

# On The Reconstruction of Interaction Networks with Applications to Transcriptional Regulation

Adam A. Margolin<sup>1,2</sup>, Ilya Nemenman<sup>2</sup>, Chris Wiggins<sup>3</sup>, Gustavo Stolovitzky<sup>4</sup>, Andrea Califano<sup>1,2,5</sup>

<sup>1</sup>Department of Biomedical Informatics, Columbia University, 622 West 168<sup>th</sup> St, New York, NY 10032

<sup>2</sup>Joint Centers for Systems Biology, Columbia University, 1150 St Nicholas Ave, New York, NY 10032

<sup>3</sup>Department of Applied Physics and Applied Mathematics, Columbia University, 500 W 120<sup>th</sup> St, New York NY 10027

<sup>4</sup>IBM Computational Biology Center, Functional Genomics Group, IBM T.J. Watson Research Center, P.O. Box 218, Yorktown Heights, N.Y. 10598

<sup>5</sup>Institute of Cancer Genetics, Columbia University, 1150 St Nicholas Ave, New York, NY 10032

Genome-wide clustering of gene expression profiles [1, 2] provides an important first step in studies of transcriptional regulatory networks. However, 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 only short range, pairwise interactions [3]. 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 direct 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. These limitations have relegated the successful application of most methods to relatively simple organisms, such as the yeast *Saccharomyces cerevisiae*. Here we introduce *ARACNE* (Algorithm for the Reconstruction of Accurate Cellular Networks), a novel information-theoretic algorithm for the reverse-engineering of transcriptional regulatory networks from microarray data that overcomes some of these critical limitations. *ARACNE* compares favorably with existing methods and scales successfully to large network sizes. The algorithm is general enough to deal with a variety of other network reconstruction problems.

**Theoretical Background:** We start by noting that with little temporal gene expression data available for higher eukaryotes, one is forced to study steady-state inter-gene statistical dependences only, which we define following the definition of [4], which builds on ideas from the Markov networks literature [5]. Briefly, by analogy with statistical physics, we write the joint probability distribution (JPD) of the stationary expressions of all genes,  $P(\{g_i\})$ ,  $i = 1, \dots, N$ , as:

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

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 [4, 6] for details).

Since the number of realistically obtainable expression profile samples,  $M$ , is rather small, it is infeasible to infer the exponential number of potential  $n$ -way interactions suggested by the expansion in Eq. (1). Rather, a set of simplifying assumptions must be made about the 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.,  $H(\{g_i\}) = \sum_i \phi_i(\{g_i\})$ , 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  $M \rightarrow \infty$  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 only,  $H(\{g_i\}) = \sum_i \phi_i(g_i) + \sum_{ij} \phi_{ij}(g_{ij})$ . 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; the goal is to discriminate such situations.

This formulation is reminiscent of spin glasses on random networks [7], particularly if the  $g_i$  are binary. In this case, the genes are the Ising spins, and truncations to the first, second, or the third order potentials are steps towards the mean field, Bethe, and Kikuchi variational approximations [8-10].

Even focusing on pairwise interactions, the problem of reverse engineering the network is still nontrivial: two genes can have nonzero correlation due to a confounding effect of a third one. That is, we may have  $P(g_i, g_j) \neq P(g_i)P(g_j)$ , but  $\phi_{ij} = 0$ , and there is no direct interaction. Since the number of potential pairwise interactions is huge (quadratic in the number of genes), uncovering such situations and removing false positive edges 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, that is

$$P(\{g_i\}) = \prod_{\langle ij \rangle} P(g_i, g_j) \prod_k P(g_k)^{1-q_k}, \quad (2)$$

where the product over  $\langle ij \rangle$  denotes the product over the nearest neighbors only, and  $q_k$  is the number of such neighbors for the  $k^{\text{th}}$  gene, then we prove that ARACNE can reconstruct it exactly for  $M \rightarrow \infty$ .

The Algorithm: ARACNE relies on a two-step process. *First*, candidate interactions are identified by estimating pairwise gene-gene mutual information (MI)  $I(g_i, g_j) = I_{ij} = \left\langle \log \left[ \frac{P(g_i, g_j)}{P(g_i)P(g_j)} \right] \right\rangle$  and by filtering them using an appropriate threshold,  $I_0$ , computed for a specific p-value,  $p_0$ , in the null-hypothesis of two independent genes. This step is basically equivalent to the Relevance Networks method [11], and, as such, suffers from critical limitations. In particular, 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) [12] -- that has 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  $g_1 \leftrightarrow \dots \leftrightarrow g_2 \leftrightarrow \dots \leftrightarrow g_3$ , and no alternative path exists between  $g_1$  and  $g_3$ ), then

$$I(g_1, g_3) \leq \min \left[ I(g_1, g_2); I(g_2, g_3) \right]. \quad (3)$$

Correspondingly, ARACNE starts with a network graph where each  $I_{ij} > I_0$  is represented by an edge  $(ij)$ . It then examines each gene triplet for which all three MIs are greater than  $I_0$  and removes the edge with the smallest value. Each triplet is analyzed irrespective of whether one of its edges has been marked for removal by a prior DPI application to a different triplet. Thus the network reconstructed by the algorithm is independent of the order in which the triplets are examined.

Theorem: If MIs can be estimated with no errors, then ARACNE (with threshold p value=1, or alternatively, only step 2) reconstructs the underlying interaction network exactly, provided this network is a tree and has only pairwise interactions.

*Proof of the Theorem:* First, notice that for every pair of nodes  $g_i$  and  $g_k$  not connected by a true direct interaction there is, at least, one other node  $g_j$  that separates them on the network tree. Applying the DPI to the  $(ijk)$  triplet leads to removal of the  $(ik)$  edge. Thus only true edges survive. Similarly, every removed edge is not present in the true network. Consider some  $(ijk)$  triplet. One of its genes, say  $g_j$ , may separate the other two. In this case the removed edge  $(ik)$  is clearly not in the true 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 edge the DPI removes is fictitious. Thus all removed edges are indirect, while all remaining edges are factual. The network is reconstructed exactly.  $\square$

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, unless we heuristically decide not to apply DPI pruning under some conditions, *every* loop with only three genes will be opened along the weakest edge). However, if loops are large, then locally the network looks like a tree. Thus, like in the corresponding discussion in statistical physics [9], algorithms designed for trees still 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. Locally tree-like structure is a good approximation for biological networks, which are believed to be sparse; consequently, the average size of a random loop is greater than a few genes, unless some yet unknown evolutionary pressure prefers tighter control loops [a notable exception is the feed forward loop, found to be over-represented in biological circuits [13]].

In the current implementation of the algorithm, we use a computationally efficient Gaussian Kernel estimator [14] to estimate MI. Given two measurement vectors  $\{x_i\}$  and  $\{y_i\}$ ,

$$I(\{x_i\}, \{y_i\}) = \frac{1}{M} \sum_i \log \frac{f(x_i, y_i)}{f(x_i) f(y_i)}, \quad (4)$$

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

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

where  $h = h(M)$  is the kernel width. This estimator is asymptotically unbiased for  $M \rightarrow \infty$ , as long as  $h(M) \rightarrow 0$  and  $[h(M)]^2 M \rightarrow \infty$ . 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  $I_{ij} > I_0$ , 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  $\bar{I}_{ij} \approx \bar{I}_{kl}$ , and the ordering of MI estimates is only weakly dependent on the choice of  $h$  and is stable even when MI itself is uncertain. 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  $O(N^3 + N^2 M^2)$ , where  $M$  is the number of samples, and  $N$  is the number of genes. This is low enough to effectively analyze networks with tens of thousands of genes. We refer the reader to [15] 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 DPI inequalities to the form  $I_{ij} \leq I_{ij}(1 - \tau)$ .

**Performance:** We analyzed ARACNE's performance on reconstructing synthetic networks proposed by [16] specifically as a benchmark for reverse engineering algorithms ([15, 17] also present applications to the Galactose metabolism network in *S. cerevisiae* and the human B cell network). The networks consist of 100 genes and 200 interactions organized in an Erdős-Rényi (random) [18] or a scale-free [19] topology, and they 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 (for the scale-free version), 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 in equilibrium).

ARACNE's performance is compared against Relevance Networks (RNs) [11] and Bayesian Networks (BNs) [5], as implemented by [20]. RNs, which are equivalent to ARACNE without the DPI step, are important to characterize the improvement associated with the introduction of the DPI, while BNs have emerged as some of the best available reverse engineering methods and provide an ideal comparative benchmark. The benchmark measures are *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 ( $N_{TP}$ ,  $N_{FP}$ ,  $N_{TN}$ , and  $N_{FN}$  stand for true/false positives/negatives). Precision vs. Recall curves (PRCs) are a better match than the more familiar ROC curves for problems where  $N_{TN}$  is far greater than  $N_{TP}$ , which is the case in large sparse networks.

PRCs are shown in the Figure below for all three comparative algorithms. We varied the MI threshold and the Dirichlet pseudocount to generate the PRCs for ARACNE/RNs and BNs respectively. ARACNE performs consistently better than BNs and RNs for both types of topologies considered. 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. Such high precision is necessary to guide experimental validation of the method's predictions. Using 1,000 samples, for both

topologies, over half of all edges can be inferred with hardly any false positives. Further, performance degrades gracefully as the sample size decreases and is highly stable with respect to the choice of the kernel width (cf. [15]).

**Summary:** ARACNE appears (a) to achieve very high precision and substantial recall, (b) to be stable with respect to the choice of parameters, and (c) to achieve substantial recall and high precision even with very few data points (125). ARACNE drastically improves network inference due to its efficiency in filtering false-positives, although it may potentially open up some loops of interacting genes and it neglects higher order interactions. In [15] we address these issues and offer suggestions for future investigation.

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. Spellman, P.T., et al., *Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization*. Mol Biol Cell, 1998. **9**(12): p. 3273-97.
3. Ma, S.-K., *Statistical mechanics*. 1985, Singapore: World Scientific.
4. Nemenman, I., *Information theory, multivariate dependence, and genetic network inference*, in *Tech. Rep. NSF-KITP-04-54, KITP, UCSB*. 2004, arXiv: q-bio/0406015.
5. Pearl, J., *Probabilistic reasoning in intelligent systems: networks of plausible inference*. 1988, San Francisco, CA: Morgan Kaufmann Publishers, Inc.
6. Nemenman, I. and N. Tishby, *An axiomatic approach to the theory of information processing in networks*. Submitted.
7. Mezard, M. and G. Parisi, *The Bethe lattice spin glass revisited*. Eur. Phys. J. B, 2001. **20**: p. 217.
8. Kikuchi, R., *A Theory of Cooperative Phenomena*. Phys. Rev., 1951. **81**: p. 988.
9. Yedidia, J., *An idiosyncratic journey beyond mean field theory*, in *Advanced Mean Field Methods: Theory and Practice*, M. Opper and D. Saad, Editors. 2001, MIT Press: Cambridge, MA.
10. Bethe, H., *Statistical Theory of Superlattices*. Proc. Roy. Soc. London A, 1935. **150**: p. 552.
11. 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.
12. Cover, T.M. and J.A. Thomas, *Elements of Information Theory*. 1991, New York: John Wiley & Sons.
13. 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.
14. Beirlant, J., et al., *Nonparametric entropy estimation: An overview*. Int. J. Math. Stat. Sci., 1997. **6**(1): p. 17-39.
15. Margolin, A., et al., *ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context*. Submitted, 2004.
16. Mendes, P., W. Sha, and K. Ye, *Artificial gene networks for objective comparison of analysis algorithms*. Bioinformatics, 2003. **19** Suppl 2: p. II122-II129.
17. Basso, K., et al., *Reverse engineering of regulatory networks in human B cells*. Nat Genet, Submitted.
18. Erdos, P. and A. Renyi, *On Random Graphs*. Publ. Math. Debrecen, 1959. **6**: p. 290-297.
19. Barabasi, A.L. and R. Albert, *Emergence of scaling in random networks*. Science, 1999. **286**(5439): p. 509-12.
20. Friedman, N. and G. Elidan, *LibB, v. 2.1*. 2004.

**Figure.** Precision/Recall for 1,000 samples generated from the Mendes network. (a) Erdős-Rényi topology. (b) Scale-free topology. PRC for (a) is slightly higher since there are fewer loops in random topologies.

