Iterative segmented least square method for functional microRNA-mRNA module discovery in breast cancer

Sungmin Rhee
Computer Science and Engineering, Seoul National University, Seoul, South Korea
Email: lars@snu.ac.kr

Sangsoo Lim
Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, South Korea
Email: sslim@snu.ac.kr

Sun Kim*
Computer Science and Engineering, Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, South Korea
Email: sunkim.bioinfo@snu.ac.kr
*Corresponding author

Abstract: MicroRNAs (miRNAs) have significant biological roles at the molecular level by regulating genes post-transcriptionally. To understand the functional effects of miRNAs in different biological contexts, it is essential to elucidate miRNA-mRNA regulatory modules (MRMs). The computational complexity for inferencing MRMs is very high due to the many-to-many relationships between miRNAs and mRNAs and inferencing MRMs is still a challenging unresolved problem. In this paper, we propose a novel iterative segmented least square method for functional MRM discovery. Our method operates in two steps: (a) grouping and ordering the miRNAs and mRNAs to build per-sample matrices representing miRNA-mRNA regulations, and (b) determining maximum sized modules from structured miRNA-mRNA matrices. In experiments with human breast cancer data sets from TCGA, we show that our method outperforms existing methods in terms of both GO similarity and cluster evaluation. In addition, we show that modules determined by our method can be used for breast cancer survival prediction and subtype classification.

Keywords: microRNA; regulatory network inference; optimisation; dynamic programming.
1 Introduction

MicroRNA(miRNA)s are small non-coding RNAs, short in length of 19–25 nucleotides. miRNAs have significant biological roles at the molecular level by regulating genes post-transcriptionally in animals and plants. The number of identified miRNAs in the human genome is about a thousand and they are known to target about 60% of protein coding genes (Friedman et al., 2009).

Target relationship between miRNA and mRNAs can be modelled as $m$ to $n$ bipartite graph since a miRNA can target multiple mRNA transcripts and mRNA transcripts can be regulated by several miRNAs. Since $m$ is much smaller than $n$, we can say that miRNAs are innately hubs in the miRNA-mRNA regulation network. Thus revealing the mechanisms of miRNA-mRNA regulation is a very important problem. miRNAs regulate not only mRNAs but also correlate with transcription factors and siRNAs (Creasey et al., 2014). However, in this study, we concentrate only on relationship between miRNA and protein coding genes and the goal is to identify a subnetwork of miRNA-mRNA or modules in an experimental condition specific manner.

Many computational methods are developed to inferencing miRNA-mRNA target relationships. A recent survey by Yoon et al. (2014) grouped existing methods into three categories. Methods and tools in the first category use sequence complementary information between miRNAs and their putative target mRNA sequences. Many well-
known algorithms such as TargetScan (Lewis et al., 2005), PicTar (Krek et al., 2005), and PITA (Kertesz et al., 2007) belong to the first category. Since these tools do not consider the expression information of miRNA and genes, tools in the first category tend to have high false positive rates and may not detect functional relationships. Hence, many computational methods in the second category use these sequence based methods for filtering spurious miRNA target pairs. Typically, methods in the second category use machine learning techniques and experimentally verified target pairs as training data. DIANA-microT (Xu et al., 2014; Reczko et al., 2012) is one of the algorithms in the second category. Tools in the third category are to detect condition-specific target relationships between miRNAs and genes. MMIA (Chae et al., 2014; Nam et al., 2009; Xin et al., 2009), PlantMirnaT (Rhee et al., 2015), and GenMiR++ (Huang et al., 2007) are methods in the third category.

Unlike the early computational methods that are developed to determine one to one target relationships, recent methods focus on identifying subnetworks or modules from miRNA regulatory networks. Xu et al. (2011) developed a method that constructs miRNA-miRNA synergistic networks and mines modules using clique percolation on a miRNA-miRNA network. For a given miRNA-miRNA pair, it is connected if and only if their co-regulating genes are enriched in terms of GO annotation and protein interaction network. Liang et al. (2016) proposed an approach based on a biclique-merging technique. From a constructed miRNA-mRNA network, the authors determine statistically significant bicliques and merge them to determine MRMs (miRNA-mRNA regulatory modules). Jin et al. (2015) proposed a method based on biclustering and Bayesian network model. They first determine gene sets that have similar expression changes across patient samples, then determine upstream miRNAs using a Bayesian network model. This method was applied to ovarian cancer and glioblastoma (GBM) data and demonstrated that 48.4% and 74.0% of modules from ovarian cancer and GBM, respectively, were enriched with cancer-related pathways. In addition, 51.7% and 71.7% of miRNAs in modules were ovarian cancer-related and GBM-related, respectively.

2 Motivation

Although several methods are developed to determine MRM structures, there are only few algorithms that directly utilise the expression pattern to identify MRMs. For example, Xu et al. (2011) did not use the expression profiles of miRNAs and genes, so it is not designed to discover functional MRMs in a specific biological context. Liang et al. (2016) used the expression pattern only to investigate the pairwise relationship between a miRNA and a gene. Hence, miRNA-gene pairs in different functional contexts can be clustered to the same module. In addition, a method by Jin et al. (2015) did not consider the requirement that miRNAs in a module also should have similar expression changes.

In our work, we define MRM as a network of miRNAs and target genes where (1) targeting miRNAs have similar expression pattern across samples, (2) targeted genes have similar expression pattern across samples, (3) and miRNAs and mRNAs have negatively correlated expression pattern. To determine miRNA-mRNA modules, it is necessary to investigate the directed bipartite graph that has source nodes of size $m$ and target nodes of size $n$ where $m$ and $n$ is the number of miRNAs and mRNAs, respectively. As the matrix representation is typically used for representing a graph, we converted the bipartite graph into an adjacent matrix to mine the module structure from the matrix. Given a matrix representation with $m$ rows and $n$ columns, the problem can
be formulated as the biclustering problem that generates biclusters, a subset of rows which exhibit similar behaviour across a subset of columns, or vice versa (Han et al., 2011). Biclustering is an extensively investigated research topic in bioinformatics. A seminal work, a node-deletion biclustering algorithm by Cheng and Church (2000), was successful in finding co-regulation patterns in yeast and human. Since then, many biclustering algorithms have been developed in bioinformatics (Oghabian et al., 2014).

However, most of the biclustering algorithms are developed for genes × experiment conditions or genes × time matrices. In this case, the experiment condition or time dimension is smaller compared to the number of genes in the gene dimension. Our problem is actually mining clusters in three dimensions, gene × miRNA × phenotype. Clustering on three dimensions is not feasible since biclustering is already known as an NP-hard problem. Thus we intentionally transform the problem into a two dimension problem of gene × miRNA where an entry in the matrix indicates negative correlations among phenotypes to model the negative regulatory role of miRNA on genes. Even with this problem formulation, direct use of existing biclustering techniques (Madeira and Oliveira, 2004) are not effective since the miRNA dimension size is much bigger, 2 order of magnitude, than the condition or time dimension. To overcome the described challenge, we designed a novel two-step approach:

1. **Grouping and ordering**: Each dimension in the matrix are grouped and ordered by hierarchical clustering. Hierarchical clustering can group miRNAs and genes that share similar functions. Also as hierarchical clustering can afford the ordering configuration and candidates for matrix segmentation, the complexity of the problem can be reduced dramatically. After hierarchical clustering of stringent cut-off, we can determine correct but fragmented gene or miRNA clusters. The main issue is how to merge up fragmented clusters correctly, which is the main topic in the next step.

2. **Defining maximum-sized modules**: Iterative Segmented Least Square method is used for matrix segmentation for module discovery. Segmented Least Square proved the effectiveness in one-dimensional segmentation problem for linear regression analysis. So we extended the segmented least square method to fit in our matrix segmentation problem.

We used the two-step method in analysing breast cancer patient samples and showed that module structures predicted by our method can be used for patient survival prediction and subtype classification.

The rest of paper is organised as follows. The process of mining functional miRNA regulatory modules is described in Section 3.2 to Section 3.4 and Section 3.1 explains the materials and data we used for analysis. The method is tested on breast cancer data to explain the survival and subtypes in Section 4.

## 3 Methods

### 3.1 Materials and data

Human breast cancer samples were used for the analysis. Expression profiles of both miRNA and mRNA are from TCGA breast cancer level 3 data (Cancer Genome Atlas Network, 2012). To determine functional MRM, we need a template network of miRNA and targets. For the template network construction, i.e., miRNA-mRNA target
Iterative segmented least square method for functional microRNA-mRNA relationship, we used both experimentally validated CLASH (Helwak et al., 2013) data, and a computationally predicted database, doRiNA (Anders et al., 2011) that is an unified databases containing CLASH experiments, TargetScan (Lewis et al., 2005), and PicTar (Kret et al., 2005). We further extended the target relationships by manually curated several databases – PITA (Kertesz et al., 2007), miRTarBase (Hsu et al., 2010), and DIANA-micro-T (Xu et al., 2014; Reczko et al., 2012).

TCGA expression data provide expression profile of total 923 miRNAs. From 923 miRNAs, we excluded miRNAs that have low abundant transcripts with expression level below 1. We also excluded the miRNAs with no putative targets. As a result, 350 miRNAs were used for analysis. In gene expression data, we also excluded the genes that are not expressed or low abundant across whole patient samples. We also excluded the genes that are not targeted by miRNAs. Then we selected genes in the cancer hallmark gene sets (Subramanian et al., 2005). As a result, total 4,149 genes were used for analysis.

We also used GoSemSim (Guangchuang et al., 2010), survival R package (Therneau, 2015), SciPy (Jones et al., 2001), and Cytoscape (Shannon et al., 2003) for further analyses.

Figure 1 Overview of the method (see online version for colours)
3.2 Overview of the method

For the discovery of functional MRMs, we formulated the problem similar to spectral biclustering analyses. This work consists of two sub-problems. First, we arranged miRNAs and mRNAs ordered in rows and columns of a matrix. The ordering of miRNAs and mRNAs reflects the fact that multiple miRNAs or mRNAs work together, rather than individually, to perform biological functions. Then, in the next step, we split the matrix into modules. But unlike typical biclustering analyses that consider only a single matrix, we consider per sample multiple matrices simultaneously. To do so, we developed an iterative optimisation method based on the dynamic programming technique. Figure 1 shows the overview of the proposed method. Details of each step are described in the following subsections.

3.3 Step 1: construction of structured miRNA-mRNA matrices

In order to find a module structure from the miRNA-mRNA matrix, the order of miRNAs and mRNAs in the matrix should be configured. We used dynamic hierarchical clustering method (Langfelder et al., 2008) for miRNA and mRNA configuration as it decides the order of miRNAs and mRNAs in the process. Dynamic hierarchical clustering algorithms were performed on the correlation matrix \( C \) that is defined as below

\[
C_{ij} = \begin{cases} 
0 & \text{if } \text{cor}(i,j) \geq 0 \text{or } p(i,j) = 0 \\
\text{cor}(i,j) & \text{else}
\end{cases}
\]  

(1)

where \( \text{cor}(i,j) \) is Pearson’s correlation coefficient of miRNA \( i \) and mRNA \( j \), and \( p(i,j) \) is an binary indicator variable whether miRNA \( i \) putatively targets mRNA \( j \) or not. Putative target information is from template miRNA-mRNA network described in Materials and Data section. After hierarchical clustering, we can expect that two miRNAs will be ordered closely if two miRNAs targets similar gene sets with similar intensity and two mRNAs are ordered closely if two mRNAs are targeted by similar miRNAs.

After the configuration of the order of miRNAs and mRNAs, matrices per each patient samples need to be defined. For a pair of miRNA \( i \) and mRNA \( j \), matrix element in sample \( s \) is defined to represent the regulating effect of miRNA \( i \) to mRNA \( j \) in sample \( s \). Let \( X = \{X^1, X^2, ..., X^S\} \) be the set of patient matrices where \( S \) is the number of samples and \( X^s \) is a matrix of patient sample \( s \). Then matrix element of \( X^s_{ij} \) is defined as below

\[
X^s_{ij} = \begin{cases} 
0 & \text{if } \text{sgn}(c_{is}) = \text{sgn}(g_{js}) \\
|c_{is}g_{js}| & \text{else if } g_{js} \geq 0 \\
-|c_{is}g_{js}| & \text{else}
\end{cases}
\]  

(2)

where \( c_{is} \) is z-score of expression value of miRNA \( i \) in sample \( s \), \( g_{js} \) is z-score of expression value of mRNA \( j \) in sample \( s \), and \( \text{sgn} \) is sign function. Sign of the value means whether gene is activated or suppressed and high absolute value means high regulating power.
3.4 Step 2: matrix segmentation algorithm

3.4.1 Problem formulation

The following notations will be used for formal description of matrix segmentation method. $B_{iy}^{m}$ is a boundary indicator variable that is set to 1 where miRNA $i$ and miRNA $i+1$ belong to different modules. The value of zero to $B_{iy}^{m}$ means miRNA $i$ and miRNA $i+1$ belongs to the same modules. $B_{iy}^{j}$ is defined similarly for mRNAs.

Then two miRNA boundary variables and two mRNA boundary variables defines a miRNA-mRNA module $M_{ij}^{y}$ where following conditions $B_{iy}^{m} = 1$, $B_{iy}^{j} = 1$, $B_{iy}^{m} = 0$ for all $i < k < i'$, $B_{ij}^{y} = 1$, $B_{ij}^{y} = 1$ and $B_{ij}^{y} = 0$ for all $j < l < j'$ are satisfied. We set a boundary variable to be value of 1 only if a hierarchical cluster is bigger than a certain size at the boundary variable. This way, we consider MRMs only at the putative boundaries, which makes the problem search space much smaller. We used cut-points of hierarchical clusters that is bigger than 10 for gene boundary variables.

3.4.2 Definition of module fitness function

To find an optimal module structure, we need to define a measure to decide whether a given module is appropriate for our module definition or not. We used the sum of squared error averaged over patient samples for a module fitness measure. Given a module $M_{ij}^{y}$, error of the module $E(M_{ij}^{y})$ is defined as following equation (3)

$$E(M_{ij}^{y}) = \frac{1}{S} \sum_{i} \sum_{j} \sum_{k} \sum_{l} (X_{kl}^{y} - \mu_{ij}^{y})^2$$

where $\mu_{ij}^{y}$ is average of $X^{y}$ values in the module $M_{ij}^{y}$ and $S$ is the number of samples.

3.4.3 Recurrence relation

Our goal is to set module boundaries that minimise the sum of the module errors as defined in equation (3). As the effect of each miRNAs to each mRNAs is highly variable, there should be a guidance on the module size and the number of modules. Of course, this is not known in advance. In this situation, a general technique is to introduce a penalty term. To handle the trade-off between the module size and the number of modules, we extended the idea proposed in the segmented least square for linear regression (Bellman, 1961). The final minimizing function is as following equation

$$f(X) = \sum_{i \in D} E(Z) + C_{m}N_{mi} + C_{g}N_{gi}$$

where $Z$ is the set of modules, $C_{m}$ is penalty term for miRNA, $C_{g}$ is penalty term for mRNA, $N_{mi}$ is the number of segments in miRNAs and $N_{gi}$ is the number of segments in mRNAs.

We can get the optimal solution for objective function in equation (4) using the recurrence relation of the solutions between sub-matrices of $X$. Let $opt(i, j)$ be the
optimal solution for sub-matrices of $X$ containing miRNAs 1, 2, ..., $i$ and mRNAs 1, 2, ..., $j$. Then we can easily induce the recurrence relation as below.

$$opt(m, n) = \min_{1 \leq i \leq m, 1 \leq j \leq n} (E_{ij} + C_{ij})$$

Then we can easily induce the recurrence relation as below.

$$1, 1 = (1, 1)_{\min}$$

$$\quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad 

Although we can still get the optimal solution using recurrence relation, it requires $O(m^3 n)$ time complexity and $O(m^2 n^2)$ space complexity where $m$ is the number of miRNAs and $n$ is the number of mRNAs. In our case, $m$ is 350 and $n$ is 4,149, which results in unrealistic computation time and memory demand.

**Algorithm 1: Iterative Segmented Least Square Algorithm**

1. Initialize $B^0$ using hierarchical clustering result
2. $m < \text{number of miRNAs}, n < \text{number of mRNAs}$
3. for $k < 1$, till convergence do
4. for $i' < 1$ to $m$ do
5. compute $E_{i' i}$ under the segment $B^0_k, ..., B^0_n$
6. end for
7. $opt(0) \leftarrow 0$
8. for $i < 1$ to $m$ do
9. $opt(i) \leftarrow 0$
10. for $i < 1$ to $m$ do
11. $opt(i) \leftarrow \min_{1 \leq i' \leq i} (E_{i' i} + C_{i' i})$
12. $bt(i) \leftarrow \arg \min_{1 \leq i' \leq i} (E_{i' i} + C_{i' i})$
13. end for
14. compute $B^m$ using backtrace information $bt$
15. for $j' < 1$ to $n$ do
16. for $j < j'$ to $n$ do
17. compute $F_{j' j}$ under the segment $H^m_{1j'}, ..., H^m_{nj'}$
18. end for
19. end for
20. $opt(0) \leftarrow 0$
21. for $j < 1$ to $n$ do
22. $opt(j) \leftarrow \min_{1 \leq j' \leq j} (E_{j' j} + C_{j' j} + opt(j' - 1))$
23. $bt(j) \leftarrow \arg \min_{1 \leq j' \leq j} (E_{j' j} + C_{j' j} + opt(j' - 1))$
24. end for
25. compute $B^0$ using backtrace information $bt$
26. end for
27. output: $B^m, B^0$

### 3.4.4 An iterative segmented least square method

To overcome the unrealistically high complexity for a genome level problem, $O(m^3 n)$ time complexity and $O(m^2 n^2)$ space complexity, we developed an iterative dynamic
Iterative segmented least square method for functional microRNA-mRNA programming. First, the indicator variables for mRNAs are initialised using hierarchical clustering result. Then the optimal solution for miRNA indicator variables are sought under the fixed mRNA indicator variables using the recurrence relation below:

\[ \text{opt}(i) = \min_{i \leq j \leq i} (E_{ij} + C_{in} + \text{opt}(i' - 1)) \]  

where \( \text{opt}(i) \) is the optimal solution for sub-matrices of \( X \) containing miRNAs 1, 2, ..., \( i \) under fixed mRNA indicator variables, \( E_{ij} \) is sum of errors of the modules between miRNA \( i' \) and miRNA \( i \). Then we fix the miRNA indicator variables and get the optimal mRNA indicator variables under fixed miRNA indicator variables as an equivalent process for mRNA indicator variables. We repeat the process until convergence. The proposed method require \( O(k(m^3 + n^3)) \) time complexity where \( k \) is the number of iteration and \( \max(O(m^2), O(n^2)) \) space complexity. Details of iterative dynamic programming are described in the algorithm 1.

Figure 2  Subtype CART using modular values. We can see that only 17 number of modular values can successfully divide the breast cancer subtypes
4 Results and discussion

4.1 miRNA-mRNA modules can explain survival and subtype of breast cancer

TCGA breast cancer samples of 1,158 patients were analysed. We empirically decided the parameter values of $C_m$ and $C_g$ to be 10 and 15, respectively. After convergence of the algorithm, the rows and columns were divided into 94 and 126 sub-clusters, respectively.

To evaluate the significance of the modules, we first calculated the per-sample scores of each module as an average of $X_{ij}$ values of pairs in each module. The calculated scores were then further for patient survival prediction and subtype classification. As a result, we generated two classification and regression trees (CART) for subtype and survival analyses (see Figure 2 and supplementary figure S1). As shown in the Figure 2, the whole cohort was divided into 17 groups and successfully classified using only 16 modular values with 79.87% of accuracy. In the case of survival analyses, 27 number of modular values divided the whole cohort into 28 subgroups. We also generated Kaplan-Meier plot using only three branches near root node from survival CART (Figure 3). KM plot result shows that patients were successfully divided into subgroups that have distinct survival using only three MRM values with a p-value smaller than 1e-19.

Figure 3 Kaplan-Meier plot using three modules including root module and two children modules of root module. Each lines denote patient groups that fall into different CART branches (see online version for colours)

4.2 Biological role of top modules in the CART

From the subtype and survival CART result, we selected modules of high importance at the top levels of the tree and investigated the biological significance of the modules
Iterative segmented least square method for functional microRNA-mRNA (Figure 4). Fifteen and four modules were selected of high importance for subtype and survival CART, respectively. Interestingly, 12 out of 15 importance modules from subtype CART are regulated by two miRNAs, hsa-miR-766 and hsa-mir-877 while four modules from survival CART have four different miRNAs.

Figure 4 High importance modules for classifying survival or subtype information of breast cancer patients. Rectangles denote miRNAs and circles denote target genes. Above two miRNA modules are merged modules of 15 subtype importance modules and below three modules are 4 survival importance modules.

In subtype related modules, hsa-miR-766 and hsa-miR-877 were selected as a distinct biomarker for breast cancer subtype in our analysis targeting the majority of module genes. Several evidence of miR-766 as cancer subtype-related were suggested (Yang et al., 2015; Tahiri et al., 2014; Li et al., 2015; De Rinaldis et al., 2013). Its expression was identified as down-regulated in both benign and malignant breast cancer cohorts compared to that of the normal samples (Tahiri et al., 2014). Fourteen miRNAs including miR-766 were also investigated by a previous study (De Rinaldis, 2013). The study focused on the prognostic evaluation of the miRNAs related to the both clinical and histopathological features such as ER negativity and Triple-negativity. miR-766 kept its prognostic significance when histopathological information such as lymphocytic infiltration, node positivity, and tumour size, was included in the multivariate prediction model. A further analysis on pathway association revealed that miR-766 was highly correlated with proliferation and EGFR signaling, rather than ECM-related pathways. However, it was overexpressed in human colorectal cancer by regulating SOX6 mRNA abundance (Li et al., 2015). This miR-766 and pathways relationship, validated by both the computational analysis and luciferase assay, reciprocally suggest the expression changes of p21 and cyclin D1, which could result in the promotion of cell proliferation without the anchorage-dependent growth characteristics. Along with hsa-miR-766, previous studies also reported hsa-miR-877 as cancer-subtype biomarker (Ujihira et al., 2015 and Qi et al., 2014). The expression level of miR-877 was significantly down-regulated in tamoxifen-resistant MCF-7 breast cancer cells as compared to standard MCF-7 cells that have no tamoxifen resistance (Ujihira, 2015). In addition, a next-generation sequencing based analysis showed deregulation of miR-877 expression level.
between metastatic melanoma and primary cutaneous melanoma samples (Qi et al., 2014). Also, target genes of two miRNAs mir-766 and mir-877 were enriched in the cancer-related pathways like MAPK-signalling pathway, adipocytokine signalling pathway, apoptosis, p53 signalling pathway, and Jak-STAT signalling pathways. These previous studies showed that deregulation of miRNAs in subtype-module is related with cancer subtypes and malignancy.

In the case of survival related modules, we found five (RHOT1, PRPF4B, MYH10, AKR1C2, CFL1) out of twelve target genes were related to cancer in previous studies. RHOT1 is a member of Rho GTPases that plays as a both tumour suppressor and oncogenes (Arpaia et al., 2012; Zhou et al., 2011). Also, RHOT1 is highly expressed in pancreatic cancer cells than paracancerous tissues and showed that down-regulation of RHOT1 could result in a significant suppression of proliferation and migration in cancer cells (Li et al., 2015). On the contrary, there is also a report that low RHOT1 expression in a pancreatic cancer cell is known to promote lymph node metastasis and associated with worse survival (Jiang et al., 2012). The pre-mRNA splicing factor 4 kinase PRP4KB correlates with HER2 amplification and affect the sensitivity of taxane treatment in breast and ovarian cancer (Corkery et al., 2015). As PRP4KB affects the taxane sensitivity, it also correlates with overall survival in ovarian cancer patients treated with taxanes with low Her2-expressing tumours (Corkery et al., 2015). MYH10 is a gene that encodes myosin II motor proteins and myosin II motor proteins play important roles in cancer cell migration (Betapudi et al., 2011). Progesterone plays an essential role in breast cancer and AKR1C2 is a member of genes that regulates progesterone metabolism. The suppression of AKR1C2 in cancer can result in significant inhibition of cellular proliferation by progesterone in cancer when compared with paired control cell lines (Ji et al., 2004). The activity of coflin (CFL1) modulates actin cytoskeleton, and it triggers tumour cells to be able to generate early metastasis. As coflin is associated with tumour invasion, high CFL1 expression levels present low overall survival rates in lung cancer patients (Castro et al., 2010). All these previous studies demonstrate that genes in survival related modules predicted in our CART analysis are correlated with prognosis of cancer cells.

4.3 Comparison and validation of module structure

To evaluate how good the module structures predicted in our study, we calculated the GO similarity of the module genes and our method was compared with a standard spectral biclustering algorithm (Kluger et al., 2003) implemented in SciPy library (Jones et al., 2001), and the leading MRM module discovery method, Mirsynergy (Li et al., 2014). GO similarity of each gene pairs within each module was calculated using Wang’s method (Wang et al., 2007) by R GoSemSim package. Then we calculated the average GO similarity score of gene pairs in each module to evaluate the modularity. Table 1 shows a comparison of module GO similarity values using our modular structure and the result of competing methods. The result shows that our modular structure has higher functional similarities between module members than those predicted by biclustering and Mirsynergy methods.
Iterative segmented least square method for functional microRNA-mRNA

Table 1  A table of comparing GO similarity scores of the proposed method, biclustering method, and Mirsynergy. Our method shows the best results in terms of 1st quantile, median, mean, and 3rd quantile.

<table>
<thead>
<tr>
<th>Methods</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean ± Std.</th>
<th>3rd Qu.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed Method</td>
<td>0.4937</td>
<td>0.5313</td>
<td>0.5341 ± 0.054</td>
<td>0.5722</td>
</tr>
<tr>
<td>Biclustering</td>
<td>0.4905</td>
<td>0.5234</td>
<td>0.5202 ± 0.077</td>
<td>0.5506</td>
</tr>
<tr>
<td>Mirsynergy</td>
<td>0.4841</td>
<td>0.5164</td>
<td>0.5140 ± 0.053</td>
<td>0.5558</td>
</tr>
</tbody>
</table>

Table 2  A table of inter and intra-cluster similarity scores of the proposed method, biclustering method, and Mirsynergy. Our method outperforms others in terms of both inter and intra-cluster similarities.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Inter-cluster similarity</th>
<th>Intra-cluster Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed Method</td>
<td>0.1065</td>
<td>0.0106</td>
</tr>
<tr>
<td>Biclustering</td>
<td>0.0133</td>
<td>0.0121</td>
</tr>
<tr>
<td>Mirsynergy</td>
<td>0.0685</td>
<td>0.0663</td>
</tr>
</tbody>
</table>

Inter and intra-cluster similarity is also calculated for the comparison of module results. Inter and intra-cluster similarity are calculated as average Pearson’s correlation coefficients of gene pairs that belong to the same cluster or different clusters, respectively. Bigger inter-cluster similarity and smaller intra-cluster similarity indicate better clustering results. Table 2 shows that our method outperforms the two methods in terms of both inter and intra-cluster similarity.

In addition, we also performed the cross-validation analysis of survival outcomes to show the robustness of our miRNA-mRNA modular structure. Area under curve (iAUC) on the receiver operating characteristic curve were calculated after ten cross-validation CART process. The high average value 0.88, in the 10-fold cross-validation iAUC result indicates the robustness of the miRNA-mRNA modular structures that our method determined.

Additionally, we compared the module size in terms of the number of genes in each module. Table 3 shows size distribution of modules by the three methods. Mirsynergy result contains only a small number of genes compared to results by the proposed method and the biclustering method. Also, the size distribution of biclusters is highly skewed than our result. In a comparison of our method and the biclustering method in terms of the skewness of the module size distribution, our method produced much tighter distribution.

Table 3  A table of comparing cluster sizes of the proposed method, biclustering method, and Mirsynergy. Our method produced tighter distribution than biclusters while containing more number of genes than the result produced by Mirsynergy.

<table>
<thead>
<tr>
<th>Methods</th>
<th>min</th>
<th>Median</th>
<th>max</th>
<th>Std.</th>
<th>Gene number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed Method</td>
<td>6</td>
<td>21</td>
<td>230</td>
<td>32.8</td>
<td>4149</td>
</tr>
<tr>
<td>Biclustering</td>
<td>1</td>
<td>16</td>
<td>519</td>
<td>58.48</td>
<td>4149</td>
</tr>
<tr>
<td>Mirsynergy</td>
<td>5</td>
<td>15</td>
<td>37</td>
<td>8.96</td>
<td>194</td>
</tr>
</tbody>
</table>
5 Conclusion

We presented a novel two-step algorithm to determine MRM module structure from miRNA and gene expression profiles. Our method first arranges a template miRNA-mRNA network into a matrix based on hierarchical clustering to incorporate co-operative characteristics of miRNAs and genes. Then an iterative segmented least square optimisation algorithm was designed and used to determine MRMs.

As shown in the analysis of TCGA breast cancer data sets, our method was able to detect MRMs that were known to be closely related to tumour malignancy and subtypes. Our method outperformed the biclustering algorithm and the state of the art MRM discovery method, Mirsynergy, in terms of GO functional similarity and clustering metrics. Based on our extensive analysis, we believe that the proposed method is effective in determining module structure from miRNA and mRNA expression data.

Acknowledgements

This research was supported by a BK21 Plus for Pioneers in Innovative Computing (Dept. of Computer Science and Engineering, SNU) funded by National Research Foundation of Korea (NRF) (21A20151113068), Collaborative Genome Program for Fostering New Post-Genome industry through the National Research Foundation of Korea (NRF) funded by the Ministry of Science ICT and Future Planning (NRF-2014M3C9A3063541), and a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI14C3405).

References


Iterative segmented least square method for functional microRNA-mRNA


