Reconstructing Directed Signed Gene Regulatory Network from Microarray Data

Peng Qiu, *Member, IEEE*, and Sylvia K. Plevritis

Abstract—**Great efforts have been made to develop both algorithms that reconstruct gene regulatory networks and systems that simulate gene networks and expression data, for the purpose of benchmarking network reconstruction algorithms. An interesting observation is that although many simulation systems chose to use Hill kinetics to generate data, none of the reconstruction algorithms were developed based on the Hill kinetics. One possible explanation is that, in Hill kinetics, activation and inhibition interactions take different mathematical forms, which brings additional combinatorial complexity into the reconstruction problem. We propose a new model that qualitatively behaves similar to the Hill kinetics, but has the same mathematical form for both activation and inhibition. We developed an algorithm to reconstruct gene networks based on this new model. Simulation results suggested a novel biological hypothesis that in gene knockout experiments, repressing protein synthesis to a certain extent may lead to better expression data and higher network reconstruction accuracy.**

I. Introduction

Microarray technologies measure the expression levels of thousands of genes simultaneously, from which we may gain insights into the gene regulatory networks that govern various cellular processes. Understanding these networks can help us to reveal the underlying mechanisms. Great efforts have been made in reverse-engineering regulatory interactions from microarray gene expression data. Examples include: Boolean networks [1], information theoretic approaches [2], Bayesian networks [3] and differential equations [4].

In addition, several systems that simulate gene networks and expression data have been developed, for the purpose of benchmarking network reconstruction algorithms [5; 6]. In these simulation systems, gene expression data are simulated by a set of coupled ordinary differential equations, where the regulatory interactions are embedded using Hill kinetics [7], a widely applied model for interacting biochemical reactions.

An interesting observation is that although many simulation systems chose to use the Hill kinetics to generate gene expression data, none of the existing network reconstruction methods were developed based on the Hill kinetics. One possible reason is that the Hill kinetics uses different mathematical forms for activation and inhibition. To reconstruct networks using the Hill kinetics, one needs to not only identify which genes are the regulators of each target gene, but also determine whether a regulator is an activator or an inhibitor. This brings additional combinatorial complexity into the already difficult problem. Such an observation motivated us to propose a new model that behaves similar to the Hill kinetics, but has the same mathematical form for both activation and inhibition. Based on this new model, we developed an algorithm REDSIGN, which REconstructs Directed SIgned Gene Network using steady state microarray gene expression data.

P. Qiu is with the Department of Bioinformatics and Computational Biology, University of Texas MD Anderson Cancer Center, Houston, TX, 77030 USA e-mail: pqiu@mdanderson.org.

S. Plevritis is with the Department of Radiology, Stanford University, Palo Alto, CA, 94305 USA email: plevriti@stanford.edu.

Manuscript received April 1, 2011; revised July 20, 2011.

II. Methods

A. Hill kinetics in existing network simulation systems

In many existing network simulation systems [5; 6], gene expression data are simulated in two steps. First, the topology of a simulated network is either randomly generated or sampled from known biological networks. In the second step, ordinary differential equations are used to simulate the dynamics of the network, where the regulatory interactions are approximated by the Hill kinetics [7].

In the Hill kinetics, if gene j is an activator of a target gene i, the activation function is $G_{ji} = \frac{x_j^n}{x_j^n + K_{ji}^n} + 1$, where x_j is the mRNA concentration of the activator j, n is the Hill-coefficient that controls the sigmoidicity of the activation function, and K_{ji} is the binding affinity. When the activator concentration x_j is 0, G_{ji} equals 1. As x_j increases to infinity, the value of G_{ji} goes to 2. If gene j is an inhibitor of gene i, the repression function is $F_{ji} = \frac{K_{ji}^n}{x_j^n + K_{ji}^n}$. As the inhibitor's concentration x_j increases from 0 to infinity, the value of F_{ji} decreases from 1 to 0. The strength of a regulatory interaction is inversely related to the binding affinity.

Given a network topology and the parameters in the Hill kinetics, gene expression data can be simulated by a group of coupled differential equations [6],

$$\frac{dx_i}{dt} = -D_i x_i + T_i \left[\prod_{j \in I_i} F_{ji}\right] \left[\beta_i + Z_i \left(-1 + \prod_{j \in A_i} G_{ji}\right)\right] \quad (1)$$

where D_i is the degradation rate constant of gene i, T_i is the maximum transcription rate. When both the activators and inhibitors are absent, the basal transcription rate of gene i is $\beta_i T_i$. I_i is the set of gene i's inhibitors, and A_i is the set of activators. Z_i is defined by $Z_i = \frac{1-\beta_i}{2^{|A_i|}-1}$, so that the whole activation term falls into the numerical range $[\beta_i, 1]$.

Equation (1) uses different mathematical forms for activation (G_{ji}) and inhibition (F_{ji}). If one wants to develop a network reconstruction algorithm based on the Hill kinetics in equation (1), the estimated activators and inhibitors need

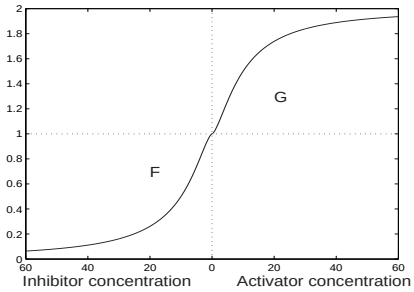


Fig. 1. Example of activation and inhibition in Hill kinetics (K=10, n=1.5).

to be handled separately, which brings in an additional layer of combinatorial complexity to the already difficult problem. This partly explains why none of the existing network reconstruction algorithms were developed based on the Hill kinetics.

B. REDSIGN

In the Hill kinetics, the influence of activators and inhibitors on their targets is modeled by sigmoid functions F_{ji} and G_{ji}. An illustrative example is shown in Figure 1. As mentioned above, reconstructing networks using these two different mathematical forms brings additional combinatorial complexity. Therefore, we consider an alternative model, $\arctan$, which generates similar sigmoid curves, but uses a unified mathematical form for activation and inhibition:

$$\frac{dx_i}{dt} = -D_i x_i + \widetilde{T}_i \left(1 + \frac{2}{\pi} \arctan\left(\sum_j w_{ji} x_j\right)\right) \quad (2)$$

where w_{ji} is positive if gene j activates gene i; w_{ji} is negative if gene j is an inhibitor of gene i; w_{ji} equals 0 if j is not a regulator of i; and the absolute value of w_{ji} controls the sigmoidicity. $\widetilde{T}_i$ is the basal transcription rate when gene i's regulators are absent. The presence of activators and inhibitors can at most double the transcription rate to $2\widetilde{T}_i$, or suppress it to 0. Although not equivalent to equation (1), the proposed model (2) is able to describe similar dynamic behaviors.

Starting from equation (2), we develop a new algorithm to REconstruct Directed and SIgn Gene regulatory Networks (REDSIGN) from steady state microarray gene expression data, i.e. wild type, gene knockout experiments, etc. At steady states, the right hand side of equation (2) equals 0,

$$-D_i x_{ik} + \widetilde{T}_i \left(1 + \frac{2}{\pi} \arctan\left(\sum_j w_{ji} x_{jk}\right)\right) = 0 \quad (3)$$

The subscript k is the index of the k'th microarray experiment. The summation runs over all the genes, where $w_{ji} = 0$ if gene j is not a regulator of gene i. Equation (3) holds for all the microarray experiments where the target gene i is not directly perturbed. From equation (3), it is easy to see that,

$$\sum_j w_{ji} x_{jk} = \tan\left(\frac{\pi D_i}{2\widetilde{T}_i} x_{ik} - \frac{\pi}{2}\right) \quad (4)$$

Denote $\alpha_i = \frac{\pi D_i}{2\widetilde{T}_i}$. Due to the numerical range of $\arctan$, we have $0 \le \alpha_i \le \frac{\pi}{x_{ik}}$. Since equation (3) holds for all the experiments where gene i is not directly perturbed, we have

$$\sum_j w_{ji} x_{jk} = \tan\left(\alpha_i x_{ik} - \frac{\pi}{2}\right) \quad (5)$$

$$0 \le \alpha_i \le \frac{\pi}{\max_k(x_{ik})} \quad (6)$$

where k belongs to the set of experiments where gene i is not directly perturbed.

Given the value of α_i, the regulators of gene i can be estimated by the following minimizing problem,

$$e = \min_{w_{ji}} \sum_k \left(\sum_j w_{ji} x_{jk} - \tan\left(\alpha_i x_{ik} - \frac{\pi}{2}\right)\right)^2 \quad (7)$$

This is essentially a least squares problem. When the total number of genes is large, additional L1- or L2- regularization terms can be added to induce sparsity and improve the robustness of least squares estimation [8]. Since α_i is unknown, we can perform a grid search, estimating the w_{ji}'s and the least squares error for each possible value of α_i. We then pick the α_i value that leads to the minimum least squares error, and use it to estimate the regulators of gene i.

In summary, given steady state expression data, REDSIGN reconstructs a gene network using the following procedure:

1. For each gene i, estimate its regulators using steps 2-4.
2. Create a grid for α_i in its numerical range (6) $\alpha_i = \frac{\pi}{\max_k(x_{ik})} \frac{s}{100}, s = 1, 2, ..., 99$
3. For each value of α_i, solve the least squares problem in (7) and compute the least squares error.
4. Set α_i to be the value that results in the smallest least squares error, and solve equation (7) again. The resulting w_{ji} describes how gene i is regulated by other genes.

The estimated w_{ji} can be arranged in a matrix form $[w_{ji}]$, where the (j, i) element describes the whether gene j regulates gene i, and if yes, whether the regulatory interaction is an activation or an inhibition. The absolute value of w_{ji} indicates the estimated strength of the regulatory interaction.

III. RESULTS

A. Simulation settings and performance metric

Simulations were performed to compare the reconstruction accuracy of REDSIGN with relevance network (RelNet) [2] and Bayesian regression (BayesReg) [3]. Following the simulation settings in [3; 6]: gene networks with 30 nodes were simulated; the in-degree distribution of each node was compact and the out-degree distribution was scale-free; no self-regulation was simulated; the type of each regulatory interaction, activation or inhibition, was randomly chosen with equal probability. Gene expression data were simulated using equation (1). Experimentally derived values for D_i and T_i were obtained from [6]. β_i and n were set to be 0.5 and 1.5, respectively. Multiplicative noise was simulated by log-normal distributions.

The values of binding affinities K_{ji} are not available for most genes in the literature. Although K_{ji}'s are important parameters that describe the strength of regulatory interactions, few existing works studied the effect of these parameters.

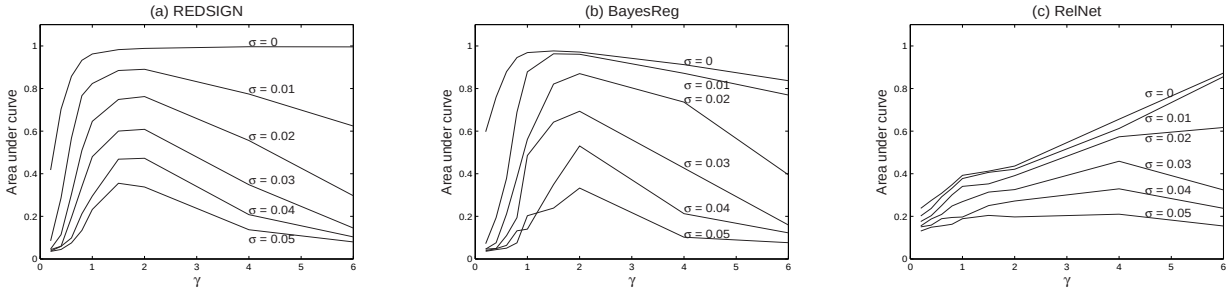


Fig. 2. Network reconstruction accuracy under different noise levels. Experiment noise is multiplicative, which follows log-normal distributions.

Therefore, we simulated data using different values of K_{ji}, to evaluate their impact on the reconstruction accuracy. According to equation (1), when the regulators are absent, the basal steady state mRNA concentration of gene j is $\beta_j T_j / D_j$. In our simulation, we set $K_{ji} = \gamma(\beta_j T_j / D_j)$, and varied γ between 0.2 and 6. When γ equals 1, gene j's binding affinities to its targets are the same as its basal mRNA concentration. In this case, gene j's effect (G_{ji} and F_{ji}) on its targets is approximately two-fold. Since the strength of a regulatory interaction is inversely related to the binding affinity, small γ leads strong regulatory interactions, while large γ implies weak regulatory interactions.

At each level of regulation strength γ and each noise level, 10 networks were simulated, each contained 30 genes and in average 45 regulatory interactions. For each network, equation (1) was used to simulate expression data for wild type and knockout experiments for each gene. A knockout experiment was simulated by holding the knocked-out gene's maximum transcription rate $T_i = 0$. Therefore, 31 microarray experiments were simulated for each network.

The reconstruction accuracy is measured by area under precision-recall curve (AUC). Denote true positives N_{TP}, false positives N_{FP}, and false negatives N_{FN}. The precision and recall are defined as $p = N_{TP}/(N_{TP} + N_{FP})$ and $r = N_{TP}/(N_{TP} + N_{FN})$. For BayesReg and REDSIGN, the reconstructed networks were obtained by thresholding the absolute values of the regression coefficients w_{ji}. If w_{ji} exceeded the threshold, it must satisfy two criteria to be a true positive: gene j regulated gene i; and w_{ji}'s sign agreed with the type of the regulatory interaction. For RelNet, the reconstructed networks were obtained by thresholding the mutual information matrix [9; 10]. Since RelNet does not recover direction or interaction type, if its (j, i) entry exceeded the threshold, this entry was considered as a true positive if genes i and j shared a regulatory interaction. One particular threshold produces a pair of precision and recall values, which corresponds to one point on the precision-recall curve. The precision-recall curves were generated by varying the threshold. The AUCs were computed and used as performance metric. The AUC of a perfect classifier equals 1. If a classifier behaves randomly, its AUC is the ratio between number of true positives and the sum of all true positives and true negatives, which is around $45/(30 * (30 - 1)) \approx 0.05$ in our simulation.

B. Effect of regulation strength on reconstruction accuracy

We applied RelNet, BayesReg and REDSIGN to reconstruct networks from the simulated gene expression data. In Figure 2, we plot the average reconstruction accuracy as a function of regulation strength. The vertical axis is the AUC, and the horizontal axis is inversely related to interaction strength. From the noise-free cases ($\sigma = 0$), we observed that REDSIGN and BayesReg consistently outperformed RelNet. More importantly, both REDSIGN and BayesReg were able to recover the directions and signs of the reconstructed interactions. Although these two methods showed similar performance, REDSIGN ran significantly faster, because REDSIGN is based on linear regression, while BayesReg is an iterative Bayesian algorithm based on the relevance vector machine which has high computational complexity [11].

From the simulations with noise, we observed an expected trend that experiment noise was inversely related to the reconstruction accuracy. At noise level $\sigma = 0.01$, BayesReg outperformed REDSIGN. However, for higher noise levels, the two methods showed similar performance. At noise level $\sigma = 0.05$, when the regulation strength $\gamma = 1 \sim 3$, REDSIGN and BayesReg outperformed the RelNet. When the regulation strength was either quite strong or quite weak, RelNet performed better.

An interesting observation is that, with the presence of noise, as γ increased and the regulation strength decreased, the reconstruction accuracy of both REDSIGN and BayesReg first increased and then decreased. The intuition is that, when γ is small and the regulatory interactions are strong, change of a regulator strongly affects not only its direct target, but also the downstream genes that its targets regulate. Strong regulatory interactions create high correlations between a regulator, its direct and indirect targets. In this case, it is difficult to distinguish indirect and direct targets. When the strength of regulatory interactions are less strong, the effect of a regulator's change does not propagate far, resulting in fewer highly correlated indirect targets, and thus, less potential false positives. On the other extreme, when the regulatory interactions are weak, the correlations between regulators and their targets may be overwhelmed by experiment noise, resulting in poor reconstruction accuracy.

Figure 2 suggests that, it is desirable that $\gamma = 1.5 \sim 2$. Given a biological network, although we can not directly control the binding affinities K_{ji} to tune γ into the desirable range, we can apply general protein synthesis inhibitor, such as Cycloheximide, to repress the synthesis of transcription factor proteins. Since in reality it is the transcription factor proteins of the regulators that modulate the expression of their target genes, this intervention effectively weakens the strength of all edges in the network. We hypothesize that in

isolated nodes	$\sigma = 0.01$	$\sigma = 0.02$	$\sigma = 0.03$	$\sigma = 0.04$	$\sigma = 0.05$
0	0.8955	0.7828	0.6288	0.4373	0.3541
15	0.8252	0.7181	0.5927	0.4332	0.3452
30	0.8097	0.6997	0.5790	0.4486	0.3338

TABLE I
AVERAGE AUC FOR NETWORKS WITH DIFFERENT NUMBER OF ISOLATED NODES AND NOISE LEVEL.

gene knockout experiments, repressing protein synthesis to a certain extent will lead to better data for more accurate network reconstruction.

C. Robustness with respect to isolated nodes

We also evaluated the robustness with respect to isolated nodes. The simulated 30-node networks with $\gamma = 1.5$ from the previous subsection were used. We added 15 or 30 isolated nodes, re-simulated gene expression data with different noise levels σ, and reported the average AUCs of REDSIGN in Table I. The additional isolated nodes reduced the reconstruction performance by less than 10%. Although the additional isolated nodes doubled the size of the network, they did not alter the dynamics and complexity of the data, and thus did not have significant impact on the reconstruction performance.

IV. CONCLUSION

We propose a new computational model for gene regulatory networks, which uses the same mathematical form to approximate the Hill kinetics of activation and inhibition relationships. Based on this new model, we developed the REDSIGN algorithm, which reconstructs gene networks from steady state expression data of wild-type and gene knockout experiments. Simulation results indicated a novel biological hypothesis that in gene knockout experiments, repressing protein synthesis to a certain extent will generate better data, based on which the underlying network can be more accurately reconstructed. This is a hypothesis to be experimentally tested. Another possible extension of this work is to include an additional set of differential equations which explicitly model the dynamics of protein synthesis and degradation. Although including the protein activities will double the size of the problem, such an extension more accurately reflects the biology of transcription regulation, and has the potential of reconstructing networks from time series data.

REFERENCES

[1] I. Shmulevich, E. R. Dougherty, S. Kim, and W. Zhang, "Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks." *Bioinformatics*, vol. 18, no. 2, pp. 261–274, 2002.

[2] A. J. Butte and L. S. Kohane, "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements," in *Pacific Symposium on Biocomputing*, vol. 4, 2000, pp. 418–429.

[3] S. Rogers and M. Girolami, "A bayesian regression approach to the inference of regulatory networks from gene expression data," *Bioinformatics*, vol. 21, no. 14, pp. 3131–3137, 2005.

[4] D. C. Weaver, C. T. Workman, and G. D. Stormo, "Modeling regulatory networks with weight matrices." *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp. 112–123, 1999.

[5] D. Marbach, T. Schaffter, C. Mattiussi, and D. Floreano, "Generating realistic in silico gene networks for performance assessment of reverse engineering methods," *Journal of Computational Biology*, vol. 16, no. 2, pp. 229–239, 2009.

[6] B. C. Haynes and M. R. Brent, "Benchmarking regulatory network reconstruction with grendel," *Bioinformatics*, vol. 25, no. 6, pp. 801–807, 2009.

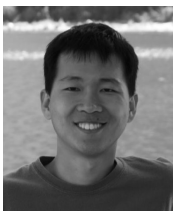
[7] A. V. Hill, "The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves." *J. Physiol.*, vol. 40, pp. iv–vii, 1910.

[8] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society B*, vol. 67, pp. 301–320, 2005.

[9] P. Qiu, A. J. Gentles, and S. K. Plevritis, "Fast calculation of pairwise mutual information for gene regulatory network reconstruction," *Computer methods and programs in biomedicine*, vol. 94, pp. 177–180, 2009.

[10] P. Qiu, A. Gentles, and S. K. Plevritis, "Reducing the computational complexity of information theoretic approaches for reconstructing gene regulatory networks," *Journal of Computational Biology*, vol. 17, no. 2, pp. 169–176, 2010.

[11] M. Tipping, "Sparse bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.



Peng Qiu (M'08) received the B.S. degree from the University of Science and Technology of China, Hefei, China, in 2003, and the Ph.D. degree from the University of Maryland, College Park, in 2007, both in electrical engineering. After spending three years as a postdoctoral fellow in the Integrative Cancer Biology Program at Stanford University, he is currently an assistant professor in the Department of Bioinformatics and Computational Biology, the University of Texas MD Anderson Cancer Center, Houston, TX. His research interests include bioinformatics, signal process and machine learning, and in particular, understanding biological progression using high-dimensional data.

Sylvia K. Plevritis is an Associate Professor in the Department of Radiology. Her research focuses on computational modeling of cancer biology and outcomes. Her work intersects the fields of medical imaging, computational biology, genomics, proteomics and medical technology assessment. Dr. Plevritis holds a Ph.D. in Electrical Engineering (Stanford, 1992) with concentration on MRI spectroscopic imaging of tumors. She also holds an M.S. in Health Services Research (Stanford, 1996), with concentration on the evaluation of cancer screening programs on reducing cancer mortality. Dr. Plevritis is the Director of the Stanford Center for Cancer Systems Biology (CCSB), the co-Director of Information Sciences in Imaging at Stanford (ISIS). Dr. Plevritis is a Principal Investigator of the Stanford Cancer Intervention Surveillance Network (CISNET) which develops mathematical models of cancer progression and evaluates the effectiveness of mammography and MRI in screening for breast cancer and CT in screening for lung cancer.