

Combining Stochastic Grammar and Semi-Supervised Learning Techniques to Extract RNA Structures with Pseudoknots

Sixin Tang^{*}

College of Computer Science and Technology, Hengyang Normal University, Hengyang, 421002, China

Abstract

To predict RNA structures with pseudoknots, traditional stochastic grammar models must collect several related labeled RNA sequences, which limits the practical application of this method. In order to use a large number of unlabeled RNA sequences effectively for structure prediction, the combination of stochastic grammar and semi-supervised learning techniques has been proposed. In these techniques, we used a small amount of labeled RNA sequences and a large number of unlabeled sequences as a training set of the prediction model. Designing a semi-supervised learning model based on the SCFG inside/outside algorithm and using a SCFG model based on the generative method as a classifier, we labeled the unlabeled RNA sequences through training and then gradually merged them into the labeled data set. This model can regulate the proportion of labeled and unlabeled sequences and finally output the structure tags sequence. Experimental results showed that this method can utilize unlabeled sequences data effectively, greatly reduce the demand for the number of related sequence samples, and improve the prediction accuracy. In addition, we measured the performance of model prediction influenced by different amounts of unlabeled sequences.

Keywords: RNA structures; semi-supervised learning; prediction accuracy; performance improvement

(Submitted on March 13, 2019; Revised on May 15, 2019; Accepted on June 15, 2019)

© 2019 Totem Publisher, Inc. All rights reserved.

1. Introduction

In recent years, an exponential growth of sequence data in RNA databases has been observed by biologists. This can be attributed to the idea of "structures decide functions", which promotes the study of RNA structures. However, the experimental methods to predict RNA structures are time-consuming and laborious; thus, studies on RNA structure prediction based on computational methods have become more important [1].

Conceptually, the models for RNA structure prediction can be divided into two parts. The first is the development of a probabilistic model for the distribution of RNA structures: $p(y | x; w)$, where x is an RNA sequence, y is a structure unit of x , and w indicates the parameter(s) of the model. These probabilistic models include the McCaskill model (McCaskill, 1990), stochastic context-free grammars (SCFGs), and hierarchical Dirichlet processes for SCFG (HDP-SCFG) (Dowell and Eddy, 2004). Conditional random fields (CRFs) RNA structure prediction is a combinational optimization problem of bases in the discrete space [2]. The folding form of single-stranded RNA depends primarily on the interaction between its constituent nucleotides, and it is also subjected to its solution environment. RNA secondary structure prediction based on computational methods first requires the calculation model of the RNA structure.

Traditional stochastic grammar models, such as the stochastic context-free grammar model (SCFG), use the comparative sequence analysis method to predict the secondary structure of RNA [3-4]. They usually require training of unknown structure sequences to the sample sets of RNA sequences with known structure (usually homologous sequences with sequences to be sequenced), and they must obtain the transfer probability of each grammar method production rule [5]. Through training, an evolutionary tree is obtained, which maximizes the probability of comparison. By using evolutionary information and stochastic context-free grammar, the generation rule that maximizes the probability of each sequence can

^{*} Corresponding author.

E-mail address: tangsix@qq.com

be found, and the corresponding secondary structure of the sequences can be deduced. It is difficult to add new features to this traditional model because it is totally dependent on the production process, causing the scalability to be worse [6]. In addition, due to the huge search space, this model mostly adopts the method of cascading search. First, the optimal solution or approximate solution is obtained under relatively simple model conditions, and then the approximate solution of the more complex model is searched in the neighborhood of the optimal solution or approximate solution. This kind of cascading search process not only is complex but also suffers from the accumulation of errors between different levels and the confusion of local optimal solutions. There are certain errors in the search, which influence the final results to a certain extent [7].

In this study, by introducing the semi-supervised model into the stochastic grammar model, a machine learning algorithm suitable for the lexicalized stochastic grammar model was proposed. The semi-supervised model was used to pre-search the lemma, reducing the target set for comparative sequence analysis and thus quickening the model search speed and improving the prediction accuracy. This helped address one of the key problems that prevented the original stochastic grammar model from being widely used [8].

Since context-free grammar is suitable for modeling the long-range correlated "ring" formed by the unpaired base region in the "stem" and stem middle of the nesting paired RNA structure, compiling the grammar form of the RNA base pairing rules can estimate which bases may form a "stem" or "ring." However, there are only four kinds of bases in an RNA sequence, while there are six basic pairing generation rules (refer to Watson-Crick pairing and G-U pairing), so there is too much ambiguity in pairing bases by using context-free grammar generation rules. In order to avoid such problems, the existing RNA sequence with known secondary structure homologous to the tested sequence, which has unknown secondary structure, is trained to obtain the using probability of each grammar generation rule. Then, the generating probability in each column of this group of RNA sequences is calculated, and the probabilities of the columns are multiplied to get the compared probabilities. An evolutionary tree is generated from the training, maximizing the compared probability, and then the evolutionary information and random context-free grammar are used to find the maximum probability generation rule of each sequence, i.e., the corresponding secondary structure of the sequence.

The traditional SCFG model is a typical supervised learning prediction model that requires the RNA sequence sample set of secondary structures. It is necessary to find the relevant sequence samples of sufficient reliable known structures. However, the existing RNA structure data set cannot meet the needs of training or evaluation of such complex models in many cases [9]. Moreover, the process of finding relevant sequence samples usually requires the participation of an expert, which is time-consuming and expensive. At present, in the RNA sequence database, the data volume of labeled structures is small and difficult to collect, but the unlabeled sequence data is abundant and easily obtained [10]. Therefore, learning from labeled samples and making full use of a large number of unlabeled sequences will greatly save computation and time costs (because the training data set required for the initial model can be significantly reduced without collecting a large number of related sequences) and improve the performance of the prediction model.

2. Methods

Semi-supervised learning is a new idea of machine learning that was proposed in recent years [11]. In this kind of machine learning, the sample set is composed of a mixture of some labeled data and mostly unlabeled data. Because a large number of unlabeled data can provide more information about joint probability distribution, adding a large number of unlabeled samples can improve the accuracy of classification prediction.

The semi-supervised learning algorithm is an algorithm that uses the mixture of labeled data and unlabeled data. Its basic idea is to use some assumptions of distribution to predict the tags of unlabeled data and then merge them into labeled data sets, so as to train new classifiers to achieve better results. Since semi-supervised learning has the characteristics of both supervised learning and unsupervised learning, statistical learning methods can still be used for unified modeling and analysis.

2.1. Identification and Division of a Lemma

In this study, we extend a semi-supervised learning approach based on stochastic grammars. The model is a hybrid of generative and discriminative models. A generative model defines the joint distribution $p(x, y)$ of an RNA sequence x and its secondary structure y , and a discriminative model defines the conditional distribution $p(y | x)$.

In the stochastic grammar of RNA secondary structure, terminal symbols refer to the character set consisting of four bases $\{a, u, g, c\}$. Production rules $P =$

$S \rightarrow aSu / USA / cSg / gSc / uSg / gSu$ (paired generation rules)
 $S \rightarrow aS / cS / gS / uS$ (generating unpaired bases on the left)
 $S \rightarrow Sa / Sc / Sg / Su$ (generating unpaired bases on the right)
 $S \rightarrow SS$ (recursive palindrome: another stem protruded on the side of the stem)
 $S \rightarrow \varepsilon$ (end)

Compared with natural language, the "word" in RNA sequence is unknown. Irrespective of any semantic information, the word can be considered as some continuous or connected basic unit or pattern, i.e., a substructure of RNA secondary structure. This study divides the possible secondary structures into three kinds: a spiral stem structure, represented by H ; a ring structure formed by many bases, represented by E ; and an isolated unpaired single-strand, represented by U . Since there are unpaired bases between two paired single-strand of the stem structure, the lemma of RNA secondary structure should be composed by nucleotides and "." Here, "." represents any kind of nucleotide, e.g., the secondary structure lemma can be $GG.CC$, $AU.AU$, or $CAUCA$. In this paper, we use "lemma" rather than "word", because the "semantic" and "syntactic" information of these lemma is not clear. They are only the modes that occur under certain conditions, such as the number of occurrences and the maximum of the mode, with empirical composition.

The above-mentioned production rules are the basic syntactic rules of RNA secondary structure [12], and the lexicalized stochastic grammar model introduces recognition rules. Its prediction of RNA secondary structure is similar to processing unknown language, which uses bottom-up analysis to deal with such language problems when only the basic component units (bases) are known and there is no knowledge of the language. The word grid model [7] is used to identify common lemmas, and the grammar rules that are fitted to the lemma rules are established according to these lemmas.

Assume $R = (r_1, r_2, \dots, r_n)$ is an RNA sequence containing n bases. For a given RNA secondary structure dictionary D (D was pre-defined through prior information or expert knowledge), R could be divided into lemma sequence $S = (S_1, S_2, \dots, S_m)$, in which S_i belongs to D and $(r_j, r_{j+1}, \dots, r_k)$ matches with S_i . Then, the word grid can be used to describe all possible divided lemmas. Word grid is a directed acyclic graph containing all possible divisions of an RNA sequence, and each route from the start node to the end node represents a possible dividing way [8]. Assume $T = (T_1, T_2, \dots, T_m)$ is the corresponding secondary structure of lemma sequence S , in which $T_j \in \{H, E, U\}$, and for each RNA sequence, (S, T) is generally not the only one. Therefore, the identification of the lemma of an RNA sequence involves finding the most probable (S, T) , which can be defined as:

$$\Gamma(R) = P(S, T | R) \quad (1)$$

According to Bayes formula and omitting the constant term, we get:

$$\Gamma(R) = \arg \max_{(S, T)} P(R | S, T) P(S, T) = \arg \max_{(S, T)} P(S, T) \quad (2)$$

Here, $\Gamma(R)$ represents the candidate parse tree. Since lemma sequence S and RNA sequence R are approximately equal, $P(R | S, T)$ can be approximated as 1. Equation (2) assumes that the occurrence of each type of secondary structure depends only on the current lemma. However, a single lemma cannot provide enough information to determine its secondary structure state. In order to make accurate predictions, enough information must be extracted. The stochastic grammar cloud model uses the cloud model to predict the state of the secondary structure of a lemma. It can introduce a variety of context information to reason and get a corrected grammar model integrated with the lemma information. Finally, the cloud droplet sampling algorithm is used to search the optimal path, and the terminal node with maximum probability is selected. Then, through backtracking, the optimal lemma sequence can be found.

2.2. Combining Stochastic Grammar and Semi-Supervised Learning

To use semi-supervised techniques to predict RNA structures, given a set of RNA sequences with known structures, we define $D^l = \{x_1, \dots, x_m\}$. Meanwhile, for a set of RNA sequences with unknown structures, we define $D^u = \{x_{n+1}, \dots, x_n\}$. If there are no identical samples in D^l and D^u , then $D^l \cap D^u = \Phi$. They are mixed together to form the training set for the model.

Let $X = \{x_1, x_2, \dots, x_m\}$ be a set of sequence samples, and each x_i is a sequence consisting of L bases. Let $Y = \{y_1, y_2, \dots, y_n\}$ be the set of structural samples of X , so y_i can be regarded as the class label of x_i . If there are undetermined elements in Y , that is, only parts of samples have been identified with the structure tag, and the tag of other samples is still to be

determined, then the goal of semi-supervised learning is to learn the rules of structure sample data Y from sequence set X . We can build a predictive model $p_\theta(y | x)$, where the conditional probability $p(y | x)$ represents the structural hypothesis of sequence x , and θ denotes the parameter set of the prediction model. The prediction model uses the joint probability density $p(x, y)$ with parameters and assumes that $p(x, y) = p(x)p(y | x)$, where $p(y | x)$ is an identifiable mixed distribution such as the Gaussian distribution. Thus, it is easy to see that $p(x)$ can affect $p(y | x)$.

Let $\Sigma = \{A, C, G, U\}$ represent a collection of terminators (bases) of the RNA sequence. Let x denote an RNA sequence of unknown structure, and then the sequence with length L is denoted as $x \in \Sigma_L$. x_i represents the i^{th} character of sequence x ($i = 1$ to L). x_{ij} represents the substring from x_i to x_j on sequence x .

Given an RNA sequence $x = cuuag$, its secondary structure can be denoted as $y = ((\cdot))$. According to the rule of base pairing of an RNA sequence, the only analytic tree of corresponding structure y is:

$$S \rightarrow cSg \rightarrow cuSag \rightarrow cuuSaa \rightarrow cuuag$$

The joint probability of the analytic tree of sequence x produced by SCFG model is:

$$p(x, \sigma) = p(S \rightarrow uSa) \times p(S \rightarrow cSg) \times p(S \rightarrow uS) \times p(S \rightarrow \varepsilon) \quad (3)$$

The conditional probability is as follows:

$$P(y | x) = \sum_{\sigma \in y} P(\sigma | x) = \frac{\sum_{\sigma \in y} P(x, \sigma)}{\sum_{\sigma' \in \Omega(x)} P(x, \sigma)} \quad (4)$$

Where $\Omega(x)$ represents all possible parse trees for sequence x .

One of the advantages of using the SCFG model is that, as a language model that describes the secondary structure of RNA, there are already many easy parameter estimation algorithms that are easy to implement. Given a data set $D = \{(x(1), y(1)), \dots, (x(m), y(m))\}$ and secondary structure $y(i)$, which has been experimentally tested, and corresponding to m RNA sequences $x(i)$, the tasks of training are to find the parameter set $\theta = \{p_1, \dots, p_n\}$ (i.e., the transition probability of each rule) and maximize the objective function value of the specified model.

The maximum likelihood method can be used to solve this problem. This method assumes that the maximum likelihood of combination between training sequences and their structures is:

$$l_{ML}(\theta | D) = \prod_{i=1}^m P(x^{(i)}, y^{(i)}; \theta) \quad (5)$$

The objective of the maximum likelihood method is to find and satisfy the following objectives:

$$\theta^* = \arg \max_{\theta} P(\theta | D) = \arg \max_{\theta} \{P(D | \theta)P(\theta)\} \quad (6)$$

There is always a solution to this constrained optimization problem for context: independent grammar without ambiguity.

Next, assuming that the likelihood of all data has been optimized, using the EM algorithm [13], the SCFG model based on the production method is taken as the classifier. The probability that unlabeled samples belong to a certain category is regarded as a set of missing parameters. As a result, supervised data and unsupervised data can be mixed together using Equation (5), where $f\theta(x)$ represents unmarked sequence data whose likelihood has been optimized.

$$L(\theta | D^u, D^l) = \sum_{x \in D^u} \log P(x, f_\theta(x) | \theta) + \sum_{(x, y) \in D^l} \log p(x, y | \theta) \quad (7)$$

2.3. The Structure Labels of RNA Structures

In order to integrate the semi-supervised learning methods into the production based stochastic grammar model, we first

define and represent the secondary structure of RNA with structural tags, construct the overall architecture of the semi-supervised learning-based stochastic grammar model, and design the prediction algorithm and steps.

RNA secondary structure is a stem-loop structure, that is, the paired bases are stacked together to form the stem (stacking pairs and helices), the unpaired bases form rings or unpaired single chains, and the rings can be divided into hairpin rings, internal rings, protruding rings, and multi-branched rings [13]. In order to classify RNA secondary structure by machine learning, the secondary structure features of RNA must be defined and described. In this study, the secondary structural features of RNA are divided into seven categories: hairpin ring, inner ring, protruding ring, stem, multi-branch ring, unpaired single chain, and surface pseudoknot.

A stem of RNA is made up of two sub-sequences. One of the sub-sequences in the sequence is closer to the 5', which is called a positive stem, and the other sub-sequence is closer to 3', which is called a negative stem. Thus, we can use the following definitions of the bases of various structures: the letter t is the base of a positive stem, the letter p stands for the base of a plane pseudoknot that is closer to 5', the letter f denotes the base in the negative stem, the letter s is the base of the plane pseudoknot that is closer to 3', the letter n denotes an unpaired base in a ring, and the letter x denotes the base of an unpaired single chain.

For example, for the following RNA sequence, the structure tag is shown in Figure 1.

5'--TTGGAACCAACAUGGAUUCAUGCUUCGGCCUGGUCGCG--3'

The recovery process is as follows: for a tag sequence with known structure, as long as the labels "t" and "p" are pressed respectively when traversing the sequence, the later "f" and "s" will necessarily match the "t" or "p" at the top of the stack. After the formation of a base pair, the corresponding label will pop up in the stack and continue to match the next base pair, so as to restore the secondary structure of RNA.

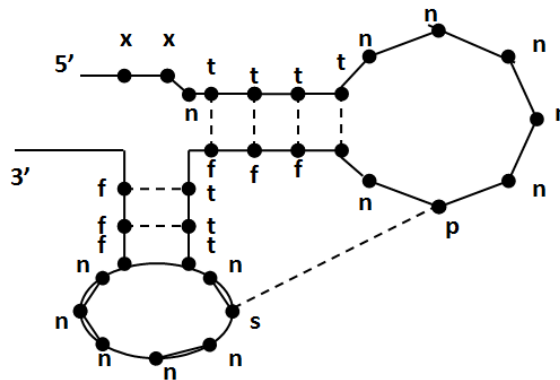


Figure 1. The structure label for RNA secondary structure

2.4. Predictive Models and Algorithmic Steps

In order to establish the initial model of semi-supervised learning, we first use the lightweight SCFG model to estimate an initial grammar G_0 , train the labeled sequence sample set, evaluate its parameters through the sequence to be sequenced, and obtain an initial classifier, i.e., the initial classification model λ_0 . Then, we use λ_0 to predict the secondary structure of the number of U unlabeled sequences.

The way to achieve this is to find an optimal marker sequence Y_i for the existing observation sequence X_i , so as to maximize the conditional probability $p(y | x)$, namely the largest conditional probability $p(Y_i | X_i, \lambda_0)$. The sequence with the highest joint likelihood will be merged into the marked sequence set, through the lexical SCFG model analysis and extraction of feature rules.

The overall design of the prediction model is shown in Figure 2.

The specific method is as follows:

(1) Establish the training sample set. According to the secondary structure representation method in Section 3.1, the structure of labeled sample RNA sequences is represented, and these sequences are taken as training sets.

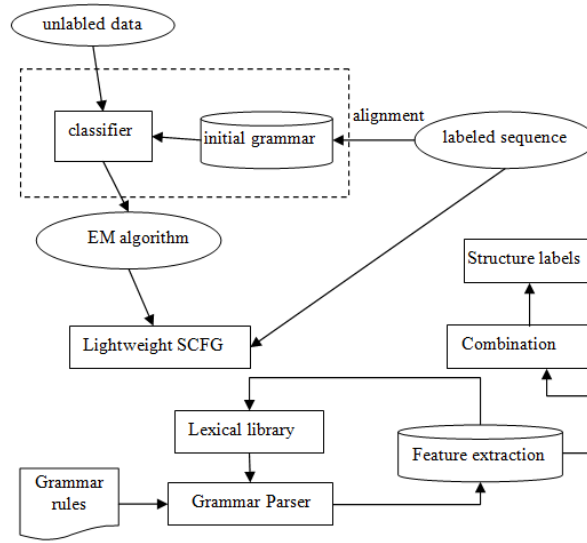


Figure 2. Overall design of prediction model

(2) Construct and solve optimization problems:

$$\sigma(x) = \arg \max_{\sigma \in \Omega(x)} P(\sigma | x, f_{\lambda}(x)) \quad (8)$$

(3) Carry out feature extraction through sequence comparison, and conduct lexical analysis and syntactic analysis for the extracted features. Then, the matching degree of the extracted features is judged. If the matching degree is greater than the given threshold value, the stem combination rule is output, and then the rules in the syntactic parser are updated according to the extracted secondary structural features.

(4) Predict the classification of the predicted samples. First, according to the base classification for paired bases and unpaired bases, the prediction samples are coded according to the same coding method as the training samples, and the input vector of the stochastic grammar model is converted.

3. Results and Analysis

To test the validity of the semi-supervised learning model of this study, we conduct a series of cross-verification experiments. We take a series of non-coding RNA sequences with known mutual secondary structures in the Rfam database [14]. (Rfam v9.1 contains seeds for multiple sequence comparisons of 1,372 non-coding RNA families, all of which are derived either from the results predicted using the covariant model or from published research literature). For each series of families, using 61 sequences, 2,758 bases are used as the training sample, 30 sequences are extracted from it, and 587 bases are tested as test samples. Among them, 1,132 sequences are selected as unlabeled sequences.

For now, the measurement of the accuracy of the learning model is usually measured in terms of the sensitivity S_n (sensitivity), its own ability, and the f-value. For the RNA secondary structure prediction model, TP can be used to represent the number of base pairs that are correctly predicted. FN is a real structure, but it does not contain the number of base pairs that are correctly predicted. FP represents the number of base pairs that do not exist in the real structure but are wrongly predicted to exist. TN represents the number of correctly predicted base pairs. The specific calculation formula is as follows:

$$S_n = \frac{TP}{TP + FN} \quad (9)$$

$$S_p = \frac{TP}{TP + FP} \quad (10)$$

$$F\text{-value} = 2 \times \frac{S_n \times S_p}{S_n + S_p} \quad (11)$$

The experiment in this study includes two aspects:

(1) In order to evaluate the effectiveness of the prediction model, the proposed semi-supervised learning model is compared with the traditional supervised learning model based on SCFG, and the performance of the semi-supervised learning model proposed in this paper is evaluated. Table 1 lists the test results of the traditional supervised learning model. Tables 2 and 3 are the test results of semi-supervised learning after adding 100 and 800 unlabeled RNA sequences, respectively (note: N represents the number of labeled RNA sequences).

Table 1. Benchmark results: prediction performance of supervised model

N	S_n	S_p	$F\text{-value}$
20	0.5553	0.6474	0.5527
100	0.5634	0.6247	0.5809
200	0.6213	0.6112	0.5657

Table 2. Prediction performance of models that have mixed with 100 unlabeled RNA sequences

N	S_n	S_p	$F\text{-value}$
20	0.5729	0.6163	0.5604
100	0.6086	0.6187	0.5882
200	0.6294	0.6198	0.5995

Table 3. Prediction performance of models that have mixed with 800 unlabeled RNA sequences

N	S_n	S_p	$F\text{-value}$
20	0.6049	0.6189	0.5879
100	0.6270	0.6195	0.6116
200	0.6394	0.6248	0.6174

It can be seen from the results that all the performance parameters of the benchmark prediction model have been improved, especially the sensitivity S_n .

(2) To evaluate the performance of the semi-supervised learning model in detail and the impact of the number of unlabeled sequences on the prediction performance, we adjust the labeled and unlabeled sequences of proportion. 122 are tagged RNA sequences as the training data set, and 20, 40, and 60 are randomly selected as three different tagged sequences of the training data set. Then, 100, 200, 400, and 800 sequences are randomly selected from 1,132 unlabeled RNA sequences as different unlabeled data sets. The F-values are shown in Figure 3.

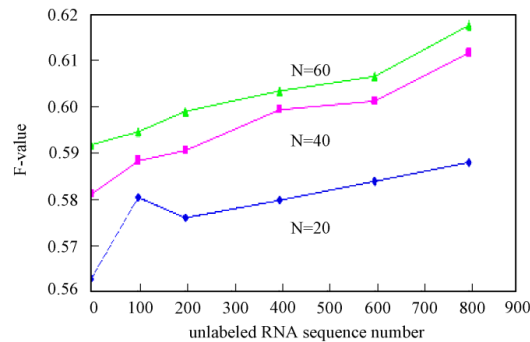


Figure 3. Comparison of F-values for different unlabeled RNA sequence numbers

It can be seen that, regardless of the size of the labeled sequence set, the performance of the prediction model can be improved after the unlabeled RNA sequence is added. The more unlabeled sequences that are added, the better the performance will be. This shows that the model can use unlabeled sequence samples effectively to improve the prediction accuracy.

4. Conclusions

In this paper, the traditional grammar based on the supervised learning model was improved. The design can integrated unlabeled data into the semi-supervised learning stochastic grammar models, apply semi-supervised learning technology to RNA secondary structure prediction, and overcome, to an extent, the comparative sequence analysis method to predict RNA secondary structure. This study effectively improved the performance of the prediction of RNA secondary structure by using unlabeled sequences through the stochastic grammar model. Experiments demonstrated that this method has better prediction, which means that it has good application value.

Acknowledgements

This work is supported by the Scientific Research Projects (No. 15C0204) of the Hunan Education Department.

References

1. K. Kappel and R. Das, "Sampling Native-Like Structures of RNA-Protein Complexes through Rosetta Folding and Docking," *Structure*, Vol. 31, No. 4, pp. 139-151, 2018
2. D. D. Song and Z. D. Deng, "Fussy Model for RNA Secondary Prediction," *Science in China Vol E: Information Science*, Vol. 37, No. 10, pp. 1285-1303, 2007
3. J. H. Liu, J. T. L. Wang, and J. Hu, "A Method for Aligning RNA Secondary Structures and Its Application to RNA Motif Detection," *BMC Bioinformatics*, Vol. 6, pp. 88-107, 2005
4. S. X. Tang, Y. B. Liu, and J. Yin, "Research Advances of Grammatical Inference of RNA Secondary Structure," *China Journal of Bioinformatics*, Vol. 4, pp. 190-192, 2008
5. M. Andronescu, A. Condon, and D. H. Mathews, "Computational Approaches for RNA Energy Parameter Estimation," *RNA*, Vol. 16, pp. 2304-2318, 2010
6. S. X. Tang, Y. Zhou, and Y. Yi, "The Application of Stochastic Grammars for RNA Secondary Structure Prediction," *Journal of Biomathematics*, Vol. 23, No. 4, pp. 735-742, 2008
7. E. P. Nawrocki, "Structural RNA Homology Search and Alignment using Covariance Models," Washington University School of Medicine, 2009
8. R. D. Dowell and S. R. Eddy, "Evaluation of Several Lightweight Stochastic Context-Free Grammars for RNA Secondary Structure Prediction," *BMC Bioinformatics*, Vol. 5, No. 1, pp. 71, 2004
9. T. Jebara, "Discriminative, Generative and Imitative Learning," Massachusetts Inst. of Technology, Media laboratory, 2001
10. E. Come, L. Oukhellou, and T. Denoeux, "Learning from Partially Supervised Data using Mixture Models and Belief Functions," *Pattern Recognition*, Vol. 42, No. 3, pp. 334-348, 2009
11. A. Tanzer, I. L. Hofacker, and R. Lorenz, "RNA Modifications in Structure Prediction – Status Quo and Future Challenges," *Methods*, Vol. 39, No. 10, pp. 23-38, 2018
12. G. Tur, D. D. Hakkani-Tur, and R. E. Schapire, "Combining Active and Semi-Supervised Learning for Spoken Language Understanding," *Speech Communication*, Vol. 45, pp. 171-186, 2005
13. S. X. Tang, Y. Zhou, and S. Zou, "The RNA Secondary Structure Prediction based on the Lexicalized Stochastic Grammar Model," *Computer Engineering & Science*, Vol. 31, No. 3, pp. 128-131, 2009
14. S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy, "Rfam: An RNA Family Database," *Nucleic Acids Research*, Vol. 31, No. 1, pp. 429-441, 2003

Sixin Tang is a lecturer in the College of Computer Science and Technology at Hengyang Normal University. His research interests include machine learning and bioinformatics.