Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2011, Article ID 506205, 8 pages
doi:10.1155/2011/506205

Research Article

Prediction of RNA-Binding Proteins by Voting Systems

C. R. Peng,[1,2] L. Liu,[1] B. Niu,[3] Y. L. Lv,[4] M. J. Li,[2] Y. L. Yuan,[5] Y. B. Zhu,[2] W. C. Lu,[2] and Y. D. Cai[6]

[1] *School of Materials Science and Engineering, Shanghai University, 149 Yan-Chang Road, Shanghai 2000721, China*
[2] *Department of Chemistry, College of Sciences, Shanghai University, 99 Shang-Da Road, Shanghai 200444, China*
[3] *College of Life Sciences, Shanghai University, 99 Shang-Da Road, Shanghai 200444, China*
[4] *University of Shanghai for Science and Technology Library, 516 Jun-Gong Road, Shanghai 200093, China*
[5] *Department of Synthesis, WuXi AppTec Co., Ltd., Shanghai 200131, China*
[6] *Institute of Systems Biology, Shanghai University, 99 Shang-Da Road, Shanghai 200444, China*

Correspondence should be addressed to W. C. Lu, wclu@shu.edu.cn and Y. D. Cai, cai_yud@yahoo.com.cn

Received 9 March 2011; Revised 12 May 2011; Accepted 26 May 2011

Academic Editor: Zoran Obradovic

It is important to identify which proteins can interact with RNA for the purpose of protein annotation, since interactions between RNA and proteins influence the structure of the ribosome and play important roles in gene expression. This paper tries to identify proteins that can interact with RNA using voting systems. Firstly through Weka, 34 learning algorithms are chosen for investigation. Then simple majority voting system (SMVS) is used for the prediction of RNA-binding proteins, achieving average ACC (overall prediction accuracy) value of 79.72% and MCC (Matthew's correlation coefficient) value of 59.77% for the independent testing dataset. Then mRMR (minimum redundancy maximum relevance) strategy is used, which is transferred into algorithm selection. In addition, the MCC value of each classifier is assigned to be the weight of the classifier's vote. As a result, best average MCC values are attained when 22 algorithms are selected and integrated through weighted votes, which are 64.70% for the independent testing dataset, and ACC value is 82.04% at this moment.

1. Introduction

Protein-RNA interactions play significant roles in a wide range of biological processes, including regulation of gene expression, protein synthesis and replication, and the assembly of many viruses [1–4]. A good knowledge of protein-RNA interactions is fundamentally important for the understanding of how proteins regulate gene expression. Machine learning and data mining methods have been widely applied in the fields of computational biology and bioinformatics [5–9], and the same principles are also applied to determine whether a protein participates in RNAbinding [10–16]. Some investigations code a protein using primary amino acid compositions [10, 11, 13, 14], and some code with protein chemical or physical properties and structural information [10–12, 14–16]. In terms of machine learning methods, support vector machine (SVM) [10, 14], artificial neural networks [17], Naive Bayes [18], and so forth, were all found in the literature to uncover the interaction between proteins and RNA. A specific study [19] was carried out to determine the interaction sites between RNA and Rev proteins of HIV-1 and EIAV, in which both protein-protein interface residues and protein-RNA interface residues were predicted, by first training the predictors using known protein-protein and protein-RNA complexes and then using the trained predictors to predict the binding sites of HIV-1 and EIAV Rev proteins.

The above reviewed papers applied a single classifier to determine the interactions between RNA and proteins. However, for a specific biological dataset, an individual classifier has its own strengths and weaknesses. Underfit or overfit of a single classifier will affect the accuracy or the generalization of the prediction performance. Thus, people are inspired to integrate multiple classifiers [20, 21], in attempts to improve the prediction/classification performance. Recently, Chen et al. [21] proposed a few voting systems for the classification (prediction) of protein structural classes. Chen et al. [21] used an unprecedented number of machine learning algorithms from Weka (http://www.cs.waikato.ac.nz/~ml/weka/) for the voting systems and realized that some of the classifiers

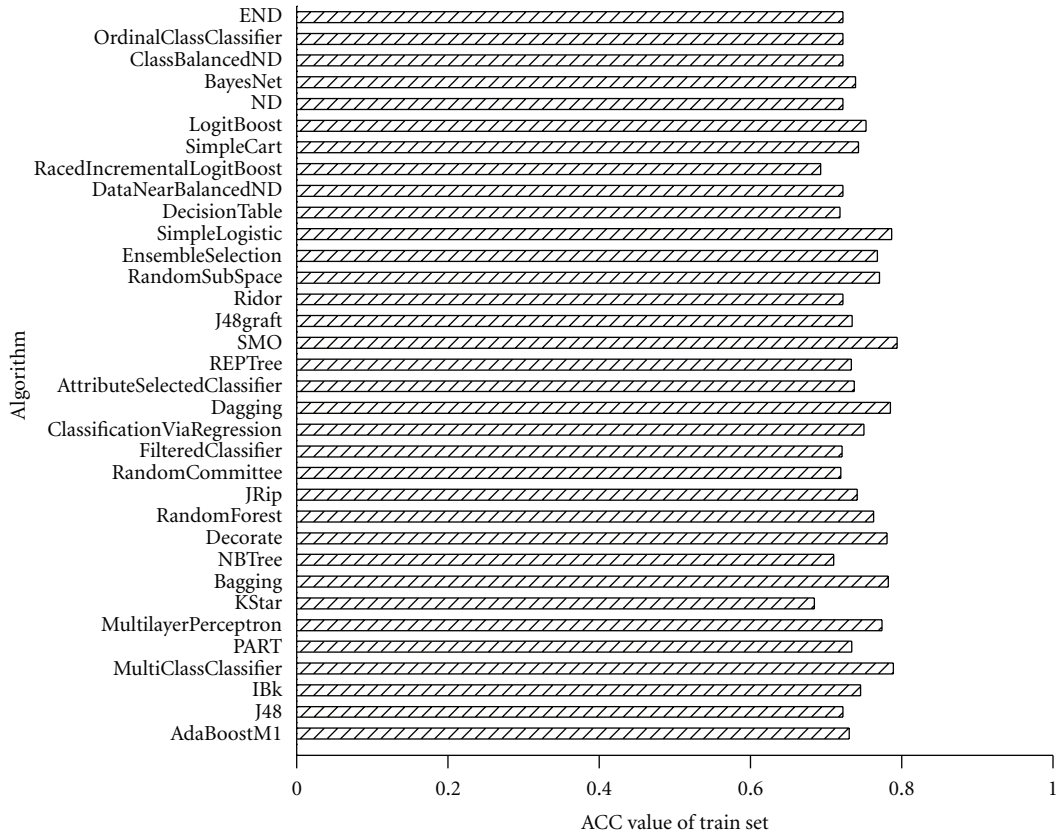

Figure 1: The average ACC values of 34 algorithms in basic training dataset.

may be redundant since they could worsen the overall classification performance if included. Therefore, mRMR (minimum redundancy maximum relevance) [22] strategy, which is originally developed for feature selection [23, 24], was transferred into classifier selection. As a result, four voting systems were developed [21]. They are simple majority voting system (SMVS), weighted majority voting system (WMVS), SMVS with algorithm selection (SMVS_AS); and WMVS with algorithm selection (WMVS_AS). In this paper, these voting systems are adopted and applied to predict the interaction between proteins and RNA.

2. Materials and Methods

2.1. Data Preparation

(i) The Rough "Positive" Dataset: Using "RNA binding" as keywords to search the SWISS-PROT database (version 54.2), 20132 proteins were retrieved. This collection was designated as "positive" dataset.

(ii) The "Contrast" Dataset: A "contrast" set of 72331 proteins was retrieved from SWISS-PROT by searching with a list of keywords which possibly imply RNA/DNA-binding functionality, using the "or" logic, which was proposed by Cai and Lin [10].

(iii) The Rough "Negative" Dataset: the proteins in the "contrast" dataset were removed from the SWISS-PROT database (it has 232345 sequence entries) and 160014 proteins were obtained to form the "negative" dataset.

(iv) The RNA-Binding Protein Dataset: protein sequences with length >6000 aa or <50 aa were removed since they might be protein complexes or protein fragments. Proteins including irregular amino acid characters such as "x" and "z" were also removed. Moreover, the redundancy among the sequences in "positive" and "negative" datasets was removed by using CD-HIT [25] and PISCES [26] program, with a threshold of 40%. As a result, 2063 and 21562 proteins were produced in nonredundant RNA-binding and "negative" datasets, respectively. To achieve data balance, datasets were built in the following manner: first all the proteins in the "positive" subset were selected as the first part. Then the proteins in the "negative" subset were randomly selected as the second part. The number of proteins selected in the "negative" subset equals that of the first part. Thirdly we combined the first part and the second part together to be total dataset; finally we randomly drew out third of that total dataset to be test dataset, the rest to be train dataset and Consequently, the RNA-binding protein training dataset of 2752 proteins and the RNA-binding protein testing dataset of 1374 proteins (see Table 1, "A" means RNA-binding protein and "B" means RNA-nonbinding protein)

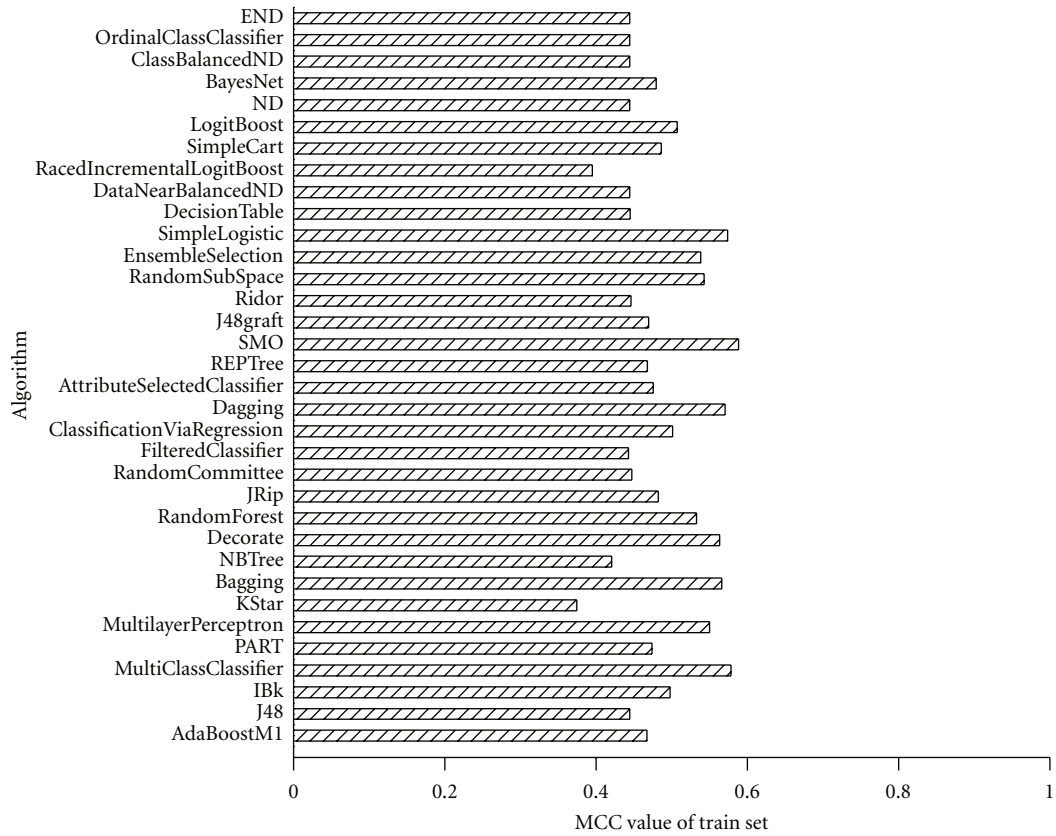

Figure 2: The average MCC values of 34 algorithms in basic training dataset.

Table 1: The distribution of proteins in training dataset and test dataset.

Dataset	A	B
Basic training dataset	1376	1376
Independent test dataset	687	687

are available in Supplementary Material (see Supplementary Material available online at doi:10.1155/2011/506205). In order to ensure the stability of the built model, we repeat these steps ten times. That is to say, we build ten train datasets and ten test datasets randomly, and all of ACC (overall prediction accuracy) value and MCC (Matthew's correlation coefficient) value in our paper are the average value.

2.2. Feature Vector. A successful classification requires an effective way to represent a protein. Under current techniques, it is not possible to know every aspect of a protein from its sequential information. However, the biological properties of the amino acids that compose a protein are known, and they may reveal some properties of a whole protein sequence. Thus, in this paper a protein is represented by amino acid compositions and the biological properties of each amino acid [14] which is one of the popular representation methods in the literature. The biological properties include hydrophobicity, predicted secondary structure, predicted solvent accessibility, normalized Van Der Waals volume, polarity, and polarizability. As a result, totally 132 features are derived, among which 112 features come from biological properties and 20 from the amino acid compositions. Detailed information of these features can be found in [14].

2.3. Machine Learning Algorithms. 34 machine learning algorithms in Weka [27] were selected and integrated using various voting systems. These algorithms are listed below.

BayesNet, DecisionTable, JRip, PART, Ridor, AttributeSelectedClassifier, Bagging, ClassificationViaRegression, Dagging, Decorate, END, EnsembleSelection, FilteredClassifier, LogitBoost, MultiClassClassifier, OrdinalClassClassifier, RacedIncrementalLogitBoost, RandomSubSpace, ClassBalancedND, ND, DataNearBalancedND, RandomCommittee, IB1, AdaboostM1, Kstar, MultilayerPerceptron, SimpleLogistic, SMO, J48, J48graft, NBTree, RandomForest, REPTree, SimpleCart.

Readers may refer to [27] for detailed introduction about these algorithms.

2.4. Ensemble Approach. Four ensemble approaches, Simple majority voting system (SMVS), weighted majority voting system (WMVS), SMVS with algorithm selection (SMVS_AS), and WMVS with algorithm Selection (WMVS_AS), are introduced briefly here. Readers may refer to [21] for the detailed information about these voting

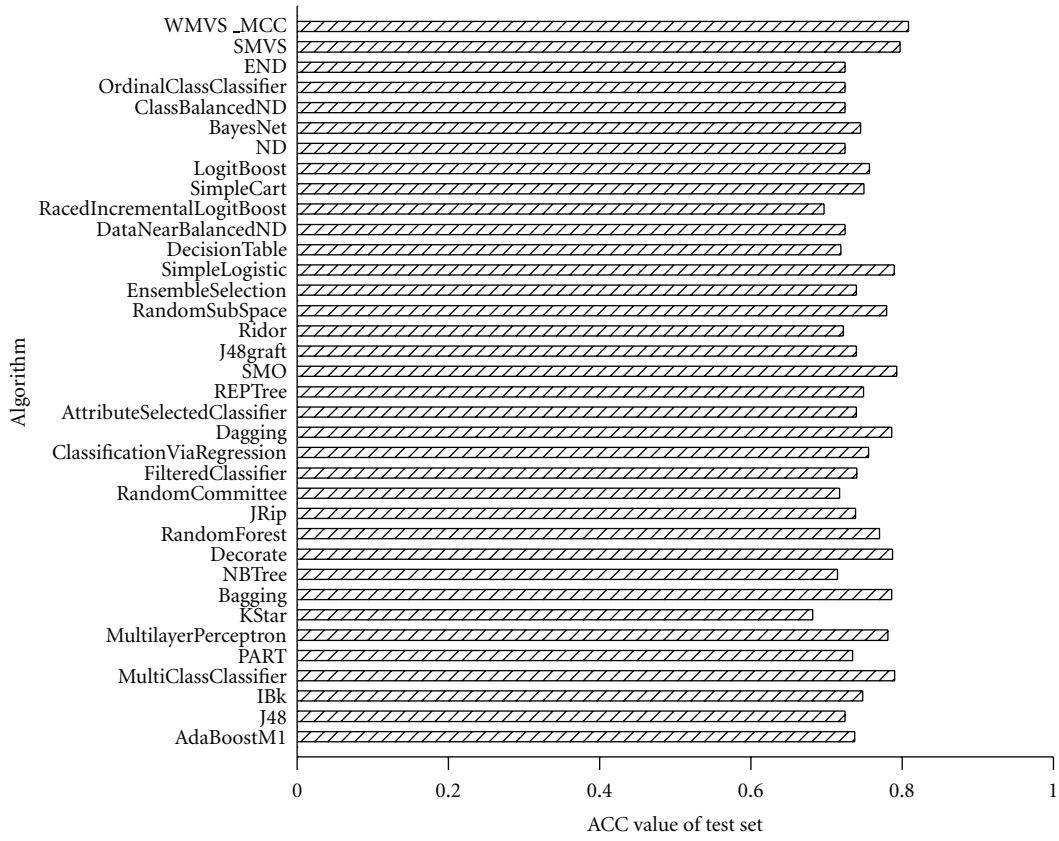

Figure 3: The average ACC values of 34 algorithms in independent test dataset (including the results of SMVS and WMVS_MCC).

systems. SMVS takes the class label that gains the majority votes as the class of a processed data. WMVS weighs each vote with the overall prediction accuracy of the corresponding classifier on a training dataset. SMVS_AS first selects some classifiers using mRMR method, and then the selected algorithms are integrated through SMVS. WMVS_AS is like the SMVS_AS to first select some classifiers using mRMR method, but then WMVS is used instead of SMVS in the integration.

3. Results and Discussion

3.1. Prediction Results of the 34 Algorithms. 34 algorithms were tested by tenfold cross-validation (10-CV) on both the basic training dataset and the independent testing dataset. The detailed outputs of 10-CV on the basic training dataset and independent testing dataset are listed in Supplementary Material.

Figures 1, 2, 3, and 4 depicted both the average values of ACC and MCC of each algorithm in basic training dataset and independent test dataset, respectively. Figures 3 and 4 also included the average values of ACC and MCC in SMVS and WMVS_MCC (WMVS based on MCC value, all of WMVS values are based on MCC value in our paper). SMO performs best on the training dataset, with 79.40% of ACC value and 58.81% of MCC value, and also SMO performs best on the testing dataset, with 79.29% of ACC value and 58.58% of MCC value. The standard deviation of ten datasets of the 34 algorithms is listed in Table 2; it seems that the results are stable.

The Matthew's correlation coefficient (MCC) is used in machine learning as a measure of the quality of binary (two-class) classifications. It takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes. The MCC can be calculated directly from the confusion matrix using the following formula:

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TN} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TP} + \mathrm{FP})}}. \quad (1)$$

In this equation, TP is the number of true positives, TN the number of true negatives, FP the number of false positives, and FN the number of false negatives.

3.2. Results of SMVS and WMVS. Average predicted results and standard deviation of SMVS and WMVS are shown in Table 3. SMVS and WMVS perform better than any individual algorithm selected in Weka, and WMVS performs a little better than SMVS. It implies that as a whole the 34 algorithms collaborate to improve the prediction accuracy through voting. The values of standard deviation also decrease significantly through voting. It implies that voting system increases the stability of prediction model.

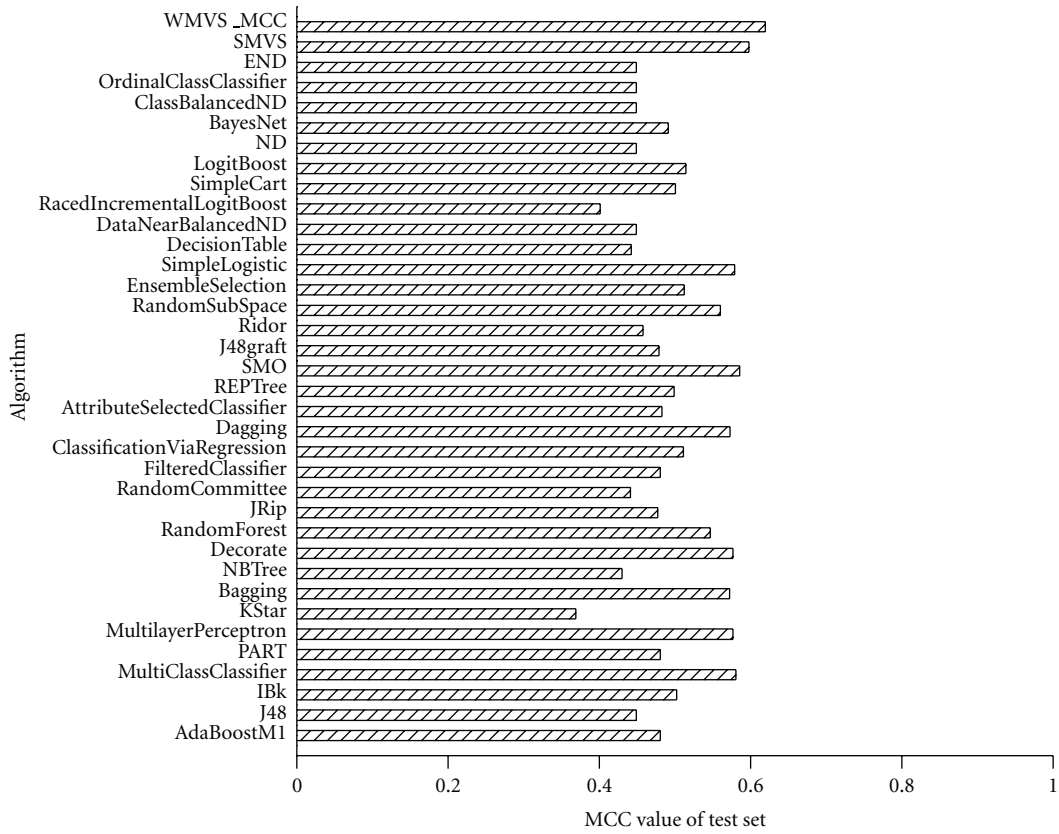

Figure 4: The average MCC values of 34 algorithms in independent test dataset (including the results of SMVS and WMVS_MCC).

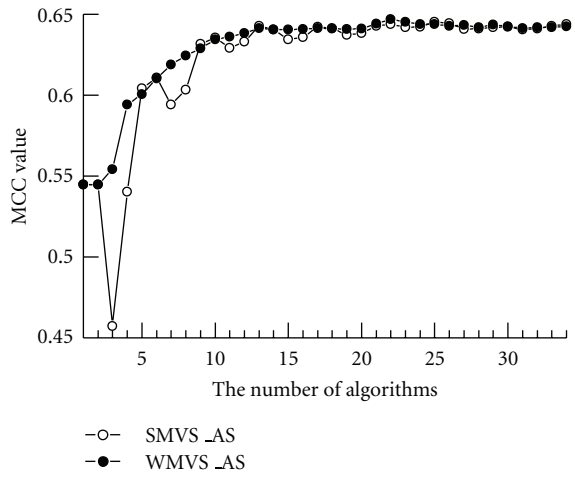

Figure 5: The average MCC value of SMVS_AS and WMVS_AS.

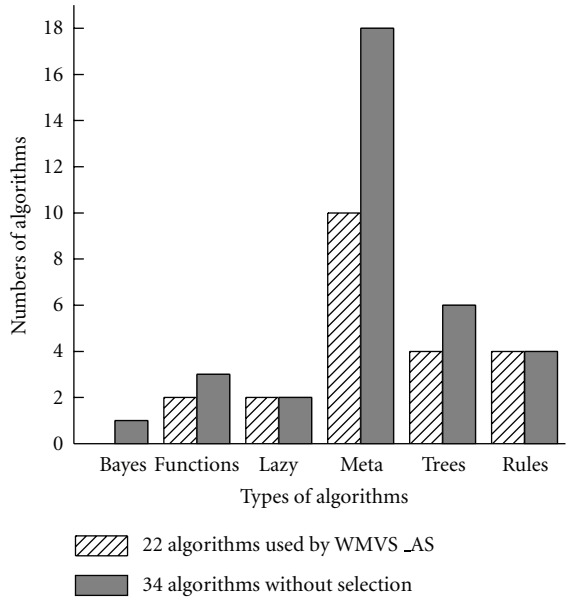

Figure 6: Distribution of algorithms.

3.3. Results of SMVS_AS and WMVS_AS. Algorithms are added into the voting system one by one according to the order of mRMR. The voting result of each added algorithm is plotted in Figure 5.

SMVS_AS and WMVS_AS achieve the highest average MCC value of 64.40% and 64.70% when the 22th algorithm is added. The curve in Figure 5 shows that WMVS_AS performs better than SMVS_AS in most cases, especially when the voting system involves an even number of algorithms. Voting systems with algorithm selection perform better than those without, indicating that some of the 34 algorithms cause a negative effect or no effect and should be excluded in

Table 2: The standard deviation of the 34 algorithms.

Algorithm	Standard deviation			
	Basic training dataset		Independent test dataset	
	ACC (%)	MCC (%)	ACC (%)	MCC (%)
AdaBoostM1	0.61	1.16	1.00	1.94
J48	0.88	1.76	1.42	2.84
IBk	0.52	1.01	1.18	2.21
MultiClassClassifier	0.60	1.21	1.04	2.09
PART	0.55	1.25	1.26	2.54
MultilayerPerceptron	1.26	2.52	2.22	3.04
KStar	0.72	1.41	1.07	2.00
Bagging	0.76	1.51	0.43	0.88
NBTree	0.82	1.64	2.04	4.09
Decorate	0.73	1.47	1.16	2.25
RandomForest	0.67	1.32	0.62	1.25
JRip	0.48	0.96	2.25	4.43
RandomCommittee	0.51	0.99	1.23	2.59
FilteredClassifier	1.11	2.22	1.16	2.32
ClassificationViaRegression	0.96	1.91	0.80	1.57
Dagging	0.70	1.38	1.00	2.00
AttributeSelectedClassifier	0.85	1.71	0.66	1.40
REPTree	0.71	1.46	1.32	2.66
SMO	0.55	1.10	1.06	2.11
J48graft	1.06	2.12	1.40	2.81
Ridor	1.01	2.14	1.70	3.44
RandomSubSpace	0.91	1.84	1.22	2.44
EnsembleSelection	0.78	1.60	1.35	2.42
SimpleLogistic	0.41	0.83	0.92	1.84
DecisionTable	0.98	2.06	1.86	3.87
DataNearBalancedND	0.88	1.76	1.42	2.84
RacedIncrementalLogitBoost	0.63	1.59	1.68	3.61
SimpleCart	0.63	1.26	1.13	2.25
LogitBoost	0.43	0.87	1.23	2.47
ND	0.88	1.76	1.42	2.84
BayesNet	0.51	1.02	1.02	2.10
ClassBalancedND	0.88	1.76	1.42	2.84
OrdinalClassClassifier	0.88	1.76	1.42	2.84
END	0.88	1.76	1.42	2.84

Table 3: The comparison of the predictors.

Predictor	Average predicted results		Standard deviation	
	ACC (%)	MCC (%)	ACC (%)	MCC (%)
Best individual algorithm	79.29	58.58	1.06	2.11
SMVS	79.72	59.77	0.76	1.49
WMVS	80.82	61.94	0.68	1.32
SMVS_AS	81.88	64.40	0.55	1.02
WMVS_AS	82.04	64.70	0.42	0.81

the voting. Thus algorithm selection is essential for a better classification performance.

3.4. Result of mRMR. In Weka version 3.5.7, the 34 algorithms are divided into Bayesian classifiers (Bayes), trees, rules, functions, metalearning algorithms (meta), and lazy classifiers (lazy). The number of algorithms of different types involved in the voting before algorithm selection and after algorithm selection is shown in Figure 6 (the number of algorithms used by WMVS_AS is average value of 22 algorithms). In terms of proportion, all adopted lazy

and rules classifiers are selected by the voting system, and around half of functions and tree classifiers are selected, indicating that there is less redundancy among these types of classifiers. The Bayes classifier is excluded, indicating that it performs negatively or has no effect in the voting. Because the number of metaclassifiers is the greatest among all types of classifiers involved, many of them are redundant and excluded from the voting. Nevertheless, more metaclassifiers remain in the voting than any other types of classifiers after the algorithm selection. On the whole, the number of classifiers of different types becomes evener after the algorithm selection, indicating that classifiers from different types tend to collaborate better in the voting than those from the same type.

4. Conclusions

To predict the interaction between proteins and RNA, we integrate a number of machine learning algorithms selected from Weka using four voting systems [21]. As a result, voting systems perform better than any single classifier, voting systems with algorithm selection perform better than those without, and weighted voting systems perform better than those without weighting. Weighted voting systems with algorithm selection achieve the best prediction results with 82.04% (ACC value) and 64.70% (MCC value) on the independent dataset.

Acknowledgments

This paper is supported by grants from the National Natural Science Foundation of China (20973108), the Key Research Program (CAS) (KSCX2-YW-R-112), Shanghai Leading Academic Discipline Project (J50101) and Systems Biology Research Foundation of Shanghai University, the National Natural Science Foundation of China (20902056), and Science Foundation of Shanghai for Excellent Young Teachers (B.37010107716).

References

[1] M. S. Jurica and M. J. Moore, "Pre-mRNA splicing: awash in a sea of proteins," *Molecular Cell*, vol. 12, no. 1, pp. 5–14, 2003.

[2] M. J. Moore, "From birth to death: the complex lives of eukaryotic mRNAs," *Science*, vol. 309, no. 5740, pp. 1514–1518, 2005.

[3] H. F. Noller, "RNA structure: reading the ribosome," *Science*, vol. 309, no. 5740, pp. 1508–1514, 2005.

[4] E. O. Freed and A. J. Mouland, "The cell biology of HIV-1 and other retroviruses," *Retrovirology*, vol. 3, pp. 77–87, 2006.

[5] Y. D. Cai and L. Lu, "Predicting N-terminal acetylation based on feature selection method," *Biochemical and Biophysical Research Communications*, vol. 372, no. 4, pp. 862–865, 2008.

[6] Y. D. Cai, Z. L. Qian, L. Lu et al., "Prediction of compounds' biological function (metabolic pathways) based on functional group composition," *Molecular Diversity*, vol. 12, no. 2, pp. 131–137, 2008.

[7] W. J. Li, K. Lin, K. Y. Fen, and Y. Cai, "Prediction of protein structural classes using hybrid properties," *Molecular Diversity*, vol. 12, no. 3-4, pp. 171–179, 2008.

[8] X. C. Xu, D. Yu, W. Fang et al., "Prediction of peptidase category based on functional domain composition," *Journal of Proteome Research*, vol. 7, no. 10, pp. 4521–4524, 2008.

[9] B. Niu, Y. Jin, L. Lu et al., "Prediction of interaction between small molecule and enzyme using AdaBoost," *Molecular Diversity*, vol. 13, no. 3, pp. 313–320, 2009.

[10] Y. D. Cai and S. L. Lin, "Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence," *Biochimica et Biophysica Acta—Proteins and Proteomics*, vol. 1648, no. 1-2, pp. 127–133, 2003.

[11] L. Y. Han, C. Z. Cai, S. L. Lo, M. C. M. Chung, and Y. Z. Chen, "Prediction of RNA-binding proteins from primary sequence by a support vector machine approach," *RNA*, vol. 10, no. 3, pp. 355–368, 2004.

[12] O. T. P. Kim, K. Yura, and N. Go, "Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction," *Nucleic Acids Research*, vol. 34, no. 22, pp. 6450–6460, 2006.

[13] L. Wang and S. J. Brown, "BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences," *Nucleic Acids Research*, vol. 34, pp. W243–W248, 2006.

[14] X. J. Yu, J. P. Cao, Y. D. Cai, T. Shi, and Y. Li, "Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines," *Journal of Theoretical Biology*, vol. 240, no. 2, pp. 175–184, 2006.

[15] J. Tong, P. Jiang, and Z. H. Lu, "RISP: a web-based server for prediction of RNA-binding sites in proteins," *Computer Methods and Programs in Biomedicine*, vol. 90, no. 2, pp. 148–153, 2008.

[16] Y. Wang, Z. Xue, G. Shen, and J. Xu, "PRINTR: prediction of RNA binding sites in proteins using SVM and profiles," *Amino Acids*, vol. 35, no. 2, pp. 295–302, 2008.

[17] E. N. Jeong and S. Miyano, "A weighted profile based method for protein-RNA interacting residue prediction," *Transactions on Computational Systems Biology IV*, vol. 3939, pp. 123–139, 2006.

[18] M. Terribilini, J. H. Lee, C. H. Yan, R. L. Jernigan, V. Honavar, and D. Dobbs, "Prediction of RNA binding sites in proteins from amino acid sequence," *RNA*, vol. 12, no. 8, pp. 1450–1462, 2006.

[19] M. Terribilini, J. H. Lee, C. H. Yan et al., "Identifying interaction sites in "recalcitrant" proteins: predicted protein and RNA binding sites in rev proteins of HIV-1 and EIAV agree with experimental data," *Pacific Symposium on Biocomputing*, pp. 415–426, 2006.

[20] A. F. Rahman, H. Alam, and M. C. Fairhurst, "Multiple classifier combination for character recognition: revisiting the majority voting system and its variations," *Document Analysis System V, Proceedings*, vol. 2423, pp. 167–178, 2002.

[21] L. Chen, L. Lu, K. Feng et al., "Multiple classifier integration for the prediction of protein structural classes," *Journal of Computational Chemistry*, vol. 30, no. 14, pp. 2248–2254, 2009.

[22] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[23] L. Liu, Y. D. Cai, W. C. Lu, K. Feng, C. Peng, and B. Niu, "Prediction of protein-protein interactions based on PseAA composition and hybrid feature selection," *Biochemical and Biophysical Research Communications*, vol. 380, no. 2, pp. 318–322, 2009.

[24] B. Niu, L. Lu, L. Liu et al., "HIV-1 protease cleavage site prediction based on amino acid property," *Journal of Computational Chemistry*, vol. 30, no. 1, pp. 33–39, 2009.

[25] W. Z. Li, L. Jaroszewski, and A. Godzik, "Clustering of highly homologous sequences to reduce the size of large protein databases," *Bioinformatics*, vol. 17, no. 3, pp. 282–283, 2001.

[26] G. Wang and R. L. Dunbrack, "PISCES: a protein sequence culling server," *Bioinformatics*, vol. 19, no. 12, pp. 1589–1591, 2003.

[27] H. Ian and E. F. Witten, *Data Mining: practical Machine Learning Tools and Techniques*, Morgan Kaufmann, San Francisco, Calif, USA, 2nd edition, 2005.