Identifying functional miRNA-mRNA modules based on hypergraph-based learning

Soo-Jin Kim[1] **Jung-Woo Ha** [2] **Byoung-Tak Zhang** [1, 2]

1. Interdisciplinary Program in Bioinformatics, Seoul National University
2. School of Computer Science and Engineering, Seoul National University
Gwanak-ro 1, Gwanak-gu, Seoul, 151-742, Korea
{*sjkim, jwha, btzhang*}*@bi.snu.ac.kr*

1 Background and problem

Analyzing functional relationships between microRNAs (miRNAs) and mRNAs is an important issue in biological process because it can give new insights into the pathogenesis mechanism of complex diseases including various cancers. Especially, miRNAs emerge recently as one of crucial molecules in post-transcriptional regulation and play critical roles in diverse processes such as tumorigenesis by regulating target mRNAs [1]. Thus, identifying a group of miRNAs and mRNAs as a module is essential for the discovery of their combinatorial effects on different physiological and pathological conditions. To investigate such issues, various approaches have been suggested to detect miRNA-mRNA regulatory relationships in recent years [2-3]. However, the discovery of relevant miRNA and mRNA regulatory modules still remains a major challenge in biology due to the complexity of the issue.

Here we introduce a hypergraph-based model for identifying miRNA-mRNA modules underlying specific regulatory conditions from expression data. The proposed model represents explicit higher-order interactions among many features, thus facilitating the analysis of complex biological phenomena. We evaluate the proposed model on a prostate cancer dataset, and show the discovery of significant miRNA-mRNA regulatory modules involved in specific cancer processes. The biological significance of the identified miRNA-mRNA modules is confirmed by gene ontology analysis and literature reviews. Figure 1 outlines the proposed approach to identifying miRNA-mRNA modules on specific regulatory conditions.

2 Hypergraph-based classifier

A hypergraph is a generalized graph suitable for representing higher-order relationships among heterogeneous features (e.g. miRNAs and mRNAs) by extending an edge to a hyperedge which can connect more than two nodes [4-5]. Formally, a hypergraph H is defined as $H = (V, E)$ where V and E denote a set of vertices, v, and hyperedges, e,

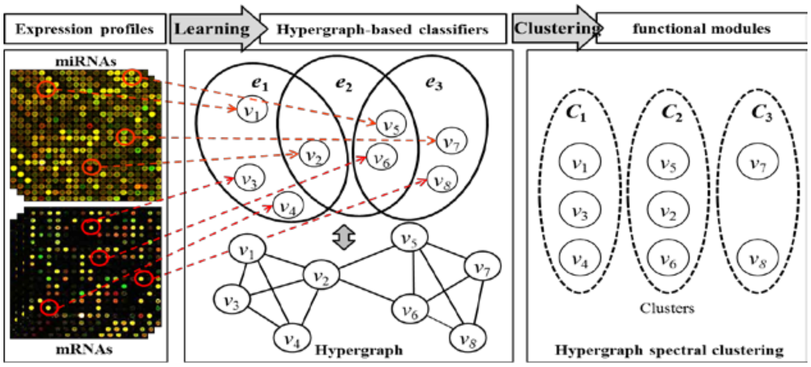

Figure 1. Flow of identifying functional miRNA-mRNA modules using hypergraph-based classifiers

respectively. Also, the degree of a vertex *v*, $d(v)$, and the degree of a hyperedge *e*, $\delta(e)$, are defined to $d(v)=\sum_{e\in E} w(e)h(v,e)$ and $\delta(e) = |e|$, respectively, where $w(e)$ is the weight of *e* and $h(v, e)$ equals 1 if *v* is an element of *e* and 0, otherwise.

In the hypergraph-based classifier, each vertex denotes the expression level of miRNA or mRNA and a hyperedge represents an arbitrary combination or set of vertices like Figure 1. Thus, the hypergraph structure represents the higher-order relational patterns of miRNAs and mRNAs in given expression profiles.

Unlike general hypergraph models [6-7], our model is built by learning distinct patterns of each class (e.g. normal and cancer) from expression profiles without other domain knowledge. Since a hyperedge is a set of miRNA and genes, in addition, each hyperedge can be considered as a building block representing the relationships between miRNAs and mRNAs with high interpretability. Thus, the proposed model is suitable for identifying relevant miRNA-mRNA regulatory modules underlying specific conditions. Furthermore, our model can be applied to the initial discovery problems where domain knowledge is not available or hard to be obtained.

When an expression level data set *D*, $D=\{d^{(n)}\}_{n=1}^{N}=\{(\mathbf{x}^{(n)},y^{(n)})\}_{n=1}^{N}$ is given where $\mathbf{x}^{(\mathbf{n})}$ is a vector of expression levels, formally, learning a hypergraph-based classifier is finding a model that maximizes the classification accuracy, H^*. For classifying the data, we define two indicator functions to determine whether a hyperedge matches a given instance or not:

$$f_t^{+}(d^{(n)},e_i)=\begin{cases}1, & \text{if } \mathbf{x}^{(n)} \text{ matches } e_i \text{ and } y^{(n)}=y_i\\ 0, & \text{otherwise}\end{cases} \quad , \tag{1}$$

$$f_t^{-}(d^{(n)},e_i)=\begin{cases}1, & \text{if } \mathbf{x}^{(n)} \text{ matches } e_i \text{ and } y^{(n)}\neq y_i\\ 0, & \text{otherwise}\end{cases} \quad , \tag{2}$$

where y_i is the label of e_i. We call (1) and (2) positive matching function and negative matching function, respectively. When a set of class label values, $Y = \{y^1,\ldots, y^M\}$, is given, then, unlabeled instance $d^{(n)}$ is classified as $\hat{y}^{(n)}$ by the ensemble of hyperedges in H^*:

1. Make E' consisting of hyperedges matched by $d^{(n)}$ in H^*.
2. Calculate the sum of weights for each label value, c_y, with all hyperedges of E' as follows:

$$c_y=\sum_{i=1}^{|E'|}\left\{w(e_i)\,f_t^{+}(d^{(n)},e_i)\delta(y^{(n)},y)\right\}, \tag{3}$$

   where $w(e_i)$ denotes the weight of the *i*-th hyperedge e_i.

3. Predict $\hat{y}^{(n)}$ as the label of $d^{(n)}$ and evaluate the performance of the model as follows:

$$\hat{y}^{(n)}=\arg\max_{y\in Y} c_y \Big/ \sum_{y\in Y} c_y \ , \tag{4}$$

Because our model represents the huge combinatorial feature space, learning the model may cause heavy computational cost. To overcome the issue, our model conserves the restricted number of building blocks with high weight only in the hypergraph, and we use mutual information in building hypergraphs for efficient learning. In learning our model, one epoch consists of three steps: 1) generating hyperedges, 2) calculating the weight of hyperedges and the objective function of the model, and 3) removing the hyperedges with low weights.

The hyperedges are generated by selecting attributes from a given training instance and combining them to a vertex set. In generating a hyperedge, the probability of selecting the *i*-th attribute is proportional to mutual information between the attribute and class label:

$$\mathrm{P}_{MI}(X_i)=\left\{MI(X_i,Y)\right\}^{\eta} \Big/ \sum_{j=1}^{m}\left\{MI(X_j,Y)\right\}^{\eta} \ , \tag{5}$$

where η is a non-negative constant for controlling the effect of mutual information. Also, the degree of a hyperedge is determined based on the distribution of the degrees of hyperedges in a hypergraph model at the *t*-th epoch, H_{t-1}.

Secondly, the weight of a hyperedge is defined to reflect the ability to discriminate the class label:

$$w(e_i) = \sum_{n=1}^{N} \left(\alpha\, f_t^{+}(d^{(n)}, e_i) - (1-\alpha)\, f_t^{-}(d^{(n)}, e_i) \right) \times \frac{|\,\bar{D}^{y^{(n)}}\,|}{|\,D^{y^{(n)}}\,|} \quad , \tag{6}$$

where D^y is the set of data instances with label y and $\alpha \in (0, 1)$ is a constant parameter to balance larger positive matching or smaller negative matching. The objective function of the hypergraph-based classifier at epoch t, F_t, consists of an accuracy term and a model size term:

$$P(Y \mid \mathbf{x}, H) \approx \sum_{n=1}^{N} \delta(y^{(n)}, \hat{y}^{(n)}) \cong \sum_{n=1}^{N} \sum_{e \in E_t} w(e) \left\{ f_t^{+}(d^{(n)}, e) - f_t^{-}(d^{(n)}, e) \right\} , \tag{7}$$

$$F_t = \sum_{n=1}^{N} \sum_{e \in E_t} w(e) \left\{ f_t^{+}(d^{(n)}, e) - f_t^{-}(d^{(n)}, e) \right\} - \lambda \cdot \frac{|H_0|}{|H_t|} , \tag{8}$$

where λ is a non-negative constant for regularizing memory size, $|H_t| = \sum_{e \in E_t} \delta(e)$. The learning terminates at the epoch when F_t stops increasing or t reaches the maximum epoch, T.

Finally, the structure of hypergraph is learned by eliminating low-weighted hyperedges and regenerating new hyperedges. The number of removed hyperedges, R_t, and the number of newly-generated hyperedges, G_t, controlled as a function of time t:

$$R_t = \frac{R_{max} - R_{min}}{\exp\left(t/\kappa\right)} + R_{min} , \tag{9}$$

$$G_t = \gamma_t \cdot R_t \quad \text{s.t.} \quad \gamma_t = \begin{cases} \left(F_{t-1}/F_t\right)^{\tau} & (\Delta F_t \ge 0) \\ F_{max}/F_t & (\Delta F_t < 0) \end{cases} , \tag{10}$$

where $\Delta F_t = F_t - F_{t-1}$ and τ is a constant for controlling the reduction of the population size.

Although each hyperedge is regarded as a building block, it is reasonable that modules are extracted with considering the overall structure of the hypergraph because the subject of classification is not each hyperedge but the ensemble of all hyperedges in the hypergraph. After learning our model, therefore, we apply hypergraph spectral clustering method [4] to modularize the structure of the learned model for identifying miRNA-mRNA modules. Let D_v and D_e denote diagonal matrices whose diagonal elements are $d(v)$ and $\delta(e)$, respectively. Also, G denotes an incident matrix whose rows and columns are genes and hyperedges, respectively. W denotes a diagonal matrix whose elements are weights of hyperedges. Then, hypergraph Laplacian L is given as follows:

$$L = I - D_v^{-1/2} G W D_e^{-1} G^T D_v^{-1/2} \; . \tag{11}$$

We use eigenvalues and eigenvectors of L to extract the functional miRNA-mRNA modules underlying a specific biological condition.

3 Experimental results and discussion

For experiments, we use the matched miRNAs and mRNAs expression datasets derived from the same patient groups having non-aggressive or aggressive types of prostate cancer [8]. The expression states are converted to binary values including high and low levels because discretized values provide more interpretable representation in analyzing the model. As shown in Table 1, hypergraph-based classifiers provide a competitive performance to SVMs and outperform decision trees and Bayesian networks. Figure 2 shows the evolution of objective function values and model size. Figure 3 depicts the distribution of hyperedge degrees as learning goes on. Despite the reduction of model size, according to Figure 2, the discriminative ability of the model dose not decrease but remains or rather increases. From

Table 1. Classification accuracy comparison of each algorithm

Algorithms	HG-classifiers	SVMs	Decision trees	Bayesian Networks
Avg. (stdev)	**0.921 (±0.014)**	**0.911 (±0.011)**	**0.819 (±0.028)**	**0.841 (±0.013)**

All methods ran 10 times using 10-fold cross-validation and averaged. Accuracy of HG-classifiers is value at maximum F_t. HG-classifiers denotes hypergraph-based classifiers.

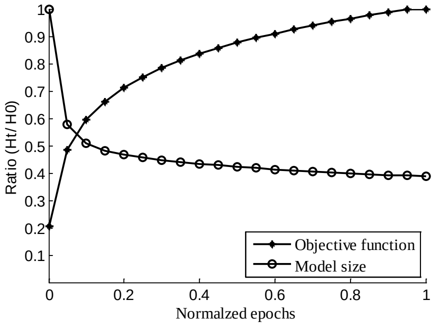

Figure 2. Change of objective function values and model size as learning proceeds.

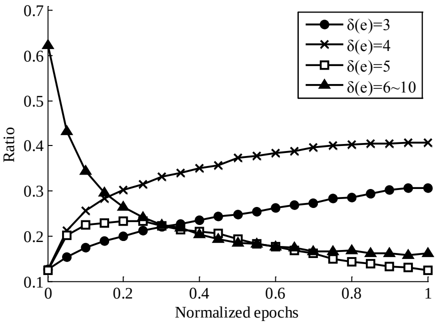

Figure 3. Distribution of hyperedge degrees as learning proceeds.

Table 2. Identified miRNA-mRNA modules and functional analysis by gene ontology (GO) [10]

No.	GO biological process terms	*p*-value	miRNAs	mRNAs
I	aging developmental process protein folding	3.98E-11 1.62E-3 1.94E-3	**hsa-miR-181c**, hsa-miR-214, **hsa-miR-221,** **hsa-miR-222,** hsa-miR-598	CANX CBFA2T3 CDK5R1
II	M phase of mitotic cell cycle response to nutrient cell cycle process cellular ketone metabolic process decidualization maternal placenta development	3.53E-6 5.25E-5 8.77E-5 2.92E-4 7.86E-3 1.00E-3	**hsa-miR-145,** **hsa-miR-331,** hsa-miR-34c, hsa-miR-431, hsa-miR-635, hsa-miR-648	ACSL3, ACTL6A, CARS, CDC2, CDC25A, CDC25C, CDC42BPA CDK6, ELF4, ESPL1 STC2, STK39, USP25

Figure 3 we observe that the ensemble of various higher-order units plays a significant role in determining the type of cancer.

Table 2 presents the identified miRNA-mRNA modules by the proposed model with GO biological process terms. We show two miRNA-mRNA modules with more biological significance in this paper. In particular, *hsa-miR-181c*, *hsa-miR-221* and *hsa-miR-222* of module I are reported in the literature [8] as having a significant association with prostate cancer. Also *hsa-miR-145* and *hsa-miR-331* consisting of module II are closely correlated to prostate cancer processes [9]. Moreover, mRNAs consisting of modules are enriched in biological process terms related to cellular mechanisms associated with the prostate cancer progression. These properties confirm the biological relevance of the identified modules.

Acknowledgment

This work was supported by the National Research Foundation (NRF) grants (2011-0016483-Videome, 2010-0018950-BrainNet), the IT R&D Program of KEIT (10035348-mLife), and the BK21-IT Program.

References

[1] Bartel, D. (2009) “MicroRNAs: target recognition and regulatory functions”, *Cell*, 136 (2): 215-33.

[2] Zhang, S. *et al.* (2011) “A novel computational framework for simultaneous integration of multiple types of genomic data to identify miRNA-gene regulatory modules”, *Bioinformatics*, 27(13):401-9.

[3] Bonnet, E. *et al.* (2010) “Module network inference from a cancer gene expression data set identifies microRNA regulated modules”, *PLoS ONE*, 5(4):e10162.

[4] Zhou, D. *et al.* (2007) “Learning with hypergraphs: clustering, classification, and embedding”, *Advances in Neural Information Processing Systems (NIPS)* 19, 1601-08.

[5] Zhang, B. -T. (2008) “Hypernetworks: a molecular evolutionary architecture for cognitive learning and memory”, *IEEE Computational Intelligence Magazine*, 3(3):49-63.

[6] Klamt, S. *et al.* (2009) “Hypergraphs and cellular networks”, *PLoS Comput. Bio.*, **5**(5):e1000385.

[7] Tian, Z. *et al.* (2009) “A hypergraph-based learning algorithm for classifying gene expression and arrayCGH data with prior knowledge”, *Bioinformatics*, 25 (21): 2831-38.

[8] Wang, L. *et al.* (2009) “Genome-wide transcriptional profiling reveals microRNA correlated genes

and biological processes in human lymphoblastoid cell lines”, *PLoS One*, 4:e5878.

[9] Sevli, S. *et al*. (2010) "The function of microRNAs, small but potent molecules, in human prostate cancer", *Prostate Cancer and Prostatic Diseases*, 13: 208-217.

[10] Zheng, Q and Wang, X. (2008) “GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis”, *Nucleic Acids Research*, 36(2):W358-63.