DP-miRNA: An Improved Prediction of precursor microRNA using Deep Learning Model

Jaya Thomas
[1]Department of Computer Science,
Stony Brook University,
Stony Brook, NY 11794, USA
[2]Department of Computer Science,
State University of New York Korea,
Incheon 406-840, Korea;
Email: jaya.thomas@sunykorea.ac.kr

Sonia Thomas
[2]Department of Computer Science,
State University of New York Korea,
Incheon 406-840, Korea;
Email: sonia.thomas@sunykorea.ac.kr

Lee Sael
[1]Department of Computer Science,
Stony Brook University,
Stony Brook, NY 11794, USA
[2]Department of Computer Science,
State University of New York Korea,
Incheon 406-840, Korea;
Email: sael@cs.stonybrook.edu

***Abstract*—MicroRNA (miRNA) are small non-coding RNAs regulating gene expression at the post-transcriptional level. Detecting miRNA in a genome is challenging experimentally and results vary depending on their cellular environment. These limitations inspire the development of knowledge-based prediction method. This paper proposes a deep learning based classification model for predicting precursor miRNA sequence that contains the miRNA sequence. The feature set consists of sequence features, folding measures, stem-loop features and statistical features. We evaluate the performance of the proposed method on human dataset. The deep neural network based classification outperformed support vector machine, neural network, naive Bayes classifiers, k-nearest neighbors, random forests as well as hybrid systems combining SVM and genetic algorithm.**

I. Introduction

MicroRNAs (miRNAs) are single-stranded small noncoding RNA typically 22 nucleotides long that regulate the translation of mRNAs. The miRNA regulates gene expression at the post transcription level by base pairing with the complementary sequence. This process hinders the translation of mRNA to proteins. The miRNA biogenesis involves number of steps. First, primary transcripts of miRNA (pri-miRNA) are transcribed often from introns of protein coding genes that are several kilobases long. The pri-miRNAs are then clopped by Rnase-III enzyme Drosha into $\sim$70 base pairs (bp) long hairpin-looped precursor miRNAs (pre-miRNAs). The exportin-5 protein transports pre-miRNAs hairpin into the cytoplasm through nuclear pore. In cytoplasm, pre-miRNAs are further cleaved by Rnase-III enzyme Dicer to produce a $\sim$20 bp double stranded intermediate called miRNA:miRNA*. A strand of the duplex with the low thermodynamic energy becomes a mature miRNA. Most mature miRNAs interact with the RNAi induced silencing complex (RISC) through base pairing of the target mRNAs regulate the expression of the genes. The miRNAs play key roles in development, cell proliferation and cell death. Thus, their deregulation has been connected with neurodegenerative disease, cancer and metabolic disorders [1].

Currently, miRBase [2] reports over 28645 miRNAs in more than 200 species, out of which over 2000 miRNA are reported for human. Informatics analysis predicts that 30% of human genes are regulated by miRNA [3]. miRNAs can be experimentally determined by directional cloning of endogenous small RNAs [4]. However, this is a time consuming process that require expensive laboratory reagents. These drawbacks motivate the application of computational approaches for predicting miRNAs.

Machine learning based methods can identify nonhomologous and species-specific miRNAs as compared to homologous search and comparative genomics approaches [2]. Distinguishing real pre-miRNAs and other pseudo hairpins is a problem that can be readily expressed as a binary classification problem. In this context, the human pre-miRNAs are labeled as +1 or positive samples whereas; pseudo hairpins are labeled as -1 or negative samples. The derived features are learned and mapped to the feature space for classification. Many approaches have been developed using naive Bayes classifier (NBC), artificial neural networks (ANN), support vector machines (SVM), and random forests (RF). SVM has been widely applied, including triplet-SVM[5], MiRFinder [6], miPred [7], microPred [8], yasMiR [9], YamiPred [10], MiRenSVM [11], MiRPara [12], etc. The other classifiers are neural network based MiRANN[13] classifier, random forest based [14] classifier.

Deep neural network (DNN) algorithm performs well in a setting where extracting features from raw data is not obvious by enabling raw data to be feed in directly. It is also performs well in a setting where number of features is very large. Whether the input is a raw data or a high dimensional feature set, DNN uses multi-layer architecture to learn multiple level of representation. This architecture automatically extracts high-level feature necessary for classification. The multiple layers in deep learning helps in processing of high data volume and exploit the complexities of data patterns. Hence, DNN have exhibited a good performance in different machine learning problems such as protein structure prediction [15], and predict splicing patterns [16].

In this work, we utilize heterogeneous features including sequence features, folding measures, stem-loop features and statistical features (z-score) to differentiate pre-miRNAs from

pseudo hairpins. The pseudo hairpins are RNA sequences, which have similar stem-loop features to pre-miRNAs but does not contain mature miRNAs. We use experimentally validated pre-miRNAs as positive examples and pseudo hairpins as negative examples to train and test the proposed method. The features of pre-miRNA and pseudo hairpins are used as input to DNN. We compared the performance of proposed DNN model against existing machine learning classifier and it achieves higher accuracy.

The main contribution of the paper are summarized as:

- Deep learning based prediction model is proposed for integrating large number of heterogeneous features for predictive analysis of pre-miRNAs from pseudo hairpins.
- Modified sampling technique is applied to address class imbalance problem.

II. Methods

A. Data

The human pre-miRNA sequence was retrieved from the miRBase 18.0 release. Similar to miPred [7] approach, the multiple loops were discarded to get 1600 pre-miRNA as positive dataset. The obtained sequence had an average length of 84 nt with minimum 43 nt and maximum 154 nt. The negative dataset consists of 8494 pseudo hairpins as the false samples. They were extracted from the human protein-coding regions as suggested by microPred [8]. The average length of the sequence is 85 nt with minimum as 63 nt and maximum as 120 nt. The different filtering criteria, including non-overlapping sliding window, no multiple loops, lowest base pair number set to 18, and minimum free energy less than 15kcal/mol were applied on these sequences to resemble the real pre-miRNA properties.

B. Feature set

The common characteristics of pre-miRNAs used for evaluation consists of sequences composition properties, secondary structures, folding measures and energy. This work adopts 58 characteristic features, which are shown useful in existing studies for predicting miRNA. The sequence characteristics include features related to the frequency of two and three adjacent nucleotide and aggregate dinucleotide frequency in the sequence. The secondary structure features from the perspectives of miRNA bio-genesis relating different thermodynamic stability profiles of pre-miRNAs. These structures have lower free energy and often contain stem and loop regions. They include diversity, frequency, entropy-related properties, enthalpy-related properties of the structure. The other features are hairpin length, loop length, consecutive base-pairs and ratio of loop length to hairpin length of pre-miRNA secondary structure. The energy characteristic associated to the energy of secondary structure includes the minimal free energy of the secondary structure, overall free energy NEFE, combined energy features and the energy required for dissolving the secondary structure.

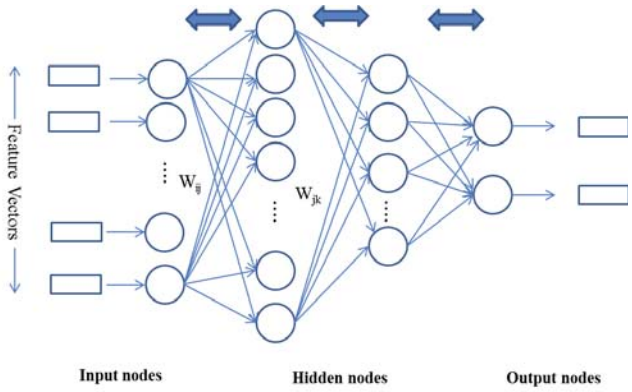

Fig. 1. A deep learning to predict miRNA with exracted features

C. Deep neural network

The proposed deep neural network (DNN) based miRNA prediction method, we call DP-miRNA, has three hidden layers, and the model is denoted as X-100-70-35-1, where X being the size of the input layer, 1 denotes the number of neuron in the output layer and the remaining values denotes the number of neurons in each hidden layer. Figure 1 illustrates the model architecture and layer-by-layer learning procedure. Different model architectures were trained using the same learning procedure but varying the number of hidden layer and nodes. Amongst the candidate network models, a better one was selected based on the classifier accuracy. The network model is pre-trained layer after the layer with the restricted Boltzmann machine (RBM). The initialization of the weight between every pair of adjacent layer is a step process that begins from the input or visible layer and completes at last hidden layer. At first, RBM learns the structure of the input data that constitutes to the activation of the first hidden layer, then the data is moved one layer down the network. Going in reverse, with each new hidden layer, the input from the previous layer is approximated by adjusting the network weights. This back and forth adjustment process is termed as Gibbs sampling, where the weights are updated by considering the difference in the correlation of the hidden activations and visible inputs. The process continues and now the first hidden layer will act as the input, which is multiplied by weights at the nodes of second hidden layer and the probability for activating the second hidden layers is calculated. This process results in sequential sets of activations by grouping features of features resulting in a feature hierarchy, by which networks learn more complex and abstract representations of data. This procedure of training a RBM can be repeated several times to create a multi-layer network. At the end, a standard feed-forward neural network is added after the last hidden layer, so the input being the activation probabilities which is used to predict the label. The resulting deep network was put together to adjust the weights using the standard back propagation algorithm to minimize the cross-entropy cost function error [17].

The deep learning network are trained with standard back propagation algorithm, with the weights adjusted using the

TABLE I
COMPARISON WITH EXISTING COMPUTATIONAL INTELLIGENCE TECHNIQUES

Classification Method	Accuracy	Sensitivity	Specificity	Geometric Mean
NBC	0.914 ±0.003	0.943 ±0.003	0.796 ±0.012	0.867± 0.006
KNN	0.908± 0.005	0.970± 0.122	0.657±0.023	0.798± 0.009
RF	0.937± 0.004	0.979±0.002	0.765±0.002	0.865±0.008
miRANN	0.917± 0.002	0.963±0.004	0.705±0.006	0.837±0.006
YamiPred	0.932±0.005	0.937±0.008	0.912±0.012	0.924±0.004
DP-miRNA	0.968± 0.002	0.973± 0.005	0.942± 0.006	0.971± 0.004

stochastic gradient descent as [18]]:

$$w_{ij}(t+1) = w_{ij}(t) + \eta \frac{\partial C}{\partial w_{ij}} \tag{1}$$

Where, $w_{ij}(t+1)$ is the weight computed at $t+1$, ∂ denotes the learning rate, and C is the cost function. For the given model, softmax is used as an activation function and the cost is computed using cross entropy. The softmax function is defined as

$$p_j = \frac{exp(x_j)}{\sum_k exp(x_k)} \tag{2}$$

Here, p_j stands for the output of the unit j, x_j and x_k denotes the total input to unit j and k respectively for the same level. The cross entropy is given by

$$C = -\sum_i d_j log(p_j) \tag{3}$$

Where d_j is the target probability for output unit j and p_j is the probability output after applying the activation function.

Another problem that we have addressed here is the class imbalance problem in miRNA predictions, as the number of negative class samples is more compared to the positives. We address this problem during the training phase by adopting a modified under sampling approach [19]. In the modified approach, we divided the majority class into subsets using k-means algorithm with k=5, and thus obtain clusters with slightly higher similarity amongst the group. These clusters are used to form different training sets by varying the ratio of majority class sample to minority class samples. Amongst the training dataset, one with higher accuracy was selected as an input to the classifier.

III. RESULTS

To evaluate the performance of the proposed classifier, we compare our method to existing state of the art miRNA classifiers. The evaluation is carried out by dividing the available data samples into training (60%), validation (20%) and testing (20%) set. The size of the input vector here is 58, i.e., the number of features used to build the model. The input data is normalized to standardizing the inputs in order to improve the training and to avoid getting stuck in local optima.

A. Performance Evaluation Metrics

The DP-miRNA model is a two class classifier, where true positive (TP) denotes the number of data samples classified as positive (real pre-miRNAs) and true negative (TN) represent correctly classified negative samples (pseudo pre-miRNAs). Similarly false positive(FP) and false negative (FN) represents the numbers of the misclassified positive and negative samples, respectively. The other measuring terms are sensitivity (SE) that measures the proportion of positives that are correctly identified accounting for the total positive samples, SE=TP/(TP+FN). Specificity (SP) measures the proportion of negatives that are correctly identified accounting for the total negative sample, SP =TN/(TN + FP). The classification accuracy (Accuracy) is proportion of correctly classified positive and negative class samples to total number of samples, Accuracy= (TP + TN)/(TP + TN + FP + FN). Another measure is geometric mean (Gm) to evaluate global classification performance, Gm=$\sqrt{SE \times SP}$.

B. Performance comparison of DP-miRNA

The DP-miRNA classifier learns more abstract features from the lower one to better summarize the pre-miRNAs and pseudo hairpins in the vector space. Table I shows a comparative result of the proposed DP-miRNA against the common machine learning approach for miRNA prediction. Considering the stochastic nature of the algorithm the output values are averaged for twenty executions. In comparison to the tested machine leaning techniques, DP-miRNA classifier shows a better performance. Another, point observed was that the modified sampling approach helped to overcome class imbalance problem as compared to random selection of data during training phase.

Further, we examined the performance of the DP-miRNA on selected twenty features that mostly represented sequence information and other thermodynamical characteristics. The feature set consist of dinucleotide frequencies AG, AU, CU, GA, UU, MFEI4, MFEI5, Positional Entropy, EAFE, Freq, dH/L, Tm, Tm/L, L, Avg_BP_stems, (G-U)/stems, (CE/L), (A-U)/stems and Statistical Z-scores zG, zQ and zSP. On the selected feature set we obtained a accuracy of 99.2%, with sensitivity and specificity high as 99.58% and 98.24% respectively. The result support the fact that features as entropy, enthalpy, minimum free energy and melting temperature are crucial for predicting miRNA [10].

IV. Conclusion

In this paper, we proposed deep learning classifier based pre-miRNA prediction method and showed performance improvement over existing methods. The proposed classifier was evaluated extensively on human dataset. The 58 features used as the input to deep learning framework included sequence conservation features, secondary structure features, and energy features of miRNA. For comparison, the dataset were generated with four biologically significant groupings and their combined set.

Acknowledgement

This research was supported by Basic Science Research Program through the RNF of Korea (NRF-2015R1C1A2A01055739), by the KEIT Korea under the "Global Advanced Technology Center" (10053204) and by the MSIP, Korea, under the "ICT Consilience Creative Program" (IITP-2015-R0346-15-1007) supervised by the IITP.

References

[1] T. M. Witkos, E. Koscianska, W. Krzyzosiak, Practical aspects of microrna target prediction, Curr Mol Med 11 (2) (2011) 99–109. doi:10.2174/156652411794859250.

[2] Y. Zhong, P. Xuan, K. Han, W. Zhang, J. Li, Improved pre-mirna classification by reducing the effect of class imbalance, BioMed Research International 2015 (2015) 1–12. doi:dx.doi.org/10.1155/2015/960108.

[3] J. S. Ross, J. A. Carlson, G. Brock, mirna: the new gene silencer, Am J Clin Pathol. 128 (5) (2007) 830–836. doi:http://dx.doi.org/10.1309/2JK279BU2G743MWJ.

[4] P. Y. Chen, H. Manninga, K. Slanchev, M. Chien, J. J. Russo, J. Ju, R. Sheridan, B. John, D. S. Marks, D. Gaidatzis, C. Sander, M. Zavolan, T. Tuschl, The developmental mirna profiles of zebrafish as determined by small rna cloning, Genes and Development 19 (11) (2005) 1288–1293. doi:10.1101/gad.1310605.

[5] C. Xue, F. Li, T. He, G. Liu, Y. Li, X. Zhang, Classification of real and pseudo microrna precursors using local structure sequence features and support vector machine, BMC Bioinformatics 6 (2005) 310. doi:10.1186/1471-2105-6-310.

[6] T. H. Huang, B. Fan, M. F. Rothschild, Z. L. Hu, K. Li, S. H. Zhao, Mirfinder: an improved approach and software implementation for genome-wide fast microrna precursor scans, BMC Bioinformatics 8 (2007) 341. doi:10.1186/1471-2105-8-341.

[7] K. L. S. Ng, S. K. Mishra, De novo svm classification of precursor micrornas from genomic pseudo hairpins using global and intrinsic folding measures, BMC Bioinformatics 23 (11) (2007) 1321–1330. doi:10.1186/1471-2105-8-341.

[8] R. Batuwita, V. Palade, micropred: effective classification of pre-mirnas for human mirna gene prediction, BMC Bioinformatics 25 (8) (2009) 989–995. doi:10.1093/bioinformatics/btp107.

[9] D. Pasaila, A. Sucial, I. Mohorianu, S. T. Pantiru, L. Ciortuz, Mirna recognition with the yasmir system: The quest for further improvements, Adv Exp Med Biol. 696 (2011) 17–25. doi:10.1007/978-1-4419-7046-6 2.

[10] D. Kleftogiannis, K. Theofilatos, S. Likothanassis, S. Mavroudi, Yamipred: A novel evolutionary method for predicting pre-mirnas and selecting relevant features, IEEE ACM Transactions on Computational Biology and Bioinformatics 12 (5) (2015) 1183–1192. doi:10.1109/TCBB.2014.2388227.

[11] J. Ding, S. Zhou, J. Guan, Mirensvm: towards better prediction of microrna precursors using an ensemble svm classifier with multi loop features, BMC Bioinformatics 14 (11) (2010) Suppl 11:S11. doi:10.1186/1471-2105-11-S11-S11.

[12] Y. Wu, B. Wei, H. Liu, T. Li, S. Rayner, Mirpara: a svm-based software tool for prediction of most probable microrna coding regions in genome scale sequences, BMC Bioinformatics 12 (2011) 107. doi:10.1186/1471-2105-12-107.

[13] M. E. Rahman, R. Islam, S. Islam, S. I. Mondal, M. R. Amin, Mirann: A reliable approach for improved classification of precursor microrna using artificial neural network model, Genomics 99 (2012) 189–194.

[14] J. Xiao, X. Tang, Y. Li, Z. Fang, D. Ma, Y. He, M. Li, Identification of microrna precursors based on random forest with network-level representation method of stem-loop structure, BMC Bioinformatics 12:165. doi:10.1186/1471-2105-12-165.

[15] M. Spencer, J. Eickholt, J. Cheng, A deep learning network approach to ab initio protein secondary structure prediction, IEEE/ACM Trans Comput Biol Bioinform. 12 (1) (2015) 103–112. doi:10.1109/TCBB.2014.2343960.

[16] M. K. K. Leung, H. Y. Xiong, L. J. Lee, B. J. Frey, Deep learning of the tissue-regulated splicing code, Bioinformatics 30 (12) (2014) i121–i129. doi:10.1109/TCBB.2014.2343960.

[17] G. E. Hinton, Training products of experts by minimizing contrastive divergence, Neural computation 14 (2002) 1771–1800. doi:10.1016/j.ygeno.2011.04.011.

[18] G. Hinton, L. Deng, D. Yu, G. Dahl, A. R. Mohamed, N. Jaitly, V. Vanhoucke, P. Nguyen, T. Sainath, B. Kingsbury, Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, IEEE Signal Processing Magazine (2012) 82–97doi:10.1.1.248.3619.

[19] S. J. Yen, Y. S. Lee, Cluster-based under-sampling approaches for imbalanced data distributions, Expert Systems with Applications 36 (3) (2009) 5718–5727. doi:10.1016/j.eswa.2008.06.108.