Clustering MicroRNAs from Sequence and Time-Series Expression

Didem Ölçer, Hasan Oğul
Department of Computer Engineering
Baskent University
Ankara,Turkey
e-mail: dtokmak@baskent.edu.tr, hogul@baskent.edu.tr

***Abstract*— Inferring co-operative actions of microRNAs is crucial for analyzing large-scale gene regulatory networks. We introduce here a probabilistic generative model to cluster microRNAs from their mature sequences and time-series expression profiles. Sequence model is defined over the distribution of k-mers, all possible k-length substrings from RNA alphabet. The expression model is built upon a spline-basis function over a Gaussian assumption. Two models are integrated to form a single likelihood. Cluster enrichment analysis has shown that the data integration over a Bayesian framework could improve the clustering ability and produce biologically more plausible patterns.**

Keywords- microRNA expression; microRNA regulation; graphical model; data integration; time-series data analysis

I. INTRODUCTION

Post-transcriptional regulation of genes is mainly directed by small non-coding RNAs called microRNAs (miRNAs). It has been shown that they are abundantly found in many organisms and affiliated with several biological processes such as development, aging and apoptosis [4][5][9][12][17]. It is proven that various diseases are associated with the abnormal behaviors of specific miRNAs [13][14][18]. Recent studies have shown that miRNAs usually operate in a co-operative manner to perform their activities [1]. This suggests that some miRNAs can form context-specific modules, i.e., cluster of entities, while regulating gene expression. Since the elucidation of gene regulatory networks comprising all actors is one of the ultimate goals of systems biology, which miRNAs are functionally similar in a certain context is high-potential knowledge for the researchers and clinicians working in this domain. Here, a functional similarity refers a common regulatory behavior in a certain context, e.g., a specific disease condition or temporal response to a stimuli.

Several features can be employed to infer functional miRNA clusters. An obvious indicator for miRNA's regulatory function is its expression profile. Its differentiation usually results with a consequential change in the expression of its target genes, thus in relevant regulatory pathways. On the other hand, a similarity between the expression profiles of two miRNAs does not necessarily imply a similarity about their genome-wide functions. Several other factors may affect the regulation of miRNAs, and therefore they may arbitrarily express in a similar way. Sequence information can also unveil the structural similarity between miRNAs since the target selection process is usually mediated by a complementarity between mature miRNA sequence and its target mRNA sequence [4]. We can easily argue that two miRNAs having similar sequences will have similar binding preferences, which lead them in target mRNA regulation. However, it was shown that a miRNA may not always be active in a certain context although its binding affinity is very high [6]. Hence, sequence information alone is not expected to give reliable results in miRNA functional similarity associations. In this study, we propose to use both information in a single model to obtain functional miRNA clusters. While designing our model, we were inspired from Kundaje et al. [11] where they combined the promoter sequence motifs with gene expression profiles to infer transcriptional modules. We adopt their model for miRNA expression profiles and propose a novel approach to integrate mature miRNA sequence into overall framework. The framework is built upon a probabilistic graphical model, which simultaneously integrates sequence and time-series expression data to infer coherent miRNA clusters. It enables to adjust and understand the contribution of each information to final cluster assignments. Experimental validation on a real biological data set demonstrates that the integration can improve the clustering ability and produce biologically more plausible patterns.

The rest of paper is organized as follows. We explain methods in Section II. A description of analysis and results can be found in Section III, followed by conclusion in Section IV.

II. METHODS

A. Probabilistic Graphical Model

The problem is to learn the functional clusters of miRNAs where their similarity is explained by a common regulatory mechanism at the transcriptional level and consequential regulatory effect in post-transcriptional level. We define a probabilistic framework which assigns the miRNAs to clusters based on two types of data for each miRNA i: its time-series expression profile, i.e., a set of temporal expression values, E_i, and a set of features representing its mature sequence, S_i. We let the variable Z_i refer to the cluster assignment of miRNA i. Since we assume that both sequence and expression of a miRNA is conditioned on its cluster assignment, following graphical model can be used:

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-260-8

$$S \leftarrow Z \rightarrow E \tag{1}$$

The joint probability distribution for a single miRNA can be written as $P(E_i, S_i, Z_i) = P(E_i|Z_i)P(S_i|Z_i)P(Z_i)$, where E_i and S_i are assumed to be conditionally independent for given cluster assignment Z_i. Having the joint probability model, the task is then to learn the model parameters that maximize the likelihood of input data for a given set of cluster assignments. Since the expression and sequence data have different characteristics in nature, two independent sub-models are provided to define their conditional probabilities.

B. *Sequence Model*

Mature miRNA sequences might be of different lengths, usually between 22-24 nucleotides. For the probabilistic model defined above, the sequence is needed to be modeled by a fixed number of numerical features, which potentially represent its regulatory behavior. We select 3-mer model for this representation. In k-mer model, defined over RNA alphabet *A={'A','G','U','C'}*, a sequence $s_1s_2...s_m$ of miRNA *i* is represented by $S_i = \{n_{i1}, n_{i2}, \dots, n_{iP}\}$ where n_{ij} denotes the count of *j*th *k*-length substring among all possible substrings composed by the same alphabet *A*, and *P* is the number of such substrings. In our case, *3*-mer representation involves $P=4^3$ distinct count values of all possible 3-length substrings from *A*. This scheme is able to consider the content of the miRNA sequence as well as the order of residues inside the sequence, which is one of the major determinants of miRNA binding. Similar representations have been successfully applied in several domains [15].

For model integration, we represent the mature sequence for miRNA i as the sparse vector S_i of count of k-mers that it contains, where S_i is indexed by all possible k-mers: $S_i = \{n_{i1}, n_{i2}, \dots, n_{iP}\}$. We let $n_i = \sum_{p=1\dots P} n_{ip}$ be the total count of k-mers observed in miRNA i. For each cluster j, we define another vector of k-mer frequencies observed in the miRNAs of same cluster; $\theta_j = (\theta_{j1}, \theta_{j2}, \dots, \theta_{jp})$, where $\sum_{p=1\dots P} \theta_{jp} = 1.$ The sub-model for miRNA sequence then becomes a multinomial model, defined by the following conditional probability;

$$P(S_i|Z_i = j, \theta_j) = \frac{n_i!}{n_{i1}!n_{iP}!} \prod\nolimits_{p=1P} \theta_{j_p}^{n_{ip}} \tag{2}$$

With the assumption that observation of each k-mer is independent from each other, the model parameter to be evaluated here is θ_j.

C. *Expression Model*

In the expression model, we define each cluster by a Gaussian distribution over spline parameters that model the common time-course behavior of its member miRNAs. This model was originally proposed by Bar-Joseph et al. [3] and successfully applied for inferring temporal gene regulatory mechanisms [3]. In our framework, each miRNA expression profile is represented by a spline curve. More formally, for each miRNA i assigned to cluster j, its expression profile is given as a function of time as $(f_1(t) \dots f_q(t))(\mu_j + \gamma_{ij})$, where $f_1(t), \dots, f_q(t)$ are spline basis functions. Here, μ_j denotes the mean of coefficients for cluster j, q denotes the number of spline control points used and γ_{ij} is the miRNA specific variation of coefficients, which is treated as a latent variable. γ_{ij} is assumed to be normally distributed with mean 0 and covariance matrix $\mathit{\Gamma}_j$. $\epsilon \sim N(0, \sigma^2)$ is the random Gaussian noise. If we have m time points of observation denoted by $t_1, \dots, t_m$, the expression profile is given as:

$$E_i = \begin{pmatrix} f_1(t_1) & \cdots & f_q(t_1) \\ & \ddots & \\ f_1(t_m) & \cdots & f_q(t_m) \end{pmatrix} \left[\begin{pmatrix} \mu_j^1 \\ \vdots \\ \mu_j^q \end{pmatrix} + \begin{pmatrix} \gamma_{ij}^1 \\ \vdots \\ \gamma_{ij}^q \end{pmatrix} \right] + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_m \end{pmatrix} \tag{3}$$

The probability of miRNA expression conditioned on any cluster assignment then becomes:

$$\begin{aligned} &P\left(E_{i,\gamma_{ij}} \middle| Z_i = j, \mu_j, \mathit{\Gamma}_j, \sigma^2\right) \\ &= (2\pi)^{-(m+q)/_2} \left|\mathit{\Gamma}_j\right|^{-\frac{1}{2}} \sigma^{-m}. e^{-1/2\sigma^2 \left(E_i - f(\mu_j + \gamma_{ij})\right)^t (E_i - f(\mu_j + \gamma_{ij})} \\ &\quad e^{-1/2\gamma_{ij}^t \mathit{\Gamma}_j^{-1} \gamma_{ij}} \end{aligned} \tag{4}$$

As suggested by Bar-Joseph et al. [3], natural cubic splines are used where the size q of spline basis is equal to the number of evenly-spaced knots. Optimal value of q and number of clusters are selected using 2-fold cross-validation to maximize the likelihood computed from the sum of the log loss function at each fold.

Parameter estimation and cluster assignment are done using a set of alternating Expectation and Maximization (EM) steps. Initial clusters are obtained by k-means algorithm using only expression data. At each E-step, the algorithm calculates $p(i|j)$, the probability of gene i being in cluster j, and the expectations for latent variable γ_{ij}. M-step updates the parameters based on the expected values. These alternating iterations repeat until the likelihood convergences. The results of last M-step reports the final miRNA clusters and corresponding parameters inferred.

III. RESULTS

We perform our experiments in a recently released miRNA expression data for transcriptome analysis on ovarian cancer [7]. The data set involves the Affymetrix measurements for expression profiles of several miRNA probes at six time points in three replicates. In the experiment, they studied the pathways and growth properties of cultured human ovarian cancer cells that are expressing luteinizing hormone receptor (LHR). Their particular interest was to understand the changes in the expression as a result of the activation of receptor by its cognate ligand, gonadotropin (LH). They used SKOV3 ovarian cancer cell line stably transfected with LHR, and investigated the response of these cells in culture following exposure to LH. They chosen the parent SKOV-3 ovarian cancer cell line, which did not express LHR, as a control in the experiments and observed the alterations in gene expression elicited by LH. Resulting data set is composed of six groups of SKOV-3 cells: LHR- (parent cell line), LHR+ (just after transfection), and LHR+ incubated with LH in four time points: 1, 4, 8, and 20 h. To pre-process the data, we average over three replicates and calculate differential expression all miRNAs with respect to LRH- condition. For integrative analysis, we remove the miRNAs with no sequence information. This gives us a dataset of differential expression profiles of 80 miRNAs at five different time points.

To assess the functional homogeneity of clusters, we use the enrichment of Gene Ontology (GO) terms [2]. We extract a collection of confirmed miRNA targets from the TarBase [16], miRecords [19], miRTarBase_MTI [10] and circuitDB [8] databases. For each cluster, we build a target set by taking corresponding miRNAs and setting the union of their targets. We then evaluate the functional enrichment of GO terms in each target set based on biological process category, using a Bonferroni-corrected hypergeometric test with an original p- value of 0.01.

Two-fold cross-validation on expression data set suggests us an optimal model with q=4 (number of spline points) and c=5 (number of clusters). We run both single model and integrated model with these parameters. To see the effect of different number of clusters, we also compile the same setup for c=10. At the end of each run, we ignore the outliers by removing the clusters having less than three miRNAs from final cluster set.

GO-enrichment test results are shown in Table 1 for sequence model, expression model and integrative model for c=5 and c=10. For each run, we report the number final clusters obtained and the percentage of clusters with at least one GO-term enriched for targets of more than two miRNAs.

According to Table 1, each single model can achieve a fairly well percentage of GO-enrichments in resulting cluster set. This implies that coherent clusters can be obtained by using either sequence or expression data. On the other hand, the table demonstrates that the integration of two different data types can remarkably increase the number of clusters enriched with significant functional GO-terms. This result obviously suggests that the data integration can improve the clustering ability and helps to obtain biologically meaningful patterns.

TABLE I. COMPARISON OF THE CLUSTERING ABILITY OF MODELS WITH REGARD TO MIRNA TARGET GENE FUNCTIONAL ENRICHMENT

	***c*=5**			***c*=10**		
	Sequence	**Expression**	**Combined**	**Sequence**	**Expression**	**Combined**
Number of clusters	5	4	4	9	8	8
Number of GO-enriched clusters	3	2	4	5	4	6
Percentage of GO-enriched clusters (%)	60	50	100	56	50	75

IV. CONCLUSION

Inferring similar miRNAs can provide valuable information for understanding regulatory mechanisms behind gene expression. Mature miRNA sequence can explain the post-transcriptional regulation of miRNAs, but it cannot give any information about how miRNA itself is regulated. Their expression values can provide some clues about how they are regulated but not about how they regulate since their differential expression might occur due to several random effects. Therefore, the integration of two information sources is essential to discover context-dependent functional miRNA clusters. This study introduces an integrative model to combine two data sources over a probabilistic framework. Two independent models are designed for each type of information, which can also be compiled to obtain only transcriptional (using expression data solely) or only post-transcriptional (using sequence data solely) miRNA groups. The experiments performed on real biological data sets reveals that employing both information can improve the explanatory power of final clusters obtained.

Our study is ongoing to validate our model on larger datasets. Assessing the effects of different parameter selections, such as k in k-mer analysis, will be another future issue.

ACKNOWLEDGMENT

This study was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under the Project 110E160.

REFERENCES

[1] A.V. Antonov, S. Dietmann, P. Wong, D. Lutter, and H.W. Mewes, “GeneSet2miRNA: finding the signature of cooperative miRNA activities in the gene lists,” Nucleic Acids Res, 37, 2009, pp. W323-W328.

[2] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, et al., “Gene Ontology: Tool for the Unification of Biology,” Nat. Genet., 25, 2000, pp.25-29.

[3] Z. Bar-Joseph, G. Gerber, D.K. Gifford, T.S. Jaakkola, and I. Simon, “A New Approach to Analyzing Gene Expression Time Series Data,” Proc. 5th RECOMB Conf., 2002, Canada

[4] D.P. Bartel, “MicroRNAs: genomics, biogenesis, mechanism, and function,” Cell, 116, 2004, pp. 281-297.

[5] D.P. Bartel, “MicroRNAs: target recognition and regulatory functions,” Cell, 136, 2009, pp. 215-233.

[6] C. Cheng and L.M. Li, “Inferring microRNA activities by combining gene expression with microRNA target prediction,” PLoS ONE, 3, 2008, pp. 1-9.

[7] J. Cui, J.B. Eldredge, Y. Xu, and D. Puett, “MicroRNA expression and regulation in human ovarian carcinoma cells by luteinizing hormone,” PLoS One, 67, e21730, 2011.

[8] O. Friard, A. Re, D. Taverna, M. De Bortoli, and D. Cora, “CircuitDB: a database of mixed microRNA/transcription factor feed-forward regulatory circuits in human and mouse,” BMC Bioinf., 11, 435, 2010.

[9] R.C. Friedman, K.K. Farh, C.B Burge, and DP. Bartel, “Most mammalian mRNAs are conserved targets of microRNAs,” Genome Res, 19, 2009, pp. 1-11.

[10] S-D. Hsu, F-M. Lin, W-Y. Wu, C. Liang, W-C. Huang, W-L. Chan, W-T. Tsai, G-Z. Chen, C-J. Lee, C-M. Chiu, and et al., “miRTarBase: a database curates experimentally validated microRNA–target interactions,” Nucleic Acids Res., 39, 2011, pp. D163-D169.

[11] A. Kundaje, M. Middendorf, F. Gao, C. Wiggins and C. Leslie, “Combining Sequence and time series expression data to learn transcriptional modules,” IEEE/ACM Transactions on Computational Biology and Bioi nformatics, 2, 2005, pp. 194-202.

[12] R.C. Lee, R.L. Feinbaum, and V. Ambros, “The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14,” Cell, 75, 1993, pp. 843-854.

[13] J. Lu, G. Getz, E.A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B.L. Ebert, R.H. Mak, A.A. Ferrando and et al. “MicroRNA expression profiles classify human cancers,” Nature, 435, 2005, pp. 834–838.

[14] S.F. Madden, S.B. Carpenter, I.B. Jeffery, H. Bjorkbacka, K.A. Fitzgerald, L.A. O'Neill, and D.G. Higgins, “Detecting microRNA activity from gene expression data,” BMC Bioinf. 11, 257, 2010.

[15] H. Oğul and E. Mumcuoğlu, “A Discriminative Method for Remote Homology Detection Based on n-peptide Compositions with Reduced Amino Acid Alphabets,” Biosystems, 87, 2007, pp. 75-81.

[16] G.L. Papadopoulos, M. Reczko, V.A. Simossis, P. Sethupaty, and A.G. Hatzigeorgion, “The database of experimentally supported targets: A functional update of TarBase,” Nucleic Acids Res., 37, 2009, pp. D155-D158.

[17] X. Peng, Y. Li, K.A. Walters, E.R. Rosenzweig, S.L. Lederer, L.D. Aicher, S. Proll and M.G. Katze, “Computational identification of hepatitis c virus associated microRNA-mRNA regulatory modules in human livers,” BMC Genomics, 10, 373, 2009.

[18] P.M. Voorhoeve, “MicroRNAs: Oncogenes, tumor suppressors or master regulators of cancer heterogeneity,” Bioc Bioph Acta, 1805, 2010, pp. 72-86.

[19] F. Xiao, Z. Zuo, G. Cai, S. Kang, X. Gao, and T. Li, “miRecords: an integrated resource for microRNA-target interactions,” Nucleic Acids Res., 37, 2009, pp. D105-D110.