World Academy of Science, Engineering and Technology
International Journal of Medical and Health Sciences
Vol:10, No:2, 2016

Prediction of MicroRNA-Target Gene by Machine Learning Algorithms in Lung Cancer Study

Nilubon Kurubanjerdjit, Nattakarn Iam-On, Ka-Lok Ng

***Abstract*—**MicroRNAs are small non-coding RNA found in many different species. They play crucial roles in cancer such as biological processes of apoptosis and proliferation. The identification of microRNA-target genes can be an essential first step towards to reveal the role of microRNA in various cancer types. In this paper, we predict miRNA-target genes for lung cancer by integrating prediction scores from miRanda and PITA algorithms used as a feature vector of miRNA-target interaction. Then, machine-learning algorithms were implemented for making a final prediction. The approach developed in this study should be of value for future studies into understanding the role of miRNAs in molecular mechanisms enabling lung cancer formation.

***Keywords*—**MicroRNA, miRNAs, lung cancer, machine learning, Naïve Bayes, SVM.

I. Introduction

MICRORNAS (miRNAs) are a class of naturally occurring, small non-coding RNA endogenous molecules of ribonucleic acid found in eukaryotic cells [1], about 21-25 nucleotides in length. MiRNAs function is to downregulate gene expression. The translational inhibition by miRNAs has been thought of as a major mechanism in animal systems while mRNA degradation regulation has been considered as a major regulatory mechanism in plants [2]. MiRNAs play crucial roles in wide range of aspects of cancer biology, such as proliferation, apoptosis, invasion, and angiogenesis [3].

References [4] and [5] used microarray analysis to reveal the significantly different miRNAs profiles in cancer cells compared with those in normal cells in the same tissue. Their work indicated that miRNAs signature profiling enabled the tumor tissue samples. Moreover, several previous research works performed microarray analysis on various types of cancers such as breast, leukemia, colon, lung [6]-[8]. Reference [9] identified that miRNA expression was correlated with specific breast cancer such as estrogen and progesterone receptor expression, tumor stage and proliferation. Besides, [10] found that the down-regulation of let-7d, miR-210 and miR-221 in ductal carcinoma *in situ* while they were up-regulated in the invasive transition. Reference [11] identified 43 miRNAs that were differentially expressed in microarrays between normal lung and non-small-cell lung cancer (NSCLC) pairs. Moreover, [12] reported that miR-30, miR-7i and miR-126 significantly down-regulated in squamous cell lung cancer. The recent studies report that miRNAs expression changes induced by cigarette smoke may be prevented by N-acetylcysteine, oltipraz, indole-3-carbinol, 5,6-benzoflavone, phenethyl isothiocyanate and budesonide [13], [14]. Previous studies reveal that lung cancer biomarkers miRNAs are the most promising because of remarkable stability and cancer-type specificity [15]. Lung cancer has a poor prognosis, therefore studies in this area are necessary to be successful in prospective clinical applications.

Many computational techniques have been discovered to predict miRNA-target gene, multiple factors are introduced to identify their target genes such as complementarily of different regions on miRNAs, binding site conservation and also target sites accessibility. Different predictive algorithms are based on different factors, therefore, integrating diverse algorithms may improve target prediction. MiRanda is an algorithm written in C for finding target genes for miRNAs. This algorithm was developed at the Computational Biology Center of Memorial Sloan-Kettering Cancer Center [16], [17]. MiRanda identifies miRNA-target genes based on sequence complementarily and conservation of target sites; whereas, PITA [18] predicts miRNA-target gene by calculate the free energies of RNA-RNA duplexes, PITA gives matching scores to multiple biding sites, therefore, optimal combination of different algorithms may improve the prediction performance. In this study, we use a combination approach for identifying miRNA target of lung cancer gene which is adopting miRNA-target prediction algorithms and also machine learning algorithms (ML).

II. Methodology

A. Data Sources

The 738 predicted lung cancer genes list was obtained from [19] and their FASTA sequence was obtained from Uniprot [20]. A whole set of human miRNAs sequence was downloaded from mirBase [21], list of 39,111 experimentally confirmed miRNA-target pairs was obtained from mirTarBase [22].

B. System Workflow

In this study, we integrated various types of approaches to identify miRNA-target gene to reveal the essential biological process of miRNA related to lung cancer. Firstly, training set was generated by inputting experimentally confirmed miRNA sequences and their target FASTA sequences into PITA and miRanda to get the prediction scores. The miRNA-target pairs

Nilubon Kurubanjerdjit and Nattakarn Iam-On are with the School of Information Technology, Mea Fah Luaung University, Chiangrai, Thailand (e-mail: sendtoopal@gmail.com, nt.iamon@gmail.com).

Ka-Lok Ng is with the Department of Bioinformatics and Biomedical Engineering, Asia University, Taichung, Taiwan and Department of Medical Research, China University Hospital, China Medical University, Taichung, Taiwan (e-mail: ppiddi@gmail.com).

that satisfied these two algorithms were filtered as a positive set. Negative set is the pairs that satisfied those two algorithms with the positive set subtracted. Secondly, test set was prepared by the two predictors, a whole set of miRNAs sequences and a set of cancer associated genes FASTA sequences were submitted into predictors. The miRNA-target pairs with two prediction scores that satisfied the predictors were extracted to be the test set. Thirdly, Naïve Bayes and Support Vector Machine (SVM) are selected to classify the final prediction results. The training set was submitted into these two classifiers with optimum parameter setting in order to build up the classification models. Next, the test set was submitted to the two classifiers models to give to final prediction result.

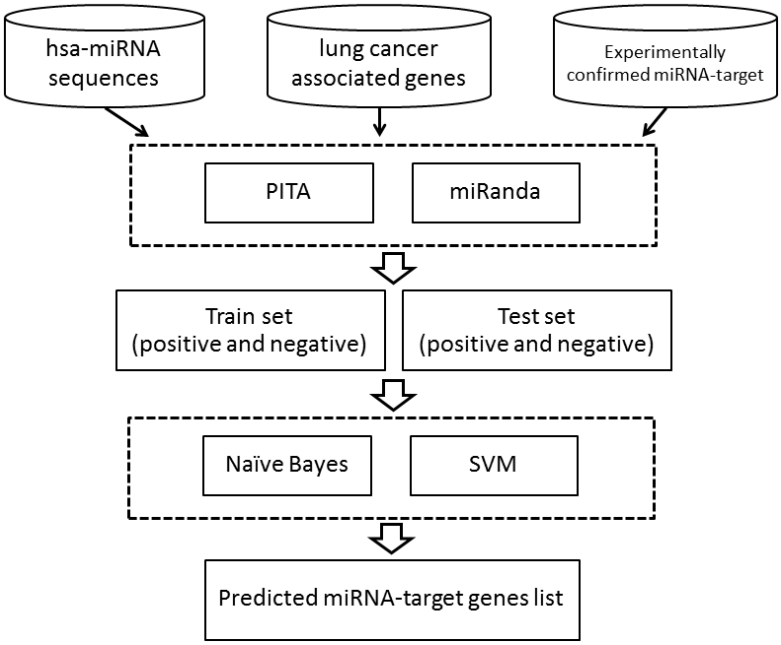

Fig. 1 System flowchart of this work

C. MiRanda and PITA Parameter Setting

Three parameters are required for miRanda execution; threshold score, MFE and scaling factor, which are given as 80, 14 kcal/mol and 2.0 respectively. The max score was observed in this study, the higher score is the better biding between miRNA and their target gene.

For PITA, sequence of miRNA and UTR were analysed by default parameter setting. In case of single binding site, the score is given by $\Delta\Delta G$ value whereas multiple binding sites, the score were determined by minimum value of its binding.

D. Training Set and Test Set Generation

A comprised set of 37,443 experimentally confirmed miRNA-target pairs was downloaded from mirTarBase. These pairs were derived from a set of 596 miRNAs and 12,104 mRNAs. These sets were processed by the two predictors; miRanda and PITA. Then, predicted miRNA-target pairs are merged. Positive set (195 pairs) are experimentally confirmed pairs that satisfied the two algorithms. Negative set is a total of 442 pairs that satisfied two algorithms with the positive set subtracted.

The test set was generated by the two predictors for 738 cancer associated genes and a whole set of human miRNA.

E. Machine Learning Process

The open-access software RapidMiner was adopted as a tool for classification in this work. Two classification algorithms; Naïve Bayes and Support Vector Machine (SVM) were selected to predict which miRNAs are likely to target which associated lung cancer genes.

For Naïve Bayes parameter setting, estimation mode is set to greedy, value of minimum band-width, Number of kernels and right is set to 0.1, 10, and 19 respectively. There are five parameters for SVM to be set; Kernel type is set to radial, Kernel gamma, Kernel cache, C value and Convergence epsilon are set to 1.0, 200, 0.5 and 0.001 respectively.

Once the two classifiers were optimized, they were adopted to evaluate the test set. MiRNA-target pairs obtained from these two classifiers were integrated and only the pairs that satisfy the two classifiers were filtered as a final result.

III. Results

A. Prediction of miRNA-Target Gene Interaction

It was found that Naïve Bayes achieves the best performance at 80.03% of accuracy and SVM achieves the best performance at 82.40% of accuracy. A total of 884 pair was predicted by Naïve Bayes and SVM. There are only novel 18 miRNA-target pairs were predicted by both classifiers, 254 novel pairs satisfied Naïve Bayes and 62 novel pairs satisfied SVM.

B. Novel miRNA Co-Regulated Target Genes

Fig. 2 depicts RASA1 is targeted by two miRNAs; miR-142-3p and miR-217. RASA1 involves in regulation of apoptosis, and regulation of programmed cell death.

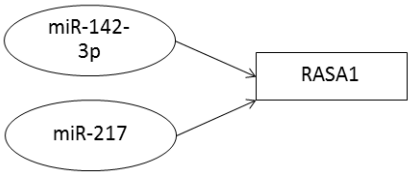

Fig. 2 Co-regulated miRNAs of RASA1

Fig. 3 depicts CD46 is targeted by two miRNAs; miR-140-5p and miR-139-5p. CD46 involves in defence response process. PAK2 involves in cell death, apoptosis, cell division and also ErbB signaling pathway recorded by KEGG.

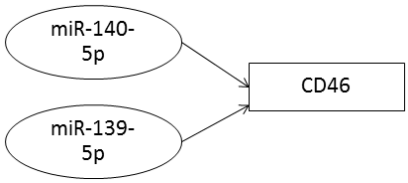

Fig. 3 Co-regulated miRNAs of CD46

C. Novel miRNA-Target Genes

This experiment found that some miRNAs directly regulate lung cancer proteins; miR133b, miR139-5b, miR-142-3p, miR-199a-5p, miR-206, miR-217 and miR-140-5p. Fig. 4 depicts miR-133 target genes. It targets to CD44 and CRK which are lung cancer protein. CD44 is lung cancer protein that involves in defense response, regulation of apoptosis, regulation of programmed cell death, and regulation of cell death. Besides, CRK involves in KEGG pathway of cancer, prostate cancer, and ErbB signalling.

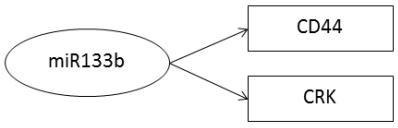

Fig. 4 MiR-133b and target genes

Fig. 5 shows miR-139-5p target genes. CD46 involves in defense response process. PAK2 involves in cell death, apoptosis, cell division and also ErbB signaling pathway recorded by KEGG.

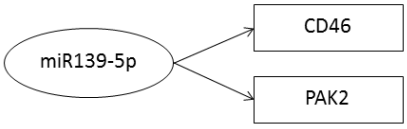

Fig. 5 Mir-139-5p and target genes

Fig. 6 shows miR-199a-5p target gene. BAG1 involves in regulation of apoptosis, and regulation of programmed cell death. MAPK9 involves in Pancreatic cancer, and ErbB signaling pathway recorded by KEGG.

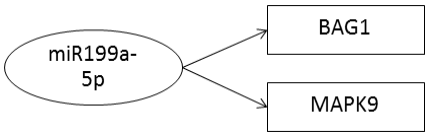

Fig. 6 MiR-199a-5p and target genes

Fig. 7 depicts miR-206 target genes. BST2 involves in regulation of protein kinase cascade. PLCG1 involves in non-small cell lung cancer and glioma pathway recorded by KEGG.

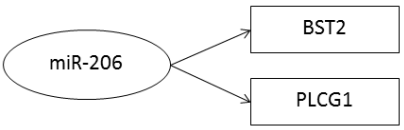

Fig. 7 MiR-206 and target genes

Fig. 8 shows miR-217 target genes. RASA1 involves in regulation of apoptosis, and regulation of programmed cell death.

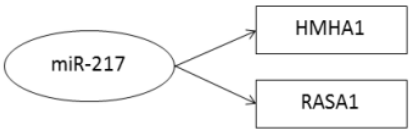

Fig. 8 MiR-217 and target genes

IV. Conclusions

miRNAs are emerging as key components in gene regulatory pathways in human cancers. They play important roles in a variety of human cellular process such as apoptosis, proliferation and programmed cell death. Our evidence indicates that miR133b, miR139-5b, miR-142-3p, miR-199a-5p, miR-206, miR-217 and miR-140-5p target to lung cancer proteins that involve in crucial cancer biological processes such as defence response, regulation of apoptosis, regulation of programmed cell death, and regulation of cell death. The approach introduced in this work and also the results should be of value for future studies to reveal the role of miRNAs in cancer study.

Acknowledgment

The work of Nilubon Kurubanjerdjit and the work of Natthakan Iam-On are supported by Mea Fah Luang University, Chiang Rai, Thailand. The work of Ka-Lok Ng is supported by the Ministry of Science and Technology of Taiwan (MOST) under grants MOST 102-2632-E-468-001-MY3 and MOST 104-2221-E-468-012, and also supported by Asia University under the grants 103-asia-06.

REFERENCES

[1] D. P. Bartel, "MicroRNAs: genomics, biogenesis, mechanism, and function", Cells, 166 (2): 281–297 (2004).
[2] X. Dai et al., "Computational analysis of miRNA targets in plants: current status and challenges", Briefings Bioinformatics, 12(2), 115-212 (2010).
[3] Y. S. Lee and A. Dutta, "MicroRNAs in cancer", Annu Rev Pathol, 4, 199-227 (2009).
[4] SM. Hammonad, "microRNA detection comes of age", Nat Methods, 3(1), 12-13 (2006).
[5] C. G. Liu, "An oligonucleotide microchip for genome-wide microRNA profiling in human and mouse tissue", Proc Natl Acad Sci USA, 101(26), 9740-9744 (2004).
[6] M. V. Iorio et al., "MicroRNA gene expression deregulation in human breast cancer", Cancer Res, 65(16), 7065-7070 (2005).
[7] H. He et al., "The role of microRNA genes in papillary thyroid carcinoma", Proc Natl Acad Sci USA, 102(52), 19075-19080 (2005).
[8] G. A. Calin, "Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers", Proc Natl Acad Sci USA, 101(9), 2999-3004 (2004).
[9] M. V. Iorio et al., "MicroRNA gene expression deregulation in human breast cancer", Cancer Res, 65(16), 7065-7070 (2005).
[10] S. Volinia et al., "Breast cancer signatures for invasiveness and prognosis defined by deep sequencing of microRNA", Proc Natl Acad Sci USA, 109(8), 111-114 (2012).
[11] N. Yanaihara et al., Unique microRNA molecular profiles in lung cancer diagnosis and prognosis, Cancer Cell, 9, 189-198 (2006).
[12] Y. Yang et al., The role of microRNA in human lung squamous cell carcinoma, Caner Genet. Cytogenet, 200, 127-133 (2010).
[13] A. Izzotti et al., Chemoprevention of cigarette smoke-induced alterations of MicroRNA expression in rat lungs, Cancer Prev.Res. (Phila, PA), 3, 62-72 (2010).
[14] A. Izztti et al., Modulation of microRNA expression by budesonide phenethyl isothiocyanate and cigarette smoke in mouse liver and lung, Carcinogenesis, 31, 894-901 (2010).
[15] G. Malgorzata et al., MicroRNA-Role in Lung Cancer, Diseas Markers, Article ID 218169, 13 (2014).
[16] A. J. Enright et al., "MicroRNA targets in Drosophila", Genome Biol, 5(R1) (2003).
[17] B. John et al., "MicroRNA Targets", PLos Biol, 2(11), e363 (2004).
[18] M. Kertesz et al., "The role of site accessibility in microRNA target recongnition", Nat. Genet, 39(10), 1278-1284 (2007).
[19] K. Nilubon et al., "Identification of Lung cancer associated protein by Molecular Complex Detection Analysis", IBBB 2015, Taiwan, Jan. 24-25 (2015).
[20] A. Bairoch et al., "The Universal Protein Resource (UniProt)", Nucl.Acids Res, 33, D154-D159 (2005).
[21] G. J. Sam et al., "miRBase: microRNA sequences, targets and gene nomenclature", Nucl.Acids Res, 34, D140-D144 (2005).
[22] DH. Sheng et al., "miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions", Nucl. Acids Res, 42(D1), D78-D85 (2014).