INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN : 2454-4752 P - ISSN : 2454-4744
VOL 2 ISSUE 1 (2016) PAGES 461-466
RECEIVED : 041/01/2016. PUBLISHED : 13/02/2016
February 13, 2016

MicroRNA EXPRESSION ANALYSIS WITH DATA MINING CONCEPTS

[1S.] Geeitha*, Assistant Professor, Mahendra Engineering College for Women, Tamilnadu, India
[2]Dr. M. Thangamani, Assistant Professor, Kongu Engineering College, Tamilnadu, India

ABSTRACT

MicroRNAs (miRNAs) involved are small non-coding RNAs that cause mRNA degradation and translational inhibition. They are important regulators of development and cellular homeostasis through their control of diverse processes. Recently, great efforts have been made to elucidate their regulatory mechanism, but the functions of most miRNAs and their precise regulatory mechanisms remain elusive. Today miRNA is considered as potential therapeutic targets and potential diagnostic biomarkers. This research tries to enumerate miRNA targets in several diseases and to explore the degree of miRNA regulations on them by implementing data mining technique.

Keywords: miRNA, Gene expression and Data mining

1. INTRODUCTION

Genome consists of protein-coding genes that code for MicroRNAs and relatives of small RNAs. Vast majority of MicroRNAs adjust other genes by binding to complementary sequences in the target gene. For the given DNA sequence, predicting MicroRNA genes and their respective targets are big issues in biomedical. In order to achieve this, it needs to analyse the MicroRNA and requires intelligent mining techniques.

1.1 MicroRNAs

MicroRNAs constitute a recently discovered class of non-coding RNAs that play key roles in the regulation of gene expression. Acting at the post-transcriptional level, these fascinating molecules may fine-tune the expression of as much as 30% of all mammalian protein-encoding genes. MicroRNA genes are transcribed by RNA polymerase II as large primary transcripts that are processed by a protein complex containing to form approximately 70 nucleotide precursor microRNA. This precursor is subsequently transported to the cytoplasm where it is processed by a second RNase III enzyme, DICER, to form a mature microRNA of approximately 22 nucleotides. The mature microRNA is then incorporated into a ribonuclear particle to form the RNA-induced silencing complex, RISC, which mediates gene silencing.

1.2 Data mining and Medical Informatics

Data mining is an essential step in the process of knowledge discovery in databases in which intelligent methods are applied in order to extract patterns. Medical informatics plays a very vital task in the use of clinical data. In such discoveries pattern recognition is important for the diagnosis of new diseases and the study of different patterns found during the classification of data. Computer assisted

information retrieval may help to support quality decision making and to avoid human error. This leads to the use of data mining in medical informatics, the database that is found in the hospitals, namely, the hospital information systems containing massive amounts of information which includes patients' information, data from laboratories which keeps on growing year after year. With the help of data mining methods, useful patterns of information can be found to detect and identify the disease.

2. RELATED WORKS

Many researches in MicroRNA is carried out by researchers and global biomedical organization. However, many challenges are still under experimental conditions.

2.1 MicroRNA Review

MicroRNA expression and sequence analysis database (mESAdb) is a regularly updated database for the multivariate analysis of sequences and expression of microRNAs provide a series of interactive analysis tools for testing the association of microRNA sequence characteristics with target gene function, human diseases and microRNA expression patterns using multivariate analyses. mESAdb is also a meta-analysis tool for comparative analysis of function and expression for microRNA lists across different organisms including human, mouse and zebrafish[1]. Human disease genes are separated which depends on the miRNA target to the disease genes in two groups: miRNA-targeted disease genes and miRNA non-targeted disease genes in order to find out the disease class which is mostly targeted by miRNA among all disease classes. A total of 301 miRNA targeted disease genes are broadly categorized into eight different disease classes according to the Human Gene Mutation Database (HGMD) [2].

Regression analysis has established AU-rich elements as the most influential genomic property that determines miRNA hits on disease genes. Jyotirmoy Das et.al. [3], investigated the cause and fate of miRNA targets on cancer genes that helps in enhancing the knowledge and medicinal improvement of cancer genes. miRNA expression data have been retrieved from Cancer Miner database which contains microarray expression data [4] for ten tissues of cancer patients. Then the expression values are averaged over tissues and then mapped the miRNA expression levels with the dataset.

2.2 Review of datamining using MicroRNA

Data mining provides support for constructing a model for managing the hospital resources which is an important task in healthcare. Using data mining, it is possible to detect the chronic disease and based on the complication of the patient disease, prioritization can be made easily so that they will get effective treatment in timely and accurate manner. Fitness report and demographic details of patients are also useful for utilizing the available hospital resources effectively. An automated tool using data mining is proposed by J.Alapont et al., for managing hospital resources such as physical and human resources Group Health Cooperative provides various healthcare services at lower cost using data mining techniques [5]. Due to analytical and descriptive ability, Data Mining is widely used in medical field. Healthcare providers utilize the data mining tools to make effective decision regarding how to enhance

the patient health, how to provide health care services at low cost and how to predict fraud in health insurance etc.. Before applying classification technique, there is a need to recognize the redundant and inappropriate attributes because these attributes act as a noise and outlier which in turn slow down the processing task. These attributes also have an adverse affect on the performance of classifier. Statistical methods are used for recognizing these attributes.

The most relevant and useful attributes can be recognized by feature selection methods which in turn enhance the performance and accuracy of classification model. In the existing system there is no single classifier which produces the best result for every dataset. In order to check the performance of classifier, a dataset is divided into two parts- training and testing. So, a classifier is selected only when it produces better performance among all classifiers. The performance of a classifier is evaluated using testing data set. But there occurs problem with testing data set. Sometimes, it becomes complex and sometimes it becomes easy to classify the testing data set. The performance of classifier depends on the testing data set. In order to avoid these problems, cross validation can be used for both training and testing the data set [6].

The existing paper emphasizes the statistical methods in the analysis of the associations of miRNA gene expression with human cancers and related clinical phenotypes. Simple statistical methods include chi-square test, correlation analysis, t-test and one-way ANOVA. Regression models include linear and logistic regression survival analysis approaches such as non-parametric Kaplan-Meier method, log-rank test and semi-parametric Cox proportional hazards models have been used for time to event data. A Multivariate method such as cluster analysis has been used for clustering samples and principal component analysis (PCA) for data mining. Bayesian statistical methods have recently made great inroads into many areas of science [7], including the assessment of association between miRNA expression and human cancers and multiple testing. A weighted rule-mining technique to rank the rules using two novel rule-interestingness measures, viz., rank-based weighted condensed support and weighted condensed confidence measures to bypass the problem. These measures are basically depended on the rank of genes. Using the rank, we assign weight to each item. RANWAR [8] generates much less number of frequent item sets than the state-of-the-art association rule mining algorithms. Thus, it saves time of execution of the algorithm. RANWAR techniques run on gene expression and methylation datasets. The genes of the top rules are biologically validated by Gene Ontologies and Kyoto Encyclopedia of Genes and Genomes pathway analyses.

The interface provided by the PuTmiR [9], web server facilitates a vital resource for analyzing the direct and indirect regulation of human miRNAs. While it is already an established fact that miRNAs are regulated by binding to their upstream region, this database might possibly help to study whether an miRNA can also be controlled by the binding to their downstream region. A data mining algorithm [10], has been used to explore the data of miRNA. Specifically, the author's aim is to provide the biologists with a tool that can support them in two challenging tasks, that is, the detection of actual miRNAs target genes and the identification of the context-specific co-associations of different miRNAs. A further contribution to the considered research consists in the ranking of the extracted biclusters on the basis of the semantic similarity between the target genes, which allows the biologists to select the most significant results easily from a biological opinion. Incremental Fuzzy Mining is tested with real expression datasets

for both classification and clustering tasks [11]. In this case, uncover hidden patterns are identified and cancer related outliers are detected. miRNA and mRNA expression for the filtering of sequence-based putative targets have been anticipated[12]. These models are useful in identifying the most prominent interactions from the databases of putative targets by dependency analysis, linear regression and Bayesian approaches.

Clustering algorithms are used for examining gene expression data. It is also intended to introduce one of the main problems in medical informatics as clustering gene expression data to the operations research community. Clustering may be performed by grouping genes over samples or samples over genes. Since the number of genes is normally thousands and many of the genes have low or invariant expression values, filtering gene expression data to reduce the dimension of the $n \times m$ matrix is often necessary. Gene interactions may be represented by graphs using an adjacency matrix [13]. Lee H et. al produced cloud workbench to develop an environment for the integrated analysis of microRNA and mRNA expression data, named BioVLAB-MMIA. The workbench facilitates computations on the Amazon EC2 and S3 resources orchestrated by the XBaya Workflow Suite. It services as computational tool to provide on-demand cloud computing resources and workflow. The author states that BioVLAB-MMIA will be an easy-to-use computing environment for researchers who plan to perform genome-wide microRNA-mRNA (gene) integrated analysis tasks [14]. Another technique implemented is an efficient multivariate filtering designed to analyze the topological properties of a co-expression network in order to identify potential relevant genes for a given disease [15]. In this research, acute myeloid leukemia, breast cancer, and diffuse large B-cell lymphoma data sets are used.

Data mining of clinical biochemical examined miRNA results for high-risk population with cancer and cardiovascular disease[16].Hence data mining techniques such as classification and clustering play a vital role in recognizing the pattern analysis in miRNA expression that acts as therapeutical treatment. Predicting the outcome of a disease is one of the most interesting and challenging tasks in developing data mining applications. The use of computers with automated tools helps in collecting the large volumes of medical data and making them available to the medical research groups [17].

REFERENCES

[1]. Koray D. Kaya, Gökhan Karakülah, Cengiz M. Yakıcıer, Aybar C. Acar and Özlen Konu, mESAdb: microRNA Expression and Sequence Analysis Database, Nucleic Acids Research, Vol.39, Pp.170–180., 2011.

[2]. Tenson PD, Ball EV, Mort M, Phillips AD, Shaw K, Cooper DN, The human gene mutation databaseand its exploitation in the fields of personalized genomics and molecular evolution. Curr Protoc Bioinformatics, Chapter.1, 2012.

[3].Jyotirmoy Das, Soumita Podder and Tapash Chandra Ghosh, Insights into the miRNA regulations in human disease genes, BMC Genomics, Vol.15, No.1, 2014.

[4]. Jacobsen A, Silber J, Harinath G, Huse J, Schultz N, Sander C, Analysis of microRNA-target Interactions across diverse cancer types. Nat Struct Mol Biol, Vol. 20, No.11, Pp.1325–1332, 2013.

[5]. C. Koh and G. Tan, Data Mining Application in Healthcare, Journal of Healthcare Information Management, vol. 19, No. 2, pp.64-72, 2005.

[6]. Divya Tomar and Sonali Agarwal, A survey on Data Mining approaches for Healthcare, International Journal of Bio-Science and Bio-Technology Vol.5, No.5, Pp. 241-266, 2013.

[7]. Keshena Wang1, Yue Pan2 and Chun Xu3 ,Statistical Modeling of MicroRNA Expression with Human Cancers, Journal of Biometrics & Biostatistics,No.10, pp.2015.

[8].Ujjwal Maulik,Saurav Mallik, Anirban Mukhopadhyay ,RANWAR: Rank-Based Weighted Association Rule Mining From Gene Expression and Methylation Data, NanoBioscience, IEEE Transactions on ,Vol.14 , No.1, pp.1-8, 2014.

[9]. Sanghamitra Bandyopadhyay and Malay Bhattacharyya,PuTmiR: A database for extracting neighboring transcription factors of human microRNAs ,BMC Bioinformatics,Vol.1,2010

[10]. Gianvito Pio, Michelangelo Ceci, Corrado Loglisci, Donato Malerba, Domenica D'Elia, The integration of microRNA target data by biclustering techniques opens new roads for signaling networks analysis, EMBET .journal, Vol.18, No.11, Pp. 3598-3613, 2012.

[11]. K. Upendra Babu, R. Rajeswari, Dr. G. GunaSekaran ,A Survey on Data Mining Of Gene Expression Data for Gene Function Prediction, International Journal of Innovative Research in Computer and Communication Engineering Vol. 3, No.14, 2015.

[12]. Ander Muniategui, Jon Pey, Francisco Planes and Angel Rubio, Joint analysis of miRNA and mRNA expression data, Briefings in Bioinformatics Advance Access, No.10, 2012.

[13]. Harun Pirim,1,* Burak Ekşioğlu, Andy Perkins, and Çetin Yüceer ' Clustering of High throughput Gene Expression Data, Computer and Operation Research, Vol.39, No.12,2013.

[14]. Lee H1, Yang Y, Chae H, Nam S, Choi D, Tangchaisin P, Herath C, Marru S, Nephew KP, Kim S, BioVLAB-MMIA: a cloud environment for microRNA and mRNA integrated analysis (MMIA) on Amazon EC2,IEEE Trans Nanobioscience. Vol. 11, No.3, Pp.266-72, 2012.

[15]. Alfredo Benso, Paolo Cornale, Stefano Di Carlo, Gianfranco Politano, and Alessandro Savino ,Reducing the Complexity of Complex Gene Coexpression Networks by Coupling Multiweighted Labeling with Topological Analysis, BioMed Research International, No.10, 2013.

[16]. Qing Ang , Wei-dong Wang , Bo-ya Zhao , Jing Li , Kai-yuan Li, Application of data mining based on clinical medicine database, International conference on Signal Processing Systems , Vol.3 , Pp. 719-723, 2010.

[17]. Shweta Kharya, Using Data Mining Techniques for Diagnosis and Prognosis of Cancer disease, International journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, Pp.55-65, 2012.

Authors Biography

Ms. S. Geeitha has completed Master of Engineering in Computer Science and Engineering in Anna University Application. Her research expertise covers Medical data mining, machine learning, cloud computing, big data, fuzzy, soft computing and ontology. She has presented 10 papers in national and international conferences in the above fields. She is currently working as Assistant Professor in Mahendra Engineering College for Women.

Dr. M. Thangamani possesses nearly 20 years of experience in research, teaching, consulting and practical application development to solve real-world business problems using analytics. Her research expertise covers Medical data mining, machine learning, cloud computing, big data, fuzzy, soft computing, ontology development, web services and open source software. She has published nearly 70 articles in refereed and indexed journals, books and book chapters and presented over 67 papers in national and international conferences in the above field. She has delivered more than 60 Guest Lectures in reputed engineering colleges on various topics. She has got best paper awards from various education related social activities in India and Abroad. She has organized many self-supporting and government sponsored national conference and Workshop in the fields of data mining, big data and cloud computing. She continues to actively serve the academic and research communities. She is on the editorial board and reviewing committee of leading research journals, which includes her nomination as the Associate Editor to International Journal of Entrepreneurship and Small & Medium Enterprises at Nepal and on the program committee of top international data mining and soft computing conferences in various countries. She is also seasonal reviewer in IEEE Transaction on Fuzzy System, international journal of advances in Fuzzy System and Applied mathematics and information journals. She has been nominated as chair and keynote speaker in international conferences in India and countries like Malaysia, Thailand and China. She has Life Membership in ISTE, Member in CSI, International Association of Engineers and Computer Scientists in China, IAENG, IRES, Athens Institute for Education and Research and Life member in Analytical Society of India. She is currently working as Assistant Professor at Kongu Engineering College at Perundurai, Erode District.