Prediction for the Domain of RNA with the Support Vector Machine *

Chu-Kai Liu, Chang-Biau Yang†, and Chiou-Yi Hor
Department of Computer Science and Engineering
National Sun Yat-sen University, Kaohsiung, Taiwan 80424

Abstract

The three-domain system is a biological classification of RNA. In bioinformatics, predicting the domain of RNA is helpful in the research of DNA and protein. By reviewing the related literature, we notice that many researches are conducted for domain prediction with only the primary structure. However, compared with the primary structure, the secondary structure of an RNA contains more discriminative information. Therefore, we propose an SVM-based prediction algorithm that considers both the features of primary and secondary structures. In our experiment, we adopt 1606 RNA sequences from RNase P, 5S ribosomal RNA and snoRNA databases. The experimental results show that our algorithm achieves 96.39%, 95.70%, and 95.46% accuracies by combining three softwares of secondary structure prediction, pknotsRG, NUPACK, and RNAstructure, respectively. Thus, our method is a new effective approach for predicting the domain of an RNA sequence.

Keywords: RNA, SVM, three-domain system, secondary structure, bioinformatics

1 Introduction

The *three-domain system* is a biological classification proposed by Woese *et al.* in 1990 [17]. They observed the results of molecular comparisons, and then they reached a conclusion that the earth's species should be categorized into the prokaryotes (archaebacteria and eubacteria) and the eukaryotes. Therefore, Woese *et al.* developed a new taxon-domain, which is above the level of *kingdom*. The three new domains are the *Archaea* (archaebacteria), *Bacteria* (eubacteria), and *Eukarya* (eukaryotes). The relationship between the three-domain system and the traditional kingdom system is illustrated in Table 1. The merit about this classification is that many kinds of RNA are included in the three-domain system, which is shown in Figure 1 [10].

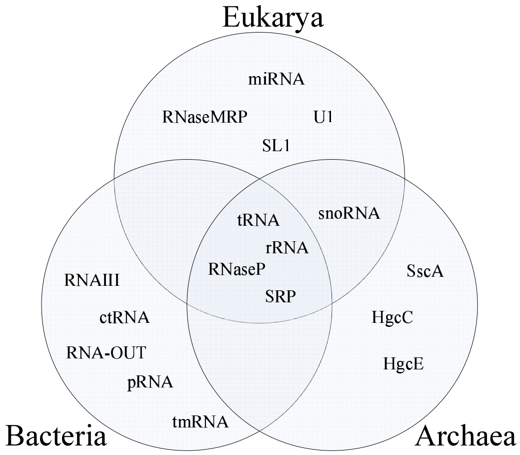

Figure 1: The RNA distribution in the three-domain system.

The secondary structure of RNA, which is constructed by hydrogen bonds, can help biologists understand the function of RNA [5]. In an RNA secondary structure, a pair of non-sequential nucleotides connected by hydrogen bonds is called a *base pair*. To form a secondary structure, there are three different base pairs A-U, G-C, and G-U.

We propose a novel approach to predict RNA domains. The experimental results show that our method is quite effective. The remaining sections of this paper are organized as follows. In Section 2, some secondary structure prediction softwares are reviewed and the machine learning tool SVM

[*] This research work was partially supported by the National Science Council of Taiwan under contract NSC 99-2221-E-110-048.

[†] Corresponding author: cbyang@cse.nsysu.edu.tw

Table 1: The relationship between the three-domain system and the kingdom system.

Domain	Archaea	Bacteria	Eukarya			
Kingdom	Archaebacteria	Eubacteria	Protista	Fungi	Plantae	Animalia

(support vector machine) is introduced. In Section 3, the proposed feature selection method is explained in detail. The experimental results are given and conclusions are drawn in Sections 4 and 5, respectively.

2 Preliminary

pknotsRG, *NUPACK* and *RNAstructure* are three powerful, well-developed, and recently widely-used software tools for predicting RNA secondary structure. We use these tools to predict the secondary structures and then the prediction results are served as features for a subsequent classification in order to increase accuracy. We introduce these tools in the following subsections.

2.1 pknotsRG

pknotsRG is a software suite for predicting RNA sequences with simple recursive pseudoknots. General pseudoknot prediction is known to be an NP-complete problem in energy-based models [12]. However, pknotsRG imposes some constraints and thus it can predict a class of simple pseudoknots in $O(n^4)$ time and $O(n^2)$ memory space [15]. There are two kinds of models available as follows:

1. **pknotsRG-mfe:** A model for computing whether the structure contains a pseudoknot or not based on the minimal free energy criterion.

2. **pknotsRG-enf:** A model manages to seek the energetically best complete structure that includes at least one pseudoknot somewhere.

The pknotsRG-mfe is more suitable for prediction of unknotted sequences. Therefore, if the sequence is dominated by unknotted structures, using pknotsRG-mfe is able to achieve a better prediction. On the other hand, the pknotsRG-enf model is more appropriate for prediction of knotted sequences [14]. We use the pknotsRG-mfe in our research.

2.2 NUPACK

The NUPACK (Nucleic Acid Package) is a software package which can predict not only RNA secondary structures but also DNA secondary structures [18]. There are three kinds of functions available as follows:

1. **Analysis:** Analyze the secondary structure of a target sequence and draw the minimum free energy (MFE) structure diagram.

2. **Design:** Design a sequence with a target secondary structure.

3. **Utilities:** Evaluate properties of a sequence, such as the number of bases and the free energy.

We use the first function, *analysis* to predict the MFE secondary structure of RNA for our research.

2.3 RNAstructure

RNAstructure is also a software tool for RNA secondary structure analysis and prediction [16]. The software has many functions for analyzing RNA sequences. We introduce some commonly used functions as follows:

1. **Fold RNA:** Predict the minimum free energy (MFE) structure of a target RNA sequence.

2. **Efn2 RNA:** Calculate the free energy of a target RNA sequence.

3. **Draw:** Draw the secondary structure diagram of a target RNA sequence.

We use the first function, *Fold RNA* to predict the MFE secondary structure of RNA in this paper.

2.4 Rfam Database

The Rfam database is a powerful database that collects a lot of RNA families [9, 10, 8]. This database uses multiple sequence alignments, secondary structures and covariance models to classify various families. It can also identify the domain of RNase P sequences.

A covariance model is a powerful tool for searching sequences in databases [6]. Existing sequence alignments can be used in covariance models to identify queried sequences. Besides, covariance models predict the secondary structure of new sequences, so that the secondary structure alignment can be performed to distinguish the family of the given sequence. We will provide more detailed explanations for this database in Section 4.

2.5 The Support Vector Machine (SVM)

A *support vector machine* (SVM) is a supervised learning method that can be used for classification or regression [1, 4]. Briefly speaking, given two groups of data elements, SVM first builds a model by some training algorithms. Then, for a given unclassified data element, SVM can predict which group the data element belong to according to the pre-trained model.

Generally, SVM uses a hyperplane to perform binary classification. In SVM, multiple classification can be achieved by combining multiple binary classifications. That is, $\frac{k(k-1)}{2}$ classifiers are constructed if k groups of data elements need to be classified.

Let the training data set be $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, $x_i \in R^d$, $y_i \in \{+1, -1\}$, for $1 \leq i \leq n$, where x_i, y_i, n, and d denote the feature vector, label, number of training samples, and dimension of the vectors, respectively. We first consider a simple problem which is *linearly separable*. That is, we can draw a line (hyperplane) $f(x) = w^T x + b$ to divide the vectors x_i's into two classes such that each sample (x_i, y_i) with $y_i = +1$ fall on one side ($f(x_i) \geq +1$), and each sample with $y_i = -1$ fall on the other side ($f(x_i) \leq -1$), as shown in Figure 2. The separating function can be formulated as follows:

$$f(x_i) = sign(w^T x_i + b) \begin{cases} \geq 0 & \text{if } y_i = +1 \\ < 0 & \text{if } y_i = -1 \end{cases} ,$$

$1 \leq i \leq n$, where $w \in R^d$ and $b \in R$.

2.6 LIBSVM

LIBSVM is a useful tool and library for SVM implementation [3]. It can facilitate the classification or regression tasks for users. In fact, the instruction manuals for LIBSVM are available on the Internet [11].

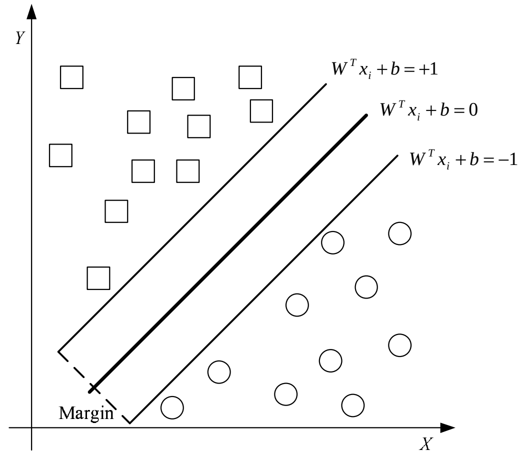

Figure 2: An illustration of a hyperplane that separates the data set into two subsets. The maximum-margin is represented by the dotted line between $w^T x_i + b = +1$ and $w^T x_i + b = -1$.

The cross-validation process is important for training data. It can avoid the *over-fitting* caused by improper training data [13]. When over-fitting occurs, the testing data would not fit the model, even if the model seems well-trained. The method of *N-fold cross-validation* is to divide the training data into N equal parts. If each part has only one training data element, then this special configuration is called the *leave-one-out cross-validation*. Different cross-validations would lead to different results. According to the study of Cawley and Talbot [2], the leave-one-out cross-validation (also called the *jackknife validation*) is more credible than general N-fold cross-validation.

3 Effective Features for the Domain Classification

In this section, we propose a method which predicts the domain of a given RNA sequence based on the three-domain system. Our method considers not only the primary structure but also the secondary structure information. Compared with the primary structure, the secondary structure of RNAs contains more discriminative information. Therefore, it is appropriate to utilize the features of the secondary structure in SVM. In the following, we explain our method, which achieves high *accuracy of classification* (AOC), and is simple and easy to implement.

In SVM, an appropriate set of features can be used to generate a good hyperplane to separate two groups of data elements. Therefore, how to extract features from RNA sequences is an important issue. Totally, six kinds of factors (each contains several features) are used in our method. These factors are *unary composition*, *binary composition*, *triplet composition*, *spaced binary composition*, *base pair frequency*, and *segmented bond frequency*. As summarized in Table 2, there are totally 113 features, which are explained in the following subsections.

3.1 The Factor of the Unary Composition

The factor of the unary composition in an RNA sequence represents the occurrence frequency of each of four bases A, C, G and U. Let the $|L|$ denote the length of sequence L, and $Num(X)$ stand for the number of base X in the given sequence. The frequency of base X is:

$$\frac{Num(X)}{|L|}, \tag{1}$$

where $X \in \{$A, C, G, U$\}$.

3.2 The Factor of the Binary Composition

The factor of the binary composition stands for the frequency of every two consecutive bases appearing in the given sequence. For example, AA, CA, AC, GU and GC are some of two consecutive bases. Therefore, we have 16 features in this factor. Let $Num(X, Y)$ denote the number of two consecutive bases X and Y. The frequency for two consecutive bases X and Y can be written as:

$$\frac{Num(X, Y)}{|L| - 1}, \tag{2}$$

where $X, Y \in \{$A, C, G, U$\}$.

3.3 The Factor of the Triplet Composition

Similar to the factor of the binary composition, the factor of the triplet composition records the frequency of every three consecutive bases in one sequence. Therefore, the triplet composition contains 64 features. Let $Num(X, Y, Z)$ denote the number of three consecutive bases X, Y and Z in the given sequence. The frequency for three consecutive bases X, Y and Z can be written as:

$$\frac{Num(X, Y, Z)}{|L| - 2}, \tag{3}$$

where $X, Y, Z \in \{$A, C, G, U$\}$.

3.4 The Factor of the Spaced Binary Composition

The factor of spaced binary composition is an extension to the binary composition, which records the frequency of each pair of the ith and the $(i+2)$th bases. Therefore, there are 16 features in this factor. Let $\hat{N}um(X, Z)$ represent the number of pairs X and Z, in which Z is the $(i+2)$th base if X is the ith base. The frequency for such X and Z can be written as:

$$\frac{\hat{N}um(X, Z)}{|L| - 2}, \tag{4}$$

where $X, Z \in \{$A, C, G, U$\}$.

3.5 The Factor of the Base Pair Frequency

The factor of the base pair frequency, which is obtained from RNA secondary structure prediction software, represents the frequency of base pairs in one sequence. Therefore, there is only one feature in this factor. Let $|L|$ represent the length of sequence L, and $Num(BP)$ represent the number of base pairs BP in the given sequence. The frequency of base pair BP is:

$$\frac{Num(BP)}{|L|}, \tag{5}$$

where $BP \in \{$A-U, G-C, G-U$\}$.

3.6 The Factor of the Segmented Bond Frequency

The factor of the segmented bond frequency means the frequency of hydrogen bonds associated with each base that occurs within each segment in the sequence. The segment is obtained by dividing a sequence into R substrings with equal length. Three different base pairs A-U, G-C, and G-U have hydrogen bonds which are considered in each segment. Let X_j denote the jth base. The frequency of hydrogen bonds of each base X is normalized by $Num(X)$ which means the number of bases X in the given sequence. The frequency of base X having hydrogen bonds in segment i is computed as follows:

We set $R = 3$ in our method. Therefore, we have $4 * 3 = 12$ features in this factor. For convenience, we let SG_1 denote the first segment and

Table 2: The six feature factors used in this thesis.

Factor	Explanation	Size
Unary composition	The occurrence frequency of each base	4
Binary composition	The occurrence frequency of the permutation of every two consecutive bases	16
Triplet composition	The occurrence frequency of the permutation of every three consecutive bases	64
Spaced binary composition	The binary composition for the ith and the $(i+2)$th bases	16
Base pair frequency	The occurrence frequency of base pairs (secondary structure)	1
Segmented bond frequency	The occurrence frequency of hydrogen bonds in each segment	12

$$SB_{Xi} = \frac{1}{Num(X)} \times \sum_{j=\frac{(i-1)|L|}{R}+1}^{\frac{(i)|L|}{R}} \begin{cases} 1 & \text{if } X_j = X \text{ and it has bonds} \\ 0 & \text{otherwise} \end{cases} ,1 \leq i \leq R, \tag{6}$$

where $X \in \{$A, C, G, U$\}$, and R represents the number of segments.

SB_{A1} represent the frequency of base A with hydrogen bond in SG_1. SG_2 represents the second segment and SB_{G2} means the frequency of base G with hydrogen bond in SG_2. Other representations can be defined similarly.

3.7 Our Method with SVM

Our method for predicting the domain of a given RNA sequence is as follows.

Step 1. Collect all primary structures (sequences) from the databases.

Step 2. Apply a software tool to predict the secondary structures from the primary structures.

Step 3. Use the method presented in Section 3.1 through Section 3.6 to extract features for the training data set.

Step 4. Use LIBSVM to train the data set produced in Step 3, and then build the trained SVM model.

Step 5. Extract the features of the target RNA for prediction.

Step 6. Predict the domain of the target RNA by using the SVM model.

4 Experimental Results

In this section, we use the statistical measurement to evaluate our experimental results, and compare them with Rfam.

4.1 Data Sets and Evaluation Criteria

To get the performance of our method, we adopt 465 RNA sequences from RNase P database, which contains 39 Archaea sequences, 366 Bacteria sequences and 60 Eukarya sequences; 1379 RNA sequences from 5S ribosomal RNA database, which contains 92 Archaea sequences, 758 Eubacteria (Bacteria) sequences and 529 Eukaryota (Eukarya) sequences; 236 RNA sequences from snoRNA database, which only contains 65 Crenarchaeotes (Archaea) sequences and 171 Euryarchaeotes (Eukarya) sequences.

However, we eliminate the homologous sequence of 5S rRNA databases and the unknown sequences of RNase P database from Rfam in our data sets. An *unknown* sequence means that Rfam cannot find out it in Rfam database. Therefore, we do not know which domain an unknown sequence is classified into. As shown in Table 3, there are 1606 RNA sequences in our data set.

In engineering statistics, the accuracy is a statistical measurement of the performance for classification tests. Since we focus on the domain prediction, the accuracy in our experiment can be written as:

$$100\% \times \frac{TD}{TD+FD}, \tag{7}$$

where *TD* (*true domain*) and *FD* (*false domain*) represent the numbers of sequences whose domains are predicted correctly and incorrectly, respectively.

Table 3: The data sets of three databases in our research.

		Three domains			Total number of each database	Average sequence length
		Archaea	Bacteria	Eukarya		
Databases	RNase P	37	352	55	444	332
	5S rRNA	61	470	395	926	119
	snoRNA	65	0	171	236	57
Total number of each domain		163	822	621	1606	169

4.2 The Accuracy of Secondary Structure Prediction

There are two methods to evaluate the accuracy of the secondary structure prediction. One is base-pair accuracy [7] and the other is bonding state accuracy. The former one is more widely adopted than the latter one.

The base-pair accuracy represents the capability that the predictor identifies real base pairs. Let $S = X_1 X_2 \ldots X_{|L|}$ be a sequence. Suppose the real partner of base X_i is X_j, where $i \neq j$, $1 \leq i \leq |L|$ and $1 \leq j \leq |L|$. If X_i is not in one base pair, let $j = 0$; otherwise, j is the position of the partner of X_i in the sequence. The same may be said, let the predicted partner of X_i be X_k. The predicted base-pair accuracy is given as:

$$100\% \times \frac{1}{|L|} \sum_{i=1}^{|L|} \begin{cases} 1 & \text{if } j = k, \\ 0 & \text{otherwise.} \end{cases} \qquad (8)$$

The bonding state accuracy represents how accurate the predictor can tell whether bases have hydrogen bonds or not. Let rh_i denote the real bonding state of base X_i. If X_i has hydrogen bonds, then $rh_i = 1$; otherwise, $rh_i = 0$. Similarly, let ph_i denote the predicted bonding state of base X_i. The predicted bonding state accuracy is given as:

$$100\% \times \frac{1}{|L|} \sum_{i=1}^{|L|} \begin{cases} 1 & \text{if } rh_i = ph_i, \\ 0 & \text{otherwise.} \end{cases} \qquad (9)$$

Since only RNase P database provides the real secondary structure information, we only show its prediction results of the three software tools in Table 4.

Table 4: The prediction accuracy of secondary structures in RNase P database with three prediction softwares.

	pknotsRG	NUPACK	RNAstructure
Base-pair accuracy	56.78%	54.51%	56.82%
Bonding state accuracy	74.03%	70.96%	72.66%

4.3 The Jackknife Test

The jackknife validation is an objective method which is more credible than the general N-fold cross-validation. We use jackknife test to evaluate the final results of our method.

In order to verify quality of the three software tools, we only use the 13 features of secondary structure explained in Section 3.5 through Section 3.6 to compare the real secondary structure of RNase P database. That is, the predicted or real secondary structures are encoded as numerical formats and served as feature vectors for SVM classification. The classification rates associated with each software tool is shown in Table 5. The model trained with the real secondary structures not only has 5% higher than those with the predicted secondary structures but also rivals the model with primary structure which contains 100 features.

Next, we combine the primary structure and secondary structure. The results are illustrated in Table 6. We perform the experiment with features, explained from Section 3.1 to Section 3.6, involved for classification. In Table 6, we show prediction results from all combinations of database sets and software tools. The database sets include RNase P, 5S rRNA and snoRNA. The three software tools include pknotsRG , RNAstructure and NUPACK. Because the average sequence length of snoRNA sequence is about 57, it may not be enough to extract statistically discriminative features. Therefore, the prediction accuracy on snoRNA database is lower than those on the other two databases. In the bottom row of Table 6, we can see that the overall accuracies of combining three databases are lower than those of individual RNase P and

Table 5: The results of jackknife experiments with both only secondary structure and only primary structure information, in which the accuracy is measured with RNase P database and three software tools for predicting RNA secondary structures.

	Primary structure only	Real secondary structure only	Three software tools		
			pknotsRG	NUPACK	RNAstructure
RNase P Database	98.42%	98.42%	93.24%	92.79%	92.79%

individual 5S rRNA databases, but higher than snoRNA database.

Although sequences used in this paper come from three different databases, we believe that they are intrinsically similar in secondary structures within each domain. Thus, in spite of being lower in overall performance, the features of secondary structure achieve a slightly better performance on 5S rRNA database. Whatever real or predicted structure is used, this may imply that the features of secondary structure indeed contain some valuable information that is not provide by the primary structure.

4.4 The Comparison with Rfam

Rfam [9, 10, 8] is a database that can also classify the domain of RNase P sequences. Our SVM model is different from Rfam in many ways. For example, the covariance model of Rfam is searching, not prediction, however, our method is to perform prediction on domain classification. The task of prediction is more difficult than search. Though the comparison is not so fair for our method, we still compare the result of our prediction method with Rfam to see how accurate our method is. The result comparison is shown in Table 7. For fair, we compare results with Rfam by eliminating the unknown sequences as explained in Section 4.1.

There are two different kinds of accuracy in Table 7. The first row is that we only show the highest accuracies of our method. The second row is the accuracies of Rfam. And then, the first column is that we ignore the unknown sequences from Rfam. The second column is that we contain all sequences of RNase P database. It is worthy to mention that our method is based on prediction rather than on searching and our accuracy is almost the same as that of Rfam.

Table 7: The comparison of accuracies of our results with Rfam.

	Without unknown sequences	All RNase P sequences
Our method	98.42%(437/444)	98.49%(458/465)
Rfam	98.87%(439/444)	94.41%(439/465)

5 Conclusion

According to the experimental results, we get a conclusion that the information of secondary structures can help us to increase the prediction accuracy of RNA domain classification. Even if users cannot provide real secondary structures, our method can alternatively use three software tools to obtain the predicted secondary structure information and then perform the subsequent domain prediction. Although the domain classification with only the predicted secondary structure information does not outperform that with only the primary structure information, however it should be noted that we only extract 13 features of secondary structure.

References

[1] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, Vol. 2, pp. 121–167, 1998.

[2] G. C. Cawley and N. L. Talbot, "Efficient leave-one-out cross-validation of kernel Fisher discriminant classifiers," *Pattern Recognition*, No. 36, pp. 2585–2592, 2003.

[3] C. C. Chang and C. J. Lin, *LIBSVM: A library for support vector machines*. National Taiwan University, No. 1, Roosevelt Rd. Sec. 4, Taipei, Taiwan 106, ROC, 2001. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

Table 6: The results of jackknife experiments, in which the accuracy is measured with three databases and three software tools for predicting the RNA secondary structure.

			Primary + three software tools			Primary + real secondary structure
		Primary	pknotsRG	NUPACK	RNAstructure	
Databases	RNase P	98.42%	98.20%	98.42%	98.42%	98.87%
	5S rRNA	98.16%	98.92%	98.60%	98.81%	N/A
	snoRNA	90.25%	87.71%	89.41%	88.56%	N/A
Combination of three databases		94.89%	96.39%	95.70%	95.45%	N/A

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

[5] A. J. Dingley and S. Grzesiek, "Direct observation of hydrogen bonds in nucleic acid base pairs by internucleotide $^2J_{NN}$ couplings," *Journal of the American Chemical Society*, Vol. 120, No. 33, pp. 8293–8297, 1998.

[6] S. R. Eddy and R. Durbin, "RNA sequence analysis using covariance models," *Nucleic Acids Research*, Vol. 22, pp. 2079–2088, 1994.

[7] X. Fu, H.Wang, W. Harrison, and R.Harrison, "RNA pseudoknot prediction using term rewriting," *International Journal of Data Mining and Bioinformatics*, 2006.

[8] P. P. Gardner, J. Daub, J. G. Tate, E. P. Nawrocki, D. L. Kolbe, S. Lindgreen, A. C. Wilkinson, R. D. Finn, S. Griffiths-Jones, S. R. Eddy, and A. Bateman, "Rfam: updates to the RNA families database," *Nucleic Acids Research*, pp. 1–5, 2008.

[9] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy, "Rfam: an RNA family database," *Nucleic Acids Research*, Vol. 31, pp. 439–441, 2003.

[10] S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S. R. Eddy, and A. Bateman, "Rfam: annotating non-coding RNAs in complete genomes," *Nucleic Acids Research*, Vol. 33, pp. 121–124, 2005.

[11] C. W. Hsu, C. C. Chang, and C. J. Lin, "A practical guide to support vector classification." http://www.csie.ntu.edu.tw/∼cjlin/papers/guide/guide.pdf, 2004.

[12] R. B. Lyngsø and C. N. Pedersen, "RNA pseudoknot prediction in energy-based models," *Journal of Computational Biology*, Vol. 7, No. 3/4, pp. 409–427, 2000.

[13] A. W. Moore, "Cross-validation for detecting and preventing overfitting." http://www.autonlab.org/tutorials/overfit10.pdf, 2007.

[14] J. Reeder and R. Giegerich, "Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics," *BMC Bioinformatics*, Vol. 5:104, pp. 1–12, 2004.

[15] 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.

[16] J. S. Reuter and D. H. Mathews, "RNAstructure: software for RNA secondary structure prediction and analysis," *Bioinformatics*, Vol. 11, pp. 1–9, 2010.

[17] C. R. Woese, O. Kandler, and M. L. Wheelis, "Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya," *Proceedings of the national academy of sciences of the united states of america*, Vol. 87, pp. 4576–4579, 1990.

[18] J. N. Zadeh, C. D. Steenberg, J. S. Bois, B. R.Wolfe, M. B. Pierce, A. R. Khan, R. M. Dirks, and N. A. Pierce, "NUPACK: Analysis and design of nucleic acid systems," *Journal of Computational Chemistry*, Vol. 32, pp. 170–173, 2011.