miRNA based Tuberculosis Diagnosis using Machine Learning Techniques

[1]Darshana Mukundan, [2]Deepthi N, [3]Ananya Bist, [4]Ifrah Abdul Khadar Ramadurg, [5]Dr. Asha T.

[1,2,3,4]UG Student, [5]Professor and Head
Department of Computer Science and Engineering,
Bangalore Institute of Technology, Bangalore, Karnataka, India.

Abstract : According to WHO Tuberculosis (TB) is one of the top 10 causes of death globally. Every year millions of people fall prey to this disease and lose their lives. The transmission of this disease occurs at a very high rate as compared to other diseases. Using the model discussed in this paper, the disease can be detected at an early stage which can result in better treatment of the disease. The proposed model performs blood-based diagnosis of tuberculosis using a graph-based approach in conjunction with a novel signature definition and analysis method. It uses statistical techniques in order to find the miRNAs of interest.

The approach used here relies on the construction of a reference map of the transcriptional signatures of a large number of subjects, which include both healthy and affected individuals. The signatures are generated using the filtered circulating miRNAs of interest.

The construction of the reference map involves the use of Minimum Spanning Tree based clustering. A new patient is diagnosed according to the relative position of their transcriptional signature on the map. The two significant aspects which make this method the preferred choice for large scale applications such as a mass screening tool, point-of-care diagnostics are: it is minimally invasive and remains persistent to lab-to-lab protocol variability, measurement errors and batch effects. This is because the method demands accuracy only in the relative ranking of miRNA species, not in their absolute values.

IndexTerms **- miRNA, transcriptional signature, biomarker, circulating miRNA.**

I. INTRODUCTION

MicroRNAs (miRNAs) are small non-coding RNA molecules consisting of twenty-two nucleotides. It contributes in RNA silencing and post-transcriptional regulation of gene expression. Thus, dysregulation of miRNAs results in malignant transformations.

We can study the contribution of the dysregulation of miRNAs towards the development of tuberculosis. For this, we need to determine the miRNA's that fail to operate as tuberculosis suppressors. To ascertain these miRNAs, we pass DNA obtained from hematologic samples of healthy and diseased individuals into a microarray, which is a genomic tool that detects the expression of thousands of genes.

To conduct a microarray analysis, DNA molecules are gathered from both an experimental and a reference sample. DNA molecules collected are transcribed into mRNAs. Complementary DNA (cDNA) are then obtained from these two mRNA samples and each sample is marked with a different colored fluorescent dye. Following this the two samples are allowed to hybridize on the microarray slide. The microarray is then scanned using a laser to measure the signal intensities (expression) of each gene. The signal intensities determine the expression value for a particular gene, where the genes with a higher value appear differently colored from the genes with a lower value. Samples having equal expression have intermediate signal intensities. The quantized data gathered through microarrays can be used to determine which DNA and subsequently RNAs are differentially expressed, which then helps us diagnose a new subject accurately.

Machine learning provides an optimal tool to incorporate and analyze data extracted from miRNA profiles for early and effective Tuberculosis diagnosis. The advantage of using machine learning techniques is that they can capture unforeseen patterns within complex data sets. Consequently, by using graph-based machine learning techniques for relevant biological use, we discuss the promise of miRNA-based therapeutics.

II. RELATED WORKS

The disease gene prediction method proposed by Ping Luo et al [1] combines protein-to-protein interaction(PPI) network, clinical RNA-seq data and Online Mendelian Inheritance in Man(OMIM) data. Based on the OMIM database a Disease Gene Network(DGN) is then constructed. This DGN is used to identify the set of non-disease gene which are then eliminated. Wei Peng et al proposes a framework ThrRWMDE that integrates multiple biological data sources to predict miRNA-disease association. This involves construction of three types of networks including miRNA similarity networks, disease similarity networks and environmental factors similarity networks. An unbalanced three random walk is implemented on the three types of networks. This helps to identify new miRNA-diseases associations or environmental factor disease associations. This is done by gathering information from neighbors in the same network or nodes in the other networks [2]. Chuan Wang et al [3] proved that clinical classification or age and gender did not help in identifying clear distinction between the samples . This result is used for further analysis in the model proposed in the paper. describes a rank based tanscriptional signature as a novel technique for diangostic biomarker definition and analysis. Mario Lauria et al provides a novel approach for RNA signature definition and analysis which addresses the issue of high sensitivity of expression profiling method to protocol varaition. This method was, tested upon multiple diseases [4]. Mario Lauria et al [5] summarises the characteristics of each subjects and then uses the circulating miRNA to compute the transcriptional signatures of an individual from the subjects. To calculate the signature for a perticular profile, the miRNA species for that profile is ranked according to the values, with higher valued receiving a better rank. This model applies the method for signature generation proposed in [4] for the diagnosis of Tuberculosis.

III. METHODOLOGY

The issue with working with real-world data is that they are usually inconsistent, incomplete, and inadequate in certain behaviors or trends, and more often than not they contain many errors. This makes data pre-processing a crucial step as it is a proven method for resolving such issues. Our input to the application are text (.txt) files obtained from Agilent Human miRNA Microarray platform. Each text file contains the miRNA data obtained by the microarray for a particular individual. While the data set is itself very diverse, proper implementation of this step ensures that redundant data and features are eliminated.

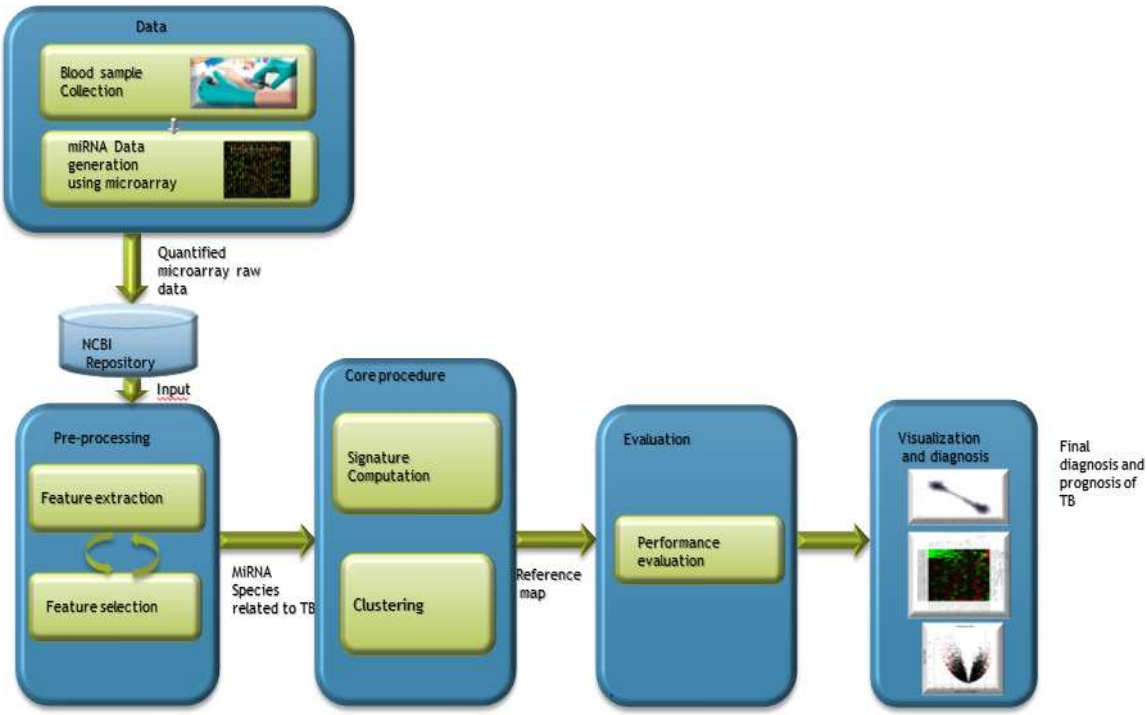

Figure 1. Architecture of the proposed model

We considered the miRNA profiles published by two separate research groups, where one set of profiles was used as the training set and the second as the test set. Data set collected from NCBI repository with GEO accession number GSE29190 (6 active TB profiles, 6 latent TB profiles and 3 healthy profiles) [Wang 2011] was used for training, and dataset with GEO accession number GSE34608 [Maertzdorf 2012] was used for testing. Implementation of the first step involves the removal of attributes like transcription_id, locus of the miRNA, etc using filtration methods. In addition, we derive additional features from the original features. Statistical techniques are then applied to obtain the miRNA's of interest. p-value of 0.08 was used to obtain 64 statistically significant miRNA's.

After filtration is done to obtain the miRNA's of relevance, the next step is signature generation in which subject specific signature is obtained by taking n1(20) most expressed and n2(20) least expressed miRNA. The Enrichment Score metric was used to represent the distance between each pair of subjects. The Enrichment score was used to generate an all-to-all distance matrix. This matrix quantifies the degree of similarity between a pair of signatures [5]. The Enrichment score is computed based on the Kolmogorov-Smirnov running sum. Both top and bottom of the signature contribute towards the calculation of the Enrichment score. The Enrichment score is therefore the average of both parts calculated separately against each of the other samples, therefore ES = (ES_top -ES_bottom)/2. The KS test determines if the miRNA between the 2 samples are randomly distributed or concentrated at the top and bottom. This test has an advantage of making no assumption about the underlying distribution of data [4].

In the map construction step, the distance matrix is employed to construct a minimum spanning tree. The MST is then partitioned into two distinct clusters based on prior knowledge about the two groups of subjects in the training data (healthy and affected). The distance between the new subject (test data) and each of the other subjects in the training data is computed using the same distance computation method previously used to compute distance matrix. This is then used to identify the node the test subject is nearest to, and since the cluster to which the nearest node belongs to is already known, it becomes the deciding parameter to determine which cluster the test subject associates to. The test subject is then assigned to the corresponding cluster. The cluster the new subject associates to gives us the diagnosis. To better represent the results obtained throughout the clustering process, data visualization is pivotal. The results including the Minimum spanning tree, clusters obtained and the position of the new subject in the clusters are plotted as networkx graphs.

IV. RESULTS

The signature generation step yields a distance matrix, represented as a heat-map shown in Figure 2. We notice that the diagonal elements are shaded black, as the distance of a subject to itself is zero. We also notice that the subjects with a high degree of similarity are darker, and that as the distance increases, the intensity of the color decreases.

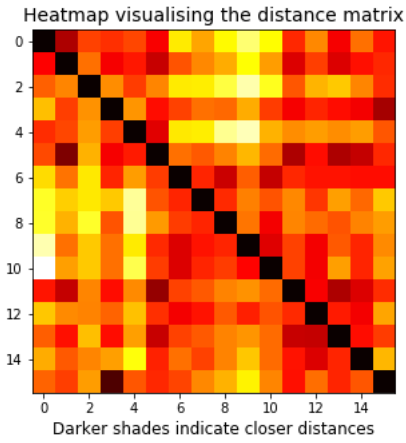

Figure 2. Heatmap of computed distance matrix

Figure 3 represents the clustering step produces two distinct clusters, namely healthy (represented in green) and affected (represented in red).

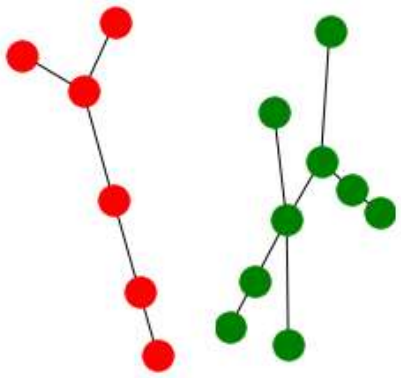

Figure 3. Cluster created using MST

When placed in the reference map as shown in the figure, the new subject (represented in yellow) associates itself with one of the two clusters. In Figure 4, the subject is seen to associate with the affected (red) cluster, and hence the diagnosis is TB positive.

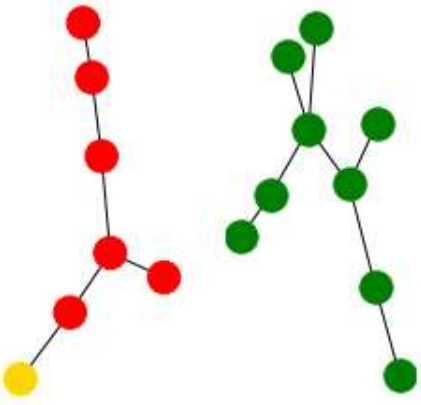

Figure 4. Diagnosis of test data based on clusters

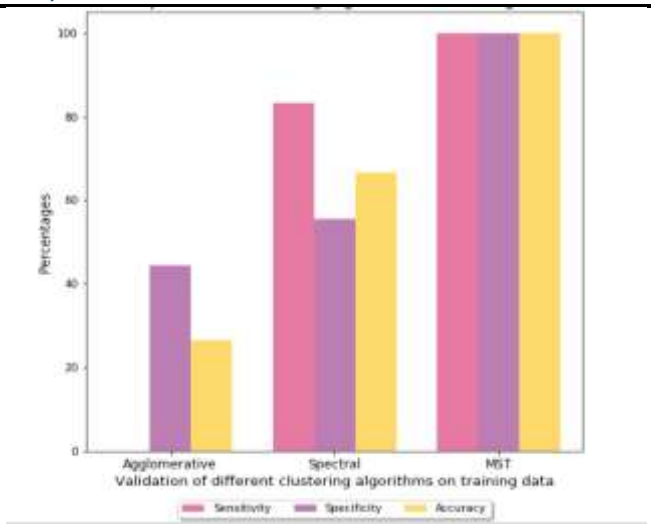

Figure 5. Evaluation of clustering algorithms

Figure 5 indicates the sensitivity,specificity and accuracy of some of the possible clustering algorithms employed for training the model. The clustering algorithms that were implemented include: Agglomerative, Spectral and Minimum Spanning Tree. The disadvantage of using agglomerative was that sensitivity was nil, while spectral provided a comparitively lower accuracy. MST based clustering was preferred because the threshold could be explicitly set, giving rise to a higher degree of accuracy and better demarcation of clusters.

V. Conclusion

The proposed model implements a four-step diagnosis method, in which the pre-processing step is the most computationally intensive. For a modest number of samples, the methodology employed for signature definition and analysis is not as computationally rigorous.The MST based clustering algorithm presented here provides better user control to obtain clear, distinct clusters. This we know from our sample experiment done on a set of subjects (active=6, latent and healthy=9, test=5). Visualization of the application and the results make it simpler and easier to comprehend. A limitation of the proposed method is that it is has been implemented and tested with a marginally small (train=15, test=5) number of subjects, and that the true prowess of the MST clustering algorithm implemented on a larger dataset cannot be predicted.

Future enhancement and application of the method discussed in this model is that the progression of the treatment can be better and easily studied by using real-time mapping of the current state with the existing reference maps. In addition, a provision to externally validate the results generated using medical diagnosis can make the system even more intelligent.

References

[1] Ping Luo, Li-Ping Tian, Jishou Ruan, Fang-Xiang Wu "Disease Gene Prediction by Integrating PPI Networks, Clinical RNA-Seq Data and OMIM Data", IEEE/ACM Transactions on Computational Biology and BioInformatics, Vol.16, Issue 1, pp:222-232, © January 2019

[2] Wei Peng, Wei Lan, Zeng Yu, JianXin Wang, Yi Pan, "A Framework for Integrating multiple biological networks to predict microRNA-disease associations", in IEEE Transactions on NanoBioScience, Vol.16, Issue 2, pp:100-107, © 2016.

[3] Chuan Wang, Shunayao Yang, Gang Sun, Xuying Tang, Shuihua Lu, Olivier Neyrolles, Qian Gao "Comparative miRNA Expression Profiles in Individuals with Latent and Active Tuberculosis", PLoS one, vol. 6,Issue 10, 2011

[4] Mario Lauria, "Rank-based transcriptional signatures: A novel approach to diagnostic biomarker definition and analysis." Systems Biomedicine, Vol.1, Issue 4, pp: 228-239, © 2013

[5] Mario Lauria ,"Rank-based miRNA signatures for blood-based diagnosis of tuberculosis", IEEE 978-1-4244-9270-1/15,© 2015

[6] Lijun Chang, Wei Li, Lu Qin, Wenjie Zhang, and Shiyu Yang, "Fast and ExactStructural Graph Clustering" pSCAN: in IEEE Transactions On Knowledge And Data Engineering, Vol. 29, No. 2, pp-387-401, ©February 2017

[7] Pedro López-Romero, Manuel A González, Sergio Callejas, Ana Dopazo, Rafael A Irizarry "Processing of Agilent microRNA array data" in BioMedCentral , doi:10.1186/1756-0500-3-18, © 2016

[8] Biplab Banerjee, Surender Varma, Krishna Mohan Buddhiraju, and Laxmi Narayana Eeti "Unsupervised Multi-Spectral Satellite Image Segmentation Combining Modified Mean-Shift and a New Minimum Spanning Tree Based Clustering Technique", IEEE Journal Of Selected Topics In Applied Earth Observations And Remote Sensing, VOL. 7, NO. 3, MARCH 2014