Academic Journal of Computing & Information Science
ISSN 2616-5775 Vol. 3, Issue 3: 43-52, DOI: 10.25236/AJCIS.030306

RNA secondary structure prediction based on long-range interaction and Support Vector Machine

Lili Jia*, Tingting Sun

Zhejiang University of Science and Technology, Hangzhou 310023, China
**Corresponding author e-mail: 118026@zust.edu.cn*

ABSTRACT. *The structure of RNA is very important in biological processes. Over the recent years, lots of machine learning method have been emerged to predict the secondary structure of RNA. In this paper, we use Support Vector Machine to predict secondary structure of RNA sequence. Meanwhile, a sequence-based method is proposed by combining a new feature representation which is based on RNA long-range interaction. We first quote E-NSSEL labels to represent the secondary structure of RNA. Combining with the definition of a new feature vector based on long-range interaction, the secondary structure of test sequence is predicted by SVM model, and the corresponding E-NSSEL sequence is consequently obtained. This sequence can be restored to secondary structure finally. The results which are obtained from RNA training and testing datasets show that this long-range-sequence-based method is superior to those method without new feature. It has higher prediction accuracy as considering the new feature. Moreover, it can predict RNA sequences with long length, which is difficult to deal with traditional folding prediction. Furthermore, it suggests that our method may provide a reliable tool for RNA secondary structure prediction, including the prediction of RNA with pseudoknots.*

KEYWORDS: *Machine learning, SVM, RNA secondary structure, long-range interaction*

1. Introduction

As the existence of an omnipresent carrier in the cell, RNA is important in many life processes. The activity of RNA molecules is determined by its secondary even third structure. Therefore, fully understanding the secondary structure of RNA molecules reveals important limitations of the molecular physical properties and functions of regulatory molecules [1-2]. Actually, the use of computational methods to predict secondary structure of RNA has been over 40 years history [3]. The typical prediction algorithms are the minimum free energy algorithm [4-7], the

ISSN 2616-5775 Vol. 3, Issue 3: 43-52, DOI: 10.25236/AJCIS.030306

Stochastic context free grammars(SCFG) [8-11] and so on. Most of the experimental predictions for base pairing of RNA sequences have achieved a more reliable method for predictions of RNA secondary structure [12-16]. However, it costs a lot of time and expense to do these experiments, especially for too long RNA sequence. Hence, the method based on machine learning provides an attractive alternative for the prediction of RNA secondary structure.

As artificial intelligence improves at anamazing speed, several attempts of machine learning method have been made to predict biological function. Meanwhile, as an important classification method of machine learning, which is called Support Vector Machine (SVM), proposed by Cortes, Vapnik and co-workers is an effective classification method [17-19]. The SVM method has been successfully applied to face recognition [20-21], prediction of time series [22-24], automatic location of video captions in extraction [25-26], RNA secondary structure prediction [27-30], etc. There are also lots of examples of classification prediction using Support Vector Machine, such as in biosynthesis, SVM is used to predict protein-protein interactions [31-32] and the Plant Root-Associated Ecological Niche of 21 Pseudomonas Species [33] and so on.

Long-range RNA structures are the interacting parts which are separated by long distances in RNAs. Throughout the tree of life, functional long-range base pairings in RNAs are widely known. There are an increasing number of reports in eukaryotic RNAs [34]. And it also contributes to human disease, including neurological disorders and other pathologies [35]. Since the long-range RNA structures are abundant and significant in RNA sequences, we define a new feature based on the long-range interaction when applying machine learning method. It will improve the prediction efficiency of secondary structure of RNA.

In this paper, we use Support Vector Machine to consider the secondary structure prediction of RNA. In addition, we define a new feature vector according to the long-range interaction in RNA, which made our prediction accuracy greatly improved. The experimental data shows that the method and the new feature are feasible and have high prediction accuracy.

ISSN 2616-5775 Vol. 3, Issue 3: 43-52, DOI: 10.25236/AJCIS.030306

2. Method and data

2.1 Support Vector Machine

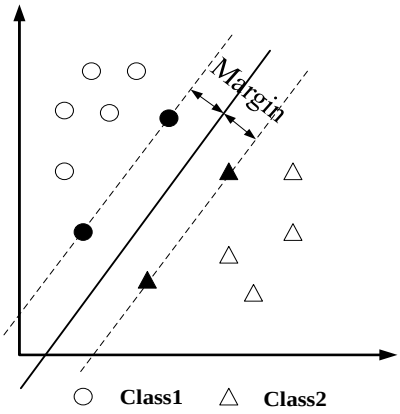

Figure. 1 The class1 and class2 are divided by Optimum Classification hyperplane.

Secondary structure predictions were generated automatically using Support Vector Machine [36]. Here we classify a pile of sequence data using SVM, especially for the two-class classification problem. As is shown in Fig. 1, This solid line represent the classification hyperplane, which is the best fit plane to separate the types of data into two categories. The criterion for "best fit" is that the distance between the line and the data on either side of the line is the largest. So our goal is to look for this hyperplane.

When we input the x_i in the equation:

$$y_i = (\omega \bullet x_i) + b$$

When $y_i > 0$ is obtained, this point belongs to the class1, conversely, this point belongs to the class2.

Support Vector Machine can also map the input vectors $\vec{X} \in R^d$ into a high dimensional feature space $\phi(\vec{X}) \in H$ and construct an optimal separating Hyperplane(OSH) [37], which maximizes the margin. The mapping is performed by a kernel function $K(x_i, x_j) = \phi(x_i) \bullet \phi(x_j)$ which defines an inner product in the space *H*.

The decision function implemented by SVM can be written as:

ISSN 2616-5775 Vol. 3, Issue 3: 43-52, DOI: 10.25236/AJCIS.030306

$f(x)=\mathrm{sgn}\{(\omega \bullet x)+b\}=\mathrm{sgn}\left\{\sum_{i=1}^{l} y_i a_i K(x_i, x_j)+b\right\}$ Where the coefficients a_i is obtained by solving the Convex Quadratic Programming problem. In the problem, extending two new regularization parameters C and Υ.

For the given dataset in this article, we choose the radial basis kernel function, the regularization parameter C=1000 and Υ=0.1 are selected for this dataset.

2.2 Data and E-NSSEL labels

RNA sequence data and the 3D-structure are obtained from PDB data bank (http://www.rcsb.org/).RNAVIEW(http://ndbserver.rutgers.edu/ndbmodule/services/download/rnaview.html) is used to get the secondary structure in detail.

The secondary structure of RNA is related with the long-range and short-range information of nucleotides. Here, we choose one nucleotide with the front and back N nucleotides as an input window. Therefore, the length of an input window is L_w=2N+1. The window lengh L_w can be changed in the prediction. Then, multiple sample inputs are obtained by moving the center position with shifting the input window at the same time. Each window corresponds to one input sample, the window with (Lw•4)bit of 0/1 encoding is used to represent the input of the window. The meaning of encoding is: each digit in the window represents a nucleotide, and a 4bit orthogonal 0/1 encoding is used to represent one of the nucleotide types (ACGU). The corresponding relationship between each nucleotide and the encoding number is:

A-1000 C-0100

G-0010 U-0001

Meanwhile, the E-NSSEL labels are used to translate RNA secondary structure into a specific way as is shown in table 1[38]. This labels can be another representation form of RNA secondary structure. Meanwile, the labels can recover the secondary structure of RNA. In the secondary structure, for a helix, there are several continuous base pairs.The helix has two sides, one side is closer to the 5' , which is called the positive side**,** the other side is closer to the 3', which is called the negative side. We use number 1 and 3 to represent these two situations respectively. For pseudoknot, the bases in positive side and the negative side is substituted as 2 and 4 respectively. But for the loop, let all the bases be 5. Thus, one RNA sequence can be written as a digit string with number 1-5. More importantly, the secondary structure is also included in the digit string. In this way, the computer can distinguish the digit and we can use the machine learning method to analyze the secondary of RNA sequence.

ISSN 2616-5775 Vol. 3, Issue 3: 43-52, DOI: 10.25236/AJCIS.030306

Table 1 The label representation method of the RNA new secondary structure elements

Name of SSE	Detail of SSE	Label of SSE
+Helix	the positive side of the helix(closer to the 5')	1
+pseudoknots	the positive side of the pseudoknot(closer to the 5')	2
-Helix	the negative side of the helix(closer to the 3')	3
-pseudoknots	the negative side of the pseudoknot(closer to the 3')	4
Loop	unpaired base	5

2.3 Long-range-sequence-based feature

Long range interaction in RNA secondary structures is significant. In this paper, considering the long range interactions, we search for the whole sequence from the beginning, and apply the RNA pairing rules (A-U, C-G) to find where there will exist base pairs.

Since one base pair is not stable in RNA secondary structure, we consider stack-based structure instead of single base pair in the searching process. The number of continuous base pairs N_{cb} is supposed. In order to find which number is more suitable for the prediction, we change the number of continuous base pairs N_{cb} from 0-5. It is considered as a new feature vector. The nucleotides in continuous base pairs are marked 1 and 2 respectively, and the remaining unmatched pairs are marked as 3, respectively. By adjusting the value of N_{cb}, the prediction accuracy of RNA secondary structure improves.

2.4 Evaluation

In this article, the accuracy evaluation we used is the precision rate, the recall rate, and the weight ratio coefficient f1-score between the precision rate and recall rate through Support Vector Machine model[39]. In the prediction of RNA secondary structure, TP is the number of base pairs which are correctly predicted; FN represents the number of base pairs which exist in the real structure but are not predicted in the result; FP indicates the number of base pairs which are not in the real structure but predicted from the model; TN indicates the number of unpaired nucleotides which are correctly predicted from the model. The calculation formulas are as follows:

$$P_{precision} = \frac{TP}{(TP + FP)}$$

$$R_{(recall)} = \frac{TP}{(TP + FN)}$$

$$f_{\beta} = (1 + \beta^2) \times \frac{precision \times recall}{\beta^2 \times precision \times recall}$$

$$\text{when } \beta = 1\text{: } f_1 = 2\frac{precision \times recall}{precision + recall}$$

ISSN 2616-5775 Vol. 3, Issue 3: 43-52, DOI: 10.25236/AJCIS.030306

Accuracy is as important as recall.

3. Results

In order to evaluate the performance of SVM classifier for RNA secondary structure prediction, we compared the results of different number of continuous base pairs N_{cb} and different window length L_w. We did quantitative experiments to find the most suitable window length L_w and number of continuous base pairs N_{cb} in the classifier.

3.1 Choose suitable L_w and N_{cb}

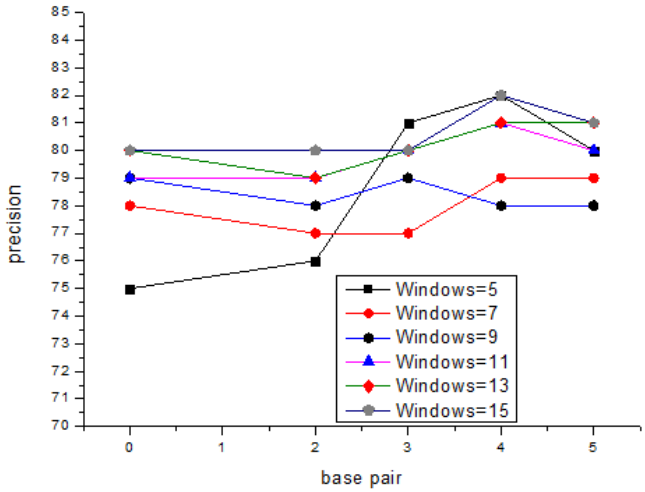

Figure. 2 The precision of RNA secondary structure by using SVM with different number of continuous base pairs (N_{cb}) and different length of windows (L_w).

In Fig. 2, when the feature of base pairs is not considered, it is found that the higher the length of window is, the higher the overall precision rate is. The predict precision when the length of windows equals to 15 is the largest. It is 80%. As there is no other feature, longer length of the window covers more information about the nucleotides. Therefore, the prediction accuracy of long window is higher than that of short window.

On the other hand, we considered the number of continuous base pairs N_{cb} as a new feature vector. It varies from 2 to 5. It is found that when the window length is 5, the precision increases first, then reaches the maximum as N_{cb}=4. The prediction precision then declines. The maximum of precision is 82%, which is larger than that of L_w=15, N_{cb}=0. It is obvious that the influence of N_{cb} to the precision for L_w=5 is

ISSN 2616-5775 Vol. 3, Issue 3: 43-52, DOI: 10.25236/AJCIS.030306

more severe than other longer L_w. The reason is that shorter L_w brings fewer information around nucleotide, so the information of base pairs will help to improve the predicting precision.

3.2 Test of RNA sequences

282 RNA sequences without pseudoknots and 37 RNA sequences with pseudoknots were taken from PDB database to be discussed in this paper. Here, we chose the length of windows L_w= 5 and the number of continuous base pairs N_{cb}= =4 to get the prediction results using SVM classifier.

Table 2 The prediction results of RNA sequences without pseudoknots

	With long-range-sequence-based feature				Without long-range-sequence-based feature			
	Precision	Recall	F1-score	Support Vector	Precision	Recall	F1-score	Support Vector
1 (+Stem)	0.67	0.69	0.68	331	0.54	0.59	0.56	37
2 (+pseudoknots)								
3 (-Stem)	0.71	0.72	0.72	348	0.57	0.49	0.52	340
4 (-pseudoknots)								
5 (Loop)	0.9	0.89	0.89	1115	0.87	0.89	0.88	1117
average/total	0.82	0.82	0.82	1794	0.75	0.75	0.75	1794

Tabel 3. The prediction results of RNA sequences with pseudo-knots

	With long-range-sequence-based feature				Without long-range-sequence-based feature			
	Precision	Recall	F1-score	Support Vector	Precision	Recall	F1-score	Support Vector
1 (+Stem)	0.69	0.53	0.6	34	0.67	0.73	0.7	33
2 (+pseudoknots)	0.83	0.71	0.77	7	0.8	0.67	0.73	6
3 (-Stem)	0.88	0.81	0.84	26	0.61	0.61	0.61	23
4 (-pseudoknots)	0.88	0.7	0.78	20	0.71	0.48	0.57	21
5 (Loop)	0.88	0.95	0.91	185	0.9	0.92	0.91	189
average/total	0.85	0.86	0.85	272	0.83	0.83	0.83	272

The Predicting precision for RNA secondary structure sequences without pseudo-knots is shown in Table 2. And Table3 presents the predicting precision. for RNA sequences with pseudo-knots.

In Table 2, when the long-range-sequence-based feature is not concerned, the average value of precision and recall are both 75%. However, when we apply long-range-sequence-based feature in the prediction, the average value of precision and recall increases to 82%. It suggests this long-range-sequence-based feature is significant in prediction of RNA secondary structure.

ISSN 2616-5775 Vol. 3, Issue 3: 43-52, DOI: 10.25236/AJCIS.030306

In Table 3, when there is no other features except E-NSSEL, the average value of precision and recall are both only 83%, and the average value of precision with long-range-sequence-based feature is 85%, and the average recall is 86%.

It shows that the influence extent of long-range-sequence-value feature to RNA sequences without pseudo-knots is larger than to RNA with pseudo-knots.

4. Conclusion

We proposed a new feature that based on long-range interaction of nucleotides for predicting RNA secondary structure. We first observed that for window length L_s=5 and number of continuous base pairs N_{cb}=4, the prediction has the highest precision comparing with other conditions. Then, under this condition, we compared the prediction results of RNA sequences with the new feature and without the new feature. From the investigation of RNA sequences without pseudo-knots, the prediction precision applying new feature is higher than those which don't consider new feature. However, in the research of RNA sequences with pseudoknots, the difference of the precisions with new feature and without new feature is not obvious in contrast to those RNA with pseudoknots.

We believe that the support vector classifiers that combine long-range-sequence-based new feature represent a powerful methodology that will form the basis for many future RNA secondary structure prediction approaches

Of wireless sensor networks, moving target tracking based on wireless sensor networks also has broad application prospects.

Acknowledgements

This research was financially supported by the National Natural Science Foundation of China (Grant nos. 20904047, 21673207, 21873087) and the Natural Science Foundation of Zhejiang Province (Grant nos. LY17A040001).

References

[1] Zuker M, Mathews D H, Turner D H. Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide [M] // RNA Biochemistry and Biotechnology. Springer Netherlands, 1999: 11-43.

[2] Ding Y, Lawrence C E. A statistical sampling algorithm for RNA secondary structure prediction. [J]. Nucleic Acids Research, 2003, 31 (24): 7280.

[3] Jing-Yuan H E. The Model Research of Support Vector Machines in the RNA Secondary Structure Prediction [J]. Computer Science, 2008, 35(4):181-183.

[4] Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information [J]. Nucleic Acids Research, 1981, 9 (1): 133-148.

ISSN 2616-5775 Vol. 3, Issue 3: 43-52, DOI: 10.25236/AJCIS.030306

[5] Sakakibara Y, Brown M, Hughey R, et al. Stochastic context-free grammars for tRNA modeling. [J]. Nucleic acids research, 1994, 22 (23): 5112.
[6] Rivas E, Eddy S R. A dynamic programming algorithm for RNA structure prediction including pseudoknots. [J]. Journal of Molecular Biology, 1999, 285 (5): 2053-2068.
[7] Zuker M. Calculating nucleic acid secondary structure [J]. Current Opinion in Structural Biology, 2000, 10 (3): 303-310.
[8] Horesh Y, Doniger T, Michaeli S, et al. RNAspa: a shortest path approach for comparative prediction of the secondary structure of ncRNA molecules. [J]. Bmc Bioinformatics, 2007, 8 (1): 366.
[9] Andrews M W. Stochastic Context-Free Grammars [J]. 2004.
[10] Knudsen B, Hein J. Pfold: RNA secondary structure prediction using stochastic context-free grammars [J]. Nucleic Acids Research, 2003, 31 (13): 3423-3428.
[11] Knudsen B, Hein J. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. [J]. Bioinformatics, 1999, 15 (6): 446-454.
[12] Searls D B. Linguistic approaches to biological sequences [J]. Computer Applications in the Biosciences Cabios, 1997, 13 (4): 333.
[13] James B D, Olsen G J, Pace N R. Phylogenetic comparative analysis of RNA secondary structure [J]. Methods in Enzymology, 1989, 180 (1): 227.
[14] Winker S, Overbeek R, Woese C R, et al. Structure detection through automated covariance search [J]. Computer Applications in the Biosciences Cabios, 1990, 6 (4): 365-371.
[15] Eddy S R, Durbin R. RNA sequence analysis using covariance models [J]. Nucleic Acids Research, 1994, 22 (11): 2079-88.
[16] Dowell R D, Eddy S R. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. [J]. Bmc Bioinformatics, 2004, 5 (1): 71.
[17] Cortes C, Vapnik V. Support-vector networks [C] // Machine Learning. 1995: 273-297.
[18] Chang C C, Lin C J. LIBSVM: A library for support vector machines [M]. ACM, 2011.
[19] Hua S, Sun Z. Support vector machine approach for protein subcellular localization prediction [J]. Bioinformatics, 2001, 17 (8): 721.
[20] Osuna E, Freund R, Girosi F. Training Support Vector Machines: an Application to Face Detection [C] // Computer Vision and Pattern Recognition, 1997. Proceedings. 1997 IEEE Computer Society Conference on. IEEE, 2002: 130-136.
[21] Zhifeng Li, Xiaoou Tang. Bayesian face recognition using support vector machine and face clustering [J]. 2004, 2: II-374-II-380 Vol.2.
[22] Rubio G, Pomares H, Rojas I, et al. A heuristic method for parameter selection in LS-SVM: Application to time series prediction [J]. International Journal of Forecasting, 2011, 27 (3): 725-739.
[23] Espinoza M, Suykens J A K, Moor B D. Short Term Chaotic Time Series Prediction using Symmetric LS-SVM Regression [J]. Proc. of the 2005

ISSN 2616-5775 Vol. 3, Issue 3: 43-52, DOI: 10.25236/AJCIS.030306

International Symposium on Nonlinear Theory and Applications (NOLTA) pages:606-609, 2005: 606-609.
[24] Lendasse A. Fast bootstrap applied to LS-SVM for long term prediction of time series [J]. 2004, 1: 705--710.
[25] Zhuang Y. Automatic Caption Location and Extraction in Digital Video Based on Support Vector Machine [J]. Journal of Computer Aided Design & Computer Graphics, 2002, 14 (8): 750-749.
[26] Liu J W, Guo Z J, Fei W U, et al. Automatic Caption Location and Extraction in Digital Video Frame Based on SVM and ICA [J]. Journal of Image & Graphics, 2003, 8 (11): 1334-1340.
[27] Zhao Y, Wang Z. Consensus RNA Secondary Structure Prediction Based on Support Vector Machine Classification [J]. Chinese Journal of Biotechnology, 2008, 24 (7): 1140-1148.
[28] Chen Z, Hong W, Wang C. RNA secondary stucture prediction with plane pseudoknots based on support vector machine [J]. Icic Express Letters, 2009, 3 (4): 1411-1416.
[29] He J, He Z, Zou D. The research of RNA secondary structure prediction based on extended NSSEL labels [C] // Intelligent Control and Automation, 2008. Wcica 2008. World Congress on. IEEE, 2008: 5396-5400.
[30] Oliveira J V D A, Costa F, Backofen R, et al. SnoReport 2.0: new features and a refined Support Vector Machine to improve snoRNA identification [J]. Bmc Bioinformatics, series [J]. 2004, 1: 705--710.
[31] Guo Y, Yu L, Wen Z, et al. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. [J]. Nucleic Acids Research, 2008, 36 (9): 3025-30.
[32] You Z H, Yu J Z, Zhu L, et al. A MapReduce based parallel SVM for large-scale predicting protein–protein interactions [J]. Neurocomputing, 2014, 145 (18): 37-43.
[33] Chien J, Larsen P. Predicting the Plant Root-Associated Ecological Niche of 21 Pseudomonas Species Using Machine Learning and Metabolic Modeling [J]. 2017.
[34] Pervouchine, D.D., Khrameeva, E.E., Pichugina, M.Y., Nikolaienko, O.V., Gelfand, M.S., Rubtsov, P.M.Mironov, A.A. Evidence for widespread association of mammalian splicing and conserved long-range RNA structures. RNA 2012, 18, 1–15.
[35] Bernat, V.; Disney, M.D. RNA Structures as mediators of neurological diseases and as drug targets. Neuron. 2015, 87, 28–46.
[36] Ukil A. Support Vector Machine [J]. Computer Science, 2002, 1 (4):1-28.
[37] Garg P, Sharma V, Chaudhari P, et al. SubCellProt: predicting protein subcellular localization using machine learning approaches [J]. Silico Biol, 2009, 9 (1-2): 35-44
[38] Saeys Y, Inza I, Larrañaga P. WLD: review of feature selection techniques in bioinformatics [J]. Bioinformatics, 2007, 23 (19): 2507-2517.
[39] Gardner P P, Giegerich R. A comprehensive comparison of comparative RNA structure prediction approaches [J]. Bmc Bioinformatics, 2004, 5 (1): 1-18.