BIOINFORMATICS

Vol. 20 no. 18 2004, pages 3565–3574
doi:10.1093/bioinformatics/bth445

Discovery of meaningful associations in genomic data using partial correlation coefficients

*Alberto de la Fuente**, *Nan Bing*†, *Ina Hoeschele and Pedro Mendes*

Virginia Polytechnic Institute and State University, Virginia Bioinformatics Institute, 1880 Pratt Drive, Blacksburg, Virginia, 24061 USA

Received on June 2, 2004; revised on July 15, 2004; accepted on July 24, 2004
Advance Access publication July 29, 2004

ABSTRACT

Motivation: A major challenge of systems biology is to infer biochemical interactions from large-scale observations, such as transcriptomics, proteomics and metabolomics. We propose to use a partial correlation analysis to construct approximate Undirected Dependency Graphs from such large-scale biochemical data. This approach enables a distinction between direct and indirect interactions of biochemical compounds, thereby inferring the underlying network topology.

Results: The method is first thoroughly evaluated with a large set of simulated data. Results indicate that the approach has good statistical power and a low False Discovery Rate even in the presence of noise in the data. We then applied the method to an existing data set of yeast gene expression. Several small gene networks were inferred and found to contain genes known to be collectively involved in particular biochemical processes. In some of these networks there are also uncharacterized ORFs present, which lead to hypotheses about their functions.

Availability: Programs running in MS-Windows and Linux for applying zeroth, first, second and third order partial correlation analysis can be downloaded at: http://mendes.vbi.vt.edu/tiki-index.php?page=Software

Contact: alf@vbi.vt.edu

Supplementary information: Supplementary information can be found at: URL to be decided

INTRODUCTION

Inferring the topology of biochemical networks, including metabolic networks and gene networks, rests mainly on the ability to distinguish direct from indirect interactions. Several methods for inference of such networks from experimental data have been proposed in the recent literature (Brazhnik *et al*., 2002). Some require very specific experimental designs (de la Fuente *et al*., 2002; Gardner *et al*., 2003), others are less stringent on experiments but rely on very specific assumptions about the underlying network topology; e.g. sometimes it is assumed that biochemical networks can be modeled as directed acyclic graphs (Friedman *et al*., 2000; Wagner, 2001). However, cyclic network structures, such as feedback loops, are ubiquitous in biology and are associated with many of the specific properties of living systems, and therefore analyses should be independent of such assumptions.

We propose a method to construct approximate undirected dependency graphs (UDGs) from large-scale biochemical data using partial correlation coefficients. UDGs are graphs in which pairs of vertices are connected by undirected edges if there is a direct dependence between them (Shipley, 2002). Because the graphs constructed by this method are undirected, many of the problems that arise in inferring networks with cycles are circumvented. Previously, methods based on the same framework have been proposed (Spirtes *et al*., 1993; Pearl, 2000; Shipley, 2002) but which are computationally intractable for large-scale data sets. The present method is simpler than those, but is able to efficiently analyze genome-sized data sets, such as from microarray experiments. The method starts by constructing networks based on correlations, where vertices are biochemical species and edges correspond to their correlation; an edge is present if the correlation between any two biochemical species is higher than a certain threshold. Since correlation is symmetrical, edges are undirected. In a second step edges for which partial correlation coefficient falls below a certain threshold are eliminated, resulting in an undirected dependency graph.

SYSTEM AND METHODS

The Pearson product moment correlation coefficient is a widely used measure of association between continuous random variables. As is well known, correlation should not be confused with causality, since many different causal relationships can correlate the same pair of variables. The use and interpretation of zero-order correlation networks in 'omics' studies has been discussed thoroughly earlier (Eisen *et al*., 1998; Steuer *et al*., 2003a,b). Although it is clear that correlation networks are not the same as the underlying causal

*To whom correspondence should be addressed.

†Present address: GlaxoSmithKline, Five Moore Drive, Research Triangle Park, North Carolina, 27709

networks, correlation is still informative about the underlying system. What causal properties can be inferred from studying correlations has been well investigated before (Spirtes *et al*., 1993; Pearl, 2000; Shipley, 2002). In this paper, we explore what can be gained from studying correlations in genomic data sets. The most important concept in this study is the partial correlation coefficient. A partial correlation coefficient quantifies the correlation between two variables (e.g. gene activities) when conditioning on one or several other variables. For example, what exactly is the correlation $r_{xy.z}$ between variables x and y conditioning on z? It is the correlation between the parts of x and y that are uncorrelated with z. To obtain these parts of x and y, they are both regressed on z. The residuals of the regression are then the parts of x and y that are uncorrelated with z. The correlation between these residuals of x and y is the partial correlation between x and y when conditioning on z. The order of the partial correlation coefficient is determined by the number of variables it is conditioned on. For example, $r_{xy.z}$ is a first-order partial correlation coefficient, because it is conditioned solely on one variable (z). Partial correlation can be calculated to any arbitrary order. Equations (1)–(3) allow the calculation of partial correlation coefficients of orders 0–2 and similar equations exist to calculate higher-order coefficients.

$$\text{zeroth-order correlation: } r_{xy} = \frac{\text{cov}(xy)}{\sqrt{\text{var}(x)\text{var}(y)}} \tag{1}$$

$$\text{first-order correlation: } r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}} \tag{2}$$

$$\text{second-order correlation: } r_{xy.zq} = \frac{r_{xy.z} - r_{xq.z}r_{yq.z}}{\sqrt{(1 - r_{xq.z}^2)(1 - r_{yq.z}^2)}} \tag{3}$$

Thus we can use partial correlation coefficients to distinguish between the correlations between two variables due to direct causal relationships from the correlations between the same two variables that originate *via* intermediate variables (sequential pathways) or directly due to other variables (common causes). Although partial correlation analysis still does not infer causal relationships, it excludes many of the possibilities, and thus is a step in the direction of causal inference. We propose to calculate up to second-order partial correlation coefficients to infer significant interactions from biochemical data sets, including those containing transcriptomics, proteomics and metabolomics data. The correlation between two variables is evaluated by conditioning on all possible pairs of other variables. If any of these pairs yields a zero partial correlation (or a correlation not significantly different from zero), that edge is removed from the correlation network. Executing this over all possible edges results in a network of putative direct interactions. We refer to the graph obtained in this way as a second-order UDG approximation. In order to obtain the exact UDG for n variables, one would potentially need to calculate all partial correlation for each order from 0 to $n - 2$, i.e. one would need to calculate correlation coefficients conditioned on every possible subset of the set of $n - 2$ other variables (Shipley, 2002). This because there can be more than two indirect paths between two variables resulting in correlation. In order to do so, one needs at least as many observations as there are variables. This is thus not possible for existing data sets containing hundreds to tens of thousands of variables but where there are only in the order of ten to hundreds of observations—a well-known problem of genomics. An alternative approach would be to subdivide the data set in smaller subsets of variables and calculate the partial correlations in these subsets conditioning on all other variables of the subset, yielding many different sub-networks (Kishino and Waddell, 2000; Shinohara *et al*., 2000; Waddell and Kishino, 2000). This approach, however, suffers from yet another problem: How to reconstruct the entire network from these sub-networks? Another approach would be to cluster the genes in a small number of clusters and find the network between the clusters (Toh and Horimoto, 2002). The resolution of the latter approach is low, since clusters can consist of many genes. Furthermore, in these approaches (Kishino and Waddell, 2000; Shinohara *et al*., 2000; Waddell and Kishino, 2000; Toh and Horimoto, 2002) the conditioning is done only on the full sets, and lower-order partial correlations are not considered—but two variables may be independent when only conditioned on a subset of the variables, while dependent when conditioned on all other variables together. Therefore, these methods are unlikely to discover all independencies in the network (see below). We propose to 'remove' the two most active paths, which we argue reduces the correlation sufficiently so that it falls below the threshold of significance, and therefore the pair will be seen as independent. This means that one only needs to calculate up to second-order partial correlation coefficients. In this case we favor a high threshold value, based on a high significance level. Then, we expect only a few false positives (Type I errors; connections inferred which do not correspond to a connection in the real network) with the drawback of a higher number of false negatives (Type II errors; connections in the real network not inferred). With this approach one is at least quite certain of the existence of the edges found, though many can be missed. Accordingly, this should not be seen as a 'network inference' approach: the goal is not to infer the network correctly, but instead to develop, with confidence, new hypotheses of interactions between biochemical components.

As an alternative to Pearson correlation, this analysis could employ Spearman rank correlation in Equations (1)–(3) [see section 3.9 in Shipley (2002)] for justification of evaluating conditional independence with test based on Pearson partial correlation, but replacing these with Spearman partial correlations). Spearman rank correlation does not depend on normality and linearity of interactions, and might therefore be better suited for biochemical networks. The accompanying software has the ability to carry out the analysis based

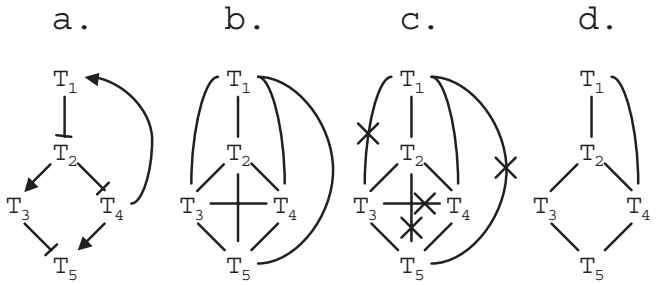

Fig. 1. Stepwise description of the method. (**a)** The actual causal network. (**b)** The UDG based on zero-order correlation. (**c)** Edges with zero partial correlation coefficients are eliminated. (**d)** The resulting UDG.

on Pearson or Spearman rank correlations, at the user's choice.

ALGORITHM

As an explanatory example we consider a simple gene network model, whose dynamics are described by the following system of simultaneous ordinary differential equations:

$$\begin{aligned}
\frac{dT_1}{dt} &= \frac{V_1}{(1+K_{T_1}/T_4)} - k_1T_1 + \theta_1T_1,\\
\frac{dT_2}{dt} &= \frac{V_2}{(1+T_1/K_{T_1})} - k_2T_2 + \theta_2T_2,\\
\frac{dT_3}{dt} &= \frac{V_3}{(1+K_{T_2}/T_2)} - k_3T_3 + \theta_3T_3,\\
\frac{dT_4}{dt} &= \frac{V_4}{(1+T_2/K_{T_2})} - k_4T_4 + \theta_4T_4,\\
\frac{dT_5}{dt} &= \frac{V_5}{(1+T_3/K_{T_3})(1+K_{T_4}/T_4)} - k_5T_5 + \theta_5T_5
\end{aligned} \tag{4}$$

Parameters V_i are maximal transcription rates, k_i are degradation rate constants and the K_i are inhibition or activation constants; all parameter values are set to unity. The T_is are transcript levels (gene activities) and the θ_i are error terms designed to simulate biological variance. The error terms are initialized by sampling from a normal distribution with zero mean and a standard deviation of 0.01, and then a steady state is calculated numerically with the software Gepasi (Mendes, 1993, 1997). We generated 1000 steady states, which differ only due to the random values of θ_i. Figure 1a depicts the network structure.

First the sample correlation matrix is calculated and a network is drawn in which the variables that have significant correlation are connected. In this example all variables are significantly correlated (given 1000 observations and using an alpha value of 0.01; see the supplementary information for details on how the minimum significant correlation was determined), giving rise to a totally connected graph (Fig. 1b). Testing the correlation between T_1 and T_2 by conditioning on all other variables individually, and on all combinations of two variables, shows that this correlation remains significant and so corresponds to a direct dependency and its edge is therefore not removed. Although T_1 and T_3 are not adjacent in the original network (Fig. 1a), they are highly correlated. This is caused by the path between T_1 and T_3 through T_2. Calculation of the partial correlation between T_1 and T_3 conditioning on T_2 ($r_{T_1T_3.T_2}$) shows that this correlation vanishes below significance (it has a value of -0.0316). This prompts the removal of the edge between T_1 and T_3 from the inferred network, and one does not continue conditioning this correlation on any other variables. In a similar way we test all other correlations, find that some are not significant, and remove the corresponding edges from the graph (Fig. 1c). Also T_2 and T_5 are correlated. There are two paths running from T_2 to T_5, one through T_3 and the other through T_4. If we condition only on T_3 the correlation is reduced to -0.314, which is still significant. Conditioning on T_4 yields a reduction to -0.164, but this is also still significant. In this case calculation of a second-order partial correlation coefficient is necessary to find that this pair is independent, i.e. $r_{T_2T_5.T_3T_4}$, which has a value of -0.0194, which is no longer significant.

Systematically testing all pair-wise correlations by conditioning first on all other variables individually and subsequently on all possible pairs, and removing edges as soon as they have non-significant partial correlation (Fig. 1c) results in the undirected dependence graph depicted in Figure 1d. It must be noted that, even for this simple example, conditioning on all other genes together as in previous approaches (Kishino and Waddell, 2000; Shinohara *et al*., 2000; Waddell and Kishino, 2000; Toh and Horimoto, 2002) will not discover the independence between T_3 and T_4. This is because of the fact that T_2 is a common ancestor of T_3 and T_4, while T_5 is a causal descendent of T_3 and T_4. Conditioning on any common causal descendent introduces a correlation between two variables that are independent conditional on their causal ancestors. Therefore, conditioning on all variables simultaneously can introduce some dependencies, which are not due to direct causal effects or common ancestors. The correlation matrix analyzed above was processed at the ASIAN (Aburatani *et al*., 2004)) website (http://eureka.ims.u-tokyo.ac.jp/asian/), at which the algorithm of (Toh and Horimoto, 2002) is implemented. The results were similar to those of Figure 1d, but with an additional incorrect edge between T_3 and T_4. This emphasizes the need to consider lower-order partial correlations first.

There are ways to direct some of the edges of this graph (Spirtes *et al*., 1993), but these are limited and we will not pursue this here. In this example, calculating two orders of partial correlation suffices to detect all independencies. For more complicated networks, higher-order correlations may be necessary. We expect that in general conditioning on two variables (disrupting two paths of influence) should be enough

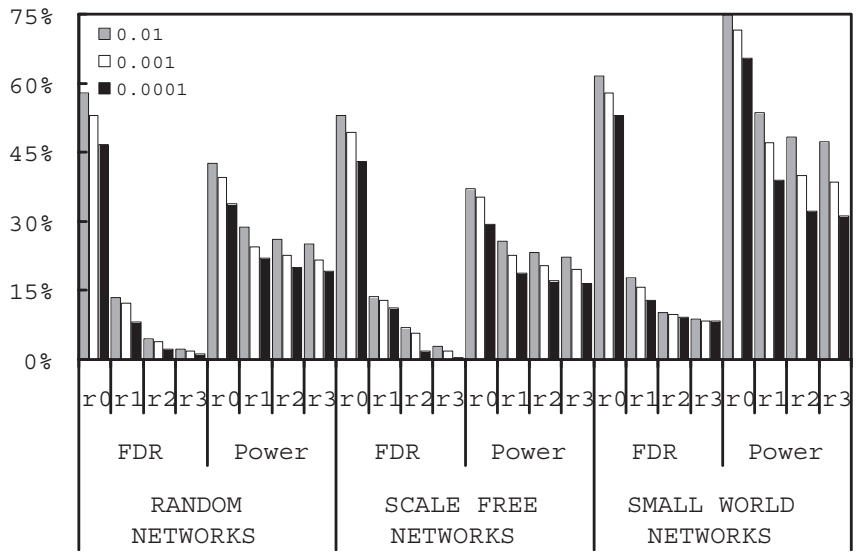

Fig. 2. Average False Discovery Rate (FDR) and power found with partial correlation analysis of different orders. r0 is zero-order partial correlation (Pearson correlation), r1, r2, and r3 are first-, second- and third-order partial correlations, respectively. Results are given at different threshold values for three different network topologies. For each topology 50 networks of 100 genes and 200 connections were analyzed.

to reduce most indirectly caused correlations below the significance threshold. By selecting a high confidence level, we ensure that conditioning on only two variables is enough to reduce the correlation below the significance threshold, but with the drawback of losing those direct connections that are weak. Controlling for more than two variables at a time would also dramatically increase the computation time and limit the application to small networks.

The theoretical basis of the present approach rests on the operation of d-separation (Pearl, 2000). In simple terms, two variables are said to be d-separated if there exists a conditioning set of variables that prevents a flow of information between the two. The definition of d-separation was originally made by Pearl (Pearl, 2000), while Shipley provides a more accessible explanation (Shipley, 2002). It has been mathematically proven for Directed Acyclic Graphs that d-separation implies statistical independence (Spirtes *et al*., 1993). In cyclic networks this might not always hold (Spirtes, 1995).

RESULTS

Results from artificial data

We evaluate the performance of the algorithm on simulated data from large artificial gene networks (Mendes *et al*., 2003). Although recently arguments were presented in support of the view that biochemical networks follow a 'scale-free' (Barabasi and Albert, 1999) topology, the actual global network architecture of biochemical networks is still largely unknown. In order to show that the present method is robust toward network topology, we tested it on different simulated network architectures. We tested 50 networks of 100 genes and 200 connections for each of three different topologies: 'random' (Erdös and Rényi, 1960; Kauffman, 1969), 'scale free' (Barabasi and Albert, 1999) and 'small world' (Watts and Strogatz, 1998). Note that these networks contain an arbitrary number of cycles. These models were constructed with the system previously described by one of us (Mendes *et al*., 2003), which is based on ordinary differential equations with non-linear kinetics similar to Equation (4). Details of the networks can be found in (Mendes *et al*., 2003) and on the web at: http://mendes.vbi.vt.edu/AGN/Century/index.html. For each gene network, mutant experiments were simulated by setting each gene's transcription rate to 50% of its original value (one at a time), so the sample size for each network is 100 (corresponding to 100 single gene perturbations). This mutant pool represents the necessary biological variance needed to apply this method. The interactions in these models are defined with non-linear kinetics (Mendes *et al*., 2003), so in this exercise we also test the performance of the method on non-normal data. For correlation of order 0, 1, 2 and 3, the False Discovery Rate (FDR) is calculated at different threshold values as well as the power. The FDR is expressed as the number of wrongly predicted edges divided by the total number of predicted edges, and the power is defined as the number of edges correctly inferred as a fraction of the total number of edges in the network.

Figure 2 summarizes the results at several threshold levels; the exact numbers and standard deviations can be found in Tables S1 and S2 in the supplementary data. Interestingly, the method performed differently on each topology, having more success with the Erdös-type random networks. A slightly higher FDR was found for the scale-free topology, but fewer

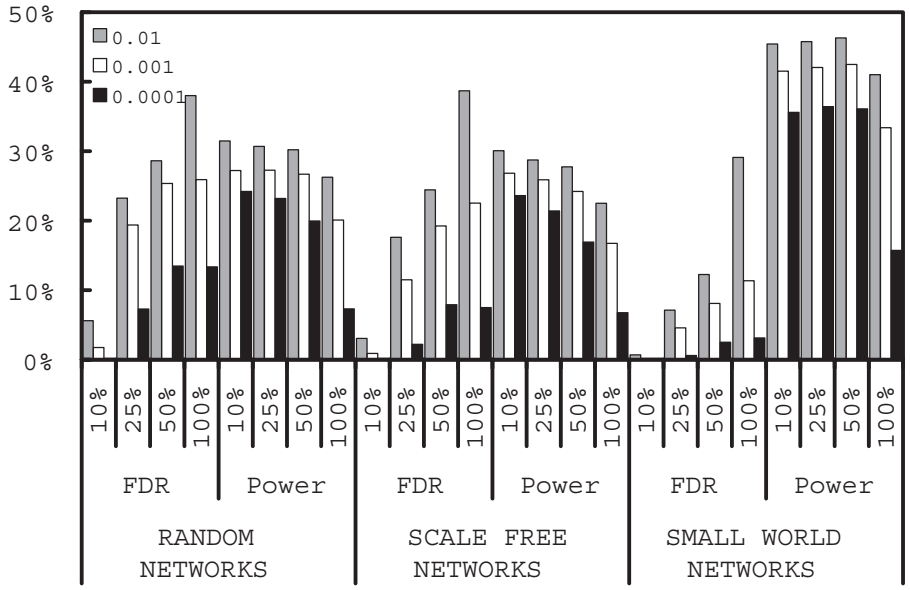

Fig. 3. Average FDR and power for three networks of different topologies, with different levels of added noise. One hundred data sets per noise level per topology were generated and analyzed with partial correlation analysis up to second order.

edges were discovered, i.e. there were more false negatives. The results for the small-world topology were not as good, although even for these there is still a considerable improvement compared to considering zero-order correlation only. The method produces many false negatives, i.e. misses many connections that exist in the network, but mostly because they represent weak interactions. Only 20 to 40% of the total number of edges was retrieved in this exercise. Figure 2 depicts the effect of increasing order in the partial correlations, and it is evident that increasing the order from two to three improves the discovery only slightly. This suggests that calculating up to second-order partial correlation at an alpha level of 0.001 is a reasonable choice for confident inference of biochemical interactions (networks).

To test the effect of experimental error on the performance of the method, we selected one network from each topology and added noise to their data sets. The method had not generated false positives without noise, so all false positives we find in the presence of noise are due indeed due to the noise. Since this method makes use of the information contained in the biological variance and covariance in data sets, its performance depends on *how large the experimental noise is relative to the biological variance*. Noise was added by sampling from a normal distribution with zero mean and a variance corresponding to 10, 25, 50 and 100% of the biological variance of each variable. A hundred data sets per noise level per topology were generated, then the method was applied with first- and second-order partial correlation, and finally the FDR and power were calculated. Figure 3 summarizes the results at several threshold levels; the exact numbers and standard deviations can be found in Tables S3 and S4 in the supplementary data. We observed that the higher the proportion of experimental variance relative to the biological variance, the more false positives the method produces. However, the FDR remains quite acceptable even at higher noise levels. This is due to the higher noise levels causing a loss in correlations corresponding to real edges (resulting in lower power), but also a similar rate of loss of false positives. Inference on the scale-free and especially the small-world topologies is quite robust to the noise level, both in terms of the FDR as well as the power.

Results from *Saccharomyces cerevisiae* microarray data

We combined budding yeast gene expression data from two previous publications (Brem *et al*., 2002; Yvert *et al*., 2003). In this data set, allelic polymorphism at multiple *loci*, resulting from a cross of two inbred lines, causes variance of expression levels. Two of us (N. Bing and I. Hoeschele, submitted for publication) have recently reanalyzed this data set in a genetical genomics (Jansen and Nap, 2001) study. From this study, based on 42 observations (Brem *et al*., 2002; Yvert *et al*., 2003), candidate causal links between genes were identified from QTL analyses of expression profiles. The resulting networks included a total of 781 genes, which form the set of genes included in the present study. For the partial correlation analysis, the data is augmented with 88 additional measurements (Yvert *et al*., 2003), giving a total of 130 observations. Each of these 130 observations corresponds to a particular offspring as a result of the cross between the two parental strains. The parental strains are homozygous for different variants at a number of sites throughout the genome. The offspring have

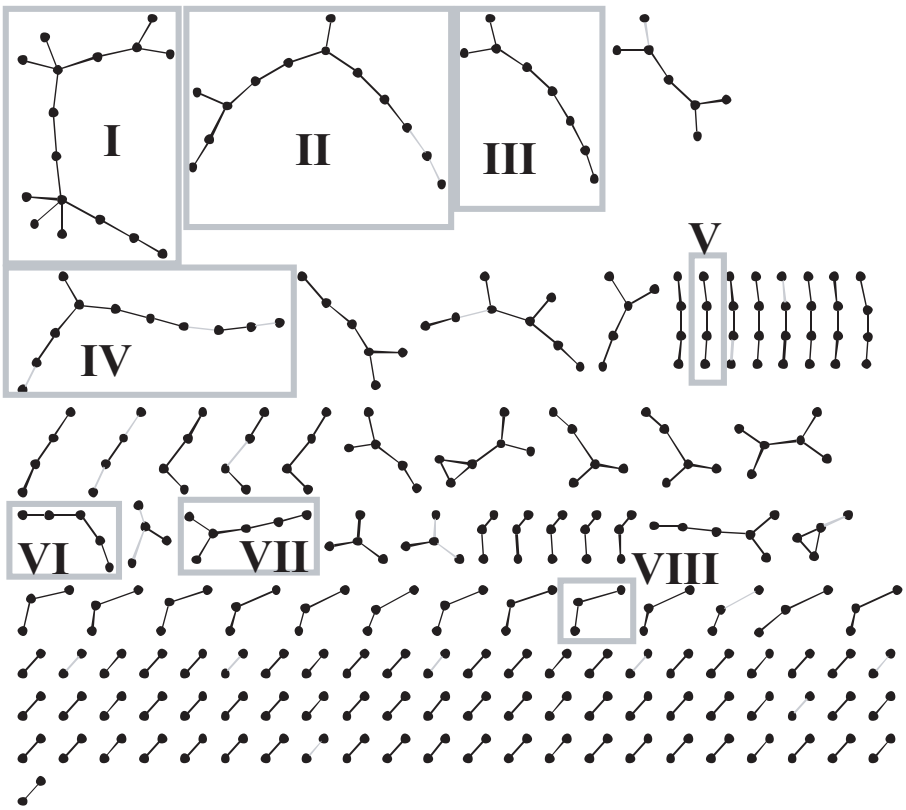

Fig. 4. The Undirected Dependence Graph comprising 374 yeast genes and 258 direct relationships, resulting from a second-order partial correlation analysis at confidence level alpha = 0.001. Black and gray edges correspond to positive and negative second-order partial correlation, respectively. Figure created with the program Cytoscape (Shannon *et al*., 2003). Boxes indicate sub-networks of known function. Box I contains mostly genes related to mitochondrial protein biosynthesis and respiration. Box II contains many genes involved in sterol and lipid biosynthesis. Box III contains many genes involved in oxidative phosphorylation and nucleotide metabolism. Box IV contains genes involved in the pheromone response pathway. Box V contains genes involved in amino acid biosynthesis, Box VI contains genes involved in protein biosynthesis, Box VII contains genes involved in cell growth and Box VIII contains three genes that code for ATP-driven ion transporters.

inherited different combinations of variants in their genomes, leading to different expression levels. The variance of gene expression among the offspring is then used to infer network structures. Figure 4 shows the second-order correlation network resulting from an alpha value of 0.001. At that threshold level, and with 130 observations, second-order correlations above the threshold of 0.348 (see supplementary information) are considered to be non-zero. The resulting network is very sparse with only 258 connections (from a total of 304,590 possible pair-wise interactions), and contains only 374 out of the 781 genes that were included in the analysis (Figure 4). The network obtained using an alpha value of 0.01 contains more connections and variables and an alpha value of 0.0001 is sparser and contains less variables (see supplementary information). The network at an alpha value of 0.001 contains many small sub-networks of just two and three genes, as well as larger sub-networks. Verifying the biological relevance of the recovered networks is difficult, since many interactions between genes are currently unknown. These gene networks are also phenomenological, i.e. many connections do not correspond to direct physical interactions between gene product and promoter elements, but to a complicated action through more complex regulatory pathways involving the proteome and metabolome (Brazhnik *et al*., 2002; de la Fuente and Mendes, 2002). One way to verify the biological relevance of the inferred networks is by using the Gene Ontology Term Finder of the *Saccharomyces* Genome Database (Christie *et al*., 2004) at http://genome-www4.stanford.edu/cgi-bin/SGD/GO/goTermFinder to investigate if the sub-networks contain a high proportion of functionally related genes. We found that most of the inferred networks had indeed high significance scores, implying that the probability of grouping them by chance is very small, thus the method seems to uncover relevant information. Most of the genes in the network of Box I (Figure 4) are related to mitochondrial protein biosynthesis and respiration. Interestingly, grouped with these genes are two uncharacterized ORFs: YHR116W, whose knockout mutant showed growth defects on non-fermentable carbon

source, and YLR253W, whose knockout mutant showed severe growth defects on minimal media or lactate. We hypothesize that these genes are involved in mitochondrial protein biosynthesis and/or respiration. The network in Box II of Figure 4 contains many genes involved in sterol and lipid biosynthesis. There is one uncharacterized ORF in this group: YLR050C. It is directly connected to ERG28/YER044C and could thus be an unknown activator of ERG28. The cascade in Box III contains many genes involved in oxidative phosphorylation and nucleotide metabolism. The genes in the network in Box IV are involved in the pheromone response pathway. Included in this group are two uncharacterized ORFs: YCR097W-A and YKL177W. Both these ORFs are classified as 'dubious' in the *Saccharomyces* Genome Database (Christie *et al*., 2004), but these results indicate that they may be involved in the pheromone response. Many other networks contain just few high-scoring genes grouped with genes involved in distinct processes. There is some support that the small sub-networks also depict potentially significant interactions; Box VIII contains three genes that code for ATP-driven ion transporters, Box V contains genes involved in amino acid biosynthesis, Box VI contains genes involved in protein biosynthesis and Box VII genes involved in cell growth. Many other sub-networks contain genes belonging to unknown biological processes; our result indicates that they may interact with each other or be co-regulated. This method finds interactions between genes and is not necessarily intended to group genes based on similar biochemical function.

DISCUSSION

We propose a simple and efficient method to organize genes in UDGs starting from large-scale biochemical data, such as obtained from microarrays. This is based on partial correlation coefficients up to order 2. These UDGs would ideally correspond to the underlying network of direct interactions between the biochemical compounds. Due to the limited amount and low accuracy of the data being generated at present, the results of this approach should be seen as an initial estimate of the real underlying network, enabling us to develop new hypotheses for interactions between biochemical components. We demonstrated the application of the method to gene expression data, but it is worth stressing that it would work equally well on data from metabolomics or proteomics.

The power of the method presented here is not very high. From studies with simulated data we found that only 20–40% of the total number of interactions were retrieved. This is due to correlation not being an ideal way to quantify biochemical interactions. Better measures for quantification of genetic interactions exist and inference methods based on these measures have higher power (de la Fuente *et al*., 2002; di Bernardo *et al*., 2004). However, these methods require specific and complicated experiments (Gardner *et al*., 2003), while the present method can be applied to a wider variety of experimental data. The present method uses observational data, which is an advantage over methods that require costly and complicated experimental setups (e.g. de la Fuente *et al*., 2002; Gardner *et al*., 2003). Observational data is obtained by measuring individuals in the same physiological state, only different due to biological variance that distributes the individuals around the mean state. Also, data obtained from experimental interventions that are not too drastic (like small temperature changes, small changes of medium) will still allow the linear approximation of biochemical interactions and can thus be used. The large set of simulated data analyzed in this paper was generated in such experimental fashion, by creating small perturbations of the expression of each gene individually. The method discovers connections with a low FDR, so although one can only find a proportion of all interactions, one can be fairly confident about the identified links.

Microarrays measure the gene expression of millions of cells simultaneously, and what is measured is thus the average (or aggregated) gene expression rather than the gene expression inside single cells (which represent the actual causal process). It was shown that the independence relations hold under aggregation for variables that interact linearly (Chu *et al*., 2003). While gene networks, and biochemical networks in general, are better characterized by non-linear kinetics, for small deviations from the mean values one can expect that the non-linear interactions can be validly approximated by linear functions, as is commonly done in systems analysis (Stucki, 1978). Therefore, we believe that the recent success of application of methods based on finding independence in gene expression data (e.g. Friedman *et al*., 2000; Gardner *et al*., 2003; Segal *et al*., 2003) (and the approach described in this paper) can be attributed to the approximate validity of the assumptions.

Similar algorithms to identify UDGs have been proposed earlier (Spirtes *et al*., 1993; Kishino and Waddell, 2000; Pearl, 2000; Shinohara *et al*., 2000; Waddell and Kishino, 2000; Shipley, 2002; Toh and Horimoto, 2002). The present method is faster and able to deal with large data sets. The reason for this scalability is that only up to second-order partial correlations are calculated, while other algorithms proceed to complete conditioning order if necessary. Although better, these other algorithms cannot be applied to data sets as large as those currently produced in 'omics' studies and as analyzed here. Another reason for the increased speed of our algorithm is that we calculate the matrix of second-order partial correlations conditioned on p and q immediately after calculating the matrix of first-order partial correlations conditioned on p, thus using the first-order correlation coefficients recursively in the calculation of the second-order coefficients, using Equation 3. Previously proposed algorithms (Spirtes *et al*., 1993; Kishino and Waddell, 2000; Pearl, 2000; Shinohara *et al*., 2000; Waddell and Kishino, 2000; Shipley, 2002;

Toh and Horimoto, 2002) compute partial correlation coefficients by creating many sub-covariance matrices and inverting them or by first applying (multiple) regression and calculating correlation between residuals. Systematically using Equations 2 and 3 prevents us from using computationally expensive tasks such as matrix inversion.

The limitation of our method to correlations of up to order two is justified on several grounds. The first reason is the usually small sample sizes in functional genomics, making higher-order coefficients unreliable. Second, our studies with simulated data indicate that little is gained by increasing to third order (Fig. 2). A practical reason is that for data sets of about 1000 variables, as analyzed here, the computation would become prohibitively intensive for higher orders. In the worst case the computational complexity is about $\mathbf{O}(n^{p+2})$, where n is the number of variables and p the order of the correlation used (see supplementary information). This worst-case scenario would occur when the network is totally connected, since it would test all conditionings to find independence. Fortunately, biochemical networks are sparse and thus most correlations will be close to zero. Once a certain (partial) correlation is evaluated as zero, no further computations are performed for this pair of variables. The results on the 781-gene data set were obtained with a Dell PC 1.90 GHz running for slightly over 2 h, but when calculating all second-order partial correlations it took almost 23 h. Supercomputing facilities and parallel implementation of the algorithm would perhaps enable extending the analysis to include higher orders of partial correlation and to analyze larger data sets. But its advantage is still put in question by our results that show little gain in information when going from order 2 to 3 (Fig. 2).

A non-zero partial correlation between two variables, X and Y, can imply several causal mechanisms.

(1) X directly affects Y, or Y directly affects X, or both
(2) X and Y are both affected by a third 'hidden' variable,
(3) X becomes conditionally correlated to a hidden variable affecting Y (or vice versa) (there is an inducing path between them).

Based on partial correlation analysis alone it is not possible to distinguish between these three possibilities. The existence of hidden variables makes the interpretation of partial correlation difficult and therefore one should include in the analysis as many biochemical variables as possible. If a partial correlation between two genes were due to mechanism 1, it would simply imply that one of them is a direct effector of the other or that the both directly affect each other. Mechanism 2 corresponds to the case that both genes are co-regulated by a common factor, which was not included in the measurements. In this case the two genes are direct neighbors in the interaction graph, while in reality they are neighbors of degree 2 in the real underlying biological network. Mechanism 3 may lead to more severe mistakes, since it can result in finding a connection between two genes that are further removed from each other due to the conditional correlation with a hidden variable.

The advantage of this method based on partial correlations over Bayesian networks is that it is conceptually simpler and requires less computational effort, when restricted to the calculation of up to second-order partial correlation. Furthermore, in the process of distinguishing between direct and indirect interactions, using the Bayesian network approach one has to propose a causal structure in order to evaluate its likelihood. This causal structure needs to be a Directed Acyclic Graph (DAG), which is in contradiction to the known structure of biochemical networks, as feedback is ubiquitous. A third advantage is that the partial correlation approach is appropriate for continuous variables, while most Bayesian network approaches require that the data be discretized, thereby losing information and posing a problem of how the discretization should be made. On the other hand, Bayesian network approaches generate directed graphs, while the approach proposed here yields an undirected graph. While there are ways to partially direct the undirected graph, these are mostly dependent on the assumption that the graph is a DAG (Spirtes *et al*., 1993; Pearl, 2000) as well. One exception, able to deal with directions in cycles has been proposed (Spirtes *et al*., 1993; Shipley, 2002) but is limited to simple cycles of two variables and yields equivalent classes of structures that cannot be distinguished based on this type of data. Recently, a first application of Structural Equation Modeling (SEM) (Bollen, 1989; Xiong *et al*., 2004) to microarray data has been presented (Bollen, 1989; Xiong *et al*., 2004). Linear acyclic SEMs are equivalent to Bayesian networks, but linear SEMs can handle cyclic structures, and non-linear SEMs might be even more suitable for biochemical networks.

The most straightforward strategy to find out directions is by studying time series (Chevalier *et al*., 1993; Arkin *et al*., 1997; Díaz-Sierra *et al*., 1999; Vance *et al*., 2002; Torralba *et al*., 2003). Once we have obtained an UDG it is straightforward to propose specific time series experiments that would help clarify the directions of the edges. Starting with the UDG will greatly reduce the number of necessary time series experiments, as compared to designing these experiments without prior information (Torralba *et al*., 2003). Similarly, the UDG can assist in reducing the number of experiments needed in a quantitative approach based on perturbation experiments and measurements of steady state responses (de la Fuente *et al*., 2002). This approach can also be used for selection of subsets of variables to include in further causal analysis, such as Structural Equation Modeling and complete partial correlation analysis, using software such as TETRAD (Spirtes *et al*., 1993).

CONCLUSION

The results obtained with the yeast data set show promise that this method is indeed useful. Not only were we able to

recover known biochemical interactions, but we were also able to hypothesize that a number of genes with unknown function could be related to specific biological functions. It remains to be seen whether these hypotheses are correct, but their generation allows one to design appropriate experiments in order to confirm or refute them. Overall this information is exactly what we need to obtain from large-scale functional genomics studies.

REFERENCES

Aburatani,S., Goto,K., Saito,S., Fumoto,M., Imaizumi,A., Sugaya,N., Murakami,H., Sato,M., Toh,H. and Horimoto,K. (2004) ASIAN: a web site for network inference. *Bioinformatics*, **16**, 2853–2856.

Arkin,A., Shen,P. and Ross,J. (1997) A test case of correlation metric construction of a reaction pathway from measurements. *Science*, **277**, 1275–1279.

Barabasi,A. and Albert,R. (1999) Emergence of scaling in random networks. *Science*, **286**, 509–512.

Bollen,K. (1989) *Structural Equations with Latent variables*. Wiley-Interscience, NY, USA.

Brazhnik,P., de la Fuente,A. and Mendes,P. (2002) Gene networks: how to put the function in genomics. *Trends Biotechnol.*, **20**, 467–472.

Brem,R., Yvert,G., Clinton,R. and Kruglyak,L. (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752–755.

Chevalier,T., Screiber,I. and Ross,J. (1993) Toward a systematic determination of complex reaction mechanisms. *J. Phys. Chem.*, 6776–6787.

Christie,K., Weng,S., Balakrishnan,R., Costanzo,M., Dolinski,K., Dwight,S., Engel,S., Feierbach,B., Fisk,D., Hirschman, J. *et al*. (2004) Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res.*, **1**, D311–D314.

Chu,T., Glymour,C., Scheines,R. and Spirtes,P. (2003) A statistical problem for inference to regulatory structure from associations of gene expression measurements with microarrays. *Bioinformatics*, **19**, 1147–1152.

de la Fuente,A. and Mendes,P. (2002) Quantifying gene networks with regulatory strengths. *Mol. Biol. Rep.*, **29**, 73–77.

de la Fuente,A., Brazhnik,P. and Mendes,P. (2002) Linking the genes: inferring quantitative gene networks from microarray data. *Trends Genet.*, **18**, 395–398.

di Bernardo,D., Gardner,T.S. and Collins,J. (2004) Robust identification of large genetic networks. *Proc. Pac. Symp. Biocomp.*, **9**, 486–497.

Díaz-Sierra,R., Lozano,J.B. and Fairén,V.J. (1999) Deduction of chemical mechanisms from the linear response around steady state. *J. Phys. Chem.*, **103**, 337–343.

Eisen,M., Spellman,P., Brown,P. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci., USA*, **95**, 9212–9217.

Erdös,P. and Rényi,A. (1960) On the Evolution of Random Graphs. *Publ. Math. Inst. Hungar. Acad. Sci.*, **5**, 17–61.

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

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

Jansen,R. and Nap,J. (2001) Genetical genomics: the added value from segregation. *Trends Genet.*, **17**, 388–391.

Kauffman,S. (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.*, **22**, 437–467.

Kishino,H. and Waddell,P.J. (2000) Correspondence analysis of genes and tissue types and finding genetic links from microarray data. *Genome Inform Ser Workshop Genome Inform.*, **11**, 83–95.

Mendes,P. (1993) GEPASI: a software package for modelling the dynamics, steady states and control of biochemical and other systems. *Comput. Appl. Biosci.*, **9**, 563–571.

Mendes,P. (1997) Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. *Trends Biochem. Sci.*, **22**, 361–363.

Mendes,P., Sha,W. and Ye,K. (2003) Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics*, **19**, Suppl 2: II122–II129.

Pearl,J. (2000) *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK.

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

Shannon,P., Markiel,A., Ozier,O., Baliga,N., Wang,J., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Shinohara,A., Iida,K., Takeda,M., Maruyama,O., Miyano,S. and Kuhara,S. (2000) Finding Sparse Gene Networks. *Genome Inf.*, **11**, 249–250.

Shipley,B. (2002) *Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations and Causal Inference*. Cambridge University Press, Cambridge, UK.

Spirtes,P. Ed. (1995) Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers, San Francisco, CA.

Spirtes,P., Glymour,C. and Scheines,R. (1993) *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, USA.

Steuer,R., Kurths,J., Fiehn,O. and Weckwerth,W. (2003a) Interpreting correlations in metabolomic networks. *Bioch. Soc. Trans.*, **31**, 1476–1478.

Steuer,R., Kurths,J., Fiehn,O. and Weckwerth,W. (2003b) Observing and interpreting correlations in metabolomic networks. *Bioinformatics*, **19**, 1019–1026.

Stucki,J.W. (1978) Stability analysis of biochemical systems. A practical guide. *Progress Biophys. Mol. Biol.*, **33**, 99–187.

Toh,H. and Horimoto,K. (2002) Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. *Bioinformatics*, **18**, 287–297.

Torralba,A.S., Yu,K., Shen,P., Oefner,P.J. and Ross,J. (2003) Experimental test of a method for determining causal connectivities of species in reactions. *Proc. Natl Acad. Sci., USA*, **100**, 1494–1498.

Vance,W., Arkin,A. and Ross,J. (2002) Determination of causal connectivities of species in reaction networks. *Proc. Natl. Acad. Sci., USA*, **99**, 5816–5821.

Waddell,P.J. and Kishino,H. (2000) Cluster inference methods and graphical models evaluated on NCI60 microarray gene expression data. *Genome Inform Ser Workshop Genome Inform.*, **11**, 129–140.

Wagner,A. (2001) How to reconstruct a large genetic network from n gene perturbations in fewer than n(2) easy steps. *Bioinformatics*, **17**, 1183–1197.

Watts,D. and Strogatz,S. (1998) Collective dynamics of 'small-world' networks. *Nature*, **393**, 440–442.

Xiong,M., Li,J. and Fang,X. (2004) Identification of genetic networks. *Genetics*, **166**, 1037–1052.

Yvert,G., Brem,R., Whittle,J., Akey,J., Foss,E., Smith,E., Mackelprang,R. and Kruglyak,L. (2003) Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nat. Genet.*, **35**, 57–64.