Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

A committee of NNA classifiers for the prediction of the binding between miRNAs and the target genes using a novel coding method

Zhisong He
Department of Bioinformatics, College of Life Sciences
Zhejiang University
HangZhou, China
jfsamery@gmail.com

Kaiyan Feng
Division of Imaging Science & Biomedical Engineering
The University of Manchester
Manchester, UK
kaiyan.feng@gmail.com

Yudong Cai*
Institute of System Biology
Shanghai University
Shanghai, China
cai_yud@yahoo.com.cn

* To whom correspondence should be addressed.

***Abstract*—We present a paper for the prediction of the bindings between microRNAs (miRNAs) and their target genes. A novel coding for the miRNAs, the binding sites (i.e. the target genes) and the flanking sequences of the binding sites is adopted to code the related information comprehensively. A feature selection method, Minimum Redundancy Maximum Relevance (mRMR), is used to filter out ineffective and redundant features. Because the data are severely imbalanced, a committee of NNA (Nearest Neighbor Algorithm) classifiers is applied to distribute the data more evenly between different classes. The final prediction results are gained through voting from the classifier committee. As a result, 83.33% positive samples are correctly identified with an overall correct prediction rate of 76.78%. The feature analysis, performed by mRMR feature selections using the classifier committee, shows that the seed region of miRNAs and the flanking sequences of the binding sites play a significant role in the regulation of miRNA binding.**

***Keywords*—miRNA target prediction, mRMR**

I. INTRODUCTION

MicroRNAs (miRNAs) are small non-coding RNAs with an approximate length of 22nt, serving as very important post-transcriptional regulators for gene expression. They bind to complementary regions of mRNAs of the genes they regulate, and with uncertain mechanisms, these bindings cause the cleavage of mRNAs or translational repressions [1, 2, 3]. Thus, miRNAs regulate a large number of functions in a cell and play an important role in cell division, development, and other biological processes [4, 5].

Over 8000 miRNAs have been discovered and their data are stored in the miRNA database – miRBase [6, 7, 8], with new miRNAs being found rapidly. However, it appears to be an obstacle to determine their functions experimentally, since these functions are deeper hidden and experimental determination is a lengthy process. Thus it would be highly beneficial if the target genes of a miRNA could be predicted accurately, which will not only help feed back to determine their target genes experimentally, but also aid to understand what affects the binding. Computational methods are well documented, e.g. the miRanda [9], TargetScan [10, 11], miTarget [12] and microTar [13], most of which first detect the potential binding sites (with a large degree of complementarity to the miRNA), and discard any sites that do not appear to be conserved across multiple species. These methods may encounter problems if the genomes of the multiple species are not clear. Furthermore, they mainly focus on the binding sites of miRNAs and ignore some important factors in the flanking contexts of the mRNAs.

In this paper, 67 positive samples with experimental validation were used, 61 of which are used for training and the remaining 6 ones are used for testing. 11990 negative samples produced by randomly coupling some miRNAs and 3'UTR regions were used. 10902 of them are used for training and the remaining 1088 samples are used for testing. With these data, we presented a novel method to predict miRNAs' binding sites. Firstly, a delicate coding system is adopted to code the related binding information as follows. The sequences of miRNAs, their binding sites and the 3' flanking sequences of these sites are coded by both 0/1 system and the frequency of each single nucleotide and dinucleotide. Several additional factors, including the complementary of miRNAs' seed regions, the whole miRNAs, and the distance between termination codon to the binding site are also used for coding. Secondly, because of the imbalance between the training set and the testing set, 11 individual classifiers based on the Nearest Neighbor Algorithms (NNAs) are used to solve the severe imbalanced data by distributing the negative samples evenly into the 11

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

classifiers. To improve the effectiveness of the coding system, a feature selection method, Incremental Feature Selection (IFS) based on Minimum Redundancy Maximum Relevance (mRMR) [14], is used to filter out the redundant and/or poor-performed features. As a result, 83.33% positive samples are correctly identified with an overall correct prediction rate of 76.78%.

II. Materials and Methods

A. Dataset

67 positive pairs of miRNA-Binding site are used in our study (see the supplemental materials 1), which are drawn from Tarbase [15], containing information of the exact binding sites of miRNA in 3' UTR of mRNAs. 11990 non-binding pairs are selected by randomly coupling some miRNAs and 3'UTR regions. The training dataset contains 61 positive pairs and 10902 negative pairs while the testing dataset contains 6 positive pairs and 1088 negative pairs.

B. Coding of miRNAs and their binding sites

Because the length of miRNAs and target sites are variable, all the sequences with length less than 25nt are extended to exactly 25nt by adding some "N" suffixes. For miRNAs, these Ns are added to the 3' terminal; while for the binding site, they are added to the 5' terminal. Then these 25nt sequences can be coded by a 100 feature vector by encoding each nucleotide and "N" using 4 binary digits as:

$$\begin{cases} "A" := 0001 \\ "U" := 1000 \\ "C" := 0010 \\ "G" := 0100 \\ "N" := 0000 \end{cases}$$

For example, sequence "AUGUUGCCAUCGUUAAGCACAGUNN" is coded by "0001 1000 0010 1000 1000 0100 0010 0010 0001 1000 0010 0100 1000 1000 0001 0001 0100 0010 0001 0010 0001 0100 1000 0000 0000", resulting in 100 features. The frequencies of each single nucleotide ("A", "C", "G", "U", "N") and di-nucleotide ("AA", "AC", "AG", "AU", "CA", "CC", "CG", "CU", "GA", "GC", "GG", "GU", "UA", "UC", "UG", "UU") are also added as features for the prediction. Thus, a 121-feature vector (121=100+5+16) is used to code the sequence of a miRNA or a binding site.

C. Coding of 3' flanking sequences of miRNA binding sites

Besides the binding site, the flanking sequences are also proven to be influential in the regulation [16]. Thus, we extract the 3' flanking sequences, with length of 25nt, of the binding site. A 3' flanking sequence is also coded into 121-feature vector as is described above.

D. Other factors affecting the miRNA regulation

The complement of the whole miRNA, especially the complement of position 2-8 (seed region) in a miRNA, is important in the binding process of miRNA to 3' UTR of mRNA. A recent research [17] shows that the position of binding site in the 3'UTR is also important for the efficiency of miRNA regulation. Thus three more features are obtained for each pair of miRNA and binding site. They are the number of base pairs in a seed region, the proportion of matched base pairs in the whole miRNA, and the position of binding site started from the 3' terminal codon of mRNA. The first two features require the situation of hybridization between miRNA and binding site. Because it is difficult to get the actual situation, RNAduplex in ViennaRNA Package [18] is used for the prediction in our study to get the predictive one.

E. Normalization

After the vectors are constructed, normalizations are made among each feature of vectors, to put all the vectors in the same scale. The normalization function is:

$$x_{ni} = \frac{x_i - \bar{x}}{s_x} \tag{1}$$

where x_i is the feature before normalization, x_{ni} is the feature after normalization. Suppose the number of samples is n,

$$s_x = \sqrt{\sum_{i=1}^{n} (x_i - x)^2 / (n-1)} \tag{2}$$

which is the standard deviation of x in all samples, and

$$\bar{x} = \sum_{i=1}^{n} x_i / n \tag{3}$$

which is the average of x in all samples.

F. Combining Vectors

Suppose R, B, F, O are the coding for the sequence of miRNA, sequence of the binding site, sequence of the 3' flanking sequence and the other factors, the combined vector can be represented as:

$$F_{RBFO} = R \oplus (k_1 \cdot B) \oplus (k_2 \cdot F) \oplus (k_3 \cdot O) \tag{4}$$

where k_1, k_2, k_3 are the weights for the each coding, removing the bias caused by the contribution difference between encoding systems.

To simplify the computation, k_1, k_2, k_3 are all fixed to be 1 in this study. After the combination, the final feature vector with totally 366 features is obtained as 121 (miRNA) + 121 (binding site) + 121 (flanking sequence) + 3 (other factors) = 366.

G. Nearest Neighbor Algorithm (NNA)

Suppose RT_1 and RT_2 are two vectors, their similarity is defined as [19]:

$$\Lambda(RT_1, RT_2) = \frac{RT_1 \cdot RT_2}{\| RT_1 \| \cdot \| RT_2 \|} \tag{5}$$

where $RT_1 \bullet RT_2$ is the dot product of RT_1 and RT_2, and $\|RT_1\|$ and $\|RT_2\|$ are the moduli of RT_1 and RT_2 respectively. The

larger $\Lambda(RT_1,RT_2)$ is, the more similar RT_1 and RT_2 are to each other. For a given vector RT, the most similar vector RTn in data set $\{RT_1,RT_2,\ldots,RT_N\}$ is found, i.e. $\Lambda(RT,RT_n)$ is the greatest, and the class of RT is assigned to have the same class as RT_n.

H. Combining multiple NNA classifiers

The size of negative samples (11990) is much larger than the positive ones (67). To overcome this imbalanced data problem, we adopt a committee of classifiers (11 NNA classifiers) to balance the *data*, as is proposed by [20]. The negative data of the training set are split into 11 parts, each of ten of which contains 993 samples, while the remaining one has 972 samples. 61 positive samples are input into every of the 11 classifiers. Thus each NNA classifier is trained by the 61 positive samples plus one part of the negative samples in the training set, and the remaining 6 positive samples and 1088 negative samples are used for testing. The final prediction results for the testing data are obtained through voting. Only when more than half (>5) classifiers predict a sample to be positive, this sample will be assigned to be positive, otherwise, it will be assigned to be negative.

I. Feature Selection

In order to improve the effectiveness of the encoding system, for each classifier, we use IFS based on mRMR to select an optimized feature set for the prediction. The optimized feature set is chosen according to its performance in Jackknife cross-validation test.

1) mRMR (Maximum Relevance, Minimum Redundancy method): Maximum Relevance, Minimum Redundancy method (mRMR) is originally developed by Peng [14] for microarray data processing. mRMR method requires the input data to be numeric vectors - each vector is taken as a mRMR feature. mRMR ranks each feature according to both its relevance to the target (highly related to the prediction accuracy) and the redundancy between the features. A "good" feature is characterized by maximum relevance with the target variable and minimum redundancy within the features. Both relevance and redundancy are defined by mutual information (MI), which estimates how much one vector is related to another. MI is defined as follows:

$$I(x,y)=\iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy \tag{6}$$

x and y are two vectors; $p(x,y)$ is the joint probabilistic density; and are the marginal probabilistic densities.

Let Ω denote the whole vector set. The already-selected vector set with m vectors is denoted by Ω_s, and the to-be-selected vector set with n vectors is denoted by Ω_t.

Relevance D of a feature f in Ω_t with a target variable c can be computed by Eq (7):

$$D=I(f,c) \tag{7}$$

Redundancy R of a feature f in Ω_t with all the features in Ω_s can be computed by Eq (8):

$$R=\frac{1}{m}\sum_{f_i\in\Omega_s}I(f,f_i) \tag{8}$$

To maximize relevance and minimize redundancy, mRMR function is obtained by integrating Eq (7) and Eq (8):

$$\max_{f_j\in\Omega_t}\left[I(f_j,c)-\frac{1}{m}\sum_{f_i\in\Omega_s}I(f_j,f_i)\right](j=1,2,\ldots,n) \tag{9}$$

Let the initial where is best performed feature, and the initial by excluding only. Eq (9) is used to obtain one vector by another in totally rounds, resulting a vector list $S=\left[f_0^{'},f_1^{'},\ldots,f_h^{'},\ldots,f_{N-1}^{'}\right]$ with the selection order where denotes at which round the feature is selected.

2) Incremental Feature Selection (IFS): The mRMR feature selection produces a list of candidate features, as is described above. However, we don't know how many features in the list should be chosen. In our study, Incremental Feature Selection (IFS) was used to determine the optimal number of features, and the optimal features.

In mRMR step, we can construct the N feature sets from ordered feature set S, where the i-th feature set is:

$$S_i=\{f_0,f_1,\ldots,f_i\}\,(0\le i\le N-1)$$

For every i between 0 and $N-1$, each NNA in the committee is constructed to make the prediction with the feature set S_i. Jackknife Cross-Validation [19] is used to obtain the prediction accuracy. As a result, we can get an IFS overall correct prediction curve with index i as its x-axis and the overall prediction accuracy as its y-axis, and an IFS positive sample correct prediction curve with index i as its x-axis and the positive sample prediction accuracy as its y-axis. The optimal feature set $S_t=\{f_0,f_1,\ldots,f_t\}$, which produces a positive sample prediction accuracy higher than 0.7, and at the same time has the highest overall prediction accuracy, is chosen as the optimized feature set with totally t featuers in it.

III. Results and Discussion

A. mRMR and prediction results

The mRMR program used in our study was downloaded from http://research.janelia.org/peng/proj/mRMR/. The parameter in mRMR was chosen to be t = 1 to discretize our data to three categorical states according to the values of: mean ± (t • std), where mean is the mean value and std is the standard deviation. The result of mRMR is a table, with the mRMR feature list defined by Eq (9) in it. Please refer to supplemental materials 2 for the feature list. It also outputs another table with MaxRel features defined by Eq (9), containing the relevance between features and the class variable. In this study, only mRMR feature list was used for the research.

IFS was used in the feature selection for each of the 11 individual NNA classifiers. The IFS curves, for both the overall

prediction accuracy and the positive sample accuracy, produced by the 11 classifiers, are shown in Figure 1 at the end of the paper. The number of features, selected for each classifier, is shown in Table 1 below.

TABLE I. THE NUMBER OF FEATURES SELECTED BY IFS FROM THE 11 CLASSIFIERS

Classifier ID	The number of features
1	8
2	42
3	9
4	7
5	77
6	8
7	41
8	41
9	243
10	76
11	7

After the training of the 11 classifiers, the pairs in test set are predicted by the classifier committee through voting. The accuracy of the prediction is shown in Table 2.

TABLE II. RESULT OF THE TEST SET

	Amount	Correctly predicted	Accuracy
Positive pair	6	5	0.833
Negative pair	1088	835	0.767
Overall	1094	840	0.768

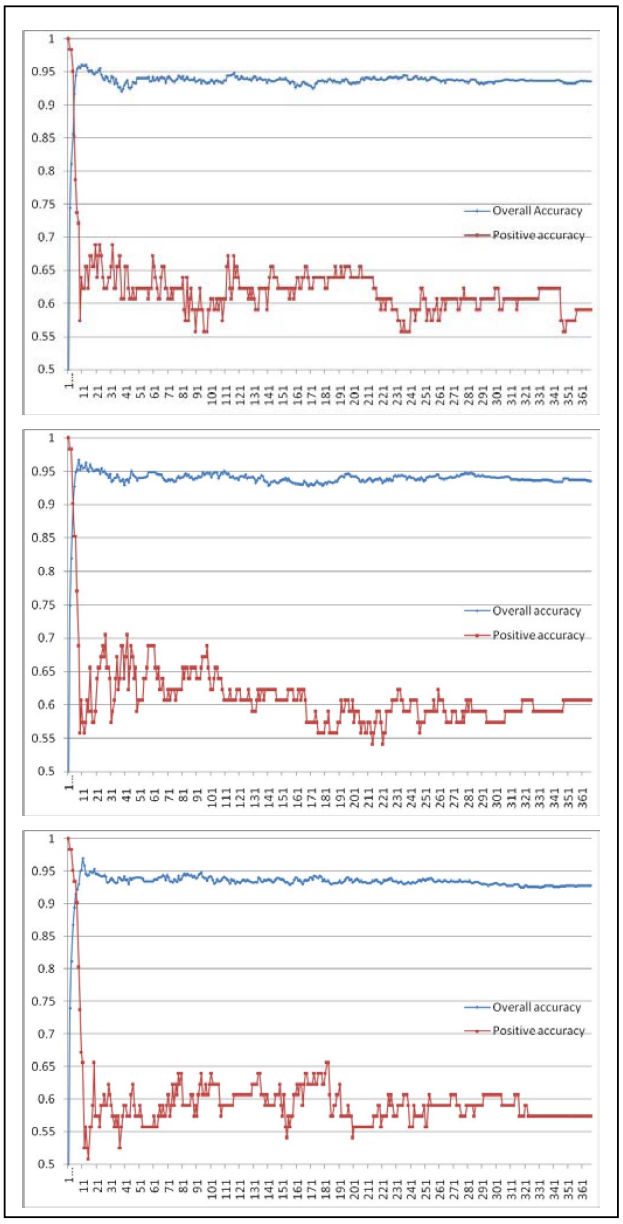
B. Discussion

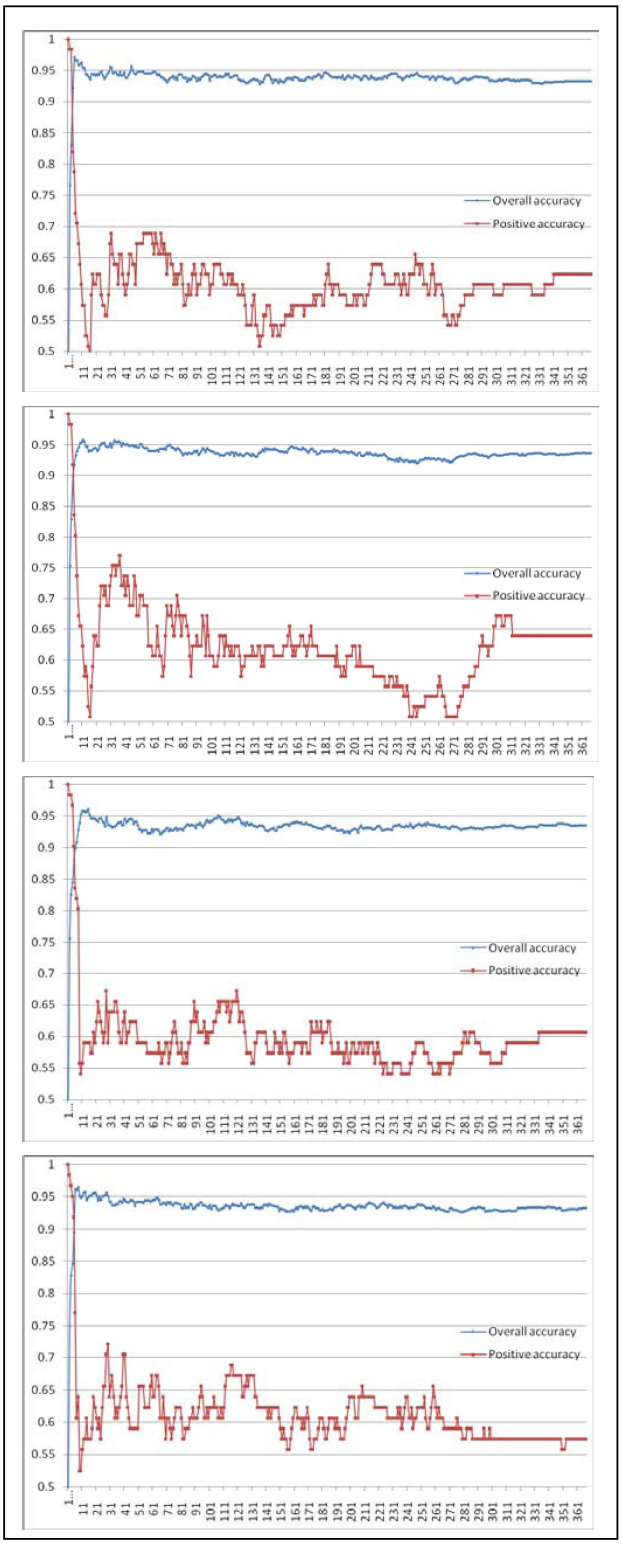
These 11 classifiers in the committee are independent from each other. However, four common features are present in all these classifiers indicating they play an important role in the process of miRNA binding. These features include the percentage of matched pairs of nucleotide in the seed region of miRNA, the first digit of the 6th nucleotide of the 3' flanking sequences, the fourth digit of the 20th nucleotide of miRNA binding site, and the first digit of the 22nd nucleotide of miRNA binding site. Among the 67 positive pairs in our study, 49 pairs have the 20th nucleotide of the binding site bound to the seed region, and 51 pairs have the 22nd nucleotide of the binding site bound to the seed region, where 20th and 22nd nucleotide of the binding sites are proven to be important in the prediction, confirming that the seed region of miRNA is influential to the miRNA binding. 27 common features, 7 of which come from the 3' flanking sequences, appear in more than 5 classifiers, plus the 6th nucleotide of the 3' flanking sequences that is shared by all classifiers, indicating that the flanking sequences of miRNA binding site may play a surprisingly important role in the regulation of miRNA binding.

IV. CONCLUSION

We present a novel method to predict the binding of miRNAs and the target genes by adopting a delicate coding system for the miRNAs, the binding sites and the flanking sequences of the binding sites. The problem of the severely imbalanced data is solved by applying a committee of NNA classifiers. As a result, we achieve a positive sample prediction accuracy of 83.3% and an overall prediction accuracy of 76.8%. The feature analysis shows that the seed region of the miRNA sequences and the flanking sequences of the binding sites play an important role in the binding of miRNAs and the target genes, demonstrating the importance of including the flanking sequences of the binding sites in the coding.

Supplemental material 1 and 2 in this paper is available upon request.


Overall Accuracy
Positive accuracy
Overall accuracy
Positive accuracy
Overall accuracy
Positive accuracy


Overall accuracy
Positive accuracy
Overall accuracy
Positive accuracy
Overall accuracy
Positive accuracy
Overall accuracy
Positive accuracy

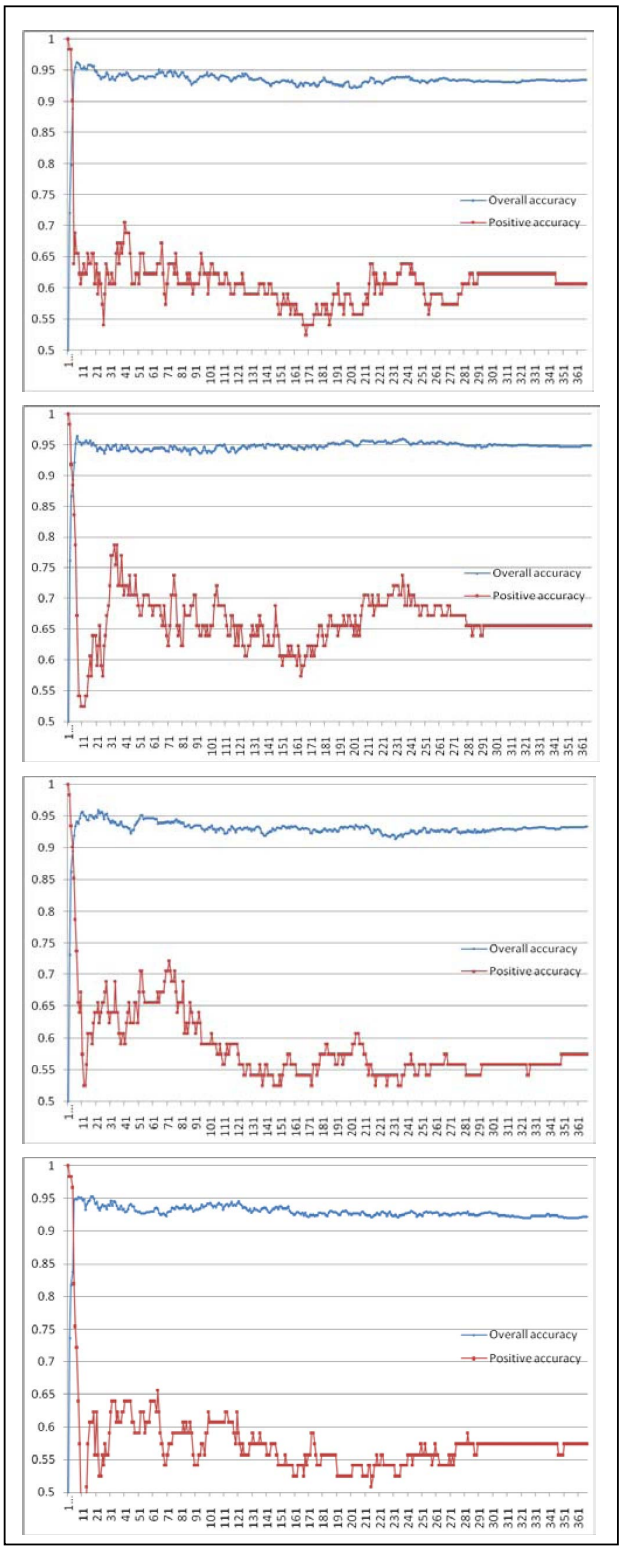


Figure 1 Accuracy of Jackknife cross-validation test of the 11 classifiers using increasing number of features. X axis denotes the number of the features, and y axis denotes the prediction accuracy.

REFERENCES

[1] D.P. Bartel, MicroRNAs: genomics, biogenesis, mechanism, and function, Cell, vol. 116, pp. 281-297, 2004.

[2] W. Filipowicz, L. Jaskiewicz, F.A. Kolb, and R.S. Pillai, Post-transcriptional gene silencing by siRNAs and miRNAs, Curr. Opin. Struct. Biol., vol. 15, pp. 331-341, 2005.

[3] E.J. Sontheimer, and R.W. Carthew, Silence from within: endogenous siRNAs and miRNAs, Cell, vol. 122, pp. 9-12, 2005.

[4] V. Ambros, The functions of animal microRNAs, Nature, vol. 431, pp. 350-355, 2004.

[5] Y. Kong, J.H. Han, MicroRNA: biological and computational perspective, Genomics Proteomics Bioinformatics, vol. 3, pp. 62-72, 2005.

[6] S. Griffiths-Jones, H.K. Saini, S. van Dongen, and A.J. Enright, miRBase: tools for microRNA genomics, Nucleic Acids Res., vol. 36, pp. D154-158, 2008.

[7] S. Griffiths-Jones, R.J. Grocock, S. van Dongen, A. Bateman, and A.J. Enright, miRBase: microRNA sequences, targets and gene nomenclature, Nucleic Acids Res., vol. 34, pp. D140-144, 2006.

[8] S. Griffiths-Jones, The microRNA Registry, Nucleic Acids Res., vol. 32, pp. D109-111, 2004.

[9] A.J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D.S. Marks, MicroRNA targets in Drosophila, Genome Biol., vol. 5, pp. R1, 2003.

[10] B.P. Lewis, I.H. Shih, M.W. Jones-Rhoades, D.P. Bartel, and C.B. Burge, Prediction of mammalian microRNA targets, Cell, vol. 115, pp. 787-798, 2003.

[11] B.P. Lewis, C.B. Burge, and D.P. Bartel, Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets, Cell, vol. 120, pp. 15-20, 2005.

[12] S.K. Kim, J.W. Nam, J.K. Rhee, W.J. Lee, and B.T. Zhang, miTarget: microRNA target gene prediction using a support vector machine, BMC Bioinformatics, vol. 7, pp. 411, 2006.

[13] R. Thadani, and M.T. Tammi, MicroTar: predicting microRNA targets from RNA duplexes, BMC Bioinformatics, vol. 7 Suppl 5, pp. S20, 2006.

[14] H. Peng, F. Long, and C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, pp. 1226-1238, 2005.

[15] P. Sethupathy, B. Corda, and A.G. Hatzigeorgiou, TarBase: A comprehensive database of experimentally supported animal microRNA targets, RNA, vol. 12, pp. 192-197, 2006.

[16] D. Didiano, and O. Hobert, Molecular architecture of a miRNA-regulated 3' UTR, RNA, vol. 14, pp. 1297-317, 2008.

[17] A. Grimson, K.K. Farh, W.K. Johnston, P. Garrett-Engele, L.P. Lim, and D.P. Bartel, MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol. Cell, vol. 27, pp. 91-105, 2007.

[18] S.R. Eddy, How do RNA folding algorithms work?, Nat. Biotechnol.,vol. 22, pp. 1457-1458, 2004.

[19] Z. Qian, Y.D. Cai, and Y. Li, A novel computational method to predict transcription factor DNA binding preference, Biochem. Biophys. Res. Commun., vol. 348, pp. 1034-1037, 2006.

[20] X.M. Zhao, X. Li, L. Chen, K. Aihara, Protein classification with imbalanced data, Proteins, vol. 70, pp. 1125-1132, 2008.