Bioinformatics Advance Access published October 17, 2010

Identifying functional miRNA-mRNA regulatory modules with correspondence latent dirichlet allocation

Bing Liu [1,2,*], Lin Liu [1], Anna Tsykin [2,3], Gregory J. Goodall [2,4], Jeffrey E. Green [5], Min Zhu [5], Chang Hee Kim [6], and Jiuyong Li [1*]

[1]School of Computer & Information Science, University of South Australia, Mawson Lakes, SA 5095, AU.
[2]Centre for Cancer Biology, SA Pathology, Adelaide, SA 5000, AU.
[3]School of Molecular & Biomedical Science, The University of Adelaide, Adelaide, SA 5005, AU.
[4]Department of Medicine, The University of Adelaide, Adelaide, SA 5005, AU.
[5]Laboratory of Cancer Biology and Genetics, National Cancer Institute, Bethesda, MD 20892, USA.
[6]Laboratory of Molecular Technology, NCI-FCRDC, Frederick, MD 21702, USA.

Associate Editor: Prof. Ivo Hofacker

ABSTRACT

Motivation: MicroRNAs (miRNAs) are small non-coding RNAs, that cause mRNA degradation and translational inhibition. They are important regulators of development and cellular homeostasis through their control of diverse processes. Recently, great efforts have been made to elucidate their regulatory mechanism, but the functions of most miRNAs and their precise regulatory mechanisms remain elusive. With more and more matched expression profiles of miRNAs and mRNAs having been made available, it is of great interest to utilize both expression profiles to discover the functional regulatory networks of miRNAs and their target mRNAs for potential biological processes that they may participate in.

Results: We present a probabilistic graphical model to discover functional miRNA regulatory modules at potential biological levels by integrating heterogeneous data sets, including expression profiles of miRNAs and mRNAs, with or without the prior target binding information. We applied this model to a mouse mammary data set. It effectively captured several biological process specific modules involving miRNAs and their target mRNAs. Furthermore, without using prior target binding information, the identified miRNAs and mRNAs in each module show a large proportion of overlap with predicted miRNA target relationships, suggesting that expression profiles are crucial for both target identification and discovery of regulatory modules.

Contact: Bing.Liu@unisa.edu.au; Jiuyong.Li@unisa.edu.au

1 INTRODUCTION

MicroRNAs (miRNAs) are non-protein-coding RNAs that are expressed from longer transcripts encoded in animals, plants, viruses, and single-celled eukaryotes (Zhao and Srivastava, 2007). They cause mRNA degradation, translational inhibition, or a combination of the two by completely or partially complementary base binding to their target mRNAs (He and Hannon, 2004).

miRNAs are pivotal regulators of development and cellular homeostasis through their control of diverse processes, including cell differentiation, proliferation, growth, mobility, and apoptosis (Du and Zamore, 2007). Consequently, dysregulation of miRNA functions may lead to diseases. Recent studies have reported differentially expressed miRNAs in diverse cancer types such as breast cancer (Iorio *et al.*, 2005), lung cancer (Yanaihara, 2006), prostate cancer (Porkka *et al.*, 2007), colon cancer (Akao *et al.*, 2007), and ovarian cancer (Yang *et al.*, 2008). Therefore, the understanding of miRNA is critical in understanding basic biological processes, elucidating the development and inhibition of pathogenesis of many diseases, and facilitating biotechnology projects.

Many computational approaches have been proposed to elucidate miRNA functions in recent years. We classify these works into three categories: i) miRNA target prediction (Bentwich *et al.*, 2005; Griffiths-Jones *et al.*, 2008; Hatzigeorgiou., 2007; Krek *et al.*, 2005), that is, to identify which mRNAs are targeted by which miRNAs; ii) discovering miRNA regulatory modules (MRMs), that is, to identify a group of co-expressed miRNAs and mRNAs, either at sequence level (Yoon and De Micheli, 2005), or by integrating sequence and expression profiles of miRNAs and mRNAs (Huang *et al.*, 2007; Joung *et al.*, 2007; Tran *et al.*, 2008; Peng *et al.*, 2009); iii) prediction of functional miRNA regulatory modules (FMRMs), which are regulatory networks of miRNAs and their target miRNAs for specific biological processes (Joung and Fei, 2009; Liu *et al.*, 2009a,b).

The identification of FMRMs is critical in understanding the biological pathways and the development and inhibition of pathogenesis of many diseases. It also has a great potential for the development of gene therapeutic treatments and miRNA based drugs (Croce, 2009).

For FMRM discovery, Liu et al. (Liu *et al.*, 2009a) proposed a method based on association rule mining. It associates the reverse expression patterns of miRNAs and mRNAs with biological conditions. A novel method is further proposed (Liu *et al.*, 2009b)

*To whom correspondence should be addressed

using Bayesian Network structure learning. This method was designed to explore all the possible miRNA regulatory patterns for biological conditions under the comparative experiment designs. These two methods are supervised methods where biological conditions are directly applied to the search for the FMRMs. An unsupervised method was proposed by Joung and Fei (2009) for FMRM discovery. It is an innovative application of the author-topic model (Steyvers *et al.*, 2004) in bioinformatics that makes use of the expression profiles of mRNAs and the putative miRNA target information, without considering the expression profiles of miRNAs. Therefore, the regulatory relationships of miRNAs and mRNAs are determined largely based on the miRNA target information which is predicted at the sequence level. Thus, it encounters difficulties in answering the question of how miRNAs regulate their target mRNAs in the identified modules.

In recent years, more and more sample matched expression data having been profiled for multiple classes of conditions or tissues with both miRNAs and mRNAs, providing the opportunity to investigate potential FMRMs systematically for various biological processes by integrating the available data. In addition, some researchers have suggested that algorithms that do not consider known targets may avoid biases (Lewis *et al.*, 2003, 2005; Bartel, 2009). Hence, it is of great interest to discover FMRMs with expression profiles of miRNAs and mRNAs without using target binding information. Therefore, in this paper, we propose a FMRM discovery method that integrates heterogeneous data sets, including expression profiles of both miRNAs and mRNAs, with or without using the prior target binding information.

Our method is inspired by the Correspondence Latent Dirichlet Allocation (Corr-LDA) (Blei and Jordan, 2003), a probabilistic graphical model that has been successfully applied to automatic image annotation with caption words. Given observations of image and caption words, Corr-LDA captures the correspondence between them by modeling topics described by both images and caption words with latent variables. In our question, FMRMs are dependent groups of miRNAs and mRNAs linked to latent functions. Our aim is to capture the correspondence between miRNAs and mRNAs, assuming that they participate in the same latent functions. Therefore, we apply the idea of annotating images with caption words to FMRM discovery by mapping topics to functional modules, images to miRNAs, and words to mRNAs, respectively.

In this work, we firstly modify the Corr-LDA and derive the solution to discover FMRMs. Then, we apply our method to mouse model expression data sets for human breast cancer research. The result shows that our model is able to capture several biologically meaningful modules. Furthermore, without using the prior target binding information, the identified miRNAs and mRNAs in each FMRMs show a large proportion of overlap with predicted miRNA target relationships, suggesting that targets and FMRMs can be predicted from expression profiles alone, and providing an independent verification of the underlying strategy.

2 METHODS

We begin with the assumption that functional modules governing miRNA and mRNA expression, which are associated with a variety of biological functions, are reflected by the data from microarray experiments. We model functional modules with latent random variables which act as a bridge between miRNAs and mRNAs. By inferring the latent variables, we can identify FMRMs.

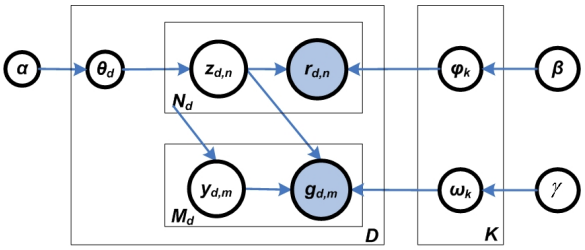

Fig. 1. Generative model of FMRM discovery. Given expression data of miRNAs and mRNAs of D samples, each sample d is a mixture of random miRNAs and mRNAs. Each miRNA $r_{d,n}$ and mRNA $g_{d,m}$ are generated from one of the K latent functional modules, selected by $z_{d,n}$.

2.1 Modeling FMRM discovery

More specifically, we model FMRMs with a probabilistic generative process. Given the K latent functions presented in the samples, our method considers miRNAs and mRNAs as observations generated from a probabilistic process over these K functions. Thus, each sample is a random mixture of miRNAs and mRNAs associated with K functional modules. By inferring the probability distributions of the latent variables, we are able to obtain the probabilities of how samples, miRNAs, and mRNAs are related to functional modules.

We depict the model in Figure 1 with a plate notation. In this notation, nodes stand for random variables (observed variables are shaded and latent ones are unshaded); edges denote conditional dependency between random variables; and plates denote replications of a substructure with the number of repetitions given in the bottom corner (either right or left side).

In Figure 1, the D samples were profiled with a set of miRNAs V and a set of mRNAs T. Random variable $r_{d,n}$ and $g_{d,m}$ denote the indexes of a miRNA and mRNA expressed in the d-th sample, respectively, with $d \in \{1, \cdots, D\}$, $n \in \{1, \cdots, N_d\}$, and $m \in \{1, \cdots, M_d\}$. N_d and M_d are the total numbers of times the miRNAs and mRNAs which are expressed in the d-th sample. Random variable $z_{d,n}$ stands for the latent functional module associating with the n-th miRNA in the d-th sample. We assume that $z_{d,n}$, $r_{d,n}$, and $g_{d,m}$ all have multinomial distributions with parameters θ_d, φ_k, and ω_k, respectively. Each parameter has a Dirichlet prior with hyperparameters α, β, and γ, correspondingly.

Without considering the putative target constraints, the generative procedure for each sample d can be illustrated by the following hierarchical sampling process: to generate the d-th sample, i) a latent module $z_{d,n}$ is drawn from its multinomial distribution θ_d; ii) then, a miRNA $r_{d,n}$ is drawn from its multinomial distribution φ_k, given the selected module $z_{d,n}$; iii) for each mRNA $g_{d,m}$, one of the miRNAs, indexed by $y_{d,m}$, is selected from $R_d = \{r_{d,n}\}$ and a corresponding mRNA $g_{d,m}$ is drawn from its multinomial distribution ω_k, conditional upon the same module that generates the selected miRNA $r_{d,n}$.

When the constraint of the putative target relationship between miRNAs and mRNAs is preferred, for each mRNA, one of the miRNAs from the set of hosting miRNAs of that mRNAs is selected, and a corresponding mRNA is drawn from the multinomial distribution of mRNAs, conditional upon the same module that generates the selected miRNA.

From the above generative process, the parameter $\mathbf{\Theta} = \{\theta_d\}$ associates samples with modules, $\mathbf{\Phi} = \{\varphi_k\}$ assigns the probability of miRNAs expressed in module $\mathbf{Z} = \{z_{d,n}\}$, and $\mathbf{\Omega} = \{\omega_k\}$ indicates the probability of mRNAs expressed in $\mathbf{Z}$ corresponding to the miRNAs. Therefore, by estimating $\mathbf{\Theta}$, $\mathbf{\Phi}$, and $\mathbf{\Omega}$, we can identify FMRMs (details in Sections 2.3 to 2.5).

Under this model, miRNAs can associate with any modules, but mRNAs may only associate with the modules that produce the miRNAs. In effect, this model captures the hierarchical notion that miRNAs are generated under specific FMRMs, and mRNAs are regulated by the miRNAs.

2.2 Data Conversion

In order to apply the above model to the expression profiles of miRNAs and mRNAs, we convert the expression values to the counts of miRNAs and mRNAs present in the samples.

Given a microarray experiment profiled D' samples, similar to Joung and Fei (2009), we considered that miRNAs and mRNAs have events of their expression in every sample that are likely to be associated with functional modules. Therefore, each miRNA or mRNA can be represented as a vector of variables, $\{s_1^+, s_2^-, \cdots, s_{D-1}^+, s_D^-\}$. It corresponds to the expression events of a miRNA or mRNA in all samples, where duplex $\{s_{2d-1}^+, s_{2d}^-\}$ indicates an over- and under- expressed miRNA or mRNA of sample d, $d \in \{1, \cdots, D'\}$, thus, $D = 2D'$. To get the integer counts $(\sigma_{2d-1,i}, \sigma_{2d,i})$ for the duplex expression status, we convert the expression value of a miRNA or mRNA of sample d with,

$$\sigma_{2d-1,i}, \sigma_{2d,i} = \begin{array}{ll} \lceil \varepsilon \cdot |e_{d,i}| \rceil, 0, & if\ e_{d,i} \geq med_d \\ 0, \lceil \varepsilon \cdot |e_{d,i}| \rceil, & if\ e_{d,i} < med_d \end{array} \quad (1)$$

where $e_{d,i}$ is the expression value of a miRNA or mRNA in the d-th sample, ε is a scaling constant, and med_d denotes the median of all miRNAs or mRNAs in the d-th sample.

Then, the counts of miRNAs and mRNAs are replaced by the indexes from the set of miRNAs, V and the set of mRNAs, T. The indexes, therefore, are the random variables $r_{d,n}$ and $g_{d,m}$ used in the model (Figure 1).

2.3 Estimating model parameters

Because the exact inference for the parameters of our model is intractable, we used the collapsed Gibbs sampling method (Liu, 1994) to estimate parameters.

This method iteratively generates samples that converge to draws from a target distribution of random variables Z through integrating out the parameters $\mathbf{\Theta}$, $\mathbf{\Phi}$, and $\mathbf{\Omega}$ for each sampling. For the d-th sample and the n-th miRNA, the sampling is expressed as a conditional probability:

$$p(z_{d,n} = k | Z_{-(d,n)}, Y_d, R_d, G_d) \propto$$
$$p(z_{d,n} | Z_{-(d,n)}) p(r_{d,n} | z_{d,n}) \prod_{m=1}^{M_d} p(g_{d,m} | y_{d,m}, Z_d) \propto$$
$$\frac{n^k_{d,-(d,n)} + \alpha_k}{(\sum_{k=1}^K n_d^{(k)} + \alpha_k) - 1} \cdot \frac{n^v_{k,-(d,n)} + \beta_v}{(\sum_{v=v_1}^{vP} n_k^{(v)} + \beta_v) - 1} \cdot \frac{m^t_{k,-(d,n)} + \gamma_t}{(\sum_{t=t_1}^{tQ} m_k^{(t)} + \gamma_t) - 1} \quad (2)$$

where $z_{d,n}$ is the current module assignment of the n-th miRNA of the d-th sample. $Z_{-(d,n)}$ is the current module assignment of all miRNAs in all samples excluding that of the n-th miRNA of d-th sample. $n^k_{d,-(d,n)}$ is the number of times that the k-th FMRM has been observed with miRNAs across samples excluding that of the n-th miRNA of the d-th sample. $n^v_{k,-(d,n)}$ is the number of times that miRNA v is assigned to the k-th FMRM excluding that of the n-th miRNA of d-th sample. $m^t_{k,-(d,n)}$ is the number of times that mRNA t is assigned to the k-th FMRM excluding the current assignment.

After sufficient sampling, the distribution of $z_{d,n}$ converges to the target distribution of $\mathbf{Z}$, then we estimate the parameters Θ, Φ, and Ω based on the values of the module assignments produced from the sampling:

$$\theta_{d,k} = \frac{n_d^k + \alpha_k}{\sum_{k=1}^K n_d^{(k)} + \alpha_k}, \quad \varphi_{k,v} = \frac{n_k^v + \beta_v}{\sum_{v=v_1}^{vP} n_k^{(v)} + \beta_v},$$
$$\omega_{k,t} = \frac{m_k^t + \gamma_t}{\sum_{t=t_1}^{tQ} m_k^{(t)} + \gamma_t} \quad (3)$$

Unlike the preceding sampling procedure, here $n_d^{(k)}$,$n_k^{(v)}$, and $m_k^{(t)}$ are calculated from the assignment results for all data without excluding the current module. Using Eq. 2 and 3, the Gibbs sampling procedure can be designed. The algorithm includes three stages: initialization, sampling, and reading out of parameters. It is provided in Supplementary File 1 – Algorithm 1.

2.4 Assigning biological conditions to modules

The parameters inferred from this model provide insights into the data sets at several levels. Θ clusters samples into modules that should relate to the biological conditions of the experiments.

We conceive a statistic model to identify the connection between biological conditions and modules. Let C be the number of biological conditions of the D samples in the data set, and c_i be the number of samples belonging to condition i, where $\sum_{i=1}^C c_i = D$. For each module, assume there are x samples among the n highest probability samples that belong to same condition i. The random variable x follows a hypergeometric distribution with parameters D, c_i, and n, denoted as

$$p(x) \sim hypergeometric(x; D, c_i, n) \quad (4)$$

We assign the biological condition i to module k when x is at a statistically significant level, for example, p-value< 0.05.

2.5 Identifying miRNAs and mRNAs for modules

The parameters $\mathbf{\Phi}$ and $\mathbf{\Omega}$ indicate the probabilities of each miRNA and mRNA participating in a FMRM. For a K-FMRMs, $\mathbf{\Phi}$ is a $K \times P$ probability matrix where the element $\varphi_{k,v}$ indicates the likelihood that miRNA v belongs to the k-th FMRM. Similarly, $\mathbf{\Omega}$ is a $K \times Q$ probability matrix where the element $\omega_{k,t}$ indicates the belief of mRNA t participating in the k-th FMRM, and Q is the number of mRNAs under investigation.

For each FMRM, we consider the top ranked miRNAs and mRNAs with the highest probabilities to be the participants of the FMRM.

2.6 Reconstructing miRNA-mRNA target relationships

We query a miRNA target database to reconstruct the target relationship of the miRNAs and mRNAs in each module. Hypothesis tests are conducted on the identified miRNAs and mRNAs to evaluate whether they are likely to have been identified by chance or not.

2.7 Function and pathway analysis of FMRMs

Function and pathway analysis of the identified FMRMs is conducted by reviewing literature and querying the Ingenuity Pathway Analysis (IPA, www.ingenuity.com) database of functional biological pathways to identify the significantly enriched functions and pathways.

3 RESULTS

In this section, we present the results and analysis of applying our model to a mouse mammary dataset (Zhu *et al.*, 2010).

The data set were profiled with 46 samples derived from 9 classes of mouse models, representing one normal type and two breast cancer subtypes: basal and luminal. The expression data were screened with 1,336 probes of miRNAs (corresponding to 334 unique miRNAs) and 22,626 probes of mRNAs. For each type of the conditions, 3-7 samples were profiled with miRNAs and mRNAs (details in Supplementary File 1).

In order to compare with the target prediction, the expression data sets of miRNAs and mRNAs were further filtered with MicroCosm Targets V5.0 (Griffiths-Jones *et al.*, 2008), and only those in MicroCosm were maintained for analysis. Consequently, 1,112 probes of miRNAs and 19,223 probes of mRNAs were used in our experiment.

3.1 Implementation

Given the above expression data of miRNAs and mRNAs, the input data for our model include a 1,112×46 matrix of miRNA expression values and a 19,223×46 matrix of mRNA expression values. In the following discussion, we do not consider the putative target information to avoid the bias probably incurred by the prior prediction (Bartel, 2009).

In the experiment, the constant ε for converting the expression values was 30. After the data conversion, the number of samples D is 92. We set the number of FMRMs, K, to 20. This value is determined by the number of sample types. Our data sets were profiled with 9 classes of mouse models. miRNAs and mRNAs could be over- or under- expressed in the samples so the number of sample types is 18. In addition, 2 extra types were added to allow the redundancy as our model could discover subtypes of classes. We set the hyperparameters α, β, and γ to 10. The number of iterations of Gibbs sampling is 2,500. These value settings are based on empirical experiments.

3.2 Associating FMRMs with biological conditions

The parameter $\mathbf{\Theta}$ obtained with our method is a 92×20 probability matrix. Referring to section 2.4, the element $\theta_{d,k}$ of $\mathbf{\Theta}$ is the belief of sample d belonging to module k. We extracted the top 5% (5) ranked samples with highest probabilities in each module, and assigned biological conditions to each module according to these samples as discussed in section 2.4. Figure 2 illustrates this mapping procedure. The 5 highest probability samples extracted from each module are arranged on the y-axis. Their associations with other modules are also shown in the map.

In order to assign biological conditions to modules at the statistically significant level, we conceived a statistical model to map modules to biological conditions by using the mouse model classes instead of tumor types directly (Table 1). From Table 1, 7 modules have been mapped to specific mouse model classes at a significant level (p-value < 0.05). These mouse models can be further mapped to two human breast tumor subtypes (Desai *et al.*, 2002; Blenkiron *et al.*, 2007; Herschkowitz *et al.*, 2007), suggesting that the identified modules are associated with those biological conditions. Other modules are clustered by samples with mixed biological conditions, suggesting that they may participate in several cellular processes.

Furthermore, the top 5% (56) ranked probes of miRNA and the top 0.1% (192) ranked probes of mRNA with the highest probabilities in each module were also extracted from the inferred parameters $\mathbf{\Phi}$ and $\mathbf{\Omega}$. They are assigned to the same biological conditions according to the modules they belong to, respectively. The detailed information of each FMRM is given in Supplementary File 2, including the miRNAs, predicated target mRNAs, and the associated biological condition for each FMRM.

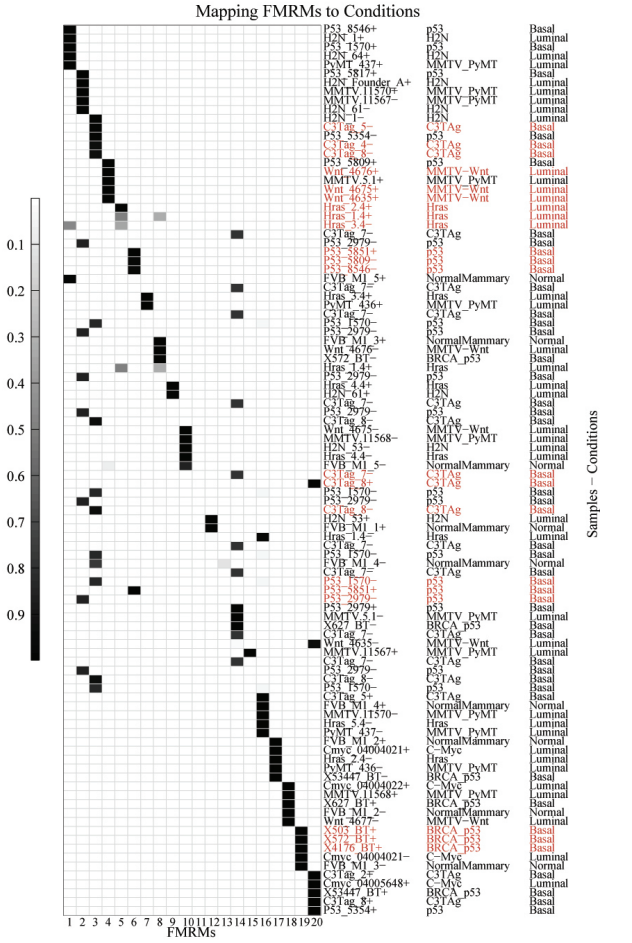

Fig. 2. Assigning biological conditions to FMRMs. The y-axis on the right side of the figure denotes sample names, mouse model types, and breast cancer subtypes in three columns. Using the parameter $\mathbf{\Theta}$, the likelihood that a particular sample is associated with a specific module, the top 5% samples associated with each module are displayed using the grey scale. These samples are considered to map modules to biological conditions. Samples may occur more than once in the y-axis because some samples are significantly associated with more than one module. Some modules, such as module-11, have only rather low probability of association with samples, and thus have nearly white shading even for their top 5 samples. Significant mapping of FMRMs to conditions is highlighted.

3.3 Target reconstruction

To reconstruct the target relationships between miRNAs and mRNAs, we use MicroCosm to link miRNAs and mRNAs identified in each FMRM. The numbers of linked miRNAs and mRNAs are given in Table 2.

To investigate whether the miRNAs and mRNAs in each module were identified by chance, we randomly selected a group of miRNAs and a group of mRNAs from MicroCosm with the same numbers as those in the identified modules, and queried how many pairs that can be linked by MicroCosm. The distribution of the number of matched pairs was estimated by a simulation which was executed 10,000 times. Illustrated with Figure S1 in Supplementary File 1, the estimated distribution shows that the numbers of target relationships of the randomly chosen miRNAs and mRNAs are significantly different from those of the identified miRNAs and mRNAs in each module (p-value < 0.05). It indicates that the miRNAs and mRNAs in each module are not identified by chance. The linked miRNAs and mRNAs of each FMRM are given in Supplementary File 3.

Table 1. Assigning biological conditions to FMRMs.

FMRM#	c_i	x	Mouse model class	Tumor subtype	p-value
3	10	3	C3TAg	Basal	0.0081
4	8	3	MMTV-Wnt	Luminal	0.004
5	10	3	Hras	Luminal	0.0081
6	14	3	p53	Basal	0.0222
11	10	3	C3TAg	Basal	0.0081
13	14	3	p53	Basal	0.0222
19	10	3	BRCA_p53	Basal	0.0081

According to Eq. 4, biological conditions are assigned to FMRMs based on a hypergeometric distribution. The significant results are given in this table. The size of population is 92, the number of each draw is 5% of the population, i.e. 5. c_i is the number of samples belonging to each condition, which include both over- and under expressed status. x denotes the observed number of samples with the assigned biological condition in each draw. FMRM# is the module number corresponding to the number in Figure 2.

Table 2. Numbers of miRNA-mRNA pairs identified in FMRMs.

FMRM#	Subtype	miRNA#	mRNA#	target pair#	p-value
3	Basal	33	190	273	1.70E-07
4	Luminal	18	190	147	3.23E-08
5	Luminal	16	191	144	2.98E-07
6	Basal	16	189	146	1.48E-06
11	Basal	17	190	122	1.13E-11
13	Basal	18	186	136	1.29E-10
19	Basal	18	188	133	2.71E-12

The miRNAs and mRNAs identified in each module are linked by MicroCosm. Compared with the number of pairs linked by MicroCosm given the same number of randomly chosen miRNAs and mRNAs, the miRNAs and mRNAs identified in each module are not identified by chance.

3.4 Functional validation of miRNAs

To further validate that the identified miRNAs are relevant to cancers, we investigated the implications of miRNAs for cancers through literature review. We built a benchmark based on the current knowledge (details in Supplementary File 4), and compared it with the miRNAs identified in the modules.

From the literature, 42 miRNAs have been validated to have implications for cancers. We identified a significant number of miRNAs covered by the benchmark shown in Table 3. The comparison shows that the miRNAs identified by our method are largely consistent with the current knowledge of miRNAs for cancers.

It is worth noting that several miRNAs, such as the *let-7* family and *miR-21*, are identified in multiple modules, suggesting they could be involved in multiple biological processes. The frequent occurrence of these particular miRNAs is consistent with their known strong association with multiple cancer types, including breast cancers. The identification of multiple modules containing different but overlapping sets of miRNAs is likely to be the consequence of activation of distinct subsets of common gene interaction networks in specific cancer subtypes. For example, Blenkiron *et al.* (2007) identified 31 miRNAs differentially

Table 3. Validation of identified miRNAs in the FMRMs.

FMRM#	Supported miRNAs	Coverage	p-value
3	*let-7a, let-7b, let-7c, let-7d, let-7e, let-7f, miR-221,miR-29a*	8/33(22.24%)	0.02641
4	*let-7a, let-7b, let-7c,let-7d, let-7e, let-7f, let-7g, let-7i, miR-21,miR-221*	10/18(55.56%)	6.68E-06
5	*let-7b, let-7c, let-7d, let-7i, miR-200b, miR-200c, miR-29a, miR-29b, miR-30c*	9/17(52.94%)	3.56E-05
6	*let-7a, let-7b, let-7c, let-7d, let-7i, miR-103, miR-21, miR-221*	8/16(50.00%)	1.76E-04
15	*let-7a, let-7c,let-7f, let-7g, miR-141, miR-19b, miR-21, miR-200a,miR-200b*	9/17(52.94%)	3.56E-05
19	*let-7a, let-7b, let-7c, let-7d, let-7e, let-7f, miR-143, miR-145, miR-21, miR-29a, miR-29b*	11/18(61.11%)	5.45E-07

The comparison shows that significant numbers of miRNAs identified in the FMRMs are relevant to cancers. From the literature, 42 miRNAs have been validated as either oncogenes or tumor suppressors (details in Supplementary File 4). Among the 334 miRNAs under investigation, a significant number of miRNAs in identified modules are supported by the current knowledge. The coverage is the percentage of the number of miRNAs in each module supported by literature. p-value is calculated by a hypergeometric probability density function at each of the numbers of miRNAs supported by the literature, using the corresponding size of the total miRNAs under investigation (334), numbers of miRNAs in each module, and numbers of miRNAs identified from the literature (42). The modules with significant supports are given in this table.

expressed between basal and luminal tumors. Among them, *let-7a*, *b*, and *f* are under-expressed in basal tumors but over-expressed in luminal tumors. These miRNAs were identified in module 3, 4, 5, and 6 using our method and show patterns that are consistent with their reported involvement in these tumors.

3.5 Functional validation of miRNA target genes

It is expected that the miRNA target genes are also relevant to the specific biological processes. To validate that the identified mRNAs are relevant to basal and luminal tumors, firstly we compared the identified mRNAs with a work conducted by Adelaide *et al.* (2007). Their results suggest the existence of potential oncogenes and tumor suppressor genes differentially associated with the basal and luminal subtype. As their results are largely consistent with many previous researches (Bergamaschi *et al.*, 2006; Chin *et al.*, 2006; Neve *et al.*, 2006), we validate our analysis based on their results.

In our results, 18 genes have been identified by Adelaide *et al.* (2007) as in Table S1 of Supplementary File 1. Among these genes, Ccdc77 identified in module-3 also is targeted by *miR-29a* and *miR-221*, Hspa14 identified in module-4 is targeted by *miR-21*, and Cox4i1 identified in module-19 is targeted by *let-7c* and *let-7e*. It further confirms that *let-7e*, *miR-21*, *miR-29a*, and *miR-221* may have important regulatory functions towards basal and luminal tumors. In addition, Rbm4b identified in module-3 is targeted by

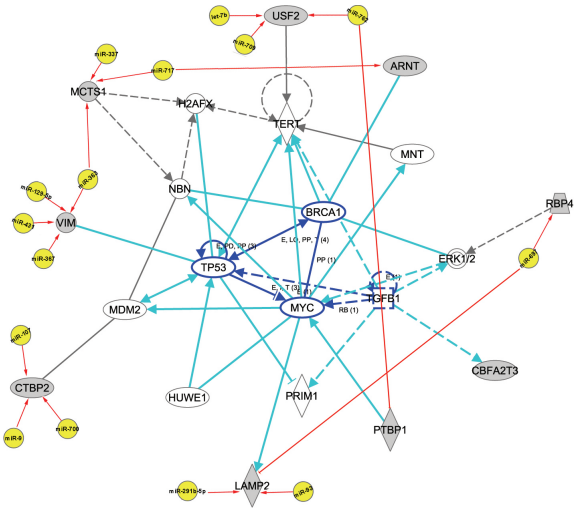

Fig. 3. A network with the function of cancer, cellular compromise, DNA replication, and repair. It is participated by a group of miRNAs and their target mRNAs identified by our method, suggesting these miRNAs and their target mRNAs have the function of cancers.

Table 4. Cancer associated genes of FMRMs.

FMRM#	mRNAs	mRNA#	*p*-value (adj.)
3	CALR, RBP4, VIM, NDUFV2, SDCBP, MCTS1, AP2S1, PRPF8, COL18A1, AK2, ARNT, RPS15	12	4.89E-03 - 2.54E-02
4	DNMT1, NF2, RRM2	3	2.13E-03 - 3.05E-02
5	CEBPB, DDX39, HSP90AB1, MT2A, NUP62, SQLE, TCP1, TRIO	8	2.01E-03 - 4.84E-02
11	DICER1, ENO1, HSP90B1, RXRB, SPRY2	5	5.26E-03 - 4.84E-02
13	IGF2R, LSM14B, NCOR2, SP110, STX5, TOR2A, ACHE, HDAC3, PARP1, POSTN, SMAD4, UBE2I, RNF6, BAK1	14	6.88E-03 - 4.56E-02

Many genes identified in FMRMs are relevant to cancers. Genes identified in FMRMs are directly assigned to diseases and disorders. The cancer related genes of FMRMs within their top five bio-functions are listed.

miR-697 and *miR-700*, Rbx1 identified in module-5 is targeted by *miR-709*, Gspt1 identified in module-11 is targeted by *miR-669c* and *miR-710*, and Cox4i1 identified in module-19 is targeted by *miR-709*. It suggests that *miR-669c*, *miR-697*, *miR-709*, and *miR-710* may also play important roles in regulating basal and luminal tumors. It is worth noting that many previously reported results were not recovered in the current study because the investigated data were profiled with mouse model while the results of Adelaide *et al.* (2007) were produced on breast cancer samples of humans.

Furthermore, we have queried the mRNAs identified in each module against the Ingenuity Pathway Analysis (IPA) Database. We specifically focused on human species as we are interested in the networks of human cancers. The networks participated by the mRNAs identified in FMRMs are highly associated with cancers. Many genes are directly related to cancers and genetic disorders. They are co-targeted by a group of miRNAs identified from our method, suggesting the identified miRNAs and their target mRNAs have implications for cancers. For example, a network participated by the miRNAs and mRNAs identified by our method are associated with cancer, cellular compromise, DNA replication, and repair (Figure 3). The networks which are explicitly associated with cancers and within the top five networks of each module are given in Table S2 of Supplementary File 1. The detailed networks are also given in Figure S2 to S6 of the Supplementary File 1. The identified genes of FMRMs, which are relevant to cancers, are given in Table 4. The results indicate that our methods effectively identified many cancer related genes. Those genes are targeted by a group of miRNAs, suggesting those miRNAs also participate in the networks of cancers.

4 CONCLUSION AND DISCUSSION

miRNAs have been regarded as one of the most important regulators. Identifying their functions and regulatory mechanisms is critical in understanding biological processes of organisms. Great efforts, in both biological experiments and computational methods, have been made to illustrate their functions. However, the precise regulatory functions of most miRNAs remain elusive due to the complexity of the regulatory mechanisms.

In this paper, we have presented a model to discover functional miRNA regulatory modules (FMRMs), which are groups of miRNAs and mRNAs for specific biological conditions. This model is inspired by the Corr-LDA, which has been used to extract the correspondence patterns from heterogeneous data. We modified Corr-LDA and derived the solution for FMRM discovery.

Our method models FMRMs with a generative process. It makes use of the expression profiles of miRNAs and mRNAs, with or without using the target relationships between miRNAs and mRNAs based on the sequence binding information. It simultaneously identifies groups of interactive miRNAs and mRNAs, which are believed to participate in specific biological functions.

We have applied this method to a mouse model data set for human breast cancer research. The method has effectively identified several modules related to breast cancer subtypes: basal and luminal. Since the data sets used were profiled from mouse tissues, many genes have been filtered out because we focus on human genes. Thus, previously reported results were not fully recovered in this work. However, a large proportion of miRNAs and mRNAs identified in the modules have been reported to have associations with basal and luminal subtypes. Many others have direct indications on cancers and genetic disorders. Furthermore, many novel associations among miRNAs, mRNAs, and biological processes have been predicted by our model. Several miRNAs and mRNAs are highly related to cancers as reported by previous works, suggesting those modules may have roles in the corresponding development processes.

Our model allows discovering the FMRMs with or without using the target relationship between miRNAs and mRNAs. Some researchers have suggested that algorithms that do not consider known targets may avoid biases (Lewis *et al.*, 2003, 2005; Bartel, 2009). Bonnet *et al.* (2010) also showed that expression profiles only can be used to infer miRNA regulatory networks. Our method

provides the flexibility of inferring FMRMs with or without target relationships of miRNAs and mRNAs. We have demonstrated this model without using the prior target prediction. The results suggest that expression profiles of miRNAs and mRNAs are crucial for both target identification and regulatory module discovery.

ACKNOWLEDGEMENT

We thank Dr. Chuxia Deng, Dr. Daniel Medina, Dr. Yi Li, Dr. Ming Yi, and Dr. Robert Stephens for providing some tumor samples of the data sets.

Funding: This research has been supported in part by the Intramural Program, Center for Cancer Research, NIH, Bethesda, MD.

Conflict of interest: None declared.

REFERENCES

Adelaide, J., Finetti, P., Bekhouche, I., Repellini, L., Geneix, J., Sircoulomb, F., Charafe-Jauffret, E., Cervera, N., Desplans, J., Parzy, D., Schoenmakers, E., Viens, P., Jacquemier, J., Birnbaum, D., Bertucci, F., and Chaffanet, M. (2007). Integrated profiling of basal and luminal breast cancers. *Cancer Res*, **67**(24), 11565–11575.

Akao, Y., Nakagawa, Y., and Naoe, T. (2007). MicroRNA-143 and -145 in colon cancer. *DNA and Cell Biology*, **26**(5), 311–320.

Bartel, D. P. (2009). MicroRNAs: Target recognition and regulatory functions. *Cell*, **136**(2), 215–233.

Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., Barzilai, A., Einat, P., Einav, U., Meiri, E., Sharon, E., Spector, Y., and Bentwich, Z. (2005). Identification of hundreds of conserved and nonconserved human microRNAs. *Nat Genet*, **37**(7), 766–770.

Bergamaschi, A., Kim, Y. H., Wang, P., Sorlie, T., Hernandez-Boussard, T., Lonning, P. E., Tibshirani, R., Borresen-Dale, A.-L., and Pollack, J. R. (2006). Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. *Genes, Chromosomes and Cancer*, **45**(11), 1033–1040.

Blei, D. M. and Jordan, M. I. (2003). Modeling annotated data. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134, New York, NY, USA. ACM.

Blenkiron, C., Goldstein, L., Thorne, N., Spiteri, I., Chin, S.-F., Dunning, M., Barbosa-Morais, N., Teschendorff, A., Green, A., Ellis, I., Tavare, S., Caldas, C., and Miska, E. (2007). MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype. *Genome Biology*, **8**(10), R214–.

Bonnet, E., Tatari, M., Joshi, A., Michoel, T., Marchal, K., Berx, G., and Van de Peer, Y. (2010). Module network inference from a cancer gene expression data set identifies microRNA regulated modules. *PLoS ONE*, **5**(4), e10162.

Chin, K., DeVries, S., Fridlyand, J., Spellman, P. T., Roydasgupta, R., Kuo, W.-L., Lapuk, A., Neve, R. M., Qian, Z., Ryder, T., Chen, F., Feiler, H., Tokuyasu, T., Kingsley, C., Dairkee, S., Meng, Z., Chew, K., Pinkel, D., Jain, A., Ljung, B. M., Esserman, L., Albertson, D. G., Waldman, F. M., and Gray, J. W. (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*, **10**(6), 529–541.

Croce, C. M. (2009). Causes and consequences of microRNA dysregulation in cancer. *Nat Rev Genet*, **10**(10), 704–714.

Desai, K. V., Xiao, N., Wang, W., Gangi, L., Greene, J., Powell, J. I., Dickson, R., Furth, P., Hunter, K., Kucherlapati, R., Simon, R., Liu, E. T., and Green, J. E. (2002). Initiating oncogenic event determines gene-expression patterns of human breast cancer models. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(10), 6967–6972.

Du, T. and Zamore, P. D. (2007). Beginning to understand microRNA function. *Cell Res*, **17**(8), 661–663.

Griffiths-Jones, S., Saini, H. K., van Dongen, S., and Enright, A. J. (2008). mirbase: tools for microRNA genomics. *Nucl. Acids Res.*, **36**(suppl_1), D154–158.

Hatzigeorgiou., A. G. (2007). Same computational analysis, different miRNA target predictions. *Nature Methods*, **4**(3), 191.

He, L. and Hannon, G. J. (2004). MicroRNAs: small RNAs with a big role in gene regulation. *Nat Rev Genet*, **5**(7), 522–531.

Herschkowitz, J., Simin, K., Weigman, V., Mikaelian, I., Usary, J., Hu, Z., Rasmussen, K., Jones, L., Assefnia, S., Chandrasekharan, S., Backlund, M., Yin, Y., Khramtsov, A., Bastein, R., Quackenbush, J., Glazer, R., Brown, P., Green, J., Kopelovich, L., Furth, P., Palazzo, J., Olopade, O., Bernard, P., Churchill, G., Van Dyke, T., and Perou, C. (2007). Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biology*, **8**(5), R76.

Huang, J. C., Babak, T., Corson, T. W., Chua, G., Khan, S., Gallie, B. L., Hughes, T. R., Blencowe, B. J., Frey, B. J., and Morris, Q. D. (2007). Using expression profiling data to identify human microRNA targets. *Nat Meth*, **4**(12), 1045–1049.

Iorio, M. V., Ferracin, M., Liu, C.-G., Veronese, A., Spizzo, R., Sabbioni, S., Magri, E., Pedriali, M., Fabbri, M., Campiglio, M., Menard, S., Palazzo, J. P., Rosenberg, A., Musiani, P., Volinia, S., Nenci, I., Calin, G. A., Querzoli, P., Negrini, M., and Croce, C. M. (2005). MicroRNA gene expression deregulation in human breast cancer. *Cancer Res*, **65**(16), 7065–7070.

Joung, J.-G. and Fei, Z. (2009). Identification of microRNA regulatory modules in arabidopsis via a probabilistic graphical model. *Bioinformatics*, **25**(3), 387–393.

Joung, J.-G., Hwang, K.-B., Nam, J.-W., Kim, S.-J., and Zhang, B.-T. (2007). Discovery of microRNA-mRNA modules via population-based probabilistic learning. *Bioinformatics*, **23**(9), 1141–1147.

Krek, A., Grun, D., Poy, M., Wolf, R., Rosenberg, L., Epstein, E., MacMenamin, P., Piedade, I., Gunsalus, K., Stoffel, M., and Rajewsky, N. (2005). Combinatorial microRNA target predictions. *Nature Genetics*, **37**, 495 – 500.

Lewis, B. P., Shih, I. h., Jones-Rhoades, M. W., Bartel, D. P., and Burge, C. B. (2003). Prediction of mammalian microRNA targets. *Cell*, **115**(7), 787–798.

Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**(1), 15–20.

Liu, B., Li, J., and Tsykin, A. (2009a). Discovery of functional miRNA-mRNA regulatory modules with computational methods. *Journal of Biomedical Informatics*, **42**(4), 685–691.

Liu, B., Li, J., Tsykin, A., Liu, L., Gaur, A., and Goodall, G. (2009b). Exploring complex miRNA-mRNA interactions with bayesian networks by splitting-averaging strategy. *BMC Bioinformatics*, **10**(1), 408.

Liu, J. S. (1994). The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, **89**, 930 – 958.

Neve, R. M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F. L., Fevr, T., Clark, L., Bayani, N., Coppe, J.-P., Tong, F., Speed, T., Spellman, P. T., DeVries, S., Lapuk, A., Wang, N. J., Kuo, W.-L., Stilwell, J. L., Pinkel, D., Albertson, D. G., Waldman, F. M., McCormick, F., Dickson, R. B., Johnson, M. D., Lippman, M., Ethier, S., Gazdar, A., and Gray, J. W. (2006). A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell*, **10**(6), 515–527.

Peng, X., Li, Y., Walters, K.-A., Rosenzweig, E., Lederer, S., Aicher, L., Proll, S., and Katze, M. (2009). Computational identification of hepatitis c virus associated microRNA-mRNA regulatory modules in human livers. *BMC Genomics*, **10**(1), 373.

Porkka, K. P., Pfeiffer, M. J., Waltering, K. K., Vessella, R. L., Tammela, T. L., and Visakorpi, T. (2007). MicroRNA expression profiling in prostate cancer. *Cancer Res*, **67**(13), 6130–6135.

Steyvers, M., Smyth, P., Rosen-Zvi, M., and Griffiths, T. (2004). Probabilistic author-topic models for information discovery. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 306–315, New York, NY, USA. ACM.

Tran, D., Satou, K., and Ho, T. (2008). Finding microRNA regulatory modules in human genome using rule induction. *BMC Bioinformatics*, **9**(Suppl 12), S5.

Yanaihara, N. (2006). Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. *Cancer Cell*, **9**(3), 189 – 198.

Yang, H., Kong, W., He, L., Zhao, J.-J., O'Donnell, J. D., Wang, J., Wenham, R. M., Coppola, D., Kruk, P. A., Nicosia, S. V., and Cheng, J. Q. (2008). MicroRNA expression profiling in human ovarian cancer: miR-214 induces cell survival and cisplatin resistance by targeting PTEN. *Cancer Res*, **68**(2), 425–433.

Yoon, S. and De Micheli, G. (2005). Prediction of regulatory modules comprising microRNAs and target genes. *Bioinformatics*, **21**(suppl_2), ii93–100.

Zhao, Y. and Srivastava, D. (2007). A developmental view of microRNA function. *Trends in Biochemical Sciences*, **32**(4), 189–197.

Zhu, M., Yi, M., Kim, C. H., Deng, C., Li, Y., D.Medina, Hunter, K., Stephen, R., and Green, J. (2010). Comprehensive genomic profiling identifies miRNA signatures associated with mouse mammary tumor subtypes. In preparation.