Sequence-Based Prediction of microRNA-Binding Residues in Proteins Using Cost-Sensitive Laplacian Support Vector Machines

Jian-Sheng Wu, and Zhi-Hua Zhou, *IEEE Fellow*

Abstract—The recognition of microRNA (miRNA)-binding residues in proteins is helpful to understand how miRNAs silence their target genes. It is difficult to use existing computational method to predict miRNA-binding residues in proteins due to the lack of training examples. To address this issue, unlabeled data may be exploited to help construct a computational model. Semi-supervised learning deals with methods for exploiting unlabeled data in addition to labeled data automatically to improve learning performance, where no human intervention is assumed. In addition, miRNA-binding proteins almost always contain a much smaller number of binding than non-binding residues, and cost-sensitive learning has been deemed as a good solution to the class imbalance problem. In this work, a novel model is proposed for recognizing miRNA-binding residues in proteins from sequences using a cost-sensitive extension of Laplacian Support Vector Machines (CS-LapSVM) with a hybrid feature. The hybrid feature consists of evolutionary information of the amino acid sequence (PSSMs), the conservation information about three biochemical properties (*HKM*) and mutual interaction propensities in protein-miRNA complex structures. The CS-LapSVM receives good performance with a *F1* score of 26.23±2.55 % and an *AUC* value of 0.805±0.020 superior to exisiting approaches for the recognition of RNA-binding residues. A web server called SARS is built and freely available for academic usage.

Index Terms—Laplacian Support Vector Machine, cost-sensitive learning, miRNA-binding residues, evolutionary information, mutual interaction propensities.

1 Introduction

MicroRNAs (miRNAs) are endogenous ~22 nt RNAs that act as important gene-regulatory roles in animals and plants by pairing to messenger RNA transcripts (mRNAs) of protein-coding genes to direct their posttranscriptional silence[1]. So far, miRNA research has revealed multiple roles in negative regulation (transcript degradation and sequestering, translational suppression) and possible involvement in positive regulation (transcriptional and translational activation) [1]. By affecting gene regulation, miRNAs are likely to be involved in most biological processes, such as in developmental timing, cell death, cell proliferation, haematopoiesis and patterning of the nervous system [2]. The process of miRNAs for silencing target mRNAs is performed by RNA-induced silencing complexes (RISCs) in which the main catalytic subunit is one of the Argonaute proteins (AGO), and miRNAs serve as a template for recognizing specific mRNA sequences [3]. Consequently, the recognition of miRNA-binding residues in RISCs can significantly improve our understanding of how miRNAs silences target genes and understanding of many related biological processes, and also provide further insights into protein functions and mechanisms of protein - miRNA specific interaction.

Recently, various computational methods have been developed to recognize RNA-binding residues in proteins. These methods can be roughly divided into two categories [4-8], *i.e.*, structure-based and sequence-based prediction methods. There are many types of RNA molecules with diverse structures, and the mechanisms of diverse RNA molecules recognizing their protein partners are often different. Thus, it is not easy to identify the actual miRNA-binding residues in proteins using the traditional RNA-binding residue prediction methods. It is desired to develop computational methods focusing on recognizing miRNA-binding residues in proteins.

Currently, there are few available structures of protein-miRNA complexes in the Protein Data Bank (PDB) database [9]. Thus, it is difficult to build a strong computational model for predicting miRNA-binding residues in proteins due to the lack of training examples. However, numerous miRNA-binding protein sequences can be obtained from the UniProt database [10]; such sequences provide abundant unlabeled instances for constructing classifiers to predict miRNA-binding residues in proteins. Because labeling the unlabeled data requires human effort and expertise, exploiting unlabeled data to help improve the learning performance has become a very hot topic during the past decade. There are two major techniques for this purpose [11, 12], *i.e.*, active learning and semi-supervised learning.

- *J.-S. Wu is with the National Key Laboratory for Novel Software Technology, Nanjing University, 210023, and the School of Geography and Biological Information of Nanjing University of Posts and Telecommunications, 210046, Nanjing, China. E-mail: wujs@lamda.nju.edu.cn.*
- *Z.-H. Zhou is with the National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China. E-mail: zhouzh@lamda.nju.edu.cn.*

Manuscript received (insert date of submission if desired). Please note that all acknowledgments should be placed at the end of the paper, before the bibliography.

Active learning [13,14] deals with methods which assume that the learner has some control over the input space, and the goal is to minimize the number of queries from human experts on ground-truth labels for building a strong learner. Semi-supervised learning [16-19] deals with methods for automatically exploiting unlabeled data in addition to labeled data to improve learning performance, where no human intervention is assumed. Transductive learning [15] is a specific type of semi-supervised learning, which tries to exploit unlabeled data automatically, but assumes that the unlabeled data are exactly the test data. Due to the expensive and time-consuming processes of experimental determination of protein structures and human effort and expertise on recognizing ground-truth miRNA-binding residues, we will focus on semi-supervised learning that tries to exploit unlabeled data without human intervention. The Laplacian SVM [20] is one of state-of-the-art semi-supervised learning methods that will be applied for building miRNA-binding residues prediction models by making use of both labeled and unlabeled data in this article.

MiRNA-binding proteins almost always contain a much smaller number of binding than non-binding residues. Learning algorithms that do not consider class-imbalance tend to be overwhelmed by the majority class and ignore the minority one [21]. However, in class-imbalance learning, the primary interest is in identifying the minor class [21]. That is, the cost of misclassifying a minor class example is usually more expensive than that of misclassifying a major one [22, 23]. The situation in our work is that there are only a few miRNA-binding residues and in practice it is more important that miRNA-binding residues will not be missed; thus, the cost of misclassifying a miRNA-binding residue (i.e., a missing of a binding) is more expensive than misclassifying a non-binding (i.e., a false alarm). In particular, cost-sensitive learning has been deemed as a good solution to the class imbalance problem, and a similar manner can be employed for learning from imbalanced data sets and learning when costs are unequal and unknown [24]. Therefore, a cost-sensitive learning scheme will be incorporated into the Laplacian SVM model to deal with the class imbalance problem for building a strong miRNA-binding prediction model.

In a word, the motivation of our work is that for the problem of recognizing miRNA-binding residues in proteins from sequences, there are insufficient labeled examples and the task suffers seriously from class-imbalance. Thus, we propose the CS-LapSVM algorithm, a cost-sensitive extension of Laplacian Support Vector Machine [25] for this task. In this paper, a hybrid feature is obtained by combining evolutionary information of the amino acid sequence (PSSMs), the conservation information about three biochemical properties (HKM) and mutual interaction propensities in protein-miRNA complex structures. The results show that our CS-LapSVM models reach a F1 score of 26.23±2.55% with an AUC value of 0.805±0.020 for recognizing miRNA-binding residues in proteins from sequences.

2 MATERIALS AND METHODS

2.1 Dataset

All protein-miRNA complex structures are collected from the Protein Data Bank(PDB) [9], and all miRNA-binding protein sequences have been downloaded from the Universal Protein Resource (UniProt) databank[10] (released by March 15, 2012) (Table 1). Then, redundancy among all protein sequences is removed by clustering analysis using the blastclust program in the BLAST package [26] from NCBI with a threshold of 25% for sequence identity. Thus, the non-redundant dataset MBP20 which contains 20 amino acid sequences (of these sequences 4 from PDB and 16 from Uniprot) is created by retaining only the longest sequence in each cluster (Table 1).

TABLE 1

The Original and Non-redundant Datasets Downloaded from the PDB and UniProt Databank.

Original				
PDB (ID)	3A6P	3ADI	3TRZ	3TS0
	3TS2			
UniProt (ID)	Q9XGW1	O04379	O04492	P92186
	Q8K3Y3	Q2KIA0	Q06413	F1LZC6
	Q8CFN5	A4UTP7	Q5R444	Q8TCS8
	Q8K1R3	Q5RCW2	Q01860	F7D1A4
	Q3MHX3	Q9BWF3	Q4R979	Q8C7Q4
	Q9BDY9	P48431	Q9GNA3	F1PAY8
	Q9U4F5	Q9GNA6	Q9GNJ2	Q9GND0
	Q8MRC7	Q9TW27	Q9TW12	E9Q6I7
	Q9NHW9	Q9W5S7	Q9NIH3	Q86LT0
	Q9U6N4			
Non-redundant				
PDB (ID_chain)	3TS0_B	3ADI_A	3A6P_A	3A6P_C
UniProt (ID)	Q9GNA6	Q8CFN5	Q4R979	Q5RCW2
	F7D1A4	O04379	Q9XGW1	O04492
	Q01860	P48431	Q9TW12	Q86LT0
	F1LZC6	P92186	E9Q6I7	Q8K3Y3

As in previous studies [5,27], an amino acid residue in a protein is defined to be a binding site if it contains at least one atom that falls within the cutoff distance of 3.5Å from any atoms of the miRNA molecule in the complex, and all other residues are labeled non-binding sites. Each instances is a segment of amino acid sequences with length l = 11. From a protein sequence with n residues, a total of $(n–l+1–r)$ instances are extracted, where l is the sliding window size and r is the number of residues that lack information about their atomic coordinates in the PDB entries. An instance is labeled positively if the central residue is miRNA-binding or negatively if the central residue is non-binding. Unlabeled instances are reached by scanning the miRNA-binding protein sequences with n residues from the UniProt databank using the same sliding window size l= 11 and a total of (n–l 1*)* instances are

extracted. Finally, the MBP20 dataset contains 61 positive, 1298 negative and 7983 unlabeled instances.

In order to build a true independent test dataset, labeled examples in the MBP20 dataset are randomly divided into two parts. The first part, which contains the two-third of labeled samples, acts as the training dataset. It is used to obtain the prediction performance of CS-LapSVM models, to analyze the contribution of various features on prediction performance and to study the impact of instance lengths on the classifiers' performance. The second part which comprises the one-third of labeled samples acts as an independent test dataset for implementing performance comparison with other methods. In order to eliminate the influence of randomly sampling on prediction performance, the process of randomly generating the training dataset and the independent test dataset is repeated five times and the results are stated by their mean and standard deviation of the performance of the five datasets.

2.2 Feature Descriptors

Nucleic acid molecules can recognize the specific structural motifs in proteins. Such motifs are more conserved in evolution and usually have preferences of some physico-chemical properties and the usage of amino acids. Therefore, it is beneficial to have a better understanding of protein-miRNA interaction and obtain novel feature descriptors for building classifiers by analyzing preferences of physico-chemical properties in miRNA-binding regions and mining their correlations among different properties. To highlight the importance of the nearest neighbor residues in determining whether a residue interacts with nucleotides and evaluate the contribution of physicochemical properties in affecting protein-miRNA interaction, we use the labels of training data to calculate the mutual interaction propensity of a residue triplet and a nucleotide during EACH round of cross validations [4]. A residue triplet is regarded as interacting with a nucleotide when its central residue is miRNA-binding. Here, the 20 kinds of amino acids are grouped into six classes based on their dipoles and side-chain volumes, namely Class *a*: Ala, Gly, Val; Class *b*: Ile, Leu, Phe, Pro; Class *c*: Tyr, Met, Thr, Ser, Cys; Class *d*: His, Asn, Gln, Tpr; Class *e*: Arg, Lys; and Class *f*: Asp, Glu [27]. Meanwhile, nucleic acids are clustered into two classes: purine *(i)* and pyrimidine *(j)*. The mutual interaction propensity is defined as follows [4]:

$$P(x,y)=\sum_{i,j} f_{i,j}(x,y)\log_2\frac{f_{i,j}(x,y)}{f_i(x)f_j(y)} \quad (2)$$

Here,

$$f_{i,j}(x,y)=N_{i,j}(x,y)/\sum_{x,y}N_{i,j}(x,y) \quad (3)$$

$$f_i(x)=N_i(x)/\sum_x N_i(x) \quad (4)$$

$$f_j(y)=N_j(y)/\sum_y N_j(y) \quad (5)$$

Where x represents a residue triplet composed of 6 classes of amino acids $(i.e.\ x\in\{(a,a,a),(a,a,b),...,(f,f,f)\})$, y is a nucleotide class $(i.e.,\ y\in\{i,j\})$, $N_{i,j}(x,y)$ is the number of residue triplet x binding with nucleotide y, $N_i(x)$ is the number of residue triplet x, and $N_j(y)$ is the number of nucleotide y. For an example with the length of l = 11 residues, 9 triplets are obtained and every triplet is represented by its corresponding values of mutual interaction propensity with two classes of nucleotides *(i, j)*.

Evolutionary information of amino acid sequences in terms of their position-specific scoring matrices (PSSMs) are generated for building miRNA-binding residue prediction models by the PSI-BLAST program[28] searched against the nonredundant (nr) dataset of amino acid sequences at NCBI with running for three iterations and the E-value threshold set to *1e-3*. The PSSM elements are scaled to the range 0–1 by the standard logistic function [29]:

$$f(x)=\frac{1}{1+\exp(-x)}. \quad (6)$$

The evolutionary information of amino acid sequences indicated by PSSMs is previously shown to improve the performance for predicting RNA-binding residues in proteins [5, 8]. However, it might miss some evolutionary information of amino acid sequences, for instance, the characteristics of the amino acid distribution and the preference of biochemical properties in some regions. Thus, new descriptors called *HKM* have been defined using the similar strategy as Wang's approach [6] to capture the conservation information about biochemical properties of miRNA-binding residues in the present study (Figure 1). For a given protein sequence p, its homologues $Hp = \{h_1, h_2, \ldots, h_j\}$ in a reference database can be retrieved and aligned to p using PSI-BLAST. Then, the sequence alignment is used to compute the mean and standard deviations of a feature for each residue a_i in the protein sequence p. In this study, three biochemical features of amino acids (*i.e* hydrophobicity, side chain pKa value and molecular mass) relevant to protein-nucleic acid interactions have been investigated. Hydrophobicity (feature *H*) plays a key role in protein folding. Hydrophobic amino acids are usually located inside proteins [6], but underrepresented at the miRNA interaction interfaces. The side chain pK_a value (feature *K*) expresses the ionization state of a residue. Because the phosphate groups of nucleic acids are negatively charged, the ionization state of amino acid side chains has influence on the interaction with miRNA molecules [6]. The value of molecular mass (feature *M*) of each amino acid is unique, and it is closely related to the volume of space occupied by the residue in protein structures [6]. MiRNA-binding residues may own the size constraint to be interacted with the interaction interface. In this manuscript, we arrive at the three descriptor values (*i.e* hydrophobicity, side chain pK_a value and molecular mass) from wang *et al.* [30]. Therefore, for each instance, the input vector contains 304 feature values, including 18 (2×9) mutual interaction propensity elements, 220 (20 × 11) PSSMs and 66 (6×11) *HKM*.

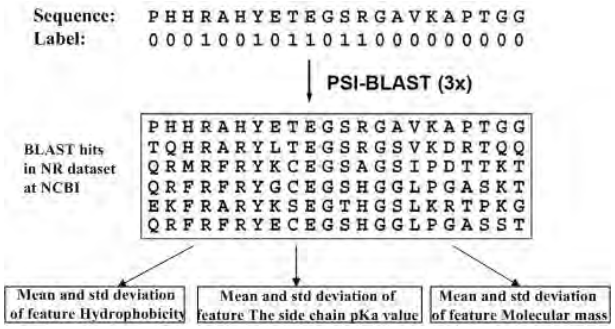

Fig. 1. Schematic diagrams for generating new descriptors called *HKM*.

2.3 Algorithms for classification

The Cost-Sensitive extension of Laplacian support vector machine, namely CS-LapSVM is used for building miRNA-binding residues prediction models in proteins from sequences in this work.

Laplacian support vector machine (LapSVM)[22] is a popular semi-supervised algorithm. It is built on two important factors. One is *manifold assumption*, *i.e.,* similar instances have similar outputs; the other is *large margin principle*, *i.e.,* the distributions of two different classes have a large margin.

Formally, given a set of training examples $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{l} \cup \{\mathbf{x}_{l+j}\}_{j=1}^{u}$ where $\{\mathbf{x}_i, y_i\}_{i=1}^{l}$ and $\{\mathbf{x}_{l+j}\}_{j=1}^{u}$ are labeled and unlabeled data, respectively. l and u are numbers of labeled and unlabeled data, respectively. LapSVM then aims to learn a decision function f such that

$$\arg\min_{f\in \mathrm{H}} \frac{1}{l}\sum_{i=1}^{l} V(x_i, y_i, f) + \gamma_A \|f\|_{\mathrm{H}}^2 + \frac{\gamma_I}{(u+l)^2}\sum_{i,j=1}^{l+u}(f(x_i)-f(x_j))^2 W_{ij}. \quad (7)$$

the following functional is minimized, *i.e.*,

Here the first term is the classification loss on labeled data, e.g., hinge loss in SVM that enforces the distributions of two different classes have a large margin; the second term prefers the decision function to be a simple classifier; while the third term enforces that similar instances have similar output according to the similarity weighted matrix W of all training instances. γ_A and γ_I are two parameters trading-off these three terms. It has been found that LapSVM is useful for many applications [31, 32].

It is evident that LapSVM is cost-blind because it does not take any misclassification cost into account. In this paper, we extend LapSVM for cost-sensitive scenarios. Specifically, for each labeled training example, the misclassification cost is incorporated into the classification loss, *i.e.*,

$$\arg\min_{f\in \mathrm{H}} \frac{1}{l}\sum_{i=1}^{l} c(y_i) V(x_i, y_i, f) + \gamma_A \|f\|_{\mathrm{H}}^2 + \frac{\gamma_I}{(u+l)^2}\sum_{i,j=1}^{l+u}(f(x_i)-f(x_j))^2 W_{ij}. \quad (8)$$

Where $c(y_i)$ corresponds to the misclassification cost for label y_i. This leads to our proposed CS-LapSVM. It can be shown that CS-LapSVM is a convex optimization whose global optimal solution can be solved efficiently. Moreover, LapSVM package is already public and thus CS-LapSVM is easy to implement because CS-LapSVM only makes a minor modification with LapSVM. As will be verified in empirical study, such a simple modification works quite well.

2.4 Measurement of classifier's performance

The *F1 score* is the harmonic mean of precision and recall (that is, $F1 = 2pr/(p + r)$, where p is precision and r is recall), which is a more stable measure, especially for data sets with huge class-imbalance [33]. In this article, the *F1 score* is the key criteria for selecting the optimal classifier during the process of training models for predicting miRNA-binding residues in proteins. The receiver operating characteristic (ROC) curve is to plot the true positive rate (sensitivity) against false positive rate (1-specificity), and the area under the ROC curve (*AUC*) is a reliable measure for evaluating classifier performance [34]. For comparison with other methods, the recall, precision, Matthew's correlation coefficient (*MCC*) [35], *F1* score and *AUC* value are used to assess the prediction performance in this study.

3 RESULTS AND DISCUSSION

3.1 Prediction performance of the CS-LapSVM method

The CS-LapSVM models are trained by a three-fold cross-validation procedure for predicting miRNA-binding residues in proteins from sequences. In this study, the parameters γ_A and γ_I in Equation (7) are fixed to 100 and 0.1, respectively. The gaussian radial basis function, *i.e., exp(-gamma(x-x$_i$)^2)*, which corresponds to the decision function f in Equation (7) is used in this paper, and the parameter *gamma* ranges from 2^(-5) to 2^5 for optimizing during cross validations. The parameter $c(y_i)$ in Equation (8) corresponds to the misclassification cost for label y_i which involves in two parts, i.e., the cost of misclassifying negative class into positive one and the cost of misclassifying positive class into negative class *(opt.cost)*. In this work, the cost of misclassifying negative class into positive one is fixed to 1, whereas the cost of misclassifying positive class into negative class *(opt.cost)* is set to 5,10,15,20 and 25 for optimizing. Indeed, the "cost ratio" is crucial, rather than the absolute cost values; setting the cost of "misclassifying negative to positive" vs. "misclassifying positive to negative" as "1 vs 5" is equal to setting them as "10 vs. 50" or "15 vs. 75". The optimal values for parameter *gamma* and *opt.cost* are 2^3 and 10 respectively after implementing a standard grid search method for cross validations. The five training datasets which are randomly generated from the MBP20 dataset as the scheme described in the Section 2.1 are used, and the results are stated by their mean and standard deviation of the performance on the five datasets.

In this article each instance for training classifiers is a segment of amino acid sequences with a certain length. We first study the impact of instance lengths on the classifiers' performance for identifying miRNA-binding residues in proteins by CS-LapSVM Models from sequences. As indicated in Table 2, when an instance has a length of 11 amino acids, the CS-LapSVM Models achieve the best performance with an *F1*-score of 26.23±2.55% and an *AUC* (area under the ROC curve) value of 0.805±0.020.

In addition, the contributions of each kind of features are also considered, see Table 3. A position-specific scor-

ing matrix (PSSM) is a commonly used representation of motifs and evolutionary information of amino acid sequences. The classifier with the PSSM feature just receives a 13.83±1.43% *F1* score and a 0.749±0.011 *AUC* value. The mutual interaction propensities are to describe preferences of some physico-chemical properties and the usage of amino acids in miRNA-binding regions and mining their correlations among different properties. The classifier only with the mutual interaction propensities obtains a 4.76±0.50% *F1* score and a 0.906±0.051 *AUC* value. The *HKM* descriptor is to capture the conservation information about biochemical properties of miRNA-binding residues. The classifier only with the *HKM* feature reaches a 14.08±2.34% *F1* score and a 0.672±0.075 *AUC* value. The combination of all features reports the best performance, indicating that the combination of all features is capable of capturing more information for discriminating miRNA-binding sites from non-binding ones.

TABLE 2
Performance Comparison Based on Different Lengths for Defining an Instance in Predicting MiRNA-binding Residues in Proteins from Sequences by a CS-LapSVM Model. The Five Training Datasets Which Are Randomly Generated from the MBP20 Dataset as the Scheme Described in the Section 2.1 Are Used, and the Results Are Stated by Their Means and Standard Deviations (*Mean ±std*) of the Performance on the Five Datasets. The Results Show that the Model Achieved the Best Prediction Performance When the Instance Length Is 11 Amino Acids.

Length	*F1* score (%)	Recall (%)	Precision (%)	MCC	*AUC*
7	24.82±1.07	61.54±4.35	15.55±0.67	0.246±0.017	0.762±0.005
9	26.28±1.28	42.19±0.86	19.08±1.63	0.232±0.012	0.796±0.009
11	**26.23±2.55**	**63.00±9.59**	**16.77±2.37**	**0.266±0.024**	**0.805±0.020**
13	23.18±0.67	54.00±7.03	14.82±0.31	0.220±0.016	0.771±0.004
15	17.61±1.26	73.33±0.00	10.01±0.82	0.185±0.016	0.757±0.005

TABLE 3
The Performance of Our CS-LapSVM Models with Various Features for Predicting MiRNA-binding Residues in Proteins from Sequences. The Five Training Datasets Which Are Randomly Generated from the MBP20 Dataset as the Scheme Described in the Section 2.1 Are Used, and the Results Are Stated by Their Means and Standard Deviations (*Mean ±std*) of the Performance on the Five Datasets.

Feature	*F1* score (%)	Recall(%)	Precision (%)	MCC	*AUC*
A	13.83±1.43	79.00±6.02	7.60±0.90	0.141±0.017	0.749±0.011
B	4.76±0.50	2.50±0.21	50.00±5.11	0.104±0.023	0.906±0.051
C	14.08±2.34	25.00±3.15	9.80±0.81	0.094±0.018	0.672±0.075
AB	23.55±10.41	36.00±3.79	19.65±11.90	0.198±0.110	0.729±0.075
AC	10.38±0.60	95.00±5.30	5.49±0.32	0.095±0.027	0.783±0.077
BC	20.49±6.41	36.00±9.12	15.93±7.11	0.169±0.066	0.693±0.039
ABC	**26.23±2.55**	**63.00±9.59**	**16.77±2.37**	**0.266±0.024**	**0.805±0.020**

A:PSSMs B: Mutual interaction propensities C:HKM

3.2 Performance comparison with other methods

In this article, several machine learning methods are applied to compare with our CS-LapSVM method. During the process of optimizing the parameters of these methods, a standard grid search method is also utilized and the *F1* score is the key selection criteria. After obtaining the optimal parameters, the five training datasets randomly generated from the MBP20 dataset described in the Section 2.1 are used and the results are stated by their mean and standard deviation of the performance on the five datasets (Table 4).

When compared with the primary LapSVM method [19] (*i.e.,* without considering the cost-sensitive problem), the results show that the prediction performance is improved after incorporating the misclassification cost for labeled examples in the LapSVM model (Table 4). The cost-sensitive semi-supervised support vector machine (CS4VM)[22] is an efficient algorithm that first estimates the label means of the unlabeled instances, and then trains the CS4VM with the plug-in label means by an efficient SMO solver. The results show that our CS-LapSVM method is superior to the CS4VM approach (Table 4). The transductive SVM (TSVM) method [36] is one of the state-of-the-art semi-supervised learning algorithms. TSVM is cost-blind and in the experiments we extend it for cost-sensitive learning as in the CS-LapSVM. It can be seen in Table 4 that our CS-LapSVM method outperforms the cost-sensitive extension of TSVM (CS-TSVM).

In addition, a supervised cost-sensitive SVM (CS-SVM) model is built by using only the labeled training examples. The results indicate that the method gives poor performance with a 0% *F1*-score when no unlabeled data is appended for building classifiers (Table 4). A similar situations also happens in the prediction using the traditional SVM method (Table 4). The main reason for poor performances reported by the CS-SVM and SVM classifiers is that all examples are forecasted as negative class due to the small size of labeled examples and a huge imbalance of positive versus negative examples in training classifiers. Thus, it can be inferred that the unlabeled instances is helpful for understanding the overall space distribution of instances and finding the optimal classification hyperplane for separating miRNA-binding from non-binding residues.

To further illustrate the impact of unlabeled instances on classifiers' performance, two foreign instance datasets (*i.e.,* siRNA-binding proteins and piRNA-binding proteins) have been used to build models based on the CS-LapSVM algorithm for predicting miRNA-binding residues. The results show that both CS-LapSVM classifiers based on the two unlabeled instance datasets deteriorate prediction performances (Table 5).

The reason why the unlabeled instances significantly contribute to the excellent performance is that the labeled examples are insufficient to reflect the overall space distribution of instances and the geometry of the marginal distribution. That is, the models only based on labeled examples are not easy to get the ground-truth classification hyperplane for separating miRNA-binding residues from non-binding. Semi-supervised learning tries to exploit unlabeled data to help improve learning performance, particularly when there are limited labeled training examples [15]. Therefore, the CS-LapSVM algorithm

is developed to exploit unlabeled instances for understanding the geometry of the marginal distribution and improving the learning performance in this manuscript.

Belkin illustrates how unlabeled instances may force us to restructure our hypotheses during learning [20]. When labeled examples are insufficient, additional unlabeled examples are helpful to exploit the geometry of the marginal distribution which should be incorporated in the regularization principle to impose structure on the space of functions in nonparametric classification or regression [20]. In addition, the models based on foreign unlabeled instance datasets deteriorate learning performances (Table 5) .One important reason for the usefulness of unlabeled instances lies in the fact that they provide some information on the data distribution which is able to help the construction of prediction model when the amount of labeled data is limited [18]. These "foreign" unlabeled instances, however, come from other data sets, are not able to provide distribution information of the original data sets. Thus, it is not strange that by using these foreign unlabeled instances, the learning process is misled and therefore leads to a worse performance.

TABLE 4
Performance Comparison with Other Algorithms. The Five Training Datasets and Independent Test Sets Which Are Randomly Generated from the MBP20 Dataset as the Scheme Described in the Section 2.1 Are Implemented, and the Performance Are Presented by Their Means and Standard Deviations (Mean ± std).CS-LapSVM：the Cost-sensitive Extension of Laplacian SVM，LSVM：Laplacian SVM, CS4VM：The Cost-Sensitive Semi-Supervised SVM，CS-TSVM: the Cost-sensitive Extension of Transductive SVM,CS-SVM：a Supervised Cost-sensitive SVM, SVM: Support Vector Machine, RF: Random Forest.

Algorithms	*F1* score (%)	Recall(%)	Precision (%)	MCC	*AUC*
CS-LapSVM	**20.87±3.46**	**57.14±8.13**	**12.77±3.86**	**0.197±0.029**	**0.812±0.026**
LSVM	15.79±2.48	71.43±9.72	8.88±2.83	0.155±0.017	0.770±0.018
CS4VM	21.05±3.78	38.1±5.20	14.55±5.11	0.175±0.026	0.751±0.067
CS-TSVM	12.5±1.44	9.52±3.89	18.18±0.93	0.101±0.012	0.687±0.038
CS-SVM	0±0	0±0	0±0	0±0	0.831±0.087
SVM	0±0	0±0	0±0	0±0	0.835±0.069
RF	15.6±4.67	41.9±8.51	9.62±3.10	11.58±6.63	0.695±0.045

TABLE 5
Performance Comparison with Two Foreign Unlabeled Instance Datasets.That is, Unlabeled Instances from siRNA-Binding Proteins or piRNA-Binding Proteins are Used to Build Models Based on the CS-LapSVM Algorithm for Predicting miRNA-Binding Residues. The Five Training Datasets Which Are Randomly Generated from the MBP20 Dataset as the Scheme Described in the Section 2.1 Are Used, and the Results Are Stated by Their Means and Standard Deviations (*Mean ±std*) of the Performance on the Five Datasets.

Unlabeled data	*F1* score (%)	Recall(%)	Precision (%)	MCC	*AUC*
miRNA	**20.87±3.46**	**57.14±8.13**	**12.77±3.86**	**0.197±0.029**	**0.812±0.026**
piRNA	13.68±2.32	38.1±4.35	8.33±1.23	0.091±0.009	0.688±0.013
siRNA	12.28±1.65	33.33±6.03	7.53±0.63	0.069±0.007	0.680±0.009

3.3 Performance Comparison with RNA-binding Residue Classifiers

Until now, there was no customized computational method for predicting miRNA-binding residues in proteins. Therefore, the RNA-binding residue prediction models are implemented for performance comparison with our CS-LapSVM method (Table 6). In this paper, three web servers, namely, PRBR [5], BindN+ [6] and BindN [30] are taken into consideration due to the same cutoff distance of 3.5Å for defining binding residues in proteins as our method. In order to ensure that the amino acid sequences for extracting training examples are non-homologous with these for testing, four sequences with known three-dimensional strutures in the MBP20 dataset (Table 1) are also used to implement the performance comparison with RNA-binding residue prediction web servers. Each sequence is evaluated by the model trained by the other three sequences in our CS-LapSVM method. The mean and standard deviation of the prediction performance of four sequences are shown in Table 6, and the results indicate that our CS-LapSVM method is obviously better than the web servers for predicting miRNA-binding residues in proteins from sequences. Currently, there are so few available protein-miRNA complex structures in the PDB database. The number of labeled examples for training our CS-LapSVM classifiers is much smaller than that for building each of the three web servers. Hence, it can be concluded that unlabeled instances significantly contribute to the excellent performance of our method for predicting miRNA-binding residues in proteins from sequences. In addition, note that these approaches, such as BindN, were not specially designed to predict the miRNA-binding residues, and thus the comparison is somewhat unfair. However, this comparison discloses that RNA-binding residue prediction approaches do not work well on miRNA-binding residue prediction, and thus we need to develop methods for miRNA-binding residue prediction.

On the other hand, Figure 2 exhibits the details about prediction of four amino acid sequences by our model in a more intuitive manner and graphical visualization is implemented by the PyMOL molecular graphics tool (http://www.pymol.org).

3.4 Application of the CS-LapSVM method on other datasets

In this article, in order to check the learning performances of our CS-LapSVM algorithm on other small sample datasets, models are built for recognizing two other types of small RNA-binding residues (*i.e.,* siRNA-binding residues and piRNA-binding residues) in proteins from sequences (Table 6). We obtain the labeled and unlabeled data for recognizing siRNA-binding residues and piRNA-binding residues using exactly the same scheme as that for miRNA-binding residues which is described in the Section 2.1 and 2.2. The results show that our CS-LapSVM models reach obvious performance improvement on predicting siRNA-binding residues or piRNA-binding residues by making use of both labeled and unlabeled data after comparing with the SVM-based and random forest-

based models only using labeled data (Table 6). We think that the CS-LapSVM algorithm perhaps has the potential to handle more problems with small-samples.

TABLE 6
Performance Comparison with Three RNA-binding Residue Prediction Web Servers (i.e., BindN+, BindN and PRBR) Due to the Same Definion for Binding Residues in Proteins as Our Method. All Classifiers Are Evaluated by Test examples from Four Amino Acid Sequences (3A6P_A, 3A6P_C, 3ADI_A and 3TS0_B) with Known Three-dimensional Strutures in Our MBP20 Dataset, and the Final Performances Are Stated Using Their Means and Standard Deviations (Mean ± SD). Examples From Each Sequence Are Assessed by the Model Trained by Examples from the Other Three Sequences in our CS-LapSVM Method. NA: No Available AUC Values for the PRBR Classifier Due to the Lack of Prediction Scores.

Classifiers	F1 score (%)	Recall (%)	Precision (%)	MCC	AUC
CS-LapSVM	30.78±9.86	77.12±33.42	21.23±8.19	0.264±0.068	0.869±0.122
BindN+[6]	26.52±21.70	48.27±39.55	18.41±15.19	0.169±0.223	0.713±0.156
BindN [30]	21.94±14.32	63.61±25.09	16.76±15.90	0.18±0.054	0.742±0.161
PRBR [5]	15.58±14.43	26.22±23.82	11.14±10.39	0.034±0.066	NA

TABLE 7
The Performance of Our CS-LapSVM Models for Recognizing siRNA-binding or piRNA-binding Residues in Proteins from Sequences. The Five Training Datasets Which Are Randomly Generated for Both siRNA-binding and piRNA-binding Residues Datasets as the Scheme Described in the Section 2.1 Are Used, and the Results Are Stated by Their Means and Standard Deviations (*Mean ±std*) of the Performance on the Five Datasets. SVM: Support Vector Machine, RF: Random Forest.

Algorithms	*F1* score (%)	Recall (%)	Precision (%)	MCC	*AUC*
siRNA-binding Residue					
CS-LapSVM	**29.41±2.12**	**66.47±3.36**	**18.91±1.68**	**0.260±0.025**	**0.787±0.020**
SVM	0.00±0.00	0.00±0.00	0.00±0.00	0.000±0.000	0.810±0.018
RF	23.47±3.67	65.19±6.87	14.44±2.81	0.189±0.032	0.728±0.032
piRNA-binding Residue					
CS-LapSVM	**28.80±6.62**	**39.00±4.18**	**24.10±9.60**	**0.212±0.081**	**0.702±0.027**
SVM	0.00±0.00	0.00±0.00	0.00±0.00	-0.018±0.025	0.523±0.084
RF	13.10±5.59	37.36±25.19	8.30±3.31	-0.002±0.104	0.552±0.076

3.5 Web server

SARS is available at http://cbi.njupt.edu.cn/SARS/SARS.htm. On the SARS web page, users can copy/paste amino acid sequences (≤4 pieces in one run of prediction and only in FASTA format) and an E-mail address is required to receive the results. The CS-LapSVM model which is used for predicting new proteins in SARS is constructed from all the labeled and unlabeled data in the MBP20 dataset. The CS-LapSVM algorithm is coded by Matlab and implemented by the generated executable file with the postfix *exe*. The program slides a window with length l = 11 amino acids along the input sequence to receive multiple segments of amino acid sequences. Each window segment is arranged as an instance and the central residue in each window is labeled as whether it binds to miRNA molecules or not. Each instance will be maped into a 304-dimention feature space which consists of 18 mutual interaction propensity elements, 220 PSSMs elements and 66 *HKM* elements. The web server outputs the prediction results of each residue which consists of its predicted label, output score.

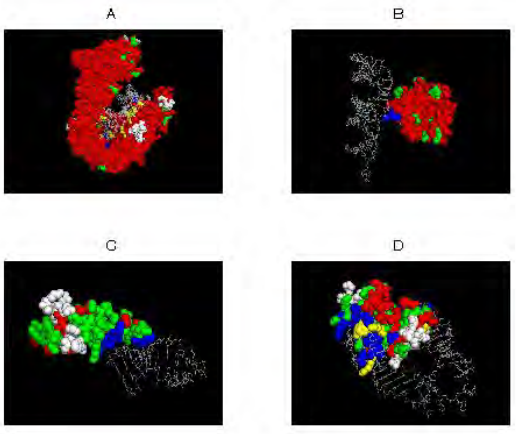

Fig. 2. Prediction results of residues within four amino acid sequences (A) the A chain of complex structure 3A6P (PDB ID), true positives (TP) 7, false negatives (FN) 17 , false positives (FP) 23 and true negatives (TN) 975 , with a 25.93% *F1* score and a 0.789 *AUC* value, (B) the C chain of complex structure 3A6P , TP 2, FN 0, FP 12 and TN 146, with a 25% *F1* score and a 0.962 *AUC* value, (C) the A chain of complex structure 3ADI , TP 6,FN 0, FP 33 and TN 22, with a 26.67% *F1* score and a 0.985 *AUC* value, (D) the B chain of complex structure 3TS0, TP 23,FN 6,FP 49 and TN 38, with a 45.54% *F1* score and a 0.741 *AUC* value. The correctly identified binding residues (TP) are in blue space fill; the correctly identified non-binding residues (TN) are in red space fill; the binding residues with negative predictions (FN) are in yellow space fill; the non-binding residues but wrongly predicted as positives (FP) are in green space fill.The total 10 residues located in the N-terminal and C-terminal of the four amino acid sequences are not used in reporting prediction performance by our model and shown in white space fill. The miRNA molecules are indicated in gray wire frame. The presentation in the format of three-dimensional structures is generated with PyMOL (http://www.pymol.org).

4 CONCLUSIONS

In this study, we build the first model for predicting miRNA-binding residues in proteins from sequences using the cost-sensitive extension of Laplacian Support Vector Machine (CS-LapSVM) method with a hybrid feature. The hybrid feature is composed of evolutionary information of amino acid sequences，the conservation information about three biochemical properties and mutual

interaction propensities in protein-miRNA complex structures. The results show that the CS-LapSVM model achieves a *F1* score of 26.23±2.55% with an *AUC* value of 0.805±0.020. From comparison with other machine learning methods, the results demonstrate that our CS-LapSVM method is the most effective for predicting miRNA-binding residues in proteins from sequences without using 3D structural information. A web server called SARS has been built for efficient online predictions. In the next step, we will attempt to capture stronger feature attributes and deal with the class-imbalance problem in order to propose a better miRNA-binding residue prediction model, and extend the model to to handle more problems with small-samples.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation of China (61073097, 61272217, 61203289),the National Fundamental Research Program of China (2010CB327903) and China Postdoctoral Science Foundation (2013T60523). The authors wish to thank the anonymous reviewers and the editor-in-chief Prof. Ying Xu for their constructive comments and suggestions.The authors wish to thank Yu-Feng Li for his helpful discussions and Nan Li and Ekhine Irurozki for their reading draft version of the paper. Z.-H. Zhou is the corresponding author.

REFERENCES

[1] D. P. Bartel, "MicroRNAs: target recognition and regulatory functions," *Cell,* vol. 136, pp. 215-233, Jan. 2009.

[2] L. He and G. J. Hannon, "MicroRNAs: small RNAs with a big role in gene regulation," *Nat Rev Genet,* vol. 5, pp. 522-31, Jul. 2004.

[3] T. M. Rana, "Illuminating the silence: understanding the structure and function of small RNAs," *Nat Rev Mol Cell Biol,* vol. 8, pp. 23-36, Jan. 2007.

[4] Z. P. Liu, *et al.*, "Prediction of protein-RNA binding sites by a random forest method with combined features," *Bioinformatics,* vol. 26, pp. 1616-1622, Jul. 2010.

[5] X. Ma, *et al.*, "Prediction of RNA-binding residues in proteins from primary sequence using an enriched random forest model with a novel hybrid feature," *Proteins,* vol. 79, pp. 1230-1239, Apr. 2011.

[6] L. Wang, *et al.*, "BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features," *BMC Syst Biol,* vol. 4, Suppl 1, pp. S3, May 2010.

[7] Y. Murakami, *et al.*, "PiRaNhA: a server for the computational prediction of RNA-binding residues in protein sequences," *Nucleic Acids Res,* vol. 38, pp. W412-416, Jul. 2010.

[8] M. B. Carson, *et al.*, "NAPS: a residue-level nucleic acid-binding prediction server," *Nucleic Acids Res,* vol. 38, pp. W431-435, Jul. 2010.

[9] H. M. Berman, *et al.*, "The Protein Data Bank," *Nucleic Acids Res,* vol. 28, pp. 235-242, Jan. 2000.

[10] R. Apweiler, *et al.*, "UniProt: the Universal Protein knowledgebase," *Nucleic Acids Res,* vol. 32, pp. D115-119, Jan. 2004.

[11] Z.-H. Zhou, "When semi-supervised learning meets ensemble learning," in *Proc. 8th International Workshop on Multiple Classifier Systems (MCS'09),* pp. 529-538, 2009.

[12] Z.-H. Zhou, "Learning with unlabeled data and its application to image retrieval," in *Proc. 9th Pacific Rim International Conference on Artificial Intelligence (PRICAI 2006),* pp. 5-10, 2006.

[13] B. Settles, "Active learning literature survey," Technical Report 1648, Department of Computer Sciences, University of Wisconsin at Madison, Wisconsin, WI, 2009.

[14] S. J. Huang, R. Jin, and Z.-H. Zhou, "Active learning by querying informative and representative examples," in *Advances in Neural Information Processing Systems (NIPS'10),* vol. 23, pp. 892-900, 2010.

[15] V. Vapnik, "Statistical learning theory," Wiley, New York, 1998.

[16] Z.-H. Zhou and M. Li, " Tri-training: Exploiting unlabeled data using three classifiers," *IEEE Transactions on Knowledge and Data Engineering,* vol. 17, pp. 1529-1541, 2005.

[17] O. Chapelle, B. Sch kopf, and A. Zien. "Semi-Supervised Learning",MIT Press, Cambridge, MA, 2006.

[18] Z.-H. Zhou and M. Li, "Semi-supervised learning by disagreement," *Knowledge and Information Systems,* vol. 24, pp. 415-439, 2010.

[19] X. Zhu, "Semi-supervised learning literature survey," Technical Report 1530, Department of Computer Sciences, University of Wisconsin at Madison, Madison, WI, 2006.

[20] M. Belkin, P. Niyogi, and V. Sindhwani , "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *Journal of Machine Learning Research,* vol. 7, pp. 2399-2434, 2006.

[21] N. V. Chawla, , N. Japkowicz, and A. Kotcz, "Editorial: special issue on learning from imbalanced data sets," *ACM SIGKDD Explorations,* vol. 6, pp. 1-6, 2004.

[22] D. Margineantu, "When does imbalanced data require more than cost-sensitive learning," in *Proceedings of the AAAI'2000 Workshop on Learning from Imbalanced Data Sets*, pp. 47-50,2000.

[23] G. M. Weiss and F. J. Provost, "Learning when training data are costly: The effect of class distribution on tree induction," *J. Artif. Intell. Res. (JAIR),* vol. 19, pp. 315-354, 2003.

[24] M. A. Maloof, "Learning when data sets are imbalanced and when costs are unequal and unknown," in *ICML'03 Workshop on Learning from Imbalanced Data Sets II*, 2003.

[25] Y. F. Li, J.T. Kwok, and Z.-H. Zhou, "Cost-sensitive semi-supervised support vector machine," in *Proc. 24th AAAI Conference on Artificial Intelligence (AAAI'10),* pp. 500-505,2010.

[26] S. F. Altschul, *et al.*, "Basic local alignment search tool," *J Mol Biol,* vol. 215, pp. 403-410, Oct. 1990.

[27] J. Wu, *et al.*, "Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature," *Bioinformatics,* vol. 25, pp. 30-35, Jan. 2009.

[28] S. F. Altschul, *et al.*, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Res,* vol. 25, pp. 3389-3402, Sep. 1997.

[29] Y. Wang, *et al.*, "Better prediction of the location of alpha-turns in proteins with support vector machine," *Proteins,* vol. 65, pp. 49-54, Oct. 2006.

[30] 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 Res,* vol. 34, pp. W243-248, Jul. 2006.

[31] L. Gómez-Chova, L. G. Camps-Valls, J.Muñoz-Marí, and J. Calpe, "Semisupervised image classification with Laplacian support vector machines," *IEEE Geoscience and Remote Sensing Letters,* vol. 5, pp. 336-340, 2008.

[32] J. Wu, *et al.*, "A semi-supervised learning based method: Laplacian support vector machine used in diabetes disease diagnosis," *Interdisciplinary Sciences: Computational Life Sciences,* vol. 1, pp. 151-155, 2009.

[33] X. B. Xue and Z.-H. Zhou, "Distributional features for text categorization," *IEEE Transactions on Knowledge and Data Engineering,,* vol. 21, pp. 428-442, 2009.

[34] C. E. Metz, "Basic principles of ROC analysis," *Semin Nucl Med,* vol. 8, pp. 283-298, Oct. 1978.

[35] D. Matthews, *et al.*, "Homeostasis model assessment: insulin resistance and β -cell function from fasting plasma glucose and insulin concentrations in man," *Diabetologia,* vol. 28, pp. 412-419, 1985.

[36] T. Joachims, "Transductive inference for text classification using support vector machines," in *In Proc. 16th Int'l Conf. Mach.Learn.,* pp. 200-209,1999.

Jian-Sheng Wu received the BS, MS and PhD degrees in bioengineering, ecology, biomedical engineering from nanchang University, east china normal university, southeast university, China, in 2000, 2004, and 2009, respectively. He joined the School of Geography and Biological Information of Nanjing University of Posts and Telecommunications, China, in 2009. Currently, He is also a postdoctoral fellow of the National Key Laboratory for Novel Software Technology in Nanjing University and a member of LAMDA group. His research interests are mainly in machine learning and bioinformatics. In these areas he has published more than 10 papers in leading journals and conference proceedings.

Zhi-Hua Zhou (S'00-M'01-SM'06) received the BSc, MSc and PhD degrees in computer science from Nanjing University, China, in 1996, 1998 and 2000, respectively, all with the highest honors. He joined the Department of Computer Science \& Technology at Nanjing University as an assistant professor in 2001, and is currently professor and Director of the LAMDA group. His research interests are mainly in artificial intelligence, machine learning, data mining, pattern recognition and multimedia information retrieval. In these areas he has published more than 100 papers in leading international journals or conference proceedings, and holds 12 patents. He has won various awards/honors including the IEEE CIS Outstanding Early Career Award, the National Science \& Technology Award for Young Scholars of China, the Fok Ying Tung Young Professorship Award, the Microsoft Young Professorship Award and nine international journals/conferences paper or competition awards. He is an Associate Editor-in-Chief of the Chinese Science Bulletin, Associate Editor of the ACM Transactions on Intelligent Systems and Technology and on the editorial boards of various other journals. He is the founder and Steering Committee Chair of ACML, and Steering Committee member of PAKDD and PRICAI. He serves/ed as General Chair/Co-chair of ACML'12, ADMA'12 and PCM'13, Program Chair/Co-Chair for PAKDD'07, PRICAI'08, ACML'09, SDM'13, etc., Workshop Chair of KDD'12, Program Vice Chair or Area Chair of various conferences, and chaired many domestic conferences in China. He is the Chair of the Machine Learning Technical Committee of the Chinese Association of Artificial Intelligence, Chair of the Artificial Intelligence \& Pattern Recognition Technical Committee of the China Computer Federation, Vice Chair of the Data Mining Technical Committee of IEEE Computational Intelligence Society and the Chair of the IEEE Computer Society Nanjing Chapter. He is a fellow of the IAPR, the IEEE, and the IET/IEE.