BIOINFORMATICS
Vol. 1 no. 1 2001
Pages 1–9

Inferring Subnetworks from Perturbed Expression Profiles

Dana Pe'er[1], Aviv Regev[2,3], Gal Elidan[1] and Nir Friedman[1]

[1]School of Computer Science & Engineering, Hebrew University, Jerusalem, 91904, Israel, [2]Department of Cell Research and Immunology, Life Sciences Faculty, Tel Aviv University, Tel Aviv, 69978, Israel and [3]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, 76100, Israel

ABSTRACT

Genome-wide expression profiles of genetic mutants provide a wide variety of transcripts measuring the response of cells to perturbations. Standard analysis of such data identifies genes that were affected by the perturbations and uses clustering to group genes of similar function. In this paper we discover a finer structure of interactions between genes, such as causality, mediation, activation and inhibition by using a Bayesian network framework. We extend this framework to correctly handle perturbations, and to identify significant substructures of interacting genes. We apply this method to expression data of *S. cerevisiae* mutations and uncover a variety of structured metabolic, signaling and regulatory pathways.
Contact: danab@cs.huji.ac.il

INTRODUCTION

Integrated molecular pathways consisting of interacting proteins, genes, and small molecules underlie the major functions of living cells. These include signal transduction and processing, regulation of gene expression and metabolism. Genome wide expression profiles allow us to gain insight into these processes. In order to obtain a wide variety of profiles, reflecting different active pathways, various perturbations and treatments are employed. Perturbation by mutation of specific genes serves a dual purpose, providing a rich variety of different profiles, while allowing us to compare a wild type profile with a mutant one and to determine the molecular effect or function of the mutated gene.

Two recent studies use such an experimental design, providing different types of analysis. **?** (**?**) compare mutant and wild type profiles to identify sets of "downstream" genes whose expression is affected by a specific mutation. **(author?)** (11) use clustering to group genes with correlated expression in different mutant strains or group entire mutant profiles. Valuable biological insight can be gained by both approaches.

In this paper, we strive to answer questions that deal with finer structure. For example, is the effect of a mutated gene on a target gene direct, or is it mediated by other genes? Which genes mediate the interactions within a cluster of genes or between clusters? What is the nature of the interaction between genes (e.g gene A inhibits gene B)?

To infer such finer relations from perturbed gene expression profiles[†] we use the framework of **(author?)** (8). In this framework, we treat the measured expression level of each gene as a random variable and regulatory interactions as probabilistic dependencies between random variables. Friedman *et al.* use *nonparametric bootstrap* to estimate the confidence of *features* of Bayesian networks learned from expression profiles. This allows them to identify pairwise relations of high confidence such as: "Genes A and B closely interact".

We extend this framework in four ways. First, we adapt and extend recent results on learning with interventions (1) to handle genetic mutations. Second we devise new, better suited, methods for discretizing the data prior to analysis. Third we define and learn new features: mediator, activator and inhibitor. Finally, we describe how to use features to construct *substructures* of strong statistical significance.

The resulting method comprises the following steps. We start by discretizing the data. Then, we apply bootstrap analysis to learn an ensemble of networks which represent potential models of the interactions between genes. We use this ensemble to extract features involving relationships between pairs and triplets of genes with high statistical confidence. We then identify statistically significant subnetworks which contain several high-confidence features. These subnetworks capture a strong statistical signal in the expression profile which is usually associated with some cellular process.

As a case study, we apply our framework for the analysis of the Rosetta Compendium of expression profiles from *Saccharomyces cerevisiae* (11).

[†]We stress that any attempt to perform this task is limited to learning relations that are represented in mRNA expression data. For example, post-translational regulation may often be missed.

Still need to add "a few sentences related to the biological results in the Introduction" Aviv?

BAYESIAN NETWORK ANALYSIS OF EXPRESSION DATA

Probabilistic Modeling of Gene Expression

Measurements of gene expression involve noise arising from the measurement technology and experimental procedures. In addition, the underlying biological processes are themselves stochastic. Thus, we choose to treat gene expression as a probabilistic process. We represent the expression level of each gene as *random variable*. The joint distribution over the set of all genes reflects the distribution of cell "states" and how these effect transcript levels . Our ultimate goal is to estimate and understand the structure of this distribution.[‡]

Most standard methods for analyzing gene expression focus on pairwise relations, such as correlation, between genes. However, biological interaction is seldom so simple, and often includes chains of mediators between two genes. By going beyond pairwise relations and exploring multi-variable interactions, we can infer more about the structure of the relationship between genes In particular, we focus on *conditional independence*. For example, if X and Y are co-regulated by Z then, while Y correlates with X, it might be that given the value of Z, Y becomes independent of X. In this case, we say that Z *separates* between X and Y. In general, such a separator can be a set of variables.

Bayesian Networks

A *Bayesian network* over a set $\mathbf{X} = \{X_1, \ldots, X_n\}$ is a representation of a joint probability distribution over $\mathbf{X}$. This representation consists of a *directed acyclic graph* (DAG) G whose vertices correspond to the random variables $X_1, \ldots, X_n$, and a parameterization which describes a conditional distribution for each variable given its immediate parents in G.

The graph G represents conditional independence properties of the distribution. These are the *Markov Independencies*: Each variable X_i is independent of its non-descendants, given its parents in G. A distribution that satisfies these independencies can be decomposed into the *product form*

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | \mathbf{Pa}_i^G), \qquad (1)$$

where $\mathbf{Pa}_i^G$ is the set of parents of X_i in G. The parameterization component of the network describes the conditional distributions $P(X_i|\mathbf{Pa}_i^G)$. Thus, the network represents the unique distribution.

The Markov independencies represented by G often imply other conditional independencies. We can determine whether G implies that X and Y are independent given $\mathbf{Z}$ by using *d-separation* (12). This is a simple graph theoretic criteria on the structure of the graph G. It turns out that two DAGs can imply exactly the same set of independencies. For example, consider graphs $X \rightarrow Y$ and $X \leftarrow Y$ over two variables X and Y. Both graphs imply that X and Y are not independent. In such a situation, we say that the two graphs are *equivalent*.

The notion of equivalence is crucial, since when we examine observations from a distribution, we cannot distinguish between equivalent graphs. Thus, we want to find the common properties of *equivalence classes* of DAGs. **(author?)** (14) show that equivalent graphs have the same underlying undirected graph but might disagree on the direction of some of the arcs. Moreover, they show that an equivalence class of network structures can be uniquely represented by a *partially directed graph* (PDAG), where a directed edge $X \rightarrow Y$ denotes that all members of the equivalence class contain the arc $X \rightarrow Y$; an undirected edge X—Y denotes that some members of the class contain the arc $X \rightarrow Y$, while others contain the arc $Y \rightarrow X$.

Learning Bayesian Networks

Given a *training set* $D = \{\mathbf{x}[1], \ldots, \mathbf{x}[M]\}$ of independent samples from an unknown distribution $P(\mathbf{X})$, we want to estimate this distribution by a network G. The common approach to this problem is to introduce a statistically motivated scoring function that evaluates each network with respect to the training data, and to search for the optimal network according to this score (9). A popular score based on Bayesian reasoning, scores candidate graphs G by their posterior probability given the data (see (10) for a complete description). We define the score $S(G : D)$ to be proportional to $P(G \mid D)$. An important characteristic of this score is that when the data is complete (no missing values) the score is *decomposable*:

$$\mathrm{S}(G : D) = \sum_i S_{local}(X_i, \mathbf{Pa}_i^G : D) \qquad (2)$$

The contribution of each variable X_i to the total score depends only on the values of X_i and $\mathbf{Pa}_i^G$ in the training instances.

$$S_{local}(X_i, \mathbf{U} : D) = \log P(\mathbf{Pa}_i = \mathbf{U}) + \log \int \prod_m P(X_i[m] \mid \mathbf{U}[m], \theta) dP(\theta).$$

[‡] We use the following notation in the remainder of the paper. We use capital letters, such as X, Y, Z, for variable names. Sets of variables are denoted by boldface capital letters $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$.

The first term is the prior probability assigned to choice of the set $\mathbf{U}$ as the parents of X_i. The second term measures the probability of the data, when we integrate over the possible parameterizations (θ) of the conditional distribution. These local contributions for each variable can be computed using a closed form equation (see (9) for details).

MODELING PERTURBATIONS INTO BAYESIAN NETWORKS

Ideal Interventions

Above we assumed that each training instance was sampled from the underlying distribution. This does not apply in genetic mutation experiments. For instance, by knocking out gene X, we replace the original molecular control on X's expression (its parents) by an external one. Thus, any consequent measurement (in which X's value is constantly set to 0) will not teach us anything about X's conditional distribution on its parents. Modeling such interventions for learning Bayesian networks involves two issues: the score function and the definition of equivalence.

Recall that the score of a DAG G, given a data set D, decomposes into a product of entities that depend on the conditional distributions $P(X|\mathbf{Pa}_X^G)$. Suppose that in a certain sample, we intervene by fixing the value of $X_i[m]$. In this sample, it is clear that we should not take into account $P(X_i[m] \mid \mathbf{Pa}_i[m])$, as the value of X_i in the sample does not depend on this distribution. However, if our intervention only modified the value of X_i, all others variables were sampled from their respective conditional distributions. We call such manipulations *ideal interventions* (13) and treat their score as follows: If we let $Int(m)$ denote the set of variables that were intervened in the m'th sample, then the modified local score is

$$S_{local}(X_i, \mathbf{U} : D) = \log P(\mathbf{Pa}_i = \mathbf{U}) + \log \int \prod_{m, X_i \notin Int(m)} P(X_i[m] \mid \mathbf{U}[m], \theta) dP(\theta).$$

See (1) for more details on this score.

This score is no longer *structure equivalent*, i.e., the score of two equivalent graphs, G and G' is no longer guaranteed to be the same. This should be expected, as interventions help us determine the direction of causality. We say that G and G' are *intervention equivalent* given interventions $I \subseteq \{X_1, \ldots, X_n\}$, if they receive the same score given a data set D where $Int(m) \subseteq I$, for all m. This notion of equivalence is more restrictive, and thus more edges in the PDAG will be directed. These include, but are not limited to, all edges entering or leaving an intervened variable X. We modified the procedure for constructing a PDAG representation from a DAG (**?**) to fit our new equivalence relation. Due to space restrictions, we omit the technical details.

Modeling Perturbations

We distinguish between two types of perturbations in gene expression data. The first type includes gene deletion and over-expression. Both imply a direct change to the expression level of the mutated gene. Formally, the random variable corresponding to this level is deterministically assigned a specific value. We model such mutations as ideal interventions, as described above.

The second class of perturbations includes temperature sensitive and kinetic mutations (**?**)) ; and the application of external conditions (e.g. environmental stress (**?**)). These perturbations do not directly determine an expression level of a specific gene, and thus cannot be modeled as ideal interventions. Still, they have an important effect on the expression level of many genes in our system, and therefore their occurence in a given sample should be indicated. We add *indicator variables* to our domain, one for each treatment type. We constrain such variables to be roots i.e. no other variables can be their parents in the network.

ZOOMING IN: IDENTIFYING FEATURES

Potential Features

In this section we focus on the following question: Can we elucidate the nature of interaction between two genes? We use the perturbed gene expression profiles to learn a Bayesian network model G and construct its corresponding PDAG U_G (taking into account the patterns of interventions) Assuming that G correctly captures the dependencies in the domain, what types of conclusions can we draw from G? We now consider several types of "queries" or "features" that can be identified from G and U_G.

Markov and Edge Relations To find if there is a direct interaction between X and Y we can query our network whether X and Y are *Markov neighbors*. Markov neighbors are variables that are not separated by any other measured variable in the domain. They include parent-child relations (one gene regulating another spouse relations (two genes that co-regulate a third), and sibling children of a hidden variable (two genes regulated by a third one, not modeled in the network, e.g. protein activation). When neither of these situations occurs, the network implies that the interaction between X and Y is indirect.

We can query whether the edge $X \rightarrow Y$ appears in U_G. Recall that this implies that X and Y are Markov neighbors (parent-child type) and that the edge between them is directed in all networks in the equivalence class of G. The existence of such a directed edge suggests that X

is a direct *cause* of Y.[§]

Separators When X and Y are indirectly dependent, we can ask what factors *mediate* this dependence. In the simple case, a single variable Z, separates X and Y. For example, the edges $X \rightarrow Z \rightarrow Y$ or the undirected edges X—Z—Y appear U_G. In the former case X affects Z who in turn affects Y, while in the latter Z might be a common cause of both X and Y.

In more complex cases, X and Y can be more distant in the graph structure (e.g Z is a common grandparent of both X and Y) and there might be more than one variable that mediates their interaction (e.g X is parent of Z_1 and Z_2, who in turn are both parents of Y). In these cases we must employ a global approach, searching for variables $\mathbf{Z}$, such that Y is independent of X given $\mathbf{Z}$ in the network. In such a situation, we say that $\mathbf{Z}$ explains all the dependencies between the two variables.

We can test such dependencies using d-seperation. More precisely, to check that two variables X and Y are independent given Z, we need to check that no path between X and Y can "pass" information when we know the value of $\mathbf{Z}$. (See **(author?)** (12) for the precise definition.) Thus, to test for d-separation we need to consider *every* path between X and Y in the network. This is computationally impractical for a large domain. We can solve this problem by recognizing that when two variables are far from each other in the network, the dependence between them diminishes significantly. Thus, in practice we check for d-separation between variables along paths of limited length.

Activation and Inhibition When X is a parent of Y, we can gain understanding of X's effect on Y. Contrary to previous cases, here we are interested in the conditional distribution $P(Y \mid \mathbf{Pa}_Y)$. Let $\mathbf{U} = \mathbf{Pa}_Y - \{X\}$. Intuitively, if $P(Y = 1 \mid X, \mathbf{u})$ increases when X transitions from -1 to 0 and then to 1 and $\mathbf{u}$ is held fixed, we say that X *activates* Y. Since all other direct influences on Y have been kept at the same state, the change in X is the explanation to the change in Y. Similarly, if $P(Y = -1 \mid X, \mathbf{u})$ increases, then X *inhibits* Y. We currently use a very strict criteria that requires the X activates/inhibits Y for every set of values $\mathbf{u}$ of $\mathbf{U}$. We are currently exploring less naive approaches that soften this requirement.

Feature Confidence

Above we assumed that the network G correctly represents the interactions in the underlying domain. How reasonable is this assumption? If we have a sufficiently large number of samples, we can be (almost) certain that the network we learn is a good model of the data (**?**). However, given only a small number of training instances, there can be many models that explain the data almost equally well. Such models can have qualitatively very different structures. We do not have confidence that one network is an accurate description of the biological mechanisms.

Therefore, instead of querying a single structure, we can examine the *posterior* probability of the feature given the data. Formally, we consider the distribution of *features*. A feature of a network is a property such as "$X \rightarrow Y$ is in the network" or "$\mathbf{Z}$ d-separates X from Y in the network". We define the feature using an indicator function $f(G)$ that has the value 1 when G satisfies the feature and value 0 otherwise. The posterior probability of a feature is

$$P(f(G) \mid D) = \sum_G f(G) P(G \mid D). \qquad (3)$$

This probability reflects our *confidence* in the feature f.

A naive way of calculating equation 3, is by enumerating all high scoring networks. Unfortunately, the number of such networks can be exponential in the number of variables, so exact computation of the posterior probability is impractical. Instead, we can estimate this posterior by sampling representative networks, and then estimating the fraction that contain the feature of interest. We can generate such networks using non-parametric bootstrap (6) or using more exact but costly MCMC simulations (**?**). **(author?)** (8) evalate the bootstrap approach in simulated data that matches the distributions observed in gene expression data. They note that the rate of false negatives is high. Thus, the fact that we do not detect high confidence for a feature, does not mean it does not exist, but rather that the data does not strongly support this feature.

RECONSTRUCTING SIGNIFICANT SUB-NETWORKS

While features provide us with important insight, our view remains limited to relations between two or three genes. However, by bringing together multiple relations our framework can offer a much broader viewpoint. It provides richer, more structured context when exploring data. We consider a small subset, U, of the variables. A *sub-network* $G[U]$ is a graph on U whose edges encode pairwise features between variables.While our a full-scale network is currently of insufficient quality and statistical significance, self-contained sub-networks can be reliably reconstructed from individual features.

These sub-networks are derived from *seeds*: small connected components of high confidence Markov features. We believe that seeds indicate biological phenomena captured by the data. We extend such seeds into connected

[§] To reach causal conclusions from a Bayesian Network a few assumptions must be made. See (13; 3) regarding the connection between Bayesian networks and causality, and (8) for a discussion of these connection in the context of gene expression.

sub-graphs with a high concentration of Markov features. This sub-graph and all features associated with its variables constitute the resulting sub-network. While an isolated feature of moderate confidence might be a false positive, it becomes significant in the context of a sub-network rich with high scoring features. Simulated data strongly supports this claim: Figure **??** compares the false positive rate in the entire network with the false positive rate in the significant sub-networks alone.

We present two approaches for extracting sub-networks from features. The first, more naive approach simply captures a defined radius around seeds, while the second approach automatically searches for an optimal sub-network based on some scoring scheme.

Naive Approach

We now define a weighted complete graph $M = (\mathbf{X}, E, W)$ over the set of random variables in our domain. The weight $W(X, Y)$ of the edge between X and Y is the confidence of the Markov relation between them (which is 0 in most cases). We further consider threshold-induced sub-graphs of M: for $t \in [0, 1]$ let $M[t] = (\mathbf{X}, E[t])$ include all edges of weight at least t.

We used the following heuristic procedure to construct sub-networks: Choose a seed U_S to be a connected component in $M[t_S]$ s.t. $|U_S| \geq k_S$ (we used $t_S = 0.75$ and $k_S = 3$). Next, expand U_S into a full sub-network $G[U_F]$ by inducing $M[t_F]$ on the subset U_F of all vertices within distance d_F from the seed U_S (we used $t_F = 0.5$ and $d_F = 1$).

In order to obtain sub-networks that represent a coherent biological processes, we operated in a semi-manual way. Loosely speaking, we merged sub-networks whose genes are known to be related to the same biological process. While results make biological sense (see below), there are a number of drawbacks to this approach: first there is no measure of quality for the resulting networks; and second, the symmetric expansion of the seed is a crude rule of thumb. We now address these issues.

Score-based Approach

We assume biologically meaningful sub-networks are *concentrated* with features. In order to measure the statistical significance of a sub-network, we formalize this notion and develop a scoring scheme for sub-networks. Denote $n = |\mathbf{X}|$ and $N = \binom{n}{2}$. We consider a null hypothesis of i.i.d. confidence levels for each edge. Define this distribution of confidence levels by the cumulative probability $\text{Freq}(c) = Prob(W(e) \geq c)$. Consider a specific subset U of size k, and set $K = \binom{k}{2}$. The chance of U inducing a sub-network with edges $e_1, \ldots, e_l$ having confidence levels better than $c_1, \ldots, c_l$, respectively is less than $\binom{K}{l} \prod_i \text{Freq}(c_i)$. Given such a set $C = \{c_i\}$, we can thus bound the expected number of such subsets by

$$B(k, C) = \binom{n}{k} \binom{K}{l} \prod_i \text{Freq}(c_i) \qquad (4)$$

For a sub-network $G[U']$, with edges $e_1, \ldots, e_l$, define $C' = \{W(e'_i)\}$, and score $G[U']$ using $-\log B(|U'|, C')$. It remains to explain how we compute the distribution $\text{Freq}(c)$. We estimate this function from the given Markov relations by $\text{Freq}(t) \equiv \frac{|E[t]|}{N}$.

We implemented a local search to find such high scoring networks. The search starts with a candidate seed U_0 which is a vertex triplet connected by high scoring ($M[t_S]$) edges. At each step we consider adding or removing a single vertex to U_i, attempting to improve the score of $M[t_F]$ induced on U_{i+1}. Optima whose score exceeds a specified threshold are considered significant.

Experience with using the score presented thus far was encouraging. However, we found that biologically meaningful sub-network are characterized not only by their concentration, but also by their typical structure: in general, such sub-networks include a few key-genes that regulate many others. This results in an uneven distribution of degrees. We can bias our scoring function to prefer such structures by replacing the $\binom{k}{l}$ term in equation **??** with the multinomial coefficient $\binom{2l}{d_1\ d_2\ \cdots\ d_k}$, where $\{d_i\}$ are the degrees of vertices in the sub-network. The rationale is that the edges are no longer arbitrary, but rather the degrees are conserved.

As a sanity check, we tried the procedure on randomized data. We reshuffled the original data-set, thus eliminating genuine dependencies between variables. We used this simulation to set a threshold markedly beyond any of the highest scoring random artifacts.

DISCRETIZING GENE EXPRESSION DATA

Due to noisy experimental procedures and measurement techniques, gene expression data must be handled with care to ensure successful application of analysis methods. A key pre-processing step is the discretization of expression levels into functional expression states (e.g. under-expressed, baseline and over-expressed). In this section we introduce a procedure to discretize the gene expression measurements. The procedure uses a model of the gene expression to guide a k-means clustering algorithm. Our procedure requires repeated gene expression measurements.

In our model, Each gene can be in a few discrete functional expression states, which relate to its activity. Given a gene's state, we model its expression level as a Normal distribution. This distribution has a gene specific variance, with each state centered around a different mean ¶ The precision of a gene can be empirically

¶We base our model on results from repeated gene expression measure-

Fig. 1. Comparison between false positive rate in entire networks verses signifi cant sub-networks alone. These rates were calculated by applying our process to simulated data

estimated from repeated measurements. In order to identify significant changes in expression levels, we devise a test on the ratio of X's gene expression between a condition and a control. In most studies a *two-fold test* is used, considering changes two-fold or higher as significant. We employ an alternative approach, based on a Bayesian procedure to estimate the posterior probability over the mean and variance for a given gene (5) and test the probability that the treated sample came from the same distribution. For lack of space we omit the technical details.

We apply our two step discretization method separately to each gene. First we use the ratio significance test from the previous paragraph to mark each measurement as an under/over expressed or baseline state if the is significantly under/over expressed or neither. This is used as the starting point for a k-means clustering algorithm, which decided both the number k of states (bins) and their specific allocation. K-means clustering is performed over the expression levels of the gene in order to assign a value $\{-1, 0, 1\}$ to each measurement according to the bin it was clustered into. K-means clustering is known to prefer round Gaussian-like bins, consistent with our model.

RESULTS

The Rosetta Inpharmatics Compendium (11) is a reference dataset compiled of 300 full-genome expression profiles obtained from 276 deletion mutants, 11 tetracyclin regulatable alleles of essential genes, and 13 chemically treated *S. cerevisiae* cultures, each compared to a baseline wild type or mock-treated culture. This wide range of interventions gives rise to a heterogenous set of expression profiles corresponding to differrent signaling, metabolic and regulatory pathways. We have set to unravel and explore these pathways using the Bayesian network framework which we developed for intervention-based expression profiles.

We chose a subset of 565 genes which included the mutated genes and genes which showed significant change in at least 4 profiles. Using the 63 control experiments for parameter estimation we discretized these genes using our k-means clustering approach. We included a root variable for the genetic strain of the control sample, as well as indicator variables for the chemical treatments. We used our bootstrap learning procedure to learn 100 networks and extracted edge, markov, separation triplet and activation features. The learning procedure is rather computationally extensive, each network is learned independently and requires approximately 3 hours CPU using a Intel III 700 processor and 500 megabytes of memory. We have developed a Pathway Explorer for the study of feature-rich areas. Entire sub-networks filtered to the desired level of confidence are visualized as directed graphs, in which extensive local information is associated with the undirected and directed edges, including confidence levels (for the Markov and edge relations), correlation coefficients, and activation and inhibition relations (when available). Important metadata (gene and protein information, expression patterns, etc) is readily available as well. We stress that no prior biological knowledge was used by our learning procedure when reconstructing the networks. The full annotated results can be viewed using Pathway Explorer at our web site: `http://www.cs.huji.ac.il/labs/compbio/Rose`. Here we focus on several examples that highlight the validity and power of our approach.

Pairwise Relations

Biological analysis of individual Markov pair relations indicates that many are supported by previous findings, and represent either a known biochemical or regulatory interaction, a shared common regulator, or functional link.

ments. Such experiments clearly a Normal distribution for the gene expression with the precision varying greatly between different genes

Strikingly, the Pearson correlation coefficient between approximately a third of these "proof-of-principle" gene pairs was lower than 0.7. Our method is capable of discovering such relations because of the *context specific* nature in which it handles the data. There are many biological processes that occur only under specific conditions. Correlation "misses" such interactions, which are only apparent in part of the samples. Scores for features are presented in the following format: (Confidence, Pearson correlation) for each such pair. Two such "proof of principle" Markov pairs are, Phosphoribosylaminoimidazole carboxylase (*ADE2*) and Phosphoribosylamidoimidazole-succinocarboxamide synthase (*ADE1*) (0.797, 0.518), which catalyze the sixth and seventh steps in the de novo purine biosynthesis pathway, respectively; and *SST2*, a (negative) regulator of the mating signaling pathway and *STE6* (0.914, 0.677) the membrane transporter responsible for the export of the "a" mating factor.

Even pair-wise relations alone succeed in providing new biological insight. For example, we studied an edge relation (0.914, 0.162) from *ESC4*, a protein involved in chromatin silencing to *KU70*, a key component of the DNA non-homologous double strand break DNA repair mechanism. This is a previously unknown link, yet we supply evidence from the known literature strongly supporting it. Other chromatin silencing genes (SIR2, 3, and 4) are necessary together with KU70 and KU80 for DNA end joining(**?**). *ESC4* also contains 6 BRCT domains, that are known to occur predominantly in proteins involved in cell cycle checkpoint functions responsive to DNA damage (**?**). Together, these facts clearly support both a functional association between the two proteins and a regulatory directed interaction (from *ESC4* to *KU70*) assigning a new (putative) regulatory function to *ESC4* in double strand break repair. Note, that a *ku70* mutant strain is included in the compendium data, while *ESC4* had not been mutated. This illustrates how our treatment of mutations aids in inferring causal relations in a counter intuitive direction. While typical analysis can only find the effect of a mutation, we find a causal source (in wild-type strains) of a mutated gene.

Separator Relations

In this section we provide an illustration of the capability of separator triplets to explain away dependencies, providing an enhanced insight into the underlying molecular architecture of pathways. We consider three genes each appearing in several undirected separator triplet relations. All three genes are well known mediators of transcriptional responses, and the genes they separate share functional roles and regulation patterns, consistent with the separator serving as a common regulator.

The first gene, *KAR4*, is a mating transcriptional regulator of karyogamy (nuclear fusion) genes, which is known to pair with the mating transcription factor Ste12p to activate genes required for nuclear fusion (**?**). *KAR4* separates several pairs of cell fusion genes (e.g. *AGA1* and *FUS1*). The second gene, *SLT2*, encodes the MAP kinase of the cell wall integrity (low osmolarity) pathway, which post-translationaly activates (by phosphorylation) the transcription factors *Rlm1p* and *Swi4/6* which in turn activate low osmolarity response genes (**?**). *SLT2* separates several pairs of cell membrane and cell wall proteins (e.g. *YSP1*) as well as previously uncharacterized one (e.g. *SRL3*). In addition, an activation relation was detected between *SLT2* and *YSP1* which is consistent with *SLT2*'s known regulatory effect. The third gene, *SST2*, is a post-translational negative regulator of the G-protein in the mating signaling pathway (**?**). *SST2* separates the mating response genes *TEC1* and STE6. Moreover, a directed inhibition edge was discovered from *SST2* to *STE6*, consistent with *SST2*'s known inhibitory role in the mating pathway.

We conclude that in all three cases, our inference has reconstructed the regulatory role in the correct molecular and functional context, revealing both transcriptional and post-translational regulators. Importantly, many of these significant three-wise relations were characterized by low and/or uniform correlation coefficients, indicating that correlation analysis and clustering would fail to identify this fine regulatory and functional structure. Our approach succeeds, through the use of conditional independence. Furthermore, since previously uncharacterized genes participated in some of these interactions (e.g. *SRL3* in *SLT2*, *YNL276W* in *KAR4*) we could assign them putative functions, probably as effectors, in cell wall integrity and cell fusion, respectively.

Sub-network analysis

The full power of our approach becomes apparent when exploring sub-networks of high confidence (> 0.75). Of 87 top scoring Markov pairs, 61 were organized[‖] within the context of 6 well-structured sub-networks, interleaved with additional lower confidence relations. Each of the sub-networks represents a coherent molecular response: mating response, low osmolarity cell wall integrity pathway, stationary phase response, iron homeostasis, amino acid metabolism along with mitochondrial function, and citrate metabolism (two are depicted in figure 1, all available at our website).

While (11) have identified some of these responses (amino acid metabolism, iron homeostasis andmating) by use of clustering, our reconstructed networks provide a much richer context for both regulatory and functional

[‖] An additional 16 relations could be explained as individual interacting gene pairs or triplets, and only 10 relations currently remain unassociated or unexplained.

analysis. For example, (11) describe a large cluster of genes associated with amino acid metabolism. In our network, we can discern at least three finer structures with high confidence. The first involves the genes *ARG1*, *ARG3* and *ARG5*, all part of the urea cycle (and its close periphery), which are known to be transcriptionally co-regulated. (**?** ; **?**). The second, for sulfate metabolism which further decomposes into two branches: one of sulfate transporters (*SUL1* and *SUL2*) and the other of sulfate assimilation (*MET3*, *MET14*, and *MET22*). The common separator for these branches is the *MET10* gene. The third and major part of the network interleaves various enzymes for amino acid metabolism (e.g. *HIS4*, *HIS5*, *LEU4*, *ILV2* and *ARG4*) with mitochondrial proteins, most prominently transporters and carriers (e.g. *BAT1*, *OAC1*, and *YHM1*). A regulatory link has been found between the general amino acid response and mitochondrial function (**?**). Thus, a large group of genes, which by correlation alone would be simply clustered together, can be organized in clear functional networks.

We use the mating response sub-network (figure 1) to illustrate the power of our method to reconstruct a coherent biological tale and raise novel biological hypotheses. We discern two distinct branches , one for cell fusion and the other for outgoing mating signaling. According to our network, the cell fusion response branch is mediated by the *KAR4* gene (see above), and includes several known cell membrane fusion genes (*FUS1*, *AGA1*, *AGA2*, *PRM1* and *FIG1*) (**?** ; **?** ; **?**) as well as two genes previously unassociated with this process (*TOM6* and *YEL059W*). The multitude of high confidence relations strongly suggests a putative role to *KAR4* not only regulation of nuclear fusion but also regulation of cell membrane fusion.

Another branch is directed from the mating signaling pathway regulator *SST2* (above). Since an *SST2* mutant has been incorporated in the compendium we could determine edge direction, and identify *SST2* as a prime regulator of several other genes (*TEC1*, *STE6*, *MFA1*) previously shown to be transcriptionally regulated by the mating pathway (**?** ; **?** ; **?**). The regulatory link from *SST2* to *KSS1* is intriguing as the two share an interaction with *MPT5*, a multicopy suppressor of transcript specific regulators of mRNA degradation in yeast (**?** ; **?**), but *KSS1* was not previously associated with the mating pathway, but rather with the (related) filamentous invasive growth response.

Some puzzling discrepancies exist in our network. The first is the absence of the main transcription factor of the pathway, *STE12*. This may be due either to loss of information by our discretization procedure or to our bias to reduce the number of false positive interactions. The second, is the marginal position of the pathway's MAP kinase, *FUS3*. This may be due to positive feedback, rendering *FUS3* both an activator and an activation target. However, despite the knockout mutation in FUS3 we have failed to identify directed regulation. We believe that larger number of repetitions for each mutation will enhance our framework's capabilities to discover such regulatory relations.

DISCUSSION AND FUTURE WORK

In this paper we extend the framework of Friedman et al (8) for analyzing gene expression measurements with Bayesian networks. We integrated into this framework a new discretization procedure and a principled way for learning with a mixture of observational and interventional data. We defined and examined new types of features which can be uncovered using our analysis method. We presented automated methods of integrating these features into structures representing biological processes. Finally, we applied these tools to analyze the Compendium data of *S. cerevisiae* mutations (11).

This analysis illustrates several aspects of our framework. First, our score significantly differs from the correlation coefficient often used by clustering methods (**?**). The advantage is two fold. On the one hand we are able to discover inter-cluster interactions between genes whose expression profiles have very low correlation. On the other hand we can uncover fine intra-cluster structure among related genes. This assists us to understand the roles of genes within a richer context. Thus, genes can be assigned putative novel roles. The use of the Pathway Explorer greatly facilitates such biological exploration. Both regulatory, metabolic and signaling components are identified, showing the potential of our approach to uncover the three major types of molecular networks. We stress that our approach cannot recover all interactions. Instead we attempt to provide the biologist with a small number of highly promising hypotheses.

The primary contribution of this paper is an automated methodology that can reconstruct biological pathways from gene expression profiles. Our initial automatic attempts were quite successful and succeeded in recovering 4 out of 6 partially handcrafted networks. Still issue of scoring biologically meaningful sub-networks opens the door further research. A possible avenue is to use prior knowledge to both improve the quality of the feature and guide the search for meaningful sub-networks. Currently our learning and inference framework is based on expression data alone, without incorporating any prior biological knowledge. While the rich biological interactions uncovered shows proof of our method's capability, our aim is to develop tools that aid the biologist in finding novel hypotheses. Therefore, it is important to introduce principled methods for incorporating prior biological knowledge into our methodology.

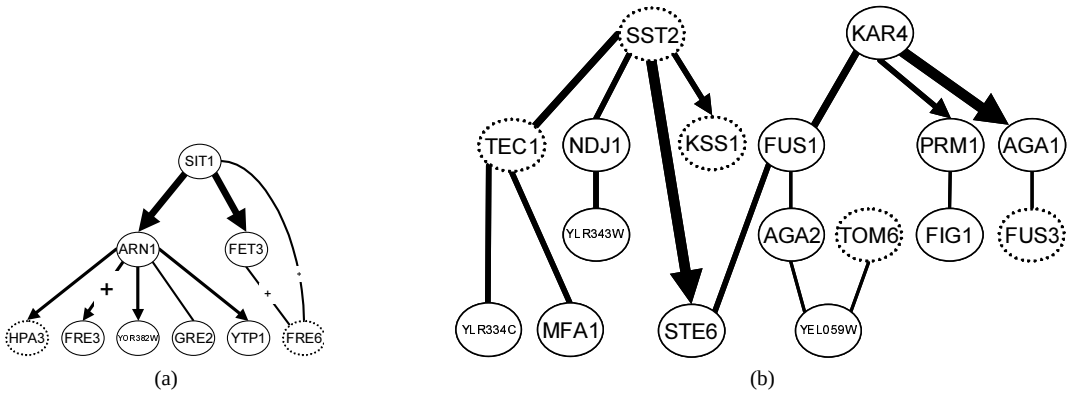

Fig. 2. Two sub-networks that visualize features discovered. (a) Iron homeostasis (b) Mating response. The width of the arc corresponds to the confidence of the feature. The edges are directed only when there is high confidence in its orientation. Nodes circled with a dashed line correspond to genes which have been mutated in some of the samples. Arcs marked by a + sign are activators, size corresponds to confidence of feature. Due to space limitations, the iron homeostasis pathway, which displays many edge triplets, is not discussed here.

Another important issue we address is that of recovering causal structure and differentiating between direct and indirect effects. The list of separator relationships from the resulting analysis shows a strong bias (at least on the Compendium data) toward mediators who are common parents. We were less successful in identifying mediators between a mutation and its indirect targets. This is partially due to the specific nature of the dataset, where only a single profile is available for each mutation. Devising better methods to find such mediators remains an important research problem.

Acknowledgements

The authors are grateful to Michal Chur, Rani Nelkin, Matan Ninio, Itsik Pe'er and Eran Segal for comments on drafts of this paper and useful discussions relating to this work. This work was supported in part by Israel Science Foundation (ISF) grant 244/99 and Israeli Ministry of Science grant 2008-1-99. The computational resources were funded by an ISF infrastructure grant. D. Pe'er was supported by an Eshkol Fellowship from the Israeli Ministry of Science. A. Regev was supported by the Colton Foundation. N. Friedman was supported by an Alon Fellowship.

REFERENCES

[1] G. Cooper and C. Yoo. Causal discovery from a mixture of experimental and observational data. In K. Laskey and H. Prade, editors, *Proc. Fifthteenth Conference on Uncertainty in Artifi cial Intelligence (UAI '99)*, pages 116–125. Morgan Kaufmann, San Francisco, Calif., 1999.

[2] G. F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

[3] G.F. Cooper and C. Glymour, editors. *Computation, Causation, and Discovery*. 1999.

[4] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 1991.

[5] M. H. DeGroot. *Probability and Statistics*. Addison Wesley, Reading, MA, 1989.

[6] N. Friedman, M. Goldszmidt, and A. Wyner. Data analysis with Bayesian networks: A bootstrap approach. In K. Laskey and H. Prade, editors, *Proc. Fifthteenth Conference on Uncertainty in Artifi cial Intelligence (UAI '99)*, pages 206–215. Morgan Kaufmann, San Francisco, Calif., 1999.

[7] N. Friedman and D. Koller. Being Bayesian about Bayesian network structure: A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, 2001. Accepted for publication. Earlier version appeared in UAI'2000.

[8] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7:601–620, 2000.

[9] D. Heckerman. A tutorial on learning with Bayesian networks. In M. I. Jordan, editor, *Learning in Graphical Models*. Kluwer, Dordrecht, Netherlands, 1998.

[10] D. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

[11] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend. Functional discovery via a compendium of expression profi les. *Cell*, 102(1):109–26, 2000.

[12] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, Calif., 1988.

[13] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge Univ. Press, 2000.

[14] J. Pearl and T. S. Verma. A theory of inferred causation. In

KR '91, pages 441–452. 1991.

[15] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search*. Number 81 in Lecture Notes in Statistics. Springer-Verlag, 1993.