# Reverse Engineering of the Yeast Transcriptional Network Using the ARACNE algorithm

Adam A. Margolin[1,2], Nilanjana Banerjee[2], Ilya Nemenman[2], Andrea Califano[1,2]

[1]Department of Biomedical Informatics, Columbia University, 622 West 168th Street, Vanderbilt Clinic 5th Floor, New York, New York 10032

[2]Joint Centers for System Biology, Columbia University, Russ Berrie Pavilion, 1150 St. Nicholas Ave, Rm 121, New York, New York 10032

## Abstract

Cellular phenotypes are determined by dynamical activity of networks of co-regulated genes. Elucidating such networks is crucial for the understanding of normal cell physiology as well as for the dissection of complex pathologic phenotypes. Recently we have shown that *ARACNE*, a novel information-theoretic algorithm for reverse engineering of transcriptional networks using microarray data, holds significant promise for the *genome-wide* analysis of mammalian networks, which had never been performed *in silico*.

In this paper we present the application of ARACNE to reverse engineering of transcriptional networks in the yeast *Saccharomyces cerevisiae*. This provides another platform for further comparisons of the new method to a variety of established ones, which have been extensively used to analyze this important model organism. Moreover, it provides an additional genome-wide interactome for the yeast using a method that is shown to produce very few false positive interactions, both *in vivo* and *in silico*. Additionally, we investigate the global topological properties of the reconstructed networks and determine that the scale-free structure suggested by existing models should be taken cautiously. Finally, analysis of the ARN1 sub-network, which has also been investigated using Bayesian Networks, shows that ARACNE is able to distinguish the key regulatory elements of this pathway from within a large cluster of co-regulated genes.

## 1. INTRODUCTION

Cellular phenotypes are determined by complex relationships among genes and their products that control the majority of cellular functions. By modeling these relationships, the whole genome can be organized into a network of genetic interactions. Understanding this organization is crucial to elucidate normal cell physiology as well as to dissect complex pathologic phenotypes. The advent of high throughput assays for monitoring gene expression profiles across entire genomes has spawned much research aimed at using this data to “reverse engineer” genetic networks by grouping together genes that exhibit similar transcriptional responses to various cellular conditions [1, 2]. While this approach has shown promise in applications such as the stratification of disease-related phenotypes [3], the organization of genes into co-regulated clusters is too coarse a representation to identify individual interactions. This is because as biochemical signals travel through cellular networks the expression of many genes that interact only indirectly may become strongly correlated. More generally, as has long been recognized in statistical physics, a long range order (that is, a high correlation among indirectly interacting random variables) can easily result from several short range, pairwise interactions [4]. Thus one cannot use correlations, or *any other* local dependency measure, as a tool for the reconstruction of interaction networks without additional assumptions.

Within the last few years a number of more sophisticated approaches for the reverse engineering of cellular networks, also called deconvolution, from gene expression data have emerged. The goal of such methods is to produce a high-fidelity representation of the cellular network topology as a graph, where genes are represented as nodes and regulatory interactions as edges. However, all available approaches suffer to some degree from various problems such as overfitting, exponential complexity, reliance on non-realistic network models, or a critical dependency on data that is only available for simple organisms. We recently introduced a novel, information theoretic algorithm, *ARACNE*, for the reverse engineering of gene regulatory network that overcomes some of these critical limitations. ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks) [5] extends upon traditional clustering approaches and reconstructs finer-grained dependencies within gene clusters by further discriminating between direct and

indirect interactions. ARACNE compares very favorably with existing reverse-engineering methods, such as Bayesian Networks and Relevance Networks, and scales successfully to the complexity of large mammalian networks. Analysis of mammalian networks from a large number of microarray profiles for normal, tumor-related, and experimentally manipulated B cells shows a significant ability to deconvolute complex networks. By analyzing a specific sub-network that includes the c-MYC proto-oncogene, we have shown that the method successfully recapitulates known targets of this transcription factor and is able to discover many previously unknown targets. The latter were biochemically validated using Chromatin Immunoprecipitation assays, with better than 90% success rate [6].

In this paper, we first review ARACNE (within its theoretical context) and then summarize an extensive comparison of ARACNE vs. Bayesian Networks and Relevance Networks using a synthetic platform of realistic complexity. Finally, we report the results of the ARACNE-based deconvolution of gene regulatory networks in the yeast *Saccharomyces cerevisiae*. This analysis provides a basis for the objective comparison of the algorithm's performance in a real biological context, where substantial background information is available to assess the accuracy of *in silico* predictions. Additionally, results of this analysis provide an additional and orthogonal "interactome" for yeast, which (based on the benchmark's results) is likely to have a low rate of false positive interactions as compared to previous microarray-based studies. This can be integrated with other models produced either experimentally or *in silico*.

## 2. Theoretical Background

Let us start by noting that (as most reverse-engineering methods) we will focus on the study of steady-state inter-gene statistical dependences only. These are defined based on the definition of [7], which builds on ideas from the Markov networks literature [8]. Briefly, by analogy with statistical physics, we write the joint probability distribution (JPD) of the stationary expressions of all genes, $P(\{g_i\})$, $i = 1,\ldots,N$, as:

$$P(\{g_i\}) = \frac{1}{Z}\exp\left[-\sum_i \phi_i(g_i) - \sum_{i,j}\phi_{ij}(g_i,g_j) - \sum_{i,j,k}\phi_{ijk}(g_i,g_j,g_k) - \mathsf{L}\right] \equiv \exp[-H(\{g_i\})] \qquad (1)$$

where $N$ is the number of genes, $Z$ is the *partition function*, $\phi_i(g_i)$ are *potentials*, and $H(\{g_i\})$ is the *Hamiltonian* that defines the system's dynamics. Then a set of variables is called *interacting* if and only if the single potential that depends exclusively on these variables is nonzero. The expansion in Eq. (1) does not define the potentials uniquely, and additional natural constraints of the Maximum Entropy type are needed to avoid the ambiguity (see [7, 9] for details).

Given the relatively small number of expression profile samples, $M$, that can be realistically obtained, it is infeasible to infer an exponentially large number of potential *n*-way interactions, as suggested by Eq. (1). Rather, a set of simplifying assumptions must be made about the variable dependency structure. Eq. (1) provides a principled and controlled way to introduce such approximations. The simplest model is one where genes are assumed independent, i.e., $H(\{g_i\}) = \sum \phi_i(g_i)$, such that the first-order potentials can be evaluated from the marginal probabilities, $P(g_i)$, which are in turn estimated from samples. As more data become available, we should be able to reliably estimate higher order marginals and incorporate the corresponding potentials progressively, such that for $M \to \infty$ the complete form of the JPD is restored. In fact, $M > 100$ is generally sufficient to estimate 2-way marginals in genomics problems, while $P(g_i, g_j, g_k)$ requires about an order of magnitude more samples. Thus, we truncate Eq. (1) at the pairwise interactions level, $H(\{g_i\}) = \sum_i \phi_i(g_i) + \sum_{i,j}\phi_{ij}(g_i, g_j)$. Within this approximation,

two genes are declared non-interacting if they are statistically independent (i.e. $P(g_i, g_j) \approx P(g_i)P(g_j)$ ), and more complex interactions are not investigated.

However, the reverse is not true, and genes may still not interact (zero potential) even if the marginals do not factorize. Therefore, even focusing on pairwise interactions, the problem of network reverse engineering is still nontrivial: two genes can have nonzero correlation due to a confounding effect of a third one. That is, we may still have $P(g_i, g_j) \neq P(g_i)P(g_j)$ while $\phi_{ij} = 0$ meaning that there is no direct interaction. Since the number of potential pairwise interactions is huge (quadratic in the number of genes), uncovering them to remove false positives presents a formidable challenge to all network reconstruction algorithms. To date, no method has been proposed to solve this issue exactly and to reconstruct an arbitrary two-way interaction network reliably from a *finite number of samples* and in a *computationally feasible time*. However, if the regulatory network can be represented as a tree, then we prove that ARACNE can reconstruct it exactly for $M \to \infty$.

## 3. ALGORITHM

ARACNE relies on a two-step process. *First*, candidate interactions are identified by estimating pairwise gene-gene mutual information (MI) $I(g_i, g_j) = I_{ij} = \left\langle \log[P(g_i, g_j) / P(g_i)P(g_j)] \right\rangle$ and by filtering them using an appropriate threshold, $I_0$, computed based on a specific p-value, $p_0$, in the null-hypothesis of two independent genes. This step is basically equivalent to the Relevance Networks method [10], and, as such, suffers from critical limitations. In particular, as previously discussed, genes separated by one or more intermediaries may be highly co-regulated without implying a direct physical interaction.

Thus, in its *second step*, ARACNE removes the vast majority of indirect candidate interactions using a well-known property of mutual information – the data processing inequality (DPI) [11] – not been previously applied to the reverse engineering of networks. The DPI states that if genes $g_1$ and $g_3$ interact only through a third gene, $g_2$, (i.e., if the interaction network is of the form $g_1 \leftrightarrow ... \leftrightarrow g_2 \leftrightarrow ... \leftrightarrow g_3$ and no alternative path exists between $g_1$ and $g_3$), then

$$I(g_1, g_3) \leq \min\left[ I(g_1, g_2); I(g_2, g_3) \right]. \tag{2}$$

Correspondingly, ARACNE starts with a network graph where each $I_{ij} > I_0$ is initially represented by an edge ( $i \leftrightarrow j$ ). It then examines each gene triplet for which all three edges are above statistical significance and removes the edge with the smallest mutual information. Triplets are analyzed irrespective of whether edges have been previously marked for removal by the DPI (based on a different triplet). Thus the network reconstructed by the algorithm is independent of the order in which the triplets are examined.

<u>Theorem:</u> If MIs can be estimated with no errors, then ARACNE (with threshold p-value of 1, or equivalently, only the second step) reconstructs the underlying interaction network exactly, provided this network is a tree and has only pairwise interactions.

*Proof of the Theorem:* First, consider that for every pair of nodes $g_i$ and $g_k$ that are not physically interacting there is at least one other node $g_j$ that separates them on the tree. Thus, applying the DPI to the (*ijk*) triplet leads to the removal of the (*ik*) edge and only edges corresponding to true interactions survive. Similarly, each removed edge cannot correspond to a true interaction. Consider some (*ijk*) triplet. One of its genes, say $g_j$, may separate the other two. In this case the removed edge (*ik*) clearly does not correspond to a true interaction in the tree. Alternatively, there may be no separating gene, and one may be able to move between any gene pair in the triplet without going through the third one. In this case none of the three edges is in the true graph, and

any DPI-removed edge would not correspond to a true interaction. Thus, all removed edges are indirect, while all remaining edges are factual. The network is reconstructed exactly. □

As we will demonstrate using a synthetic dataset, the introduction of the DPI results in a remarkable reduction of false positive interactions with minimal impact on false negative ones. However, the algorithm is not guaranteed to reconstruct correct networks if loops are present (in fact, *every* loop with only three genes will be opened along the weakest edge). However, if loops are large, then the network has essentially a local tree structure and the algorithm will work well. This is because nodes in a network generally decorrelate rather quickly, and interactions over more than a few separating edges are weak, reducing the impact of large loops. While a local tree-like structure is a reasonable first-order approximation for biological networks, there will be many violations. For instance, the feed forward loop (a three-gene loop) has been found to be over-represented in biological circuits [12]. In practice, as it will be shown, even in the presence of tight regulatory loops and complex network topology, ARACNE performs remarkably well and outperforms comparable reverse engineering algorithms.

In the current implementation of the algorithm, we use a computationally efficient Gaussian Kernel estimator [13] to estimate MI. Given two measurement vectors $\{x_i\}$ and $\{y_i\}$,

$$I\left(\{x_i\},\{y_i\}\right)=\frac{1}{M}\sum_i \log\frac{f\left(x_i,y_i\right)}{f\left(x_i\right)f\left(y_i\right)}, \tag{3}$$

where $f(x,y)$ and $f(x)$ are Gaussian kernel density estimators defined as:

$$f\left(x,y\right)=\frac{1}{2\pi h^2 M}\sum_i \exp\left\{-\frac{\left(x-x_i\right)^2+\left(y-y_i\right)^2}{2h^2}\right\},$$

$$f\left(x\right)=\frac{1}{\sqrt{2\pi}hM}\sum_i \exp\left\{-\frac{\left(x-x_i\right)^2}{2h^2}\right\}$$

Here, $h=h(M)$ is the kernel width. This estimator is asymptotically unbiased for $M\to\infty$, as long as $h(M)\to 0$ and $[h(M)]^2 M\to\infty$. Unfortunately, for finite $M$, the bias strongly depends on the choice of $h(M)$, and the correct choice is not universal. However, ARACNE's performance does not depend directly on the accuracy of the MI estimate, but rather on the accuracy of the estimation of MI ranks: to test if MI is statistically significant or to apply DPI, one only needs to check if $I_{ij}\geq I_0$, or if $I_{ij}>I_{ik}$, respectively; that is, only to rank MI estimates. It turns out that for fixed $h$ the bias tends to cancel out, especially for $\overline{I}_{ij}\approx\overline{I}_{kl}$, and the ordering of MI estimates is only weakly dependent on the choice of $h$. Thus it is accurately estimated even when MI is not. Thus selecting a single "ensemble best" value of *h* rather than searching for the best kernel width for each estimate (a computationally intensive operation) impedes performance very little.

With such a choice, ARACNE's complexity is $O(N^3+N^2M^2)$, where, as always, $M$ is the number of samples, and $N$ is the number of genes. This is allows to effectively analyze networks with tens of thousands of genes. We refer the reader to [6] for details of selection of the kernel width as well as the other adjustable parameter, the DPI tolerance, $\tau$, which can be used to further minimize the impact of potential MI estimation errors by transforming the exact form of the DPI inequality to the form $I_{ij}\leq I_{ik}(1-\tau)$. We note that $\tau=0$ corresponds to the strict implementation of the DPI, while $\tau=1$ corresponds to Relevance networks, where no DPI is applied. Values in between explore the range between the two extremes. In particular, in [6], we

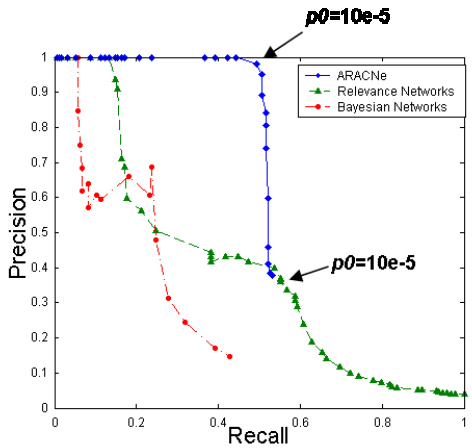


**Figure 1** Precision vs. Recall for 1,000 samples generated from the Mendes networks. PRCs for ARACNE are consistently better than for the other algorithms. That is, for any reasonable precision (i.e. $>40\%$), ARACNE has a significantly higher recall than the other methods, and its precision reaches ~100% at significant recall values. To isolate the performance gain associated with introduction of the DPI, points on the PRCs for ARACNE and RNs are indicated corresponding to $p_0 = 10^{-5}$ (the value yieding $<0.5$ expected false positives for 4,950 potential interactions). Using this p-value, ARACNE correctly infers 97 out of 194 true interactions, with on 2 false positives. By contrast, Relevance Networks infer 113 true connections with 234 false positives. Therefore, the DPI eliminates 232 false candidate interactions at the expense of only 15 true positives. Moreover, as shown, the statistical significance threshold yielding near optimal performance can set a-priori, by choosing a p-value (calculated from the background distribution described earlier) yielding a low expected number of false positives for the sample size in question.

show that for tolerances up to about 20% the number of true positives inferred by the algorithm can be increased with very little expense in terms of false positives.

# 4. RESULTS

We benchmark our algorithm by comparing its performance to that of Relevance Networks (RNs) [10] and Bayesian Networks (BNs), as implemented by [14]. RNs are basically equivalent to ARACNE without the DPI and thus are representative of a class of statistical/information theoretical which define two-way probabilistic measures of gene correlation to distinguish potential interactions from background noise. This comparison is important to assess the performance gain associated with the introduction of the DPI. BNs have emerged as some of the most widely used reverse engineering methods and provide an ideal comparative benchmark. We compare these three algorithms' ability to reconstruct a realistically implemented synthetic network, and to identify key regulatory elements in a yeast iron homeostasis pathway. We conclude by analyzing the global topological properties of the whole-genome regulatory network reconstructed by ARACNE.

### A Synthetic Network Benchmark

***Validation Framework:*** We have benchmarked *ARACNE* against two other algorithms using a synthetic networks model proposed by [15] as a realistic platform for the objective comparison of reverse engineering algorithms. These networks contain 100 nodes and 200 interactions organized in a scale-free topology [16], and evolve according to a multiplicative Hill dynamics. Such networks present a formidable challenge to reconstruction algorithms due to (a) their realistic complexity, (b) the presence of many regulatory loops, (c) the presence of a few highly interconnected genes, and (d) the biologically motivated non-linear transcriptional dependencies among genes. To generate synthetic microarrays, we randomly vary the efficiency of gene synthesis and degradation reactions for each synthetic sample at the beginning of each simulation. This models the sampling of a population of distinct cellular phenotypes at random time points (but at equilibrium). Reconstruction of these networks is evaluated by calculating *recall*, $N_{TP}/(N_{TP}+N_{FN})$, and *precision*, $N_{TP}/(N_{TP}+N_{FP})$, which, respectively, measure the fraction of true interactions correctly inferred by the algorithm and the fraction of genuine interactions among all predicted ones ($N_{TP}$, $N_{FP}$, $N_{TN}$, and $N_{FN}$ stand for true/false positives/negatives).

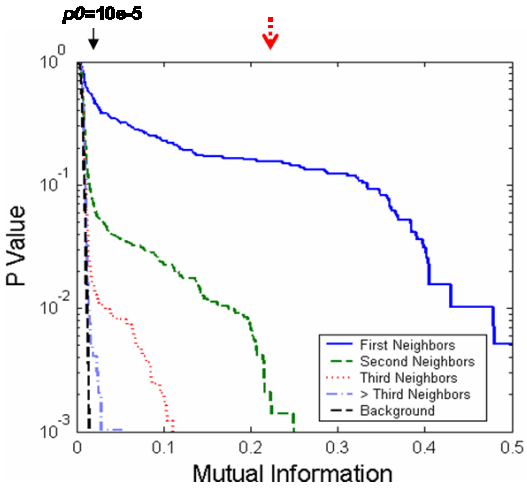


**Figure 2** Here we plot the log of the empirical probability that MI for a given separation between genes is above some value (in nats) marked on the horizontal axis. For both topologies, high MI values are significantly more probable for closer genes. Statistical significance thresholds of $10^{-5}$ for the background MI distribution, corresponding to $I_0 = 0.0175$ nats, is marked on each graph. As shown, this threshold retains a large number of indirect candidate interactions, and there is no threshold that would be able to separate indirect and direct interactions; a threshold that eliminates most of the former (red arrows) also eliminates the majority of the latter.

Precision vs. Recall curves (PRCs) are a better match than the more familiar ROC curves for problems where the number of true negatives is far greater than that of true positives, which is the case in large sparse networks.

***Performance Analysis:*** PRCs are shown in Figure 1 for all three comparative algorithms. These curves were obtained by changing the MI threshold and the Dirichlet pseudocounts, respectively for ARACNE/RNs and BNs. As is obvious from the figure, ARACNE performs consistently better than BNs and RNs, and achieves remarkably low false positive rates. Such high precision is necessary to guide experimental validation of the method's predictions. Using 1,000 samples, over half of all edges can be inferred with hardly any false positives. As shown, for a given p-value, introduction of the DPI produces a dramatic increase in precision with hardly any impact in recall, indicating that the DPI is highly efficient in filtering false candidate interactions with minimal impact on true positives. The reason for this success can be understood by considering the distribution of MIs as a function of the length of the shortest path connecting each gene pair (degree of connectivity). As shown in Figure 2, there is no unique threshold that can separate direct and indirectly interacting genes, as is evident by the high false positive rates in the PRC for Relevance Networks. However, MI is larger for directly interacting genes and decreases rapidly with this distance, raising the possibility of eliminating a substantial number of distant indirect associations by imposing a slightly conservative threshold that will eliminate only a few true interactions, while connections with enriched mutual information due to indirectly interacting genes can be eliminated a-posteriori via the DPI. Moreover, as discussed in [5], ARACNE's performance remains stable as the sample size decreases. In particular, the number of true positives decays gracefully while the number of false positives remains extremely low. For this data set and a sample size of only 125 synthetic microarrays, ARACNE still recovers 46 true interactions with only 3 false positives.

**Analysis of Regulatory Networks in *S. cerevisiae***

To further benchmark our method in a real biological context, we compare the three methods on the reconstruction of regulatory networks in *S. cerevisiae*, an organism for which many biochemical interactions have been validated experimentally, and for which several reverse engineering studies have been considered to infer such interactions. We applied our method to two *S. cerevisiae* gene expression data sets. One data set monitors response of yeast cells to diverse environmental transitions [17]. The second data (Rosetta Compendium) captures yeast transcriptional response corresponding to nearly 300 diverse mutations and chemical treatments [18].

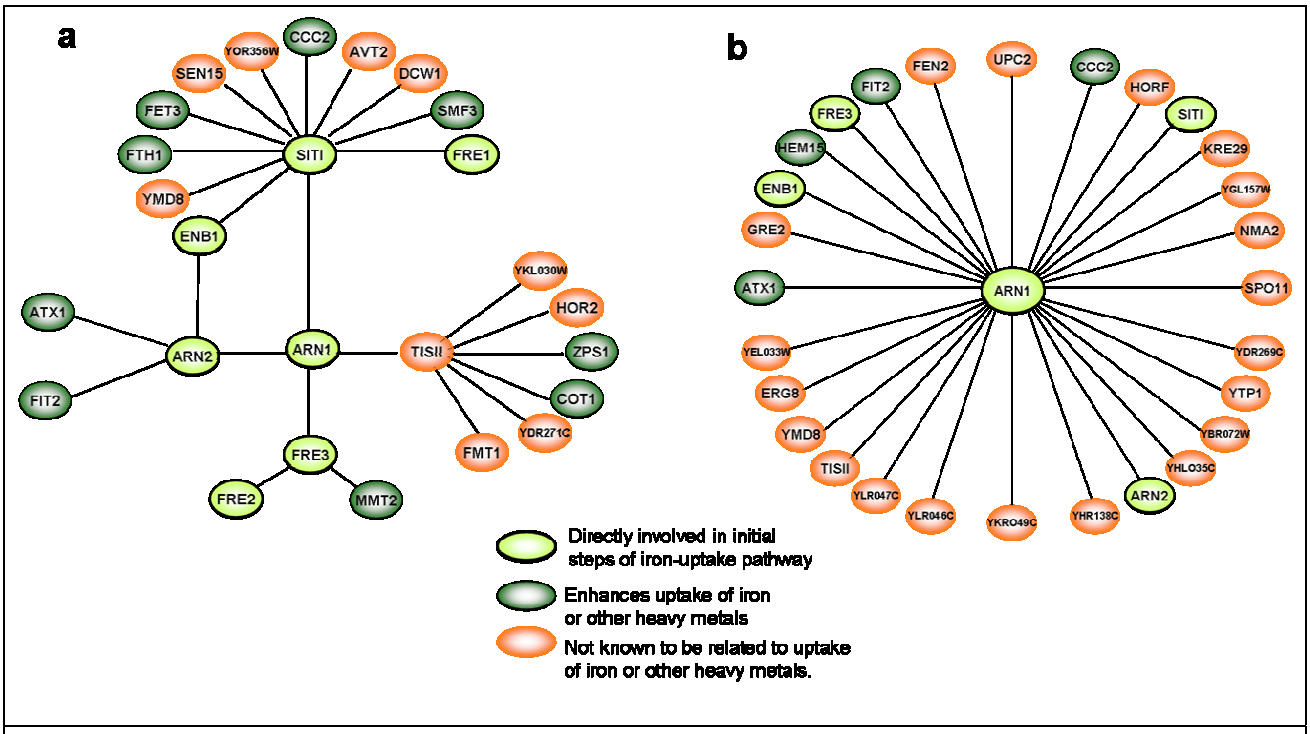


Figure 3 **(a)** ARN1-centered network with first and second neighbors. The first neighbors are enriched in the two classes of proteins that are involved in the intial steps of iron-uptake: siderophore transporters (SIT1, ARN2) and reductase (FRE3) (nodes shown in bright green). Many second neighbors enhance uptake of iron or other heavy metals (nodes shown in dark green). **(b)** Arn1-centered network with 15% tolerance. Several direct and indirect players of iron-uptake are identified but there are many neighbors which are not known to be related to iron homeostais.

***Analysis of the ARN1 Iron Homeostasis Pathway:*** We examine the reconstruction of an iron homeostasis pathway centered around the ARN1 gene that was previously used to test a Bayesian Networks approach to network reconstruction [19], using the dataset described in [18].

*ARN1-mediated iron uptake:* Although iron is one of the most abundant elements, ionic forms of iron, especially iron (III), its most common state, are very insoluble under physiological conditions. Evolution of life has depended on development of effective methods for its assimilation. The yeast Saccharomyces cerevisiae can use two different high-affinity mechanisms to take up iron from the extracellular medium [20, 21]. The reductive mechanism involves the reduction of extracellular ferric chelates via the inducible plasma membrane reductases Fre1p, Fre2p and Fre3p. The other method is a non-reductive mechanism which utilizes high-affinity transporters to shuttle ferric complexes into the cells prior to any reduction step. These transporters, encoded by ARN1, ARN2/TAF1, ARN3/SIT1 and ARN4/ENB1, each uptake specific forms of iron-bound siderophores. Another family of genes FIT1, FIT2 and FIT3 enhance iron uptake but are not essential for the process [22].

It is clear that the first response to iron deprivation is modulated by two groups of proteins: the plasma membrane reductases (FRE1, FRE2 and FRE3) and ferrous transporters (ARN1, ARN2/TAF1, ARN3/SIT1).

*Reconstruction Performance:* In **Figure 3**.a, we show the first and second neighbors of ARN1 generated by ARACNE with 0% tolerance. Remarkably, ARACNE is able to capture the two direct groups players in the initiation of the iron uptake pathway – the plasma membrane

reductase (FRE3) and the ferrous transporters (ARN3/SIT1, ARN2/TAF1). The secondary players that enhance iron uptake, or enhance uptake of other metals like copper or zinc are abundant in the second neighbors (e.g. FIT1, CCC1, FET3, MMT2, ATX1 and COT). One ferrous-transporter missing among the first neighbors is ARN4/ENB1. But ARN4/ENB1 does have ARN2/TAF1 and ARN3/SIT1 as its first neighbors. The ARN1 network generated by ARACNE is richer in genes directly related to iron uptake than an earlier ARN1 network reverse-engineered using BNs [19]. While three of the four ARN1 first neighbors inferred by ARACNE are essential for and directly related to iron- uptake, only two out of six of the ARN1 first neighbors of the BN approach fall in that category. Moreover, ARACNE is able to capture the modular structure of the initial steps of iron-uptake pathway (the transporter and reductase activity) more clearly than the BN approach.

When the tolerance is relaxed to 15%, the ARN1-centered network recruits several additional first neighbors interactions (28), as shown in **Figure 3**.b. While some of the relevant iron-uptake genes are present – the network is now crowded with many genes from complementary pathways and it is difficult to assess the hierarchy of the iron-uptake pathway. With a complete (100%) relaxation of tolerance, which is equivalent to the Relevance Networks approach, ARN1 has over 200 first neighbors and the importance of iron homeostasis becomes undetectable.

***Analysis of the Global S. cerevisiae Network:*** ARACNE was able to infer whole-genome interaction networks for the Rosetta and the Stress Response data respectively (at the DPI tolerance of 15%). As ARACNE is believed to produce very few false positives even with this tolerance, the disparity between the numbers of inferred interactions is probably due to a different nature of the expression data in the datasets. The Rosetta data consists of responses to various knockouts and drugs, which only lead to relatively small perturbations of the entire system (according to [23], the steady state of a weakly perturbed Yeast transcriptional network is very close to that of the unperturbed one). As mutual information is bounded from above by variability (as measured by the entropy), such robustness to perturbations decreases the number of reconstructable edges. On the other hand, the stress response set consists of time series data from a macroscopic biological behavior (motion of the organism). This involves substantial variation of many expressions and provides the necessary dynamic range for discovery of the interactions.

We summarize the inferred networks by their global connectivity properties in Figure 4. The results show a power-law tail in the relationship between the number of genes, *n*, in the networks and their number of interactions, *k*, which extends over two orders of magnitude in *n*. This can be interpreted as an evidence for a scale free structure for underlying networks [16], with few hubs and many weakly connected genes. However, for the stress response data there also is a peak at $k \approx 20$ in the degree distribution. Thus such identifications should be made with an extreme caution (see Discussion).

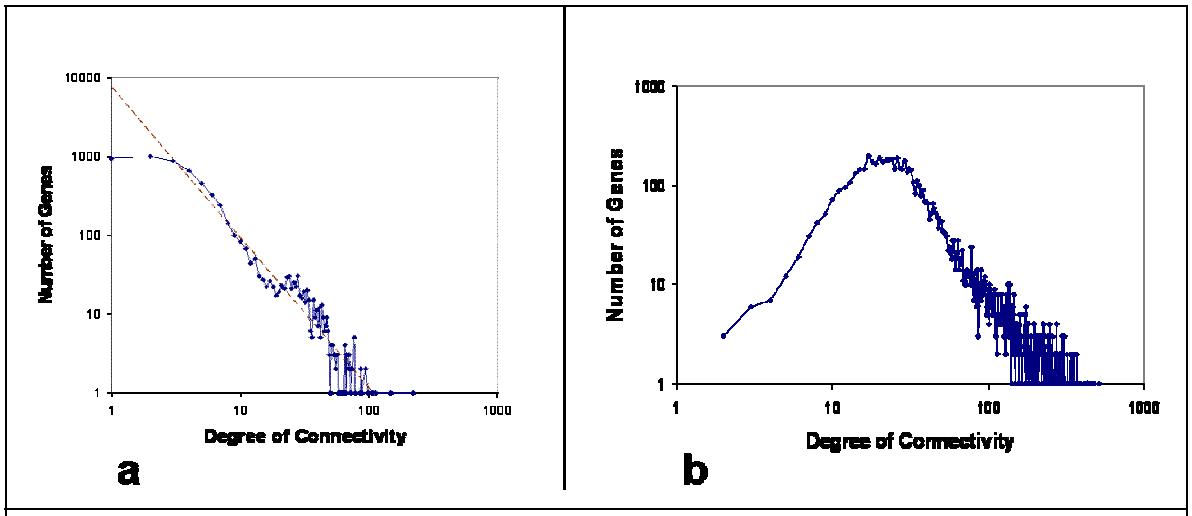


**Figure 4.** Distribution of nodes with a specific number of incident edges (degree of connectivity) in log-log scale using a tolerance $\varepsilon = 0.15$ for the (**a**) Rosetta and (**b**) stress response dataset. There is some evidence that the Rosetta dataset possesses a scale-free topology with the power law of . On the other hand, the graph of the degree distribution for the stress response dataset, which also has a power law tail has a clear peak at about 20 incident edges. These results persist for the tolerance down to 0%, suggesting clear biological differences between the datasets.

## DISCUSSION

Inferring functional dependencies between genes from co-regulated expression profiles has presented significant challenges for reverse engineering algorithms that rely on statistical correlations between genes since even without direct interactions genes may be highly co-regulated. However, due to inherent stochasticity in biochemical reactions, mutual information between genes decreases as genes become more distantly related. We have developed an algorithm that exploits this property of transcriptional networks and uses the data processing inequality to infer intricate dependencies within co-regulated gene clusters. Through synthetic data analysis we have shown that this approach is highly effective in distinguishing between direct and indirect interactions, and that it offers significant performance improvements over widely used Bayesian Networks algorithms. By analyzing regulatory relationships in *S. cerevisae*, we demonstrated that ARACNE refines the dependency structure within a large class of co-regulated genes centered around the iron transporter ARN1, and identifies only key "first responders" in this pathway as being directly related to ARN1.

Furthermore, application of ARANCE to the Rosetta dataset hints at possible scale free structure of the yeast transcriptional network. Such result would agree with previous studies which have shown that many real biological networks [24] [25] [26] [27] [28], including even the Yeast transcriptional network , possess the scale free property. However, analysis of the stress response data set shows strong preference towards about 20 connections per node and probably does not agree with the scale free hypothesis. Thus decision regarding the underlying structure of the yeast network would be premature. In fact, one reason for the discrepancy between the data sets could be that the stress response data corresponds to a particular biological phenomenon, while the Rosetta data set is a compendium of various responses. Thus, in the first case, we reconstruct the network that is responsible for a particular phenomenon, while in the second case the reconstructed network is an average. Since the full network is never active simultaneously, it is unclear if the overall scaling behavior is biologically relevant. Further, it is well known that

combination of networks with well defined, but different, scales can produce a scale free structure, and this may be the explanation of the observed discrepancy. We emphasize again that definite conclusions are premature at this stage. The question may be analyzed by using ARACNE to study networks conditional on particular phenomenological characteristics.

1. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14863-8.
2. Tamayo, P., et al., *Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.* Proc Natl Acad Sci U S A, 1999. **96**(6): p. 2907-12.
3. Alizadeh, A.A., et al., *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.* Nature, 2000. **403**(6769): p. 503-11.
4. Ma, S.-K., *Statistical mechanics*. 1985, Singapore: World Scientific.
5. Margolin, A.A., et al., *ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context.* E-print q-bio.MN/0410037, 2004.
6. Basso, K., et al., *Reverse engineering of regulatory networks in human B cells.* Nat Genet, *Submitted*.
7. Nemenman, I., *Information theory, multivariate dependence, and genetic network inference*, in *Tech. Rep. NSF-KITP-04-54, KITP, UCSB*. 2004, arXiv: q-bio/0406015.
8. Pearl, J., *Probabilistic reasoning in intelligent systems: networks of plausible inference*. 1988, San Francisco, CA: Morgan Kaufmann Publishers, Inc.
9. Nemenman, I. and N. Tishby, *An axiomatic approach to the theory of information processing in networks*. *Submitted*.
10. Butte, A.J. and I.S. Kohane, *Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements.* Pac Symp Biocomput, 2000: p. 418-29.
11. Cover, T.M. and J.A. Thomas, *Elements of Information Theory*. 1991, New York: John Wiley & Sons.
12. Mangan, S. and U. Alon, *Structure and function of the feed-forward loop network motif.* Proc Natl Acad Sci U S A, 2003. **100**(21): p. 11980-5.
13. Beirlant, J., et al., *Nonparametric entropy estimation: An overview.* Int. J. Math. Stat. Sci., 1997. **6**(1): p. 17-39.
14. Hartemink, A.J., et al., *Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks.* Pac Symp Biocomput, 2001: p. 422-33.
15. Mendes, P., W. Sha, and K. Ye, *Artificial gene networks for objective comparison of analysis algorithms.* Bioinformatics, 2003. **19 Suppl 2**: p. II122-II129.
16. Barabasi, A.L. and R. Albert, *Emergence of scaling in random networks.* Science, 1999. **286**(5439): p. 509-12.

17. Gasch, A.P., et al., *Genomic expression programs in the response of yeast cells to environmental changes.* Mol Biol Cell, 2000. **11**(12): p. 4241-57.
18. Hughes, T.R., et al., *Functional discovery via a compendium of expression profiles.* Cell, 2000. **102**(1): p. 109-26.
19. Pe'er, D., et al., *Inferring subnetworks from perturbed expression profiles.* Bioinformatics, 2001. **17 Suppl 1**: p. S215-24.
20. Philpott, C.C., et al., *The response to iron deprivation in Saccharomyces cerevisiae: expression of siderophore-based systems of iron uptake.* Biochem Soc Trans, 2002. **30**(4): p. 698-702.
21. Lesuisse, E., et al., *Siderophore uptake and use by the yeast Saccharomyces cerevisiae.* Microbiology, 2001. **147**(Pt 2): p. 289-98.
22. Protchenko, O., et al., *Three cell wall mannoproteins facilitate the uptake of iron in Saccharomyces cerevisiae.* J Biol Chem, 2001. **276**(52): p. 49244-50.
23. Li, F., et al., *The Yeast Cell-Cycle Network Is Robustly Designed.* Proc. Natl. Acad. Sci. USA, 2004. **101**: p. 4781.
24. Jeong, H., et al., *The large-scale organization of metabolic networks.* Nature, 2000. **407**(6804): p. 651-4.
25. Jeong, H., et al., *Lethality and centrality in protein networks.* Nature, 2001. **411**(6833): p. 41-2.
26. Babu, M.M., et al., *Structure and evolution of transcriptional regulatory networks.* Curr Opin Struct Biol, 2004. **14**(3): p. 283-91.
27. Madan Babu, M. and S.A. Teichmann, *Evolution of transcription factors and the gene regulatory network in Escherichia coli.* Nucleic Acids Res, 2003. **31**(4): p. 1234-44.
28. Shen-Orr, S.S., et al., *Network motifs in the transcriptional regulation network of Escherichia coli.* Nat Genet, 2002. **31**(1): p. 64-8.