BIOINFORMATICS

Vol. 18 Suppl. 1 2002
Pages S258–S267

Minreg: Inferring an active regulator set

Dana Pe'er[1], Aviv Regev[2, 3] and Amos Tanay[4]

[1]School of Computer Science & Engineering, Hebrew University of Jerusalem, [2]Department of Cell Research and Immunology, Life Sciences Faculty, Tel Aviv University, [3]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science and [4]School of Computer Science, Tel Aviv University

Received on January 24, 2002; revised and accepted on April 1, 2002

ABSTRACT

Regulatory relations between genes are an important component of molecular pathways. Here, we devise a novel global method that uses a set of gene expression profiles to find a small set of relevant active regulators, identify the genes that they regulate, and automatically annotate them. We show that our algorithm is capable of handling a large number of genes in a short time and is robust to a wide range of parameters. We apply our method to a combined dataset of *S. cerevisiae* expression profiles, and validate the resulting model of regulation by cross-validation and extensive biological analysis of the selected regulators and their derived annotations.

Keywords: gene expression; gene regulation; gene networks; machine learning.

INTRODUCTION

Pathways of interacting proteins and genes underlie the fundamental functions of living cells. A major promise of high-throughput methods, such as gene expression profiling (DeRisi *et al.*, 1997), is that they will enable us to reconstruct molecular pathways. This paper focuses on an important aspect of this task: Reconstruction of regulatory relations between genes using gene expression data. Previous attempts at network reconstruction have been restricted to *local* relations and focused on particular pathways (Friedman *et al.*, 2000; Pe'er *et al.*, 2001; Tanay and Shamir, 2001). Here, we attempt to reconstruct regulatory relations on a *global* scale, while building on the principles of local modelling of regulatory interactions developed by Friedman *et al.* (2000).

Since only a relatively small part of the genome is directly involved in transcriptional regulation, our approach focuses on a small subset of regulators. Accordingly, we define the following task: Given a set of *candidate regulators*, we wish to find a small sub-set of *active regulators*, which control the processes that take place in a given set of experiments, identify the genes that they regulate (their *regulatees*), and characterize both sets. Specifically, we search for genes which are both known to participate in regulation and are globally predictive of the expression of many other genes in the data set.

We formulate this task as a precise combinatorial optimization problem, devise an efficient approximation algorithm that solves it, and prove a lower bound on the quality of the solution, under certain weak assumptions. Our algorithm has several important features that allow us to use it on a global scale. First, our implementation is extremely fast and can provide a model over thousands of genes within minutes. Second, our model's parsimony provides statistical robustness, as demonstrated by its ability to correctly predict the expression values for entire arrays using only those of a small set of active regulators. Finally, our framework is general and can integrate additional data types including DNA binding locations (as shown below) and regulatory motifs.

We devise a method that uses the resulting model to automatically assign biological function to the discovered active regulators based on the properties of their regulatees. The analysis uses the Gene Ontology (The Gene Ontology Consortium, 2000) annotations and derives *p*-values for its predictions, facilitating the generation of hypotheses on both the biological process regulated by the gene and the molecular function employed in the regulation process. We further determine the logic (activation or inhibition) of this regulation.

We applied our approach to several expression datasets in yeast (Hughes *et al.*, 2000; Gasch *et al.*, 2000; Spellman *et al.*, 1998), as well as DNA binding data (Simon *et al.*, 2001). First, we correctly predicted the function of many known regulators both at the level of the biological process and the molecular functions they regulate. Second, we assigned detailed functional annotation to previously uncharacterized transcription factors and signaling molecules. Third, we investigated individual regulator-regulatee relations in detail, finding evidence that our model can capture transcriptional and post-translational regulation. Finally, we incorporated binding data into our model and achieved global improvement for genes both with and without binding information.

REGULATION MODEL

In order to detect a small sub-set of active regulators and their respective regulatees, we first model the *local* relation between a regulatee and its active regulators and evaluate it based on their mutual information score. We then seek a collection of such relations that optimizes an overall score, while adhering to certain *global* constraints. We apply two constraints on our regulating genes: We use prior biological information to limit our search to a set of candidate regulators; and we require that the union of active regulator sets for all regulatees be of small cardinality.

Local regulation model

Due to the noisy nature of both technology and biology we treat gene expression as a probabilistic process. Following Friedman *et al.* (2000) we assign each gene a random variable that represents its level of expression[†]. We represent gene regulation by a directed regulation graph (the *network*), whose nodes correspond to genes. Each arc connects a *regulator* to its *regulatee*. We denote by P_X the set of *active regulators* of the gene X and by $\mathcal{X}_r$ the set of *regulatees* of a regulator r. Our model has a mechanistic nature: the expression level of each regulatee gene is a probabilistic function of its active regulators.

In order to find a network model that best explains the data, we use a *scoring function* to measure how well a set $\mathbf{Y}$ of regulators predicts the expression level of gene X, and solve the optimization problem of finding the highest scoring model. An example of such a scoring function is *mutual information*[‡].

Constraining the global structure

When trying to infer a network model from expression data, one finds that the number of possible solutions is prohibitively large. Therefore, we make a number of biologically-motivated assumptions, which significantly reduce the space of possible models. Not only does this reduction in search space lead to a more efficient search algorithm, it also enhances the statistical robustness of the results.

Our first assumption exploits prior biological knowledge to limit our set of candidate regulators $\mathcal{C}$ to the list of known and putative regulators of a given organism. Thus, our learning process focuses on finding which candidate regulators are *actively* regulating other genes, instead of finding which genes function as regulators altogether. Unlike previous approaches (Tanay and Shamir, 2001), which limit themselves to transcription factors, we expand our set of candidates to proteins involved in different aspects of regulation, namely transcription factors, signal transducers and protein kinases[§].

To justify this approach, we stress that in order to capture a regulation event in gene expression data, we must observe changes in the expression of both the regulator and the regulatee. Unfortunately, while transcription factors directly control transcription, the factors themselves are frequently regulated post-translationally, and their expression levels are often too low to allow reliable detection with microarrays. In such cases, we cannot observe the change in their regulatory activity in gene expression profiles. On the other hand, signaling molecules, which are often enzymes, are expressed at significantly higher levels and may be regulated transcriptionally (primarily by positive feedback). Thus, we can capture regulation relations indirectly, by the change in the expression levels of the signaling molecule (which in turn regulates a transcription factor) and its indirect target regulatee genes.

Our second set of assumptions is related to the structure of our regulation graph. We limit the maximal in-degree of each node in the graph. This assumption is commonly made by modellers of gene networks (Akutsu *et al.*, 1998; Friedman *et al.*, 2000; Tanay and Shamir, 2001). Finally, we allow only a small number of genes (the active regulators) to have an out-degree greater than zero. This is our central structural constraint and the key to our robustness at a global scale. This constraint is both biologically and computationally motivated. First, it adheres to the assumption that only a small fraction of the genome is directly involved in regulating transcription and that each such 'master regulatory gene' may affect the transcription of many other genes. Second, it assists in achieving statistical robustness: Only when a gene consistently scores high as a parent of many genes, we believe it indicates a true signal, while an occasional high score as a parent of a single gene is attributed to spurious chance.

Formal problem definition

We can now formally define the *Best Regulator Set* problem:

PROBLEM 1. Best Regulator Set
We are given as input:

- *A set of genes $\mathcal{X}$ and a set of m samples: $\mathcal{M} = M_1, \ldots, M_m$ over $\mathcal{X}$.*
- *A set of candidate regulators: $\mathcal{C}$*
- *A local scoring function: Score : $\mathcal{X} \times 2^{\mathcal{C}} \rightarrow Real$*

[†] We will use the term 'gene', instead of 'random variable denoting the gene's expression' throughout the paper.

[‡] Mutual information is defined as
$I(X, \mathbf{Y}) = \sum_{x,\mathbf{y}} P(X = x, \mathbf{Y} = \mathbf{y}) \log \frac{P(X=x,\mathbf{Y}=\mathbf{y})}{P(X=x)P(\mathbf{Y}=\mathbf{y})}$

[§] We used keywords related to transcription factors, signal transducers and protein kinases to obtain a set $\mathcal{C}$ of 456 candidate regulators from SGD (Cherry *et al.*, 2001) and YPD (Costanzo *et al.*, 2001).

- *Constants: d - the maximal indegree and k - the maximal size of the regulator set* **R**

For a candidate set of active regulators **R** *we define the following scoring function:*

$$F(\mathbf{R}) = \sum_{X \in \mathcal{X}} \max_{P \subset \mathbf{R}, |P| \leq d} Score(X, P) \qquad (1)$$

$F(\mathbf{R})$ *measures the quality of the optimal regulation model in which only* **R** *are regulators. The goal of our optimization problem is to find a small set* $\mathcal{R}$ *of regulators, which maximize this score:* $\mathcal{R} = \text{argmax}_{\mathbf{R}|\mathbf{R} \subset \mathcal{C}, |\mathbf{R}| \leq k} F(\mathbf{R})$

By treating this scoring function as an *oracle* to the optimization algorithm, we detach our combinatorial optimization problem from our probabilistic model, allowing us to reuse our algorithmic framework with other local scoring methods. In this paper we use mutual information as our scoring function. By maximizing the mutual information between regulators and regulatees, we are minimizing the conditional entropy of the regulatees. Thus, given the values for the set of active regulators, we minimize our uncertainty when predicting the values for the rest of the genes.

APPROXIMATION ALGORITHM

We now propose a greedy algorithm that searches for the best regulator set and its corresponding graph structure. Such an approximation algorithm is justified as our problem is NP-complete (see our website). We describe various implementation details which lead to speedy performance even over thousands of genes. We characterize the conditions under which we can guarantee the quality of the resulting approximation and justify them by empirically validating that these conditions hold for existing gene expression datasets.

Greedy Algorithm

We propose the following simple greedy algorithm: Begin with an empty set of active regulators and at each iteration add the candidate regulator that gives the largest gain. Continue to iterate until no candidate provides a significant additional contribution to the score. A crucial point is to correctly define the gain of a given regulator at each iteration. Note, that we calculate mutual information between a variable and its regulating **set**. Therefore, when considering a new candidate regulator c as a parent for a regulatee gene X, what we measure is not how much information c holds on X, but how much additional information c holds, which is not already contained in X's current regulator set. Since we must also adhere to the maximal indegree constraint and maintain $|P_X| < d$, in many cases c can be included into P_X only by removing another regulator from P_X. Such a swap is effective only if it leads to an improvement in X's local score. The active regulator chosen at each iteration is the candidate c which provides the most improvement when summing over all genes in $\mathcal{X}$.

Minreg Algorithm
set $R = \emptyset$, $F = 0$
do $(i = 1 \ldots)$ **set** $F' = F$ {
  //For each iteration find $c^* = argmax_{c \in \mathcal{C}} F(c|R)$
  foreach $c \in \mathcal{C}$ { **set** $R' = R \cup c$, $F'' = 0$
    foreach $X \in \mathcal{X}$ { **set** $\mathbf{P}_X = \emptyset$
      //greedily approximate $\max_{P \subset R', |P| \leq d} Score(X, P)$
      for $j = 1 \ldots d$ {
        $\mathbf{P}_X = \mathbf{P}_X \cup argmax_{p' \in R' \setminus \mathbf{P}_X} Score(X, \mathbf{P}_X \cup p')$ }
      $F'' += Score(X, \mathbf{P}_X)$ }
    if $F'' > F'$ **set** $c^* = c$, $F' = F''$ }
  $R = R \cup c^*$, $F = F'$ }
until $\forall c F(c|R) <$ threshold

Fig. 1. Overview of the Minreg algorithm. The algorithm consists of two nested greedy loops. The external loop finds the optimal set R of k regulators. For each $X \in \mathcal{X}$, an internal loop finds an optimal set of parents P_X.

We formulate the notion of best improvement by borrowing terminology from the field of economic utilities. We define a *valuation* function f_x for each gene $x \in \mathcal{X}$ as the function that measures the local score of x given a regulator set **R**: $f_x(\mathbf{R}) = \max_{P \subset \mathbf{R}, |P| \leq d} Score(X, P)$. The global valuation of a regulator set **R** is now defined as $F(\mathbf{R}) = \sum_{x \in \mathcal{X}} f_x(\mathbf{R})$. Note, that our function F is defined on subsets of $\mathcal{C}$. We define the *marginal utility* of adding a regulator set **C** to an already chosen regulator set **R** with regard to the function f_x as: $f_x(\mathbf{C}|\mathbf{R}) = f_x(\mathbf{C} \cup \mathbf{R}) - f_x(\mathbf{R})$. In a similar way, we define the marginal utility of **C** at $\mathbf{R} \subset \mathcal{C}$ with regard to F. Therefore, at each step our greedy algorithm chooses the single element with the largest marginal utility ($\text{argmax}_{c \in \mathcal{C}} F(c|R)$). An overview of our *Minreg* algorithm is presented in Figure 1.

Several details in Minreg's implementation allow us to quickly generate a model over thousands of genes. First, we have implemented a very efficient routine to calculate mutual information over discrete distributions. However, since other *Score* functions can be computationally intensive, the algorithm calculates and caches $Score(X, \mathbf{Y})$ only when needed. The pseudo-code presented in Figure 1 requires $dk^2|\mathcal{C}||\mathcal{X}|$ calculations of score, thus our algorithm is linear in $|\mathcal{X}|$ and quadratic (assuming $\mathcal{C} \subseteq \mathcal{X}$) in $|\mathcal{C}|$.

We also employ a number of heuristic tricks which in practice lead to substantial speed-up over the naïve algorithm. First, we use branch and bound: The potential candidates $\mathcal{C}$ are stored in a heap sorted by Improve(c), which is the previously calculated marginal utility of c. We traverse the candidate set starting from candidates that were recently found to have large marginal utility, and avoid the evaluation of candidates for which the recent

improvement Improve(c) is smaller than a factor of α of the best $F(c|R)$ found in the current iteration. Second, we do not recalculate $f_X(c|\mathcal{R})$ for each c and each X in each iteration. Rather, the function f_X is re-evaluated only after X's parent set P_X^* changes. This is particularly effective in later iterations where changes in P_x^* are rare.

Approximation bounds

The success of the greedy algorithm largely depends on the properties of the scoring function. Our algorithm fails in the situation where the marginal valuations $F(c_1|\mathcal{R})$ and $F(c_2|\mathcal{R})$ are both low, but $F(c_1 \cup c_2|\mathcal{R})$ is significant. In this case, neither c_1 nor c_2 are attractive enough to be chosen by our greedy algorithm, although by including both c_1 and c_2 in $\mathcal{R}$ the final score would increase over that derived by the greedy algorithm.

We formalize this intuition by introducing α-*modular* scoring functions. Following Lehmann *et al.* (2001), we define an α-*modular* function as follows:

DEFINITION 1. *A function f is α-modular ($\alpha \geq 1$) if and only if* $\forall z, \mathbf{A}, \mathbf{R}$ *the following holds:*

$$f(\mathbf{A} \cup z|\mathbf{R}) \leq f(\mathbf{A}|\mathbf{R}) + \alpha f(z|\mathbf{R}). \quad (2)$$

One might consider α as some measure on the convexity of f over the space of subsets: for larger α more can be gained by joining sets together. If we assume that the empirical distribution of our data is such that F is an α-modular function, we can prove a guaranteed lower bound on the quality of our approximation (see website):

THEOREM 1. *Denote by F_{opt} the optimal solution for the Best Regulator Set problem and denote by F_{minreg} the solution derived by the minreg algorithm (Figure 1). For distributions in which F is α-modular and monotone increasing it is guaranteed that:*

$$(\alpha + 1)F_{minreg} \geq F_{opt} \quad (3)$$

We empirically validated that we can safely assume small α-modularity of F (i.e., good approximation factor). Our validation scheme tests an equivalent but less intuitive formulation of α-modularity, namely that F is an almost decreasing function ($\forall T, \forall S \subset T, \forall z \notin T, \alpha F(z \mid S) \geq F(z \mid T)$) (Lehmann *et al.*, 2001), by randomly sampling such sets. Our procedure consisted of iteratively building up $\mathcal{R}$ by randomly choosing the regulator at each iteration k, and calculating $F(c|\mathcal{R}_k)$ for all $c \in \mathcal{C}$. We applied this procedure 1000 independent times to the data. Indeed, our calculations consistently showed that the marginal valuation of each $c \in \mathcal{C}$ is an almost decreasing function of the iteration. In the worst case $\frac{F(c|\mathcal{R}_{k+1})}{F(c|\mathcal{R}_k)}$ was 1.4.

PREDICTIVE POWER OF REGULATORS

We evaluated our algorithm on a dataset containing 358 samples combined from the Hughes *et al.* (2000) and Gasch *et al.* (2000) datasets. These datasets measure *S. cerevisiae* expression profiles under a wide variety of cellular conditions[¶]. The data was normalized and discretized to 3 values: *down-regulated*, *no change* and *up-regulated*, as proposed by Tanay and Shamir (2001). We automatically filtered non-informative genes (i.e., those that remain almost unchanged across the conditions), and remained with 3622 variables in $\mathcal{X}$. We bound the maximal indegree to 3 and assumed a maximal α-factor of 2 (defines depth of search in heap). Our current implementation of the Minreg algorithm requires 19 minutes on an Intel III 1GHz processor and detected a set $\mathcal{R}$ of 45 active regulators among a set $\mathcal{C}$ of 456 candidate regulators.

Robustness: Our algorithm is very robust to different choices of parameters, including discretization methods, thresholds for the stopping criteria and non-informative gene filtering, as well as parameters for the maximal α-factor and the maximal indegree d. Almost invariably, we obtained robust regulator sets: an intersection of over 80% between the top-20 regulator sets[‖] of almost any two runs with different parameters. Regulatee robustness was evaluated qualitatively: Significantly over-represented GO annotations in these sets (see below) also behaved robustly over the different variations.

Consistency of Prediction: More importantly, we performed cross-validation analysis, testing the ability of our active regulators to predict the expression levels of their regulatees. Recall that in our model each $X \in \mathcal{X}$ is a probabilistic function of its active regulators (i.e., each instantiation, $\mathbf{y}$, of the active regulators defines a distribution $P(X|\mathbf{Y} = \mathbf{y})$). Given a data sample in which $X = x$ and $\mathbf{Y} = \mathbf{y}$, the probability that our model assigns $X = x$ is $P(X = x|\mathbf{Y} = \mathbf{y})$. The parameters for the distributions $P(X|\mathbf{Y} = \mathbf{y})$ are estimated using maximum likelihood on the training data. However, small sample size often leads to an overfit model, which attains excellent performance on the training data, but performs poorly on samples not encountered during the training process.

We evaluated the performance of our algorithm using 5-fold cross validation. The data was randomly partitioned into a test set (containing 20% of the samples) and a training set containing the remaining samples. Minreg inferred a model containing 46 active regulators, using only the training set as its input. For each test sample, the values of all 3576 regulatees (genes) were predicted based on the values of the active regulators in that sample and

[¶] Hughes *et al.* (2000) contains 276 deletion mutants from various functional classes and Gasch *et al.* (2000) contains 82 samples of responses to 12 different stress conditions.

[‖] We defined an order on the regulators based on the size of their regulatee sets in the final model.

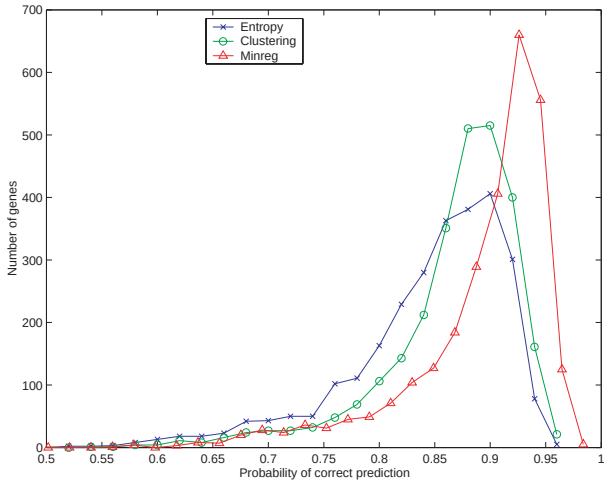

Fig. 2. Cross validation of the predictive capabilities of our model on test data. The graph measures the number of genes correctly predicted at each probability. We compare our model (triangles) to the null model (crosses) that uses the empirical distribution of each gene and to a model based on cluster representatives (circles).

the regulation model inferred from the training data. We calculated the overall precision of the model by comparing our predicted values with the actual measured ones.

We compared Minreg's prediction capabilities of several different models. As a baseline we used the empirical distribution of each gene as its own predictor. Since most of the genes remained unchanged most of the time, even this simple predictor scored well (Figure 2, crosses). As competition to our Minreg algorithm we generated 45 clusters using standard k-means clustering (Duda and Hart, 1973) and randomly chose from within each cluster a gene $r \in \mathcal{C}$ as its 'active regulator'. For each cluster we calculated $P(X|r)$ and used this as our predictor. While cluster representatives somewhat improved the prediction (0.06 log-loss/instance, Figure 2, circles), our Minreg algorithm clearly provided the best predictions (0.11 log-loss/instance, Figure 2, triangles). Several important advantages of Minreg account for its predictive success over clustering. These include capturing 'mechanistic' regulation using mutual information (vs simpler co-expression in clustering), identifying both activatory (correlated) and inhibitory (anti-correlated) regulators, and finding combinatorial logic and non-linear interactions between a gene and its active regulators (impossible in clustering where the same gene cannot belong to more than one cluster). In conclusion, our cross-validation demonstrates that most of the information in an entire array can be captured by a small set of master predictors.

BIOLOGICAL INTERPRETATION

In this section we describe how to extract biologically meaningful features from the results of our core algorithm. We distinguish between *local* features (e.g., Gene 1 regulates Gene 2) and *global* ones (e.g., Gene 2 regulates cell wall organization). By using multiple local relationships, we devise a method to answer questions on the global role of regulators, such as 'What is the biological process or molecular mechanism that a gene regulates?'.

While the basic building block of our model is the local *regulated* relationship, the main power of our framework lies in global interpretation. Individual local relations must be treated with caution, since they may rather indicate spurious artifacts or co-regulation. As an alternative, we gain global biological robustness by annotating each active regulator according to prominent features of its *entire regulatee set*. We devise a fully automated methodology, based on the Gene Ontology(GO)**, to map an active regulator r with specific biological processes (e.g., cell wall organization) and molecular functions (e.g., amino acid transporter) that are significantly over represented in its set of regulates, $\mathcal{X}_r$. We use the hypergeometric tail distribution to calculate a p-value that measures the probability of the given set to contain such a high concentration of a given GO term under a uniform null model. For a set of interest we systematically traverse all the nodes in the GO tree, searching for significantly associated GO terms. We developed an extension of the GENESYS system (Tanay and Shamir, 2001) for the visualization of the resulting model as well as its annotation and filtering according to GO-based classifications.

We validated our results in two ways. First, we compared our derived annotations of active regulators to their known functions as reported in the literature. Indeed, our results corresponded well to previous findings. For example, the associations for 8 of the top 10 active regulators provided 'proof of principle' (success or partial success, Table 1) by coinciding with well-known functional roles of these genes. Of the remaining two active regulators, we were able to assign a putative role to one previously uncharacterized gene, but failed to identify the correct role of the other. Importantly, the competing 'cluster representative' approach (GO annotation by cluster) did not yield a similar success (3/10 'sucesses', 7/10 'no support', for more details see website), further emphasizing the uniqueness of the Minreg approach. Second, we used the significant GO terms in order to focus our attention on individual regulatees which are more likely to represent true signal.

** Gene Ontology (The Gene Ontology Consortium, 2000) provides a well-defined vocabulary for the annotation of molecular function, biological process and cellular location. It is embedded in a tree hierarchy which permits queries at different levels of granularity. The entire yeast genome has been annotated using GO (Cherry *et al.*, 2001).

Numerous individual regulator-regulatee relations of interest were examined, which lent support to our global analysis. Below we provide a number of detailed examples that demonstrate some of the abilities of our method. More results and gene lists by GO term will be available on our website.

Our functional assignment can be highly discriminative and comprehensive, providing us with an elaborate characterization of the active regulator gene based on its regulatee set. For example consider SLT2, the MAP kinase that activates the cell wall integrity pathway. Our GO term derived annotation correctly predicts the biological process which SLT2 regulates (cell wall organization and biogenesis), the molecular function of the effectors (cell wall structural proteins and endopeptidases), the molecular mechanism of regulation (protein kinase cascade) and associated, cross-talking modules (mating, cell-cell fusion). Inspection of individual regulatee genes further supported this annotation.[††] Most importantly, Minreg found SLT2 to regulate RLM1, the main transcription factor of the low-osmolarity cell-wall integrity pathway. RLM1 is phosphorylated and activated *post-transcriptionally* by SLT2 (an event that cannot be observed in expression data), and, in turn, directly mediates SLT2's activatory effect. Indeed, several of the aforementioned regulatees (PST1, CRH1, BOP1, KTR2, GSC2, YPS3, PTP2) are known or putative RLM1 targets and promoter sequence analysis of SLT2's regulatees indicated a highly significant Rlm1 motif ($P = 2.85\text{e}{-05}$).

Our method can also correctly piece together the joint regulation of a specific biological process by several active regulators. For example, we detected several GATA transcription factors (GAT1, UGA3, DAL80) that are known to participate in the regulation of the nitrogen starvation response. All three regulators were associated with correct biological processes ('nitrogen starvation response' for GAT1 and DAL80, 'urea cycle' and 'nitrogen metabolism' for UGA3). However, the molecular functions they regulate only partially overlapped, highlighting their specific roles in a common response. Thus, only DAL80 regulates allantoin pathway genes (DAL1, 2, 3 and 7, consistent with the specific effect of dal80-null mutant on allantoin catabolism (Chisholm and Cooper, 1982)), while only GAT1 regulates protein biosynthesis and ribosome biogenesis, another molecular component of the response. Our model also captured *inter-regulator links* correctly: we found that DAL80 regulates GAT1, in accordance with recent findings on exactly such direct transcriptional regulation (Cunningham *et al.*, 2000).

In certain cases, our method allows us to assign function to previously uncharacterized regulators (Table 2), or to expand the functional assignment of known ones. For example, it associated the protein kinase YOL128C with the GO terms 'Nitrogen starvation response', 'response to external stimulus', and 'cell cycle control'. Among YOL128C's specific regulatees we find numerous cell cycle genes and regulators.[‡‡] Thus, we postulate that YOL128C acts as a cell cycle regulator in response to starvation signals. This prediction is further supported by YOL128C's homology to the meiosis regulator GSK3.

In other cases, our functional assignment may be misleading. For example, APG1, a signaling molecule involved in induction of autophagy after nutrient limitation, is strongly associated with 'protein biosynthesis' and 'structural proteins of the ribosome'. Although APG1 is not known to regulate this process, APG1 is regulated by TOR proteins, which are known to regulate **both** ribosome biogenesis and autophagy (Raught *et al.*, 2001). Tor1p itself is regulated post-transcriptionally, as reflected by its unchanged expression in most of the arrays. In the absence of a TOR signal in the data, we are capturing APG1 as its 'replacement' in our model, reflecting co-regulation rather than true regulation. Another TOR1 target TF, GAT1, shows a similarly strong association to 'protein biosynthesis' for the same reason. Note, that APG1's true role in autophagy is supported by other GO terms, including O-glycosyl hydrolases, as well as by specific autophagy genes (AUT2, AUT4) in its regulatee set.

Regulatory logic

Taking our approach a step further, we can identify the logic (activation or inhibition) of regulation. As before, rather than characterizing individual regulatory relations we focus on the regulation of an entire process or response, assuming that a given regulator exerts a coordinated effect on a subset of regulatees. Utilizing the mechanistic nature of our model we first annotate each individual relation with one of 3 logic values: activation, inhibition or unknown. Then, similarly to the method employed for functional annotation, we look for subsets of regulatees of the same GO class with a significant enrichment of a particular logic. Full details of the method are available at our website.

[††]Among the regulatees are known and putative structural cell wall proteins (CIS3, SED1, TIR1, PIR3, PIR1, BOP1, PAU1, PAU6, PRY2, DAN1, DAN3, TIR1, CRH1, PST1), cell wall enzymes (ASP3-2, BGL2), cell wall aspartic endopeptidases (YPS3, YPS5, YPS6), enzymes and other genes involved in cell wall biogenesis (MYO3, KTR2, KRT6, GSC2, CHS1) and transcriptional regulators known to affect cell wall maintenance (RLM1, ECM22, TUP1). Many of these genes were previously reported to be transcriptionally regulated by various stress conditions, such as heat- and cold-shock or changes in osmolarity. Most of the identified mating and cell-cell fusion genes (AFR1, CDC1, PRM8, PRM5, PRM10, CMP2, SCW10, PTP2, GIC2) are cell-wall related through biological process (budding, stress) or localization (membrane or cell wall proteins).

[‡‡]e.g., RSC3, CDC22, CDC25, NET1, KCC4, NIP29, DMC1, ULP1, PCL6, TEM1, ELM1, CHS3, SPA2

Table 1. Functional annotation of top-10 active regulators sorted by p-value of most significant GO term. For each active regulator, an existing annotation (adapted from YPD) and known GO annotations (SGD) and keywords)YPD) are compared to significant GO terms within its regulatee set (derived GO terms). Five of the derived annotations fully match known ones (Success), three match part of the known annotation (Partial Success), one is a novel functional assignment for a previously uncharacterized TF (novel function) and one is completely inconsistent with known functions (no support)

Regulator (# regulatees)	Concise Annotation	SGD (S) GO annotations and YPD (Y) keywords for biological process	Minreg's derived GO terms (score, #genes)	Verdict
SST2 (99) signaling	Negative regulator of the mating pheromone signaling pathway by binding to Gpa1p and desensitizing it to pheromone	Adaptation to mating signal (S,Y) Signal transduction (S)	Mating (0,17) Signal transducer (8.9e-05,7)	**Success**
MET28 (254) TF	Transcriptional activator of sulfur amino acid metabolism	Sulfur amino acid biosynthesis (S) Transcription regulation from Pol II promotor (S,Y)	Amino acid metabolism (0,24) Sulfur utilization (1.2e-05,6)	**Success**
GAT1 (125) TF	GATA zinc finger transcription factor that activates genes needed to use non-preferred nitrogen sources	Amino-acid metabolism (Y) Other metabolism (Y) Pol II transcription (Y) Cell Stress (Y)	Protein biosynthesis (0,30) Amino acid metabolism (0.000106,10) Hydrolase, acting on carbon-nitrogen (but not peptide) bonds in linear amides (0.005208,3) Nitrogen starvation response (0.008642,2)	**Success**
TEA1(141) TF	Ty1 enhancer activator of the Gal4p-type family of DNA-binding proteins	Pol II transcription (Y)	Amino acid metabolism (0,15) Nitrogen metabolism (7.9e-05,5) Urea cycle intermediate metabolism (0.000679,4)	**Novel function**
UGA3(73) TF	Transcriptional activator for 4-aminobutyric acid (GABA) catabolic genes	Amino acid metabolism (Y) Pol II transcription (S,Y)	Amino acid biosynthesis (1e-06,8) Urea cycle intermediate metabolism (5.4e-05,4) Nitrogen metabolism (7.8e-05,4)	**Success**
APG1 (168) Signaling	Serine/threonine protein kinase involved in the induction of autophagy after nutrient limitation	Autophagy (S) Meiosis (Y) Protein degradation (Y) Vesicular transport (Y)	Protein biosynthesis (0,29) Structural protein of ribosome (0,24) Hydrolase, hydrolyzing O-glycosyl compounds (0.000288,6)	**Partial Success**
SLT2 (245) Signaling	Serine/threonine (MAP) kinase involved in the cell wall integrity (low-osmolarity) pathway.	Signal transduction (S,Y) Cell stress (Y) Cell wall maintenance (Y) Protein amino acid phosphorylation (S)	Cell wall structural protein (4e-06,6) Cell wall organization and biogenesis (5.4e-05,11) Protein kinase cascade (0.017959,2)	**Success**
TPK1 (603) Signaling	Catalytic subunit of cAMP-dependent protein kinase 1, protein kinase A or PKA.	Signal transduction (Y) Aging (Y) RAS protein signal transduction (S) Pseudohyphal growth (S)	Ribosome biogenesis (7e-06,34) Protein biosynthesis (3.3e-05,58) Structural protein of ribosome (0.000128,46)	**Partial Success**
SIP4 (265) TF	Transcriptional activator of gluconeogenic genes through CSRE elements.	Transcription factor (S,Y)	Structural protein of ribosome (1.1e-05,28) Protein biosynthesis (6.9e-05,31)	**No support**
TEC1 (104) TF	Transcriptional activator, involved with Ste12p in pseudohyphal formation	Differentiation (Y) Pseudohyphal growth (S)	Mating (3.2e-05,9) Pheromone response (9.2e-05,6) Cell surface receptor linked signal transduction (0.013253,3)	**Partial Success**

The regulation logic we unraveled was often consistent with that reported in the literature, supporting the validity of our approach. For example, we found that MET28 activates the biological processes of 'threonine and methionine amino acid metabolism' as well as 'sulfur utilization'. Similarly our method correctly predicted the activatory roles of SLT2 and UGA3. All these findings are consistent with the known roles of these regulators (see Table 1). Importantly, the same regulator may assume different logical roles with respect to different processes or functions. For example, we found that GAT1 activates several amino acid metabolic processes, but inhibits 'protein biosynthesis' and 'ribosome biogenesis'. This is fully consistent with GAT1's known role in mediating the nitrogen starvation response, which involves both increase in amino acid metabolism and concomitant transcriptional down-regulation of ribosomal proteins and other biosynthetic genes (Kuruvilla *et al.*, 2001).

Table 2. Novel functional assignment for previously uncharacterized regulators with the Minreg algorithm. For each active regulator, significant GO terms were derived based on their regulatee set. These served as a basis for functional annotation of the regulator. In some cases (e.g., YOL128C, see text), the annotation was independently supported by external data

Regulator (# regulates)	Concise annotation	Significant GO terms (score, # genes)	Suggested novel annotation
TEA1 (141) TF	Ty1 enhancer activator of the Gal4p-type family of DNA-binding proteins	Amino acid metabolism (0,15) Nitrogen metabolism (7.9e-05,5) Urea cycle intermediate metabolism (0.000679,4)	**Regulation of amino acid biosynthesis (nitrogen utilization?)**
KIN1 (292) Sig	Serine/threonine kinase. Null mutant is viable and **shows no obvious phenotype**	Protein biosynthesis (0.000182,32) Cytokinesis (0.003699,6) Budding (0.005838,8) GTPase activator (0.019626,3) Cell cycle (0.034484,18)	**Cell cycle regulation (budding and cytokinesis)**
YOL128C (170) Sig	Protein kinase. **Unknown biological process and molecular function.**	Cell communication (0.000362,14) Nitrogen starvation response (0.000944,3) Response to external stimulus (0.00426,8) Cell cycle control (0.020422,5) Cell cycle (0.031847,12) Signal transduction (0.034232,6)	**Cell cycle regulation, perhaps in response to certain starvation signals**
KIN82 (127) Sig	Putative serine/threonine protein kinase. **Biological process unknown.**	Nucleotide metabolism (0.002014,3) Primary active transporter (0.003399,6) Mating-type specific transcriptional control (0.008909,2)	**Potential participant in the mating response pathway**

Our derived logic is particularly useful for the annotation of previously uncharacterized proteins. For example, we found that YOL128C positively regulates the 'nitrogen starvation response', while negatively regulating 'cell cycle control'. This indicates that the gene may regulate a stress response by inhibiting cell cycle progression under starvation conditions.

Regulatory logic must be interpreted with care, especially when the regulator is a signaling molecule. For example, we found that SST2 positively regulates mating and transmembrane receptors, while in fact it is a negative regulator of both (Table 1). To explain this discrepancy, note that SST2 is activated by STE12, the main mating TF, along with most of the mating pathway genes, while it exerts its inhibitory effect on the pathway only later, after a time delay. Thus, rather than representing its own (negative) regulatory role, SST2 serves as a representative of its fellow, co-activated mating genes. We believe that in such cases (mostly involving signalling molecules), we cannot reconstruct the regulatory logic from steady-state expression profiles. Still, we conclude that correct logic may often be derived by our analysis.

LOCATION ANALYSIS

A major advantage of our method is its applicability to various data types. Clearly, learning regulation solely from expression measurements is limited and much can be gained by incorporating data from diverse sources. Here, we demonstrate the flexibility of our framework by integrating gene expression with gene binding data obtained by genome-wide location analysis (Ren *et al.*, 2000). Genome-wide location analysis is a new high-throughput experimental approach that measures the binding of transcription factors to the promoter regions of an entire genome. The results ('binding data') are given in the form of a matrix: The entry $L_{i,j}$ represents the probability that TF i binds to the regulatory region of gene j. We devised a scoring function, $Score_b$, that integrates both expression and binding data, such that our model should now explain both types of observations. Thus, evidence of DNA binding of Y to X's promoter should raise the information Y contains on X. We do this by extending our alphabet and adding *psuedo-samples* for each TF with available binding data. When $i \in \mathbf{Y}$ is shown to bind to X the mutual information $I(X, \mathbf{Y})$ gains an additive factor. We control the weight of the binding location data relative to the expression data through a parameter b.

As a test, we used the Spellman *et al.* (1998) cell-cycle expression dataset (76 measurements of 800 genes in synchronized yeast cultures) in combination with binding data for 9 cell-cycle transcription factors (Simon *et al.*, 2001). In order to evaluate the contribution of the binding information we compared a run using no binding data to one incorporating such data with a low weight. When no binding data was used, 20 active regulators were chosen: 11 are known cell-cycle regulators, 5 regulate the mating and budding responses (both active in the dataset), 4 regulators are irrelevant, and 1 is unknown. Thus, as before, Minreg performs well in identifying biologically active regulators in the dataset. When the binding data was included, we observed some additional improvements: 5 additional cell cycle regulators were identified (to a total of 16), while 3 irrelevant ones were removed (leaving only 1 such regulator). Importantly, the additional cell

cycle regulators included both 3 genes for which binding data existed and 2 genes for which it did not. Thus, incorporating the binding data (even with a low weight) provided global improvement to the model for genes both with and without binding information.

DISCUSSION

In this paper we present a novel framework for inferring regulatory relationships from genome wide measurements, based on a fast algorithm that finds a small set of global active regulators. The global and robust nature of our model is demonstrated by our success in two challenging tasks. First, we succeeded in predicting significant portions of a microarray sample based solely the measurements for a small number of genes. Second, we associated comprehensive functional annotation to regulators, using a fully automated method, whose success we validated by detailed biological analysis. Our analysis shows that most of our high scoring assignments for known genes are correct, and that novel roles can be assigned to previously uncharacterized transcription factors and signaling proteins.

Our proposed method lies between molecular network modelling (Friedman *et al.*, 2000; Pe'er *et al.*, 2001; Tanay and Shamir, 2001) and global methods for the analysis of gene expression (e.g., clustering (Eisen *et al.*, 1998; Sharan and Shamir, 2000) and PCA (Alter *et al.*, 2000; Holter *et al.*, 2000)), and offers a number of advantages over both approaches. On the one hand, while previous network modelling approaches are computationally intensive and can only handle a limited number of genes, our Minreg algorithm is capable of quickly producing a biologically significant regulation network over *thousands* of genes, thus allowing it to scale to mammalian genomes. Furthermore, Minreg's robustness, speed and ease-of-use, together with the GO-based annotation and visualization framework, make it directly accessible to a wide community of biologists. On the other hand, the regulation structure we learn goes beyond clustering and PCA by capturing combinatorial logic and non-linear behaviour, both of which are important in biological systems. Indeed, Minreg's predictive superiority over clustering was clearly demonstrated.

Our framework can be extended in several directions: First, we are further developing other scoring functions, in order to treat other local models of regulation (e.g., continuous data) and additional sources of data (e.g., regulatory sequence motifs and mutational data). Second, we are extending the analysis of regulation logic in order to better model and capture combinatorial effects between regulators. Finally, we wish to adapt methods developed by Elidan *et al.* (2002) to identify the existence of 'hidden regulators' and their regulatee sets. This would allow us to better handle the frequent event of post-transcription regulation on transcription factors themselves (e.g., Tor1), and to enhance the quality of models reconstructed from expression data.

Acknowledgements

The authors are grateful to Nir Friedman, Rani Nelken, Noam Nisan, Itsik Pe'er and Ron Shamir for comments on drafts of this paper and useful discussions. D.Pe'er was supported by an Eshkol Fellowship. A.Regev was supported by the Colton Foundation.

REFERENCES

Akutsu,S., Kuhara,T., Maruyama,O. and Minyano,S. (1998) Indentification of gene regulatory networks by strategic gene disruptions and gene over-expressions. *SODA*. ACM-SIAM.

Alter,O., Brown,P. and Botstein,D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *PNAS*.

Cherry *et al.*,J.M. (2001) Saccharomyces genome database, http://genome-www.stanford.edu/Saccharomyces/.

Chisholm,G. and Cooper,T.G. (1982) Isolation and characterization of mutants that produce the allantoin-degrading enzymes constitutively in *Saccharomyces cerevisiae*. *Mol. Cell Biol*.

The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.

Costanzo,M.C. *et al.* (2001) Ypd, pombepd, and wormpd: model organism volumes of the bioknowledge library, an integrated resource for protein information. *Nucleic Acids Res.*.

Cunningham,T.S., Rai,R. and Cooper,T.G. (2000) The level of dal80 expression down-regulates gata factor-mediated transcription in *Saccharomyces cerevisiae*. *J. Bacteriol*.

DeRisi,J., Iyer,V. and Brown,P. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **282**, 699–705.

Duda,R.O. and Hart,P.E. (1973) *Pattern Classification and Scene Analysis*. Wiley, New York.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *PNAS*, **95**, 14 863–14 868.

Elidan,G., Pe'er,D., Friedman,N. and Koller,D. (2002) Discovering hidden variables: an information-theoretic approach. *Proceedings of the Eighteenth National Conference on Artificial Intelligence*.

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

Gasch,A.P. *et al.* (2000) Genomic expression program in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.

Holter,N. *et al.* (2000) Fundamental patterns underlying gene expression profile: simplicity from complexity. *PNAS*.

Hughes,T.R. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

Kuruvilla,F.G., Shamji,A.F. and Schreiber,S.L. (2001) Carbon- and nitrogen-quality signaling to translation are mediated by distinct gata-type transcription factors. *PNAS*.

Lehmann,B., Lehmann,D. and Nisan,N. (2001) Combinatorial auc-

tions with decreasing marginal utilities. *Ecomonic Comerence*.

Pe'er,D., Regev,A., Elidan,G. and Friedman,N. (2001) Inferring subnetworks from perturbed expression profiles. *ISMB'01*.

Raught,B., Gingras,A.C. and Sonenberg,N. (2001) The target of rapamycin (tor) proteins. *Proc. Natl Acad. Sci. USA*.

Ren,B. *et al.* (2000) Genome-wide location and function of dna binding proteins. *Science*, **290**, 2306–2309.

Sharan,R. and Shamir,R. (2000) CLICK: a clustering algorithm with applications to gene expression analisys. *ISMB'00*.

Simon,I. *et al.* (2001) Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, **106**, 697–708.

Spellman,P.T. *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Tanay and Shamir,A. (2001) Computational expansion of genetic networks. *ISMB'01*.

Website, http://www.cs.huji.ac.il/labs/compbio/minreg.