

Constructing and Analyzing a Large-Scale Gene-to-Gene Regulatory Network—Lasso-Constrained Inference and Biological Validation

Mika Gustafsson, Michael Hörnquist, and Anna Lombardi

Abstract—We construct a gene-to-gene regulatory network from time-series data of expression levels for the whole genome of the yeast *Saccharomyces cerevisiae*, in a case where the number of measurements is much smaller than the number of genes in the network. This network is analyzed with respect to present biological knowledge of all genes (according to the Gene Ontology database), and we find some of its large-scale properties to be in accordance with known facts about the organism. The linear modeling employed here has been explored several times, but due to lack of any validation beyond investigating individual genes, it has been seriously questioned with respect to its applicability to biological systems. Our results show the adequacy of the approach and make further investigations of the model meaningful.

Index Terms—Biology and genetics, time series analysis, network problems, gene network, network inference, Lasso, yeast, validation, outdegree.

1 INTRODUCTION

CURRENT advances in microarray technologies make it possible to measure mRNA-levels for thousands of genes simultaneously today. Also, large-scale measurements of protein levels are gradually becoming feasible, as well as results on two-hybrid measurements on protein-protein interactions and genome-wide data for DNA binding proteins. These processes have emphasized the need for computational biology in order to get as much information out of such measurements as possible.

One way to handle these data is to infer, or reverse engineer, gene regulatory networks from temporal data. Although still somewhat speculative, researchers are exploring the boundaries for what kind of inference is possible. There are many approaches for network formation, ranging from Boolean circuits to very complicated nonlinear spatial models; see [1] and the references therein. Most models use only transcript data, whereas some incorporate other chemical constituents as well. A model based on mRNA-data only is nothing but an effective network of gene-to-gene interactions. This might look too simplistic in view of the complete network, which includes metabolites, proteins, etc., but it can be thought of as a projection onto the space of genes only. Thus, by focusing only on transcript data, the networks obtained are not biochemical regulatory networks, but gene-to-gene networks where many physical connections

between macromolecules might be hidden by short-cuts, i.e., many intermediate units in regulatory cascades might be hidden [2].

A special class of regulatory network models, which has gained some popularity, is the one of linear, time continuous models. Of course, no one claims there is a linear relationship between the units in a “real” regulatory network. Instead, the working hypothesis is that linear equations can at least capture the main features of the network. The main argument is that many functions can be quite accurately approximated around a specific working point with their linearization. Thus, it can provide a good starting point for further considerations.

A key problem for all models is, however, shortage of data. The number of genes is, in general, much larger than the number of measurements and different authors have taken somewhat different avenues to remedy this obstacle. For the linear continuous model based on transcript data, the first study we are aware of was by D’haeseler et al. [3] and focused on a subset of less than 100 genes that were believed to be interrelated. Their problem was still underdetermined, and they interpolated the data in order to achieve more, simulated, measurements. However, more measurements in the same time-series is an ineffective way of increasing the information content in the data [4]. Another early study was by van Someren et al. [5], who clustered genes into the same number of groups as they had measurements and, thus, obtained a mathematically well-posed problem. Still another approach was explored by Holter et al. [6], who formed networks among the principal components of the data. The biological interpretation of these networks are, however, not totally clear. A more biologically motivated study was performed by Yeung et al. [7], who assumed that the resulting network should be

• The authors are with the Department of Science and Technology, Linköping University (Campus Norrköping), SE-601 74 Norrköping, Sweden.
E-mail: {mikgu, micho, annlo}@itn.liu.se.

Manuscript received 8 Oct. 2004; revised 28 Apr. 2005; accepted 6 May 2005; published online 31 Aug. 2005.

For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-0167-1004.

sparse and that way got a unique solution. Somewhat in the same spirit, van Someren et al. [8] conducted a systematic study on how to incorporate prior knowledge into the inference procedure. One genome-wide inference with almost the same basic dynamical model as we explore was conducted by Dewey and Galas [9]. They considered the whole genome of yeast with more than 6,000 genes and formed a network by taking the solution which minimized the L^2 -norm of the coefficients and set to zero all matrix elements below a certain threshold. However, the resulting network had connections for only 143 genes and although they justify that their result makes biological sense on a small scale, they lack a large-scale analysis. In the same spirit, but with slightly different models, we finally mention the works by Chen et al. [31] and Segal et al. [10]. They both show biologically relevant results, but lack a systematic large-scale analysis. We will discuss the relations between these approaches and our method at the end of this paper.

In the present paper, we utilize one statistical method, the Lasso [11], to reverse engineer a network among ORFs (“Open Reading Frames,” hereafter referred to as “genes”) in the so-called extended Spellman data set. The network is subsequently analyzed with respect to its biological validity. In Section 2, we introduce the data set and the preprocessing steps that are necessary before any inference can be made. In Section 3, the actual network model is introduced and we infer its parameters. Thereafter, in Section 4, we perform an analysis of the obtained graph, especially the outdegrees, both with respect to topological and biological properties. The biological analysis is based on information from *all* known genes, as compiled by the Gene Ontology (GO) database [12]. In Section 5, we have a discussion about the relation between our work and other inference procedures and biological validations. Finally, in Section 6, we conclude the article with a brief discussion about the relevance of work of this kind and give a short outlook.

2 DATA SET AND PREPROCESSING

The extended Spellman data set is one of the most referenced sources of microarray data and contains measured mRNA levels of 6,178 genes for the yeast *S. cerevisiae*, presented as logarithms of the fraction between the measured level and a reference level. The measurements of interest for us are carried out through one or more periods of the cell cycle in four time series—Alpha, CDC15, and Elutrition from [13] and CDC28 from [14]—with different synchronization procedures. The total number of experiments in all series is 73, divided as 18, 24, 14, and 17 microarray experiments for each series. Each measurement is carried out in isolation, making it hard to estimate any error bars. However, it is generally believed that these can be huge, as other experiments with the same technology show [15].

The missing data in this set are here estimated by the procedure proposed in [16] that is based on a weighted K-nearest neighbor method (so-called *KNNimpute*¹). Essentially, it consists of selecting genes with expression profiles similar—in the Euclidean distance—to the gene of interest to

impute missing values. The number of neighboring genes used to estimate the missing values is here 15. A weighted average of the values of the experiments of the neighboring genes is computed: The contribution of each gene is weighted by similarity of its expression to that of the gene with the missing value. In [16], it is shown that the *KNNimpute* method outperforms other procedures commonly used to estimate missing values such as a singular value decomposition-based method, row average, and filling missing values with zeros. Finally, we center and normalize the expression data to have zero mean and unit variance.

The sampling periods for the four series are unequal and, thus, it is suitable to consider a time-continuous model. For such a model, the time derivatives are convenient for the inference. They are here obtained by spline interpolation of the original data. However, since ordinary splines are notoriously ill-behaved for curves obtained from experiments with large errors, we make use of the heuristic method of so-called taut splines [17] in order to achieve curves that are faithful to the measured data but still do not oscillate too wildly. A taut spline is a cubic spline with the extra constraint that there should be no extraneous inflection points. From [17], we quote: “If the broken line interpolant to the data is convex (concave) on the interval $[\tau_{r-1}, \tau_{s+1}]$, then a ‘good’ interpolant should be convex (concave) on the interval $[\tau_r, \tau_s]$.” This is achieved by inserting extra knots in a manner explicitly described in [17, pp. 303-314], where an algorithm for the implementation is also provided. We utilize that algorithm with the parameter value $\gamma = 3.1$, which means that we also allow for added knots in intervals where inflection points are permitted. Since we essentially know nothing about the boundaries, we turn to the so-called “not-a-knot” boundary condition, i.e., we let the first and last interior knot be nonactive. Eventually, the needed time derivatives are obtained by differentiation of the spline interpolants. This is trivially obtained from the algorithm provided in [17], since the output of the algorithm is the coefficients of the interpolating polynomial.

3 MODEL SELECTION

We consider a linear, time-continuous dynamical model of the form

$$\frac{dx_i(t)}{dt} = \sum_{j=1}^N w_{ij}x_j(t) + \epsilon_i. \quad (1)$$

Here, $x_i(t)$ can be any real number and denotes the logarithm of the ratio values of mRNA of gene i at time t , and $N = 6,178$ denotes the number of genes. The coefficient w_{ij} is the effect of gene j on gene i and is time-invariant. A linear model is extremely tractable from a computationally and data demanding point of view. The calculations simplify considerably and the amount of data needed for the inference procedure is order of magnitudes smaller than for nonlinear systems. However, the relevance of such a simple model can be seriously questioned and, whenever resorting to this kind of description, there has to be at least an a posteriori justification of the choice of model with respect to the outcomes. We will return to this issue in Section 4.

1. The program *KNNimpute* is freely available at <http://smi-web.stanford.edu/projects/helix/pubs/impute/>.

From a biological point of view, it makes sense to assume that the matrix w is sparse [7]. This means that we should prune the model somehow and perform a model selection. There are many different ways of choosing a subset of best predictors in such a linear model [18]. For example, the so-called Mallows's C_p or the Akaike Information Criterion can be used. However, both of these methods require extensive searches among all possible combinations of predictors, which, for three predictors among 6,178, already means approximately 40 billions different combinations. Bootstrap methods, which often are considered "better" when the number of experiments is small, are even more computationally demanding. Hence, we adopt a pragmatic approach here to the model selection problem and utilize a method which is possible to run on a personal computer (with a Pentium IV processor). The network is inferred by minimizing the residual sum of squares with an extra constraint on the L^1 -norm of the coefficients (the Lasso [11]). The hyperoctahedral form of the constraint for the Lasso makes it more likely that coefficients should become identical zero [11], hence, it acts as a combined subset selection procedure and regularization scheme. This means it is motivated both from a biological point of view, by making the matrix sparse, and from a numerical point of view, by regularizing the equations. It is even computationally possible to handle within the framework of standard numerical linear algebra [19]. However, this choice hides several problems of model selection, e.g., its relation to issues like conditional independence testing is unclear. With present data, however, we cannot ask for too much and concentrate instead on justifying our approach afterwards.

Explicitly, the numerical problem we solve takes, for each gene i , the form

$$\begin{aligned} \{\hat{w}_{ij}\}_{j=1}^N = \arg \min_{\{w_{ij}\}_{j=1}^N} \sum_{k=1}^K \left(\sum_{j=1}^N w_{ij} x_j(t_k) - \frac{dx_i}{dt}(t_k) \right)^2 \quad (2) \\ \text{subject to } \sum_{j=1}^N |\hat{w}_{ij}| \leq \mu_i. \quad (3) \end{aligned}$$

Note that (2) by itself corresponds to ordinary least squares (OLS) and, without the constraint in (3), the sum of squares can be zero in an infinite number of different ways. The Lasso constraint (3) both makes the solution unique and puts many of the coefficients to exactly zero. This can be compared with the more standard use of "ridge-regression," where the constraint instead is in the form of an upper limit on the sum of the *squares* of the coefficients [20]. In that case, however, there is no bias for values becoming exact zero and, hence, the procedure is not suitable for model selection although it works fine for regularization.

By iterating this procedure for all genes, we obtain the weight matrix w row by row, each element w_{ij} being optimized only once and together with the others regulating the same gene i . Hence, we end up with a weighted directed network. Each microarray measurement is supposed to have been performed at time t_k here and the experiments are numbered $k = 1, \dots, K$. The time derivatives are obtained from direct differentiation of the interpolants, as described in Section 2 above.

For small enough constraint parameters μ_i , the solution is unique [21]. Indeed, the Lasso *can* give unique solutions up to the case when the number of obtained nonzero predictors equals the number of measurements. To choose these constraint parameters, we first have to get an idea of the order of magnitude of the nonzero coefficients. To this end, we search for the solution of the minimization problem of the residual sum of squares in (2) (without the Lasso hyperoctahedron constraint (3)) which also minimizes the L^2 -norm of the coefficients $\{w_{ij}\}_{j=1}^N$. These minimized norms,

$$\mu_i^{(2)} = \left(\sum_{j=1}^N (\hat{w}_{ij})^2 \right)^{1/2}, \quad (4)$$

are used as the base-lines against which we measure the size of the L^1 -constraints μ_i . The values of $\mu_i^{(2)}$ are easily obtained by a singular value decomposition [22]. In this paper, we utilize the values $\mu_i = 0.1\mu_i^{(2)}$. This value of the coefficient, 0.1, is not the result of any fine-tuning, but instead somewhat arbitrary and results from our desire to keep the matrix w sparse. We have varied the coefficient in the interval [0.1, 0.8] and did not find any qualitatively different results. With this choice, the solution presented in this article is unique, which both means that the number of nonzero coefficients cannot exceed the number of measurements and that there are no points in the w -space that give the same minimum [21].

A more systematic approach to obtain the constraint parameters μ_i would of course be to perform some kind of cross-validation [23]. However, since the present data are supposed to be extremely noisy and the model is crude, the prediction errors can be large and there is an apparent risk of overfitting with a too large value of the constraint. Further, it seems reasonable from a biological point of view to keep the indegrees relatively low, since many genes are known to be controlled by a relatively small number of other genes.

4 ANALYSIS AND BIOLOGICAL VALIDATION

4.1 Topological Properties

To analyze the topological properties of the network, we focus on the adjacency matrix A , obtained from the weight matrix w as

$$A_{ij} = \begin{cases} 0 & \text{if } \hat{w}_{ji} = 0 \\ 1 & \text{otherwise.} \end{cases} \quad (5)$$

Hence, we obtain an unweighted directed graph, i.e., a digraph.

For this graph, we have indegrees, $k_j^{\text{in}} = \sum_{i=1}^N A_{ij}$, varying between unity and eight. We attribute this narrow distribution as a possible artifact of the Lasso procedure because the sum of the modulus of the coefficients is forced not to exceed a specific value and, hence, it is natural that there is no large spread in their number of nonzero values.² More interesting are the outdegrees, $k_i^{\text{out}} = \sum_{j=1}^N A_{ij}$, and

2. However, we note that a similar consideration for a biochemical network for yeast, obtained by Milo et al. [24], gives a similar distribution of indegrees.

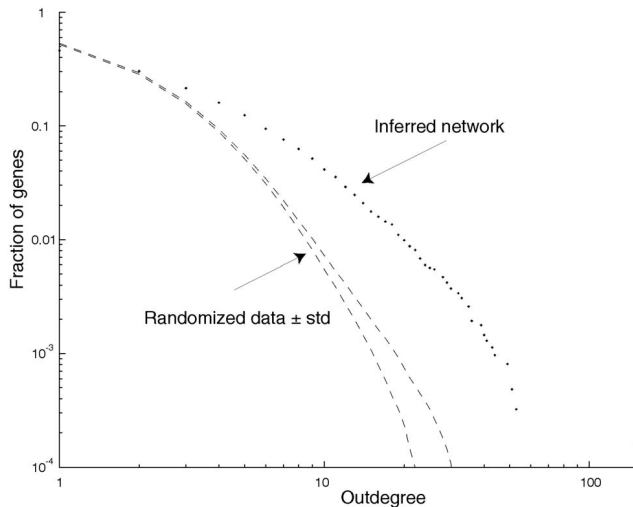


Fig. 1. Integrated distribution of outdegrees for the inferred network (dots). The dashed curves correspond to the variation of one standard deviation from the null hypothesis. Further, there are 3,331 nodes with outdegree zero, which cannot be seen in the figure.

their distribution, which is depicted as the upper curve in Fig. 1. For visualization purposes, we plot the integrated degree distribution. For example, we see in Fig. 1 how 5 percent of the genes have an outdegree *exceeding* 10. The dashed curves correspond to the variation of one standard deviation of randomized data³ and show that the obtained outdegree distribution for the ordered data is much broader than would be expected from only the distribution of the array data themselves. This broad distribution means that there are a few genes that control many others. This is what can be expected from a biological point of view and gives a first *indication* of the relevance of our inference procedure. For some parts, the distribution even follows approximately a power-law, which sometimes is called “scale-free distribution.” This subject has attracted much attention recently, see, for example, [25], but also much confusion [26], and we refrain from exploring possible implications thereof.

4.2 Biological Properties

To further study the biological relevance of the inferred network, we return to the distribution of outdegrees in Fig. 1. The gene *RRN5* has the highest outdegree, 148. According to the *Saccharomyces Genome Database*, SGD [27], it is involved in transcription of rDNA by RNA polymerase I. A systematic deletion gives an inviable organism. The genes *YHL026C* and *YJR079W* have the second and third highest outdegrees, 52 and 51, respectively. According to SGD, the organism is still viable after a systematic deletion of each gene. The molecular functions of the genes are unknown, as are the biological processes in which they are involved. A list of all genes and their ranks can be found in the supplementary material.

3. We have scrambled the expression values, both between times and between genes, in order to obtain a result with no bias but still preserve most of the statistics for the data. The derivatives are recalculated from the randomized data in Fig. 1, but the graph looks the same also when they are randomized keeping their original distribution.

However, looking up each single gene in the list of genes with high outdegrees is not necessarily fruitful because not more than half of the genome is annotated and we often obtain only the information that a specific gene has an unknown function, as seen above. Instead, we concentrate on groups of genes, each group consisting of the neighbours (with respect to outgoing edges) to a gene with outdegree larger than unity. Such a group should, to some extent, correspond to a biological process within the organism. In order to test that hypothesis, we turn to the GO database [12]. This is a database where the genes are arranged in a directed acyclic graph (DAG), according to three different ontologies. One ontology is for biological processes, and we utilize this one. The tree structure associates with each gene all ontology terms at and above it. Hence, we can query the database to search for the most common term for a group of genes.

To each ontology term which is assigned to the central gene in a group and to each group of genes, we assign a p -value showing how unlikely it is to find that specific term when we take into consideration how common the term is and the size of the group of genes.⁴ Here, we keep the term associated with the lowest p -value and discard the rest. However, testing each ontology term for every group makes it very probable that unlikely events should occur (due to multiple testing) and we need another way to judge the probability that our results should have occurred by chance. By randomizing all genes and recalculate the least p -values for groups of the actual sizes many times, we obtain a mean and a standard deviation for a null hypothesis.⁵ These values translate into standard Z -scores, i.e., into the number of standard deviations that the obtained result deviates from the null hypothesis. In Fig. 2, we show these numbers for all groups of genes ordered according to their outdegree rank. The widths of the bars are proportional to the number of genes which belong to the group, i.e., to the outdegree k_i^{out} . As is clearly seen in the figure, not all groups are known to correspond to a biological process, but still many of them are significant with more than three standard deviation. These groups are marked in Fig. 2 with the ontology term which is associated with the group when the outdegree of the central gene exceeds unity. We note as encouraging that, for the groups with negative Z -scores, the values are always within two standard deviations. We also depict the cumulative Z -score, calculated as $\sum_{i=1}^k Z_i / \sqrt{k}$, where Z_i is the Z -score for group i . This curve varies approximately between three and five standard deviations and, hence, it shows that our detected groups often have a biological significance.

Another way to study the biological relevance of the inferred network is to examine specific known classes of genes in order to see if there is any over or under representation of any group. It seems reasonable to associate the nodes with high outdegrees with genes involved in transcription, e.g., transcription factors, although the edges in the obtained network must not be

4. The underlying distribution is hypergeometric, which, in some cases, can be approximated by a binomial distribution. The inherent calculator in the database seems to use the latter distribution, even when it is not valid, and we recalculate all probabilities.

5. We do not utilize the p -values directly since their order of magnitude may differ significantly. Instead, we make use of the negative logarithms.

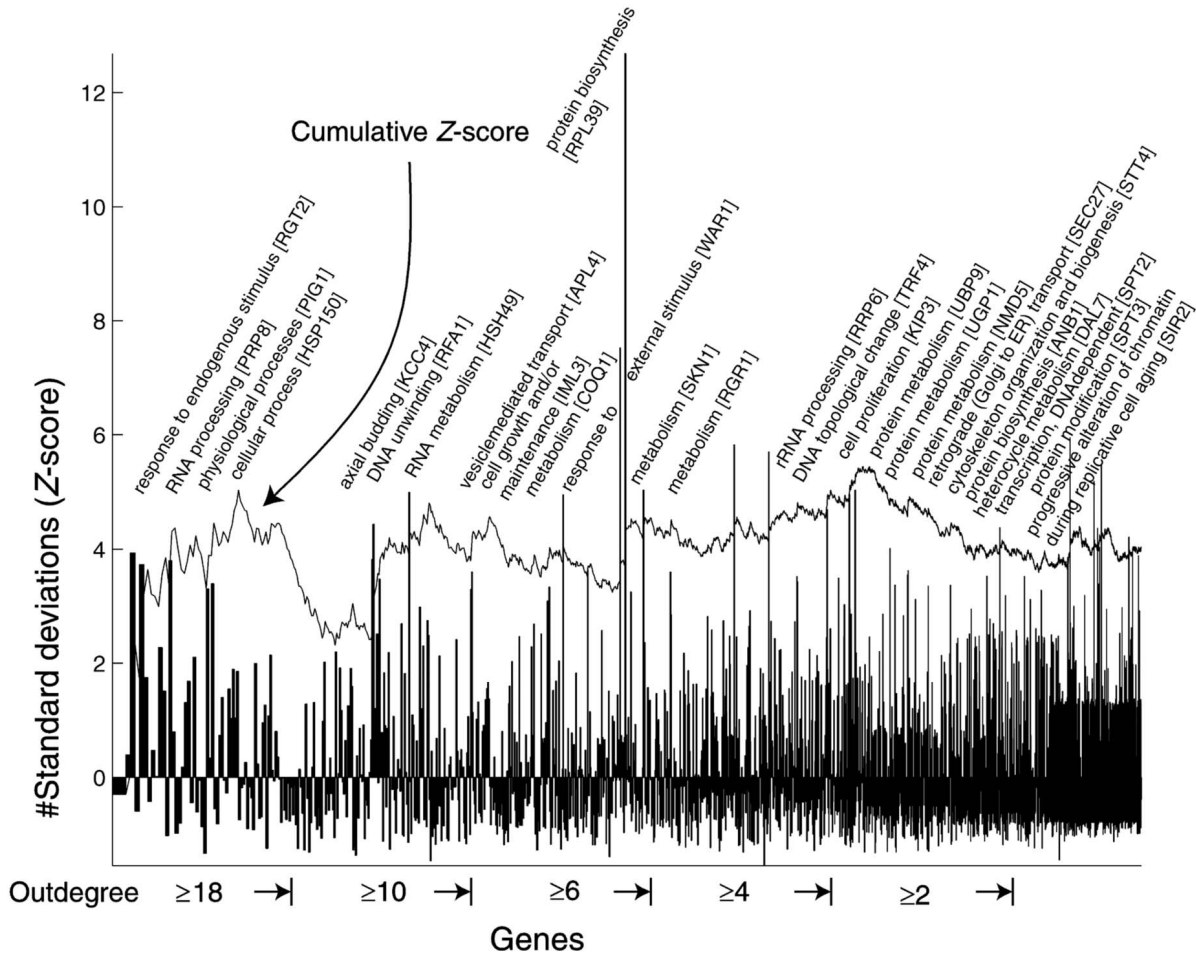


Fig. 2. Deviation from the null hypothesis, in units of standard deviations (Z -scores), for all groups of genes and their cumulated effect. The widths of the bars correspond to the number of genes in the actual group. All groups with a significance above three standard deviations and outdegree larger than or equal to two are labeled with their ontology term (unless it is unknown). A list of *all* genes, their name, degree, Z -score, and ontology term can be found in the supplementary material.

interpreted as biochemical interactions only.⁶ However, a previous study based on 273 single gene-deletion experiments for the same kind of yeast as we have did not show any such correlation [28]. Still, though, the conjecture makes sense, and we investigate if the genes of yeast known to be involved in transcription are overrepresented among the genes with highest outdegrees. In order to do so, we exploit the procedure proposed in [29], which will here be briefly reviewed.

We rank the genes according to their outdegree, giving the highest rank (i.e., rank number one) to the gene with the highest outdegree. From the GO database [12], we obtain a classification of each gene with rank r whether it codes for a product involved in transcription, $C_r = T$, or not, $C_r = NT$ (a third alternative is that it is unknown, i.e., $C_r \neq T$ and $C_r \neq NT$, simultaneously). From these data, we form the *cumulative excess* of genes which are known to be involved in transcription,

$$\Delta_r = N_C^r - H_r^0, \quad (6)$$

where $N_C^r = \#\{r' : C_{r'} = T, r' \leq r\}$ is the number of genes with the actual property and rank less than or equal to r , as a function of rank r . The number we subtract, H_r^0 , is the expected number of genes coding for products involved in transcription under the null hypothesis that they are uniformly distributed in outdegree rank, i.e.,

$$H_r^0 = \frac{N_T^{\text{all}} \cdot N_{T \text{ or } NT}^r}{N_{T \text{ or } NT}^{\text{all}}}. \quad (7)$$

All genes are ranked, so $r = 1, \dots, N$. Here, $N_{T \text{ or } NT}^r = \#\{r' : C_{r'} = T \text{ or } C_{r'} = NT, r' \leq r\}$ is the number of classified genes with rank less than or equal to r . The notation $N_T^{\text{all}} = \#\{r' : C_{r'} = T\}$ is the *total* number of genes known to code for products involved in transcription, and $N_{T \text{ or } NT}^{\text{all}} = \#\{r' : C_{r'} = T \text{ or } C_{r'} = NT\}$ is the *total* number of classified genes.

In Fig. 3, we show Δ_r , the cumulative excess, as function of rank, r . The *slope* of the curve corresponds to the excess of genes involved in transcription.⁷ The curve in Fig. 3 shows a clear excess of genes coding for products involved

6. Even if many of the known transcription factors in yeast are known to have not too many targets, they are still regulators, and our inference procedure can connect genes not directly linked in a regulatory cascade, as discussed in the introduction.

7. In principle, the slope is a more interesting entity than the cumulated excess, but it turns out to be less suitable for visualization [29].

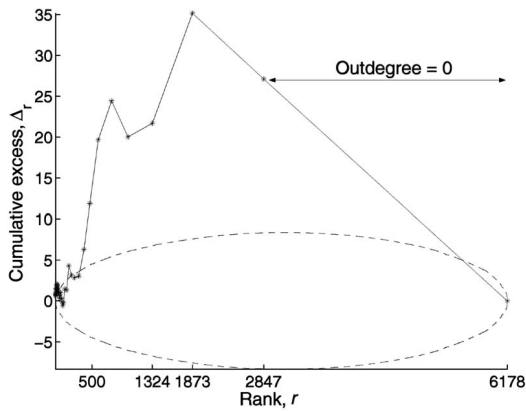


Fig. 3. Cumulated excess of genes coding for products involved in transcription, ranked according to their outdegrees. The dashed curves correspond to one standard deviation under the null hypothesis that they are uniformly distributed among all genes. If that would be the case, the curve Δ_r should be a straight line with constant value zero. We see a clear deviation from the null hypothesis of at most 4.8 standard deviations. The straight lines in the actual curve correspond to sets of genes with the same outdegree and whose order within the set is thereby arbitrary.

in transcription among the nodes with high outdegrees, more precisely, with ranks between 400 and 2,000. To see this, we also depict in the figure the curves corresponding to plus and minus one standard deviation under the null hypothesis (dashed curves). The ratio between the observed deviation and the standard deviation translate into Z -scores. We have a signal of 4.8 standard deviations for the first 737 genes and 4.6 standard deviations for the first 2,000 genes.

To shed further light on the use of the cumulative excess, we repeat the analysis above, but instead of considering genes involved in transcription, we concentrate on genes associated with transport and genes associated with the cell cycle, respectively. The latter is expected to have an over-representation of the genes with high outdegrees because the data are originally collected during the cell cycle and other genes might not be differentially expressed at all, while the former, to the best knowledge of the present authors, should be quite close to the null hypothesis. In Fig. 4, we see how these hypotheses are verified. The excess of cell cycle associated genes (squares) is almost as pronounced as the one for genes associated with transcription and there is a signal of more than 4.6 standard deviations. For the genes involved in transport (triangles), the excess (or rather the deficit) is almost within one standard deviation from the null hypothesis that the genes are uniformly distributed with respect to outdegree rank.

Hence, because of the analysis above, we claim that the obtained distribution of genes with respect to outdegrees is very far from accidental and has biological relevance.

5 DISCUSSION

We have presented the application of a specific inference procedure to the reverse engineering problem of a gene-to-gene regulatory network from temporal data. Here, we discuss the relation between our approach and some previous important work. Especially, we emphasize that, although there are many different approaches to this

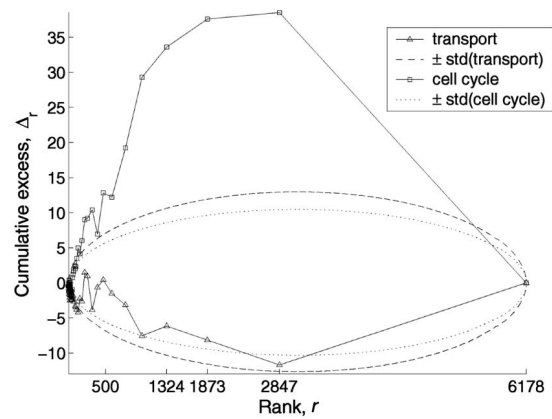


Fig. 4. Cumulated excess of genes coding for products involved in transport (triangles) and cell cycle (squares), respectively. The dashed and dotted curves indicate one standard deviation from the null hypothesis that the genes are uniformly distributed. It is clearly seen that there is an excess of cell cycle associated genes among the ones with high outdegree, while we cannot find any kind of accumulation of genes related to transport. This is in accordance with what could be expected and, hence, support the validity of the inferred network.

specific system identification problem, the complexity of the studied system makes it virtually impossible to judge “theoretically” which method is most appropriate, beyond the issue of pure computational complexity, but, instead, we have to resort to an a posteriori justification.

A main difficulty encountered when reverse engineering gene networks on the basis of mRNA expression levels is represented by the limited number of measurements available. In this context, standard model estimation methods based on system identification [30] cannot be applied as they usually require a great number of data. Methods have been proposed to solve the problem of reverse engineering in the presence of shortage of data by imposing additional constraints. In the work by D’haeseleer et al. [3], they start with exactly the same linear model as we have. In order to circumvent the problem of shortage of data, though, they turn to cubic interpolation on the expression level data, which makes the numerical procedure at least possible, but not feasible [4].⁸ Although [3] works with real data, there is no attempt to justify their approach beyond looking at some edges in the network. We have used another spline interpolation scheme, but, here, the splines are not used as a remedy for shortage of data, which makes our approach qualitatively different.

Further, there are similarities between our approach and the one by Yeung et al. [7] which deserve some attention. Especially, we have identical models as starting points for our inference procedures. In [7], a family of candidates is determined by using a singular value decomposition on the experimental data. However, they search for values of the coefficients w_{ij} within the space of values that result in the sum of squares in (2) becoming identically zero.⁹ By applying an L^1 -norm in that search, the resulting network turns to be sparse and, hence, they have another model selection procedure, which is not a special case of the one

⁸ They also had the same problem as we with unequally sampled time-series, which was another motivation for their use of interpolation.

⁹ In our case, this sum cannot be zero because of the constraint (3) being active.

we present here. However, they never apply their procedure to data from a real organism, which leaves it without justification.

One whole genome large-scale inference with almost the same model as the one we consider is the one investigated by Dewey and Galas [9]. They start with a model where the expression levels in one microarray experiment at some time t_k are a function of all expression levels in a previous experiment at time t_{k-1} , which means that their method is suitable only for time-series data where the samples are taken with a constant time difference. Both linear and nonlinear models are developed and applied to real data from *S. cerevisiae*, but the gene network they eventually analyze is based on the linear part. Effectively, they minimize the sum of squares in (2) and choose the solution which minimizes the sum $\sum_j w_{ij}^2$. The resulting matrix elements which are smaller than an arbitrarily chosen threshold will give zeroes in the corresponding adjacency matrix and ones otherwise, in a manner that [7] argues is questionable. However, Dewey and Galas [9] put the threshold so high that only connections among 143 genes are determined. Although interesting, the result is somehow restrictive. Another similar, but not identical, model is the one by Chen et al. [31]. They start with linear equations, but the expression levels are squashed by sigmoid-like functions. They apply their procedure to 2,119 genes in one of the data sets we utilize and present an extensive discussion about 20 of these. This should be compared with our approach, where we do not discard *any* of the 6,178 genes a priori and, subsequently, perform an analysis with respect to *all* annotated genes according to the GO database [12].

Finally, the work by Segal et al. [10] presents a biological validation on the same scale as we have (2,355 genes from *S. cerevisiae*), and they utilize the same GO database. However, their procedure requires a large precompiled set of candidate regulatory genes and the formalism relies on hierarchical clustering and the optimization of a Bayesian score to maximize the model's fit to the data. Hence, this approach differs from ours both with respect to input and formalism.

6 CONCLUSIONS AND OUTLOOK

In summary, a regulatory gene-to-gene network is inferred from temporal microarray data from yeast by a specific inference procedure. By studying the simplified network where we have discarded the weights of the links, we find a distribution of outdegrees which is broad. The existence of nodes with high outdegrees by chance is improbable, but reasonable from a biological point of view. Especially, we also find biologically relevant groups of genes around many of the genes (see the supplementary material for a detailed list), as well as a clear excess of genes coding for products involved in transcription among the genes with high outdegrees. These results show that the inferred network has biological relevance and, hence, that the use of linear methods is valid, at least for obtaining large-scale properties.

We stress that our method provides an insight of gene regulatory networks on a large-scale, but it cannot be used to get detailed information of small subnetworks related to biological functions. Especially, when turning to smaller

models, the stochastic nature of biological processes has to be taken into account. For example, Bayesian networks seem to be well-suited for this task [32], although they need more experimental data as well as larger computer resources than our inference presented here [33].

It is important to keep in mind that the network we obtain is not a network of direct interactions between various units in the cell, such as the DNA strands, mRNA-molecules, proteins, metabolites, etc. Instead, it is the network obtained by projecting all such interactions onto a space consisting of genes only. This makes it somewhat hard to compare with known interactions in the literature; indeed, two units related by only one link in our network can in the classical description with all intermediate steps included be quite separated.

A potential use of this kind of gene-to-gene network is to group the found structure into communities, beyond the simple version of only considering nearest neighbors employed here. These communities would then correspond to various biological processes and are hypotheses for which context each gene belongs to. This is then an important step from the normal clustering of data (see, e.g., [34], [35]) and gives the possibility to pick up more aspects than just covariation. We are currently planning for such a study, which will be reported elsewhere.

Our results show that the linear regime still has many promises for the analysis of gene expression data and must not be discarded.

ACKNOWLEDGMENTS

The authors would like to thank Kasper Eriksen for helpful discussions and some preliminary data sets to analyze. The Center for Industrial IT at Linköping University (CENIIT), the Swedish research council (VR), and the Carl Trygger Foundation (CTS) are acknowledged for financial support.

REFERENCES

- [1] H.D. Jong, "Modeling and Simulation of Genetic Regulatory Systems: A Literature Review," *J. Computational Biology*, vol. 9, no. 1, pp. 67-103, 2002.
- [2] P. Brazhnik, A. de la Fuente, and P. Mendes, "Gene Networks: How to Put the Function in Genomics," *Trends in Biotechnology*, vol. 20, pp. 467-472, 2002.
- [3] P. D'haeseleer, X. Wen, S. Fuhrman, and R. Somogyi, "Linear Modeling of mRNA Expression Levels During CNS Development and Injury," *Proc. Pacific Symp. Biocomputing*, R.B. Altman, A.K. Dunker, L. Hunter, T.E. Klein, and K. Lauderdaule, eds., vol. 4, pp. 41-52, 1999.
- [4] M. Hörnquist, J. Hertz, and M. Wahde, "Effective Dimensionality for Principal Component Analysis of Rat CNS Expression Data," *BioSystems*, vol. 71, pp. 311-317, 2003.
- [5] E.P. van Someren, L.F.A. Wessels, and M.J.T. Reinders, "Linear Modeling of Genetic Networks from Experimental Data," *Proc. Eighth Int'l Conf. Intelligent Systems for Molecular Biology*, pp. 355-366, 2000.
- [6] N.S. Holter, A. Maritan, M. Cieplak, N.V. Fedoroff, and J.R. Banavar, "Dynamical Modeling of Gene Expression Data," *Proc. Nat'l Academy of the Sciences USA*, vol. 98, pp. 1693-1698, 2001.
- [7] M.K.S. Yeung, J. Tegnér, and J.J. Collins, "Reverse Engineering Gene Networks Using Singular Value Decomposition and Robust Regression," *Proc. Nat'l Academy of the USA*, vol. 99, pp. 6163-6168, 2002.
- [8] E.P. van Someren, L.F.A. Wessels, E. Backer, and M.J.T. Reinders, "Multi-Criterion Optimization for Genetic Network Modeling," *Signal Processing*, vol. 83, pp. 763-775, 2003.
- [9] T.G. Dewey and D.J. Galas, "Dynamic Models of Gene Expression and Classification," *Functional and Integrative Genomics*, vol. 1, pp. 269-271, 2001.

- [10] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman, "Module Networks: Identifying Regulatory Modules and Their Condition-Specific Regulators from Gene Expression Data," *Nature Genetics*, vol. 34, no. 2, pp. 166-176, 2003.
- [11] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *J. Royal Statistical Soc., Series B*, vol. 58, no. 1, pp. 267-288, 1996.
- [12] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock, "Gene Ontology: Tool for the Unification of Biology," *Nature Genetics*, vol. 25, pp. 25-29, 2000. www.geneontology.org.
- [13] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and D. Futcher, "Comprehensive Identification of Cell Cycle-Regulated Genes of the Yeast *Saccharomyces Cerevisiae* by Microarray Hybridization," *Molecular Biology of the Cell*, vol. 9, pp. 3273-3297, 1998.
- [14] R.J. Cho, M.J. Campbell, E.A. Winzler, L. Steinmetz, A. Conway, L. Wodicka, T.G. Wolfsberg, A.E. Gabrielian, D. Landsman, D.J. Lockhart, and R.W. Davis, "A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle," *Molecular Cell*, vol. 2, pp. 65-73, 1998.
- [15] T. Ideker, V. Thorsson, J.A. Ranish, R. Christmas, J. Buhler, J.K. Eng, R. Bumgarner, D.R. Goodlett, R. Aebersold, and L. Hood, "Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Networks," *Science*, vol. 292, pp. 929-934, 2001.
- [16] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R.B. Altman, "Missing Value Estimation Methods for DNA Microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520-525, 2001.
- [17] C. deBoor, *A Practical Guide to Splines*. New-York: Springer, 1978.
- [18] E.I. George, "The Variable Selection Problem," *J. Am. Statistical Assoc.*, vol. 95, no. 452, pp. 1304-1308, 2000.
- [19] Å. Björck, *Numerical Methods for Least Squares Problems*. Philadelphia: SIAM, 1996.
- [20] N.R. Draper and H. Smith, *Applied Regression Analysis*, third ed. New York: Wiley, 1998.
- [21] M.R. Osborne, B. Presnell, and B.A. Turlach, "On the Lasso and Its Dual," *J. Computational and Graphical Statistics*, vol. 9, no. 2, pp. 319-337, 2000.
- [22] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes*. Cambridge Univ. Press, 1996.
- [23] M.R. Segal, "Regression Approaches for Microarray Data Analysis," *J. Computational Biology*, vol. 10, no. 6, pp. 961-980, 2003. <http://www.liebertonline.com/doi/abs/10.1089/106652703322756177>.
- [24] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network Motifs: Simple Building Blocks of Complex Networks," *Science*, vol. 298, no. 5594, pp. 824-827, 2002. <http://www.sciencemag.org/cgi/content/abstract/298/5594/824>.
- [25] *Handbook of Graphs and Networks, From the Genome to the Internet*, S. Bornholdt and H.G. Schuster, eds., Weinheim: Wiley, 2003.
- [26] L. Li, D. Alderson, R. Tanaka, J.C. Doyle, and W. Willinger, "Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications (Extended Version)," Technical Report CIT-CDS-04-006, California Institute of Technology, Pasadena, 2005.
- [27] K. Dolinski, R. Balakrishnan, K.R. Christie, M.C. Costanzo, S.S. Dwight, S.R. Engel, D.G. Fisk, J.E. Hirschman, E.L. Hong, L. Issel-Tarver, A. Sethuraman, C.L. Theesfeld, G. Binkley, C. Lane, M. Schroeder, S. Dong, S. Weng, R. Andrada, D. Botstein, and J.M. Cherry, "Saccharomyces Genome Database," www.yeastgenome.org, Sept. 2003.
- [28] D.E. Featherstone and K. Broadie, "Wrestling with Pleiotropy: Genomic and Topological Analysis of the Yeast Gene Expression Network," *BioEssays*, vol. 24, pp. 267-274, 2002.
- [29] K.A. Eriksen, M. Hörnquist, and K. Sneppen, "Visualization of Large-Scale Correlations in Gene Expressions," *Functional and Integrative Genomics*, vol. 4, pp. 241-245, 2004.
- [30] L. Ljung, *System Identification: Theory for the User*, second ed. Englewood Cliffs, N.J.: Prentice-Hall, 1999.
- [31] H.-C. Chen, H.-C. Lee, T.-Y. Lin, W.-H. Li, and B.-S. Chen, "Quantitative Characterization of the Transcriptional Regulatory Network in the Yeast Cell Cycle," *Bioinformatics*, vol. 20, no. 12, pp. 1914-1927, 2004. <http://bioinformatics.oupjournals.org/cgi/content/abstract/20/12/1914>.
- [32] M. Zou and S.D. Conzen, "A New Dynamic Bayesian Network (DBN) Approach for Identifying Gene Regulatory Networks from Time Course Microarray Data," *Bioinformatics*, vol. 21, no. 1, pp. 71-79, 2005. <http://bioinformatics.oupjournals.org/cgi/content/abstract/21/1/71>.
- [33] D. Husmeier, "Sensitivity and Specificity of Inferring Genetic Regulatory Interactions from Microarray Experiments with Dynamic Bayesian Networks," *Bioinformatics*, vol. 19, no. 17, pp. 2,271-2,282, 2003. <http://bioinformatics.oupjournals.org/cgi/content/abstract/19/17/2271>.
- [34] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein, "Cluster Analysis and Display of Genome-Wide Expression Patterns," *Proc. Nat'l Academy of the Sciences USA*, vol. 95, pp. 14,863-14,868, Dec. 1998.
- [35] S. Datta and S. Datta, "Comparisons and Validation of Statistical Clustering Techniques for Microarray Gene Expression Data," *Bioinformatics*, vol. 19, no. 4, pp. 459-466, 2003.



Mika Gustafsson received the MSc in mathematics from Stockholm University in 2003 and the MSc in physics from Linköping University, Sweden, in the same year. Now, he is working on a PhD dissertation in applied mathematical physics at Linköping University, focusing mainly on analysis of networks.



Michael Hörnquist is an associate professor of theoretical physics at Linköping Institute of Technology, Linköping University, Sweden. His research comprises various aspects of biological physics, such as systems biology and DNA dynamics.



Anna Lombardi received the PhD degree in computer and automation engineering from Università di Firenze, Firenze, Italy. She is currently a senior lecturer of automatic control at Linköping University, Sweden. Her research focuses on computational biology with particular interest in modeling of gene networks and clustering algorithms applied to networks.

► For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.