Effective Enrichment of Gene Expression Data Sets

Utku Sirin, Utku Erdogdu, Faruk Polat
Department of Computer Engineering
Middle East Technical University
06800 Ankara, Turkey
Email: {usirin,utku,polat}@ceng.metu.edu.tr

Mehmet Tan
Department of Computer Engineering
TOBB University of Economics and Technology
Ankara, Turkey
E-mail: mtan@etu.edu.tr

Reda Alhajj
Department of Computer Science
University of Calgary,
Calgary, Alberta, Canada
Email: alhajj@ucalgary.ca

***Abstract*—Ever-growing need for gene-expression data analysis motivates studies in sample generation due to the lack of enough gene-expression data. It is common that there are thousands of genes but only tens or rarely hundreds of samples available. In this paper we attempt to formulate the sample generation task as first building alternative Gene Regulatory Network (GRN) models, second sampling data from each of them and then filtering the generated samples using metrics that measure compatibility, diversity and coverage with respect to original data set. We constructed two alternative GRN models using Probabilistic Boolean Networks and Ordinary Differential Equations. We developed a multi-objective filtering mechanism based on the three metrics to assess the quality of the newly generated data. We presented a number of experiments to show effectiveness and applicability of the proposed multi-model framework.**

***Keywords*-gene expression data; sample generation; multiple perspectives; learning; gene regulation modeling; probabilistic boolean networks; ordinary differential equations;**

I. Introduction

One of the most important problems for gene expression datasets is the size of data. It is normally skewed, i.e., there are thousands of genes per sample but the number of samples is very small. In other words, if we denote the gene expression data as a $n \times s$ matrix where n is the number of genes and s is the number of samples, then $n >> s$.

The practical outcomes of this fact are numerous. The most important one could be articulated as follows; even when the limited number of samples draw an accurate picture of the expression levels, having so few samples would lead to low confidence in the results of any computational method [1], [2], [3], [4]. Actually, several domains, including health informatics and molecular biology suffer from the scarcity of the data to be used in inferring some common characteristics to be used for effective knowledge discovery.

There are both direct and indirect approaches proposed for data enrichment. Indirect approaches such as the ones in [5], [6], [7], [8] try to estimate the required sample size firstly, and then apply experimental techniques to utilize the existing data in case of limited sample size. The study in [9], on the other hand, is a direct approach. They have applied different types of generative models and try to enhance the available data by combining the simulated results from the different generative models. In [9], the sample selection criteria are weak and the mechanism is based on single objective function using linear combination of the metrics that cannot be transformed to each other. Furthermore, in the experimental evaluations, the samples used for model building have also been used in the metric evaluations. In other words, the training and test sets were identical, which makes experimental justification weak.

In this work, thereby, we have enhanced the model building, sample selection and experimental evaluation steps. We have selected two basic *generative* models that exhibit different characteristics. The first model is the Probabilistic Boolean Network (PBN) [10] and the second model is an Ordinary Differential Equations(ODEs) formulation of gene regulation systems [12]. For the sample selection phase, each generated sample is evaluated by three well-defined metrics. These metrics are calculated both using the *training* data as in [9] and using a *test* data which is not used in model building. Final samples are determined by a multi-objective selection mechanism. This mechanism determines the quality of the generated samples separately on all metrics and then rank them in a multi-objective way. In the last step, the highest scoring samples are selected for inclusion in the newly generated dataset.

The rest of this paper is organized as follows. Section II describes the formulation of the two generative models. Section III explains the defined evaluation metrics. Section IV contains the experimental evaluations to justify the effectiveness of our proposed sample generation method. Section V concludes our work and points out possible further

studies.

II. The Employed Generative Models

We constructed two generative models, Probabilistic Boolean Network and Ordinary Differential Equations models of GRNs. The underlying framework can be augmented with other generative models as well, but it is recommended to integrate into the framework only models that are least dependent on the already covered models. This way, the newly integrated models will be more effective towards contributing to a more robust approach.

A. Probabilistic Boolean Network Model

Probabilistic Boolean Networks (PBNs) have been proposed by Shmulevich et al. [10] for specifically modeling gene regulatory networks. They are probabilistic versions of the Boolean Networks introduced by Kauffman [11]. Each node in the PBN is associated with multiple boolean functions, a specific wiring diagram for each function and a probability distribution over the set of boolean functions. The value of each node is calculated by randomly selecting one of the boolean functions associated with it. The variables of the functions are determined by the wiring diagram.

Shmulevich et al. [10] proposed a method for deducing the parameters of PBN. This method uses a coefficient of determination (COD) measure for possible boolean functions [13]. The COD of a boolean function gives us relative decrease in the error when the value of the node is estimated via the boolean function instead of a constant estimator. The COD of boolean function $f_k^{(i)}$ for node X_i can be shown as θ_k^i and it can be formulated as

$$\theta_k^i = \frac{\epsilon_i - \epsilon(X_i, f_k^{(i)}(\boldsymbol{X}_k^{(i)}))}{\epsilon_i} \tag{1}$$

where ϵ_i is the error for constant estimation of X_i, $\boldsymbol{X}_k^{(i)}$ is the vector of nodes that are variables of $f_k^{(i)}$ and $\epsilon(X_i, f_k^{(i)}(\boldsymbol{X}_k^{(i)}))$ is the error for estimating X_i with $f_k^{(i)}$. Among all possible wirings and boolean functions it is possible to choose the best estimators for the given node and assign them probability values as

$$c_k^{(i)} = \frac{\theta_k^i}{\sum_{m=1}^{l_i} \theta_m^i} \tag{2}$$

where $c_k^{(i)}$ is the probability value assigned to $f_k^{(i)}$ and l_i is the number of estimators chosen for X_i.

In order to compute parameters of the PBN, we specify the first $[1, s-1]$ samples as inputs and the last $[2, s]$ samples as outputs of the PBN. Hence, for each sample at time t, the output o of the PBN is the sample at $t+1$. Thereby, we tried to identify the PBN that fits best to the given samples.

We have limited the number of possible functions and the number of variables for each function as 3. This restriction is mandatory, since without any restriction, the number of possible variables is in the order of factorial in terms of the number of genes, and the number of possible functions for each node is in the order of exponential in terms of the number of genes.

Once the parameters of the PBN are computed, the network can be easily used for generating new data by simply running the constructed PBN k times recursively. Also, in order to confirm diversity in the produced data, we use perturbation. While running the network, there is a small probability that the value of a node will change.

B. Ordinary Differential Equations Model

Modeling gene regulatory systems by using ordinary differential equations(ODEs) is one of the oldest and common methods. Among different approaches constructing the ODE model for gene regulation, we have used TSNI algorithm due to its prevailing properties on other regulation modeling algorithms [12]. TSNI is able to cover time series data and able to determine the external perturbations to the system automatically from the available data.

The main task in gene regulation modeling using ODEs is to find the regulation functions for each gene. TSNI algorithm assumes the regulation functions have the form

$$\dot{x_i}(t_k) = \sum_{j=1}^{N} a_{ij} x_j(t_k) + \sum_{l=1}^{P} b_{il} u_l(t_k) \tag{3}$$

where $1 \leq i \leq N, 1 \leq k \leq M$ and N is the number of genes, M is the number of samples in the existing data, P is the number of external perturbations to the system. Here $x_j(t_k)$ is the gene expression level of gene j at time t_k, a_{ij} is the effect of gene j on gene i, b_{il} is the effect of l^{th} external perturbation to gene i and $u_l(t_k)$ is l^{th} external perturbation to the system at time t_k.

If we combine all differential equations in a single matrix equation, we can rewrite equation (3) as

$$\dot{X}(t_k) = A * X(t_k) + B * U(t_k) \tag{4}$$

In Equation (4), the only unknowns are the regulation matrix A and the perturbation matrix B. $\dot{X}(t_k)$ can be approximated as $X(t_{k+1})$ and it is the last $M-1$ data points, $X(t_k)$ is first $M-1$ data points and $U(t_k)$ is the $M-1$ perturbations that are done to the system for each time t_k. To be able to solve the Equation 4, it is required that $M \geq N + P$. Since for most of the datasets this condition does not hold, TSNI applies Principal Component Analysis (PCA) to the Equation 4 which reduces the dimensions of the unknown matrices to manageable sizes. Then, it is fairly easy to solve the Equation 4 and obtain the unknown matrices as in [12].

There are two important points about TSNI algorithm, the number of principal components (PCs) to be considered and

the number of external perturbations to the system. Although both of the parameters can be adjusted by the user, in thus study we set the number of PCs as 2 and the number of external pertubations as 1 as in [12].

After finding the regulation and perturbation matrices of the differential equation model, it is easy to generate new samples from the model by just simulating the system of differential equations numerically.

III. Evaluation Metrics

This section describes the metrics that we defined to determine the most valuable samples from the pool of generated samples. We have defined three metrics, *compatibility*, *diversity* and *coverage* to measure different aspects of the generated data.

Compatibility measures how much the newly generated samples resembles to the original samples. For each newly generated sample, it is the average Euclidean distance to all samples in the original data set.

Diversity measures how much different each newly generated sample from the existing samples. We have calculated the entropy of each sample in the original dataset and sum the differences. This forms a basis for the total information held by the original dataset having M samples. Then, we added the newly generated sample to the original dataset and calculate the total information again for the dataset of $M+1$ samples. By dividing the latter value of total information to the former one, we get a ratio representing the diversity value of each newly generated sample.

Coverage measures how the newly generated samples cover the sample space. For each newly generated sample, it is the average Euclidean distance to all other newly generated samples. If a single sample is created, the value of the coverage metric is set to the maximum of the normalization interval.

For each newly generated sample, compatibility, diversity and coverage values are calculated, forming a vector. We use a **multi-objective selection mechanism** to sort these samples using the notion of (strict) dominance. For the non-dominance case, we select randomly from equally dominating samples.

IV. Experimental Resuts

Experimental evaluation is done by using different real life biological datasets. The first dataset is the gene expression profile of metastatic melanoma cells [14] which is composed of 7 genes and 31 samples [15]. The second dataset is the previously selected set of 25 genes related to cell cycle [16] of *Saccharomyces cerevisiae*. This data is available from Spellman *et al.* [17], consisting of 25 genes and 77 samples in total. The last dataset is siRNA disruptant dataset in human umbilical vein endothelial cells (HUVECs) [18]. It has 379 *Rel/NFkB*-associated genes and 400 samples.

The details of the evaluation metrics are discussed in Section III. As compatibility and coverage metrics are distance based metrics, their results are mapped into $0-100$ interval and diversity results are remained as is. For evaluation semantic, the higher values for coverage is desired and for compatibility and diversity metrics, there is a balance. For higher values of compatibility, higher values of diversity, for lower values of compatibility, lower values of diversity is desired.

Each step of each experiment is repeated 10 times to decrease the effects of the randomization. The reported results are the average values over repetitions.

A. Experiments Based on Whole Data

This section describes the experiments on number of generated samples, where the performance is evaluated based on the training sets. We have used melanoma and yeast datasets. We have run our system 50 times, produced $10, 20, ..., 500$ new samples and plot the results of the metrics.

The compatibility, diversity and coverage values are shown in Figures 1, 2 and 3, respectively.

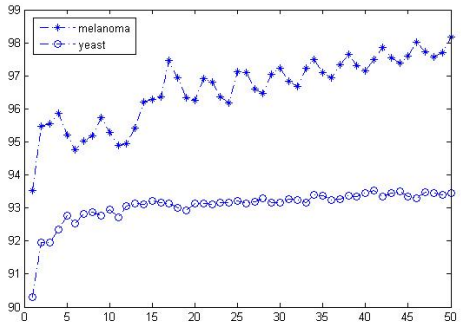

Figure 1. Compatibility values for different number of samples produced

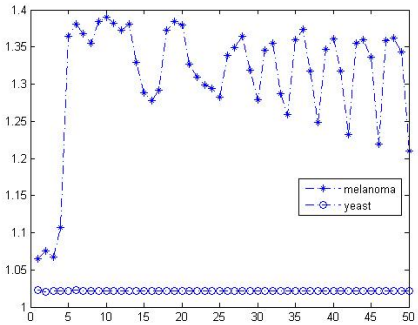

Figure 2. Diversity values for different number of samples produced

For both datasets, the results show that compatibility values increase as the number of generated samples increases. It is mainly because of the fact that the system produces more

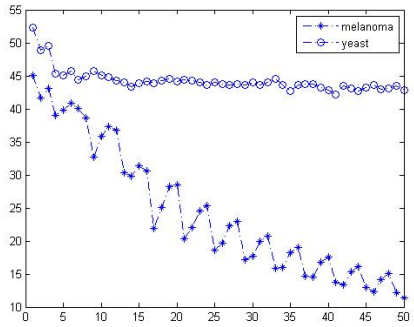

Figure 3. Coverage values for different number of samples produced

similar samples as the number of produced samples increases relative to the original sample set. As we run our system for generating more samples, system produces samples that are converging to the original ones.

The results for diversity values are not similar to the ones in compatibility values. It remains stable as the number of produced samples increases. This shows that the diversity of the produced samples is almost at the same level for each generated sample set. For the melanoma dataset, each new sample holds 30% more information with respect to the original dataset. Note that the diversity value would be 1.0 if the information contents of the original dataset and the set of generated samples are same. This result poses the fact that the data generated by our system is not only very close to the original melanoma dataset but also contains different information than the original dataset, which verifies the high quality of the generated data. For the yeast dataset, on the other hand, the diversity value is much less. However, we can still say the newly generated samples carry new information since the diversity values are always greater than one.

The results for coverage completes the evaluation of the generated datasets. For the data generated from melanoma dataset the coverage values decrease as the number of produced sample increases. This is consistent with compatibility values meaning that the generated data is getting closer to each other. We can observe similar situation for the yeast dataset case but having proportionally greater values. That means, indeed, although generated data from yeast dataset does not hold much new information, its coverage is fairly good so that it can be seen as a successful result.

B. Experiments Based on Separated Data

In the previous section we have evaluated the metrics with respect to the original datasets. Although this comparison gives fair results and represents the quality of the generated data reasonably, its confidentiality is weak since we use the original datasets not only for *training* and but also for *testing* purposes.

In this section, thereby, we have divided the yeast and HUVECs siRNA disruptant datasets into two parts, *training* and *test* sets. For yeast dataset we have used the first 50 sample as training set and last 27 samples as test set. For HUVECs siRNA disruptant dataset we have used the first 300 samples as training set and last 100 samples as test set. We have also calculated metric results relative to the training set as in the previous section to be able to see the difference.

In the Figure 4 and 5, the plots show the compatibility and entropy values based on training and test sets. The system is run for generating 50 samples and the compatibility and diversity values of each sample are plotted. It can be seen that the compatibility values are higher, and the entropy values are lower with respect to the test set. This result justifies the claim about the low confidence level of the metric results based on training set. Comparison with the training set represents the generated data as more closer and less diverse although it is less closer but much diverse in reality.

In the Figure 6 and 7 we have examined the difference between the metric values based on training and test sets wrt number of generated samples. We have generated $10, 20, ..., 500$ many samples from the separated training sets. For each generated sample set we have calculated compatibility and diversity values based on both training and test sets. By subtracting the results of training set from the results of the test set, we obtain the difference values for all generated sample sets.

According to Figure 6 and 7, the compatibility difference is negative since newly generated results are always closer to the training set than the test set. This difference is acceptable until a reasonable value. Because we do not want the generated data to be very close to the original data. On the other hand, the positive results of entropy is because the newly generated data always carries more information with respect to test set than training set. This result justifies the high quality of the data as it always holds new information with respect to the originally available and unseen test data. In the Figure 7, the diversity results of the newly generated data differs from that of Figure 6. It increases as the number of generated samples increases meaning that when we generate more and more samples we are always very close to the originally available test data and always carrying new, even more and more information relative to this available test data. It is a very striking result, in fact. Because it can be stated that, computationally, we are able to generate many new samples just like generating original samples. The complex internal dynamics of the gene regulation can be simulated successfully by superposing different methods and generating data as if it were generated originally by the complex internal dynamics.

V. Conclusions and Future Work

In this work, we attempted to solve an important problem occurring in many different areas such as health informatics

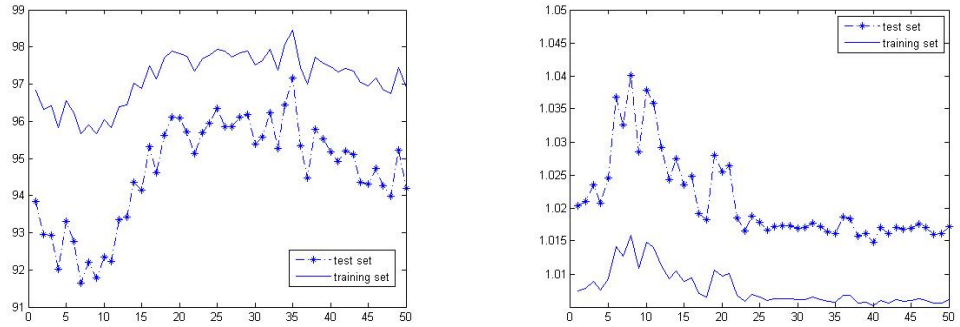

Figure 4. Compatibility and diversity values based on training and test set for each generated sample from Yeast

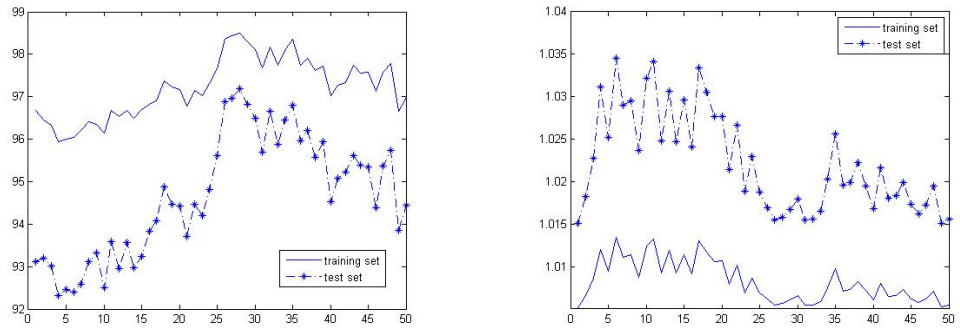

Figure 5. Compatibility and diversity values based on training and test set for each generated sample from HUVECs

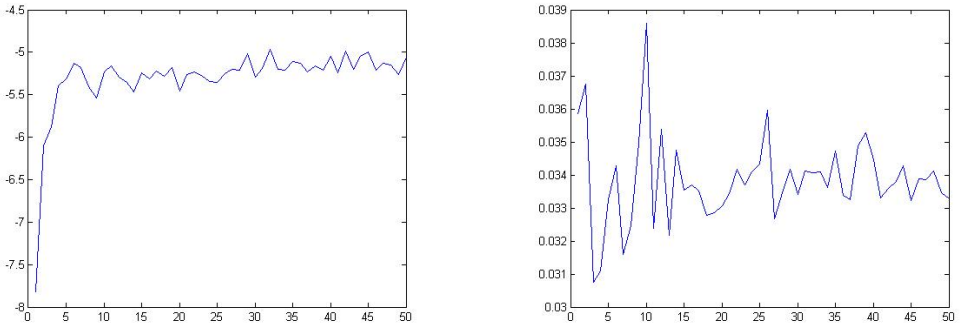

Figure 6. The difference between compatibility and diversity values based on training and test set from Yeast

or molecular biology, the sample size problem.

Experimental results demonstrate that the newly generated samples are so valuable that they can just be treated as originally available data. Moreover, the power of computational methods is verified and the practical result of simulating the very complex gene regulation dynamics is done successfully.

As a future work, the framework may be enhanced by integrating more generative models. It will improve the quality of the produced samples and robustness of the system. Moreover, the produced samples may be studied under a pre-determined analysis task for verifying the effectiveness of our system. Furthermore, determining a bound for the required sample size for generating qualified gene expression data may also be investigated.

REFERENCES

[1] G. Piatetsky-Shapiro, T. Khabaza, and S. Ramaswamy, "Capturing best practice for microarray gene expression data analysis," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003, pp. 407–415.

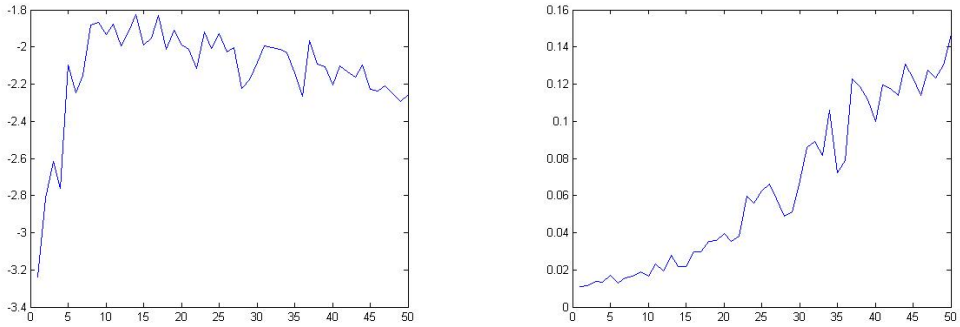

Figure 7. The difference between compatibility and diversity values based on training and test set from HUVECs

[2] W. Chen, H. Lu, M. Wang, and C. Fang, "Gene expression data classification using artificial neural network ensembles based on samples filtering," *Proceedings of the International Conference on Artificial Intelligence and Computational Intelligence*, vol. 1, pp. 626–628, 2009.

[3] H. Franz, C. Ullmann, A. Becker, M. Ryan, S. Bahn, T. Arendt, M. Simon, S. Paabo, and P. Khaitovich, "Systematic analysis of gene expression in human brains before and after death," *Genome Biology*, vol. 6, no. 13, p. R112, 2005.

[4] M. Kim, S. B. Cho, and J. H. Kim, "Mixture-model based estimation of gene expression variance from public database improves identification of differentially expressed genes in small sized microarray data," *Bioinformatics*, vol. 26, no. 4, pp. 486–492, 2010.

[5] M. Lee and G. Whitmore, "Power and sample size for DNA microarray studies," *Statistics in Medicine*, vol. 21, no. 23, pp. 3543–3570, 2002.

[6] W. Pan, J. Lin, and C. Le, "How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach," *Genome Biol*, vol. 3, no. 5, pp. 0022–1, 2002.

[7] M. van Iterson, P. 't Hoen, P. Pedotti, G. Hooiveld, J. den Dunnen, G. van Ommen, J. Boer, and R. Menezes, "Relative power and sample size analysis on gene expression profiling data," *BMC Genomics*, vol. 10, no. 1, p. 439, 2009.

[8] M. Schilling, T. Maiwald, S. Bohl, M. Kollmann, C. Kreutz, J. Timmer, and U. Klingmller, "Quantitative data generation for systems biology: the impact of randomisation, calibrators and normalisers," *IEE Proceedings on Syst Biol (Stevenage)*, vol. 152, no. 4, pp. 193–200, 2005.

[9] U. Erdoğdu, M. Tan, R. Alhajj, F. Polat, D. Demetrick, and J. Rokne, "Employing machine learning techniques for data enrichment: Increasing the number of samples for effective gene expression data analysis," in *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on*, nov. 2011, pp. 238 –242.

[10] 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.

[11] S. A. Kauffman, *The Origins of Order: Self-Organization and Selection in Evolution*, 1st ed. Oxford University Press, USA, June 1993.

[12] M. Bansal, G. D. Gatta, and D. Di Bernardo, "Inference of gene regulatory networks and compound mode of action from time course gene expression profiles," *Bioinformatics*, vol. 22, no. 7, pp. 815–822, Apr. 2006. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btl003

[13] E. Dougherty, S. Kim, and Y. Chen, "Coefficient of determination in nonlinear signal processing," *Signal Processing*, vol. 80, no. 10, pp. 2219–2235, 2000.

[14] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, and et al., "Molecular classification of cutaneous malignant melanoma by gene expression profiling." *Nature*, vol. 406, no. 6795, pp. 536–40, 2000.

[15] A. Datta, A. Choudhary, M. L. Bittner, and E. R. Dougherty, "External control in markovian genetic regulatory networks," *Mach. Learn.*, vol. 52, no. 1-2, pp. 169–191, Jul. 2003. [Online]. Available: http://dx.doi.org/10.1023/A:1023909812213

[16] A. Bernard and A. Hartemink, "Informative structure priors: Joint learning of dynamic regulatory networks from multiple types of data," in *Pacific Symposium on Biocomputing 2005 (PSB05)*, A. R., D. A.K., H. L., J. T., and K. T., Eds. World Scientific: New Jersey, 2005.

[17] P. Spellman, G. Sherlock, M. Zhang, V. Iyer, K. Anders, M. Eisen, P. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle regulated genes of yeast saccharomyces cerevisiae by microarray hybridization." *Mol. Biol. Cell*, vol. 9, pp. 3273–3297, 1998.

[18] D. Hurley, H. Araki, Y. Tamada, B. Dunmore, D. Sanders, S. Humphreys, M. Affara, S. Imoto, K. Yasuda, Y. Tomiyasu, K. Tashiro, C. Savoie, V. Cho, S. Smith, S. Kuhara, S. Miyano, D. S. Charnock-Jones, E. J. Crampin, and C. G. Print, "Gene network inference and visualization tools for biologists: application to new human transcriptome datasets," *Nucleic Acids Research*, 2011.