Computational regulatory network construction from microRNA and transcription factor perspectives

Sungmin Rhee
School of Computer Science and Engineering, Bioinformatics Institute
Seoul National University
Seoul, Korea
lars@snu.ac.kr

Jinwoo Park
School of Computer Science and Engineering, Bioinformatics Institute
Seoul National University
Seoul, Korea
jw.park.bioinfo@gmail.com

Sun Kim
School of Computer Science and Engineering, Bioinformatics Institute
Seoul National University
Seoul, Korea
sunkim.bioinfo@snu.ac.kr

ABSTRACT

As more genomic and epigenomic data becomes available, it has become possible to construct biological networks from the omics data. Among the biological networks, understanding gene regulatory mechanisms is very a important research problem that can reveal condition-specific, e.g., disease-specific, gene regulatory mechanisms. In this paper, we review the current development in the study of constructing gene regulatory networks from microRNA and transcription factor (TF) perspectives. TFs and microRNAs play crucial role in gene regulatory networks since they regulate tens to hundreds of genes, which can be seen naturally as hubs in the network. This review consists of three parts. The first part summarizes recent works on TF regulatory network reconstruction in two sections, one on TF network reconstruction using time series gene expression data and the other on TF network construction by incorporating prior knowledge. The second part is about microRNA network construction in two sections, one on methods based on seed sequence matching and the other on the integrated analysis of gene and microRNA expression data sets. The last part summarizes recent works on the integration of both TF and microRNA with target genes, which is a much more challenging research problem.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Theory

Keywords

MicroRNA, Transcription factor, Target prediction, Regulatory network

1. INTRODUCTION

Genome-wide data such as ChIP-based assay, ChIP-chip and more recently ChIP-seq, data for transcriptional networks or whole genome transcriptome data have been accumulating rapidly. Such genome wide data can be effectively used to construct or infer gene regulatory networks. Gene regulation is a dynamic process in cells, responding to stimuli by various agents such as drugs, bacteria, virus, and harsh weather conditions. Thus understanding the dynamics of the networks is very helpful studying biological mechanisms including diseases.[7] When analyzing the network, one of the most important part of the analysis is to identify the core or hub of the network. TFs and microRNAs (miRNAs) regulate up to several hundreds of genes, so they can be seen naturally as the core of the gene regulatory networks. Thus, in this survey, we classified gene regulatory networks into three categories: TF involved gene regulatory networks, miRNA involved gene regulatory networks and gene regulatory network involving both TF and miRNA simultaneously. There have been many studies for the first and second categories, but study on the last category is just in the beginning stage, probably because of its complexity of performing integrated analysis.

2. TF INVOLVED GENE REGULATORY NETWORKS

There are many computational methods to infer transcriptional regulatory networks. Survey based on the network model architecture is the most popular way [4][7]. However, our survey is based on the types of data used to infer networks so that the survey can be more practical for users. Network construction methods in this survey are classified into two categories: time-series gene expression data based approach and prior knowledge based approach.

2.1 Time-series gene expression data based approach

Time-series gene expression data is necessary to understand biological processes since the biological processes are dynamic and often time dependent [1]. Thus many genome-wide gene expression data are time series data after predesigned stimuli given, for example, drug treatment.

Ernst et al. introduced the DREM algorithm to model dynamic gene regulatory events and it used input-output hid-

den Markov model to integrate time series expression data with static ChIP-chip or motif data [2]. It takes a binary matrix of predictions of TF-gene regulatory interactions and time-series log-ratio gene expression data against the unstressed control as inputs. The algorithm models expression patterns as series of bifurcation and assigns genes to paths according to each gene's expression pattern. Then, to each bifurcation points, DREM assigns TFs that regulate the genes. This method was used to construct networks using yeast response data and recovered many of the known gene interactions. It also predicted unknown interactions that were validated experimentally. DREM 2.0 [14] is the most recent version.

Li et al. introduced DELDBN [12] that integrated ordinary differential equation (ODE) models with the dynamic Bayesian network analysis. Steady-state equation (1) is applied to short sampling time interval data and dynamic state equation (2) is applied to long sampling time interval data.

$$x_i(t+1) = \sum \beta_{ij} X_j(t) \tag{1}$$

$$\frac{x_i(t+1) - x_i(t)}{\Delta t} = \sum \beta_{ij} X_j(t) \tag{2}$$

Where $x_i(t)$, $x_j(t)$ are the expression level of gene i, gene j at time t respectively and β_{ij} is the effect of gene j on gene i. Then DELDBN uses local causality based dynamic Bayesian network analysis to learn through the above two equations. Unlike other Bayesian network analysis, DELDBN uses a low time complexity algorithm and it is scalable to infer large networks. An *in vivo* benchmark data set from yeast was used to demonstrate the performance of the algorithm, showing the highest sensitivity and accuracy in comparison with other approaches. To show the scalability, DELDBN inferred the BRCA1 network using the human Hela cell time series gene expression data that is larger than the typically used yeast data set.

Song et al. introduced KELLER [16], a kernel-reweighted logistic regression method, to infer the latent time-evolving network of gene interactions. With the assumption that time-evolving networks change smoothly, similarity between networks measured in a close time interval is higher than networks measured in a far time interval. Therefore, the problem of estimating dynamic networks can be reduced to inferring a series of static networks by aggregating temporally adjacent networks by reweighting them. To evaluate the algorithm, a microarray gene expression data of *Drosophila melanogaster* with 66-step time series in a full life cycle of 4028 genes was used. The analysis focused 588 genes that were known to be related in the developmental process. KELLER successfully inferred time-evolving network of *Drosophila melanogaster* and showed that many genes had diverse functionalities that were different at each stages.

2.2 Prior knowledge based approach

Although the cost for generating genome-wide data is decreasing rapidly, it is still difficult to obtain enough omics data in one experiment due to the limited budget and time. More importantly, knowledge obtained from many previous studies are valuable. Thus developing computational methods that can incorporate prior knowledge is very important.

Li et al. [11] introduced an network-constrained regularization method that integrated prior knowledge with the form of networks like pathways. Predictors in the model are expression levels of genes with underlying network structures that can be obtained from prior knowledge. It presents a network-constrained penalty which is aggregated form of the lasso penalty and the penalty of Laplace matrix. Network-constrained regularization criterion is defined as follows,

$$L(\lambda_1, \lambda_2, \beta) = (\mathbf{y} - \mathbf{X}\beta)^{\mathbf{T}}(\mathbf{y} - \mathbf{X}\beta) + \lambda_{\mathbf{1}}|\beta_{\mathbf{1}}| + \lambda_{\mathbf{2}}\beta^{\mathbf{T}}\mathbf{L}\beta \tag{3}$$

Where $\mathbf{y}$ is response vector, $\mathbf{X}$ is design matrix, $\lambda_1 > 0$ and $\lambda_2 > 0$ are user defined constants, $\beta_{\mathbf{1}} = \sum_{j=1}^{p} |\beta_j|$ is the L_1-norm which leads sparseness of the result and $\beta^T L\beta$ leads smoothness. With the criterion, estimator $\hat{\beta} = argmin_\beta\ L(\lambda_1, \lambda_2, \beta)$ is obtained. The algorithm was applied to the microarray gene expression data for glioblastoma. With two independent groups of clinical data, one was used for training samples and the other was used for the test set. As a prior knowledge based data, 33 KEGG regulatory pathways were used and the goal was to find disease related subnetworks. The analysis successfully discovered subnetworks that were known to be related with glioblastoma.

Greenfield et al. [3] developed two methods that incorporated prior knowledge for the analysis of time series or static gene expression data to infer dynamic gene regulatory networks. Both methods used the same ordinary differential equation model below.

$$\frac{dx_i}{dt} = -\alpha_i x_i + \sum_p \beta_{i,p} x_p, i = 1, ..., N \tag{4}$$

Where x_i is gene, $\alpha > 0$ is the first order degradation rate, β is a set of parameters to be estimated and P_i is the set of potential regulators for x_i. It is from the assumption that a gene is regulated in proportion to the amount of regulators and the gene itself. Based on this equation, two methods, Modified Elastic Net (MEN) and Bayesian Best Subset Regression (BBSR), are proposed. MEN is a modified form of existing regression application called *Elast-Net* and BBSR is a Bayesian regression based approach with Zellner's g prior. It was shown that the proposed methods by utilizing prior knowledge were tolerant to the errors in the expression data.

3. MIRNA NETWORK INFERENCE

The miRNA network inference problem is to infer a network of miRNA and mRNA of protein coding genes. Genes targeted by miRNAs are down regulated since miRNA interferes with coding genes at the transcription and translation levels. The miRNA network inference problem can be largely divided into two sub-topics. The first one considers only sequence paring information between miRNA and mRNA since miRNAs interfere with mRNAs by hybridization, i.e., sequence pairing. This sequence only prediction

method can be a good way for finding putative targets of miRNAs but they usually have very high false positive rates. The second method incorporates expression profiles of miRNAs and mRNAs for the miRNA network construction. The main idea of the second method is to utilize negative relationship between miRNA expression level and mRNA expression level when a target relationship holds.

3.1 Sequence pairing algorithms for target finding

It is well known that miRNA binds to a reverse complement sequence in the 3′UTR region of mRNA and the corresponding mRNA is degraded. In eukaryotes, a seed region of miRNA at the five prime end site that matches with mRNA is a predominant factor of mRNA repression. This information can be used for algorithms of finding uncovered miRNA target mRNAs.

TargetScan[10] utilizes that many miRNAs and their target sites are conserved across the multiple species (human, mouse, pufferfish). Short sequences of 2-7 nucleotides of miRNA are defined as the seed region and they are matched perfectly to 3′ UTR. After the perfect matching, a thermodynamics based binding score between miRNAs and putative targeted regions of mRNAs is calculated and it is used to rank the targeted genes. The final selection of targets are determined by a pre-selected rank threshold and a binding score threshold.

PicTar[9], like TargetScan, uses the conserved target site information across species and the thermodynamics based binding score for target inferencing. First, PicTar locates all possible target sites using nuclMap. Putative target sites are filtered out if the free energy between the miRNA and the targeted mRNA is higher than a preset cutoff value. When there are multiple binding sites in 3′ UTR region of mRNA, PicTar uses the maximum likelihood score to sort out true target sites. The score is based on the posterior probabilities of the binding sites that are targeted by the miRNAs compared to background of 3′ UTR regions that are not putative miRNA binding sites. This score can sort out competition between different miRNAs on the same regions and can reduce the false positive rate by considering background probability of 3′ UTR region.

Kertesz et al.[8] proposed another thermodynamics-model based approach, PITA, that consideed secondary structure opening energy for finding miRNA target recognition. Like TargetScan, PITA looks for perfect matches in the seed region and calculates binding score between the miRNA and it's putative target mRNA. In addition, PITA considers the site accessibility of target sites. miRNA and their target mRNA can have a secondary structure and should be unpaired so that miRNA can attach to mRNA and then repress the transcription of mRNA. This structure based condition is enforced by calculating the free energy that is required to unpair the secondary structure of target sites and miRNA's secondary structure. RNAFold is used for this calculation. The final miRNA target interaction score is computed as the difference between the binding score and the secondary structure opening score.

3.2 Analysis with expression profiles

Although binding sites of miRNAs and their target regions of mRNAs have similar sequences, the sequence analysis alone cannot solve the high false-positive rate problem since core sequences are very short. miRNAs repress mRNA at both transcription and translational level. For the transcriptional repressing, mRNA transcript expression decreases as expression levels of miRNA that targets the transcript increase. Thus a miRNA:mRNA pair that shows a strong negative correlation in their expression level have a high probability of being a genuine target pair. This negative correlation information is embedded in several computational methods.

MMIA[18] is a method for the integrated analysis of miRNA and RNA expression data. The integrated analysis is performed in two steps. The first step is to identify "differentially" expressed miRNAs by clustering analysis. In the second step, only genes that are targeted by differentially expressed miRNAs are considered. The gene set is further reduced by using sequence based target finding algorithms such as TargetScan, PITA and PicTar, and also by using negative correlation information between miRNA and mRNA expression levels. MMIA divides the miRNA and mRNA's expression data to three clusters: a down-regulated group, an up-regulated group and an unchanged group. It predicts miRNA:mRNA target pairs when the miRNA belongs to a down(up)regulated group and mRNA belongs to a up(down)regulated group. This approach can assure that finding genuine and actual working miRNA:mRNA pairs but also misses many actual miRNA:mRNA pairs whose expression is not significantly up(down) in the cell.

Muniategui et al.[13] proposed a linear model for indicating the degree of miRNA's repressing mRNA transcription. When sequence based algorithms report that K miRNA are predicted to target mRNA j and c_{jk} is an indicator that miRNA k putatively target j-th mRNA, a model as below is used:

$$x_j = \sum_{k=1}^{K} \beta_{jk} c_{jk} z_k + x_j^0 + \epsilon_j$$

where ϵ_j is an error term and x_j^0 is an logarithm of the expression values when no miRNA targets the mRNA. With this model, Lasso regression is used to finding β values minimizing below equation.

$$\min_{\beta_j, x_j^0} \{ \| x_j - \sum_{k=1}^{K} \beta_{jk} c_{jk} z_{jk} - x_j^0 \|_2 + \lambda_j * \sum_{k=1}^{K} |\beta_{jk} c_{jk}| \}$$

Lasso regression with a constraint that β should be non-positive for indicating only down-regulation of miRNA effect and $\lambda_j * \sum_{k=1}^{K} |\beta_{jk} c_{jk}|$ is the penalty term for enforcing the sparsity of solution.

GenMir++[5] uses a linear model for expected mRNA expression values based on the equation below:

$$E[x_{gt} | \{s_{gk}\}, \{z_{kt}\}, \Lambda, \mu_t, \gamma_t] = \mu_t - \gamma_t \sum_k \lambda_k s_{gk} z_{kt}, \lambda_k > 0$$

when x is mRNA expression values, s_{gk} is an indicator that k miRNA targets g mRNA, μ is the background expression of mRNA, γ is the tissue scaling factor. Using this linear model, a Bayesian network model is proposed. In the

Bayesian network model, target transcript expression level x is dependent on a tissue scaling parameter, the miRNA expression level, a regulatory weight, an indicator variable for whether miRNA k truly targets transcript g and this indicator is dependent on an indicator variable for miRNA k putatively targets transcript g. $P(S|X, Z, C, \Theta)$ is estimated using the Bayesian inference and the expectation-maximization technique.

Joung et. al. [6] proposed a module based target finding algorithms using the co-evolutionary machine learning approach and the estimation-of-distribution algorithm (EDA). The detection of modules, (M', T') between miRNA set and mRNA set that best fit in terms of the fitness function (see below) needs to consider all subsets of M' and T' which is computationally infeasible. Thus an evolutionary algorithm was used to find the optimal solution. The fitness function is:

$$F(M', T') = \alpha BS_{M'T'} + \beta EC_{M'} + \gamma EC_{T'} + VOL$$

when $BS_{M'T'}$ is a mean binding score between all pairs of M' and T', $EC_{M'}$ and $EC_{T'}$ are the expression coherence scores of M' and T' each, and VOL is the volume term to prevent finding a solution with one or two miRNA and mRNAs. Given the module fitness function, a co-evolutionary approach is used to find an optimal solution. For miRNA and mRNA, two populations are managed and learned in the context of each other. Individuals are selected based on the fitness function from the two populations and the probability vector is updated. The probability vector denotes the probability of choosing a miRNA or mRNA to a miRNA:mRNA target module. The updated probability vector is used to generate new population.

4. TF-MIRNA INTEGRATED ANALYSES

A gene can be regulated by both miRNA and TF, thus inferencing target relationship should consider TF and miRNA simultaneously.

Shalgi et. al.[15] integrated widely used algorithms for the miRNA target detection and the TF target detection to construct regulatory networks. They analyzed these constructed networks to find local (network motif, hub genes) and global (connectivity distributions) architectures in the network. For the network construction, TargetScan and PicTar were used for miRNA target detection and TRANSFAC was used for TF target finding. They used miRNAs and their target genes that were conserved between 4 species (human, mouse, rat and dog). To reduce TF-gene interaction candidates, only TF promoter regions conserved in orthologous genes from mouse and rat were used. Then they calculated hypergeometric p-value to finds significant miRNA-TF co-occuring pairs. They compared the constructed network with random models, detecting target hubs genes. In constructing randomized network, they preserved the number of genes per miR but shuffled the assigned genes randomly to each miR. The analysis result showed that when miRNA-TF works cooperatively, TF tends to be regulated by miRNA or TF regulates the miRNA forming feed-forward loops.

Sun et. al. [17] identified TF-miRNA regulatory networks consisting of 3-node FFL(feed-forward network) and 4-node FFLs in glioblastoma (GBM). First, GBM related genes and GBM related miRNAs from previous studies were collected. Human TFs were extracted from TRANSFAC. Then TargetScan was used for finding miRNA-TF/gene repression, and MATCH^{TM} was used for TF-gene/miRNA interaction. Co-regulated relationship among genes were predicted using the ARACNE software. The process collected TF-miRNA pairs that cooperatively regulate the same target genes using a cumulative hypergeometric test in a similar fashion to the method used in R. shalgi et. al. [15]. Based on the false discover rate, co-regulating pairs were further filtered out. The proposed method was able to find GBM specific network components.

The techniques that we surveyed so far built computational frameworks by utilizing existing tools as components. Zacher et al[19] developed a joint Bayesian inference approach. Indicator variables are used for MiRNA functional activities (S) and TF functional activities (T). MiRNA functional activities influence miRNA expression and the mRNA expression. TF functional activities also influence mRNA expression in the constructed model. Gene expression levels are approximated by a linear combination of miRNA and TF activities as an equation below:

$$o_{jlc}|S, T, b_j, \omega, v_j^2 \; N(b_j + \sum_{k \in miRNA(j)} s_{kc}\omega_{kj} + \sum_{k \in TF(j)} t_{kc}\omega_{kj}, v_j^2)$$

where o_{jlc} is expression for gene j in l-th replicate of experimental condition c, ω_{kj} are relative influences of miRNA and TF regulators, b_j are reference expression level of mRNA j, S and T are functional activities of miRNA and TF. The limma algorithm and MCMC sampling were used to estimate the parameters w, S, T. This is a comprehensive modeling technique that considered expression profiles of both TF and miRNA. However, the proposed model did not consider the fact that miRNAs can repress TFs or TFs can influence miRNAs.

5. DISCUSSION

We reviewed the recent development in constructing regulatory networks of miRNA and TF. By nature, miRNA and TF are hubs that regulate up to hundreds of genes, thus they are very important for the correct inference of biological networks. We categorized computational techniques in three groups: TF-involved networks, miRNA-involved networks, and TF-miRNA integrated networks. Although there have been many successful studies for constructing networks using omics data, techniques for inferring regulatory networks needs much more efforts. First more accurate methods for component tools need to be developed. Examples include methods for more accurate miRNA target or methods for TF target prediction. Second these component tools or techniques needs to be incorporated into coherent computational models, e.g. a joint Bayesian inference approach by Zacher et al[19].

6. ACKNOWLEDGMENTS

This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (No.NRF-2012M3C4A7033341) and by a grant from the Next-Generation BioGreen 21 Program (No.PJ009037022012), Rural Development Administration, Republic of Korea.

7. REFERENCES

[1] Z. Bar-Joseph, A. Gitter, and I. Simon. Studying and modelling dynamic biological processes using time-series gene expression data. *Nature Reviews Genetics*, 13(8):552–564, August 2012.

[2] J. Ernst, O. Vainas, C. T. Harbison, I. Simon, and Z. Bar-Joseph. Reconstructing dynamic regulatory maps. *Molecular Systems Biology*, 3(74), January 2007.

[3] A. Greenfield, C. Hafemeister, and R. Bonneau. Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. *Bioinformatics*, 29(8):1060–1067, March 2013.

[4] M. Hecker, S. Lambecka, S. Toepferb, E. van Somerenc, and R. Guthkea. Gene regulatory network inference: Data integration in dynamic modelsâĂŤa review. *Biosystems*, 96(1):86–103, April 2009.

[5] J. C. Huang, Q. D. Morris, and B. J. Frey. Bayesian inference of microrna targets from sequence and expression data. *Journal of Computational Biology*, 14(5):550–563, 10 2007.

[6] J.-G. Joung, K.-B. Hwang, J.-W. Nam, S.-J. Kim, and B.-T. Zhang. Discovery of microrna-rna modules via population-based probabilistic learning. *Bioinformatics*, 23(9):1141–1147, March 2007.

[7] G. Karlebach and R. Shamir. Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*, 9(10):770–780, October 2008.

[8] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul, and E. Segal. The role of site accessibility in microrna target recognition. *Nature Genetics*, 37(5):1278–1284, 2007.

[9] A. Krek, D. Grün, M. N. Poy, R. Wolf, L. Rosenberg, E. J. Epstein, P. MacMenamin, I. da Piedade, K. C. Gunsalus, M. Stoffel, and N. Rajewsky. Combinatorial icrorna target predictions. *Nature Genetics*, 37(5):495–500, 05 2005.

[10] B. P. Lewis, I. hung Shih, M. W. Jones-Rhoades, D. P. Bartel, and C. B. Burge. Prediction of mammalian microrna targets. *Cell*, 115(7):787 – 798, 2003.

[11] C. Li and H. Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182, May 2008.

[12] Z. Li, P. Li, A. Krishnan, and J. Liu. Large-scale dynamic gene regulatory network inference combining differential equation models with local dynamic bayesian network analysis. *Bioinformatics*, 27(19):2686–2691, August 2011.

[13] A. Muniategui, R. Nogales-Cadenas, M. Vázquez, X. L.Araguren, X. Agirre, A. Luttun, F. Prosper, A. Pascual-Montano, and A. Rubio. Quantification of mirna-mrna interactions. *PLoS ONE*, 7(2):e30766, Feburary 2012.

[14] M. H. Schulz, W. E. Devanny, A. Gitter, S. Zhong, J. Ernst, and Z. Bar-Joseph. Drem 2.0: Improved reconstruction of dynamic regulatory networks from time-series expression data. *BMC systems biology*, 6(104), August 2012.

[15] R. Shalgi, D. Lieber, M. Oren, and Y. Pilpel. Global and local architecture of the mammalian micrornaâĂŞtranscription factor regulatory network. *PLoS Comput Biol*, 3(7):e131, 07 2007.

[16] L. Song, M. Kolar, and E. P. Xing. Keller: estimating time-varying interactions between genes. *Bioinformatics*, 25(12):i128–i136, January 2009.

[17] J. Sun, X. Gong, B. Purow, and Z. Zhao. Uncovering microrna and transcription factor mediated regulatory networks in glioblastoma. *PLoS Comput Biol*, 8(7):e1002488, 07 2012.

[18] F. Xin, M. Li, C. Balch, M. Thomson, and M. Fan. Computational analysis of microrna proïňĄles and their target genes suggests signiïňĄcant involvement in breast cancer antiestrogen resistance. *Bioinformatics*, 25(4):430–434, December 2008.

[19] B. Zacher, K. Abnaof, S. Gade, E. Younesi, A. Tresch, and H. Fröhlich. Joint bayesian infrence of condition-specific mirna and transcription factor activities from combined gene and microrna expression data. *Bioinformatics*, 28(13):1714–1720, May 2012.