# Discerning static and causal interactions in genome-wide reverse engineering problems

M. Zampieri, N. Soranzo and C. Altafini
SISSA-ISAS, International School for Advanced Studies
via Beirut 2-4, 34014 Trieste, Italy

January 9, 2008

## Abstract

**Background** In the past years devicing methods for discovering gene regulatory mechanisms at a genome-wide level has become a fundamental topic in the field of system biology. The aim is to infer gene-gene interactions in a more sophisticated and reliable way through the continuously improvement of reverse engineering algorithms exploiting microarray technologies.

**Motivation** This work is inspired by the several studies suggesting that co-expression is mostly related to "static" stable binding relationships, like belonging to the same protein complex, rather than other types of interactions more of a "causal" and transient nature (metabolic pathway or transcription factor–binding site interaction). Discerning static relationships from causal ones on the basis of their characteristic regulatory structures and in particular identifing "dense modules" with protein complex, and "sparse modules" with causal interactions such as those between transcription factor and corresponding binding site, the performances of different network inference algorithms in artificial and real networks (derived from *E.coli* and *S.cerevisiae*) can be tested and compared.

**Results** Our study shows that methods that try to prune indirect interactions from the inferred gene networks may fail to retrieve genes co-participating in a protein complex. On the other hand they are more robust in the identification of transcription factor–binding sites dependences when multiple transcription factors regulate the expression of the same gene. In the end we confirm the stronger co-expression regarding genes belonging to a protein complex than transcription factor–binding site, according, also, to the effect of multiple transcription factors and a low expression variance.

## Introduction

In the field of systems biology, high throughput measurements are giving the possibility to investigate the mechanisms that regulate the behavior of the cells in a genome-wide manner. The ability to use such broad information to infer interaction between genes is the first step towards a comprehensive understanding of a biological system, in terms of genes functions, "partner genes", conditions for activation and dynamical behavior. The reconstruction of a gene network [2, 6, 10] is a very challenging problem since biological systems are difficult to perturb and perturbations/experiments are tipically much less than the number of dynamical variables composing the system. If we also consider the highly non-linear content characterizing the transcriptional expression it is easy to understand why the problem of reverse engineering collections of experimental data is underdetermined, and why such a big effort has been put in proposing new algorithms to overcome these limitations.

Last but not least, the process of bringing a DNA sequence into its corresponding final gene product is made by several steps. The different regulation layers are neglected by the vast majority of the reverse engineering algorithms, which instead consider only expression coregulation as an index of general "functional" relationship. Many studies [27, 26, 22] underline the lack of correlation between protein abundance and mRNA level, and only with reliable technologies to measure simultaneously gene expression, protein profiles and metabolite quantities, an accurate description of cell behaviour can be achieved [12]. Despite these considerations, recent works demonstrate that gene expression correlation is the most significant index among several others like ontological information, sequence similarity, protein localization and domain structure, to infer putative protein–protein interactions [30].

In this paper, we try to understand what are the topologies that reverse engineering algorithms are able to reconstruct when applied in a completely unsupervised manner. Several methods exist to treat the observed data relying on more or less sophisticated statistical analysis of gene expression profiles and modeling frameworks, like Bayesan networks [33, 17, 14] and boolean networks [19, 34], or linear and non-linear ordinary differential equations (ODEs) [39]. We focus here on two other classes of algorithms, called relevance networks and graphical models. They are computationally more treatable than most of the methods mentioned above and can therefore be applied in a truly genome-wide context. They consist essentially in computing a two-point similarity measure between gene pairs, similarity which is then used to weight the edges of a graph. The highest is the weight, the most likely the two genes interact in some way. The simultaneous usage of different information sources/technologies, like gene annotations, chromatin immunoprecipitation chips [41] (to unveil direct protein–DNA interactions) or yeast-two-hybrid experiments (to detect protein–protein interactions) can reduce the error rate and give hints on how to choose the "best" weight cutoff [15, 38], even if such information is affected by multiple error types, uncertainty and is far from being exhaustive. The similarity measures used to asses co-expression between gene

pairs are Pearson correlation (P) [4], mutual information (MI) [5], partial Pearson correlation (CP) [7], conditional mutual information (CMI), and graphical gaussian model (GGM) [8]. The first two metrics account respectively for a linear and non-linear relationship, while the remaining ones prune the inferred network from the putative false positives generated by the first two. The pruning is performed by means of a conditioning operation on the two-point measure. Conditioning which can depend on a single third gene (CP and CMI) or on the remaining $n-2$ genes (GGM), see [36] for details. Relevance networks and graphical gaussian models have been extensively used in recent years [23] and their results have been validated experimentally, for example in [3] where putative MYC interacting genes have been identified processing expression profiles of human B cells. This analysis was based on a similarity index related to CMI [24]. In this paper a comparison between the two classes of similarity metrics, direct and conditional, is performed within the aim of analyzing their ability to infer regulatory networks characterized by different connectivity degrees, that for simplicity we denote: "sparse module" and "dense module" in both artificial and real networks. In the first one multiple genes are interconnected with sparse edges, while in the second all the nodes of a set are mutually connected (Fig. 1(c,d)). These two different regulatory motifs are meant to represent two gene interactions macrocategories: one describing a cause–effect relationships, like the direct transcriptional activation due to transcription factors, the other linked to a more "static" associations like coparticipation in a protein-complex, where gene products have to be expressed in a constant stoichiometric ratio to perform their proper functions [28, 29]. The artificial network is meant to enable the evaluation under controlled conditions, like a well defined topology and known kinetics governing the system (see Methods). In the two cases of real data the identification of true positive (TP) edges relies on the real physical networks of protein complexes (PCs) and transcription factor–binding site (TF–BS) relationships collected from the literature. We choose two simple organisms, a prokaryote and an eukaryote, in order to test the consistency of the two regulatory structures for the different algorithms. For these two organisms most of the PC and TF–BS have been annotated and large collections of gene expression profiles can be gathered from online repositories. To compare the inference powers of the different algorithms in respect to the two types of regulatory modules, we ranked the weights of each similarity matrix and look at the percentage of TPs in the most significant percentile. For protein complexes we iterate the comparison while increasing their size, whereas in the case of TF–BS we increase the number of TFs acting on the same BS. In addition we apply a simple clustering algorithm to the inferred graphs and analyze how well the clusters match the PCs. This procedure allows us to make an unbiased comparison between different metrics, overcoming the problem of choosing the "best" threshold.

The two different regulatory structures are related to the stability and to the dynamical variation of an interaction between nodes. The contribution to the average connectivity degree of the gene network from the two regulatory motifs is also different. The prediction of PC memberships emerged with the advent of large scale experiments and the publication of several biological networks.

Most of these studies use protein–protein interaction databases to identify subsets of proteins having many more interactions among themeselves than with the rest of the network. Inferred multimolecular proteins rely on the identification of highly connected subgraphs [37]. Searching for defective cliques has been shown to be a good predictor [40], although computationally very intensive. Starting from this observation, the complexity of an organism can be deduced not only by the regulation complexity at the level of transcripts (see Fig. 1 (b)) [21, 20], but also looking at how the size of the dense modules, representing PCs (see Fig. 1 (a)), increases. Going from unicellular prokaryote (*E.coli*) and eukaryote (*S.cerevisiae*) to mammals (human, rat and mouse), the distribution of annotated PCs shows an heavier tail towards bigger complexes. The same happens looking at the combinatorial effect of multiple TFs. In Yeast for example the largest complex is the cytoplasmic ribosome accounting for 81 genes, while in *E.coli* it is the flagellum complex composed of 24 genes, suggesting that a complex organism can show an higher contribution of dense modules in their regulatory structure. For complex organisms, this particular aspect, in addition to the observation that co-expression stands often times for stable binding [16, 35], becomes an important issue when reverse engineering expression data, in particular in light of the fact that inference requires large compendia of expression profiles.

## Materials and methods

The model we used to generate artificial gene expression datasets is the reaction kinetics-based system of coupled non-linear continuous time ODEs introduced in [25]. The sparse module, representing influence on the transcription of each gene due to the other genes, is described by a random matrix of adjacencies, superimposed to a matrix of densely connected subsets of nodes representing the stable modules (see Fig. 5 SUPPLEMENTARY??). The rate law for the mRNA synthesis of a gene is obtained by multiplying together the sigmoidal-like contributions of the genes identified as its inhibitors and activators. Consider the $i$-th row of $A$, $i = 1, \ldots, n$, and choose randomly a sign to its nonzero indexes. Denote by $j_1, \ldots, j_a$ the indexes with assigned positive values (activators of the gene $x_i$) and with $k_1, \ldots, k_b$ the negative ones (inhibitors of $x_i$). The ODE for $x_i$ is then

$$\frac{dx_i}{dt} = V_i \prod_{j \in \{j_1,\ldots,j_a\}} \left(1 + \frac{x_j^{\nu_{i,j}}}{x_j^{\nu_{i,j}} + \theta_{i,j}^{\nu_{i,j}}}\right) \prod_{k \in \{k_1,\ldots,k_b\}} \frac{\theta_{i,k}^{\nu_{i,k}}}{x_k^{\nu_{i,k}} + \theta_{i,k}^{\nu_{i,k}}} - \lambda_i x_i, \tag{1}$$

where $V_i$ represent the basal rate of transcription, $\theta_{i,j}$ (respectively $\theta_{i,k}$) the activation (resp. inhibition) half-life, $\nu_{i,j}$ (resp. $\nu_{i,k}$) the activation (resp. inhibition) Hill coefficients (in our simulations: $\nu_{i,j}$, $\nu_{i,k} \in \{1, 2, 3, 4\}$), and $\lambda_i$ the degradation rate constants. In the multiple in silico experiments the perturbations of the system are performed by means of random initial conditions, plus "gene knockouts" (obtained setting to 0 the expression of the selected gene in the corresponding differential equation). A gaussian measurement noise is added to corrupt the output. The number of

experiments has been chosen to be comparable with the real case, where we have a ratio experiments/genes of approximately one to six, while the number of complexes and their size have been sampled from a log-normal distribution with a maximum size of 50 genes composing a complex. This choice is consistent with what observed in the real organisms and gives rise to a manageable number of genes in the simulation of gene expression profiles (2154).

**Data collected** We downloaded the $M^{3D}$ "Many Microbe Microarrays Database" (build E_coli_v3_Build_1) [9] for *E.coli* (445 experiments for 4345 genes). For *S.cerevisiae* we compiled a collection of microarrays containing experiments performed with cDNA chips (958 experiments for 6203 ORF). Both datasets were preprocessed and normalized prior to network inference. PC network for yeast was downloaded from the MPACT subsection of the CYGD database at MIPS [11]. Only the complexes annotated from the literature and not those obtained from high throughput experiments (according to the MIPS classification scheme these last are labeled "550") were considered to limit the high rate of false positive. PC sizes for human, rat and mouse were downloaded from CORUM database [31], while for *E.coli* from the EcoCyc website [18]. We obtained TF–BS networks from the *RegulonDB* database, version 5.6, for *E.coli* [32], and from a recent collection [1] for *S.cerevisiae*.

**Similarity measures** Let $m$ be the number of experiments available and $n$ the number of genes. Assume $X_i$ and $X_j$, $i, j = 1, \ldots, n$, are random variables representing the genes, and $x_i(\ell)$, $x_j(\ell)$, $\ell = 1, \ldots, m$, their sample measurements. The matrices of edges weights are computed using the following five algorithms, see [36] and references therein for details:

- Pearson correlation:
$$R(X_i, X_j) = \left| \frac{E[(x_i - \bar{x}_i)(x_j - \bar{x}_j)]}{\sqrt{v_i v_j}} \right|, \tag{2}$$
where $\bar{x}_i$, $v_i$ and $\bar{x}_j$, $v_j$ are means and variances of $x_i$ and $x_j$ over the $m$ experiments and $E[\cdot]$ denotes expectation.

- Partial Pearson correlation
$$R_{C_1}(X_i, X_j) = \min_{k \neq i,j} \left| \frac{R(x_i, x_j) - R(x_i, x_k)R(x_j, x_k)}{\sqrt{(1 - R^2(x_i, x_k))(1 - R^2(x_j, x_k))}} \right|. \tag{3}$$

- Graphical gaussian method
$$R_{C_{\text{all}}}(X_i, X_j) = \left| \frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}} \right|, \tag{4}$$
where $\Omega = (\omega_{ij})$ is $R^{-1}$ if $R^{-1}$ exists, it is the small-sample estimate of [33] when $R$ is not full-rank. De facto, $R_{C_{all}}$ is computed by means of the R package GeneNet version 1.0.1, available from CRAN (http://cran.r-project.org).

- Mutual information

$$I(X_i;\, X_j) = \sum_{\phi_i,\, \phi_j \in \mathcal{H}} p(\phi_i, \phi_j) \log \frac{p(\phi_i, \phi_j)}{p(\phi_i) p(\phi_j)}, \tag{5}$$

  where $p(\phi_i)$ is the probability mass function $p(\phi_i) = Pr(X_i = \phi_i)$, $\phi_i$ in the alphabet $\mathcal{H}$, and likewise for the joint probability function $p(\phi_i, \phi_j)$.

- Conditional mutual information

$$I_C(X_i; X_j) = \min_{k \neq i,j} \sum_{\phi_i,\, \phi_j,\, \phi_k \in \mathcal{H}} p(\phi, \phi_j, \phi_k) \log \frac{p(\phi_i, \phi_j | \phi_k)}{p(\phi_i | \phi_k) p(\phi_j | \phi_k)}. \tag{6}$$

**Clustering** Only the edges in the most significant percentile are retained and the resulting graph is decomposed using a simple hierarchical clustering algorithm, with weighted average linkage as cost of merging, and taking as number of clusters the number of cuts of size 1 (i.e. of bipartite partitions of the graph joined by a single edge). This procedure should allow to identify the most connected components. These are further tested against dense modules/PCs.

## Results

**Artificial dataset** This procedure used to construct the artificila network is such that dense regulatory modules are numerous enough to compare the inference power among the different algorithms in a statistically relevant manner. Our results (Fig. 2(a)) show that small dense modules are similarly reconstructed by all the metrics up to a critical size, beyond which conditioned metrics perform worse. According to the graph illustrating the percentage of TP, dense modules with more than ten nodes are better identified by P and MI while the worst metric is CP. The clustering procedure (performed on the similarity matrix) reflects the same behaviour and in fact the best results are obtained for direct measures (P, MI). The main reason is the different topology emerging from the application of the 5 metrics (Table. 1 (SUPPLEMENTARY ???)). The graphs deriving from CMI, CP and GGM are sparser, with fewer connected components and almost all genes have at least one edge. On the contrary, the number of nodes without edges in the top percentile is drastically increasing using P and MI while the number of clusters follows an inverse relation. In Fig. 4 the percentage of complexes completely contained in: one cluster, two clusters, 3 clusters and more than 3 are shown.

***E.coli* dataset** Owing to the different genome organization and architecture, in prokaryotes regulatory mechanisms are much simpler than in eukaryotes. Genes are organized in transcriptional units, with one promoter for many consecutive genes, a feature absent in monocystic eukaryotic DNA. *E.coli* has only a few large complexes and also the combinatorial regulation of transcription is lower, so we expect the different algorithms to have more similar performances. We calculate for

both "templates" the matrices of pairwise similarities with the five different metrics and plot the percentage of TPs for the most significant percentile, for increasing sizes of the PCs (Fig. 2(b)) and combinatoriality of TFs (Fig. 3(a)). PCs are identified slightly better by the two direct metrics, although the number of relatively large complexes is too low to have statistical significance. What really changes is the ability to recover PCs with a clustering algorithm. The different performances emerging from the clustering (Fig. 4(b)) indicate that the highest correspondence between PCs and clusters are provided by P and MI. An example of that is given by the flagellum complex accounting for 24 genes. If the clustering procedure is performed by means of P and MI, the complex belongs entirely to a single cluster, which contains also other genes functionally related to the flagellum, like chemotactic genes and other genes involved in flagellar biogenesis and motility. Instead for CMI, CP and GGM the complex belongs respectively to 14, 6 and 24 clusters. Regarding TF–BS relationships, we expect the ability in recovering true interactions to be inversely proportional to the multiplicity of TFs. This is particularly true for the algorithms performing well on low multiplicity TF (P, MI and GGM) while CMI has a counterintuitive positive trend for multiregulated targets.

***S.cerevisiae* dataset** In *S.cerevisiae* fig. 2(c) shows clearly that for small complexes the performances of conditioned correlations are comparable with those of P and MI, up to a critical size above which the inference power of CMI, GGM and CP remains almost constant while the direct metrics increase their percentage of TPs. The results are consistent with the ones obtained for the artificial data. Qualitatively the results on the two organisms are the same, although the percentages of TP are higher in the simpler one. In addition, the critical size of the dense modules for which conditioned similarities start to fail is similar to the one obtained in the artificial network and *E.coli*, suggesting an intrinsic peculiarity of such similarity metrics. The clustering performances (Fig. 4(c)) for the five algorithms are coherent with those of the *E.coli* and artificial networks and once again give better results for the P and MI metrics. If we move to the network of TF–BS (Fig. 3(c)) we immediately notice that all the three conditioned metrics perform better than the direct ones, although in absolute terms results are worse than for *E.coli*. The reason for the low inference power regarding TF–BS can be that a single gene is regulated by multiple TFs acting in a differently combinatorial way in different environmental conditions, or that TFs do not show the large variations in expression, that can be seen for the corresponding regulated genes, but instead keep their expressions at low basal levels (Fig. 7 SUPPLEMENTARY???).

## Discussion

Comparing genome-wide networks inferred by means of different similarity metrics is not a simple task. Relying on the most significant percentile for the five similarity matrixes is a reasonable choice, suitable for our purposes. The reported results show that indeed different reverse engineering algorithms have performances which are tailored to different "characteristic" regulatory modules.

PCs are characterized by a very stable binding and give rise to a sort of post-transcriptional regulation, where gene products have to be expressed in a constant stoichiometric ratio and are mutually dependent one from each other, features absent in cause–effect relationships such as transcriptional activation. For the network generated with the model and the two real ones we tested the ability to recover dense modules/protein complexes of different size. The two real datasets has been tested also in respect to the TF–BS networks. Several important observations emerge from the results:

The critical size of a dense module for which direct similarity measures begin to perform better than the corrispective conditioned ones, for the dense modules, is about 10 on both artificial and real data. The dense modules that characterize PC are better cpatured by direct similarity measures, especially for large dense modules. This is almost the same in both organisms, in spite of the different complexity and the low experiments/genes ratio . On the contrary the conditional similarity measures are more suited to deal with causal dependecies such as TF–BS interactions, especially when the combinatorial complexity of the regulation increases. The ability to recover TF–BS interactions is roughly inversely proportional to the number of TF regulating a gene. Conditioned metrics are more robust (in .....) this multiplicity effect of TFs. eNeedless to say the inference power of all the algorithms is higher in the simpler organism, for both PC and TF–BS networks. This reflects the more complex eukaryote regulation, as deducible also from Fig. 1(a,b). The expression levels of TFs are often characterized by a lower variance than the corresponding BSs thereby complicating the problem further.

## Conclusion

The predictive power of a reverse engineering algorithm is clearly a function of several aspects. First of all system complexity, data quality and numerosity. In addition inference power depends on the type of interaction and the associate topology. Showing as we do in this paper that indeed the algorithms yield different performances coherently with the features they are meant to extrapolate from the data (direct for static and stable interactions, conditional for causal interactions) is already a significant and encouraging observation.

## References


[1] S. Balaji, M.M.Babu, L. Iyer, N. Luscombe, and L. Aravind. Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *J. Mol. Biol.*, 360:213–27, 2006.

[2] M. Bansal, V. Belcastro, A. Ambesi-Impiombato, and D. di Bernardo. How to infer gene networks from expression profiles. *Mol Syst Biol*, 3(78), 2007.

[3] K. Basso, A. A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera, and A. Califano. Reverse engineering of regulatory networks in human B cells. *Nat. Genet.*, 37:382–390, 2005.

[4] A. J. Butte and I. S. Kohane. Unsupervised knowledge discovery in medical databases using relevance networks. *Proc. AMIA Symp.*, pages 711–715, 1999.

[5] A. J. Butte and I. S. Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.*, pages 418–429, 2000.

[6] H. De Jong. Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, 9:67–103, 2002.

[7] A. de la Fuente, N. Bing, I. Hoeschele, and P. Mendes. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, 20(18):3565–3574, 2004.

[8] D. Edwards. *Introduction to Graphical Modelling*. Springer, 2000.

[9] J. Faith, M. Driscoll, V. A. Fusaro, E. Cosgrove, B. Hayete, F. S. Juhn, S. J. Schneider, and T. Gardner. Many microbe microarrays database: uniformly normalized affymetrix compendia with structured experimental metadata. *Nucleic Acids Research*, page to appear, 2007.

[10] T. S. Gardner and J. J. Faith. Reverse-engineering transcriptional control networks. *Physics of Life Rev.*, 2:65–88, 2005.

[11] U. Güldener, M. Münsterkötter, M. Oesterheld, P. Pagel, A. Ruepp, H. W. Mewes, and V. Stümpflen. Mpact: the mips protein interaction resource on yeast. *Nucleic Acids Res*, 34(Database issue):436–441, Jan 2006.

[12] V. Hatzimanikatis and K. H. Lee. Dynamical analysis of gene networks requires both mrna and protein expression information. *Metab Eng*, 1(4):275–281, Oct 1999.

[13] M. J. Herrgård, M. W. Covert, and B. U. Palsson. Reconciling gene expression data with known genome-scale regulatory network structures. *Genome Res*, 13(11):2423–2434, Nov 2003.

[14] D. Husmeier. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, 19:2271–2282, 2003.

[15] D. Hwang, A. G. Rust, S. Ramsey, J. J. Smith, D. M. Leslie, A. D. Weston, P. de Atauri, J. D. Aitchison, L. Hood, A. F. Siegel, and H. Bolouri. A data integration methodology for systems biology. *Proc Natl Acad Sci U S A*, 102(48):17296–17301, Nov 2005.

[16] R. Jansen, D. Greenbaum, and M. Gerstein. Relating whole-genome expression data with protein-protein interactions. *Genome Res*, 12(1):37–46, 2002.

[17] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein. A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data. *Science*, 302:449–453, 2003.

[18] P. D. Karp, M. Riley, S. M. Paley, A. Pellegrini-Toole, and M. Krummenacker. Eco cyc: encyclopedia of escherichia coli genes and metabolism. *Nucleic Acids Res*, 27(1):55–58, Jan 1999.

[19] S. A. Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *J Theor Biol*, 22(3):437–467, Mar 1969.

[20] D. S. Lee and H. Rieger. Comparative study of the transcriptional regulatory networks of e. coli and yeast: structural characteristics leading to marginal dynamic stability. *J Theor Biol*, 248(4):618–626, Oct 2007.

[21] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J. B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298(5594):799–804, Oct 2002.

[22] P. Lu, C. Vogel, R. Wang, X. Yao, and E. M. Marcotte. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol*, 25(1):117–124, Jan 2007.

[23] S. Ma, Q. Gong, and H. J. Bohnert. An arabidopsis gene network based on the graphical gaussian model. *Genome Res*, 17(11):1614–1625, Nov 2007.

[24] A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Favera, and A. Califano. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics*, 7(Suppl 1):S7, 2006.

[25] P. Mendes, W. Sha, and K. Ye. Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics*, 19(Suppl 2):ii122–ii129, 2003.

[26] J. R. Newman, S. Ghaemmaghami, J. Ihmels, D. K. Breslow, M. Noble, J. L. DeRisi, and J. S. Weissman. Single-cell proteomic analysis of s. cerevisiae reveals the architecture of biological noise. *Nature*, 441(7095):840–846, Jun 2006.

[27] L. Nie, G. Wu, and W. Zhang. Correlation of mrna expression and protein abundance affected by multiple sequence features related to translational efficiency in desulfovibrio vulgaris: a quantitative analysis. *Genetics*, 174(4):2229–2243, Dec 2006.

[28] M. Nomura. Regulation of ribosome biosynthesis in escherichia coli and saccharomyces cerevisiae: diversity and common principles. *J Bacteriol*, 181(22):6857–6864, Nov 1999.

[29] R. J. Planta. Regulation of ribosome synthesis in yeast. *Yeast*, 13(16):1505–1518, Dec 1997.

[30] Y. Qi, Z. Bar-Joseph, and J.Klein-Seetharaman. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins*, 63:490–500, 2006.

[31] A. Ruepp, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, M. Stransky, B. Waegele, T. Schmidt, O. N. Doudieu, V. Stümpflen, and H. W. Mewes. Corum: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res*, Oct 2007.

[32] H. Salgado, S. Gama-Castro, M. Peralta-Gil, E. Diaz-Peredo, F. Sanchez-Solano, A. Santos-Zavaleta, I. Martinez-Flores, V. Jimenez-Jacinto, C. Bonavides-Martinez, J. Segura-Salazar, A. Martinez-Antonio, and J. Collado-Vides. RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res.*, 34(Database issue):D394–397, 2006.

[33] J. Schäfer and K. Strimmer. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764, 2005.

[34] I. Shmulevich, E. R. Dougherty, S. Kim, and W. Zhang. Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, 18(2):261–274, Feb 2002.

[35] N. Simonis, J. van Helden, G. N. Cohen, and S. J. Wodak. Transcriptional regulation of protein complexes in yeast. *Genome Biol*, 5(5), 2004.

[36] N. Soranzo, G. Bianconi, and C. Altafini. Comparing relevance network algorithms for reverse engineering of large scale gene regulatory networks: synthetic vs real data. *Bioinformatics*, 23:1640–1647, 2007.

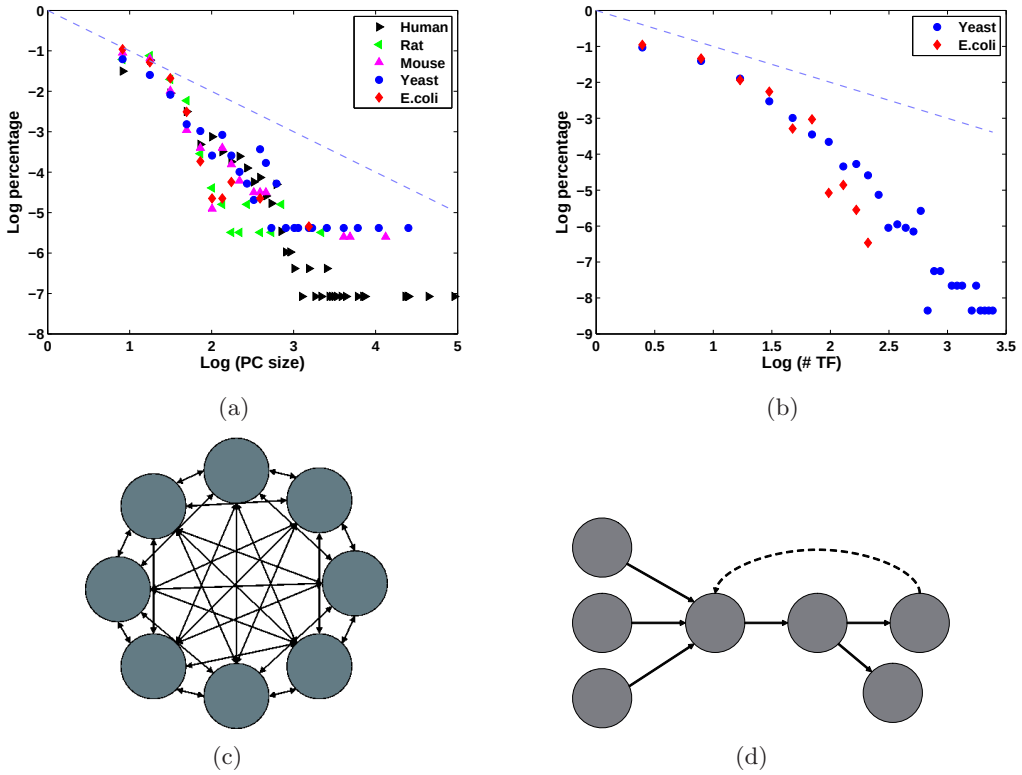
[37] V. Spirin and L. A. Mirny. Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci U S A*, 100(21):12123–12128, Oct 2003.

[38] E. Voit, A. R. Neves, and H. Santos. The intricate side of systems biology. *Proc Natl Acad Sci U S A*, 103(25):9452–9457, Jun 2006.

[39] M. K. S. Yeung, J. Tegnér, and J. J. Collins. Reverse engineering gene networks using singular value decomposition and robust regression. *Proc Natl Acad Sci U S A*, 99(9):6163–6168, 2002.

[40] H. Yu, A. Paccanaro, V. Trifonov, and M. Gerstein. Predicting interactions in protein networks by completing defective cliques. *Bioinformatics*, 22(7):823–829, Apr 2006.

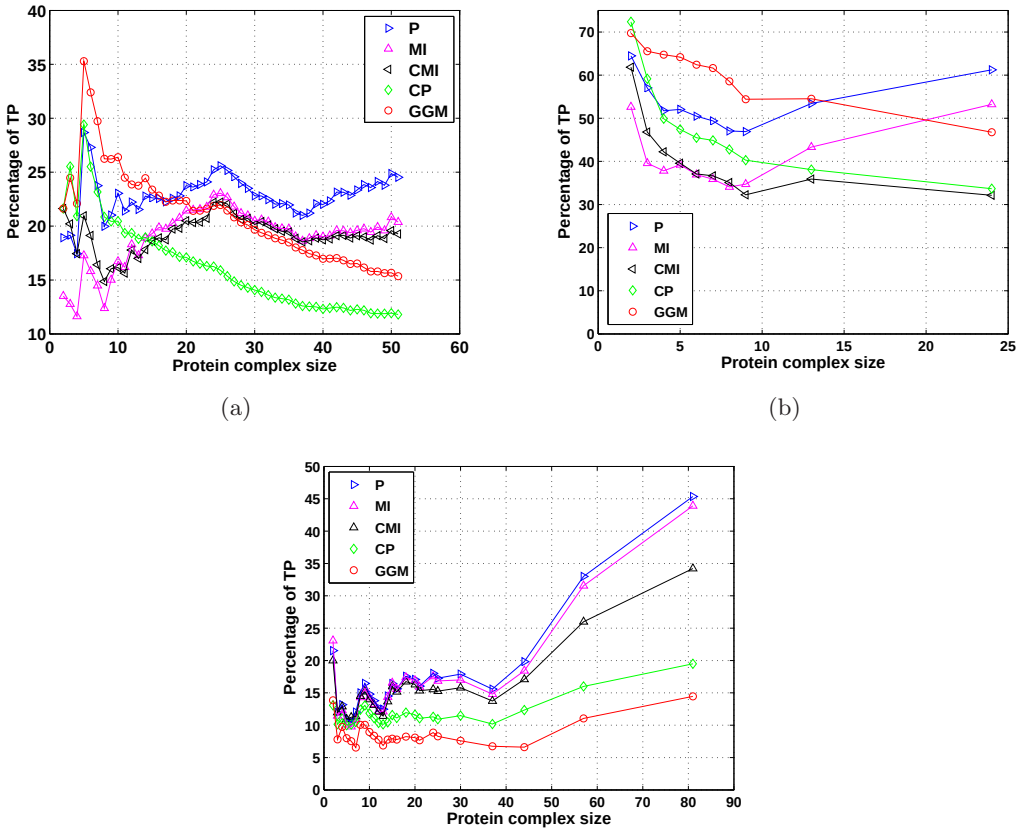
[41] Z. Zhang and M. Gerstein. Reconstructing genetic networks in yeast. *Nat Biotechnol*, 21(11):1295–1297, Nov 2003.




(a) (b)

(c) (d)

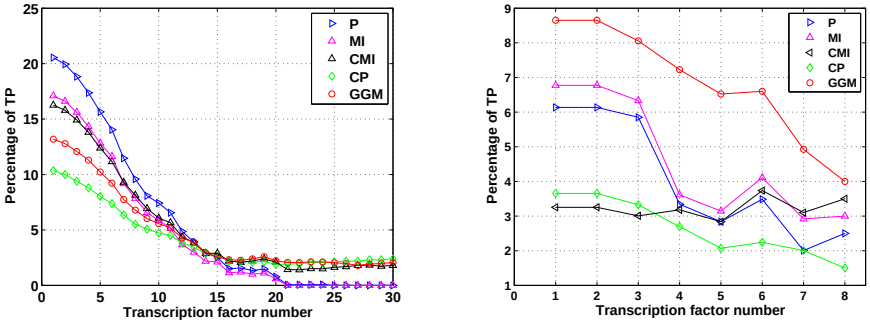
Figure 1: **Regulatory motifs and PC size density.** Log scale distribution of PC size (a), and number of TFs per gene (b) for different organisms. Scheme of the two regulatory motifs: (c) dense module, where all nodes are mutually connected (many feedback loops), (d) sparse module, accounting for only a few feedback loops and multi regulated genes.




(a) (b)

(c)

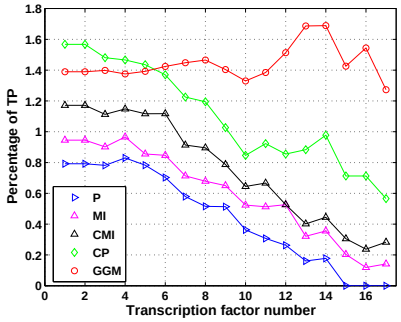
Figure 2: **Dense modules and PCs.** Percentage of TP edges for the five different correlation metrics for increasing size of the PCs, in: a) artificial dataset b) *E.coli* and c) *S.cerevisiae*. In all three cases, considering the percentage of TPs for the whole PC network the five different metrics can be ranked in the same order: P, MI, CMI, CP, GGM.




(a)

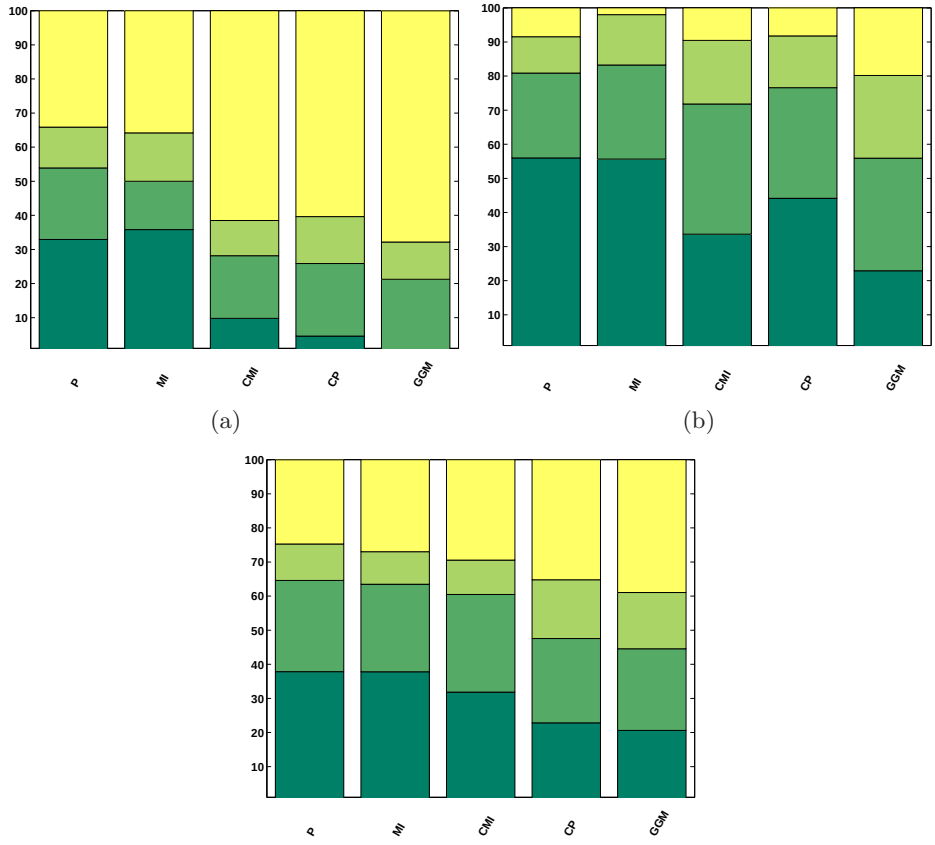



(b)




(c)

Figure 3: **Combinatorial transcription regulation.** Percentage of TPs for TF–BS network increasing the number of TF for the same BS, in: a) *E.coli* and b) *S.cerevisiae*. In *E.coli* the direct measures seem to be getting less effective as the multiplicity of TF increases (data are not conclusive however). In *S.cerevisiae*, although small in absolute terms, the conditioned measures outperform the direct ones regardlessof the combinatorial size of the TF.




(a) (b)

(c)

Figure 4: **Clustering dense modules/PC.** For (a) artificial network (b) *E.coli* and (c) *S.cerevisiae* the green scale represents the percentage of PCs belonging to a single cluster (dark green), two clusters, three clusters and more than three (yellow). This statistic highlights the drastic difference in clustering for the five similarity metrics, with the best unique correspondence in all datasets given by P and MI.

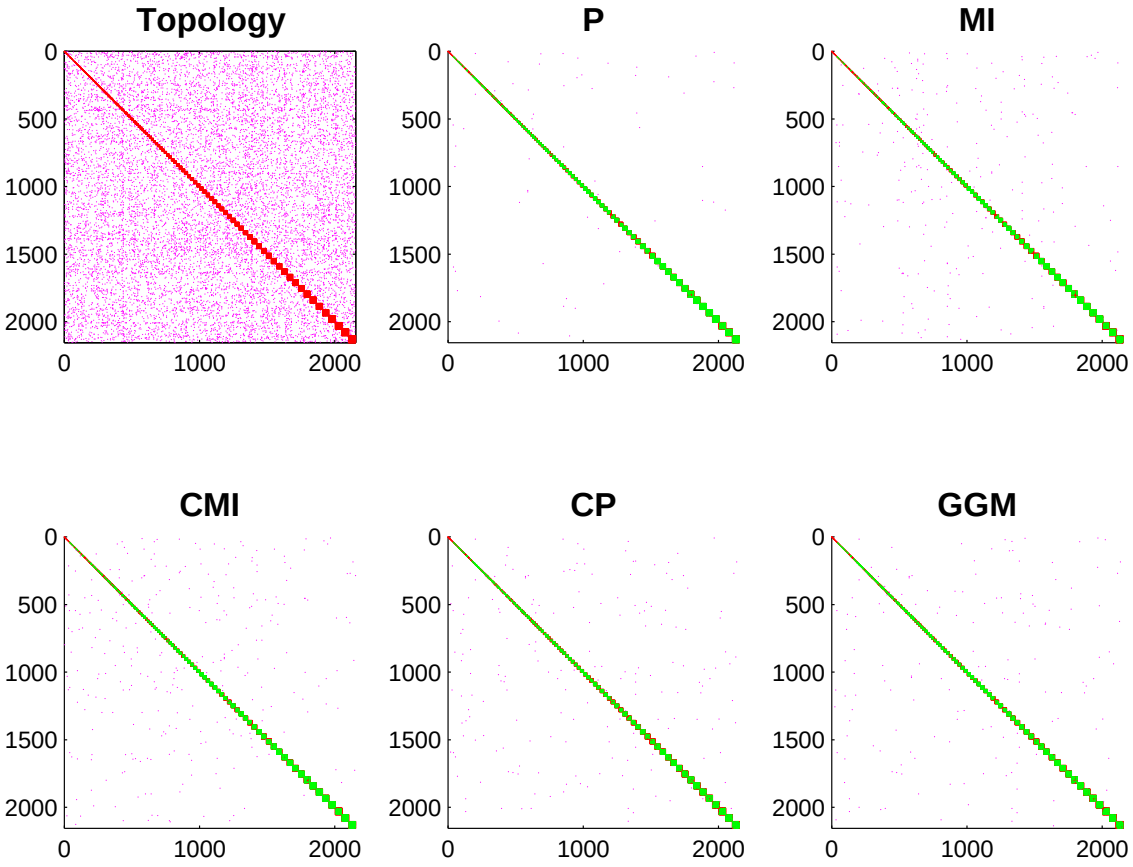
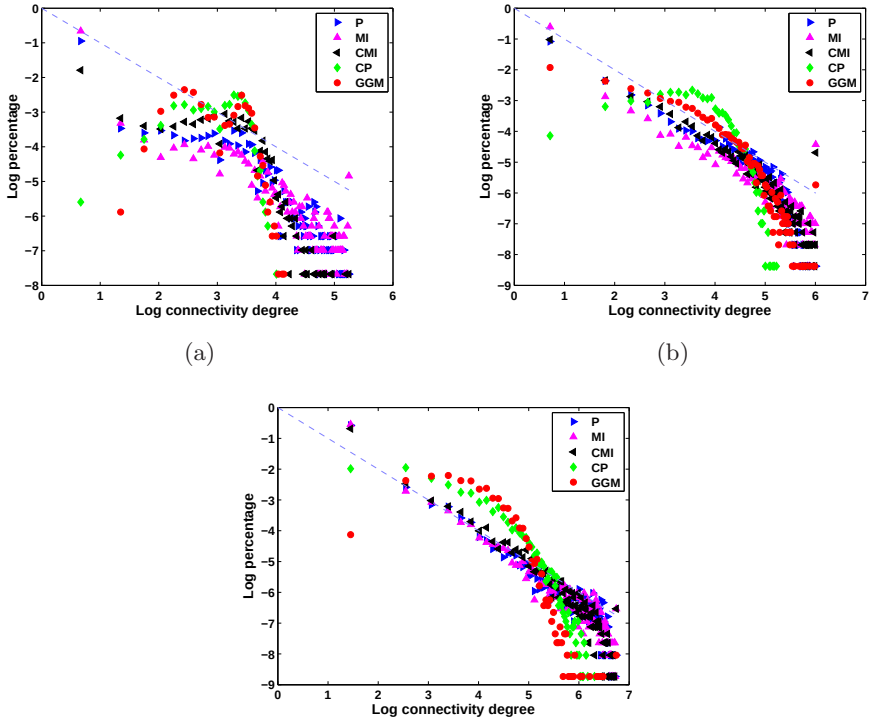


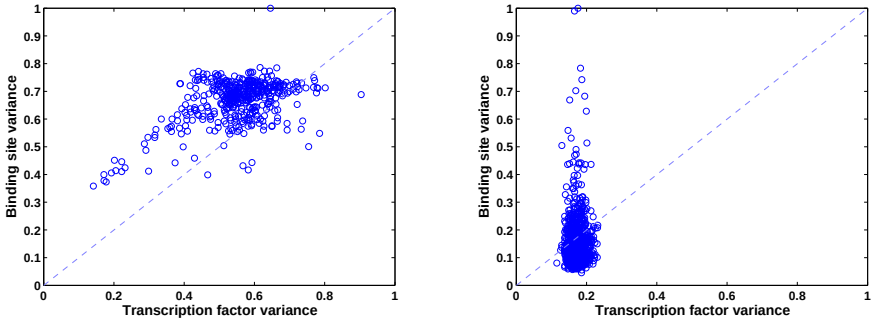

Figure 5: **Reconstruction on the artificial network: sparse versus dense modules.** The topology of the artificial network is shown in the left upper panel. In magenta are edges forming the sparse module, in red those of the dense modules. The reconstruction of the dense modules is fairly accurate, unlike for the sparse module. The other panels represent the TP edges for the two regulatory modules (green and magenta), and red for the missed dense module edges, for the five different similarity metrics.




(a) (b)

(c)

Figure 6: **Connectivity degree distribution of the reconstructions.** The percentage of nodes ($y$-axis) for the respective connectivity degree, in log scale, for the top percentile in 5 metrics for (a) artificial network (b) *E.coli* and (c) *S.cerevisiae*.




(a) (b)

Figure 7: **TF versus BS variance.** Scatter plots representing the TF against BS expression variance. In both organisms ((a) *E.coli* and (b) *S.cerevisiae*) BS expression is, most of the times, broader than corrisponding TF. Notice how especially in *S.cerevisiae* all TF have low variance and how most paires TF–BS edges live in the low variance corner.

Table 1: **Staitistics for the clusterization of dense modules/PC shown in Fig. 4.**The number of disconnected nodes (e.g. nodes having zero edges in the top percentile) and clusters, according to the number of bipartite partitions joined by a single edge, are represented in the first two columns. The remaining columns report the statistics shown in Fig. 4 (e.g. "1-2" shows the number of PC associated to two different clusters). The final column contains the number of dense modules/PC used

Clusters versus PC in artificial network

| | disconnected nodes | clusters | 1 - 1 | 1 - 2 | 1 - 3 | $1 \geqslant 4$ | PC |
|---|---|---|---|---|---|---|---|
| P | 659 | 167 | 55 | 35 | 20 | 57 | 167 |
| MI | 983 | 348 | 53 | 21 | 21 | 53 | 148 |
| CMI | 263 | 1794 | 17 | 32 | 18 | 107 | 174 |
| CP | 1 | 333 | 8 | 37 | 24 | 105 | 174 |
| GGM | 1 | 2141 | 0 | 37 | 19 | 118 | 174 |

Clusters versus PC in *E.coli*

| | disconnected nodes | clusters | 1 - 1 | 1 - 2 | 1 - 3 | $1 \geqslant 4$ | PC |
|---|---|---|---|---|---|---|---|
| P | 644 | 537 | 126 | 56 | 24 | 19 | 225 |
| MI | 1684 | 682 | 113 | 56 | 30 | 4 | 203 |
| CMI | 597 | 2053 | 74 | 84 | 41 | 21 | 220 |
| CP | 0 | 129 | 102 | 75 | 35 | 19 | 231 |
| GGM | 142 | 3805 | 52 | 75 | 55 | 45 | 227 |

Clusters versus PC in *S.cerevisiae*

| | disconnected nodes | clusters | 1 - 1 | 1 - 2 | 1 - 3 | $1 \geqslant 4$ | PC |
|---|---|---|---|---|---|---|---|
| P | 2085 | 1101 | 92 | 65 | 26 | 60 | 243 |
| MI | 2359 | 1618 | 87 | 59 | 22 | 62 | 230 |
| CMI | 1668 | 1833 | 79 | 71 | 25 | 73 | 248 |
| CP | 52 | 1229 | 61 | 66 | 46 | 94 | 267 |
| GGM | 0 | 5275 | 55 | 64 | 44 | 104 | 267 |