2010 IEEE International Conference on Data Mining Workshops

An integrative scoring approach to identify transcriptional regulations controlling lung surfactant homeostasis

Minlu Zhang[1,2], Chunsheng Fang[1,2], Yan Xu[2,3], Raj K. Bhatnagar[1], Long J. Lu[1,2]
[1]*Department of Computer Science, University of Cincinnati, Cincinnati, OH, 45221, USA*
[2]*Division of Biomedical Informatics, Cincinnati Children's Research Foundation, Cincinnati, OH, 45229, USA*
[3]*Division of Pulmonary, Cincinnati Children's Research Foundation, Cincinnati, OH, 45229, USA*
E-mail: {zhangml@mail.uc.edu, fangcg@mail.uc.edu, yan.xu@cchmc.org, raj.bhatnagar@uc.edu, long.lu@cchmc.org}

***Abstract*—Transcriptional regulatory network identification is both a fundamental challenge in systems biology and an important practical application of data mining and machine learning. In this study, we propose a semi-supervised learning-based integrative scoring approach to tackle this challenge and predict transcriptional regulations. Our approach out-performs a state-of-the-art label propagation method and reaches AUC scores above 0.96 for three datasets from microarray experiments in the validation. A map of the transcriptional regulatory network controlling lung surfactant homeostasis was constructed. The predicted and prioritized transcriptional regulations were further validated through experimental verifications. Many other predicted novel regulations may serve as candidates for future experimental investigations.**

Keywords- semi-supervised learning; lung surfactant homeostasis; integrative scoring; transcriptional regulation identification; transcriptional regulatory networks

I. INTRODUCTION

Molecular network analysis is an important field of practical applications of data mining and machine learning. A fundamental challenge in the post-genomic era of systems biology is to decode transcriptional regulatory networks in complex organisms [1-2], where transcriptional factor proteins (TFs) control the expression of their target genes (TGs) by either activation or inhibition.

The identification of the transcriptional regulatory network controlling lung surfactant homeostasis is a problem worthy of investigation. Pulmonary surfactant, whose deficiency is associated with premature birth, lung infection or injury, is essential for the functionality of the lung [3]. Surfactant homeostasis involves multiple biological processes, including the synthesis assembly, trafficking, storage, secretion recycling and degradation of surfactant proteins and lipids. Although the structures and functions of pulmonary surfactant proteins have been extensively studied, little is known regarding the transcriptional regulations controlling surfactant lipid homeostasis [4].

Previous works on inferring regulatory networks are mainly based on microarray experiments, including differential equations [5], linear models [6-7], Boolean networks [8], and Bayesian network-based approaches [9-11]. Differential equations and stochastic models simulate dynamics of regulatory systems with detailed descriptions, but the model complexity of differential equations restricts these methods to small-scale and often well-studied regulatory systems. Linear models are robust and scalable for large datasets, but cannot capture possible non-linear relationship. Boolean network is the easiest model for predicting regulatory networks from gene expression data. However, information loss occurs when transforming gene expression levels into Boolean values. Bayesian network-based methods can infer causal relationship between TFs and TGs, but these methods are more computationally complex compared with simpler models.

Semi-supervised learning is well-suited for regulatory relationship inference. Regulatory networks are well-structured with high modularity [12]. TFs tend to form modules to co-regulate downstream TGs, and TGs tend to form modules to be co-regulated [4]. In addition, only a small number of known TF-TG regulations (labeled data) exist due to the expense of carrying out experimental validations, and the rest of the TF-TG regulations (unlabeled data) need to be inferred based on their similarity to known regulations.

In this study, we propose an integrative scoring approach to predict and prioritize transcriptional regulations controlling surfactant homeostasis in the lung. We performed the following data integration: on one hand, the similarity between labeled and unlabeled data is determined based on analytic results from independent and complementary resources, including gene expression profiling, protein interaction, functional annotation, promoter and literature mining. On the other hand, the integrative approach realizes the cluster assumption of semi-supervised learning, i.e., nearby points or points on the same structure should have similar properties [13-15], based on the weighted connectivity between precomputed clusters of TFs and TGs. The advantage of an integrative approach is that each individual type of evidence is often incomplete and error-prone. We compared the performance of the approach on real-life lung TF-TG data with a state-of-the-art semi-supervised label propagation method, MINProp [16], which performs among the best in bipartite link inference. By applying the proposed approach, a map of the lung surfactant homeostasis regulatory network critical for the functionality of the lung is constructed, with focuses on the roles of key TFs in the network.

978-0-7695-4257-7/10 $26.00 © 2010 IEEE
DOI 10.1109/ICDMW.2010.110

II. METHODS

A. Problem Formulation

We consider the problem of predicting new edges in a partially known TF-TG bipartite network with individual edge weights indicating strength of associated evidence. More precisely, using graph theory notations, given a vertices set of TFs $V_t = \{t_1, t_2, \ldots, t_m\}$, a vertices set of TGs $V_g = \{g_1, g_2, \ldots, g_n\}$,, and an edge set of bipartite links connecting TFs and TGs $E_{tg} = V_t \times V_g$, each edge e_{ij} ($0 \le i \le m$, $0 \le j \le n$) between t_i and g_j is associated with an initial weight μ_{ij} ($0 \le \mu_{ij} \le 1$). High initial weights indicate direct or high-confidence evidence of the corresponding TF-TG regulations, while low weights indicate indirect or little known evidence. The goal is to predict and prioritize possible TF-TG links that are not captured by the initial weights.

Such a learning problem can be formulated into a semi-supervised learning framework, where a small number of edges with high initial weights correspond to positively labeled data, and the vast majority of the edges remain unlabeled. Based on the cluster assumption of semi-supervised learning [13-15], edges connecting similar vertices on each side of the bipartite network should have similar confidence, and the labels of unlabeled edges can be predicted based on their similarity to labeled edges. In our problem, TFs sharing expression and functional similarity are likely to form functional modules to co-regulate the same group of TG(s), and TGs sharing expression and functional similarity are likely to be co-regulated by the same TF(s). An unlabeled TF-TG pair tends to be a true regulation if many of the similar TFs to the query TF regulate many of the similar TGs to the query TG. More specifically, an edge e_{ij} connecting a TF t_i and a TG g_j is likely to have a high confidence if a large fraction of TFs similar to t_i are connected to a large fraction of TGs similar to g_j with high weights (Fig. 1).

B. An Integrative Scoring Approach

We propose to solve the TF-TG bipartite edge prediction and prioritization problem by an integrative scoring approach. The goal is to predict a confidence score for each pair of TF-TG bipartite link. The integrative approach consists of three steps:

(1) Given a set of TFs and TGs represented by a matrix between TFs and TGs with each value $Score(t_i, g_j)$ in the matrix denoting the similarity or degree of association between the corresponding TF t_i and TG g_j, cluster TFs and TGs respectively based on the matrix to obtain a set of l TF clusters Ct_p ($0 \le p \le l$) and a set of k TG clusters Cg_q ($0 \le q \le k$).

(2) A *Support* score is defined between each TF cluster Ct_p and each TG cluster Cg_q:

$$Support(Ct_p, Cg_q) = \frac{\sum_{t_i \in Ct_p} \sum_{g_j \in Cg_q} Score(t_i, g_j)}{|Ct_p| \cdot |Cg_q|},$$

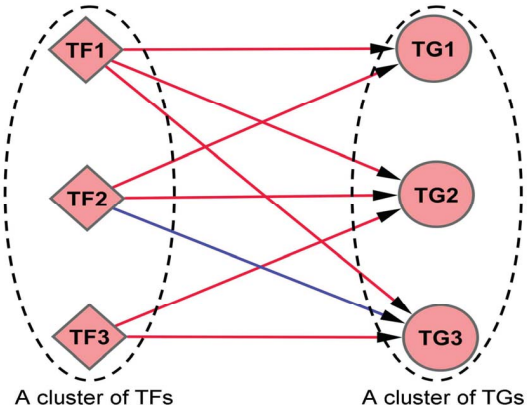

Figure 1. TF-TG regulation prediction. Suppose three TFs similar to each other form a cluster, and three TGs similar to each other form a cluster. Directed edges indicate regulations. Red edges are edges of high initial weights, and blue edges indicate regulations with some evidence. Note that an implicit 0-weighted edge exists between TF3 and TG1, indicating no evidence between the pair. Based on the cluster assumption of semi-supervised learning [13, 15], because the two clusters are densely connected with high-confidence edges, low-weight edges with few known evidence (for example the one between TF2 and TG3) may be a true regulation, and edge with no previous evidence (for example the one between TF3 and TG1) might as well be a true regulation. The figure is drawn by using Cytoscape [26].

where $|Ct_p|$ is the number of TFs in Ct_p, $|Cg_q|$ is the number of TGs in Cg_q, $t_i \in Ct_p$, $i \le |Ct_p|$ and $g_j \in Cg_q$, $j \le |Cg_q|$. A *Support* score can be calculated for each cluster pair Ct_p and Cg_q using the above equation.

(3) An *Evidence* score is then defined between each TF t_i and each TG g_j:

$$Evidence(t_i, g_j) = \frac{\sum_{g_j \in Cg_q} Score(t_i, g_j)}{|Cg_q|} \cdot \frac{\sum_{t_i \in Ct_p} Score(t_i, g_j)}{|Ct_p|},$$

$$\cdot Support(Ct_p, Cg_q) \cdot L(t_i, g_j) \cdot I(t_i)$$

where $L(t_i, g_j)$ is calculated by scaling $Score(t_i, g_j)$into [0.5, 1], $I(t_i)$ is a normalized relative TF importance ranging from 0.8 to 1.2. An *Evidence* score can be calculated for each TF-TG pair.

TFs and TGs are likely to form modules in regulation [4], and the first step of the approach finds all TF clusters and TG clusters that correspond to TF modules and TG modules by standard clustering algorithms such as hierarchical clustering. The values of the initial weight matrix are calculated by combining evidence from multiple well-established data sources, including transcription factor binding sites information, expression similarity, functional similarity, and co-citation from literature (see Data Preparation section). In the second step, the *Support* score describes the weighted connectivity between a cluster of TFs and a cluster of TGs, which borrows the idea from association rule mining [17]. The value of *Support* ranges from 0 to 1, and a higher *Support* score indicates more significant correlation between a pair of TF-TG clusters. Given a threshold of *Support*, for example 0.25, satisfying TF-TG cluster pairs are extracted as correlated cluster pairs.

The example illustrated in Fig. 1 is a TF-TG cluster pair of high *Support*. In the third step, *Evidence* describes the possibility of a true positive TF-TG relationship according to the integrated information. The first factor of *Evidence* denotes the connectivity between a t_i from its cluster Ct_p and all genes in their cluster Cg_q, the second factor measures the connectivity between a g_j from its cluster Cg_q and all TFs in cluster Ct_p, the fourth factor implies the direct evidence normalized from the initial weight between t_i and g_j, and the fifth factor $I(t_i)$ denotes the relative importance of t_i in our analysis. All factors are equally weighted in the scoring function.

C. Comparison Method

A recent paper by Hwang and Kuang proposed a novel algorithm named MINProp that can propagate labels among multiple heterogeneous networks [16]. MINProp models both homogeneous networks and the bipartite heterogeneous cross-domain networks, e.g., disease-gene network, as affinity matrices. Given an initial label for a node in a homogeneous network, the algorithm iteratively propagates the label information through bipartite links among heterogeneous networks as well as edges in homogeneous networks based on calculated affinity matrices. The confidence values of having the label for each node in each network are updated iteratively until convergence. A node is then assigned a label with the highest corresponding confidence value. The convexity of this algorithm guarantees convergence within a low number of steps, and the bipartite associations can be unraveled. MINProp achieves promising results on gene-to-disease phenotype association prediction, which is a bipartite link prediction and prioritization problem similar to the one tackled in this study. Therefore we used MINProp as a comparison to evaluate the performance of both methods.

III. Experiments

A. Data Preparations

We adopted the following procedure to pre-process the data before applying the integrative scoring approach.

1) Microarray data process: 194 microarray samples from 27 independent microarray experiments related to the lung under normal conditions were collected from multiple investigators in Cincinnati Children's Hospital Medical Center. Based on a p-value cutoff of 0.05 by unpaired two-group Student's t-test, 1498 genes were found significantly differentially expressed in response to the gene perturbations in at least five experimental conditions.

2) Clustering on expression data: 29 gene clusters were identified by fuzzy heuristic partition [18], 26 of which have over-represented functions with p-value < 0.01, and 3 of which have enriched function "lipid biosynthesis / metabolism / transport" with "lipid metabolism" as the predominant functional class. The three groups of TFs and TGs form three datasets for the experiments, henceforth referred to as D1, D2, and D3. D1, D2, and D3 contain 42 TFs and 313 TGs, 45 TFs and 54 TGs, 45 TFs and 203 TGs, respectively.

3) Evidence retrieval from various data sources: for each dataset (D1, D2, and D3), evidence for TF-TG regulations was extracted based on four types of data sources. Transcription factor binding sites information, expression similarity, functional similarity, and co-citation from literature were used.

- The existence or over-representation of specific transcription factor binding sites (TFBS) on a TG provides evidence for its regulatory relationship with a TF. Over-represented TFBS in the evolutionarily conserved regions (ECR) were extracted in 3kb upstream as well as proximal promoter regions (1.2kb) genomic sequence. In addition, the frequency of the extracted TFBS was counted. The relative importance of a TFBS was determined by the average ranking order of the ECR, promoter, and frequency analysis, which serve as the fifth factor of the *Evidence* scoring function.
- Based on the assumption that genes with similar functional annotations are likely involved in similar biological processes, the functions of TFs and TGs were extracted from DAVID (http://david.abcc.ncifcrf.gov/; [19]) for each dataset, and kappa similarity was used to calculate the functional similarity.
- Gene expression profiles of TFs and their TGs are usually correlated, and co-regulated TGs by the same TF(s) are also likely co-expressed [20-21]. Therefore, we calculated the expression correlation by Pearson correlation based on the microarray data.
- Limited experimentally verified TF-TG regulations, together with co-citation information between TFs and TGs, were extracted from databases such as Transfac [22], PReMod [23], Genomatix (Eldorado), and Ingenuity knowledge base.

4) Clustering on the integrated matrix: each of the four types of evidence represented by a TF-TG matrix is normalized, combined by summation, and scaled to form an integrated evidence matrix. Each value in the matrix, denoted as $Score(t_i, g_j)$, corresponds to a weight in the bipartite graph between TFs and TGs. Hierarchical clustering with complete linkage was then applied on the integrated matrix to extract TF and TG clusters. For D1, D2, and D3, 16, 18, and 14 TF clusters and 57, 19, 57 TG clusters were identified, respectively.

B. Performance Evaluation and Comparison

To evaluate the predicted TF-TG regulations by the integrative approach, an independent reference set of TF-TG regulations was manually curated from literature. Overall, 31, 24, and 41 retrieved regulations regarding several predicted key TFs in the lung (sterol regulatory element binding transcription factor SREBP, CCAAT/enhancer binding protein CEBP, forkhead box protein HNF3, v-ets erythroblastosis virus E26 oncogene homolog ETS, globin transcription factor GATA, and interferon regulatory factor IRF) respectively were used as the gold-standard positive set. A gold-standard negative set was constructed by extracting TGs with no TFBS information regarding these TFs, yielding

600, 140, and 433 regulations for D1, D2, and D3, respectively.

Using the top 30% predicted TF-TG regulations as positive predictions and the bottom 70% as negative predictions, precision/recall/specificity are summarized in Table I for the integrative scoring method and the comparison label propagation method (Table I). Overall, the excellent performance indicates the effectiveness of our integrative scoring method and supports the assumption that TFs and TGs do form modules in regulations [4]. Among the three reference datasets D1, D2, and D3, the integrative scoring approach yielded 66, 7, and 36 false positive predictions, respectively. This suggests that even though TFs may form modules to regulate TG modules, exceptions may exist. The only false negative prediction was made in D2 between TF HNF3A and TG *Sftpa1*, with a corresponding *Support* score of 0.06, indicating that little evidence exists between the TF cluster of HNF3A and TG cluster of *Sftpa1*. The corresponding receiver operator characteristic (ROC) curves and area under curve (AUC) scores are shown in (Fig. 2). Both methods perform well in terms of prediction, and the integrative scoring method clearly out-performs the label propagation method. The results are consistent when using the top 20% as the cutoff (Table I).

C. *A Map of the Regulatory Network Controlling Lung Surfactant Homeostasis*

Using the top 30% of *Evidence* score as the cutoff, a regulatory network controlling lung surfactant homeostasis can be constructed by combining results from three related datasets. The whole regulatory network contains 98 TFs, 411 TGs, and 6831 TF-TG regulations. Predicted key TF SREBP is a hub in the whole predicted network with a large number of TGs. An SREBP-centered subnetwork is then built by selecting TGs of SREBP that are co-regulated by other upstream TFs of SREBP (Fig. 3). As indicated by the predictions, SREBP together with 16 other TFs, such as CEBP and HNF3, forms modules and co-regulates a total of 76 downstream TGs.

D. *Experimental Validations*

Predicted TF-TG regulations by the integrative scoring approach were verified by experimental validations including promoter reporter assays, transgenic mouse models, and literature confirmation by text mining. Due to the expense and duration of experimental validations, several predicted key TFs including SREBP, CEBP and HNF3 were focused on.

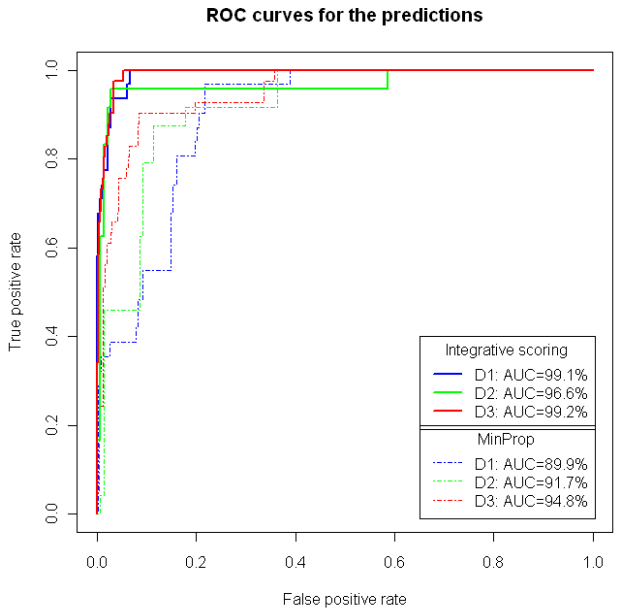

Figure 2. TF ROC curves for the predictions by the integrative scoring method and MINProp [16]. ROC curves show that both methods perform well on the dataset with relatively high AUC scores. Different colors indicate different datasets. The integrative scoring method clearly outperforms MINProp algorithm for the predictions. For the ROC curve by MINProp, the diffusion parameter $\alpha = 0.1$ was used. The performance is largely the same when $\alpha = 0.2$. The figure is drawn by ROCR package of R [27].

Gene promoter reporter assays were carried out for selected TF-TG pairs based on several criteria: (1) the top ranked gene targets of SREBP, CEBP and HNF3 by the integrative scoring method; (2) tissue and cell specificity: TGs should exist in lung epithelial type II cell and have subcellular localization in endoplasmic reticulum or Golgi; (3) functional annotations related to lipid surfactant homeostasis; (4) novelty: TGs should be novel targets of TFs that have not been reported. Based on these criteria, a small set of candidate genes, including *Elovl1*, *Slc34a2,* and

TABLE I. PREDICTION EVALUATION.

Method, Dataset	**The top 30% as cutoff**			**The top 20% as cutoff**		
	Precision	***Recall***	***Specificity***	***Precision***	***Recall***	***Specificity***
Integrative scoring, D1	**0.32**	**1**	**0.89**	**0.448**	**0.968**	**0.938**
MINProp, $\alpha = 0.1$, D1	0.15	0.968	0.883	0.174	0.806	0.802
MINProp, $\alpha = 0.2$, D1	0.149	0.968	0.882	0.179	0.806	0.808
Integrative scoring, D2	**0.767**	**0.958**	**0.95**	**0.846**	**1**	**0.89**
MINProp, $\alpha = 0.1$, D2	0.328	0.917	0.679	0.415	0.917	0.779
MINProp, $\alpha = 0.2$, D2	0.338	0.917	0.693	0.333	0.917	0.757
Integrative scoring, D3	**0.532**	**1**	**0.917**	**0.641**	**1**	**0.947**
MINProp, $\alpha = 0.1$, D3	0.281	0.927	0.776	0.402	0.902	0.873
MINProp, $\alpha = 0.2$, D3	0.279	0.927	0.774	0.402	0.902	0.873

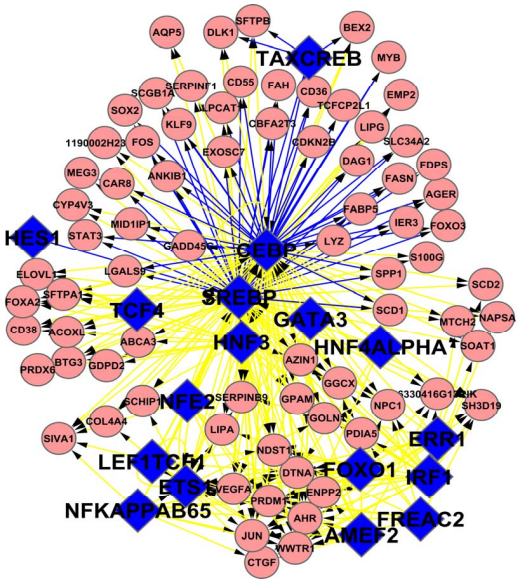

Figure 3. A SREBP-centered subnetwork of the predicted regulatory network. 17 TFs and 76 TGs exist in the SREBP-based subnetwork. Each TG is co-regulated by SREBP and at least one other TF. A salmon round node is a TG, and a blue diamond node is a TF. Arrows indicate regulations. Blue edges indicate regulations predicted in a single dataset. Yellow edges indicate regulations predicted in multiple datasets.

Zdhhc3, are selected. Experimental results show that, consistent with our predictions, CEBPA and SREBP activate *Elovl1* and *Slc34a2* promoters, while *Zdhhc3* is regulated by SREBP but not CEBPA (Table II).

Transgenic mouse models represented by mRNA expression data of differentially expressed genes under the perturbation of specific TFs provide *in vivo* evidence of TF-TG regulations. Three independent sets of microarray data describing differentially expressed genes after selective deletion of *Scap* (SREBP), *Foxa2* (HNF3), and *Cebpa* (CEBPA) from respiratory epithelial cells were used for validation [24]. Genes with a high *Evidence* score (> 0.55) were used as positive prediction, and genes with a low confidence score (< 0.45) were used as negatives. Predicted positive TGs showed significant differential expression in response to the corresponding TF deletion (p-value = 1.0e-8, 3.7e-3, and 1.6e-5 for SREBP, HNF3, and CEBPA, respectively). In addition, within the top 100 predicted TGs, 35, 25, and 21 were significantly repressed to the response of deletion of CEBP, SREBP, and HNF3 in the lung *in vivo*, respectively.

Literature confirmation by text mining provides another approach to validate the predicted TF-TG pairs. Experimentally verified TF-TG regulations were retrieved from the entire PubMed database by the text-mining tool MedScan [25]. All experimentally confirmed SREBP, HNF3, and CEBPA targets were ranked in the top 5%, 10%, and 30% of the predictions, respectively. For CEBPA targets, 86% were also ranked in the top 10% of the predictions. The comparatively weaker performance for CEBPA targets is possibly due to the fact that CEBP (CCAAT/enhancer binding proteins) include multiple TFs as members that bind to CEBP binding sites, and TGs of

TABLE II. TEST RESULTS BY PROMOTER REPORTER ASSAYS.

TF\TG	*Elovl1*	*Slc34a2*	*Zdhhc3*
SREBP	Activation; predicted positive	Activation; predicted positive	Activation; predicted positive
CEBPA	Activation; predicted positive	Activation; predicted positive	No response; predicted negative

CEBPA may not be targets of other members of CEBP family.

Taken together, the consistency from the three independent experimental validations indicates the validity of the predicted results. Other top predicted novel TF-TG regulations may serve as candidate regulations for further experimental verifications.

IV. DISCUSSION

In this paper, we proposed an integrative scoring approach to predict and prioritize TF-TG regulations controlling lung surfactant homeostasis. Based on the cluster assumption, the scoring method takes the clustering information of each side of the bipartite network and initial weighted bipartite edges representing limited knowledge of bipartite links as input and predicts an *Evidence* score indicating the possibility of the corresponding bipartite edge. The experiments demonstrated the high performance of our integrative approach as compared to a well-performed label propagation method. In addition, the complexity of calculating the *Evidence* score is proportional to the bipartite edge number after the clustering as pre-processing. Therefore the scoring approach is scalable for genome-scale datasets. Some of the predicted TF-TG links were shown to be supported by further experimental validations, and many of the rest may serve as validation candidates.

REFERENCES

[1] N. Guelzim, *et al.*, "Topological and causal structure of the yeast transcriptional regulatory network," *Nat Genet,* vol. 31, pp. 60-3, May 2002.

[2] T. I. Lee, *et al.*, "Transcriptional regulatory networks in Saccharomyces cerevisiae," *Science,* vol. 298, pp. 799-804, Oct 25 2002.

[3] J. Johansson and T. Curstedt, "Molecular structures and interactions of pulmonary surfactant components," *Eur J Biochem,* vol. 244, pp. 675-93, Mar 15 1997.

[4] Y. Xu, *et al.*, "A systems approach to mapping transcriptional networks controlling surfactant homeostasis," *BMC Genomics,* vol. In press., 2010.

[5] M. de Hoon, *et al.*, "Inferring gene regulatory networks from time-ordered gene expression data using differential equations," *Discovery Science, Proceedings,* vol. 2534, pp. 267-274, 2002.

[6] E. P. van Someren, *et al.*, "Linear modeling of genetic networks from experimental data," *Proc Int Conf Intell Syst Mol Biol,* vol. 8, pp. 355-66, 2000.

[7] E. P. van Someren, *et al.*, "Genetic network modeling," *Pharmacogenomics,* vol. 3, pp. 507-25, Jul 2002.

[8] P. Li, *et al.*, "Comparison of probabilistic Boolean network and dynamic Bayesian network approaches for inferring gene regulatory networks," *BMC Bioinformatics,* vol. 8 Suppl 7, p. S13, 2007.

[9] Z. Bar-Joseph, "Analyzing time series gene expression data," *Bioinformatics,* vol. 20, pp. 2493-503, Nov 1 2004.

[10] S. Kim*, et al.*, "Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data," *Biosystems,* vol. 75, pp. 57-65, Jul 2004.

[11] S. Imoto*, et al.*, "Combining microarrays and biological knowledge for estimating gene networks via bayesian networks," *J Bioinform Comput Biol,* vol. 2, pp. 77-98, Mar 2004.

[12] M. Zhang*, et al.*, "Molecular network analysis and applications," in *Knowledge-based bioinformatics: from analysis to interpretation*, G. Alterovitz and M. Ramoni, Eds., ed: Wiley, 2010.

[13] M. Belkin and P. Niyogi, "Semi-supervised learning on Riemannian manifolds," *Machine Learning,* vol. 56, pp. 209-239, Jul-Sep 2004.

[14] J. Weston and C. Leslie, "Semi-supervised protein classification using cluster kernels," *Advances in Neural Information Processing Systems 16,* vol. 16, pp. 595-602, 1621, 2004.

[15] D. Y. Zhou*, et al.*, "Learning with local and global consistency," *Advances in Neural Information Processing Systems 16,* vol. 16, pp. 321-328, 1621, 2004.

[16] T. Hwang and R. Kuang, "A heterogenous label propagation algorithm for disease gene discovery," *SIAM International Conference on Data Mining,* p. 12, 2010.

[17] R. Agrawal and J. C. Shafer, "Parallel mining of association rules," *Ieee Transactions on Knowledge and Data Engineering,* vol. 8, pp. 962-969, Dec 1996.

[18] L. Fu and E. Medico, "FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data," *BMC Bioinformatics,* vol. 8, p. 3, 2007.

[19] G. Dennis, Jr.*, et al.*, "DAVID: Database for Annotation, Visualization, and Integrated Discovery," *Genome Biol,* vol. 4, p. P3, 2003.

[20] T. Ideker*, et al.*, "Integrated genomic and proteomic analyses of a systematically perturbed metabolic network," *Science,* vol. 292, pp. 929-34, May 4 2001.

[21] D. J. Allocco*, et al.*, "Quantifying the relationship between co-expression, co-regulation and gene function," *BMC Bioinformatics,* vol. 5, p. 18, Feb 25 2004.

[22] V. Matys*, et al.*, "TRANSFAC: transcriptional regulation, from patterns to profiles," *Nucleic Acids Res,* vol. 31, pp. 374-8, Jan 1 2003.

[23] V. Ferretti*, et al.*, "PReMod: a database of genome-wide mammalian cis-regulatory module predictions," *Nucleic Acids Res,* vol. 35, pp. D122-6, Jan 2007.

[24] V. Besnard*, et al.*, "Deletion of Scap in alveolar type II cells influences lung lipid homeostasis and identifies a compensatory role for pulmonary lipofibroblasts," *J Biol Chem,* vol. 284, pp. 4018-30, Feb 6 2009.

[25] S. Novichkova*, et al.*, "MedScan, a natural language processing engine for MEDLINE abstracts," *Bioinformatics,* vol. 19, pp. 1699-706, Sep 1 2003.

[26] P. Shannon*, et al.*, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Res,* vol. 13, pp. 2498-504, Nov 2003.

[27] T. Sing*, et al.*, "ROCR: visualizing classifier performance in R," *Bioinformatics,* vol. 21, pp. 3940-1, Oct 15 2005.