miRNA-miRNA and Disease-Disease similarity Identification and Disease Detection using Combinatorial Prioritization Algorithm

JIJITH MP
Student: Computer Science And Engineering Department
Mar Athanasius College of Engineering
Kothamangalam, India
jjijithmp1@gmail.com

JOBY GEORGE
Professor: Computer Science And Engineering Department
Mar Athanasius College of Engineering
Kothamangalam, India
jobygeo@hotmail.com

Abstract— Increasing evidences have indicated that microRNAs (miRNAs) are functionally related to the event and progression of assorted advanced human diseases. However, the roles of miRNAs in multiple biological processes or varied diseases and their underlying molecular mechanisms still haven't been totally understood nevertheless. Meanwhile, target illness have to be compelled to be disclosed for a few new microRNAs with none glorious target disease association data as new microRNAs ar discovered annually. Therefore, process strategies for microRNA-disease association prediction have gained heaps of analysis interest. Considering the constraints in previous process strategies, here developing the model of kernel based mostly network similarity integration (KBNSI). It integrate the results of each the miRNA-miRNA association and also the illness illness association similarity then it'll integrates the each results, then sight the illness

Keywords— Measuring the similarity, miRNA and disease association, disease-related miRNAs, disease-disease association, miRNA associations, miRNA similarity, disease similarity, similarity matrix

I. INTRODUCTION

Generally microRNAs (miRNA) are terribly tiny endogenous and non-coding RNAs and its size concerning twenty two NT long. MiRNAs ar plays major vital role in several biological processes, together with the cell development, differentiation, proliferation, apoptosis, and in cellular sign Presently microRNA regulate many alternative style of process, which incorporates the cell death, differentiation, development, and diseases [1–3]. several studies associated with microRNA found that it play a vital role in tissue development [5–7], cellular sign networks [4], and cell growth [8], they're additionally related to totally different verity of diseases [9, 10], together with cardiopathy [15], carcinoma [11, 12], cell malignant neoplastic disease [16] and carcinoma [13, 14]. If the abnormality in microRNA causes the malady, then the abnormal microRNA and therefore the connected malady ar associated by the causative relationship, thus our aim to predict the microRNA-disease association. Predicting the microRNA-disease associations has emerged as a vital strategy in understanding the malady mechanisms [17], as an example, the freeing of explicit microRNAs can ends up in caspase-mediated cell death communication pathways and conjointly cell cycle regulation in willcérer [18]. Major analysis by exploitation the biological experimentation has determined that an oversized variety of associations between diseases and miRNA. Evaluating the databases like miRCancer, miR2Disease, HMDD, dbDEMC are in-built order to produce an appropriate platform for looking through an experiment verified miRNA-disease associations. the info sets like HMDD and miR2Disease square measure the gathering of through an experiment supported human miRNA-disease associations, and manually retrieved data on the premise of the literature. Info of miRCancer stores data that associated with miRNA-cancer associations, they're extracted by exploitation the rule-based text mining mechanism. additionally, dbDEMC knowledge base contains data that associated with completely differentially expressed miRNAs in fourteen human cancers by exploitation the importance analysis of various microarrays so as to retrieve the miRNAs that have different organic phenomenon levels in cancers after they compared with traditional tissues. These completely different databases function a solid foundation for prognosticative analysis of miRNAs in human diseases.

Evaluating the experimental identification of disease-related miRNAs, the major downside associated with this is time intense and pricey, researchers planned a spread of machine ways as vital complementary ways in which so as to predict miRNA-disease associations. machine ways in the main aim to retrieve the foremost promising disease-related miRNAs for future experimental examination so as to cut back the experimental time and price. The key downside concerned in miRNA-disease association illation is similarity calculation between the miRNA and illness similarity. These machine ways square measure divided into 2 categories:[18] machine-learning-based ways [27-31] and network-based[19-27] methods Network-based ways typically[is often] wont to predict miRNA-disease associations by considering the hypothesis that functionally connected miRNAs square measure usually related to phenotypically similar diseases. This hypothesis was planned by Lu et al. Basing on this hypothesis, Jiang et al [18] made a brand new practical association miRNA network i.e., an individual's phone miRNA network. For any given illness,
they computed the similarity score between all human miRNAs in these networks so prioritized of these miRNAs in keeping with similarity score. The top-ranked miRNAs square measure expected because the potential illness miRNAs. However, the matter with this model is it uses solely the neighboring data of every miRNA and it powerfully depends the anticipated miRNA-target interactions, and thereby manufacturing a false-positive and a false-negative results, it will influence the ultimate prediction accuracy. Shi et al.[21] presented a machine framework for characteristic miRNA-disease associations and more they made a bipartite miRNA-disease network for consistently analyzing the world properties of miRNA regulation of every illness genes. From these completely different experimental analyses, they reach the conclusion that the majority of the diseases within the same co-regulated module belongs to constant class. Their work helps to extending the previous hypothesis. However, this technique is appropriate for restricted application attributable to the low accuracy of target prediction and therefore the several association between disease–gene of miRNA-target interactions square measure unknown. By considering the weighted k most similar neighbors, HDMP [22] was planned for predicting disease-related miRNA associations. It absolutely was wont to judge the perform similarity between miRNAs by considering illness terms and therefore the constitution similarity between every illness genes, similarly as by assignment the upper weight to members of the miRNA cluster. However, this technique solely considers native network similarity live and disregards diseases with none well-known connected miRNAs Recently, Zou et al [27] have bestowed new technique KATZ, that uses the practical similarity score for denoting the associations on the premise of the various lengths between the illness nodes and miRNA but, the performance of this technique is comparatively poor on the sparse well-known associations.

Machine-learning-based ways square measure in the main used for the aim of resolution the matter by up the prediction performance and classification accuracy. Jiang et al.[29] introduce a Naïve mathematician model for ranking candidate disease-related miRNAs by exploitation genomic knowledge integration technique. This technique powerfully depends on datasets of miRNA-target interactions and, disease–gene associations however over half human diseases square measure still unknown. to tell apart the positive miRNA-disease associations from the negative miRNA-disease associations, Jiang et al. planned a support vector machine mechanism. This technique extracting the options supported constitution similarity knowledge and miRNA-target knowledge. In this approach considering the idea that miRNAs involved in an exceedingly specific growth constitution that show aberrant regulation of their target genes, Xu et al.[30] prioritized novel sickness miRNAs square measure on the premise of the miRNA target-deregulated network technique. The common drawback of this 2 same strategies is that the diseases don't demonstrate adequate applied mathematics confidence associated negative coaching samples consisting of non-association between miRNAs; throughout every observation the shortage of a miRNA-disease association in an exceedingly biological experiment doesn’t directly indicate absence of such an association. Chen et al.[28] developed new regular statistical procedure for every miRNA-disease association (RLSMDA) so as to search out the potential miRNA candidates for a particular sickness. This technique could be a semi-supervised technique that integrates disease–disease similarity dataset, miRNA–miRNA useful similarity network and illustrious disease-miRNA associations. Despite it’s a decent prediction performance for diseases with or while not connected miRNAs, however this technique doesn’t contemplate the topological info of every miRNA network. The same strategies have 3 main inexpedient. First, some approaches square measure unable to predict isolated disease-related miRNAs. Second, negative samples square measure tough to get for a few machine learning strategies. Third, some strategies square measure inefficient at cross-validation. so as to beat this limitations, our purpose to develop a kernel primarily based network similarity integration technique. it’s a good technique to infer a bipartite graph by multiple information supply integration [34, 35]. miRNAs and diseases square measure appointed to 2 domains and therefore the contiguity matrix is employed to represent the illustrious associations between miRNA and diseases. This matrix performs the kernel primarily based nonlinear spatial property by mistreatment input kernel matrices and therefore the common projection matrix.

II. MATERIALS

A. Data set

The info that area unit utilized in this study includes knowledge on malady linguistics similarities, famous human miRNA-disease associations, and miRNA–miRNA useful similarities. Maintaining the Integrity of the Specifications. The linguistics associations between totally different maladies area unit downloaded from malady metaphysics info [28] There area unit 8632 disease terms and 7232 associations be- tween diseases. The malady terms area unit represented as di- rected acyclic graph (DAG), wherever nodes represent malady terms and links represent associations between malady terms. The Medical Subject Headings (MeSH) de- scriptors area unit exacted from National Library of medication. The MeSH descriptors area unit represented in a hierarchy DAG, in that nodes denote diseases and edges denote associations between two dis- eases.

B. miRNA-miRNA and Disease-Disease Similarity

Here, considering two styles of miRNA-miRNA similarity; miRNA-miRNA purposeful similarity and miRNA- miRNA sequence similarity. The miRNA purposeful similarity is calculated supported misim technique [29]. The miRNA-miRNA purposeful similarity network is delineated as medium frequency matrix. The entity MF(i,j) in row i and column j representing the similarity

1382
score between miRNAs i and j, so as to see the sequence similarity between 2 miRNAs. This technique uses Emboss-Needle tool (Needleman-Wunsch alignment algorithm) [32]. The miRNA sequence similarity is wide wont to discovery and verify the miRNA-miRNA interaction identification [30, 31]. The parameters of Emboss-Needle were such because the default price (Matrix = EDNAfull, Gap open = ten, Gap extend = zero). The malady purposeful similarity and malady linguistics similarity live) used to measure the association between two diseases [29, 33], conniving malady purposeful similarity relies on the idea that similar diseases tend to be connected with similar genes [33]. Therefore, cistron perform similarity will be wont to live malady similarity calculation. This technique uses two types of malady linguistics similarity to live the similarity between 2 diseases in this study: DSSM [29] and DSSD [33], the primary one uses Mesh information whereas the opposite one uses malady metaphysics data-based. the info for renowned human miRNA-disease were collected from (http://www.cuilab.cn/hmdd, Jun-14-2014 Version). Here it performs a filtering mechanism for removing duplicated associations and people associations whose malady failed to have a connected MeSH tree range or whose malady gens couldn’t be mapped to the MeSH information. once activity filtering method, finally received 5425 high-quality through an experiment verified human miRNA-disease associations that consisting of 495 miRNAs associations and 381 diseases associations within the dataset. The ensuing matrix AS indicates miRNA-disease associations and AS(i,j) = 1 means there exists a minimum of a valid association for every miRNA i and malady j; if there’s no valid associations between them then, AS(i,j) = 0.

1) Disease directed acyclic graph

In our planned methodology, a useful similarity score is employed for every sickness combine it absolutely was calculated supported the hypothesis that miRNAs with similar functions are connected to similar diseases. Our planned system try and improves the calculation methodology, every diseases ar mapped to the MeSH info and their MeSH heading or descriptors or downloaded. every MeSH heading displays a tree structure of a gradable organization. This tree structure of every sickness is delineate as graph called a directed acyclic graph (DAG). The nodes of the tree represents diseases whereas the every edges represent the link among the parent node and their youngsters nodes. If the hierarchy of every node is higher, the additional general its which means is. Otherwise, the hierarchy of every node is lower.

2) Kernal Based Matrix Factorization

This is a good technique for infer bipartite graph by multiple information supply integration [38, 39]. Here miRNA and diseases square measure assign to completely different domain. The contiguity matrix is employed to representing the better-known associations between the miRNA and diseases. during this matrix resolving it’ll be factorise the every of the output that similar to the factorizing latent output. The result from this output from every stage is integrated with the assistance of KBNSI.

III. METHODS

Based on the thorough experiment valid info from miRNA-disease association network and 2 common assumptions, at the start reconstructing the miRNA and malady similarity networks, and utilized the KBNSI to predict potential miRNA-disease associations, one in every of the belief is miRNAs with similar functions square measure commonly related to phenotypically similar diseases and contrariwise, [13, 31] and therefore the different is diseases with similar functions square measure usually having similar linguistics descriptions and contrariwise [19] The KBNSI contains four processes. First, it’ll calculates the linguistics similarity score of diseases in step with the linguistics tree structure. Second, it calculates miRNA–miRNA purposeful similarities supported the linguistics similarity score of diseases. A miRNA purposeful network was engineered on the idea of those calculations. Third, it calculates the similarity score of maladys to reconstruct a malady similarity network by considering the malady linguistics similarities and disease similarities of noted miRNA-disease associations. Fourth, it integrates the malady similarities, miRNA similarities, and noted miRNA-disease associations to predict potential associations between miRNAs and diseases.

1) Measurement of disease semantic similarities

Some researchers have calculated the similarity measures of sickness by considering the hierarchical data structure of disease linguistics. [19, 32] In this planned work, the linguistics similarity live for sickness gens ar developed supported the knowledge from Wang et al. [19] however it’s not constant values. For each sickness A will be diagrammatical as a directed a cyclic graph, DAG(A) = (A, TA, EA), wherever atomic number 73 representing the set of all root nodes of A as well as A itself and EA contains the set of corresponding links of A additionally the contribution of root node t to A will diagrammatical as follows

\[
D_A(t) = \begin{cases} 
1 & \text{if } t = A \\
\max \{\Delta \times D_A(t') \mid t' \in \text{children of } t\} & \text{if } t \neq A.
\end{cases}
\]

(1)

Here \(\Delta\) denotes the semantic contribution factor for edges \(E_A\) linking disease \(t\) with its corresponding child disease \(t'\). Semantic value of each disease A is represented as follows:
The semantic similarity of each disease A and disease B is represented by,

\[ DV(A) = \sum_{t \in T_A} D_A(t), \quad (2) \]

The semantic similarity of each disease A and disease B is represented by,

\[ DD(A, B) = \sum_{t \in T_A \cap T_B} D_A(t) + \sum_{t \in T_A \cap T_B} D_B(t) \]

\[ \frac{2 \times \min \{DV(A), DV(B)\}}{DV(A) + DV(B)}. \quad (3) \]

Here, \( t \) denotes the illness of each in Ta and Tb. \( DA(t) \) represents the linguistics worth of illness \( t \) corresponds to illness A and \( DB(t) \) represents the linguistics worth of illness \( t \) akin to illness B. The linguistics similarity score between this 2 diseases not solely depends on the amount of common diseases however conjointly to the whole linguistics relations worth. If the whole variety of common diseases will increase and also the total linguistics worth of common diseases is additionally become higher, consistent with the very best score.

2) Measurement of miRNA functional similarity

This method defines disease set as \( DS_i = \{d_1, d_2, \ldots, d_n\} \), it is associated with miRNA \( j \). The related score for each disease \( d \in DS_i \) and set \( DS_j \) is represented as follows:

\[ DM(d, DS_j) = \max_{i \in DS_i} \{DD(d, DS_j(t))\}. \quad (4) \]

The maximum similarity of malady \( d \) and maladys in \( DS_j \) is outlined because the connected score between disease \( d \) and for every miRNA \( j \). The miRNA–miRNA operate similarity matrix outlined by metric linear unit, where \( MM(i,j) \) I representing row parts and column \( j \) expresses the purposeful similarity score between every miRNA \( i \) and miRNA \( j \). By evaluating the contribution of the similarity diseases, the purposeful similarity of \( MM(i,j) \) is set as follows:

\[ SMD_j = \sum_{d \in DS_i} DM(d, DS_j) \]

\[ MM(i,j) = \frac{SMD_i + SMD_j}{|DS_i| + |DS_j|}. \quad (5) \]

Here \( SMD_i \) denoting the similarity score of every miRNA \( i \) and every illness set \( DS_j \), and \( SMD_j \) representing the similarity score of miRNA \( j \) and illness set \( DS_i \). \( |DS_i| \) contains the quantity of noted diseases related to miRNA \( i \), and \( |DS_j| \) contains the quantity of noted diseases related to miRNA \( j \).

3) Reconstruction of a disease similarity network

The sickness similarity network was reconstructed by evaluating the sickness similarities of noted miRNA-disease associations and sickness linguistics similarities. By assuming[36,37] that a lot of common miRNAs of every sickness try has a lot of similar they’re, outlined by the sickness similarity price of every noted disease-miRNA association by evaluating the matrix \( AS \) and Jaccard similarity mensuration may be drawn by

\[ DAS(i,j) = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}. \quad (6) \]

Considering the matrix \( AS \), for every sickness \( i \) and sickness \( j \), here its count the whole variety of normally associated miRNAs of every sickness \( i \) and sickness \( j \), and so defines it as a matrix \( M11 \). Similarly, the whole variety of miRNAs that ar solely related to sickness \( i \) is represent within the matrix \( M01 \), the matrix \( M10 \) represents the whole variety of miRNAs that ar solely related to sickness \( j \), the whole variety of miRNA gens that don't seem to be related to neither i nor j is forgotten. For a precise sickness try, the similarity is about to zero if the whole variety of miRNAs related to these two diseases. By reconstructing the sickness similarity network as follows:
\[
SD(i,j) = \frac{DD(i,j) + DAS(i,j)}{2}, \quad (7)
\]

Here \(SD(i,j)\) representing the ultimate illness similarity worth of every illness \(i\) and every illness \(j\), during this equation, if additional similar illness \(i\) and illness \(j\) within the identified association network and higher the illness linguistics similarity between every of them, the upper their similarity worth is. Here hypothesise that the illness linguistics similarity for every illness is as vital because the illness similarity calculated by referring the identified association network. Thus, it give constant weight is given to make the illness similarity measuring.

4) KBNSI for miRNA-disease associations

In our planned KBNSI integrates the miRNA and illness vector house score for calculative the potential association between every miRNA-disease association scores .This methodology conjointly uses the cos similarity to see the vector house score. Evaluating the miRNA vector house, the similarity score between every miRNA \(i\) and every one miRNAs is outlined during a vector called VMMi, and MMI, area unit wont to representing the values. Likewise, the similarity between the associations of illness \(j\) and every one miRNAs is delineate as a vector named as VDj, and ASj is employed to represent it. Here \(I\) representing the row and \(j\) representing the column

\[
VD_j = AS_j
\]

\[
VMM_i = MM_i
\]

The miRNA space score value is determined as follow

\[
NSIM_M(i,j) = \frac{VMM_i \cdot VD_j}{\|VMM_i\| \|VD_j\|}, \quad (8)
\]

here \(VMM_i \cdot VD_j\) is representing the real number between vector VMMi and VDj; \(\|VMM_i\|\) is denoting the norm of vector VMMi, \(\|VD_j\|\) is that the norm of vector price of VDj, the trigonometric function similarity of vector VMMi and VDj is delineate in KBNSI_M(i,j). Obviously, for every of the smaller angle between VMMi and VDj, then larger the vector area score KBNSI_M(i,j).

For each of the upper abstraction similarity of miRNA \(i\)-associated miRNAs within the miRNA–miRNA similarity network. association between miRNA \(i\) and unwellness \(j\) is become higher. Similarly, for the upper abstraction similarity of unwellness \(j\)-associated with every miRNAs within the celebrated miRNA-disease network is, the larger the association between every miRNA \(i\) and every unwellness \(j\) is. Evaluating the unwellness vector area, the similarity between the associations of miRNA \(i\) and every one diseases is delineate as a vector VMi. It might use ASi, (\(i\) representing \(ith\) row of matrix AS) to represent it. Similarly, the similarity between unwellness \(j\) and every one diseases is delineate as vector VSDj, and therefore the vector SDj (\(j\) representing the \(jth\) column of matrix SD) wont to outline it.

\[
VSD_j = SD_j
\]

\[
VM_i = AS_i
\]

The disease space score value is determined by

\[
NSIM_D(i,j) = \frac{VM_i \cdot VSD_j}{\|VM_i\| \|VSD_j\|}, \quad (9)
\]

here \(VM_i \cdot VSD_j\) is representing the scalar product of vectors VMi and VSDj; \(\|VM_i\|\) is representing the norm of vector VMi, \(\|VSD_j\|\) defines the norm of vector VSDj. KBNSI_D(i,j) is representing the trigonometric function similarity of the vectors VMi and VSDj. Notably, for the bigger vector house score KBNSI_D(i,j), then smaller the angle between VMi and VSDj. Obviously, if the bigger association of miRNA \(i\) and malady \(j\), for every higher spatial similarity of miRNA \(i\)-associated diseases within the glorious miRNA-disease network. Likewise, for the upper spatial similarity of the malady \(j\) associated maladys in disease similarity network is, then the bigger the association of miRNA \(i\) and malady \(j\). Finally, the malady house score worth and therefore the miRNA house score values square measure integrated along as follows

\[
KBNSI(i,j) = \alpha \times KBNSI_M(i,j) + (1 - \alpha) \times KBNSI_D(i,j) \quad (10)
\]
where $\alpha$ is a parameter that used to balance the contributions from the two space similarities, $\alpha \in (0,1)$. KBNSI(i,j) in ith row and jth column is the prediction-related score of miRNA i to for each disease j. In order to determine a suitable $\alpha$ value, the different $\alpha$ values that vary from 0.1 to 1 were evaluated by the experiments. The following fig.1 shows that the highest prediction performance when the value of $\alpha$ is 0.5, it is achieved by our proposed KBNSI method.

Fig. 1 resulting AUC value obtained by KBNSI

IV. RESULT

By victimisation the leave-one-out cross validation (LOOCV) on miRNA-disease associations so as to judge the prophetic performance of our planned methodology KBNSI, every of the illustrous miRNA-disease association area unit overlooked successively because the take a look at sample for KBNSI, and therefore the remaining illustrous miRNA-disease associations area unit taken as a coaching set. By varied the edge A receiver operative characteristic (ROC) curve was planned, supported this our methodology calculated the worth of space underneath curve (AUC), within the receiver operative characteristic, the vertical and horizontal axes area unit actuality positive rate and false positive rate at totally different thresholds. The sensitivity refers to the proportion of take a look at in every miRNAs with ranking on top of of the given threshold, whereas the specificity refers to the proportion of associations below to the edge. The prediction performance is best once the United Self-Defense Force of Colombia is nearer to one. Avoid combining SI and metric system units, like current in amperes and magnetic flux in oversets. This typically ends up in confusion as a result of equations don’t balance dimensionally. If you need to use mixed units, clearly state the units for every amount that you just use in Associate in Nursing equation. The superior performance of our methodology will be contributed to some factors. Firstly, our methodology integrates totally different knowledge sources. In existing strategies, the miRNA and malady similarities area unit calculated supported Mesh dis-ease dataset which can lead to biases, so as to overcome back this limitation, miRNA sequence, sequence operate and malady linguistics info area unit) integrated to measure pairwise-miRNA similarity and pairwise-disease similarity. Secondly, the KBNSI utilized in this paper is a good methodology to integrate multiple kernels that area unit obtained from totally different knowledge sources. Last however not least, KBMF-MDI has powerful prophetic ability for diseases with few illustrous connected miRNAs. V. conclusion

One of the foremost tough challenge is that to grasp molecular mechanism among the diseases comparison with miRNAs. so as to predict the potential microRNA-disease associations by evaluating the procedure ways will give an excellent support for experimental studies associated with microRNAs. during this study, projected the KBNSI to predict the miRNA-disease associations by group action the sickness similarities, better-known miRNA-disease associations and miRNAs similarities. The KBNSI methodology obtained a highest terrorist organization of zero.9475 in leave one out cross validation methodology, moreover, case studies of colon neoplasms prostate, breast, and were enforced, and 19, 17, and twenty miRNAs within the high prediction list were confirmed By evaluating this results, it demonstrate that KBNSI will effectively perform the identification of potential disease-related miRNAs. KBNSI additionally performs a well in predicting the isolated diseases. The results incontestable that the performance of the KBNSI is superior compared to the opposite existing prediction ways. The KBNSI may be a good biological tool that may be extended to analysis on predicting and preventing the various human disorders

REFERENCES


B. Yang, H. Lin, J. Xiao et al., “The muscle-specific microRNA miR-1 regulates cardiac arrhythmogenic potential by targeting GJA1 and KCNJ2,” Nature Medicine, vol. 13, no. 4, pp. 486–491, 2007. View at Publisher · View at Google Scholar · View at Scopus


Q. H. Jiang, Y. Y. Hao, G. H. Wang, L. R. Juan, T. J. Zhang, M. X. Teng, Y. L. Liu and Y. D. Wang, BMC Syst. Biol., 2010,


P. Xuan, K. Han, M. Guo, Y. Guo, J. Li, J. Ding, Y. Liu, Q. Dai, J. Li, Z. Teng and Y. Huang, PLoS One, 2013, 8, e70204


S. Bandypadhyay, R. Mitra, U. Maulik and M. Q. Zhang, Silence, 2010,


JOBY GEORGE recived the B.Tech degree in computer science and engineering from mg university,kerala and post graduation degree in computer science and engineering from iit bombay.Currently working as Professor at ma college of engineering.His research area includes artificial intelligence,datastructure and image processing

JUJITH MP recived the B.Tech degree in computer science and engineering from calicut university,Kerala,India. He is currently pursuing post graduation degree in computer science and engineering from MA college of engineering,Kerala,India.His research interest includes Bioinformatics ,data mining and cloud computing.