BIOINFORMATICS

Vol. 24 ISMB 2008, pages i172–i181
doi:10.1093/bioinformatics/btn155

Combinatorial influence of environmental parameters on transcription factor activity

T.A. Knijnenburg[1,2,*], L.F.A. Wessels[1,3] and M.J.T. Reinders[1,2]

[1]Information and Communication Theory Group, Department of Mediamatics, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Mekelweg 4, 2628 CD Delft, [2]Kluyver Centre for Genomics of Industrial Fermentation and [3]Bioinformatics and Statistics, Department of Molecular Biology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands

ABSTRACT

Motivation: Cells receive a wide variety of environmental signals, which are often processed combinatorially to generate specific genetic responses. Changes in transcript levels, as observed across different environmental conditions, can, to a large extent, be attributed to changes in the activity of transcription factors (TFs). However, in unraveling these transcription regulation networks, the actual environmental signals are often not incorporated into the model, simply because they have not been measured. The unquantified heterogeneity of the environmental parameters across microarray experiments frustrates regulatory network inference.

Results: We propose an inference algorithm that models the influence of environmental parameters on gene expression. The approach is based on a yeast microarray compendium of chemostat steady-state experiments. Chemostat cultivation enables the accurate control and measurement of many of the key cultivation parameters, such as nutrient concentrations, growth rate and temperature. The observed transcript levels are explained by inferring the activity of TFs in response to combinations of cultivation parameters. The interplay between activated enhancers and repressors that bind a gene promoter determine the possible up- or downregulation of the gene. The model is translated into a linear integer optimization problem. The resulting regulatory network identifies the combinatorial effects of environmental parameters on TF activity and gene expression.

Availability: The Matlab code is available from the authors upon request.

Contact: t.a.knijnenburg@tudelft.nl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Transcription factors (TFs) mediate the activation or repression of gene expression by binding specific regulatory sequences (motifs) in gene promoters. The combinatorial interactions of multiple TFs play an essential role in transcriptional regulation. A classical example is *Escherichia coli*'s lactose system, where the *lac* operon is expressed only if the concentration of TF CRP is high and that of TF LacI is low. Presently, many studies have revealed an important role for combinatorial interactions between different TFs in establishing the complex patterns of gene expression (Balaji *et al.*, 2006). The advent of high-throughput genomic measurement techniques enabled the application of genome-wide computational approaches aimed at inferring these regulatory relations. Sequence data, microarray gene expression data and ChIP–chip TF binding data have been integrated in many different ways to derive regulatory networks. Several approaches fit expression data using linear regression models, where the predictors are the TFs, i.e. their binding potential or number of motifs in a gene promoter (Bussemaker *et al.*, 2001; Gao *et al.*, 2004; Nguyen and D'haeseleer, 2006). The effect of multiple TFs on gene expression is modeled as the weighted sum of the contribution of individual TFs. Combinatorial regulation by TFs, i.e. synergistic or antagonistic effects of multiple TFs on gene expression, are not incorporated into these models. Most methods that do include combinatorial effects limit the scope to TF pairs, e.g. (Bonneau *et al.*, 2006; Chang *et al.*, 2006; Das *et al.*, 2004; Yu *et al.*, 2006). Bonneau *et al.* employ continuous versions of logic functions (OR, AND and XOR) of the activities of TF pairs as additional predictors in the regression model. Although, in principle, these methods can be extended to model the combinatorial effects of more than two TFs, the model will be too complex to reliably estimate its parameters given the currently available data. Segal *et al.* (2003) and Yeang and Jaakkola (2006) present quite different approaches to the problem of combinatorial regulation in transcription networks. Segal *et al.* constructed regulatory networks by building decision trees. Genes are grouped into regulatory modules, which are defined by a hierarchical decision tree, where the decisions at the nodes of the tree are based on the expression levels of TFs. In Yeang and Jaakkola, a TF is characterized as an enhancer or a repressor, being either necessary or sufficient to cause up- or downregulation of a gene. The combinatorial function of all TFs that can bind a gene promoter is modeled as the consensus prediction of the individual TFs. It should be noted that these two approaches, as well as many of the abovementioned ones, rely on the often incorrect assumption that the activity of a TF can be derived from the expression of the gene that codes for the TF.

So far, regulatory networks have been presented as graph structures showing the (combinatorial) regulatory effect of TFs on individual genes, modules of similarly expressed or otherwise related genes or on other TFs. The extracellular signals that trigger the activation or deactivation of TFs are usually not part of the generated network. Yet they could provide more direct and trustworthy evidence to infer TF activity than other signals, such as the gene expression of a TF. Three main reasons for

[*]To whom correspondence should be addressed.

their exclusion can be identified. First, many studies on yeast are based on shake-flask cultures, where parameters like growth rate and nutrient availability are continuously changing and cannot be controlled or accurately measured. Consequently, conditions can not be accurately defined. Second, very often research questions are approached from a single perspective, i.e. a condition of interest is compared to a reference condition. Differential gene expression is then attributed to the difference between the condition of interest and the reference condition. These approaches ignore combinatorial effects of growth parameters, the presence of which have been established by various studies, e.g. Castrillo *et al.* (2007); Knijnenburg *et al.* (2007); Regenberg *et al.* (2006). That is, if the measurements were repeated using a different medium composition or temperature, chances are that a different set of differentially expressed genes would be identified. Thus, these approaches only model the *differences* between growth conditions, and not the growth conditions *themselves*. Note that this strategy is implicitly incorporated into two-channel microarray measurements, which output the gene expression ratio between the condition of interest and the reference condition. Third, when combining different microarray experiments, differences in mRNA extraction protocols, microarray platform and possibly normalization and summarization algorithms, add to the already large amount of unquantified heterogeneity amongst experimental conditions (Bammler *et al.*, 2005; Tan *et al.*, 2003).

The context dependency of regulatory networks has been identified and acknowledged in many studies. For example, in Bar-Joseph *et al.* (2003) annotation data are employed to identify the biological context in which the inferred regulatory interactions are assumed to take place. In Luscombe *et al.* (2004) condition-specific regulatory networks were derived. In this case, condition-specific refers to one of five phenomena (cell cycle, sporulation, DNA damage, stress response or diauxic shift), which were investigated with five different microarray datasets. Myers and Troyanskaya (2007) propose a Bayesian approach for context-sensitive integration of diverse genomic data. Note however, that in these approaches, the *precise* environmental conditions under which the microarray measurements were taken are *not* included in the model.

In this work we do incorporate the actual cultivation parameters in the computational framework and use this information to infer combinatorial regulation by TFs. The work is based on a yeast transcriptome compendium, comprised of 170 microarray measurements (Knijnenburg *et al.* manuscript in preparation). These measurements encompass 55 unique growth conditions with a variable number of independent biological replicates per condition. All cultivations were performed in chemostat fermentors under steady-state conditions. In a chemostat, culture broth (including biomass) is continuously replaced by fresh medium at a fixed and accurately determined dilution rate. When the dilution rate is lower than $\mu_{\max}$, the maximal specific growth rate of the micro-organism, a steady-state situation will be established in which the specific growth rate equals the dilution rate. In such a steady-state chemostat culture, μ is controlled by the (low) residual concentration of a single growth-limiting nutrient. Across the 55 different conditions, there are nine varying cultivation parameter types, including limiting element, growth rate, carbon source, aeration and temperature. Each type can assume a unique set of values. For example, in a given experiment, the employed limiting element is either carbon, nitrogen, sulfur, phosphorus, zinc or iron. Thus, each condition is characterized by a configuration of settings of these nine cultivation parameter types (Fig. 1).

In order to model the effects of the cultivation parameters on gene expression while explicitly incorporating TFs, we follow a two-step procedure. An overview of this procedure is presented in Figure 2. First, we apply a forward stepwise regression strategy to quantify the (combinatorial) effect of these environmental parameters on gene expression. The regression is performed for each gene individually. Figure 1 depicts the results of the regression analysis for one particular gene. The influence of a cultivation parameter on the expression of a gene is represented by its regression weight. These weights are discretized by mapping non-zero elements to 1 or -1, depending on the sign of the weight. Given that changes in gene expression levels as observed across different environmental conditions can be attributed to changes in the activity of TFs, we aim to infer the activity of TFs as a function of the cultivation parameters. This forms the second step of our approach. The goal is to estimate $\mathbf{M}$, such that $\widehat{\mathbf{R}}$ is the optimal approximation of the discretized regression coefficients in $\mathbf{R}$. The elements of $\mathbf{M}$ are -1, 0 or 1 and indicate whether a TF is activated as an enhancer (1) or a repressor (-1) under a (combinatorial) cultivation parameter. Additionally, each TF has a particular generic enhancer strength and a repressor strength. In the procedure we employ auxiliary matrix $\mathbf{T}$, which is derived from ChIP–chip experiments and literature and indicates whether a TF can bind a gene promoter. To decide whether a gene is upregulated, downregulated or not affected by a particular cultivation parameter, indicated by a 1, -1 and 0 in $\widehat{\mathbf{R}}$, respectively, we use the following rules concerning transcriptional regulation: if there is at least one active enhancer in a gene promoter, then the gene *can be* upregulated. If there are only active enhancers in a gene promoter, then the gene *is* upregulated. Similar rules apply to the repressors. If there are both active enhancers and repressors in a gene promoter, we compare total enhancer strength, which is the sum of the strengths of the activated enhancers, with its repressor counterpart. When the enhancer strength is larger than the repressor strength, the gene is upregulated. The gene is downregulated when the repressor strength exceeds the enhancer strength. Figure 2c visualizes the active TFs that bind the gene promoters of genes g1, g2 and g3 under cultivation parameter A. From $\mathbf{M}$ we deduce that three TFs are activated; α and β are enhancers, δ is a repressor. From $\mathbf{T}$ we deduce that α binds all three promoters, β binds the g2 and g3 promoters and δ only binds the promoter of g3. Gene g1 and g2 are upregulated, since only active enhancers bind the promoters. For gene g3, the repressor strength of TF δ exceeds that of the sum of the two enhancers, thereby downregulating the gene. The concept of TF strength enables the inference of hierarchical or combinatorial effects amongst TFs that bind a gene promoter. The inference algorithm is translated into a linear mixed integer optimization problem and solved accordingly. Both the elements of $\mathbf{M}$ as well as the TF strengths are estimated, such that the predicted gene regulation in $\widehat{\mathbf{R}}$ maximally corresponds with the discretized regression coefficients in $\mathbf{R}$. The abovementioned rules become constraints in the optimization problem. See the Methods section for details. The resulting model identifies the combinatorial influence of cultivation parameters on TF activity and gene expression. Furthermore, it infers the combinatorial regulatory code of multiple TFs in gene promoters.

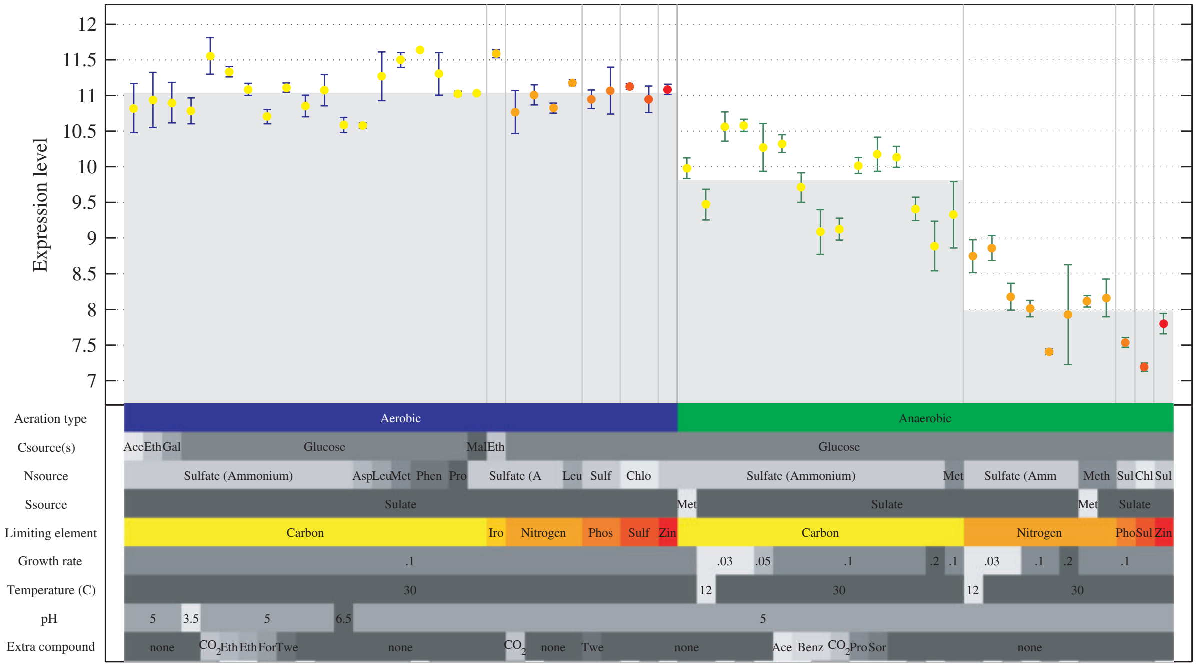

Fig. 1. Expression levels of a gene (*COX5A*) across the 55 cultivation conditions. The colored matrix is a schematic representation of the settings of the nine cultivation parameter types across the 55 conditions. The colored lanes indicate the cultivation parameter types that are employed to order the experiments, in this case, aeration type and limiting element. The regression model which models the gene expression as a function of the cultivation parameters, selected one single effect, i.e. aeration type, and one combinatorial effect, i.e. aeration type anaerobic together with limiting element carbon. The reconstructed expression pattern based on these two effects is indicated by the shaded area.

2 METHODS

2.1 Microarray data

The *Saccharomyces cerevisiae* laboratory reference strain CEN.PK 113-7D (*MAT*a) was grown in chemostat fermentors under 55 different conditions. For each condition, a variable number of independent biological replicates was performed, although mostly three, summing up to 170 microarray measurements. Across the 55 conditions, nine different cultivation parameter types can be identified. A cultivation parameter type, e.g. the carbon source, is described as a categorical variable and contains two or more categories, e.g. the used carbon source can be either maltose, glucose or ethanol. Each condition is characterized by a specific combination of these categories across the nine cultivation parameter types. Figure 1 presents an overview of the relevant categories assumed by the parameter types per condition. Sampling of the chemostat cultures, probe preparation and hybridization to single-channel Affymetrix GeneChip YG S98 microarrays was performed as previously described (Piper *et al.*, 2002). Chip quality control, condensing probe intensities to gene expression levels and normalization was performed using GeneData Refiner Array. The RMA algorithm was used to derive the log2 scale measure of the expression levels (Irizarry *et al.*, 2003). Quantile normalization was applied to normalize between arrays (Bolstad *et al.*, 2003).

2.2 Inferring the influence of cultivation parameters on gene expression

A design matrix was created, containing both main (or single) effects and interaction (or combinatorial) effects: each category within each cultivation parameter type is represented by a binary indicator column with 170 entries. These columns represent the main effects, which indicate, for each array, under which category of a particular cultivation parameter type, the yeast was grown. Interaction effect columns were obtained by applying the logic AND function to all possible pair-wise combinations of main effect columns. Redundant columns and columns containing only zeros were removed, resulting in 112 columns, of which 37 represent main effects and 75 represent interaction effects. These data are stored in the binary $[A\times C]$ design matrix $\mathbf{D}$. Here, A equals 170 and is the number of arrays. C equals 112 and is the number of (combinatorial) cultivation parameters.

A forward stepwise ordinary least squares regression strategy was applied to each gene individually:

$$\mathbf{y}=\mathbf{X}\theta+\epsilon \qquad (1)$$

Here, $\mathbf{y}_i$ denotes the measured gene expression level of a particular gene for array i, with $i=1,\ldots,A$; $\mathbf{X}$ is the predictor matrix, θ represents the regression coefficients and ϵ the error, which is assumed to be independent zero-mean normally distributed. Initially, $\mathbf{X}$ contains only the intercept, i.e. a column of A ones. In an iterative fashion, columns from $\mathbf{D}$ are added to $\mathbf{X}$. For this we apply a leave-one-out cross validation (LOOCV) scheme, where a single sample is used for testing, while the remaining $(A-1)$ samples are used for training the regression model. This is repeated such that each sample is used once as test data. The column from $\mathbf{D}$, with the smallest root-mean-squared (RMS) LOOCV error and absolute regression coefficient larger than one, is selected and added. The iterative process of adding columns is discontinued when the P-value, as output by a t-test that determines whether the regression coefficient significantly differs from zero, exceeds $0.05/C$. To prevent the inclusion of spurious combinatorial effects, the following strategy is applied: when a combinatorial effect column is selected, we check whether the addition in explained variance is larger than the addition in explained variance when adding the two main effect columns that constitute the combinatorial effect. Only in the cases where this is true, we add the combinatorial effect column. Otherwise the two main effect columns are added, provided that they satisfy the P-value threshold and their absolute regression coefficients are larger than one.

Note that only coefficients larger than 1 or smaller than -1 are allowed. In terms of the absolute expression measure, this means we only take into

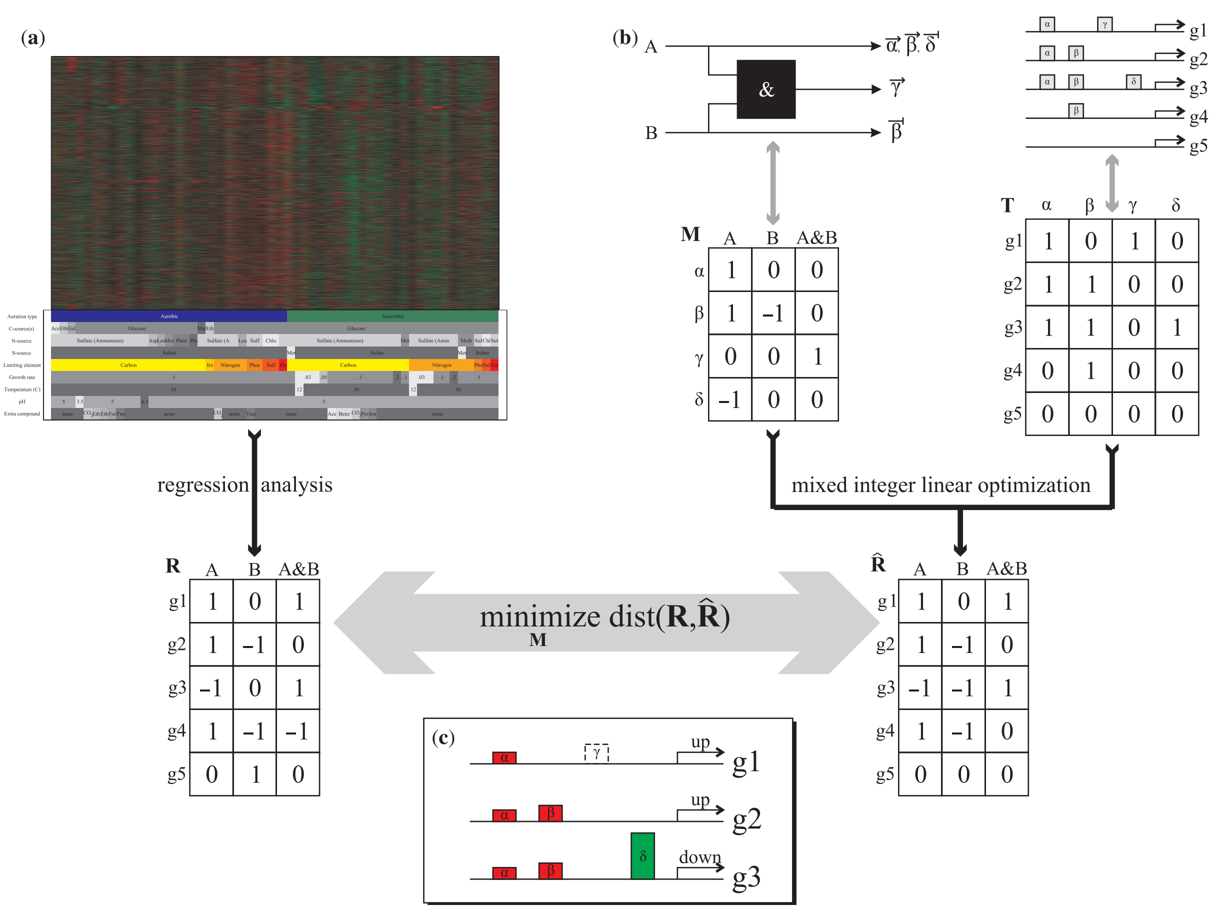

Fig. 2. Schematic overview of the approach. The goal is to build $\widehat{\mathbf{R}}$, the optimal approximation of the discretized regression coefficients in **R**. (**a**) The coefficients in **R** are derived from a regression analysis, which assesses the influence of cultivation parameters on gene expression by employing these parameters as predictors in the regression model. The discretization procedure maps non-zero regression weights to 1 or -1, depending on their sign. (The schematic representation of **R** is given for five genes and three cultivation parameters.) (**b**) The elements of $\widehat{\mathbf{R}}$ are determined by **T** and **M**. **T** is fixed and indicates binary TF binding potential to gene promoters. The elements of **M** are estimated and indicate the activity of TFs as enhancers or repressors under the different (combinatorial) cultivation parameters. A logic circuit derived from **M** is graphically depicted above the representation of **M**. (**c**) Visualization of the active TFs on the gene promoters of genes g1, g2 and g3 under cultivation parameter A. Enhancers are depicted as red boxes; repressors are depicted as green boxes. (TF γ can bind the promoter of g1, but is not active under A.) The height of a box indicates the enhancer or repressor strength. The strength of a particular enhancer or repressor is the same for all genes. A gene is upregulated when its activator strength, i.e. the sum of the heights of the red boxes, is larger than the repressor strength, which equals the sum of the heights of the green boxes. Downregulation is inferred in the opposite situation. See text for details.

account expression differences of 1-fold change or more. (The expression data are on log2 scale.) So, we focus on the cases where a cultivation parameter has a large influence on expression.

Finally, the regression coefficients for all yeast genes are discretized and put in **R** $([G\times C]\in\mathbb{Z}[-1,0,1])$, where G is the number of yeast genes. The discretization procedure maps positive coefficients to 1 and negative coefficients to -1. **R** is quite sparse since for most of the genes only two or three columns from **D** were selected as significant predictors.

2.3 TF binding data

For 111 TFs we extracted their known regulatory sites from TRANSFAC (Wingender *et al.*, 2000) and ChIP–chip data (Harbison *et al.*, 2004; MacIsaac *et al.*, 2006) (no conservation, binding P-value cutoff 0.001). These gene–TF pairs were put in the binary $[G\times F]$ TF-binding matrix **T**, where 1 indicates that a TF can bind a gene promoter. F equals 111 and is the number of TFs.

2.4 Inferring TF activity and TF strengths

The goal of our optimization problem is to infer the activity of TFs as a function of cultivation parameters, such that we can optimally explain the regression coefficients, which were distilled from the observed gene expression data. These TF activities form tertiary matrix **M** $([F\times C]\in\mathbb{Z}[-1,0,1])$. A non-zero element in **M** indicates that a TF is activated under a cultivation parameter and either acts as an enhancer (1) or a repressor (-1). Other data used in the optimization problem are: TF binding matrix **T** $([G\times F]\in\mathbb{Z}[0,1])$, discretized regression coefficient matrix **R** $([G\times C]\in\mathbb{Z}[-1,0,1])$ and its reconstructed version $\widehat{\mathbf{R}}$ $([G\times C]\in\mathbb{Z}[-1,0,1])$. First, from the tertiary matrix $\widehat{\mathbf{R}}$ two binary matrices with the same dimensions, $\widehat{\mathbf{R}}^+$ and $\widehat{\mathbf{R}}^-$, are derived. $\widehat{\mathbf{R}}^+$ has non-zero entries, where $\widehat{\mathbf{R}}$ contains 1s, and thus indicates the elements, where genes are upregulated under influence of a particular cultivation parameter. $\widehat{\mathbf{R}}^-$ has non-zero entries, where $\widehat{\mathbf{R}}$ contains -1s, and thus indicates the downregulated elements. A similar procedure is undertaken for tertiary matrix **M**, thereby obtaining $\mathbf{M}^+$, which contains the active enhancers and $\mathbf{M}^-$, which contains the active repressors. Now, all variables consist of binary integers (and are restricted to remain binary integers).

The objective function for the optimization problem is as follows:

$$\text{minimize} \sum_{\forall g,c \in I^+} [\mathbf{R}_{gc} - \widehat{\mathbf{R}}^+_{gc}] + \sum_{\forall g,c \in I^-} [-\mathbf{R}_{gc} - \widehat{\mathbf{R}}^-_{gc}] + \\ + \lambda \sum_{f=1}^{F} \sum_{c=1}^{C} [\mathbf{M}^+_{f,c} + \mathbf{M}^-_{f,c}] \qquad (2)$$

where I^+ is the set of index pairs referring to the elements where $\mathbf{R}$ is 1, and similarly, I^- refers to the negative elements of $\mathbf{R}$. Thus, we only try to explain the non-zero elements of $\mathbf{R}$, which represent the large expression changes due to the influence of the cultivation parameters. The zero elements of $\mathbf{R}$ do not only contain cases where there is no change in expression, but they contain the whole spectrum of no change in expression up to moderately large changes in gene expression. Therefore, we do not want to enforce TFs to be deactivated because of these zero elements. The last term of Equation (2) restricts the model complexity by penalizing the number of activated TFs. Parameter λ can be interpreted as the number of non-zero elements in $\mathbf{R}$ that a TF needs to help explain in order for it to be activated. Below, the constraints of the optimization problem are stated. These constraints are linear in $\mathbf{M}^+$, $\mathbf{M}^-$, $\widehat{\mathbf{R}}^+$ and $\widehat{\mathbf{R}}^-$, which are the variables in the system. In the appendix a detailed explanation for constraints **c5**, **c8** and **c12** is given.

The first two constraints are straightforward. Constraint **c1** states that a TF cannot be an active repressor and an active enhancer at the same time. Constraint **c2** states that a gene cannot be upregulated and downregulated at the same time.

$$\textbf{c1:} \quad \mathbf{M}^+_{fc} + \mathbf{M}^-_{fc} \leq 1 \qquad \forall f,c$$

$$\textbf{c2:} \quad \widehat{\mathbf{R}}^+_{gc} + \widehat{\mathbf{R}}^-_{gc} \leq 1 \qquad \forall g,c$$

Constraint **c3** states that if there is at least one active enhancer in a gene promoter, i.e. the inner product is positive then the gene *can be* upregulated, i.e. the regression coefficient can be 1. Constraint **c4** is the analogue constraint for the case of active repressors. Constraint **c5** forces a gene to be either upregulated or downregulated, when there is at least one active enhancer or one active repressor in the gene promoter.

$$\textbf{c3:} \quad \langle \mathbf{T}_{g\cdot}, \mathbf{M}^+_{\cdot c} \rangle \geq \widehat{\mathbf{R}}^+_{gc} \qquad \forall g,c$$

$$\textbf{c4:} \quad \langle \mathbf{T}_{g\cdot}, \mathbf{M}^-_{\cdot c} \rangle \geq \widehat{\mathbf{R}}^-_{gc} \qquad \forall g,c$$

$$\textbf{c5:} \quad \langle \mathbf{T}_{g\cdot}, \mathbf{M}^+_{\cdot c} \rangle + \langle \mathbf{T}_{g\cdot}, \mathbf{M}^-_{\cdot c} \rangle \leq F \cdot (\widehat{\mathbf{R}}^-_{gc} + \widehat{\mathbf{R}}^+_{gc}) \qquad \forall g,c$$

To decide upon upregulation or downregulation when multiple active enhancers and repressors bind a promoter, four continuous variables are introduced: $\mathbf{S}^+$ and $\mathbf{S}^-$; both ($[F \times C] \in \mathbb{R}[0,F]$) and $\widetilde{\mathbf{S}}^+$ and $\widetilde{\mathbf{S}}^-$; both ($[F \times 1] \in \mathbb{R}[1,F]$). $\mathbf{S}^+_{fc}$, represents the 'strength' of TF f as an enhancer under cultivation parameter c. $\mathbf{S}^+_{fc}$ is zero when $\mathbf{M}^+_{fc}$ is zero, i.e. when f is not activated as an enhancer under c. This rule is stated in constraint **c6**. $\mathbf{S}^+_{fc}$ equals the generic TF strength for f, $\widetilde{\mathbf{S}}^+_f$, when $\mathbf{M}^+_{fc}$ is one. Thus, the strength of a TF f is the same for all genes under the cultivation parameters, where the gene is activated (and zero otherwise). This rule is stated in constraints **c7** and **c8**. Analogue rules apply for $\mathbf{S}^-$ and $\widetilde{\mathbf{S}}^-$. The corresponding constraints **c9**, **c10** and **c11** are omitted for brevity.

$$\textbf{c6:} \quad \mathbf{S}^+_{fc} \leq F \cdot \mathbf{M}^+_{fc} \qquad \forall f,c$$

$$\textbf{c7:} \quad \mathbf{S}^+_{fc} \leq \widetilde{\mathbf{S}}^+_f \qquad \forall f,c$$

$$\textbf{c8:} \quad \mathbf{S}^+_{fc} - \widetilde{\mathbf{S}}^+_f \geq F \cdot (\mathbf{M}^+_{fc} - 1) \qquad \forall f,c$$

Constraint **c12** states that when the sum of the strengths of active enhancers that bind a gene promoter is larger than its repressing counterpart, the gene is upregulated. Constraint **c13** encloses the reverse scenario. Note that if an identical set of enhancers and repressors is active on a promoter, this will lead to the same reconstructed regression coefficient for any gene and under any cultivation parameter.

$$\textbf{c12:} \quad \langle \mathbf{T}_{g\cdot}, \mathbf{S}^+_{\cdot c} \rangle - \langle \mathbf{T}_{g\cdot}, \mathbf{S}^-_{\cdot c} \rangle \geq (F^2 + F^{-2}) \cdot \widehat{\mathbf{R}}^+_{gc} - F^2 \qquad \forall g,c$$

$$\textbf{c13:} \quad \langle \mathbf{T}_{g\cdot}, \mathbf{S}^-_{\cdot c} \rangle - \langle \mathbf{T}_{g\cdot}, \mathbf{S}^+_{\cdot c} \rangle \geq (F^2 + F^{-2}) \cdot \widehat{\mathbf{R}}^-_{gc} - F^2 \qquad \forall g,c$$

The optimization problem is implemented within the MATLAB environment and executed using the MOSEK optimization toolbox with standard settings for mixed integer optimization. Given constraints **c1** to **c13**, MOSEK estimates variables $\mathbf{M}^+$, $\mathbf{M}^-$, $\widehat{\mathbf{R}}^+$, $\widehat{\mathbf{R}}^-$, $\mathbf{S}^+$, $\mathbf{S}^-$, $\widetilde{\mathbf{S}}^+$ and $\widetilde{\mathbf{S}}^-$ such that the optimization function in Equation (2) is minimized.

3 RESULTS

3.1 TF activity in response to changes in oxygen and carbon presence

The regulatory network inference algorithm is run on a subset of the data. In particular, we focus on oxygen and carbon; two environmental factors, which have a large and well studied effect on the transcriptional program of *S.cerevisiae*. Four cultivation parameters are selected, i.e. aeration type, carbon-limitation and the combinatorial cultivation parameters, carbon-limited aerobic growth and carbon-limited anaerobic growth. Note that aeration type is actually a cultivation parameter type that assumes two values, i.e. aerobic growth and anaerarobic growth. Since these are mutually redundant, only aerobic growth was included in the regression model and subsequent optimization algorithm. (Downregulation under aerobic growth and upregulation under anaerobic growth are interchangeable.) There are 40 genes, which are influenced by at least two of these four cultivation parameters, i.e. there are 40 rows in $\mathbf{R}$ with at least two non-zero elements in the four columns of interest. These 40 genes are bound by 46 different TFs. In this experiment λ is set to two. The algorithm correctly inferred the regression coefficients of 58 of the 84 (70%) non-zero elements in $\mathbf{R}$. A particularly large concentration of incorrectly predicted values appears toward the bottom of $\widehat{\mathbf{R}}$, where zeros are predicted while the true expression coefficients are non-zero. See Figure 3d. This stems from the fact that the promoters of these genes have almost no motifs for the activated TFs, in which case the model cannot explain the up- or downregulation.

3.1.1 Inferred TF activity In total, nine different TFs were activated across the four cultivation parameters, some under more than one cultivation parameter. Three of these TFs, HAP1, HAP2/3/4 and ROX1, have a significantly larger strength, when compared to the others. See Figure 3a, b. The large strength indicates their dominating effect on transcriptional regulation. If one of these TFs is active and binds the promoter, it will determine the direction of transcriptional regulation. For example under aerobic conditions (Aer) the promoter of gene *PAU3* (the tenth gene from the bottom in Fig. 3c) is bound by one active enhancer, i.e. YAP7, and one active repressor, i.e. ROX1. Since the repressor strength of ROX1 is (much) larger than the enhancer strength of YAP7, the gene is (correctly) predicted to be downregulated. Interestingly, in the resulting network for this data, the TF strength of ROX1 equals 45.9995, which is very close to the maximum value of 46, the number of TFs F. However, this number is slightly smaller than

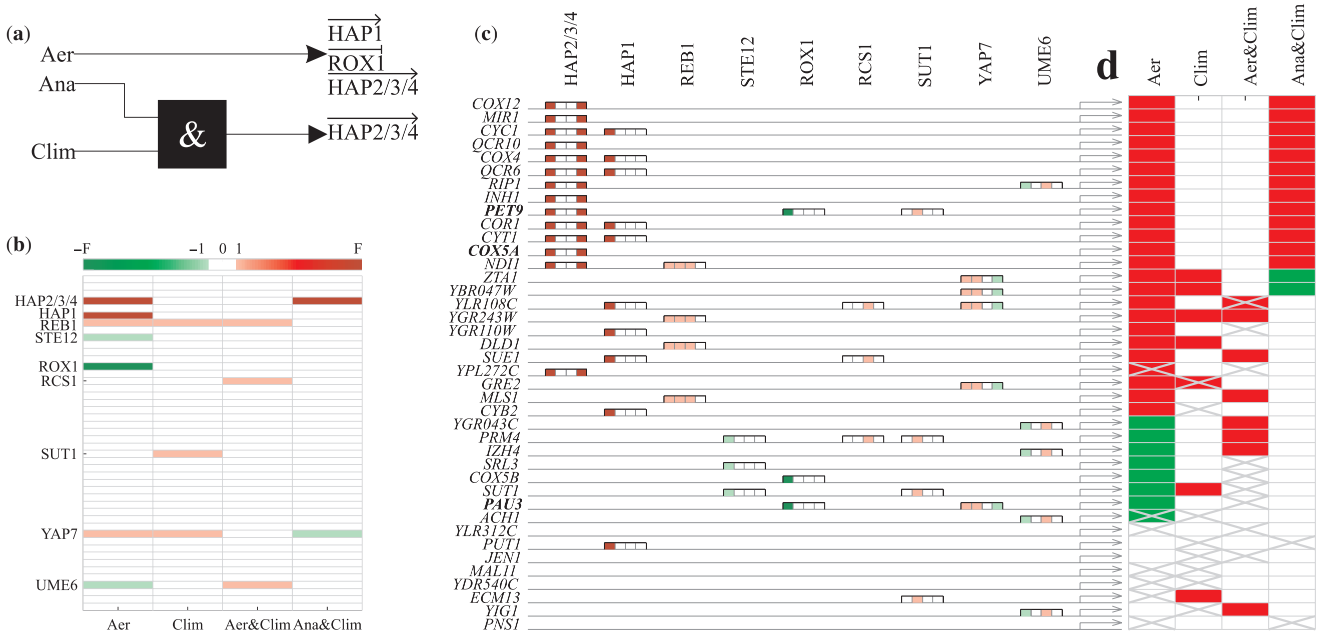

Fig. 3. Overview of the results obtained for the oxygen and carbon limitation data. (**a**) Inferred influence of cultivation parameters aerobic growth (Aer), anaerobic growth (Ana) and carbon limitation (Clim) on TF activity. Only the three dominating TFs are reported. (**b**) Representation of **S**, indicating the strength of the activated TFs under each of the four cultivation parameters. Enhancers are depicted in red; repressors are depicted in green. (**c**) Representation of **T**, indicating which gene promoters can be bound by the activated TFs. The enhancer or repressor strengths for the four cultivation parameters are visualized by the colored blocks inside the rectangle that represents a binding site. (**d**) Representation of $\widehat{\mathbf{R}}$, indicating the inferred regression coefficients. Upregulation is indicated by red; downregulation is indicated by green. Incorrectly inferred elements are marked with a gray cross. White boxes without a cross are the zero elements of **R**. These elements are not part of the optimization scheme.

the strength of HAP2/3/4 which has the maximal strength of 46. This difference can be attributed to gene *PET9* (the ninth gene from the top in Fig. 3c). Both HAP2/3/4 and ROX1 can bind the *PET9* promoter. To ensure that this gene is upregulated when grown aerobically, as was deduced from the regression analysis, the active enhancers should have a larger strength than the active repressors. Therefore, the strength of ROX1 is set a bit smaller than the strength of HAP2/3/4, however, still large enough to dominate other active enhancers.

3.1.2 Regulation of gene expression by oxygen The role of the three dominant TFs in the regulation of gene expression by oxygen is widely reported in the literature. Both HAP1 and the HAP2/3/4 complex activate genes in response to heme, which is synthesized only in the presence of oxygen (Zitomer and Lowry, 1992). TF ROX1 is needed for the repression of hypoxic or heme-repressed genes under aerobic conditions (Lowry and Zitomer, 1988). Also, the relation between carbon source and the HAP2/3/4 complex has been investigated. The HAP2 and HAP3 proteins enable DNA binding of the complex, whereas HAP4 contains the transcriptional activation domain. The synthesis of the activator subunit HAP4 is regulated by the carbon source. More specifically, the expression of HAP4 is repressed by glucose, *S.cerevisiae*'s preferred carbon source (Forsburg and Guarente, 1989). Tai *et al.* (2005) reports that HAP4 mRNA is present in carbon-limited cultivations even under anaerobic conditions, where HAP4 has no obvious role. We can corroborate and even further substantiate these findings with the observation that the HAP4 protein is an activator under carbon-limited anaerobic conditions. Note that all genes, which are upregulated under carbon-limited anaerobic growth are also upregulated under aerobic growth. See the top 13 genes in Figure 3c. The expression profile of one of these genes, *COX5A*, across all conditions is depicted in Figure 1. This expression profile is typical for all the 13 members of this group. It shows that these genes are most highly expressed when grown aerobically. Yet, in the anaerobic case, where the expression is in general lower, these genes show different expression behavior in carbon-limited growth compared to other nutrient limitations. That is, these genes have a higher expression level in carbon-limited cultivations, where there is hardly any glucose, compared to the situation, where glucose is abundant.

Also, for the other TFs, which are activated according to the inference algorithm, evidence is found in literature. For example REB1, which acts as an enhancer under three cultivation parameters, is a RNA polymerase I enhancer binding protein as well as an activator for many genes transcribed by RNA polymerase II (Ju *et al.*, 1990). STE12 is known to activate genes associated with pseudohyphal (low oxygen) growth (Norman *et al.*, 1999). SUT1 is reported to encode a glucose transporter (Weierstall *et al.*, 1999), however SUT1 also has a putative role in the regulation of some hypoxic genes (Regnacq *et al.*, 2001). In general, the precise regulatory role of these TFs in (an)aerobiosis and response to the carbon source is not known. The results of this analysis provide hints for elucidating the regulatory mechanisms of these factors.

3.1.3 Setting λ Parameter λ, which restricts the model complexity by penalizing the number of activated TFs, is chosen using a 5-fold CV scheme. The genes are divided into five parts, where consecutively four parts are used for training and one part is

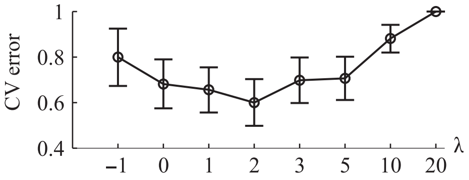

Fig. 4. CV errors for different values of λ.

used for testing. The **M** and **S** matrices, which are computed on the training set, are applied to the test set to obtain the reconstructed regression coefficients for the test set, $\widehat{\mathbf{R}}^{\text{test}}$. The error on the test set is defined as:

$$Err = \frac{1}{J} \sum_{\forall g,c \in I} \left| \mathbf{R}^{\text{test}}_{gc} - \widehat{\mathbf{R}}^{\text{test}}_{gc} \right| \qquad (3)$$

where I is the set of index pairs referring to the non-zero elements of $\mathbf{R}^{\text{test}}$ and J the number of these non-zero elements. The CV scheme is repeated 10 times. Figure 4 depicts the average error over all CV runs. For small values of λ, many TFs are activated in order to approximate the regression coefficients. Clearly, this strategy is prone to overfitting, which is also illustrated by the large CV error. For large values of λ, activating a TF is severely penalized, such that only a few TFs will be activated. (For $\lambda = 20$, no TF is activated and every element of $\widehat{\mathbf{R}}^{\text{test}}$ is zero). The high CV error in this case, indicates that a lot of true regulation is missed. The optimal λ will be found between these extremes. In this experiment, $\lambda = 2$ led to the smallest CV error and was therefore selected.

3.2 Transcriptional regulation of nitrogen metabolism

Across the conditions of the compendium, yeast was grown on six different nitrogen sources. This inspired the second experiment, where we analyzed the transcriptional regulation of the genes that comprise the nitrogen compound metabolism category of GO biological processes (Ashburner *et al.*, 2000). A total of 119 of these genes are influenced by at least one cultivation parameter and bound by one of 78 different TFs. In total, there are 68 cultivation parameters that cause up- or downregulation of at least one of these 119 genes. The resulting transcription regulation network (with $\lambda_{\text{opt}} = 2$) revealed the activation of 14 different TFs under 28 different cultivation parameters, of which 11 are combinatorial. Figure 5 depicts the network for the cultivation parameters, which are most straightforwardly related to nitrogen metabolism, i.e. the different nitrogen sources, nitrogen as growth limiting element and combinatorial effects involving these cultivation parameters.

The six different nitrogen sources can be dichotomized into preferred and non-preferred nitrogen sources. The preferred nitrogen sources are asparagine (Asn) and ammonium [in ammonium sulfate (AS)]. Proline (Pro), phenylalanine (Phe), methionine (Met) and leucine (Leu) are non-preferred (or poor) nitrogen sources (Boer *et al.*, 2007; Magasanik and Kaiser, 2002). In *S.cerevisiae*, the use of nitrogen sources is controlled by a transcriptional regulation mechanism known as nitrogen catabolite repression (NCR). When a good nitrogen source is present, NCR shuts down the pathways for the use of poor nitrogen sources. NCR is mediated by a four-member family of GATA-binding TFs: GLN3, GAT1, DAL80 and GZF3 (Hofman-Bang, 1999). In the absence of a good nitrogen source, GLN3 is activated and in turn activates the transcription of NCR-sensitive genes. Indeed, for three of the four non-preferred nitrogen sources, GLN3 acts as an enhancer. When methionine is the nitrogen source, the MET31/32 complex is activated. This complex controls the biosynthesis of sulfur containing amino acids (Blaiseau *et al.*, 1997). (Methionine is also used as a sulfur source.) In the case of leucine, two additional TFs are activated; LEU3 and GCN4, the two key regulators in the regulation of branched-chain amino acid metabolism (Boer *et al.*, 2005). The inferred role of GCN4 as an activator in the presence of a poor nitrogen source and as a repressor in the presence of good nitrogen sources corroborates the work of Sosa *et al.* (2003). It further supports the fact that NCR is not solely achieved through the action of the abovementioned family of GATA factors, but conceivably also through GCN4.

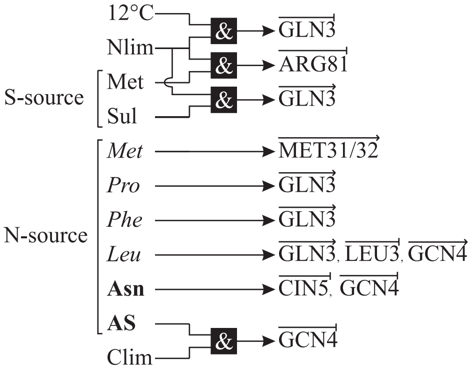

Fig. 5. Inferred TF activity derived from genes, which are involved in nitrogen metabolism. Preferred nitrogen sources are printed in bold; non-preferred nitrogen sources are printed in italic style. Abbreviations for the nitrogen and sulfur sources are explained in the text.

3.2.1 Missing and dubious TF activity Remarkably, the other three GATA factors, GAT1, DAL80 and GZF3, are not part of generated network. Inspection of the TF binding data in the promoters of the 119 nitrogen metabolism genes revealed that GAT1, DAL80 and GZF3 bind only 3, 4 and 0 genes, respectively. This could indicate that their targets are not transcriptionally regulated under the influence of the cultivation parameters. However, this observation should also be related to the ChIP–chip data. From TRANSFAC, we extracted many TF–gene pairs, which are not present in the ChIP–chip data. This indicates that not all TF targets are detected by the ChIP–chip experiments. Furthermore, Gao *et al.* (2004) estimate that 40% of the ChIP–chip TF targets are non-functional. Obviously, this complicates regulatory network inference. Another dubious result was identified when analyzing the cases in which two or more TFs were active on a promoter. In this experiment, there are 72 such cases, of which 10 are unique. Amongst the most frequent cases, we found the combinatorial regulation of TFs, which have already been reported in literature, e.g. the interplay between LEU3 and GCN4 (Boer *et al.*, 2005) and that of CBF1 and GCN4 (O'Connell *et al.*, 1995). Also, GLN3 and GCN4 were found activated together in a set of nine gene promoters. These nine genes were upregulated under two cultivation parameters, i.e. sulfur limitation and zinc limitation, where both GLN3 and GCN4 are enhancers. However, under another cultivation parameter, i.e. where leucine is used as a nitrogen source, the *same* genes were *downregulated*, where now GLN3 acts as a repressor (which is stronger than enhancer GCN4). These results seem

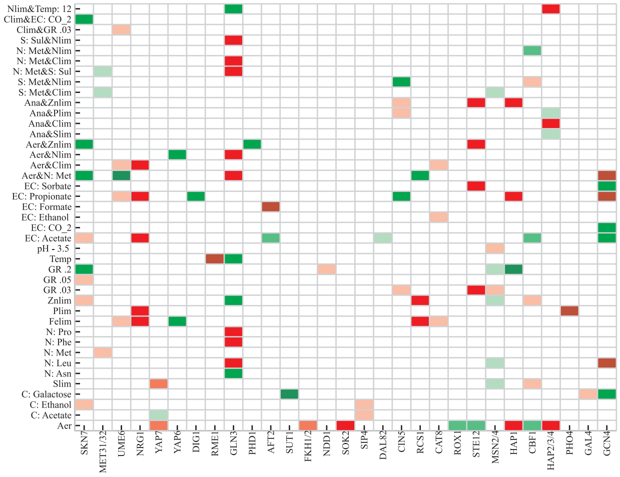

Fig. 6. Representation of **S** for the regulatory program inferred using the compendium. Color coding is identical to Figure 3b.

implausible and imply that this regulation pattern should involve another TF, which might not be present in the employed TF binding data set. Preliminary experiments with artificial datasets have demonstrated that especially missing TFs (simulated by removing colums from **T**) can have a large negative effect on the ability to reconstruct the correct regulatory network (results not shown).

3.3 Compendium analysis

The algorithm was also run on the complete compendium for all genes that are up- or downregulated under at least two cultivation parameters and for all cultivation parameters that influence the expression of at least 10 genes ($G=766, C=67, F=101, \lambda_{\text{opt}}=5$). In the resulting regulatory network, 41 (61%) cultivation parameters activated at least one of the TFs, resulting in 29 (29%) different activated TFs in total. See Figure 6. Network inference on the complete dataset allows for a more rigorous and unbiased estimation of the regulatory program. It reveals confounding factors, with respect to the previously discussed programs, which were based on a subset of the data. For example, the regulatory program of GATA factor GLN3, as discussed before, is also depending on other (combinatorial) cultivation parameters, e.g. zinc limitation and nitrogen limitation at low temperature. These results offer interesting leads, however the combinatorial regulation of TFs, as inferred by this analysis, becomes complicated. There are up to four active TFs on gene promoters. This calls for an automated procedure that uses these inferred TF activities and accompanying strengths to derive logic rules, in which the influence of multiple TFs on transcriptional regulation is formalized.

Note that the inference algorithm was run on a selection of genes and cultivation parameters. The number of variables and constraints in optimization problem is $4FC+2GC$ and $7FC+6GC$, respectively, which becomes quite large for the complete dataset. It is yet unclear (due to computation time) if converge is reached for the dataset with all genes and cultivation parameters.

4 DISCUSSION

The transcriptional program of a cell is largely determined by its extracellular environment. The accurate measurement of environmental parameters, e.g. with chemostat cultures, have inspired several approaches that analyze the (combinatorial) effect of environmental parameters on gene expression. In this study, we have, for the first time demonstrated how environmental parameters can be employed to derive transcriptional regulation networks. In these networks, the cultivation parameters form the signals that trigger the activation or deactivation of TFs. Since many TFs are regulated post-transcriptionally, this approach seems more natural than the often employed strategy of deducing the TF activity from the mRNA expression of TFs. The inference algorithm was translated into a linear optimization problem, solvable without having to rely on greedy and/or heuristic search strategies.

The combinatorial regulatory code of multiple TFs that are able to bind a promoter, is modeled using the linearly weighted sum of inferred enhancers and repressor strengths. Previous approaches have also modeled gene expression as a linearly weighted sum of TF contributions, e.g. Gao *et al.* (2004). The main improvement of our method is the fact that the activity of TFs can be explicitly turned on or off, and that the inference algorithm optimizes this choice with respect to the direction of regulation, i.e whether a gene is up- or downregulated. This strategy enables the inference of combinatorial effects between TFs. For example, a repressor, which interacts directly with the TATA binding protein, thereby completely blocking transcription independent of the possible enhancers that bind the promoter, would acquire a strength that is larger than the sum of the strengths of all enhancers that can bind the promoter. Thus, the repressor, when active, will cause downregulation of the gene, thereby nullifying the influence of the enhancers. This is in contrast with the linear regression strategies, where these enhancers would still have influence on the gene expression level.

Additional validation experiments indicate that more pairs of TFs, which are simultaneously active according to our approach, are found to co-occur in PubMed abstracts when compared to TF pairs uncovered with Gao *et al.* (2004). This difference can be attributed to the fact that we decompose the expression in terms of cultivation parameters, and analyze these cultivation parameters separately. When using only the expression data itself, some cultivation parameters (such as aeration type) can have a much larger influence than others, thereby dominating the expression pattern and thus controlling which TFs are found to be the most significant, leading to less diversity in activated TFs (and thus fewer TF pairs). An overview of this comparison is published as Supplementary Material.

A future challenge lies in the integral interpretation of the inferred regulatory networks, which must be accompanied by a computational approach that derives logic rules, which are able to describe the interplay of multiple TFs on gene promoters.

ACKNOWLEDGEMENTS

Funding: T.A.K. and M.J.T.R. are part of the Kluyver Centre for Genomics of Industrial Fermentation, which is supported by the Netherlands Genomics Initiative (NGI). L.F.A.W. is part of the Cancer Genomics Centre, which is supported by the Netherlands Genomics Initiative (NGI).

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Balaji,S. *et al.* (2006) Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *J. Mol. Biol.*, **360**, 213–227.

Bammler,T. *et al.* (2005) Standardizing global gene expression analysis between laboratories and across platforms. *Nat. Methods*, **2**, 351–356.

Bar-Joseph,Z. *et al.* (2003) Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.*, **21**, 1337–1342.

Blaiseau,P.L. *et al.* (1997) Met31p and met32p, two related zinc finger proteins, are involved in transcriptional regulation of yeast sulfur amino acid metabolism. *Mol. Cell Biol.*, **17**, 3640–3648.

Boer,V.M. *et al.* (2005) Contribution of the saccharomyces cerevisiae transcriptional regulator leu3p to physiology and gene expression in nitrogen- and carbon-limited chemostat cultures. *FEMS Yeast Res.*, **5**, 885–897.

Boer,V.M. *et al.* (2007) Transcriptional responses of saccharomyces cerevisiae to preferred and nonpreferred nitrogen sources in glucose-limited chemostat cultures. *FEMS Yeast Res.*, **7**, 604–620.

Bolstad,B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Bonneau,R. *et al.* (2006) The inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol.*, **7**, R36.

Bussemaker,H.J. *et al.* (2001) Regulatory element detection using correlation with expression. *Nat. Genet.*, **27**, 167–171.

Castrillo,J. *et al.* (2007) Growth control of the eukaryote cell: a systems biology study in yeast. *J. Biol.*, **6**, 4.

Chang,Y.-H. *et al.* (2006) Identification of transcription factor cooperativity via stochastic system model. *Bioinformatics*, **22**, 2276–2282.

Das,D. *et al.* (2004) Interacting models of cooperative gene regulation. *Proc. Natl Acad. Sci. USA*, **101**, 16234–16239.

Forsburg,S.L. and Guarente,L. (1989) Identification and characterization of hap4: a third component of the ccaat-bound hap2/hap3 heteromer. *Genes Dev.*, **3**, 1166–1178.

Gao,F. *et al.* (2004) Defining transcriptional networks through integrative modeling of mrNA expression and transcription factor binding data. *BMC Bioinformatics*, **5**, 31.

Harbison,C.T. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

Hofman-Bang,J. (1999) Nitrogen catabolite repression in saccharomyces cerevisiae. *Mol. Biotechnol.*, **12**, 35–73.

Irizarry,R.A. *et al.* (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Ju,Q.D. *et al.* (1990) Reb1, a yeast DNA-binding protein with many targets, is essential for growth and bears some resemblance to the oncogene myb. *Mol. Cell Biol.*, **10**, 5226–5234.

Knijnenburg,T.A. *et al.* (2007) Exploiting combinatorial cultivation conditions to infer transcriptional regulation. *BMC Genomics*, **8**, 25.

Lowry,C.V. and Zitomer,R.S. (1988) Rox1 encodes a heme-induced repression factor regulating anb1 and cyc7 of saccharomyces cerevisiae. *Mol. Cell Biol.*, **8**, 4651–4658.

Luscombe,N.M. *et al.* (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, **431**, 308–312.

MacIsaac,K.D. *et al.* (2006) An improved map of conserved regulatory sites for saccharomyces cerevisiae. *BMC Bioinformatics*, **7**, 113.

Magasanik,B. and Kaiser,C.A. (2002) Nitrogen regulation in saccharomyces cerevisiae. *Gene*, **290**, 1–18.

Myers,C.L. and Troyanskaya,O.G. (2007) Context-sensitive data integration and prediction of biological networks. *Bioinformatics*, **23**, 2322–2330.

Nguyen,D.H. and D'haeseleer,P. (2006) Deciphering principles of transcription regulation in eukaryotic genomes. *Mol. Syst. Biol.*, **2**, 2006.0012.

Norman,T.C. *et al.* (1999) Genetic selection of peptide inhibitors of biological pathways. *Science*, **285**, 591–595.

O'Connell,K.F. *et al.* (1995) Role of the saccharomyces cerevisiae general regulatory factor cp1 in methionine biosynthetic gene transcription. *Mol. Cell Biol.*, **15**, 1879–1888.

Piper,M.D.W. *et al.* (2002) Reproducibility of oligonucleotide microarray transcriptome analyses. An interlaboratory comparison using chemostat cultures of saccharomyces cerevisiae. *J. Biol. Chem.*, **277**, 37001–37008.

Regenberg,B. *et al.* (2006) Growth-rate regulated genes have profound impact on interpretation of transcriptome profiling in saccharomyces cerevisiae. *Genome Biol.*, **7**, R107.

Regnacq,M. *et al.* (2001) Sut1p interaction with cyc8p(ssn6p) relieves hypoxic genes from cyc8p-tup1p repression in saccharomyces cerevisiae. *Mol. Microbiol.*, **40**, 1085–1096.

Segal,E. *et al.* (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.*, **34**, 166–176.

Sosa,E. *et al.* (2003) Gcn4 negatively regulates expression of genes subjected to nitrogen catabolite repression. *Biochem. Biophys. Res. Commun.*, **310**, 1175–1180.

Tai,S.L. *et al.* (2005) Two-dimensional transcriptome analysis in chemostat cultures. combinatorial effects of oxygen availability and macronutrient limitation in saccharomyces cerevisiae. *J. Biol. Chem.*, **280**, 437–447.

Tan,P.K. *et al.* (2003) Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.*, **31**, 5676–5684.

Weierstall,T. *et al.* (1999) Cloning and characterization of three genes (sut1-3) encoding glucose transporters of the yeast pichia stipitis. *Mol. Microbiol.*, **31**, 871–883.

Wingender,E. *et al.* (2000) Transfac: an integrated system for gene expression regulation. *Nucleic Acids Res.*, **28**, 316–319.

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

Yu,X. *et al.* (2006) Genome-wide prediction and characterization of interactions between transcription factors in saccharomyces cerevisiae. *Nucleic Acids Res.*, **34**, 917–927.

Zitomer,R.S. and Lowry,C.V. (1992) Regulation of gene expression by oxygen in saccharomyces cerevisiae. *Microbiol Rev.*, **56**, 1–11.

APPENDIX

A detailed explanation for constraints **c5**, **c8** and **c12** is given.

$$\textbf{c5:} \quad \langle \mathbf{T}_{g\cdot}, \mathbf{M}^{+}_{\cdot c} \rangle + \langle \mathbf{T}_{g\cdot}, \mathbf{M}^{-}_{\cdot c} \rangle \leq F \cdot (\widehat{\mathbf{R}}^{-}_{gc} + \widehat{\mathbf{R}}^{+}_{gc}) \qquad \forall g, c$$

$\langle \mathbf{T}_{g\cdot}, \mathbf{M}^{+}_{\cdot c} \rangle$ is the inner product of row g from binary matrix $\mathbf{T}$ and column c from binary matrix $\mathbf{M}^{+}$ and indicates the number of active enhancers that binds a gene promoter. $\langle \mathbf{T}_{g\cdot}, \mathbf{M}^{-}_{\cdot c} \rangle$ indicates the number of active repressors in the promoter. The sum of these two terms is an integer between 0, when no active TFs bind the promoter, and F, when all TFs are activated and bind the promoter. The right side of constraint **c5** can be either 0, when both $\widehat{\mathbf{R}}^{-}_{gc}$ and $\widehat{\mathbf{R}}^{+}_{gc}$ are 0, or F, when one of the two equals 1. (Note that because of constraint **c2** the $\widehat{\mathbf{R}}$ coefficients cannot both be 1.) When the sum of the two inner products is zero, both $\widehat{\mathbf{R}}$ coefficients can be 0 or one of them can be 1, since $0 \leq 0$ and $0 \leq F$. However, if there is at least one active enhancer or repressor that binds the promoter, i.e. the sum of the inner products, denoted by x, is positive, then one of the $\widehat{\mathbf{R}}$ coefficients must be 1, since $x \nleq 0$ and only $x \leq F$ holds. Consequently, constraint **c5** forces a gene to be either upregulated or downregulated, when there is at least one active enhancer or one active repressor in the gene promoter.

$$\textbf{c8:} \quad \mathbf{S}^{+}_{fc} - \widetilde{\mathbf{S}}^{+}_{f} \geq F \cdot (\mathbf{M}^{+}_{fc} - 1) \qquad \forall f, c$$

Constraint **c6** ensures that $\mathbf{S}^{+}_{fc}$, the strength of TF f under cultivation parameter c, is 0, when the $\mathbf{M}^{+}_{fc}$ is 0, i.e. when f is not activated under c. In this case constraint **c8** becomes $-\widetilde{\mathbf{S}}^{+}_{f} \geq -F$, which is always satisfied, since $\widetilde{\mathbf{S}}^{+}_{f}$, the general enhancer strength of f is at most F. In the case that the TF is activated, i.e. $\mathbf{M}^{+}_{fc}$ is 1, constraint **c8** becomes $\mathbf{S}^{+}_{fc} \geq \widetilde{\mathbf{S}}^{+}_{f}$. Together with constraint **c7**, which states that $\mathbf{S}^{+}_{fc} \leq \widetilde{\mathbf{S}}^{+}_{f}$, it forces $\mathbf{S}^{+}_{fc}$ to be equal to $\widetilde{\mathbf{S}}^{+}_{f}$ in the case that $\mathbf{M}^{+}_{fc}$ is 1.

$$\textbf{c12:} \quad \langle \mathbf{T}_{g\cdot}, \mathbf{S}^{+}_{\cdot c} \rangle - \langle \mathbf{T}_{g\cdot}, \mathbf{S}^{-}_{\cdot c} \rangle \geq (F^{2} + F^{-2}) \cdot \widehat{\mathbf{R}}^{+}_{gc} - F^{2} \qquad \forall g, c$$

$\langle \mathbf{T}_{g\cdot}, \mathbf{S}^{+}_{\cdot c} \rangle$ is the inner product of row g from binary matrix $\mathbf{T}$ and column c from continuous matrix $\mathbf{S}^{+}$ and indicates the sum of the strengths of all active enhancers that can bind the promoter of g under cultivation parameter c. $\langle \mathbf{T}_{g\cdot}, \mathbf{S}^{-}_{\cdot c} \rangle$ indicates the total repressor strength. From constraints **c3**, **c4** and **c8** we know that if there are no active TFs that can bind the promoter, both inner products as well as $\widehat{\mathbf{R}}^{+}_{gc}$ are 0. In that case, constraint **c12** becomes $0 \geq -F^{2}$ and is satisfied. In the case that there is at least one active enhancer or repressor that binds the promoter, the difference between the inner products can range from $-F^{2}$, when all F TFs are active, bind the promoter and act as repressors with the maximal strength of F, to F^{2}, when the enhancer strength is at its maximum. If we want to call a gene upregulated, i.e. $\widehat{\mathbf{R}}^{+}_{gc}$ is 1, than constraint **c12** becomes: $\langle \mathbf{T}_{g\cdot}, \mathbf{S}^{+}_{\cdot c} \rangle - \langle \mathbf{T}_{g\cdot}, \mathbf{S}^{-}_{\cdot c} \rangle \geq F^{-2}$. Here, F^{-2} plays the role of a small positive number. Consequently, a gene can only be upregulated, i.e. $\widehat{\mathbf{R}}^{+}_{gc}$ can only be 1, when the enhancer strength is larger than the repressor strength.