Large-scale Reverse Engineering by the Lasso

Mika Gustafsson,* Michael Hörnquist,† and Anna Lombardi‡
Department of Science and Technology, Linköping university (Campus Norrköping), SE-601 74 Norrköping, Sweden
(Dated: February 21, 2006)

We perform a reverse engineering from the "extended Spellman data", consisting of 6178 mRNA levels measured by microarrays at 73 instances in four time series during the cell cycle of the yeast *Saccharomyces cerevisae*. By assuming a linear model of the genetic regulatory network, and imposing an extra constraint (the Lasso), we obtain a unique inference of coupling parameters. These parameters are transfered into an adjacent matrix, which is analyzed with respect to topological properties and biological relevance. We find a very broad distribution of outdegrees in the network, compatible with earlier findings for biological systems and totally incompatible with a random graph, and also indications of modules in the network. Finally, we show there is an excess of genes coding for transcription factors among the genes of highest outdegrees, a fact which indicates that our approach has biological relevance.

PACS numbers: 89.75.-k, 89.75.Hc, 05.65.+b
Keywords: reverse engineering, Lasso, networks

Advances in microarray technologies make it today possible to measure mRNA-levels for thousands of genes simultaneously. 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, genetic regulatory networks from temporal data. Although still somewhat speculative, researchers are exploring the boundaries for what kind of inference that is possible. There are many approaches for network formation, ranging from Boolean circuits to very complicated non-linear spatial models, see [1] and 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. By focusing only on transcript data, the networks obtained are not biochemical regulatory networks, but phenomenological 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, which has gained some popularity, is the linear, continuous model. 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 are, 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 hundred 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 vanSomeren *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. A more biologically motivated study was performed by Yeung *et al.* [7], who assumed that the resulting network should be sparse and that way got a unique solution. Somewhat in the same spirit, vanSomeren *et al.* [8] conducted a systematic study on how to incorporate prior knowledge into the inference procedure. Finally we mention the only large-scale inference we are aware of. It was conducted by Dewey and Galas [9], who considered the whole genome of yeast with more than 6000 genes, and formed the 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 only for 143 genes, and although they justify that their result makes biological sense on a small scale, they lack a large-scale analysis.

In the present letter, we utilize one statistical method, the Lasso [10], to reverse engineer a network among ORFs ("Open Reading Frames, hereafter referred to as "genes") in the so-called extended Spellman dataset. This dataset

is one of the most referenced sources of microarray data and contains measured mRNA levels of 6178 genes for the yeast *S. cerevisae*, 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 [11], and CDC28 from [12]—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.

The missing data in this set are estimated by the procedure proposed in [13]. 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. Finally, we center and normalize the expression data to have zero mean and unit variance.

We consider a linear, continuous model of the form

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

Here $x_i(t)$ denotes the logarithm of the ratio values of mRNA of gene i at time t, and $N = 6178$ denotes the number of genes. The coefficient w_{ij} is the effect of gene j on gene i and does not depend on time.

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). The hyperoctahedronal form of the constraint makes it more likely that coefficients should become identical zero. Explicitly, it takes the form

$$\hat{w}_{ij} = \arg\min_{w_{ij}} \quad \sum_{k=1}^{K} \left(\sum_{j=1}^{N} w_{ij} x_j(t_k) - \frac{dx_i}{dt}(t_k) \right)^2 \tag{2}$$

$$\text{subject to} \quad \sum_{j=1}^{N} |\hat{w}_{ij}| \leq \mu_i \quad \text{for } i = 1, \cdots, N. \tag{3}$$

Each microarray measurement is here supposed to have been performed at time t_k and there are K experiments. The time derivatives are obtained by spline interpolation of the original data, where we make use of so-called taut splines [14] in order to achieve curves that are faithful to the measured data but still do not oscillate too wildly. By this procedure, the problem factorize and we can perform the minimization for each gene i separately.

For small enough constraint parameters μ_i, the solution is unique. To choose these parameters, we first solve the combined minimization problem of the residual sum of squares in (2) together with the minimization of the L^2-norm of the coefficients. These values,

$$\mu_i^{(2)} = \left(\sum_{j=1}^{N} (\hat{w}_{ij})^2 \right)^{1/2} \qquad \text{for } i = 1, \cdots, N, \tag{4}$$

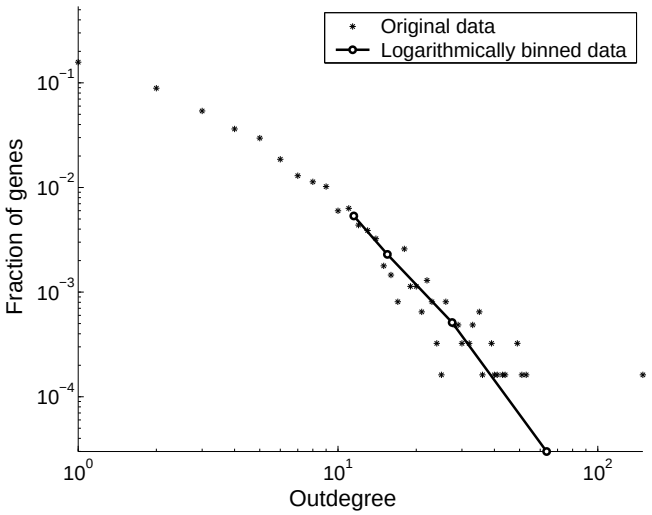

FIG. 1: Distribution of outdegrees for the inferred network. This distribution is incompatible with one from a random graph. Further, there are 3331 nodes with outdegree zero.

are used as base-lines against which we measure the size of the L^1-constraints. Here we utilize the values $\mu_i = 0.1\mu_i^{(2)}$. However, varying the coefficient from the value 0.1 does not result in large changes in the chosen subsets. With this choice, the presented solution in this letter is unique [15].

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

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

Hence we obtain an unweighted digraph.

For this digraph, we obtain a distribution of indegrees varying between unity and eight. This we attribute as an 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 non-zero values. More interesting is the distribution of outdegrees, which is depicted in Fig. 1. The distribution is very skew, totally incompatible with a Poisson distribution of a random graph [16]. For the present distribution of indegrees, the probability that any gene gets an outdegree that exceeds 50 by accident is less than 10^{-40} given that the edges are independently uniformly distributed.

Our obtained distribution of outdegrees does not follow a power-law for more than one decade, i.e., it is not scale-free; a property that many other biological networks seem to have [17]. Still, the distribution is broad, and we expect it to be robust as scale-free networks have been proven to be—both topologically [18] and dynamically [19].

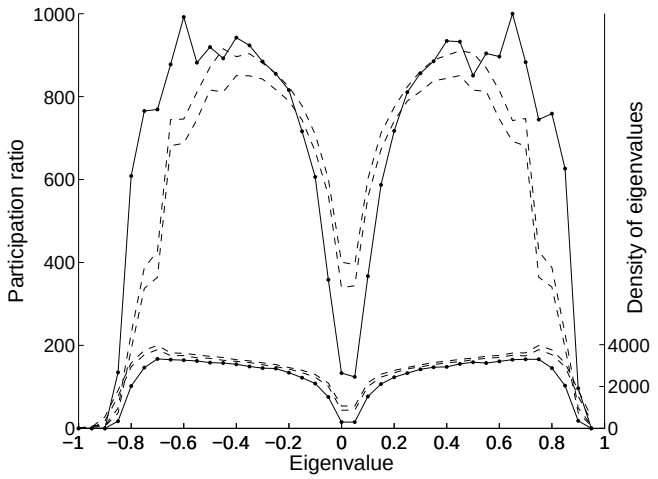

FIG. 2: Participation ratios for the normalized eigenvectors (upper curves) and density of eigenvalues (lower curves) versus eigenvalues of the transpose of the matrix T from (6). The results of the rewired networks, dashed curves, are the variation of one standard deviation around the mean of 20 randomized networks where the degree distribution has been kept constant and thus comprise a null hypothesis. The eigenvalues equal to unity, as well as the zero eigenvalues, have been discarded here. The results are binned for clarity into bins of size 0.05. There is a more modular structure in the yeast network than could be expected from a random system.

We also search for possible modules in the obtained network. To perform this search, we employ the formalism in [20], where the authors unraveled modules in the internet. Here we explore the network in undirected form, i.e., we study the corresponding undirected graph obtained by making the adjacent matrix (5) symmetric. The participation ratio of each (normalized) eigenvector to the transpose of the matrix

$$T_{ij} = \begin{cases} 0 & \text{if nodes } i \text{ and } j \text{ not are adjacent} \\ 1/k_j & \text{otherwise,} \end{cases} \qquad (6)$$

where k_j is the degree of node j, gives an estimate of the size of the corresponding module. The eigenvalue itself corresponds to how tightly connected the module is. In Fig. 2 we see how the participation ratios vary with the eigenvalues and also the density of eigenvalues. As a null hypothesis, we depict the variation within one standard deviation for values obtained from randomized networks with the same degree distribution as the actual yeast network, as described in [21]. We see in the figure how there are some modules in the yeast network, by the fact that there are quite high participation numbers, compared with the randomized networks, for eigenvalues around 0.8. A closer analysis of these modules will be published elsewhere.

To 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, 149. According to the Saccharomyces Genome Database, SGD [22], 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, 53 and 51, respectively. According to SGD, the organism is still viable after a systematic deletion of each. The molecular functions of the genes are unknown, as are the biological processes in which they are involved.

It does not seem too unrealistic to associate the nodes with high outdegrees with transcription factors, although the edges in the obtained network must not be interpreted as biochemical interactions only. 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 [23]. Still, though, the conjecture makes sense, and we investigate if the (known) transcription factors of yeast are overrepresented among the genes with highest outdegrees. In order to do so, we exploit the procedure proposed in [24].

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 [25] we obtain a classification of each gene whether it codes for a transcription factor or not (or if it is unknown). From these data, we form the *cumulative excess* of genes which code for transcription factors,

$$\Delta_r = \#\{\mathrm{TF} \leq r\} \; - \; n_r \frac{\#\{\mathrm{TF}\}}{M}, \qquad (7)$$

as a function of rank r. All genes are ranked, so $r = 1, \cdots, N$. Here $\#\{\mathrm{TF} \leq r\}$ is the number of genes known to code for transcription factors and n_r is the number of classified genes, both in the set of genes with rank $\leq r$, $\#\{\mathrm{TF}\} = 308$ is the total number of genes known to code for transcription factors, and finally $M = 3294$ is the total number of classified genes.

The number we subtract is the expected number of genes coding for transcription factors under the null hypothesis that they are uniformly distributed in outdegree ranks. In Fig. 3 we show Δ_r, the cumulative excess, as function of rank. The *slope* of the curve corresponds to the excess of transcription factors, i.e., the deviation from the null hypothesis. [26]

The curve in Fig. 3 shows a clear excess of genes coding for transcription factors among the nodes with high outdegrees, between 400 and 2000, approximately. To see this, we depict in the figure also 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 standard Z-scores. We have a signal of 4.8 standard deviations for the first 737 genes, 3.2 standard deviations for the first 1000 genes and 4.6 standard deviations for the first 2000 genes. Hence, we claim that the obtained

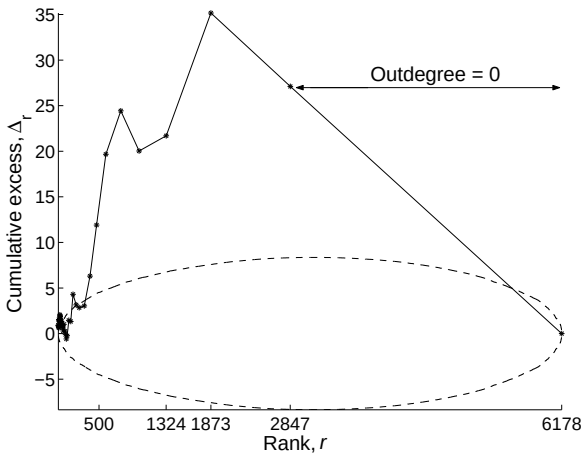

FIG. 3: Cumulated excess of genes coding for transcription factors, ranked according to their outdegrees. The dashed curves correspond to one standard deviation under the null hypothesis that the transcription factors are uniformly distributed among the classified genes. The straight lines correspond to sets of genes with the same outdegree, and whose order within the set thereby is arbitrary.

distribution of genes is very far from accidental.

In summa, we have presented the application of a specific inference procedure to the reverse engineering problem of an effective regulatory network from temporal data. By studying the simplified network where we have discarded the weights of the links, we find a distribution of outdegrees which is broad, as well as modules in the network. The existence of nodes with high outdegrees by chance is improbable. A closer look shows that the most connected node in the network represents a gene which is involved in transcription. Especially, we also find a clear excess of genes coding for transcription factors among the genes with high outdegrees. Given the simplicity of our approach, utilizing only linear models and transcript data, the method works reasonably well.

We thank Kasper Eriksen for helpful discussions and some preliminary datasets to analyze. The center for industrial IT at Linköping university, CENIIT, and the Swedish research council, VR, are acknowledged for financial support.

* Electronic address: mikgu@itn.liu.se
† Electronic address: micho@itn.liu.se
‡ Electronic address: annlo@itn.liu.se

[1] H. D. Jong, J. Comp. Biol. **9**, 67 (2002).
[2] P. Brazhnik, A. de la Fuente, and P. Mendes, Trends in Biotech. **20**, 467 (2002).
[3] P. D'haeseleer, S. Liang, and R. Somogyi, in *Pacific Symposium on Biocomputing*, edited by R. B. Altman, A. K. Dunker, L. Hunter, T. E. Klein, and K. Lauderdaule (World Scientific Publishing Co, Singapore, 1999), vol. 4, pp. 41–52.
[4] M. Hörnquist, J. Hertz, and M. Wahde, BioSystems **71**, 311 (2003).
[5] E. P. van Someren, L. F. A. Wessels, and M. J. T. Reinders, in *Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB00)* (AAAI, La Jolla, California, 2000), pp. 355–366.
[6] N. S. Holter, A. Maritan, M. Cieplak, N. V. Fedoroff, and J. R. Banavar, Proc. Nat. Acad. Sciences USA **98**, 1693 (2001).
[7] M. K. S. Yeung, J. Tegnér, and J. J. Collins, Proc. Nat. Acad. Sciences USA **99**, 6163 (2002).
[8] E. P. van Someren, L. F. A. Wessels, E. Backer, and M. J. T. Reinders, Signal Proc. **83**, 763 (2003).
[9] T. G. Dewey and D. J. Galas, Funct. Integr. Genomics **1**, 269 (2001).
[10] Tibshirani, J. R. Stat. Soc. B **58**, 267 (1996).
[11] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and D. Futcher, Mol. Biol. Cell **9**, 3273 (1998).
[12] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, et al., Mol. Cell. **2**, 65 (1998).
[13] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, Bioinformatics **17**, 520 (2001).
[14] C. deBoor, *A Practical Guide to Splines* (Springer, New-York, 1978).
[15] M. R. Osborne, B. Presnell, and B. A. Turlach, J. Comp. Graph. Stat. **9**, 319 (2000).
[16] B. Bollobás, *Random Graphs* (Cambridge University Press, Cambridge UK, 2001), 2nd ed.
[17] M. Newman, SIAM Review **45**, 167 (2003).
[18] R. Albert, H. Jeong, and A.-L. Barabási, Nature **406**, 378 (2000).
[19] M. Aldana and P. Cluzel, Proc. Nat. Acad. Sciences USA **100**, 8710 (2003).
[20] K. A. Eriksen, I. Simonsen, S. Maslov, and K. Sneppen, Phys. Rev. Lett. **90**, 148701 (2003).
[21] S. Maslov, K. Sneppen, and U. Alon, in *Handbook of Graphs and Networks*, edited by S. Bornholdt and G. Schuster (Wiley, Weinheim, 2003), chap. 8, pp. 168–198.
[22] 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, et al., *Saccharomyces genome database*, `www.yeastgenome.org/`, visited September 4, 2003.
[23] D. E. Featherstone and K. Broadie, BioEssays **24**, 267 (2002).
[24] K. A. Eriksen, M. Hörnquist, and K. Sneppen (2003), submitted, available at `www.itn.liu.se/~micho/research/Visualization.pdf`.
[25] 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, et al., Nat. Gen. **25**, 25 (2000), `www.geneontology.org/`, visited December 21, 2003.
[26] In principle, the slope is a more interesting entity than the cumulated excess, but it turns out to be less suitable for visualization [24].