Master's Thesis

Inferring the Human microRNA Functional Similarity in Latent Factor Space

Jinha Jeong (정 진 하)

Department of Computer Science and Engineering

Pohang University of Science and Technology

2016


POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

잠재 요소 공간상에서 인간 마이크로RNA의 기능적 유사도 추론

Inferring the Human microRNA Functional Similarity in Latent Factor Space

Inferring the Human microRNA Functional Similarity in Latent Factor Space

by

Jinha Jeong
Department of Computer Science and Engineering
Pohang University of Science and Technology

A thesis submitted to the faculty of the Pohang University of Science and Technology in partial fulfillment of the requirements for the degree of Master of Science in the Computer Science and Engineering

Pohang, Korea
06. 09. 2016
Approved by
Hwanjo Yu (Signature)
Academic advisor

Inferring the Human microRNA Functional Similarity in Latent Factor Space

Jinha Jeong

The undersigned have examined this thesis and hereby certify that it is worthy of acceptance for a master's degree from POSTECH

06. 09. 2016

Committee Chair	Hwanjo Yu	(Seal)
Member	Joon Hee Han	(Seal)
Member	Sanguk Kim	(Seal)

MCSE 정 진 하. Jinha Jeong
20142132
Inferring the Human microRNA Functional Similarity in Latent Factor Space,
잠재 요소 공간상에서 인간 마이크로RNA의 기능적 유사도 추론
Department of Computer Science and Engineering , 2016, 32p, Advisor : Hwanjo Yu. Text in English.

ABSTRACT

MicroRNAs (miRNAs) are small non-coding RNA molecules that have important functions in many biological processes. Unfortunately, there are no functional annotations for most miRNAs. Thus, to predict the potential functions of miRNAs, the strategy of inferring functional similarities among them has gained attention. The existing methods measure the functional similarities by applying a *best-match average strategy*, which based on "miRNA – related object information" and "between – object information". However, none of them take account of *transitive relations* among miRNAs, or integrate several types of information harmoniously. This paper proposes a new model, MFSim, based on *Matrix Factorization* (MF). By mapping miRNAs and related objects to a *joint latent factor space*, the model can take account of the transitive relations among miRNAs. The experimental results show that the proposed model is superior to a comparison method in all evaluation methods. The model is also flexible; it can integrate additional information harmoniously by mapping it in a shared latent factor space.

This characteristic will make the model easy to be extended further.


POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교

Contents

List of Tables

List of Figures

I. Introduction

MicroRNAs (miRNAs) are small non-coding RNA molecules that are about 22 nucleotides in length and are evolutionarily conserved [1]. These molecules are involved in post-transcriptional regulation by binding to complementary regions of miRNA [1]. Since their existence was revealed in 1993 [2], the number of discovered miRNAs has increased dramatically [3]. miRNAs have important functions in many biological processes including development [4], proliferation [5], apoptosis [6] and immune system regulation [7].

Unfortunately, the functions of only few miRNAs are partially known and most lack functional annotations [8]. Thus, to predict their potential functions, the strategy of inferring functional similarities among them has gained attention [9]. Because of the lack of miRNA functional annotation, one can infer the similarities indirectly with a set of "miRNA – related object information". One method [10] infers the similarities based on "miRNA – disease association information"; another [11] uses "miRNA – target gene information". Another method uses both information and produces weighted sum of the each functional similarity [8]. Among several methods, the method based on "miRNA – disease association information" (MISIM) [10] is widely used due to its reliability and usefulness.

Although the existing methods are widely used and successfully infer functional similarities, the methods share two obvious limitations. First, all the existing methods cannot take account of *transitive relations* among miRNAs, in inferring the functional similarities [8, 9, 10, 11], i.e., if miRNAs m_x and m_z are related and m_z and m_y are also related, then the miRNAs m_x and m_y may also have some relation through those relations. Furthermore, not only by m_z,

the relation between m_x and m_y can be influenced by a set of other miRNAs, which related to both m_x and m_y. The existing methods cannot basically utilize such relationship, because they measure the similarities by applying a *best-match average strategy* only between two target miRNAs [8, 9, 10, 11].

Another limitation is that, to the best of our knowledge, no method can integrate several types of information harmoniously. For example, MISIM [10] only utilizes disease related information. Thus, functional similarities between a miRNA and the existing miRNAs cannot be inferred by MISIM unless any association between the miRNA and its related disease is discovered. The problem is that discovering "miRNA – disease associations" experimentally requires much time and effort. A method in [8] utilizes both disease related information and miRNA's target genes information, but the method simply computes the weighted sum of two similarities, which are independently inferred based on each type of information.

Here, we propose a novel *Matrix Factorization* (MF) based model, MFSim, which is a latent factor model. The core concept of the latent factor model is that only a small number of latent factors characterize latent patterns of an entity, or correspond to coefficients of the patterns [12]. Among many latent factor models, MF is widely used in various applications and is a very active research area in machine learning community. Applications include document clustering [13], bioinformatics [14] and recommender system [15].

The proposed model has several strengths in comparison with the existing methods. By mapping miRNAs and related objects to a *joint latent factor space*, the model can take account of the transitive relations naturally. It has this ability because latent factor vectors for miRNAs and related objects are shared and learned simultaneously, while the associations among them are constructed. In addition, the model is flexible; the model can integrate additional information by

mapping it in the shared latent space, which facilitates us to extend the model easily. Above all, the functional similarities inferred by our model are more reliable than those inferred by a compared method in all evaluation methods.

II. Materials and Methods

2.1 Human miRNA – disease association

Our basic essential data comes from a human "miRNA – associated disease database" [16], HMDD v2.0, and we represent the data as a binary matrix $\mathbf{R}$. The data was released June 20, 2013 and contains experimentally-supported "miRNA – disease association information".

We preprocess the data similar to MISIM [10]. First, we de-duplicate the initial $10,381$ records into $6,448$ unique pre-miRNA – disease records, including 578 pre-miRNAs and 380 diseases. Then we adjust the disease name in HMDD if they differ from standard MeSH disease terms. Different pre-miRNAs that produce the same mature-miRNA are treated as single pre-miRNAs. For example, both has-mir-376a-1 and has-mir-376a-2 are treated as has-mir-376a. The number of final miRNA – disease pairs is $5,401$, including 495 merged miRNA and 362 diseases. We denote the miRNAs as a set $M = \{m_1, m_2, \ldots, m_{495}\}$, and the diseases as a set $D = \{d_1, d_2, \ldots, d_{362}\}$.

This information can be represented in a binary matrix $\mathbf{R} = (r_{ij})_{|M|\times|D|} \in \{0,1\}^{|M|\times|D|}$ (Table 1), where $|M|$ is the number of miRNAs and $|D|$ is the number of diseases. The i^{th} row of $\mathbf{R}$ corresponds to miRNA m_i and the j^{th} column of $\mathbf{R}$ corresponds to disease d_j. An element $r_{ij} \in \mathbf{R}$ is 1 if m_i is associated with d_j, and 0 otherwise. Factorizing $\mathbf{R}$ yields miRNA matrix $\mathbf{M}$ that contains the latent factor vectors for the miRNAs, and disease matrix $\mathbf{D}$ that contains the latent factor vectors for the diseases.

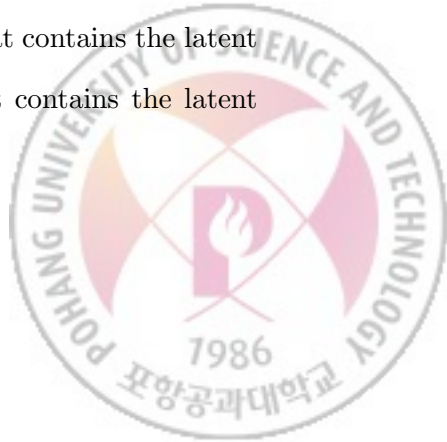


Table 1: Basic statistics of matrix $\mathbf{R}$

Proportion of 1	5401 / 179190 **(3.014%)**
Average # of related diseases	10.911 / a miRNA
Average # of related miRNAs	14.920 / a disease

2.2 Semantic similarities among diseases

We use semantic similarities among diseases based on the Medical Subject Headings (MeSH) database as an additional information when factorizing $\mathbf{R}$. The database comes from the National Library of Medicine (`http://www.nlm.nih.gov/`) and it provides a hierarchically-organized terminology for indexing and cataloging of biomedical information. One can measure the semantic similarity between two diseases by using this information, and we reproduce exactly the same procedure as MISIM [10].

The semantic similarities among diseases can be represented in a similarity matrix $\mathbf{S} = (s_{jl})_{|D|\times|D|} \in [0,1]^{|D|\times|D|}$ (Table 2), where $s_{jl} = S(d_j, d_l)$ is a semantic similarity between diseases d_j and d_l.

Table 2: The number of elements in **S** for each range of values

	Range of values	# of elements
# of 0 elements	0 (Nonexistent disease in MeSH)	7030 (10.759%)*
(78.473%)	0 (No common DAG b/w diseases)	44245 (67.71%)
# of non-0 elements (21.527%)	<0.1	9999 (15.303%)
	<0.2	2202 (3.370%)
	<0.3	859 (1.315%)
	<0.4	446 (0.683%)
	<0.5	224 (0.343%)
	<0.6	133 (0.204%)
	<0.7	98 (0.150%)
	<0.8	64 (0.098%)
	<0.9	41 (0.063%)

*Proportion of associated elements per total # of elements (65341) in **S**

2.3 MISIM and the best-match average strategy

MISIM [10]is the one of the most successful algorithms to infer functional similarity; the functional similarities inferred by MISIM are widely used in many miRNA related applications. The method is based on "miRNA – disease association information" and "disease semantic similarity information", which is derived from MeSH descriptors. This method also considers the contributions from similar diseases that are associated with two different miRNAs, respectively [10]. To be precise, first, the maximum similarity between one disease 'dt' and one disease group 'DT', $S(dt, DT)$, is defined as:

$$S(dt, DT) = \max_{1 \le i \le k} \left(S(dt, dt_i) \right), \tag{1}$$

where $DT = \{dt_1, dt_2, \ldots, dt_k\}$ is a set of diseases and $S(dt, dt_i)$ is the semantic similarity between diseases dt and dt_i. Then, the functional similarity of two

miRNAs m_x and m_y inferred by MISIM, $MISIM\,(m_x, m_y)$, is defined as:

$$MISIM\,(m_x, m_y) = \frac{\sum_{1\le i\le m} S\,(dt_i^x, DT_y) + \sum_{1\le j\le n} S\left(dt_j^y, DT_x\right)}{m+n} \tag{2}$$

where $DT_x = \{dt_1^x, dt_2^x, \ldots, dt_m^x\}$ is a set of diseases, which are related with m_x, and $DT_y = \{dt_1^y, dt_2^y, \ldots, dt_n^y\}$ is a set of diseases, which are related with m_y. m and n are the size of the sets DT_x and DT_y, respectively. In summary, the method calculates the average of the maximum similarities between each related disease of one miRNA and each related disease group of another miRNA.

This strategy, so-called *best-match average strategy*, is not only used in MISIM, but also used in the other existing methods [8, 9, 11] with different data. However, the strategy has several limitations; it utilizes only one "miRNA – related object information" (e.g., miRNA – disease association information), necessarily requires "between – object information" (e.g., disease semantic similarity information) and cannot take account of transitive relations at all.

2.4 Proposed model (MFSim)

2.4.1 Objective function and confidence level

We use *Matrix factorization* (MF) technique to map miRNAs and diseases to a k-dimensional joint latent factor space. Specifically, given a $|M| \times |D|$ binary matrix $\mathbf{R}$, a low-rank MF approach seeks to approximate the matrix $\mathbf{R}$ as a product of k-rank factors,

$$\mathbf{R} \approx \mathbf{M}^T\mathbf{D}, \tag{3}$$

where $\mathbf{M} \in \mathbb{R}^{k\times|M|}$ and $\mathbf{D} \in \mathbb{R}^{k\times|D|}$. Here, i^{th} column vector $\mathbf{m}_i \in \mathbb{R}^k$ of $\mathbf{M}$ corresponds to latent factors of miRNA m_i, and j^{th} column vector $\mathbf{d}_j \in \mathbb{R}^k$ of $\mathbf{D}$ corresponds to latent factors of disease d_j. Given $r_{ij} \in \mathbf{R}$, to learn the parameters

$\mathbf{m}_i$ and $\mathbf{d}_j$, we introduce an objective function,

$$\min_{\mathbf{M},\mathbf{D}} \frac{1}{2} \left\{ \sum_{i=1}^{|M|} \sum_{j=1}^{|D|} c_{ij} \left(r_{ij} - \mathbf{d}_j^T \mathbf{m}_i \right)^2 + \lambda_1 \left\| \mathbf{M} \right\|_F^2 + \lambda_2 \left\| \mathbf{D} \right\|_F^2 \right\}. \tag{4}$$

The optimization problem in eq. (4) minimizes the sum-of-squared-error on every element r_{ij}, with two regularization terms. This objective function will be extended in section 2.5.

Here, c_{ij} is a confidence level and has an important meaning. To be precise, we consider $r_{ij} \in \mathbf{R}$. A value of $r_{ij} = 1$ implies that miRNA m_i is associated with disease d_j; i.e., we are highly confident in the association. However, $r_{ij} = 0$ can mean either miRNA m_i is not associated with disease d_j, or the association has not been identified yet; i.e., we are not as confident in the association as we were in the case of $r_{ij} = 1$. For this reason, we assign different confidence levels depending on the value of r_{ij};

$$c_{ij} = \begin{cases} 1, & \text{if } r_{ij} = 1 \\ w, & \text{otherwise} , \end{cases} \tag{5}$$

where $0 \leq w \leq 1$. w is a hyper-parameter that should be tuned according to the data.

2.4.2 Interpretation of the model

In our model, miRNAs and diseases are mapped to a joint latent factor space of dimensionality k, where *miRNA – disease interactions* are modeled as inner products of corresponding vectors. For a given disease d_j, the elements of $\mathbf{d}_j$ measure the extent to which the disease possesses that factor positively or negatively. For a given miRNA m_i, elements of $\mathbf{m}_i$ measure the influence of a corresponding latent factor of disease d_j. Consequently, $\mathbf{d}_j^T \mathbf{m}_i$ captures the interaction between miRNA m_i and disease d_j; this interaction implies the strength of the association between the miRNA and the disease.

2.4.3 Inferring functional similarities

The functional similarity between miRNAs m_x and m_y can be directly measured by *cosine similarity* between $\mathbf{m}_x$ and $\mathbf{m}_y$. According to the low-rank approximation approach, $\mathbf{d}_j^T \mathbf{m}_i$ is close to $r_{ij} \in \mathbf{R}$, which indicates the association between a miRNA m_i and a disease d_j. This means that if two vectors $\mathbf{m}_x$ and $\mathbf{m}_y$ point to a similar direction in latent space, the distributions of "miRNA – disease associations" for the miRNAs m_x and m_y show analogous form; i.e., if $\mathbf{m}_x$ and $\mathbf{m}_y$ are close in cosine similarity, then the miRNAs m_x and m_y have strong correlation in terms of "miRNA – disease association". Furthermore, the functional similarities among miRNAs can be inferred by considering related diseases [10]. Consequently, if $\mathbf{m}_x$ and $\mathbf{m}_y$ are close in cosine similarity, we can say that the miRNAs m_x and m_y are close in functional similarity as well. We measure cosine similarities among all possible miRNA pairs from matrix $\mathbf{M}$, then use these cosine similarities to quantify functional similarities.

However, because the range of cosine similarity is $[-1, 1]$ and the range of functional similarity is $[0, 1]$, some technique is required to scale from one to another. We avoid this problem by simply converting negative similarities to 0, regardless of scaling. We will discuss about this problem in section 4.1.

2.5 Extension of the model

2.5.1 Additional term for integrating disease semantic similarity information

In this section, we integrate "disease semantic similarity information", which comes from the MeSH descriptor, into the model. If diseases d_j and d_l have a large semantic similarity s_{jl}, then intuitively their characteristics are likely to be similar. Moreover, according to the interpretation in section 2.4.2, each disease can be characterized by its associated latent factor vector. Thus we can

think that if s_{jl} is large, each factor of $\mathbf{d}_j$ tends to be directly proportional to a corresponding factor of $\mathbf{d}_l$. In contrast if s_{jl} is small, $\mathbf{d}_j$ and $\mathbf{d}_l$ may have little correlation. Because of this argument, we try to minimize the difference between s_{jl} and the cosine similarity between two vectors $\mathbf{d}_j$ and $\mathbf{d}_l$:

$$\sum_{j=1}^{|D|} \sum_{l=j+1}^{|D|} e^{s_{jl}} \left(s_{jl} - \frac{\mathbf{d}_j \cdot \mathbf{d}_l}{\|\mathbf{d}_j\|_2 \|\mathbf{d}_l\|_2} \right)^2, \tag{6}$$

where e is the Euler's number (exponential growth constant) and $e^{s_{jl}}$ is a weight that exponentially amplifies large-similarity relations; this is because the number of semantic similarities grows exponentially as their values decrease (Fig. 1). Here the scale difference problem occurs again, but we avoid the problem by directly mapping the cosine similarity between two vectors to range of $[0, 1]$, which is half of the range of cosine similarity.

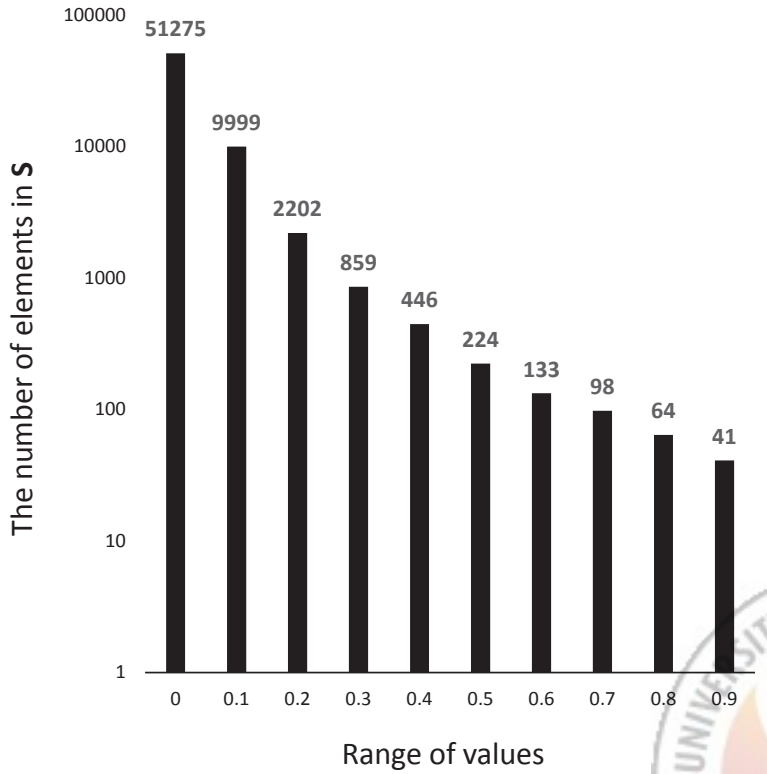


Figure 1: The number of elements in **S** for each range of values

We get the final objective function from eq. (4) and (6) as

$$\min_{\mathbf{M},\mathbf{D}} \frac{1}{2} \left\{ \sum_{i=1}^{|M|} \sum_{j=1}^{|D|} c_{ij} \left(r_{ij} - \mathbf{d}_j^T \mathbf{m}_i \right)^2 + \alpha \sum_{j=1}^{|D|} \sum_{l=j+1}^{|D|} e^{s_{jl}} \left(s_{jl} - \frac{\mathbf{d}_j \cdot \mathbf{d}_l}{\|\mathbf{d}_j\|_2 \|\mathbf{d}_l\|_2} \right)^2 + \lambda_1 \|\mathbf{M}\|_F^2 + \lambda_2 \|\mathbf{D}\|_F^2 \right\}, \quad (7)$$

where α is an important hyper-parameter that controls the influence of the "disease semantic similarity information". For example, "disease semantic similarity information" is never included in the model if $\alpha = 0$. α should be tuned according to the data.

In summary, we factorize matrix $\mathbf{R}$ and produce two matrices $\mathbf{M}$ and $\mathbf{D}$, while cosine similarities among column vectors in $\mathbf{D}$ approximate to their corresponding disease semantic similarities. These two learning processes are carried out simultaneously and we infer the miRNA functional similarities by measuring cosine similarities among column vectors in $\mathbf{M}$, after the learning processes are complete (Fig. 2).

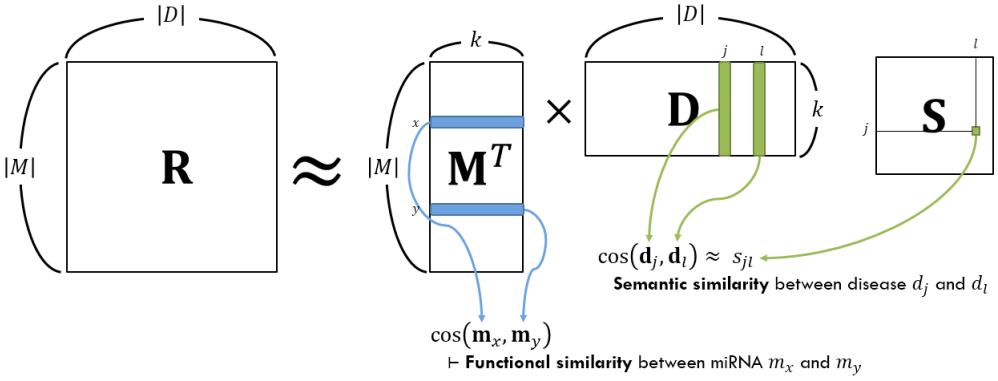


Figure 2: Illustration of overall processes in MFSim (explained in the text)

2.5.2 Update procedure

A local minimum of the objective function given by eq. (7) can be found by performing gradient descent in latent factor vectors $\mathbf{d}_j$ and $\mathbf{m}_i$, respectively:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{m}_i} = \sum_{j=1}^{|D|} c_{ij} \left(\mathbf{d}_j^T \mathbf{m}_i - r_{ij} \right) \mathbf{d}_j + \lambda_1 \mathbf{m}_i, \tag{8}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{d}_j} = \sum_{i=1}^{|M|} c_{ij} \left(\mathbf{d}_j^T \mathbf{m}_i - r_{ij} \right) \mathbf{m}_i + \lambda_2 \mathbf{d}_j + \alpha \sum_{l=j+1}^{|D|} \left\{ e^{s_{jl}} \left(s_{jl} - \frac{\mathbf{d}_j \cdot \mathbf{d}_l}{\|\mathbf{d}_j\|_2 \|\mathbf{d}_l\|_2} \right) \left(-\frac{1}{\|\mathbf{d}_l\|_2} \right) \left(\frac{d_l}{\|\mathbf{d}_j\|_2} - \frac{\mathbf{d}_j \cdot \mathbf{d}_l}{\|\mathbf{d}_j\|_2^3} \mathbf{d}_j \right) \right\}. \tag{9}$$

However, the update procedure on each latent factor vector should be performed carefully because of the cosine similarity term in eq. (6). Since the latent factor vectors for diseases are intertwined with cosine similarity constraint in our model, we updated the vectors of $\mathbf{M}$ and $\mathbf{D}$ sequentially as follows;

$i = 1$ to $|M|$:

$$\mathbf{m}_i \leftarrow \mathbf{m}_i + \eta \left\{ \sum_{j=1}^{|D|} c_{ij} \left(r_{ij} - \mathbf{d}_j^T \mathbf{m}_i \right) \mathbf{d}_j - \lambda_1 \mathbf{m}_i \right\},$$

$j = 1$ to $|D|$:

$$\mathbf{d}_j \leftarrow \mathbf{d}_j + \eta \left\{ \sum_{i=1}^{|M|} c_{ij} \left(r_{ij} - \mathbf{d}_j^T \mathbf{m}_i \right) \mathbf{m}_i - \lambda_2 \mathbf{d}_j + \alpha \sum_{l=j+1}^{|D|} \left\{ e^{s_{jl}} \left(s_{jl} - \frac{\mathbf{d}_j \cdot \mathbf{d}_l}{\|\mathbf{d}_j\|_2 \|\mathbf{d}_l\|_2} \right) \frac{1}{\|\mathbf{d}_l\|_2} \left(\frac{d_l}{\|\mathbf{d}_j\|_2} - \frac{\mathbf{d}_j \cdot \mathbf{d}_l}{\|\mathbf{d}_j\|_2^3} \mathbf{d}_j \right) \right\} \right\}.$$

III. Result

Direct measurement and comparison of the accuracy of inferred functional similarities among miRNAs are not feasible because functional annotations for miRNAs do not exist, and the ground truth of the functional similarities has not been established. Instead, the *reliability* of the functional similarities can be measured indirectly by using four different evaluation methods, which are based on miRNA family, miRNA cluster, miRNA expression similarity and the fraction of common target genes, respectively [10].

3.1 Comparison with MISIM

We compare the reliability of MFSim only with MISIM [10] because we use human "miRNA – disease association information" and "diseases semantic similarity information". Other methods are not our competitors because they all use similar inferring strategy with MISIM (Eq. 2), while using other data sources [8, 9, 11]. Experimental results of MFSim are derived from a representative similarity matrix, which is obtained by averaging a hundred trials and by cutting off negative similarities. In experiments, we fix parameters to $k = 35$, $\eta = 0.03$, $w = 0.5$, $\alpha = 1$, $\lambda_1 = 0.01$ and $\lambda_2 = 1$. The results show that MFSim is superior to MISIM in all evaluation methods.

3.1.1 Similarity distribution

Before comparing the reliability of functional similarities inferred by MISIM and MFSim, we first compare similarity distributions of functional similarities inferred by MISIM and MFSim. Both distributions show that miRNA pairs with low functional similarity outnumber miRNA pairs with high functional similarity

(Fig. 3). The biggest difference between the two distributions is the proportion of 0 entries.

The proportion of 0 entries is $\sim$10.4% in the similarity distribution inferred by MISIM, but $>$ 41% in the similarity distribution inferred by MFSim. This difference comes from the similarity transformation scheme introduced in section 2.4.3, which simply cuts off negative similarities and converts to 0. The similarity distribution largely depends on various similarity transformation schemes and we expect that the similarities inferred by MFSim, whose negative similarities are cut off, give us compact but informative relationships among miRNAs.

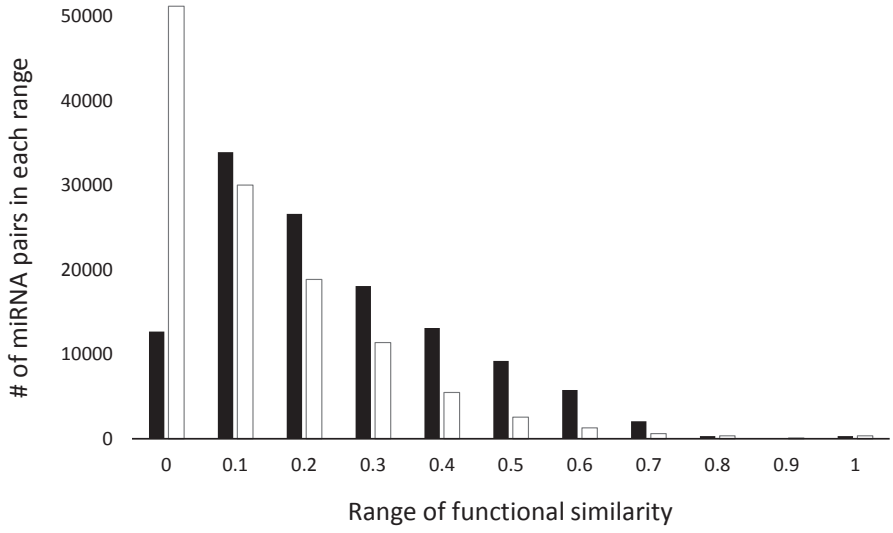


Figure 3: Similarity distributions inferred by MISIM (black bar) and MFSim (white bar)

3.1.2 miRNAs in the same family or cluster show high functional similarity

miRNA family is a set of miRNAs that have similar mature miRNA sequences and completely identical seed regions. Therefore, miRNAs of the same family tend to show high functional similarities among them [10]. Thus, we

can say that the reliability of functional similarities increases as the difference between within-family similarity and between-family similarity increases. We download miRNA family data from TAM [17] and investigate the reliability of functional similarities by calculating the difference between within-family similarity and between-family similarity inferred by MISIM and MFSim, respectively (Table 3). MFSim shows larger difference between within-group similarity and between-group similarity than MISIM.

Table 3: Within-family similarity, between-family similarity and their difference

	MISIM	MFSim
Within-family similarity	0.42771	0.32646
Between-family similarity	0.21164	0.08844
Difference	**0.21607**	**0.23802**

miRNA cluster is another type of miRNA set that is defined according to genome location. A cluster of miRNAs is usually transcribed and expressed synchronously, and functions in coordination. Thus, as is the case for a miRNA family, miRNAs of the same cluster also tend to show high functional similarities among them [10]. We download miRNA cluster data from TAM [17] and investigate the reliability of functional similarities by calculating the difference between within-cluster similarity and between-cluster similarity inferred by MISIM and MFSim, respectively (Table 4). Again, MFSim shows larger difference between within-group similarity and between-group similarity than MISIM.

Table 4: Within-cluster similarity, between-cluster similarity and their difference

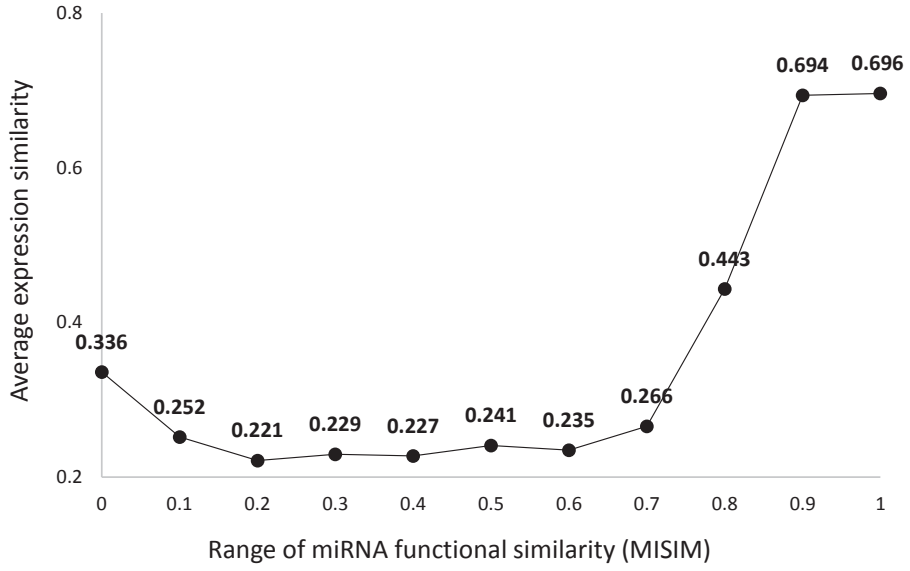
	MISIM	MFSim
Within-cluster similarity	0.38644	0.29670
Between-cluster similarity	0.21470	0.09259
Difference	**0.17174**	**0.20411**

3.1.3 miRNA functional similarity is correlated with expression similarity

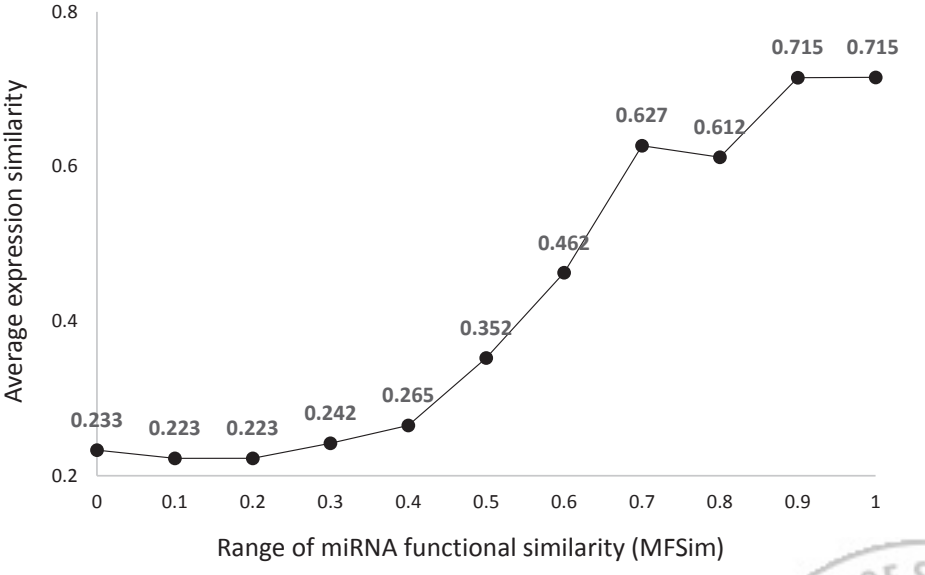
Since miRNAs with similar functions tend to have similar expression profiles [10], we investigate how miRNA functional similarity inferred by MISIM and MFSim is related to expression similarity. We obtain the miRNA expression data from 40 normal tissues [18], which is also used in [10] to validate the correlation between miRNA functional similarities inferred by MISIM and miRNA expression similarities.

First, we use absolute Pearson's Correlation Coefficient (PCC) as the measure for expression similarities among miRNAs. Then, we calculate PCC between the expression similarities and the functional similarities inferred by MISIM. We also calculate PCC between the expression similarities and the functional similarities inferred by MFSim and compare them. We observe that the correlation ($r = 0.15195$, PCC) between the functional similarities inferred by MFSim and the expression similarities is stronger than correlation ($r = 0.02335$, PCC) between the functional similarities inferred by MISIM and the expression similarities.

We further group miRNA pairs into 10 different groups according to their functional similarities in steps of 0.1 and calculate average expression similarities in each group for MISIM and MFSim, respectively (Fig. 4). Clearly, the functional similarities inferred by MISIM and MFSim are correlated with the expression similarities. However, the functional similarities inferred by MFSim are more correlated ($r = 0.95703$, PCC, Fig. 4B) with the expression similarities than those inferred by MISIM ($r = 0.73653$, PCC, Fig. 4A).



(A) MISIM



(B) MFSim

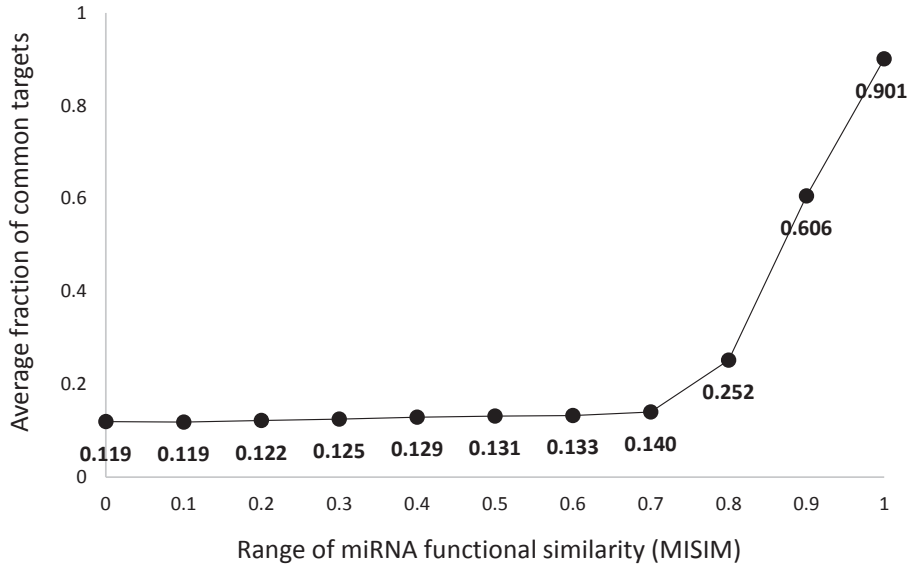
Figure 4: Average expression similarity of each miRNA pair group according to their functional similarities inferred by (A) MISIM and (B) MFSim

3.1.4 miRNA functional similarity is correlated with the fraction of their common targets

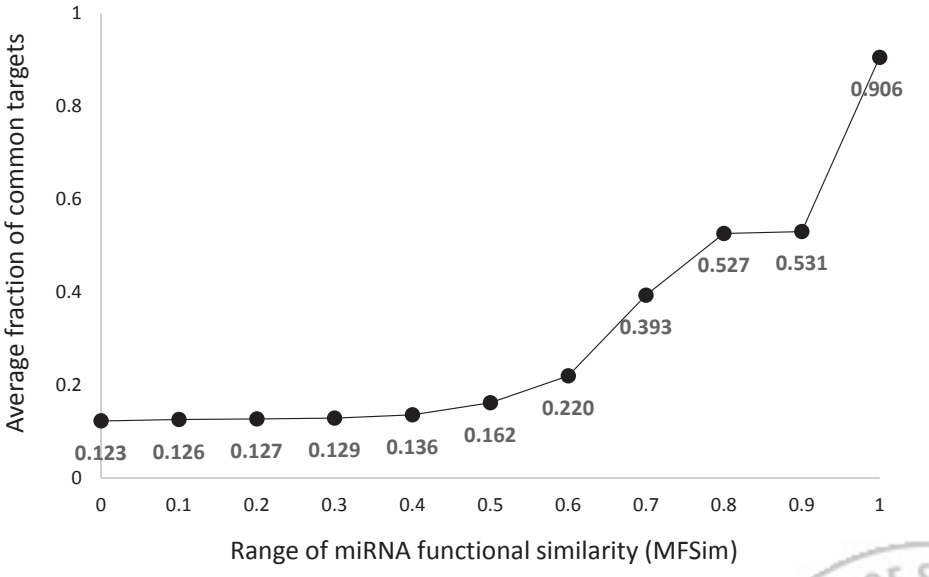
Since miRNAs manifest themselves their functions by regulating target genes, miRNA pairs that have a larger fraction of common target genes are expected to have higher functional similarities [10]. However, only few target genes of miRNAs are experimentally supported so we decided to use miRNA targets predicted by *in silico* prediction. Among several algorithms for predicting the targets, we utilize the target genes predicted by miRanda [19], which is available online (`http://www.microrna.org/microrna/home.do`). We download the predicted target sites of human miRNAs, which are conserved and show good mirSVR scores [19].

Regardless of the functional similarity inferring algorithm, miRNA pairs which have a large fraction of common targets also show high functional similarities. In detail, we observe that correlation ($r = 0.17839$, PCC) between the functional similarities inferred by MFSim and the fraction of common targets is stronger than correlation ($r = 0.1113$, PCC) between the functional similarities inferred by MISIM and the fraction of common targets.

We also group miRNA pairs into 10 different groups according to their functional similarities in steps of 0.1 and calculate the average fraction of common targets in each group for MISIM and MFSim, respectively (Fig. 5). Both graph show positive correlation between the functional similarities and the fraction of common targets, but the functional similarities inferred by MFSim are more correlated ($r = 0.86651$, PCC, Fig. 5B) with the fraction of common targets than those inferred by MISIM ($r = 0.72492$, PCC, Fig. 5A).



(A) MISIM



(B) MFSim

Figure 5: Average fraction of common targets of each miRNA pair group according to their functional similarities inferred by (A) MISIM and (B) MFSim

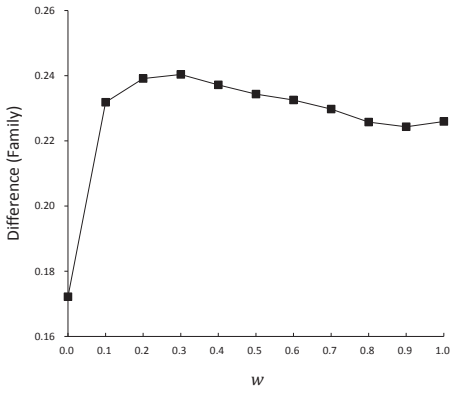
3.2 Impact of hyper-parameters

Besides the comparison between MISIM and MFSim, we also investigate the impact of hyper-parameters, especially w, α and s_{jl}. We vary one specific parameter in 3.2.1 and 3.2.2, while other parameters are fixed to $k = 35$, $\eta = 0.03$, $w = 0.5$ (in section 3.2.2), $\alpha = 1$ (in section 3.2.1), $\lambda_1 = 0.01$ and $\lambda_2 = 1$. Every result is the average of a hundred trials.

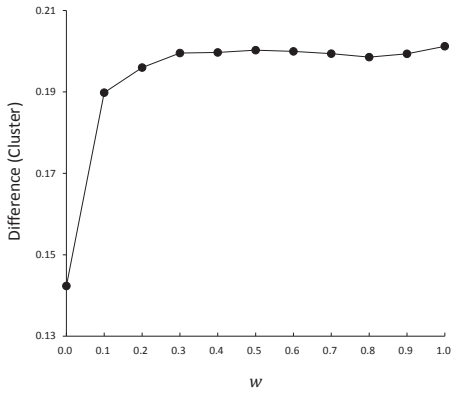
3.2.1 $w\,(0 \leq w \leq 1)$

w is the confidence level for unobserved associations between miRNAs and diseases (Eq. 5). In section 2.4.1, we introduced the parameter because we were not as confident in the unobserved associations as we were for the case of observed associations between miRNAs and diseases. Thus, we anticipate that the results would not be the best in the case of $w = 1$. In contrast, we also anticipate that the results would not be the best in the case of $w = 0$, because some unobserved associations are likely to be truly non-existent associations, and to ensure good results, such associations should not be ignored. We vary w from 0 to 1 in steps of 0.1 and obtain four evaluation results (Fig. 6) based on each evaluation method. In all the experiments, the larger the values, the better the results.

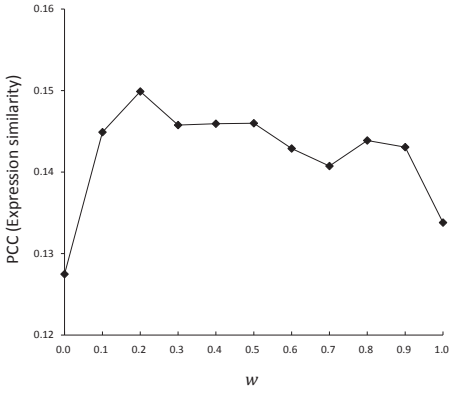
Even though the optimal value of w varies for each evaluation method, three experiments, except the difference between within-cluster similarity and between-cluster similarity (Fig. 6B), show the best results when w is between 0 and 1 (Fig. 6A, 6C and 6D). It is worth noting that all the results are much better when any small w is considered than when $w = 0$. This implies that the unobserved associations are not as useful as the observed associations, but still useful in inferring miRNA functional similarities.



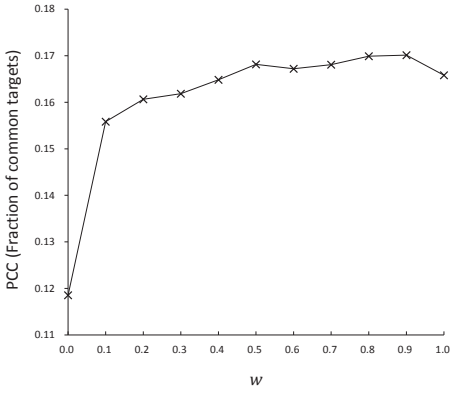
(A) Difference between within-family similarity and between-family similarity



(B) Difference between within-cluster similarity and between-cluster similarity



(C) PCC between expression similarity and functional similarity

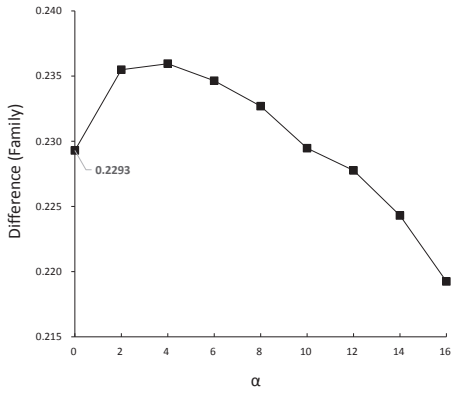


(D) PCC between fraction of common targets and functional similarity

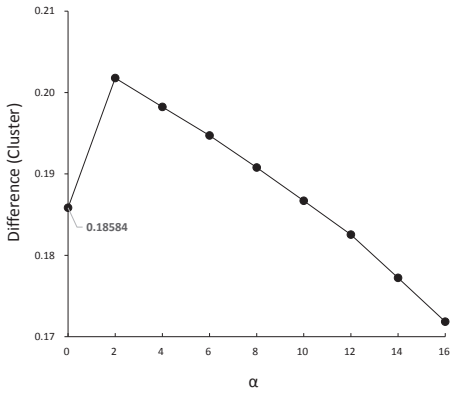
Figure 6: Impact of different w on reliability of the functional similarities inferred by MFSim in terms of the four evaluation methods

3.2.2 $\alpha\,(0 \leq \alpha)$

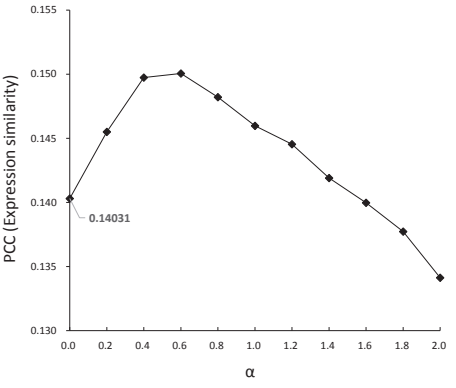
α is the hyper-parameter that controls the influence of "disease semantic similarity information" (Eq. 7). To investigate the impact of α, we vary α from 0 to 16 in steps of 2 in two evaluation methods (Fig. 7A and 7B), and from 0 to 2 in steps of 0.2 in two other evaluation methods (Fig 7C and 7D).



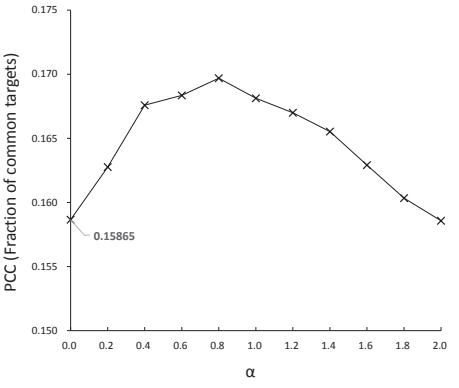
(A) Difference between within-family similarity and between-family similarity



(B) Difference between within-cluster similarity and between-cluster similarity



(C) PCC between expression similarity and functional similarity



(D) PCC between fraction of common targets and functional similarity

Figure 7: Impact of different α on reliability of the functional similarities inferred by MFSim in terms of the four evaluation methods

The results are better for $\alpha > 0$ than for $\alpha = 0$, which are natural, because we utilize the additional information to improve the results. A surprising observation is that even when the "disease sematic similarity information" is not utilized in MFSim (i.e., $\alpha = 0$), the functional similarities inferred by MFSim are more reliable than those inferred by MISIM (Fig. 7) in all the results; we expect that considering transitive relations among miRNAs improves the reliability of the functional similarities in MFSim. It is worth noting that all the results degrade as α increases beyond some point (Fig. 7). This drop signifies importance of a balance between "miRNA – disease association information" and "disease semantic similarity information".

3.2.3 $s_{jl}\,(0 \leq s_{jl} \leq 1)$

s_{jl} is a semantic similarity between diseases d_j and d_l. The proposed model utilizes all possible s_{jl} with the weight $e^{s_{jl}}$, whatever the value of s_{jl} is (Eq. 6). However, it does not tell anything about the different impact of "disease semantic similarity information" on the reliability, according to the values of s_{jl}. To investigate this, we perform two experiments; one with $e^{s_{jl}} = 0$, in case of $s_{jl} = 0$ and the other with $e^{s_{jl}} = 0$, in case of $s_{jl} > 0$.

Table 5: Reliability of the functional similarities inferred by MFSim with partial additional information

	Family	Cluster	Expression	Targets
MFSim (No disease info.)	**0.22930**	**0.18584**	**0.14031**	**0.15865**
MFSim (only $s_{jl} > 0$ info.)	0.22771	0.19312	0.14245	0.16179
MFSim (only $s_{jl} = 0$ info.)	0.23870	0.19846	0.14286	0.16224
MFSim (Proposed model)	**0.23433**	**0.20026**	**0.14598**	**0.16813**

The results show that each information, which is $s_{jl} > 0$ and $s_{jl} = 0$ respectively, improves the reliability of the functional similarities but not much

better than the case of using both information together, except the difference between within-family similarity and between-family similarity (Table 5). However, $s_{jl} = 0$ information improves the reliability of the functional similarities more than $s_{jl} > 0$ information. The reason seems that the number of 0 semantic similarities are much larger than the number of $0 >$ semantic similarities (Table 2).

IV. Discussion

MFSim is based on matrix factorization technique. It considers the transitive relations among miRNAs through their related diseases. The results show that the similarities inferred by MFSim are more reliable than those inferred by a compared method in terms of the four evaluation methods. MFSim also proves its flexibility and potential to be extended, by inferring the similarities without "disease semantic similarity information" (Eq. 4), or by integrating the information in the shared latent space harmoniously (Eq. 7). However, although MFSim shows several strengths, the model has a theoretical weakness, so-called *scale difference problem* suggested in section 2.4.3 and 2.5.1. Here, we discuss about the problem and also contemplate further extension of the model.

4.1 Scale difference problem

In sections 2.4.3 and 2.5.1, we encountered problems of scaling cosine similarities among vectors. More specifically, we needed to convert cosine similarities calculated by given vectors into the range of $[0, 1]$ in section 2.4.3 and needed to map cosine similarities of the range of $[-1, 1]$ to semantic similarities among diseases, which is the range of $[0, 1]$ in section 2.5.1. Although each case has a unique aspect, we adopt a naïve strategy in both cases to solve the problems; i.e., we just ignore the range of cosine similarity < 0.

The main reason that we adopt such a strategy is related to fundamental characteristics of cosine similarity. Assume that we have two vectors that are 2-dimensional and can have any real value as their elements. Then they can point to the same direction, be perpendicular to each other or point to the exactly

opposite directions. Therefore the range of cosine similarity between them can be $[-1, 1]$. However if we have three vectors, the maximum angle at which vectors are away from each other is 120°. As the number of vectors increases, the maximum angle among the vectors get narrow. Of course the angle can be extended as the dimension increases, but most pairs of vectors never point to the opposite directions if the number of vectors is > 2. In contrast, every vector can point to the same direction, regardless of their numbers and the dimensionality of the space that they occupy. In summary, cosine similarities among many vectors, not between two vectors, have an *imbalance property* in its distribution of possible values.

In section 2.5.1, we mapped cosine similarity of two latent factor vectors to the semantic similarity of two corresponding diseases (Eq. 6). We have 362 diseases and $\sim$80% of diseases pairs have 0 semantic similarities (Table 2). Each disease is represented as a 35-dimensional latent factor vector $\mathbf{d}_j \in \mathbb{R}^{35}$. If we want to map -1 cosine similarity to 0 semantic similarity, then $\sim$80% of total pairs of disease latent vectors should point to the opposite directions whatever scaling techniques we use, which is clearly infeasible. For that reason, we only use positive range of the cosine similarity.

Converting inferred cosine similarities to the range of $[0, 1]$ in section 2.4.3 is another problem and we have performed several schemes for the scaling. However, in many experiments, we notice that if we treat the negative cosine similarities with the same weights as positive cosine similarities (e.g., linear scaling) the results of the experiments are substantially bad. We suspect that the *imbalance property* of the cosine similarity affects the results anyhow. This problem should be investigated further.

4.2 Further extension

Even though MFSim has the potential to be extended with additional information, we actually have not used more than two data sources yet. However, MFSim should be enriched further by considering other "miRNA – related object information", for two reasons.

The first reason is that the evaluation criterion is not 'accuracy', but 'reliability'. This implies that even if the similarities inferred by MFSim shows better reliability than those inferred by MISIM, unless the results are sufficiently better, we cannot guarantee that the functional similarities inferred by MFSim is always closer to the ground truth than those inferred by MISIM. This decreases the necessity of using the new method instead of the old one (e.g., MISIM), and the meaning of the experiment results can be undervalued. Thus, reliability of the functional similarities among miRNAs should be greatly improved to justify switching from the old methods to the new one.

The second reason is that the functional similarities among miRNAs cannot be always inferred based on a single data source. For example, a functional similarity between two miRNAs cannot be inferred if they have no relation in the data. However, we can infer the similarity between them based on other information, if we utilize more than two data sources. Thus, we should integrate as much information as possible into the model and this is possible by mapping the information in the shared latent factor space.

Of course, integrating additional information is not trivial. To increase the reliability of the similarities, the model should be modified in a sophisticated way. Many models in recommender systems utilize content information to increase the accuracy of predicted ratings, or to solve the cold start problem that is the main bottleneck of MF based recommender systems. We may be able to get inspiration from those models in integrating additional information into MFSim.

V. Conclusion

This study proposes a novel model, MFSim, to infer functional similarities among miRNAs. The model can take into account transitive relations among miRNAs through their related diseases. Consequently, functional similarities inferred by MFSim are more reliable than those inferred by a compared method in terms of the four evaluation methods. MFSim also proves its flexibility and potential to be extended further with additional information. Thus, MFSim will be a useful framework for integrating several types of information harmoniously.

요 약 문

마이크로RNA (miRNA)는 작은 크기의 논코딩 RNA 분자로, 많은 생물학적 과정에서 중요한 기능들을 담당하고 있다. 불행히도, 대부분의 마이크로RNA에 대한 기능적 주석은 현재 존재하지 않는다. 따라서 그들 사이의 기능적 유사도를 추론하여 마이크로RNA의 잠재적인 기능들을 예측하는 방식이 관심을 끌고 있다. 현재 존재하는 추론 방법들은 "마이크로RNA – 관련 객체 정보"와 "해당 객체들 사이의 정보"를 기반으로 하는, '최고-매치 평균 전략'을 사용하여 기능적 유사도를 측정한다. 하지만 이들 중, 마이크로RNA 간의 '추이 관계'를 고려하거나 여러 가지 다른 형태의 정보들을 조화롭게 통합하는 방법은 존재하고 있지 않다. 본 논문에서는 '행렬 분해' (Matrix Factorization)를 기반으로 하는 새로운 추론 모델, MFSim 을 제안한다. 이 모델은 마이크로RNA와 관련 객체들을 '공통된 잠재 요소 공간' (Joint latent factor space)상에 사상함으로써 마이크로RNA 간의 추이 관계를 고려할 수 있다. 실험에서는 제안한 모델이 비교 방법보다 모든 평가 측면에서 우수함을 볼 수 있다. 본 모델은 또한 유동적인 특징을 가지고 있다. 즉, 추가적인 정보를 기존의 잠재 요소 공간상에 공유되도록 사상함으로써 해당 정보를 조화롭게 모델에 통합할 수 있으며 이러한 특징은 추후에 모델이 쉽게 확장 될 수 있도록 만들어 준다.

References

[1] D. P. Bartel, "MicroRNAs: Genomics, Biogenesis, Mechanism, and Function", Cell, vol. 116, pp. 281-297, 2004.

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

[3] A. Kozomara and S. Griffiths-Jones, "miRBase: annotating high confidence microRNAs using deep sequencing data", Nucleic Acids Res., vol. 42, pp. D68-D73, 2014.

[4] E. Wienholds, M. J. Koudijs, F. J. M. van Eeden, E. Cuppen and R. H. A. Plasterk, "The microRNA-producing enzyme Dicer1 is essential for zebrafish development", Nature Genet, vol. 35, pp. 217-218, 2003.

[5] I. Manni, S. Artuso, S. Careccia, M. G. Rizzo, R. Baserga, G. Piaggio and A. Sacchi, "The microRNA miR-92 increases proliferation of myeloid cells and by targeting p63 modulates the abundance of its isoforms", The FASEB Journal, vol. 23, pp. 3957-3966, 2009.

[6] Y. Wang and C. G. L. Lee, "MicroRNA and cancer-focus on apoptosis", J. Cell. Mol. Med., vol. 13, pp. 12-23, 2009.

[7] L. F. Lu and A. Liston, "MicroRNA in the immune system, microRNA as an immune system", Immunology, vol. 127, pp. 291-298, 2009.

[8] M. Y. Sasazaki and J. C. Felipe, “A New Ontology-Based Method for Functional Composed Comparison of microRNAs”, IEEE Computer-Based Medical Systems (CBMS), pp. 258-263, 2015.

[9] J. Meng, D. Liu and Y. Luan, “Inferring plant microRNA functional similarity using a weighted protein-protein interaction network”, BMC Bioinformatics, vol. 16, p. 360, 2015.

[10] D. Wang, J. Wang, M. Lu, F. Song and Q. Cui, “Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases”, Bioinformatics, vol. 26, pp. 1644-1650, 2010.

[11] G. Yu, C. L. Xiao, X. Bo, C. H. Lu, Y. Qin, S. Zhan and Q. Y. He, “A new method for measuring functional similarity of microRNAs”, JIOMICS, vol. 1, pp. 49-54, 2011.

[12] N. Srebro, T. and Jaakkola, “Weighted Low-Rank Approximations”, Proceedings of the 20th International Conference on Machine Learning (ICML), pp. 720-727, 2003.

[13] W. Xu, X. Liu and Y. Gong, “Document Clustering Based On Non-negative Matrix Factorization”, Proceedings of the 26th International Conference on Information Retrieval (SIGIR), pp. 267-273, 2003.

[14] J. P. Brunet, P. Tamayo, T. R. Golub and J. P. Mesirov, “Metagenes and molecular pattern discovery using matrix factorization”, Proceedings of the National Academy of Sciences, vol. USA 101, pp. 4164–4169, 2004.

[15] Y. Koren, “Matrix factorization techniques for recommender systems”, IEEE Computer, vol. 42, pp. 30-37, 2009.

[16] Y. Li, C. Qiu, J. Tu, B. Geng, J. Yang, T. Jiang and Q. Cui, "HMDD v2.0: a database for experimentally supported human microRNA and disease associations", Nucleic Acids Res., vol. 42, pp. D1070-D1074, 2013.

[17] M. Lu, B. Shi, J. Wang, Q. Cao, Q. Cui, "TAM: A method for enrichment and depletion analysis of a microRNA category in a list of microRNAs", BMC Bioinformatics, vol. 11, p. 419, 2010.

[18] Y. Liang, D. Ridzon, L. Wong and C. Chen, "Characterization of microRNA expression profile in normal human tissues", BMC Genomics, vol. 8, p. 166, 2007.

[19] B. John, A. J. Enright, A. Aravin, T. Tuschl, C. Sander and D. S. Marks, "Human MicroRNA Targets", PLoS Biol., vol. 2, p. e363, 2004.

Curriculum Vitae

Name : Jinha Jeong

Education

2007. 3. – 2014. 2. Department of Computer Science and Engineering, Chung-Ang University (B.S.)

2014. 3. – 2016. 8. Department of Computer Science and Engineering, Pohang University of Science and Technology (M.S.)


POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
1986
포항공과대학교