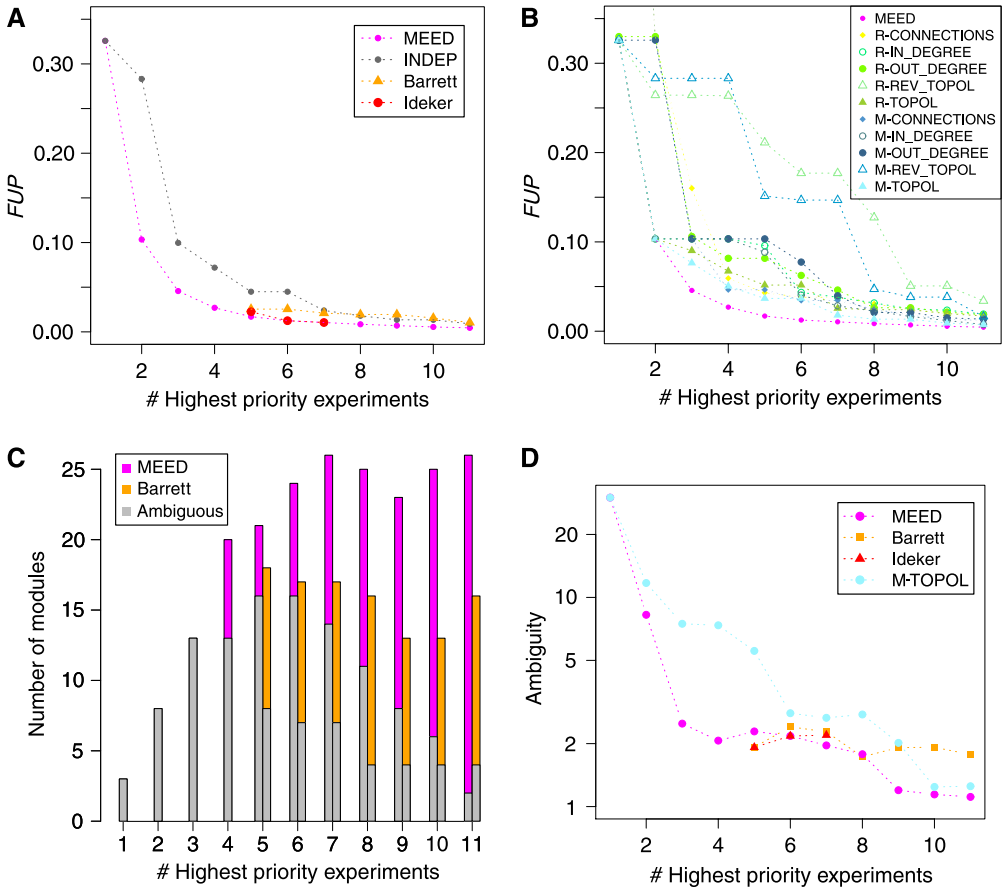


Figure 3 Comparative performance on the yeast signaling model: *FUP* scores and ambiguity of expansion. MEED (plotted in magenta) is compared with INDEP (gray), network-based methods, as well as two extant ED approaches (Barrett and Palsson (2006)—orange; Ideker *et al* (2000)—red). As the two extant methods take as input results of expansion using the first four experiments proposed by MEED, their report starts from the fifth experiment. The method of Ideker *et al* (2000) reaches its stop criterion already after choosing three experiments (fifth to seventh experiment). *x*-axis in all plots (A–D): the number of highest priority experiments. For comparison with MEED, we present up to eleven experiments chosen by the other methods. (**A**, **B**) *FUP* scores. *y*-axis: the *FUP* score, measuring the ability of the experiments to distinguish between regulatory programs (only the results for $FUP < 0.35$ are reported). With the lowest *FUP* for every number of highest priority experiments, MEED outperforms all alternative methods. The best performing of the network-based methods is M-TOPOL. (**C**) Regulatory modules. *y*-axis: the number of modules identified in expansion. The proportion of ambiguous modules is marked in gray. In comparison with the method of Barrett and Palsson (2006), more modules are obtained using the same number of highest priority experiments proposed by MEED (see Supplementary Figure S5 for similar analysis for M-TOPOL and Ideker *et al* (2000)). (**D**) Ambiguity of expansion. *y*-axis: ambiguity score (i.e., the average number of regulatory programs per gene; plotted in log scale). With lower ambiguity score for most numbers of highest priority experiments, MEED outperforms M-TOPOL and the method of Barrett and Palsson (2006) on the yeast model.

experimental data. Figure 3D indicates that MEED outperforms M-TOPOL and (except when the six highest priority experiments are used) the extant methods with respect to ambiguity scores. In Supplementary Figure S6, we use the ambiguity score to show the specificity of the set of experiments chosen by MEED for the particular yeast model. Taken together, the presented results indicate practical applicability as a strong advantage of MEED, which performs comparably or better than the extant approaches although it does not require the data from each chosen experiment to propose the next one (Supplementary Table S4).

Next, we validate the expansion of the yeast network by conducting expansion with additional experiments on top of the eleven experiments suggested by MEED. In this way, we test the stability of gene assignment, that is, whether with more experiments there is a dramatic rearrangement of genes between regulatory modules, or whether the genes are added to or removed from the modules. Our rationale is that a stable gene assignment provides evidence for the correctness of expansion results. Supplementary Figure S4A shows the total number of genes assigned to modules across different numbers of utilized experiments. The initial five highest priority experiments filter out majority of genes. After the 11 experiments proposed using MEED, using additional ones in expansion only slightly decreases the total number of assigned genes. A large fraction of those genes, which are assigned using the experiments proposed by MEED and remain assigned using extended experiment lists, is assigned to the same regulatory modules (Supplementary Figure S4B). Therefore, there is only little rearrangement between the modules when more experiments are used.

Regulatory modules in the yeast signaling model

To assess the biological findings resulting from application of our framework to the yeast signaling model, we focused

further analysis on the obtained regulatory modules. As small modules could have been generated at random, given the large number of potential regulatory programs, we restricted the analysis to fourteen modules containing at least seven genes. Figure 4 presents a map of the expansion, including the identified regulatory modules, their regulatory programs, predicted profiles and the expression matrices of the target genes. The map clearly shows high agreement between predicted profiles and observed profiles. Cases of disagreement (e.g., observed and predicted responses to the second experiment, *sko1* mutant, in two regulatory modules, inhibited by Kss1/Fus3 or Ste12, respectively) show faults in our understanding and incompleteness of the yeast signaling pathway model.

The expansion analysis provides detailed hypotheses regarding the regulatory mechanisms downstream of the yeast signaling model. To evaluate the identified regulatory modules with respect to known mechanisms, we listed eight relevant transcriptional mechanisms based on a comprehensive review (Hohmann, 2002; Supplementary Table S5). The known mechanisms include four single-regulator programs and four combinatorial regulations (not considered in this study). All four single-regulator mechanisms were detected by our analysis (activation by Msn2/4, activation by Ste12, inhibition by Sko1 and activation by Hot1—here ambiguous with Msn1 and Hog1), confirming the quality of our predictions.

In a number of cases, well-characterized target genes were identified by the expansion analysis, thereby serving as positive controls. For example, our analysis indicates that CTT1 and HSP12 are activated by Msn2/4 and FUS1, FUS3 and FIG1 are activated by Ste12, both consistent with the known transcriptional control of these target genes. As the transcriptional network underlying the measured expression data is not known, it is difficult to evaluate our results systematically. We therefore used the well-characterized gene targets reviewed by Hohmann (2002) as a repository of

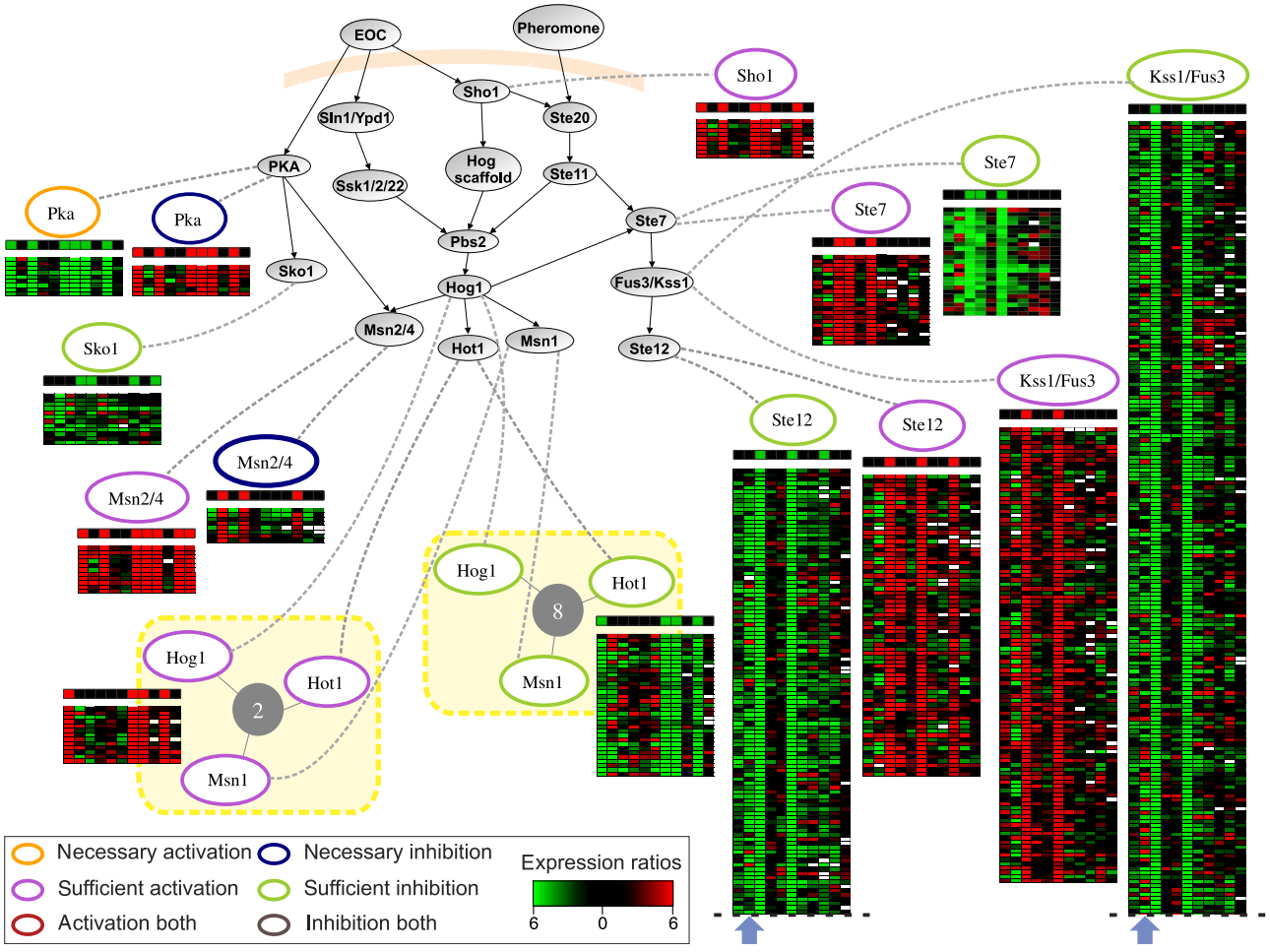


Figure 4 Expansion of the yeast signaling model using the experiments proposed by MEED. The yeast model is depicted in the center of the figure. The identified modules are presented, with additional dashed edges connecting the regulators in the pathway to their regulatory programs (nodes labeled with regulators and having a boundary color-coded according to their regulation function). The ambiguous modules, highlighted with dashed yellow squares, are presented as gray-filled nodes, labeled with their size and connected by edges to all their matching regulatory program nodes. The two ambiguous modules were subject to an additional MEED iteration, which succeeds to distinguish their regulatory programs using only two additional experiments. Matrices showing the expression measurements of target genes (rows) across the eleven experiments proposed using MEED (columns) are presented only for the modules that contain at least seven genes. The columns of the expression matrices are ordered from left to right according to the order proposed by MEED. For clarity, only subsets of the large Ste12 and Kss1/Fus3 matrices are shown. The predicted profiles appear as separate rows above the matrices. For most modules, the expression profiles agree well with the predicted profiles. Blue arrows exemplify experiments in which majority of the module genes disagree with the predicted profile.

regulatory relations (Supplementary Table S5). In total, out of sixteen target genes, known to be regulated by a single transcription factor, eight genes have been assigned correctly and no gene has been assigned to a wrong regulatory module. As combinatorial regulation was not taken into consideration in our analysis, we expect that target genes with more than one known regulator will not be assigned to any of the regulatory modules. Indeed, all six combinatorially regulated target genes did not match any of the regulatory programs (for a detailed comparison between known regulatory relations and our results, see Supplementary Table S5).

Interestingly, four kinases, including Kss1/Fus3, PKA, Sho1 and Ste7, were identified as gene regulators (Figure 4). The hypothesized regulation might be explained by an indirect influence on the target genes through alternative signaling pathways and downstream transcription factors that are not part of the model. Several such alternative pathways are known but were omitted from the model. For example, PKA regulates transcription through the transcription factors Msn2/4 and Sko1 (part of the model) or through Adr1, Rap1 and Crz1 (not modeled; Hohmann, 2002; Yoshimoto *et al*, 2002), Kss1/Fus3 mediates transcription through the Far1 kinase independently of Ste12 (Nern and Arkowitz, 1999), and the Sln1/Ypd1 kinases (which have a small module and therefore were not included in Figure 4) regulate an alternative hypo-osmotic stress pathway through the transcription factor Skn7 (not modeled; Hohmann, 2002). There is no known alternative pathway downstream the signaling molecules Sho1 and Ste7. Our results suggest that these signaling molecules have an indirect effect on gene expression through an additional pathway, independent of the model.

We evaluated all fourteen modules to test whether the proteins encoded by the target genes had a related function or a shared transcriptional regulation. To that end, we scored each module according to its enrichment in GO annotations (using the Ontologizer tool designed by Bauer *et al*, 2008) and sets of transcription targets identified by protein–DNA binding experiments (Harbison *et al*, 2004; Pokholok *et al*, 2006; Zeitlinger *et al*, 2003, computed using a hypergeometric test). Out of the four large modules (containing at least 100 target genes), three modules obtained enrichments below *P*-value threshold 0.001 (Bonferroni corrected; Figure 5). All other modules did not obtain significant enrichment, probably because of their small size (each of these modules contains less than 26 genes, including genes that were not annotated yet).

The enrichment analysis supports and provides insights into the identified modules. For example, it justifies the division of the genes downstream of the mating pathway into two activation modules: a module activated by the transcription factor Ste12 and a module activated by the kinases Kss1/Fus3. According to our enrichment analysis, the genes activated by Ste12 are characterized by several annotations, which are all related to the known functionality of Ste12 as a key transcription factor of the mating pathway (Figure 5). However, the Kss1/Fus3 targets are not enriched in any of these annotations, confirming that Ste12 does not control those targets. To provide additional evidence that the two transcriptional modules are distinct, we performed promoter sequence analysis using the Amadeus tool (Linhart *et al*, 2008). The known binding motif of Ste12 was highly enriched in the module under *sufficient activation* by Ste12 (*P*-value $<10-12$), whereas the module under *sufficient activation* by Kss1/Fus3 was not enriched with this motif. Taken together, our analysis provides evidence for transcriptional regulation by Kss1/Fus3, independently of Ste12 control.

We next asked what is the regulatory pathway mediating *sufficient activation* control by Kss1/Fus3 on its gene targets. Kss1 and Fus3 have no preferential binding to the promoters of the Kss1/Fus3 module (Pokholok *et al*, 2006; data not shown), ruling out the possibility that Kss1/Fus3 have a direct effect on their targets. One potential indirect transcriptional control by Kss1/Fus3 is mediated through the kinase Far1, which mediates cell-cycle arrest in response to pheromone, independently of Ste12. However, our module is not enriched in cell-cycle annotations (Figure 5), indicating that Far1 is unlikely to mediate the observed gene activation downstream of Kss1/Fus3. As more experimental investigations of the pathway connectivity become available, the mechanisms by which Kss1/Fus3 control its targets should be further revealed.

Ambiguity networks and iterative experimental design

To facilitate the inspection of ambiguous modules in a given expanded model, we devised the concept of an *ambiguity network*. Recall that an unambiguous module matches exactly one regulatory program, and an ambiguous module matches strictly more than one program. An ambiguity network is a graph whose nodes represent regulatory programs that matched one of the regulatory modules. One additional node is added for each ambiguous module, labeled by the number of genes it contains. There are edges between the ambiguous module nodes and their matching regulatory program nodes. In this way, the ambiguity network highlights the ambiguous modules and provides details on their size and the alternative regulation hypotheses.

Figure 6 compares two ambiguity networks for two sets of regulatory modules that differ significantly in their ambiguity score. The networks were generated based on the yeast model expansion using two groups of five and six highest priority experiments from the experiment list proposed by M-TOPOL. Adding the sixth experiment (knockout of Pbs2 in high osmotic stress) lowers the ambiguity of the identified regulatory modules (compare Figure 3D). Such a strong drop of ambiguity score can be explained by the fact that with the added experiment, the ambiguous modules either: (i) match fewer regulatory programs, or (ii) contain fewer genes. As an example of the former case, using the five highest priority experiments, the expansion procedure identifies one of the ambiguous modules to be controlled by seven regulatory programs. With the sixth experiment added, this module is replaced by two, matching four and three regulatory programs, respectively (Figure 6, red rectangles). As an example of the latter case, consider the largest ambiguous module containing 3233 genes in expansion performed using five experiments. With the sixth experiment added, this module is replaced by two smaller modules. These modules match three regulatory

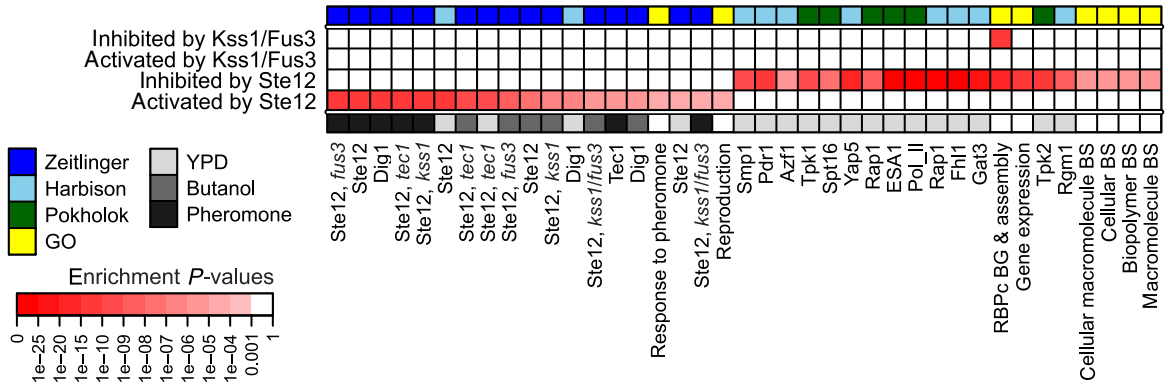


Figure 5 Functional coherence of identified regulatory modules. Enrichment of the target genes from each of four large identified modules (rows) in various experiments (columns). Significant enrichment (Boferroni-corrected hypergeometric *P*-value; indicated by shades of red) represents distinct behavior of the genes in a module compared with the rest of the genome. Enrichment *P*-values in TF–DNA binding targets (Zeitlinger *et al*, 2003; Harbison *et al*, 2004; Pokholok *et al*, 2006) and gene ontology annotation (GO, Ashburner *et al*, 2000) are reported. The different data sets and experiments' environmental conditions are color-coded above and below the matrix, respectively. The profiles used for the enrichment tests were not part of our original dataset. RPBc, ribonucleoprotein complex; BG, biogenesis; BS, biosynthesis.

programs each and contain only 307 and 677 genes (Figure 6, blue rectangles).

Our framework can be used in iterations of the MEED algorithm and expansion procedure. Experiments chosen by MEED from the restricted set of 25 candidate experiments do not distinguish all regulatory programs in the yeast model. Five groups of regulatory modules remain undistinguished (Supplementary information S7). Accordingly, expansion performed using these experiments generates two ambiguous modules (the remaining three groups of regulatory programs are not predicted to control any modules). The ambiguous modules match three regulatory programs each (the regulators Hog1, Msn1 and Hot1 as *sufficient inhibitors* and the same regulators as *sufficient activators*, shown in Figure 4). MEED was re-applied to distinguish between pairs of regulatory programs within these groups. In this iteration, the set of candidate experiments was not limited to the 25 available experiments, but included all experiments possible for the yeast model. Choosing from this set, the algorithm proposed two additional experiments: overexpression of Msn1 and of Hot1, both exposed to low osmotic stress and without pheromone treatment. Carrying out these two experiments and using them in expansion together with the previous eleven experiments (Figure 2) is expected to produce only unambiguous modules. This shows that MEED can be applied to resolve ambiguity in an existing expanded model: First, it can suggest additional experiments by considering experiments that were already carried out, and second, it is able to propose new experiments specifically for the undistinguished regulatory programs.

Discussion

This paper presents a general framework for discovering regulatory modules downstream of a studied signaling pathway. The main goal of extant systems biology frameworks is to create and improve a model of a given system under study. Our framework opens an opportunity to expand such an optimized model with downstream regulatory modules. Our results provide an indication for the good performance of MEED on random networks and the yeast signaling model.

Experiments chosen by the MEED algorithm from a set of candidates can be carried out in a lab and then given as input to the expansion procedure. If the candidate experiments distinguish all regulatory programs, using the experiments selected by MEED in expansion will result in a set of unambiguous modules. Ambiguous modules can be obtained in the case when only part of the experiment list suggested by MEED is used in expansion or when the candidate experiments do not distinguish all regulatory programs. In such case, it is possible to analyze the ambiguity network and specify ambiguous modules that are subject to additional MEED iterations (see section 'Ambiguity networks and iterative experimental design'). This follows the widely accepted iterative framework for biological discovery in systems biology (Ideker *et al*, 2001; Kitano, 2002), with the specific application of experimental design for discovering transcriptional regulation downstream of a given pathway.

MEED does not suggest all experiments necessary for high-confidence assignment of genes to regulatory modules. Rather, it tries to minimize the number of experiments required to distinguish the input list of regulatory programs. Therefore, in practice, model expansion will benefit not only from utilizing extra biological and technical repeats of the suggested experiments, but also from extending the economical list provided by MEED with additional available experiments. First, the new experiments will bring new evidence to refine the assignment of genes to modules. Second, they can be used to validate expansion results. In our study, on adding experiments beyond the eleven proposed by MEED, the total number of assigned genes remains of the same order of magnitude. Moreover, only a small fraction of the genes is rearranged between the modules (see Supplementary Figure S4 and section 'Expansion of the yeast signaling model'). This provides strong support for the robustness of the assignment of genes to modules downstream of the yeast signaling network.

Our modeling formalism was chosen to fit the available biological knowledge. In contrast to detailed modeling

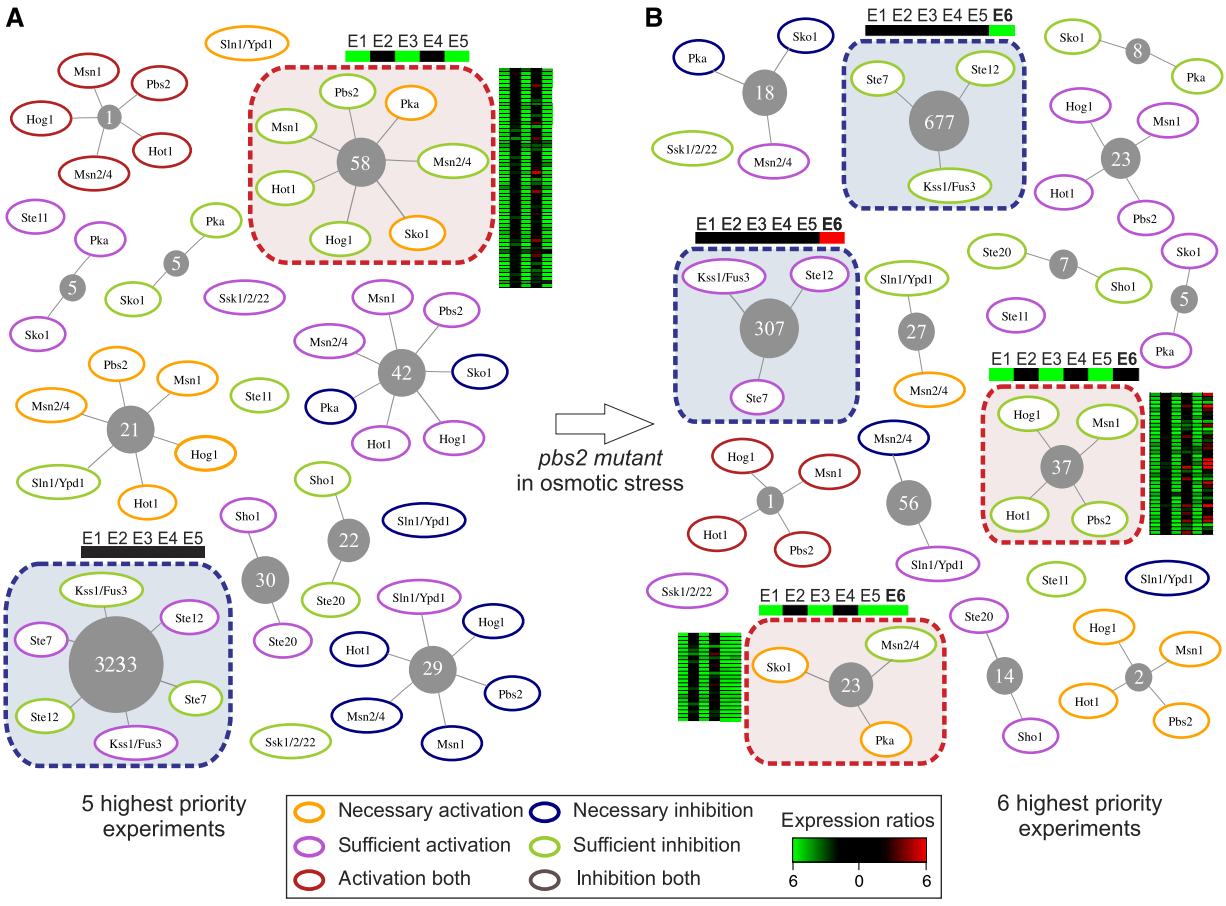


Figure 6 Illustrating expansion results with ambiguity networks. Ambiguity networks for regulatory modules obtained in expansion of the yeast model using the first five (**A**) and six (**B**) experiments on the list proposed by M-TOPOL (i.e., A and B differ only by one additional sixth experiment from the list). The ambiguity network provides a detailed insight into the ambiguous modules. Each white-filled node represents a regulatory program matching one of the identified modules. It is labeled with its regulator, and has a boundary color-coded according to its regulation function. Unambiguous modules are presented only by their unique matching regulatory program, without indicating their size. Ambiguous modules are presented as gray-filled nodes, labeled with their size and connected by edges to all their matching regulatory program nodes. Exemplary modules (highlighted with dashed squares) are shown together with their predicted profile (colored vector above the square). Dashed red: an ambiguous module controlled by seven regulatory programs containing a large set of genes in A is replaced in B by two smaller ambiguous modules controlled by four and three regulatory programs, respectively. The two modules differ in the gene response to the additional sixth experiment. Matrices showing expression profiles of the target genes (rows) across the experiments (columns) are plotted next to the modules. Dashed blue: A large ambiguous module whose genes did not respond in any of the first five experiments (the corresponding predicted profile is filled with black in A). Using the sixth experiment, the large module is replaced by two smaller ones in B. One module contains genes that were downregulated in the sixth experiment, whereas another contains genes that were upregulated (can be seen in green versus red entries in the predicted profiles of the modules). A large group of genes, whose expression has not changed in the sixth experiment, does not match any profile and therefore is not contained in any regulatory module.

approaches (e.g., ODE modeling), the logical model does not require setting a large amount of parameters, which are unknown for most signaling reactions. Other semiqualitative/qualitative modeling approaches, for example, Boolean networks (Kauffman, 1969; Glass and Kauffman, 1973), linear dynamical systems (Roweis and Ghahramani, 1999), s-systems (Savageau, 1969a, b, 1970), Hopfield nets (Hopfield, 1982) or qualitative differential equations (de Jong, 2002), are dynamic modeling approaches that require time-course data. Here, unlike these approaches, we assume that the regulatory relations are discrete logical functions and the model describes the steady state of the system, thereby enabling to utilize single time-point expression measurements.

In the proposed framework, there is a distinction between the model-based experimental design and data-based expansion procedure: The MEED algorithm selects the experiments independent of the data and relies only on the non-stochastic model predictions of discrete states reflecting responses of putative regulatory targets. The stochastic nature of the data is considered only in the expansion, once the measurements from the experiments proposed by MEED are available. We expect that based on the proposed framework, it will be possible to develop techniques handling probability model formalisms, such as a Bayesian network model, which represent the prior belief in the logical functions (as implemented in Gat-Viks and Shamir (2007)).

In this contribution, we considered only regulatory programs with single regulators and experiments with perturbations of one molecule. Our approach is general and can be extended to investigate combinatorial control by taking into account regulatory programs with multiple regulators and experiments with more than one perturbed variable (analyzed previously by, e.g., Kaufman *et al*, 2005; Nelander *et al*, 2008). The MEED algorithm, which is linear in the number of

regulatory programs (Supplementary information S10), will scale to the enlarged problem, with the condition that only a small selection of a vast number of all combinatorial possibilities is considered. For example, for two regulators and three possible states of the variables, the number of all possible regulation functions is $3^{(3^2)} = 19683$. Already in case of single regulator programs, we choose six biologically relevant regulation functions (out of 27 possible). Applying the same selection criteria, one could consider only a handful of biologically relevant combinatorial functions (e.g., adapting the combinatorial schemas proposed by Buchler *et al*, 2003; Yeang and Jaakkola, 2006).

The results of the expansion procedure must be interpreted with caution. First, MEED and the expansion procedure rely strongly on prior knowledge encoded in the model, and therefore can fail when the assumed network topology or logical relations are wrong. To overcome this obstacle, the model of the signaling pathway should be corrected using a refinement procedure (Gat-Viks and Shamir, 2007) before applying our framework. Second, measurement errors may distort the observed profiles, and consequently, the assignment of genes to regulatory modules in expansion. Third, to avoid superfluous assignment of genes, the regulatory programs should reflect all biologically relevant means of transcriptional control (see Supplementary Figure S4 for details).

Lastly, our analysis is limited to the regulation of immediate gene expression response that is secondary to signaling. We rely on the assumption that the system state can be explored before transcriptional feedback mechanisms are activated and affect the pathway. Indeed, in our case study, the yeast signaling model does not include slower temporal processes such as feedback loops, and is integrated with expression profiles measured shortly after stimulation, during the immediate gene expression response. Our results (see section 'Regulatory modules in the yeast signaling model') suggest that our modeling assumptions are appropriate for this system. In the future, we hope that the methodology can be extended to handle slower temporal processes analogously to the construction of dynamic Bayesian networks (DBN; Perrin *et al*, 2003) from steady state Bayesian networks. As in DBN, we expect that MEED and the expansion procedure can be generalized from the steady state model to the dynamic model.

MEED has several important benefits: first, it builds on qualitative knowledge formalized in a simple logical model. Therefore, it enables to focus the experimental investigation on any biological system with prior understanding of its signaling pathways. Second, the algorithm has the ability to choose experiments without access to high-throughput experimental data. Third, MEED is able to take into account dependencies between the experiments and select a whole list of required experiments at once. Finally, MEED may consider all possible or only a restricted set of candidate experiments (e.g., due to experimental cost and technical limitations). Our results show that MEED can significantly reduce the amount of experimental effort required to elucidate regulatory mechanisms downstream of a given pathway. Moreover, we showed that even having a predefined set of perturbed molecules, an experimenter can significantly benefit from consulting MEED with regard to possible environmental stimulations and the type of genetic perturbations. Taken together, our approach opens the way to practical experimental design based on well-established qualitative biological knowledge.

Materials and methods

Predictive logical model

Our ED and expansion framework is based on a predictive model, which formalizes the available knowledge on a given signaling pathway. The definitions of the model are given in section 'Outline of our framework' (see Gat-Viks *et al*, 2004 for more details). In our analysis (see sections 'Experimental design' to 'Ambiguity networks and iterative experimental design'), we assume that all variables may have three possible states, and that all zero-indegree variables are stimulators.

Recall that an experiment is given by an assignment of states to the stimulators (called stimulation), a set of perturbed variables, and their state assignment (perturbation state). In this study, we consider only experiments, in which either none or exactly one variable is perturbed. The perturbed variable cannot be a stimulator. Assuming that k perturbation states are possible for each variable, having s stimulators and p variables that can be perturbed in a given model, the number of all possible experiments is $k^s(pk+1)$.

Given an experiment, a *predicted model state* is an assignment of states to all variables in the model so that (i) the stimulators are assigned their stimulation and the perturbed variables are assigned their perturbation state, and (ii) the state of each other variable is consistent with the states of its regulators. In other words, the state of each variable equals the output of its regulation function when applied on its regulators' states. In this way, each predicted model state corresponds to a steady state of the system in a given experiment. The state assigned to each variable by a given predicted model state is called a *predicted state* of the variable. In case of an acyclic model, each experiment has exactly one possible predicted model state, giving a unique predicted state for each variable. However, for a model whose structure contains cycles, it is possible to obtain zero, one or several possible predicted model states (see Gat-Viks *et al* (2004) for computational analysis and Supplementary Figure S7A for an example).

Regulatory programs and their predicted profiles

Recall that a regulatory program is defined by a regulator and a regulation function (see section 'Outline of our framework'). Applying the regulation function to the regulator's predicted state in a given experiment, we obtain a state reflecting the response of a potential target gene to this regulatory program. This state is called predicted response. A predicted profile for a regulatory program is a vector of predicted responses to a list of experiments.

In a general case, one experiment may define a number of predicted model states, giving several predicted states per regulator. Therefore, each regulatory program might have a number of predicted profiles, determined by the different combinations of predicted model states in a list of experiments (exemplified in Supplementary Figure 7B top). We assume that under a given experiment, the biological system reaches the steady state corresponding to only one predicted model state. Before designing and carrying out experiments, we cannot anticipate which combination of their predicted model states will be reached. This problem is taken into account as part of the MEED algorithm (see section 'Distinguishing regulatory programs' and 'The MEED algorithm'). However, the expansion procedure (described in section 'Outline of our framework') requires as input a single predicted profile for each of the regulatory programs. The single profile can be obtained by choosing the predicted model states that have the best fit with the experimental measurements (for a detailed algorithm, see Gat-Viks *et al*, 2004).

Distinguishing regulatory programs

We start by defining how a given set of regulatory programs R is distinguished by a predicted model state, next we extend the definition for a single experiment (which might induce several predicted model states), and finally, we generalize by stating how R is distinguished by a set of experiments. Recall that a predicted model state m^e for a given experiment e assigns to each regulator its predicted state. From these states we can compute the predicted responses for the regulatory programs in R. In this way, each predicted model state induces a natural *partition* of the set of regulatory programs. The partition contains two regulatory programs in the same block if and only if they have the same predicted response. Regulatory programs contained in different blocks of this partition are said to be *distinguished by the predicted model state* m^e (exemplified in Supplementary Figure 7B middle).

An experiment may in general define a number of predicted model states. We consider a partition $T(e)$ of the set of regulatory programs R induced by an experiment e as the supremum over the set of partitions induced by its predicted model states. The regulatory programs contained in different blocks of $T(e)$ are called *distinguished by the experiment e* (see Supplementary Figure 7B middle for an example and Supplementary information S11 for discussion). If an experiment has no predicted model states, this experiment is not informative and its partition includes only one block containing all regulatory programs.

A pair of regulatory programs is *distinguished by a list of experiments* $E\{e_1,\ldots,e_n\}$ if and only if they are distinguished by at least one of its experiments. Equivalently, we say that E distinguishes between regulatory programs that are contained in separate blocks of a partition $S(E)=T(e_1)\cap\ldots\cap T(e_n)$ (exemplified in Supplementary Figure 7B, bottom). The partition for an empty set of experiments is a full, one-block partition containing all regulatory programs. Regulatory programs contained in the same block of the partition $S(E)$ are not distinguished by any of the experiments, whereas regulatory programs in different blocks are distinguished by at least one experiment. We say that an experiment list E *distinguishes all* regulatory programs, if its corresponding partition $S(E)$ contains only single-element blocks. Note that if a given list of experiments E, distinguishes between two regulatory programs, their predicted profiles will be different (i.e., have at least one different predicted response, see Supplementary information S11) in all possible combinations of predicted model states for E. In particular, they will be different for the steady states the biological system has reached under the experiments. This feature is crucial for our framework. It assures that by maximizing the number of distinguished regulatory programs MEED maximizes also the diversity of predicted profiles used in the expansion procedure.

The MEED algorithm

The MEED algorithm aims to select an economical subset of given candidate experiments that can be used for unambiguous expansion of a given model. First, MEED calculates the set of pairs of regulatory programs, which are distinguished by the candidate experiments. Next, it tries to select the smallest subset of the candidate experiments, which distinguishes between the same regulatory programs. The decision version of this problem is NP-complete (proposition 1 in Supplementary information S1). To obtain a practical solution, MEED implements a greedy approximation algorithm for this problem, as detailed below.

MEED evaluates the ability of a list of experiments to distinguish regulatory programs using an *entropy score*. Assume that a given list of experiments E induces a partition $S(E)$ of a set of r regulatory programs into C disjoint blocks. The score is defined as

$$H(E) = -\sum_{c=1}^{C} \frac{n_c}{r} \log\left(\frac{n_c}{r}\right)$$

where n_c is the number of regulatory programs in block c, $1\leqslant c\leqslant C$. If all regulatory programs are distinguished by the list of experiments E, then $C=r$ and the corresponding score is $H(E)=\log(r)$. If all regulatory programs are undistinguished by E, there is only one block in the partition, $C=1$ and $H(E)=0$. Intuitively, the higher the entropy score, the higher the ability of the list of experiments to distinguish between the regulatory programs. Accordingly, *entropy gain* is given by $H(E\cup e)-H(E)$, where $S(E\cup e)=S(E)\cap T(e)$ (i.e., the additional experiment introduces a finer partition of the set of regulatory programs). Entropy gain evaluates how much the joint ability to distinguish between regulatory programs will improve when the experiment e is added to the list of experiments E.

MEED outputs an ordered list of chosen experiments. In each greedy step, the algorithm extends the current experiment list E (identified in the previous steps) with one additional experiment e, which provides the highest entropy gain (Supplementary information S10). In this way, the chosen experiment results in the most 'uniform' partition of the set of regulatory programs. The algorithm approximates the size of the optimal solution (i.e. the number of proposed experiments) by the factor $1+\ln(r)+\ln(\log(k))$, where r is the number of regulatory programs and k is the number of states each model variable can have (proposition 2 in Supplementary information S2). The approximation holds for both cyclic and acyclic models.

The *FUP* score

We report the performance of a list of experiments E with the *fraction of undistinguished pairs (FUP)*. The score is given by the proportion of regulatory program pairs undistinguished by E out of all possible pairs of regulatory programs:

$$FUP(E) = \frac{\sum_c n_c(n_c-1)}{r(r-1)}$$

where n_c is the size of the c-th block of the corresponding partition $S(E)$ of the set of r regulatory programs. $FUP(E)$ attains values between 0 (all regulatory programs are distinguished) and 1 (no pair of regulatory programs is distinguished). The more pairs of regulatory programs are distinguished by a given list of experiments, the smaller its *FUP* score. Unlike the ambiguity score (see section 'Expansion of the yeast signaling model'), which evaluates the results of expansion utilizing experimental data, *FUP* evaluates a given ED method based only on model predictions.

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

Acknowledgements

We are grateful to Igor Ulitsky for valuable help on logical modeling. We also thank Roman Brinzanik, Marta Łuksza and Marcel Schulz for enriching discussions and comments on this paper. ES was supported by the SFB 618 grant of the Deutsche Forschungsgesellschaft (DFG), and partially by the PBZ-MniI-2/1/2005 grant. JT was supported by the PBZ-Mnil-2/1/2005 grant. IG-V was supported by funds from a Max–Planck Research Award.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Akutsu T, Kuhara S, Maruyama O, Miyano S (1998) A system for identifying genetic networks from gene expression patterns produced by gene disruptions and overexpressions. *Genome Inform Ser Workshop Genome Inform* **9:** 151–160

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald

M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25:** 25–29

Barrett CL, Palsson BO (2006) Iterative reconstruction of transcriptional regulatory networks: an algorithmic approach. *PLoS Comput Biol* **2:** e52

Bauer S, Grossmann S, Vingron M, Robinson PN (2008) Ontologizer 2.0—a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics* **24:** 1650–1651

Bolouri H, Davidson EH (2002) Modeling transcriptional regulatory networks. *Bioessays* **24:** 1118–1129

Buchler NE, Gerland U, Hwa T (2003) On schemes of combinatorial transcription logic. *Proc Natl Acad Sci USA* **100:** 5136–5141

Bussemaker HJ, Li H, Siggia ED (2001) Regulatory element detection using correlation with expression. *Nat Genet* **27:** 167–171

Chua G, Morris QD, Sopko R, Robinson MD, Ryan O, Chan ET, Frey BJ, Andrews BJ, Boone C, Hughes TR (2006) Identifying transcription factor functions and targets by phenotypic activation. *Proc Natl Acad Sci USA* **103:** 12045–12050

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

Gardner TS, di Bernardo D, Lorenz D, Collins JJ (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301:** 102–105

Gat-Viks I, Shamir R (2007) Refinement and expansion of signaling pathways: the osmotic response network in yeast. *Genome Res* **17:** 358–367

Gat-Viks I, Tanay A, Shamir R (2004) Modeling and analysis of heterogeneous regulation in biological networks. *J Comput Biol* **11:** 1034–1049

Glass L, Kauffman SA (1973) The logical analysis of continuous, non-linear biochemical control networks. *J Theor Biol* **39:** 103–129

Hahn JS, Hu Z, Thiele DJ, Iyer VR (2004) Genome-wide analysis of the biology of stress responses through heat shock transcription factor. *Mol Cell Biol* **24:** 5249–5256

Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature* **431:** 99–104

Hartemink AJ (2005) Reverse engineering gene regulatory networks. *Nat Biotechnol* **23:** 554–555

Hohmann S (2002) Osmotic stress signaling and osmoadaptation in yeasts. *Microbiol Mol Biol Rev* **66:** 300–372

Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci USA* **79:** 2554–2558

Ideker T, Galitski T, Hood L (2001) A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet* **2:** 343–372

Ideker TE, Thorsson V, Karp RM (2000) Discovery of regulatory interactions through perturbation: inference and experimental design. *Pac Symp Biocomput* **5:** 305–316

Kauffman S (1969) Homeostasis and differentiation in random genetic control networks. *Nature* **224:** 177–178

Kaufman A, Keinan A, Meilijson I, Kupiec M, Ruppin E (2005) Quantitative analysis of genetic and neuronal multi-perturbation experiments. *PLoS Comput Biol* **1:** e64

King RD, Whelan KE, Jones FM, Reiser PG, Bryant CH, Muggleton SH, Kell DB, Oliver SG (2004) Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature* **427:** 247–252

Kitano H (2002) Systems biology: a brief overview. *Science* **295:** 1662–1664

Linhart C, Halperin Y, Shamir R (2008) Transcription factor and microRNA motif discovery: the Amadeus platform and a compendium of metazoan target sets. *Genome Res* **18:** 1180–1189

Mnaimneh S, Davierwala AP, Haynes J, Moffat J, Peng WT, Zhang W, Yang X, Pootoolal J, Chua G, Lopez A, Trochesset M, Morse D, Krogan NJ, Hiley SL, Li Z, Morris Q, Grigull J, Mitsakakis N, Roberts CJ, Greenblatt JF *et al* (2004) Exploration of essential gene functions via titratable promoter alleles. *Cell* **118:** 31–44

Nachman I, Regev A, Friedman N (2004) Inferring quantitative models of regulatory networks from expression data. *Bioinformatics* **20:** i248–i256

Nelander S, Wang W, Nilsson B, She QB, Pratilas C, Rosen N, Gennemark P, Sander C (2008) Models from experiments: combinatorial drug perturbations of cancer cells. *Mol Syst Biol* **4:** 216

Nern A, Arkowitz RA (1999) A Cdc24p-Far1p-Gbetagamma protein complex required for yeast orientation during mating. *J Cell Biol* **144:** 1187–1202

O'Rourke SM, Herskowitz I (2004) Unique and redundant roles for HOG MAPK pathway components as revealed by whole-genome expression analysis. *Mol Biol Cell* **15:** 532–542

Perrin BE, Ralaivola L, Mazurie A, Bottani S, Mallet J, d'Alche-Buc F (2003) Gene networks inference using dynamic Bayesian networks. *Bioinformatics* **19:** ii138–ii148

Pokholok DK, Zeitlinger J, Hannett NM, Reynolds DB, Young RA (2006) Activated signal transduction kinases frequently occupy target genes. *Science* **313:** 533–536

Roberts CJ, Nelson B, Marton MJ, Stoughton R, Meyer MR, Bennett HA, He YD, Dai H, Walker WL, Hughes TR, Tyers M, Boone C, Friend SH (2000) Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* **287:** 873–880

Roweis S, Ghahramani Z (1999) A unifying review of linear Gaussian models. *Neural Comput* **11:** 305–345

Savageau MA (1969a) Biochemical systems analysis. I. Some mathematical properties of the rate law for the component enzymatic reactions. *J Theor Biol* **25:** 365–369

Savageau MA (1969b) Biochemical systems analysis. II. The steady-state solutions for an n-pool system using a power-law approximation. *J Theor Biol* **25:** 370–379

Savageau MA (1970) Biochemical systems analysis. 3. Dynamic solutions using a power-law approximation. *J Theor Biol* **26:** 215–226

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

Tong S, Koller D (2001) *Active Learning for Structure in Bayesian Networks, 17th International Joint Conference on Artifical Intelligence (IJCAI)*. Seattle, Washington: Morgan Kaufmann

Vatcheva I, de Jong H, Bernard O, Mars NJI (2006) Experiment selection for the discrimination of semi-quantitative models of dynamical systems. *Artificial Intelligence* **170:** 472

Yeang CH, Jaakkola T (2006) Modeling the combinatorial functions of multiple transcription factors. *J Comput Biol* **13:** 463–480

Yeang CH, Mak HC, McCuine S, Workman C, Jaakkola T, Ideker T (2005) Validation and refinement of gene-regulatory pathways on a network of physical interactions. *Genome Biol* **6:** R62

Yoshimoto H, Saltsman K, Gasch AP, Li HX, Ogawa N, Botstein D, Brown PO, Cyert MS (2002) Genome-wide analysis of gene expression regulated by the calcineurin/Crz1p signaling pathway in *Saccharomyces cerevisiae*. *J Biol Chem* **277:** 31079–31088

Zeitlinger J, Simon I, Harbison CT, Hannett NM, Volkert TL, Fink GR, Young RA (2003) Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. *Cell* **113:** 395–404

