A Tool Preference Selection Method for RNA Secondary Structure Prediction with SVM[1]

Chiou-Yi Hor[2], Chang-Biau Yang[2,3], Chiou-Ting Tseng[2], and Hung-Hsin Chen[2]
[2]Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan

Abstract— *Prediction of RNA secondary structures has drawn much attention from both biologists and computer scientists. Many useful tools have been developed for this purpose, with or without pseudoknots. These tools have their individual strength and weakness. As a result, we propose a tool preference selection method which integrates three prediction tools pknotsRG, RNAstructure and NUPACK with support vector machines (SVM). Our method starts with extracting features from the target RNA sequences, and adopt the information-theoretic feature selection method for feature ranking. We propose a method to combine feature selection and classifier fusion, namely incremental mRMR. The test data set contains 720 RNA sequences, where 225 pseudoknotted RNA sequences are obtained from PseudoBase, and 495 nested RNA sequences are obtained from RNA SSTRAND. Our method serves as a preprocessing way in analyzing RNA sequences before the RNA secondary structure prediction tools are employed. Experimental results show that our method improves not only the classification accuracy, but also the base-pair accuracies.*

Keywords: *RNA, SVM, Fusion, Feature Selection, Tool Preference*

1. Introduction

An RNA secondary structure is the *fold* of the given sequence. The sequence is folded due to bonds between non-adjacent nucleotides. These bonded nucleotide pairs are called *base pairs*. Three possible combinations of nucleotides may make a base pair: A-U, G-C, and G-U, where A-U and G-C are called *Watson-Crick pairs* and G-U is called the *Wobble pair*. The *RNA secondary structure prediction problem* is to identify the folding configuration of a given RNA sequence.

The methods for predicting RNA secondary structure could be roughly categorized into two types. They are based on thermodynamics [26], [27], [31], and comparative approaches [5], [29], respectively. Since these tools resort to different criteria, each of them has its own metric and weakness. With their variety, we propose a tool preference selection method that integrates these software in order to improve prediction capability. Our method is based on the machine learning approach. It includes feature extraction, feature selection and classifier combination methods. The features are first extracted from the given sequence and then these features are input into the classifier to determine the most suitable prediction software. In this paper, the state of the art feature selection, mRMR [21], is employed to identify the important features and *SVM (support vector machine)* [1], [8] is used as the basis classifier. To further improve prediction accuracies, we propose a multi-stage classifier combination method. Instead of selecting features independently, our classifier combination method takes the output of classifiers in the previous stages into consideration. Thus, the method guides the feature selector to choose features that are most relevant to the target class label while least dependent on what has been learned by the ensemble. The experimental results shows that our tool preference selection method can improve both the classification and base-pair prediction capability.

The rest of this paper is organized as follows. In Section 2, we will give a more detailed description for the RNA secondary prediction tools used in our paper. In addition, we also briefly describe SVM, and some prediction software. We introduce our feature extraction method in Section 3. Section 4 presents the feature relevance and selection. In Section 5, we focus on how to integrate multiple classifiers. Our experimental results and conclusions are given in Sections 6 and 7, respectively.

2. Preliminaries

2.1 Support Vector Machines

Support vector machine (SVM) [6], [30] is a well-established technique for data classification. Given a training set of n-dimensional instances and label pairs $(\mathbf{x_i}, y_i)$, for $1 \leq i \leq N$ where $\mathbf{x_i} \in \mathbf{R^n}$ and $y \in \{-1, +1\}$, the SVM solves the following optimization problem:

$$\begin{aligned} \min_{\mathbf{w},b,\xi_\mathbf{i}} \quad & \tfrac{1}{2}\mathbf{w^T}\mathbf{w} + C\sum_{i=1}^{N}\xi_i, \\ subject\ to \quad & y_i(\mathbf{w^T}\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0. \end{aligned} \tag{1}$$

The function ϕ maps the training vectors $\mathbf{x}_i$ into a higher dimensional space, namely feature space. SVM finds a linear separating hyperplane with normal vector $\mathbf{w}$ and offset b that constitutes the maximal margin in the feature space. The symbols ξ_i and C represent the slack variable and the penalty of errors in the optimization problem. To describe the similarity between vectors in the feature space, the kernel function is defined. In this paper, we adopt *radial basis function* (RBF) as it yields best results.

[1]This research work was partially supported by the National Science Council of Taiwan under contract NSC-98-2221-E-110-062.
[3]Corresponding author: cbyang@cse.nsysu.edu.tw.

2.2 pknotsRG

Some researchers proposed algorithms for predicting pseudoknotted structure based on thermodynamics [18], [31]. Since predicting arbitrary pseudoknotted structures in thermodynamic way is NP-complete [25], Rivas and Eddy [27] thus took an alternative approach, which is based on the dynamic programming algorithm. Their method mainly focuses on some classes of pseudoknots and the complexity is of $O(n^6)$ in time and of $O(n^4)$ in space for the worst case. Based on Rivas's system (pknots) [27], another prediction software tool, *pknotsRG* [26], was developed. The idea is motivated by the fact that H-type pseudoknots are commonly observed in RNA sequences. Hence, by setting some proper restrictions, pknotsRG can reduce the required prediction time to $O(n^4)$ for predicting pseudoknotted structures.

2.3 NUPACK

Dirks and Pierce [10] presented an alternative algorithm which is based on the partition function. Because the partition function gives information about the melting behavior for the secondary structure under the given temperature [19], the base-pairing probabilities thus can be derived accordingly.

2.4 RNAStructure

RNAStructure was developed by Mathews *et al.* [17] and it is also based on the dynamic programming algorithm. The software incorporates chemical modification constraints into the dynamic programming algorithm and makes the algorithm to minimize free energy. Since both chemical modification constraints and free energy parameters are considered, the software works reasonably better than those that adopt only free energy minimization scheme.

3. Features

3.1 The RNA Primary Structure Feature

In this paper, we adopt the RNA primary structure features from our previous studies [7]. The feature groups are composed of the compositional factor, bi-transitional factor, distributional factor, tri-transitional factor, potential base-pairing factor, nucleotide proportional factor, sequence specific score and segmental factor. The number of features is 146.

3.2 The Spaced Bi-gram Factor

Unlike the tri-transitional factor, the *spaced bi-gram factor* [13] ignores its middle nucleotide type. Therefore, there are 16 possible combinations for the $X?Y$ pattern, where X, $Y \in$ {A, U, G, C}.

3.3 The Asymmetry of Direct Complementary Triplets

According to the Watson-Crick model, two strands of DNA form a double helix which are bonded together only between specific pairs of nucleotides. These pairs are A and T as well as G and C. In this sense, we say they are complementary[14]. The asymmetry of direct-complementary triplets (ADCT) measures the average difference numbers between mutually direct complementary triplets, XYZ and $X'Y'Z'$, in a sequence, where $X, Y, Z, X', Y', Z' \in$ {A, U, G, C}. The number of features for ADCT is 3 .

3.4 The Sequence Moment

For a 2D image I with pixel intensity $I(x, y)$, the image moment M_{ij} of order $(i + j)$ [12] is defined as

$$M_{ij} = \sum_x \sum_y x^i y^j I(x, y). \tag{2}$$

According to the definition, M_{00} is the mass of the image and the centroids are $\overline{y} = M_{01}/M_{00}$ and $\overline{x} = M_{10}/M_{00}$, respectively. The central moment, which is translational invariant, is defined as

$$\mu_{ij} = \sum_x \sum_y (x - \overline{x})^i (y - \overline{y})^j I(x, y). \tag{3}$$

To represent an image that is invariant to both scale and translation changes, the central moment should be properly divided by μ_{00}. Thus, the scale and translation invariant moment is given as follows:

$$\eta_{ij} = \frac{\mu_{ij}}{\mu_{00}^{1+\frac{i+j}{2}}}. \tag{4}$$

Since the RNA sequence S is composed of a series of nucleotides, it can be regarded as an 1D image, that is $I(k, 1)$. In order to represent the distribution of different kinds of nucleotides, the scale and translation invariant moments are calculated. An RNA sequence is first converted into the $I_X(k, 1)$ format (binary bit string), where $X \in \{A, U, G, C\}$. The element $I_X(k, 1)$ is equal to 1 if $S(k)$ is X. Otherwise, $I_X(k, 1)$ is equal to 0. After this conversion, the scale and translation invariant moment of order i for a specific nucleotide X is given as follows:

$$\eta_i(X) = \frac{\mu_i(X)}{\mu_0^{1+\frac{i}{2}}(X)}. \tag{5}$$

The above moment of order 1 denotes the centroid and it is always equal to zero due to centralization. In this paper, we calculate the moment up to order 4 for each kind of nucleotide and thus there are total 12 moments for each sequence.

3.5 The Spectral Properties

Fourier transform (FT) [22] is commonly used to explore the pattern in the frequency domain. The converted series, which contains only 0 and 1 levels, can be considered as a signal and thus also applicable to FT scenario. Let's assume $I_X(k, 1)$ is decomposed into several sinusoidal signals with $F_X(f)$ as their coefficients according to

$$F_X(f) = \sum_{k=1}^{|S|} I_X(k,1) exp(-\frac{2\pi k f i}{|S|}), \quad (6)$$

where $1 \leq f \leq |S|$. The $|F_X(f)|$ represents the magnitude of a frequency f and thus it represents the intensity for a specific spectral. The total energy E_X is defined as

$$E_X(S) = \sqrt{\sum_{f=1}^{|S|} ||F_X(f)||^2}. \quad (7)$$

The spectral entropy for a given nucleotide is:

$$H_X(S) = \sum_{f=1}^{|S|} \frac{|F_X(f)|}{E_X(S)} \log \frac{|F_X(f)|}{E_X(S)}. \quad (8)$$

The spectral inertia for a given nucleotide is:

$$J_X(S) = \sum_{f=1}^{|S|} f^2 \frac{|F_X(f)|}{E_X(S)}. \quad (9)$$

The position at which the maximal spectral energy for a given nucleotide occurs and its corresponding energy percentage is:

$$P_X(S) = arg \max_{f=1}^{|S|} |F_X(f)|/|S|. \quad (10)$$

$$M_X(S) = \max_{f=1}^{|S|} |F_X(f)|/E_X(S). \quad (11)$$

In addition to separate encoding of A, G, C, U nucleotides, the spectral properties can also be considered simultaneously. That is, A, G, C, U are encoded into $0001, 0010, 0100, 1000$ respectively. Thus, the above calculation can be applicable, but the length of the binary bit string becomes $4|S|$. The spectral entropy, inertia, maximal position and maximal energy percentage features are calculated for A, G, C, U and $ACGU$ encodings. Hence, there are 20 spectral features.

3.6 The Wavelet Features

Wavelet transform [11], [22] is a technique that decomposes a signal into several components. It is believed that the wavelet analysis provides the information that might be obscured by Fourier analysis. In this paper, the maximal overlap DWT (MODWT) is adopted and let the maximal scale level be J. Since the MODWT is an energy-preserving transform, the energy is unchanged after transformation.

$$\frac{1}{N}|\mathbf{p}|^2 = \frac{\mathbf{1}}{\mathbf{N}} \sum_{\mathbf{j=1}}^{\mathbf{J}} |\mathbf{q_j}|^2 + |\mathbf{r_{J+1}}|^2, \quad (12)$$

where p represents the original signal and each $|\mathbf{q_j}|^2/\mathbf{N}$ represents the decomposed variance at scale of j. To obtain useful information for classification, we first convert the sequence into a signal $I_X(k,1)$, where $X \in \{A, U, G, C, ACGU\}$. Then we decompose the variance of the original signal with different wavelet scales. Because all of the sequences have length greater than 16, we select maximal $J = 4$. Consequently, it totally yields 20 features.

3.7 The 2D-dynamic representation

The 2D graphical representation [2], [3] was proposed by Bielinska-Waz *et al.*, which adopts 2D graphical methods to characterize nucleotide sequences. In their study, they used nucleotides to generate a walk on the 2D graph. The walks are made as follows: A=(-1,0), G=(1,0), C=(0,1) and T=(0,-1). For RNA sequences, nucleotide T is replaced by U. After finishing the walk, a 2D-dynamic graph is generated, which can be regarded as an image. The mass of point (x,y) is determined by how many times the walk stops there. In this paper, the two principal moments of inertia, orientation, eccentricity, x and y centroids, μ_{02}, μ_{03}, μ_{11}, μ_{12}, μ_{13}, μ_{20}, μ_{21}, μ_{22}, μ_{23}, μ_{30}, μ_{31}, μ_{32} and μ_{33} of the 2D-dynamic graph descriptors are used. The number of features is 19.

3.8 The Protein Features

The genetic code [20] is the set of rules by which nucleotides is translated into proteins. These codes define mappings between tri-nucleotide sequences, called codons, and amino acids. Since three nucleotides are involved for translation, this constitutes a 4^3-versus-20 mapping to common amino acids. The first codon is defined by the initial nucleotide from which the translation process starts. There are three possible positions to start translation, each of which yields a different protein sequence. As a result, we usually say that every nucleotide sequence can be read in three reading frames.

Once a nucleotide sequence is converted into a protein sequence, its 125 PSI (protein sequence information) features [9] can be extracted accordingly. Since there are three possible ways for conversion, it yields 375 PSI features.

3.9 The Co-occurrence Factor

The co-occurrence factor [16] represents the distribution that two nucleotides occur simultaneously at a given offset within an given range in a sequence S. Let the central nucleotide at position k be X and half window size be h. The co-occurrence factor counts the occurrence of (X, Y) by:

$$C_{XY}(S) = \sum_{i=k-h, i\neq k}^{k+h} \begin{cases} 1 & \text{S(i)=Y} \\ 0 & \text{Otherwise} \end{cases} \quad (13)$$

Since the minimal length is 21 among all sequences, we set h to 10. To count for longer sequences, the co-occurrence is calculated with a sliding window scheme. There are 10 distinct co-occurrence patterns to count, which are $AA, AC, AG, AU, CC, CG, CU, GG, GU, UU$. This is because the pattern XY and YX are regarded as symmetric

Table 1: Summary of feature sets.

ID	Feature	Dimension
1	The Compositional Factor	4
2	The Bi-transitional Factor	16
3	The Distributional Factor	20
4	The Tri-transitional Factor	64
5	The Spaced Bi-gram Factor	16
6	The Potential Base-pairing Factor	3
7	The Asymmetry of Direct-Complementary Triplets	3
8	The Nucleotide Proportional Factor	12
9	The Potential Single-stranded Factor	3
10	The Sequence Specific Score	1
11	The Segmental Factor	20
12	The Sequence Moment	12
13	The Spectral Properties	20
14	The Wavelet Features	20
15	The 2D-dynamic representation	19
16	The Protein Features	375
17	The Co-occurrence Factor	10
18	The 2D graphical representation	24
	Total	642

and thus only one is considered. Consequently, 10 features are obtained.

3.10 The 2D graphical representation

The 2D graphical representation [23], [24] was proposed by Randic *et al.*, which also adopts 2D graphical methods to characterize nucleotide sequences. Two kinds of matrices are used to characterize the 2D graph quantitatively. They are M/M and L/L matrices. In Randic's search, they use the leading eigenvalues of M/M and L/L matrices to characterize nucleotide sequences. Although there are $4 \times 3 \times 2 \times 1$ possibilities, we only consider 12 cases. This is because the curve generated by an order is just a vertical flip of the other one that is generated by a reverse order. Consequently, there are 24 features for the 2D graphical representation.

The total number of features is 642, which are summarized in Table 1.

4. Feature Relevance and Selection

4.1 Feature Relevance

In information theory, *entropy* is a measurement of the uncertainty that is associated with a random variable [4]. *Mutual information* $I(X,Y)$ quantifies the dependence between the joint distribution of X and Y, and it is defined as

$$I(X,Y) \quad = \quad H(X) + H(Y) - H(X,Y), \qquad (14)$$

where $H(\cdot)$ denotes the entropy and $H(X,Y)$ is the joint entropy of X and Y. If we associate X and Y with the features and class label, mutual information can be regarded as the relevance measure between these two items.

4.2 Feature Selection

The feature relevance constitutes the basic idea for feature ranking and selection. We adopt mRMR (minimal redundancy and maximal relevance) feature selection method. Most feature selection methods select top-ranking features based on F-score or mutual information without considering relationships among features. mRMR [21] manages to accommodate both feature relevance with respect to class label and dependency among features. The strategy combines both the maximal relevance and the minimal redundancy criterion. In order to take the above two criteria into consideration and to avoid an exhaustive search, mRMR adopts an incremental search approach. That is, the rth selected feature should satisfies

$$X_r \quad = \quad \{\max_r \left[I(X_j,Y) - \frac{1}{r-1} \sum_{X_i \subset \mathbf{X_{r-1}}} I(X_j,X_i) \right] | \\ X_j \subset \mathbf{X} - \mathbf{X_{r-1}}\}. \qquad (15)$$

5. Classifier Combination

5.1 Majority Vote

The *majority vote* (MAJ) [28], [15] assigns an unknown input $\mathbf{x}$ to the most representative class among classifiers' outputs. Given each classifier's output as an m-dimensional binary vector $(d_{i,1}, d_{i,2}, \ldots, d_{i,m}) \in \{0,1\}^m, 1 \leq i \leq L$, where $d_{i,j} = 1$ if $\mathbf{x}$ is labelled as class j by the classifier i, otherwise $d_{i,j} = 0$. The majority vote picks up class c among L classifiers if

$$c = arg \max_{1 \leq j \leq m} \sum_{i=1}^{L} d_{i,j}. \qquad (16)$$

The disadvantage of the original majority vote cannot handle conflicts from classifiers of even number. In this paper, we take the idea of weighted majority vote, that is, if classifiers' predictions are not equally accurate, then we assign the more competent classifiers more power in making the final decision. We use the true positive classification rates as voting weights. This is because the original weighted majority vote, whose voting weights lie nonlinearly within $\pm\log(0.99/0.01) \times constant = \pm 4.6 \times constant$, our modified majority vote shrinks the voting weights linearly between 0 and 1. This would avoid the situation that the system is dominated by one single classifier.

5.2 Incremental Feature Selection

In this paper, we try to combine classifier outputs so that the classification accuracy can be improved. We adopt mRMR feature selection method to select features incrementally and use these features to build classifiers. Because our primary goal is to build an ensemble of high classification rate, we believe that the feature selection procedures after the first stage should take the classifier preference from the previous stages into consideration. Our strategy is to let the outputs of previous classifiers serve as preselected features so that the subsequently selected features will be as much relevant to the target label as possible while as least dependent on these outputs as possible.

For mRMR feature selection after the first stage, the feature rankings are calculated as follows:

Procedure Incremental mRMR feature selection.

Input: X: all available feature set, where $|X| = n$.
X_s: selected feature subset.
L_c: outputs of the built classifiers.
m: the number of features to select.
Output: Selected features, a subset of $X - X_s$.
Step 1: Convert the outputs of the built classifiers into numeric variables $\mathbf{X_c}$. Let the number of built classifiers be c.
Step 2: Perform mRMR feature rankings with the following formula:

$$X_r = \{\max_r [I(X_j, Y) - \frac{1}{r-1+k*c} \sum_{X_i \subset \{\mathbf{X_{r-1}} \cup \mathbf{X_c}\}} I(X_j, X_i)]| X_j \subset \{\mathbf{X} - \mathbf{X_{r-1}} - \mathbf{X_s}\}\} \quad (17)$$

where k is a weighted factor to denote how much the information of the built classifiers should be taken into consideration.
Step 3: Output $\mathbf{X_m}$=$\{X_1, X_2, \ldots, X_m\}$.

6. Experimental results

6.1 Data sets

The experimental data sets are obtained from PseudoBase and RNA SSTRAND websites. We retrieve all PseudoBase and RNA SSTRAND tRNA sequences and their secondary structure information. The original numbers are 705 and 300, respectively. The sequences are then fed into pknotsRG, RNAstructure and NUPACK for secondary structure prediction. To determine which software is the most suitable for a given sequence, we adopt the base-pair accuracy for evaluation.

Given a sequence $S = a_1 a_2 \ldots a_N$, suppose the real partner of a nucleic base a_i is a_j, where $j \neq i$, $1 \leq i \leq N$ and $0 \leq j \leq N$. If a_i is unpaired, $j = 0$. Otherwise, $1 \leq j \leq N$. Let the predicted partner of a_i be a_k. The predicted base-pair accuracy for a single sequence is:

$$Accuracy = 100\% \times \frac{1}{N} \sum_{i=1}^{N} \begin{cases} 1, & \text{j=k} \\ 0, & \text{Otherwise} \end{cases} \quad (18)$$

For each sequence, we calculate the base-pair accuracy given by the three software and assign a class label, which corresponds to the preferred software, to the sequence. The labels are nu, pk and rn, which associate with NUPACK, pknotsRG and RNAstructure, respectively. Since our goal is to apply the machine learning approach to identifying the most prominent software for prediction, we remove the sequence that any two softwares have identical highest accuracies in order to avoid ambiguity. Hence, the final numbers are 495 for RNA SSTRAND tRNA and 225 for PseudoBase database. The number of numbers in each tool preference classes given in Table 2.

6.2 Feature selection

Before applying mRMR processes, each feature variable is discretized into three states at the positions $\mu \pm \sigma$, where μ is the mean value and σ is the standard deviation. The discretized values takes -1 if it is less than $\mu - \sigma$, 1 if larger than $\mu + \sigma$, and 0 if otherwise. We used the Leave-One-Out (LOO) cross-validation method for testing. By default, mRMR selects 50 most prominent features. Hence, at each stage, we obtain 50 features to train classifiers. We combine the classifiers incrementally. The methods for classifier combination is the modified weighted majority vote (WMAJ). In order to compare the difference between feature selection with and without considering classifier outputs in previous stages, we perform both experiments. The weighted factor for the modified mRMR is set to 1.

In Table 3, the experiments combine mRMR feature selection methods and WMAJ classifier fusion methods. Two feature selection configurations are compared, which are feature selection with and without considering outputs from the previous stages. For the former one, the selected features are first used to train classifiers. Then, the data are predicted by the classifiers to produce output classes. These outputs are served as artificial features for the subsequent stages. Consequently, the following feature selection procedure will encourage unselected features, which are most relevant while least dependent on the classifier outputs. Thus, the original mRMR are modified to learn incrementally. For feature selection without considering outputs from the previous stages, once features are selected, we just exclude them and perform the original mRMR procedure in the next stage. It is observed that if feature selection procedures takes the classifier preference from the previous stages into consideration, the classification rates keep elevated. This may be due to the additional discriminant information is involved.

Table 4 shows the classification, base-pair prediction ac-

Table 2: Number of sequences in each tool preference class.

rn	nu	pk	Total
212	149	359	720

Table 3: Classification Accuracies of Various Fusion Configurations.

		WMAJ			
		1	1+2	1+2+3	1+2+3+4
a.	mRMR	68.3	68.8(+0.5)	68.2(-0.6)	67.8(-0.4)
b.	imRMR	68.3	69.3(+1.0)	69.6(+0.3)	70.2(+0.6)

imRMR: Incremental mRMR.

Table 4: The classification and base-pair prediction accuracies of various configurations.

Configuration	Features#	Classification accuracy %	Base-pair accuracy %
mRMR	50	68.3	72.9(+4.1)
imRMR+WMAJ	50 × 4	70.2	73.0(+4.2)
All	642	66.3	72.2(+3.4)

imRMR: Incremental mRMR.

curacies and numbers of selected features under different configurations. The figures behind the base-pair accuracy is the percentage above the baseline accuracy, which is achieved by the the most prominent software pknotsRG. It is shown that applying all features for prediction tool selection can achieve 72.2% base-pair accuracy, which is higher than the baseline accuracy. If the feature selection tool, mRMR, is adopted, the accuracies can be improved. Furthermore, once the feature selection tools are combined with their tailored fusion methods, the results are improved again. The best base-pair accuracy achieved is 73.0%.

7. Conclusions

In this paper, we propose a tool preference selection method, which can be used for the RNA secondary structure prediction. Our method is based on the machine learning approach. That is, the preferred tool can be determined by more than one classifier. The tool selection starts by extracting features from the RNA sequences. Then, the features are input into the classifier or ensemble to find out the most suitable tool for prediction. We adopt mRMR feature selection tools so as to identify the most discriminant features. Although these tools are proved to be powerful, it still requires users to specify the number of features to be picked up. Hence, we adopt the default settings and devise data fusion methods tailored to the feature selection. The classifiers are thus trained with selected features incrementally. The number of combinations is determined implicitly by the fusion methods, which could be the diversity of classifiers' outputs or cross-validation accuracies. The experiments reveal that our tool selection method for the RNA secondary structure prediction works reasonably well, especially combined with the up-to-date feature selection method and their 'custom-built' fusion strategies. The best achieved base-pair accuracy is 73.0%, which is significantly higher than those of any previous prediction software.

References

[1] A. Ben-Hur, D. Horn, H.T.Siegelmann, and V. Vapnik, "Support vector clustering," *Machine Learning*, vol. 2, pp. 125–137, 2001.

[2] D. Bielinska-Waz, T. Clark, P. Waz, W. Nowak, and A. Nandy, "2d-dynamic representation of dna sequences," *Chemical Physics Letters*, vol. 442, pp. 140–144, 2007.

[3] D. Bielinska-Waz, W. Nowak, P. Waz, A. Nandy, and T. Clark, "Distribution moments of 2d-graphs as descriptors of dna sequences," *Chemical Physics Letters*, vol. 443, pp. 408–413, 2007.

[4] D. J. C., *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[5] L. Cai, R. L. Malmberg, and Y. Wu, "Stochastic modeling of RNA pseudoknotted structures: a grammatical approach," *Bioinformatics*, vol. 19, pp. 66–73, 2003.

[6] C. C. Chang and C. J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

[7] C.-H. Chang, C.-B. Yang, Y.-H. Peng, and C.-Y. Hor, "Accuracy improvement for rna secondary structure prediction with svm," *Proc. of the 13th Conference on Artificial Intelligence and Applications*, pp. 526–533, 2008.

[8] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[9] C. H. Q. Ding and I. Dubchak, "Multi-class protein fold recognition using support vector machines and neural networks," *Bioinformatics*, vol. 17, pp. 349–358, 2001.

[10] R. M. Dirks and N. A. Pierce, "An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots," Wiley InterScience (www.interscience.wiley.com), Wiley Periodicals, Inc., 2004.

[11] R. Gupta, A. Mittal, and K. Singh, "A time-series-based feature extraction approach for prediction of protein structural class," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2008, 2008.

[12] M.-K. Hu, "Visual pattern recognition by moment invariants," *IRE transactions on information theory*, vol. 8, no. 2, pp. 179–187, 1962.

[13] C.-D. Huang, C.-T. Lin, and N. R. Pal, "Hierarchical learning architecture with automatic feature selection for multiclass protein fold classification," *IEEE Transaction on nanobioscience*, vol. 2, no. 4, pp. 221–232, 2003.

[14] Y. Z. Jun Wang, "Characterization and similarity analysis of dna sequences based on mutually direct-complementary triplets," *Chemical Physics Letters*, vol. 425, pp. 324–328, 2006.

[15] L. Kuncheva, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine Learning*, vol. 51, no. 2, pp. 181–207, 2003.

[16] L. Leydesdorff and L. Vaughan, "Co-occurrence matrices and their applications in information science: Extending aca to the web environment," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 12, pp. 1616–1628, 2006.

[17] D. Mathews, M. Disney, J. Childs, S. Schroeder, M. Zuker, and D. Turner, "Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of rna secondary structure," *Proceedings of the National Academy of Sciences of the United States of America*, 2004.

[18] H. Matsui, K. Sato, and Y. Sakakibara, "Pair stochastic tree adjoining grammars for aligning and predicting pseudoknot RNA structures," *Bioinformatics*, vol. 21, no. 11, pp. 2611–2617, 2005.

[19] J. S. McCaskill, "The equilibrium partition function and base pair binding probabilities for RNA secondary structure," *Biopolymers*, vol. 29, pp. 1105–1119, 1990.

[20] M. Nirenberg, P. Leder, M. Bernfield, R. Brimacombe, J. Trupin, F. Rottman, and C. O'Neal, "Rna codewords and protein synthesis, vii. on the general nature of the rna code," vol. 53, no. 5, 1965, pp. 1161–1168.

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

[22] D. B. Percival and A. T. Walden, *Wavelet Methods for Time Series Analysis*. Cambridge University Press, 2000.

[23] M. Randic, V. Marjan, N. Lers, and D. Plavsic, "Novel 2-d graphical representation of dna sequences and their numerical characterization," *Chemical Physics Letters*, vol. 368, pp. 1–6, 2003.

[24] M. Randic, M. Vracko, N. Lers, and D. Plavsic, "Analysis of similarity/dissimilarity of dna sequences based on novel 2-d graphical representation," *Chemical Physics Letters*, vol. 371, pp. 202–207, 2003.

[25] J. Reeder and R. Giegerich, "Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics," *BMC Bioinformatics*, vol. 5, pp. 104–116, 2004.

[26] J. Reeder, P. Steffen, and R. Giegerich, "pknotsRG: RNA pseudoknot folding including near-optimal structures and sliding windows," *Nucleic Acids Research*, vol. 35, pp. 1–5, 2007.

[27] E. Rivas and S. R. Eddy, "A dynamic programming algorithm for RNA structure prediction including pseudoknots," *Journal of Molecular Biology*, vol. 285, pp. 2053–2068, 1999.

[28] D. Ruta and B. Gabrys, "A theoretical analysis of the limits of majority voting errors for multiple classifier systems," *Pattern Analysis & Applications*, vol. 5, no. 4, pp. 333–350, 2002.

[29] F. Tahi, "A fast algorithm for RNA secondary structure prediction including pseudoknots," in *Proceedings of the Third IEEE Symposium on BioInformatics and BioEngineering*, Bethesda, Maryland, USA, 2003, pp. 11–17.

[30] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[31] M. Zuker, "Mfold web server for nucleic acid folding and hybridization prediction," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3406–3415, 2003.