# MicroRNA-Disease Association Prediction by Known Nearest Neighbour Algorithm

Gill Varghese Sajan
Student, Computer Science and Engineering
Mar Athanasius College of Engineering, Kothamangalam
Ernakulam, Kerala, India. Pin: 686666
gillv522@gmail.com

Joby George
Professor, Computer Science and Engineering
Mar Athanasius College of Engineering, Kothamangalam
Ernakulam, Kerala, India. Pin: 686666
jobygeo@hotmail.com

**Abstract— MicroRNAs are a kind of non coding ribonucleic acids which are about 22nt nucleotides in length. Increasing confirmations shows that microRNAs perform important jobs in different diseases. The identification of disease related microRNAs will be helpful in exploring the underlying pathogenesis of the diseases. Proper treatment plans can be made by the doctor, once he is notified with all possible diseases that are related with each microRNAs. Experimental determination of diseases associations with microRNAs could be time-consuming and costly. Computational method can be an efficient alternative to identify potential diseases related to each microRNAs. Here we present an approach which computes the missing associations that exists between microRNAs and diseases using known nearest neighbor algorithm.**

**Keywords- miRNA; disease; association; prediction**

## I. INTRODUCTION

MicroRNAs are a class of single stranded endogenous non coding Ribonucleic acids that are almost 22 nucleotides long. The first microRNA ever to be discovered was lin-4. It was discovered 20 years ago. Since then many microRNAs have been commented on in different species by utilizing exploratory and computational strategies. Growing evidences shows that microRNAs play important functions in several biological processes like cell development, apoptosis, proliferation, differentiation etc. Regulating expression of disease genes is the way, the microRNAs exhibit their functions. The abnormality, disregulation or disfunctioning of microRNA biogenesis might cause different diseases, including illness that are inherited, cancers, issues in nervous system etc. Therefore, the identification of microRNAs related to each diseases will be useful for exploring the disease pathogenesis and in planning appropriate treatments.

## II. MATERIALS AND METHOD

Disease-disease similarities and microRNA-microRNA similarities are key components in microRNA-disease association prediction models. Pairwise topological similitudes among disease or microRNA are normally estimated utilizing cosine similarity or Gaussian interaction profile kernel similarity dependent on topological characteristics of microRNAs or diseases. Be that as it may, expectation prediction models utilizing these similitudes are not powerful enough.

### *A. Materials*

The microRNA information are stored in several publically available datasets. A comprehensive microRNA database called miRbase [1] contains sequences and hairpin structure of microRNAs. The HMDD [2] is a database which provides data regarding microRNA deregulation in different human diseases. MiR2Disease [3] is a physically curated learning base of tentatively proved human microRNA-disease associations.

### *B. Disease-Disease Similarity*

Using the MeSH dataset [4], every diseases could be represented as directed acyclic graph. The MeSH dataset provides a method for disease classification and can help in the studies of the relationship between diseases. In the DAG of each disease, each nodes represents diseases and links represent node relationships. Just a single sort

of relationship exists to connect a child and parent node. The relationship that exist between them are defined using 'is-a' relationship. To define the location of a disease in the MeSH graph, each diseases have at least one address in the DAG, called as codes. The parent node's code is appended behind the child's addresses and are used to define the codes of the child node. For example, there exists two possible addresses for breast neoplasms(C04.588.180 and C17.800.090.500). The corresponding parent nodes are C17.800.090 breast diseases and C04.588 neoplasms by site as shown in Fig.1.

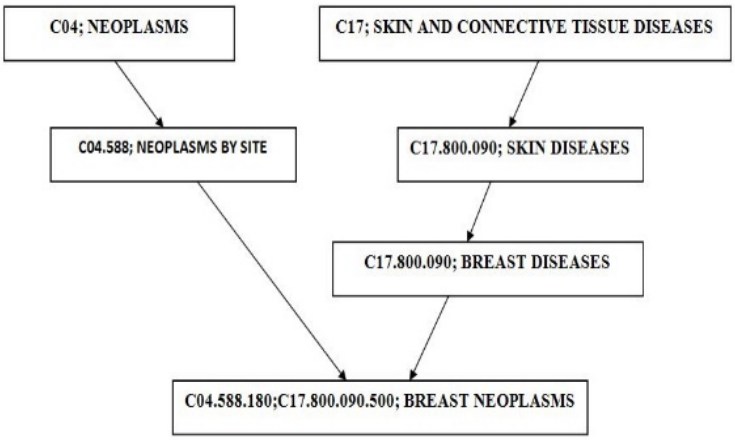


Figure 1. DAG of disease "Breast Neoplasms"

A disease 'A' can be represented as a directed acyclic graph, $DAG_A$ = (A, $T_A$, $E_A$), where $T_A$ represents the set of every ancestor nodes of 'A' including node 'A' itself and $E_A$ is the set of all corresponding links. The semantic contribution value representing the relation of a disease t to disease 'A' is represented by $D_A(t)$, which can be calculated as follows:

$$D_A(t) = \begin{cases} 1, & if\ t = A \\ \{0.5 \times D_A(t') | t' \in children\ of\ t\}, & if\ t \neq A \end{cases} \quad (1)$$

In the directed acyclic graph of 'A', disease 'A' is the most specific disease and therefore its contribution to its own semantic value is taken as one whereas its ancestor nodes which are located farther from node 'A' are more general denominations. So these ancestor nodes contribute much less to the semantic value of node 'A'. So using Equation (1), disease A's semantic value can be calculated as:

$$SemVal(A) = \sum_{t \in T_A} D_A(t) \quad (2)$$

By applying the basic principle that the diseases sharing larger part of their directed acyclic graphs tend to possess a higher semantic similarity [5], the similarity between different diseases can be calculated. The relative locations of two diseases in the MeSH disease directed acyclic graph are used for the calculation of the semantic similarity between them. The similarity value between any two diseases is calculated as follows:

$$S^d(A, B) = \frac{\sum_{t \in T_A \cap T_B} D_A(t) + D_B(t)}{SemVal(A) + SemVal(B)} \quad (3)$$

where $D_A(t)$ and $D_B(t)$ are the semantic values representing relationship of disease 't' to disease 'A' and disease 'B' respectively. Equation (3) is used to calculate the semantic similarity value between two diseases using the locations of those diseases in directed acyclic graphs and their relationships with their corresponding ancestor diseases.

### *C. MicroRNA-MicroRNA Similarity*

The contributions from similar diseases which are associated with two microRNAs are used to precisely calculate the functional similarity [6] between them. Therefore semantic similarity among a disease and a group of disease is to be defined first. For instance, let 'd' be one disease and let 'D' be one disease group, e.g. D = {$D_1$, $D_2$,..., $D_k$}. The similarity between d and D is given by S(d, D), which is the maximum similarity that exists between the disease 'd' and the disease group 'D'. It is calculated as:

$$S(d, D) = max_{1 \leq i \leq k} S(d, D_i) \quad (4)$$

For clarity, to calculate the functional similarity between two microRNAs, $m_1$ and $m_2$ assume $D_1$ represents the related diseases of $m_1$ and $D_2$ represents the related diseases of $m_2$. $D_1$ and $D_2$ have m and n distinct diseases respectively. The computation of functional similarity of two microRNAs will need to consider each and every diseases in $D_1$ as well as in $D_2$. Hence the similarity between two microRNAs is computed using:

$$S^m(m_1, m_2) = \frac{\sum_{1 \le i \le m} S(d_{1i}, D_2) + \sum_{1 \le j \le n} S(d_{2j}, D_1)}{m + n} \quad (5)$$

For having more desirable understanding of microRNAs it is necessary to construct a reliable microRNA functional network. MISIM is a trustworthy computation measure of microRNA similarity. Using MISIM the microRNA functional network can easily be constructed. The pairwise MISIM coefficients for a list of microRNAs are computed. Then a threshold value for MISIM coefficients is decided. Now microRNA pairs with MISIM coefficient which is greater than or equal to the threshold value will possess a direct link between them. And finally a functional network for the microRNAs is constructed.

### *D. Heterogeneous Network Construction*

The microRNA-microRNA similarity network and disease-disease similarity network which are constructed using the process of fixing a threshold for the similarity values in the previous sections acts as the subnets for the computation of the heterogeneous networks. After the computation of microRNA-microRNA similarity and disease-disease similarity, the next step is to combine them using the known associations to form bipartite graph. The graphs is microRNA-disease heterogeneous network. The two subnets namely, the disease similarity network and microRNA similarity network, could now be connected using an experimentally proved microRNA-disease interactions to form a heterogeneous network of microRNAs and diseases. HMDD dataset [7] is used to get the experimentally verified associations between microRNAs and diseases.

Suppose that $S^d = S^d(i,j)_{i=1,j=1}^{n,n}$ and $S^m = S^m(i,j)_{i=1,j=1}^{m,m}$ are adjacency matrices representing the disease similarity network and microRNA similarity network respectively. Similarly $Y = Y(i,j)_{i=1,j=1}^{m,n}$ represents the microRNA-disease interaction network, where n and m are the numbers of disease and microRNA entities. An example of the microRNA-disease heterogeneous network is given in Fig.2.

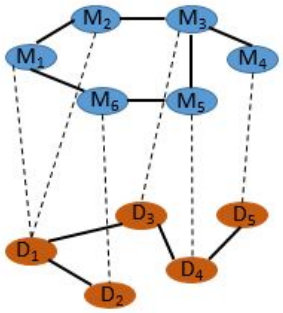


Figure 2. MicroRNA-Disease Heterogeneous Network

## III. UNKNOWN ASSOCIATION PREDICTION

### *A. MicroRNA-Disease Association*

In this section the prediction of undiscovered microRNA-disease interactions using the known associations is the main aim, hence a new feature vector called "interaction profile" for both diseases and microRNAs are introduced. Let $Y = R^{m,n}$ represent an adjacency matrix constituting the microRNA-disease associations whose rows have m microRNAs and columns have n diseases, where $Y_{ij} = 1$ if microRNA $m_i$ have a link with disease $d_j$; otherwise $Y_{ij} = 0$. The $i^{th}$ row vector of Y, represented by $Y(m_i)$, is the interaction profile for microRNA $m_i$ and the $j^{th}$ column vector of Y, represented by $Y(d_j)$, is the interaction profile for disease $d_j$. A microRNA or disease being known means that their profiles have at least one interaction; whereas they being new means that they have no interactions in their profile.ie., for new diseases or microRNAs, the interaction profiles will completely zero vectors. The matrix in Fig.3 represents the known interaction profile representations of microRNAs and diseases. In the Figure there exists m microRNAs represented by $M_1, M_2, \ldots, M_m$ and n diseases represented by $D_1, D_2, \ldots, D_n$.

| | $D_1$ | $D_2$ | $D_3$ | | $D_n$ |
|---|---|---|---|---|---|
| $M_1$ | 0 | 1 | 0 | ... | 0 |
| $M_2$ | 1 | 0 | 1 | ... | 0 |
| $M_3$ | 0 | 1 | 1 | ... | 0 |
| | ... | ... | ... | ... | ... |
| $M_m$ | 0 | 1 | 1 | ... | 1 |

Figure 3. Interaction Profiles

Most of the non interactions or 0's in matrices Y are unknown cases which can actually be true interactions.i.e. they can be false negative values. The non interacting microRNA-disease pairs in Y are actually missing edges. Hence a matrix updation process called Known Nearest Neighbours(KNN) could be utilized to calculate the interaction likelihood score for these non interacting pairs by making use of their known neighbours.ie., the procedure tries to replace the values of the $Y_{ij}$ equals 0, by a value from 0 to 1 using the algorithm given below:

### *B. Known Nearest Neighbour Algorithm*

Input: Adjacency Matrices $Y \in R^{m \times n}$, $S^m \in R^{m \times m}$ and $S^d \in R^{n \times n}$ & decay term 'T'.
Output: Modified matrices Y.
Algorithm:

1: For p=1 to n do
1.1: mn = knownNeighbour(p, $S^m$)
1.2: k = length(mn)
1.3: For i=1 to K do
1.3.1: $w_i = T^{i-1} S^m$ (p,$mn_i$)
1.4: end for
1.5: $Q_m = \sum_{i=1}^{k} S^m(p, mn_i)$
1.6: $Y_m(p) = (\sum_{i=1}^{k} w_i Y(mn_i))/Q_m$
2: end for
3: For q=1 to m do
3.1: dn = knownNeighbour(q, $S^d$ )
3.2: k = length(dn)
3.3: For j=1 to K do
3.3.1: $w_j = T^{j-1} S^d$(p,$dn_j$)
3.4: end for
3.5: $Q_d = \sum_{i=1}^{k} S^d(p, dn_j)$
3.6: $Y_d(p) = (\sum_{j=1}^{k} w_j Y(dn_j))/Q_d$
4: end for
5: $Y_{md} = (Y_m + Y_d)/2$
6: Y = max(Y, $Y_{md}$)
7: return Y

The updated Y matrix acts as new interaction profile,ie., nearest neighbour interaction profile for microRNAs and diseases. These features are used to construct the prediction models. The Y matrix can be used for the construction of resultant prediction model. The matrix Y has its rows as microRNAs and columns as diseases and the values represents microRNA-disease associations.ie., it specifies how likely is the microRNA in that row would be associated with the disease in each column. Many of the zero values in the matrix (formed using known associations) have been replaced with values within the range 0 and 1,ie., more diseases will be associated with each microRNAs. The matrix Y act as the microRNA-disease association model.

## IV. EXPERIMENTS AND RESULTS

To evaluate the performance of the proposed method in predicting microRNA-disease associations, the strategy called leave one out cross validation was adopted in our experiments. Initially we computed the various diseases that are associated with each microRNAs. Then to check the predictive performances of our method we applied leave one out cross validation.ie., for each microRNAs in the dataset, it was considered as a test

microRNA once and its association information was deleted. The remaining microRNAs were taken as the training dataset. The predicted indications for the test microRNA were ranked according to the final result received after KNN. That is we computed the prediction efficiency of our system by comparing them with experimentally proved associations. For each specific ranking threshold, if the weight of an inferred microRNA-disease association was above the threshold, it was regarded as a true positive. Otherwise, it was regarded as a false positive. True positive rate (TPR) and false positive rate (FPR) were calculated by varying threshold values. The prediction ability of the method was represented using the area under the curve (AUC) value. The leave one out cross validation procedure was used to test the prediction performance of this system and experimental outcomes exhibit that when leave one microRNA out cross validations were implemented, an average AUC value of 0.67 was received in microRNA-disease association predictions when KNN algorithm was applied. These results indicated that reliable microRNA-disease association prediction results could be achieved by our method.

## V. CONCLUSION

It is still a great challenge in understanding molecular mechanisms of diseases. Elucidating the complex molecular mechanisms of diseases can be helpful for exploring disease pathogenesis and designing appropriate and effective treatments. Some studies show a disease can be viewed as the consequence of perturbation of regulation network instead of the abnormality of single gene product. Non-coding RNAs, such as miRNAs, act as vital regulation factors of gene expression. Expanding confirmations describes that non coding RNAs have cozy association with different diseases. Predicting unknown disease-miRNA associations helps to decipher disease pathogenesis. This study presented a known nearest neighbour algorithm for disease-miRNA interaction prediction. Initially we calculated microRNA-microRNA and disease-disease similarities and formed two adjacency matrices. Then utilized the known nearest neighbour algorithm to consolidate these outcomes to foresee unknown associations. In the computational trials, our method can produce good performances.

## REFERENCES

[1] Z. Yang, F. Ren, C. Liu, S. He, G. Sun, Q. Gao, L. Yao, Y. Zhang, R. Miao, Y. Cao, et al. dbDEMC: a database of differentially expressed miRNAs in human cancers, BMC Genomics, vol. 11, no. Suppl 4, p. S5, 2010.

[2] 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, no. 1, pp. D1070-1074, 2014

[3] Q. Jiang, Y. Wang, Y. Hao, L. Juan, M. Teng, X. Zhang, M. Li, G. Wang and Y. Liu miR2Disease: a manually curated database for microRNA deregulation in human disease, Nucleic Acids Res, vol. 37, no. 1, pp. D98-D104, 2009.

[4] I.Lee,U.M.Blom,P.I.Wang and J.E.Shim, Prioritizing candidate disease genes by networkbased boosting of genome-wide association data, Genome Res,vol. 21, pp.1109-1121.

[5] L.Cheng,J.Li,P.Ju,J.Peng,Y.Wang, SemFunSim:a new method for measuring disease similarity by integrating semantic and gene functional association, PLoS One,vol. 9,no. 6,2014.

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

[7] D. P. Bartel, MicroRNAs: Genomics, biogenesis, mechanism, and function, Cell, vol. 116, no. 2, pp. 281297, Jan. 2004.