Predicting small RNAs in bacteria via sequence learning ensemble method

Wen Zhang[1*], Jingwen Shi[2], Guifeng Tang[1], Wenjian Wu[3], Xiang Yue[4], Dingfang Li[2*]

1. School of Computer, Wuhan University, Wuhan 430072, China
2. School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China
3. Electronic Information School, Wuhan University, Wuhan 430072, China
4. International School of Software, Wuhan University, Wuhan 430072, China

zhangwen@whu.edu.cn, shijingwen@whu.edu.cn, gftang@whu.edu.cn,

westonwu@whu.edu.cn, tommy96@whu.edu.cn, whudfli@163.com

*Corresponding Author

Abstract*—Bacterial small non-coding RNAs (sRNAs) play important roles in various physiological processes, and predicting sRNAs is an important task. In this paper, we develop a computational method for the sRNA prediction by using sRNA sequence-derived features. We investigate a variety of sRNA sequence-derived features, and evaluate the usefulness of features for the sRNA prediction. Then, we develop the sequence learning ensemble method, which uses the linear weighted sum of outputs from the individual feature-based predictors to predict sRNAs, and the genetic algorithm is adopted to optimize the parameters in the ensemble system. In the computational experiments, we compile a balanced dataset and four imbalanced datasets, and evaluate our method on these datasets by using 5-fold cross validation. The sequence learning ensemble method can achieve AUC scores greater than 0.9, and outperforms existing state-of-the-art sRNA prediction methods. In conclusion, the proposed method has a great potential for sRNA prediction. The source codes, datasets and supplementary are available in http://www.bioinfotech.cn/BIBM2017/SLEM.*

Keywords—sRNA; sequence-derived features; genetic algorithm; ensemble learning

I. INTRODUCTION

Non-coding RNA is an RNA molecule that is not translated into a protein. Small non-coding RNAs (sRNAs) were discovered fifty years ago in bacteria, owing to its high abundance during normal growth of cell [1]. In recent years, sRNAs have attracted more and more attention.

The sRNAs usually exist in bacteria, and have a typical size of 50-500 nucleotides [2]. sRNAs act as functional RNAs rather than encoding proteins, i.e., base-pairing with RNAs and DNAs or regulating protein-protein interactions by binding to proteins. Therefore, sRNAs can control physiological processes in bacteria, including growth, development, cell proliferation, differentiation, metabolic reactions and carbon metabolism [3].

Since sRNAs in bacteria have different functions, the identification of sRNAs is the prerequisite for understanding biological mechanisms. Wet methods for identifying sRNAs include microarrays, co-purification with proteins, and functional genetic screens [4], and the high throughput techniques. However, these methods are costly and time-consuming. Researchers also attempted to develop computational methods for the sRNA prediction. The traditional computational methods utilized comparative genomics technique and free energy technique to predict sRNAs. Comparative genomics methods aligned sequence or structural homology to known sRNAs from different bacteria to identify sRNAs. Rivas [5] identified non-coding RNAs in E. coli by comparative genomics. Free energy-based methods distinguished sRNAs from randomized sequences by scoring structural alignment and the free energy change when sRNA sequences transformed into ordinary structure. Zuker [6] predicted the second structure of sRNA by minimizing the folding free energy change.

Recently, machine learning methods were introduced to the sRNA prediction. Tjaden [7] integrated primary sequence data, transcript expression data and conserved RNA structure information to predict sRNAs in bacteria via Markov models. Arnedo [8] presented an integration methodology to identify bacterial sRNAs by incorporating different existing sRNA prediction methods. Carter [9] utilized the composition information of sRNA sequences to train support vector machine (SVM) models and neural network (NN) models. Barman [10] used tri-nucleotide composition of sequences to construct SVM-based models. Although several machine learning-based methods have been proposed for sRNA prediction, there is still room for improving performances.

As discussed above, several sRNA sequence-derived features, such as the mono-nucleotide composition, di-nucleotide composition and tri-nucleotide composition, have been used to predict sRNAs. sRNA sequence feature-based prediction models produce high-accuracy performances, indicating that sRNA sequences bring important information for sRNAs prediction. As far as we known, sequence-derived features have been used to successfully solve lots of bioinformatics problems [11-19].

In this study, we develop a computational method named "the sequence learning ensemble method (SLEM)" for the sRNA prediction by using sRNA sequence-derived features. Existing methods usually utilized one or two features to construct models, but there are a great number of sequence-derived features which characterize sequences. We investigate a variety of sequence-derived features, and conduct the comprehensive study to evaluate their usefulness

for the sRNA prediction. Then, the sequence learning ensemble method uses the linear weighted sum of outputs from the individual sRNA feature-based predictors to predict sRNAs, and the genetic algorithm is adopted to optimize the parameters in the ensemble system. We compile one balanced dataset and four imbalanced datasets from the experimentally validated sRNAs of Salmonella Typhimurium LT2 (SLT2), and use them to evaluate the proposed method. In the computational experiments, SLEM produces an AUC score of 0.950 on the balanced dataset, and achieves AUC scores of 0.951, 0.949, 0.956 and 0.958 on four imbalanced datasets, respectively. Moreover, SLEM outperforms existing sRNA prediction methods.

II. Materials and Methods

A. Datasets

In this study, we adopt the benchmark dataset of 193 experimentally verified sRNAs of Salmonella Typhimurium LT2 (SLT2) [20]. The dataset was ever used by Barman [10] and Arnedo [8] for sRNA prediction. The complete genome sequence of SLT2 can be downloaded in NCBI (http://www.ncbi.nlm.nih.gov/nuccore/16763390?report=fasta), and the start and the end position information of the specific SLT2 sRNA can be obtained in [21]. After removing redundant sRNAs, we can obtain 182 experimentally verified sRNAs, which are real sRNAs. We shuffle the complete genome sequence with EMBOSS shuffleseq program [22], and then extract sequence fragments utilizing the start and the end position information of real sRNAs. These sequence fragments are used as pseudo sRNAs.

The real sRNAs and pseudo sRNAs are adopted as positive instances and negative instances, respectively. We construct one balanced dataset which has the same number of negative instances as positive instances; we also compile imbalanced datasets which have more pseudo sRNAs than real sRNAs, and the ratios of positive instances to negative instances are respectively 1:2, 1:3, 1:4 and 1:5.

B. sRNA Sequence-derived Features

As far as we know, lots of RNA sequence-derived features have been proposed to describe the characteristics of sequences. In this work, we consider various RNA sequence-derived features, and they are briefly introduced below.

k-spectrum profile: k-spectrum profile is also known as the k-mer profile. The spectrum profile describes repeated patterns of sequences. There are totally 4^k types of k-length contiguous subsequences. Given a sequence x, the k-spectrum profile is defined as $f_k^{spe}(x) = (c_1, c_2, \ldots c_{4^k})$, where c_i is the occurrence frequency of corresponding k-length contiguous subsequences. Spectrum profile has been widely adopted in biological applications [9-12].

Mismatch profile: (k, m)-mismatch profile is similar to k-spectrum profile but allowing up to $m(m < k)$ mismatches in the exact k-length contiguous subsequences[23]. Given a sequence x, the (k, m)-mismatch profile is defined as $f_k^{mis}(x) = (\sum_{j=0}^{m} c_{1j}, \sum_{j=0}^{m} c_{2j}, \ldots, \sum_{j=0}^{m} c_{4^k j})$, where c_{ij} denotes the occurrence frequency of the ith k-length contiguous subsequence with j mismatches.

Reverse compliment k-mer (k-RevcKmer): the feature is a kind of deformation of k-mer [23], and it takes the reverse complement of RNA into consideration. Given a sequence x, the reverse complement k-length contiguous subsequences will be removed after generating k-mer, then the occurrence frequencies of the remaining k-length subsequences are calculated to constitute a feature vector.

Pseudo nucleotide composition features: the feature contains occurrences of different di-nucleotides or tri-nucleotides as well as their physicochemical properties [23]. There are four types of pseudo nucleotide composition features: parallel correlation pseudo di-nucleotide composition (PCPseDNC), parallel correlation pseudo tri-nucleotide composition (PCPseTNC), series correlation pseudo di-nucleotide composition (SCPseDNC), and series correlation pseudo tri-nucleotide composition (SCPseTNC). The pseudo nucleotide composition features have a parameter λ representing the highest counted rank of the correlation along a sequence. More details about pseudo nucleotide composition features are described in [11, 23].

TABLE I. sRNA SEQUENCED-DERIVED FEATURES

Feature group	Index	Feature	Dimension	Parameter
Spectrum profile	F1	1-spectrum profile	4	No parameter
	F2	2-spectrum profile	16	No parameter
	F3	3-spectrum profile	64	No parameter
	F4	4-spectrum profile	256	No parameter
	F5	5-spectrum profile	1024	No parameter
Mismatch profile	F6	(3, m)-mismatch profile	64	m : the max mismatches
	F7	(4, m)-mismatch profile	256	m : the max mismatches
	F8	(5, m)-mismatch profile	1024	m : the max mismatches
Reverse compliment k-mer	F9	1-RevcKmer	2	No parameter
	F10	2-RevcKmer	10	No parameter
	F11	3-RevcKmer	32	No parameter
	F12	4-RevcKmer	136	No parameter
	F13	5-RevcKmer	528	No parameter
Pseudo nucleotide composition	F14	PCPseDNC	$16+\lambda$	λ : the highest counted rank
	F15	PCPseTNC	$64+\lambda$	λ : the highest counted rank
	F16	SCPseDNC	$16+6\times\lambda$	λ : the highest counted rank
	F17	SCPseTNC	$64+12\times\lambda$	λ : the highest counted rank

Features used in this paper are demonstrated in Table I.

C. The Sequence Learning Ensemble Method

As the above discussion, we extract a variety of sequence-derived features from sRNA sequences. Although diverse features bring diverse information, features may also bring noises and redundant information. How to make use of useful features is critical for accurately predict sRNAs. In recent years, the machine learning studies show that combining the outputs of multiple predictors improves the prediction performance and reduces the generalization error. Inspired by the idea, we develop the sequence learning ensemble method (SLEM) for the sRNA prediction.

Our SLEM method has two critical components: the sequence learning and the ensemble strategy. Firstly, we consider various sequence-derived features, which can guarantee the diversity of sRNA characteristics, and we construct individual sequence feature-based prediction models based on individual sRNA sequence-derived features by using machine learning classifiers. Here, we adopt a popular machine learning classifier: random forest (RF). Further, we adopt the individual sequence feature-based models as base predictors, and design an ensemble strategy to combine the outputs of base predictors for the sRNA prediction.

The ensemble strategy is introduced as follows. Given N features, we construct N base predictors $f_i(i = 1,2, \ldots N)$ on the training set. For a new RNA sequence x, its prediction probability of being predicted as a real sRNA by the base predictors f_i is represented as $f_i(x)$. The final predicting score of the sequence x is given by

$$F(x) = \sum_{i=1}^{N} w_i f_i(x)$$
$$\sum_{i=1}^{N} w_i = 1,\ w_i \geq 0 \quad (1)$$

The weights $w_i(i = 1,2,\cdots,N)$ are free parameters in (1), and we utilize intelligent optimization algorithm: the genetic algorithm (GA) to optimize weights. GA is an evolutionary algorithm which simulates the biological evolution, and its ability for the optimization problems has been proved in lots of applications [12, 14].

We randomly generate 1000 weight vectors as the candidate solutions, and encode these candidates into chromosomes as the initial population. For a chromosome, the fitness score is the AUC score of the ensemble model in (1) using the corresponding weights, which is evaluated on the validation set. In each generation, the chromosomes are updated by three operators, i.e., selection, crossover and mutation. The selection probability, crossover probability and mutation probability are dynamically adjusted according to the fitness scores of chromosomes [24]. After 500 generations of update, we determine optimal weights in (1). Finally, the ensemble system makes predictions for the testing set.

III. Results and Discussion

A. Evaluation Metrics

5-fold cross validation (5-CV) is adopted to estimate performances of prediction models. We use several performance metrics to evaluate performance of the proposed method: accuracy (ACC), sensitivity (SN), specificity (SP), and AUC score. The AUC score assesses the performance regardless of any threshold. Thus, we adopt the AUC score as the primary metric.

B. Parameters Settings

As shown in Table I, we consider various sRNA sequence-derived features, and several features have parameters. We have to discuss how to set parameters in the computational experiments.

The mismatch profile has the parameter m for the max mismatches and k for the length of contiguous subsequences, and m is usually less than one-third of k. Since we consider (3, m)-mismatch profile, (4, m)-mismatch profile and (5, m)-mismatch profile, we have to set m = 1 for them.

The pseudo nucleotide composition features (PCPseDNC, SCPseDNC, PCPseTNC and SCPseTNC) have the parameter λ. L denotes the length of the shortest sRNA sequence. In our SLT2 sRNA dataset, the shortest sRNA sequence has 45 nucleotides, i.e. $L = 45$. The length distribution analysis of sRNA sequences is provided in the supplementary. λ is an integer that ranges from 0 to $L - 2$ in PCPseTNC and SCPseTNC; λ ranges from 0 to $L - 3$ in PCPseTNC and SCPseTNC. We construct prediction models based on PCPseDNC, SCPseDNC, PCPseTNC and SCPseTNC by using different parameter values, and 5-CV results on the balanced dataset are provided in the supplementary. The results show that $\lambda = 1$ leads to the greatest AUC scores for PCPseTNC and SCPseTNC; $\lambda = 11$ and $\lambda = 13$ lead to the greatest AUC scores for PCPseDNC and SCPseDNC. Accordingly, we use these values for PCPseDNC, SCPseDNC, PCPseTNC and SCPseTNC in the following study.

C. Feature Evaluation

For the comprehensive study, we consider sRNA sequence-derived features in Table I and Random Forest for the sRNAs prediction. We implement 5-fold cross validation to evaluate models based on one balanced dataset and four imbalanced datasets in the section “Datasets”.

We construct individual feature-based models by using RF on the balanced dataset and four imbalanced datasets. We compare individual feature-based models and test the influences of ratios of positive instances vs. negative instances on performances of prediction models. As demonstrated in Table II, among all these features, 4-RevcKmer (F13) and 5-RevcKmer (F13) perform best, and both achieve AUC score of 0.940. Furthermore, sequence-derived features may produce similar performances on the balanced dataset and imbalanced datasets, indicating these sequence information is robust to the data ratio. In general, most features can lead to high-accuracy performances on benchmark datasets, and 4-spectrum profile (F4), 3-RevcKmer (F11), 4-RevcKmer (F12), 5-RevcKmer (F13) and PCPseTNC (F15) features have better performances than other features in sRNA prediction. Different features have different performances, and can bring diverse

information. Thus, all features are adopted to build the final prediction models.

TABLE II. THE PERFORMANCES OF INDIVIDUAL FEATURE-BASED MODELS CONSTRUCTED BY RF ON THE BENCHMARK DATASETS

Index	AUC					ACC				
	Balanced	Imbalanced				Balanced	Imbalanced			
	1:1	1:2	1:3	1:4	1:5	1:1	1:2	1:3	1:4	1:5
F1	0.683	0.706	0.729	0.724	0.741	0.629	0.728	0.799	0.835	0.865
F2	0.826	0.841	0.856	0.866	0.866	0.763	0.794	0.841	0.869	0.887
F3	0.904	0.911	0.917	0.926	0.930	0.823	0.827	0.863	0.876	0.890
F4	0.922	0.931	0.927	0.934	0.931	0.856	0.842	0.854	0.869	0.883
F5	0.914	0.899	0.873	0.866	0.863	0.848	0.831	0.844	0.863	0.880
F6	0.767	0.797	0.819	0.832	0.843	0.708	0.777	0.829	0.854	0.876
F7	0.880	0.893	0.905	0.912	0.922	0.802	0.816	0.852	0.873	0.892
F8	0.917	0.923	0.928	0.934	0.939	0.840	0.836	0.858	0.874	0.889
F9	0.639	0.649	0.664	0.683	0.689	0.608	0.689	0.749	0.803	0.832
F10	0.842	0.838	0.863	0.873	0.877	0.771	0.800	0.843	0.871	0.892
F11	0.923	0.921	0.933	0.938	0.941	0.847	0.866	0.883	0.898	0.905
F12	0.940	0.947	0.946	0.953	0.955	0.874	0.875	0.884	0.896	0.908
F13	0.940	0.928	0.923	0.926	0.921	0.876	0.862	0.875	0.893	0.904
F14	0.900	0.885	0.885	0.884	0.883	0.829	0.814	0.843	0.871	0.887
F15	0.928	0.920	0.922	0.925	0.919	0.852	0.848	0.874	0.885	0.897
F16	0.905	0.895	0.896	0.889	0.888	0.826	0.836	0.860	0.876	0.893
F17	0.903	0.900	0.901	0.905	0.898	0.814	0.827	0.866	0.884	0.901

D. Performances of Sequence Learning Ensemble Method

In this section, we evaluate the performances of the sequence learning ensemble method (SLEM) which integrates diverse sRNA sequence-derived features. We implement 5-fold cross validation to evaluate our proposed SLEM on the balanced dataset and four imbalanced datasets.

As shown in Table III, SLEM achieves the 5-CV AUC score of 0.950 on the balanced dataset, and outperforms the best-performed base predictor which uses 5-RevcKmer feature (F13) and produce the AUC score of 0.940. SLEM performs similarly on imbalanced datasets, and achieves AUC scores of 0.951, 0.949, 0.956, 0.958 on the datasets with imbalance ratios 1:2, 1:3, 1:4 and 1:5, respectively. SLEM also improves prediction performances over individual feature-based predictors on the four imbalanced datasets. The results demonstrate that SLEM can effectively combine various features to produce high-accuracy performances.

TABLE III. THE PERFORMANCES OF SLEM ON THE BALANCED AND IMBALANCED DATASETS

Dataset	Ratio	AUC	ACC	SN	SP
Balanced	1:1	0.950	0.893	0.863	0.923
Imbalanced	1:2	0.951	0.861	0.615	0.984
	1:3	0.949	0.873	0.513	0.993
	1:4	0.956	0.885	0.445	0.996
	1:5	0.958	0.898	0.405	0.997

E. Comparison with Existing sRNA Prediction Methods

As far as we know, two state-of-the-art sRNA prediction methods, Carter' s method [9] and Barman's method [10], are proposed recently to predict sRNAs. The two methods utilized sequences information to develop machine learning-based prediction models. Carter used composition information of sRNA sequences, including mono-nucleotide composition and di-nucleotide composition, to train both SVM and NN for the sRNA prediction. The mono-nucleotide composition and the di-nucleotide composition are actually the 1-spectrum profile and 2-spectrum profile, and SVM models produced better performances than NN models. Barman adopted tri-nucleotide composition of sequences, which is the 3-spectrum profile in our method, to build SVM-based prediction models. Thus, we adopt these methods as benchmark methods for comparison.

We compare our proposed SLEM with Carter's sRNA prediction method [9] and Barman's sRNA prediction method [10] on the balanced dataset and four imbalanced datasets. These five datasets we compiled are the same as Barman's datasets, so we directly adopt Barman's prediction results reported in[10]. Carter used 1-spectrum profile and 2-spectrum profile to train SVM-based prediction models, and these models are used as the individual feature-based predictors in our ensemble models. Therefore, we can easily implement Carter's method for comparison. All models are evaluated by 5-CV.

As shown in Table IV, SLEM achieves AUC score of 0.950 and ACC of 0.893 on the balanced dataset, better than Barman's method (AUC score of 0.938 and ACC of 0.882) and Carter's method (AUC score of 0.566 and ACC of 0.511). Our proposed SLEM also achieves both high sensitivity and specificity value of 0.863 and 0.923, respectively. Moreover, SLEM outperforms Carter's method and Barman's method on four imbalanced datasets. There are several reasons why SLEM has excellent prediction performances. Firstly, diverse sRNA sequence-derived features can guarantee the information diversity. Secondly, the ensemble method provides an efficient way to improve predicting performances over individual feature-based predictors. Finally, GA automatically adjusts the weights to achieve the best prediction results.

TABLE IV. PERFORMANCES MEASURES OF DIFFERENT METHODS ON BALANCED SLT2 DATASETS

Dataset	Ratio	Method	AUC	ACC	SN	SP
Balanced	1:1	Carter's method	0.566	0.511	0.264	0.758
		Barman's method	0.938	0.882	0.846	0.918
		SLEM	0.950	0.893	0.863	0.923
Imbalanced	1:2	Carter's method	0.602	0.678	0.033	1.000
		Barman's method	0.937	0.884	0.851	0.916
		SLEM	0.951	0.861	0.615	0.984
	1:3	Carter's method	0.619	0.757	0.030	1.000
		Barman's method	0.944	0873	0.818	0.927
		SLEM	0.949	0.873	0.513	0.993
	1:4	Carter's method	0.627	0.805	0.025	1.000
		Barman's method	0.944	0.874	0.818	0.929
		SLEM	0.956	0.885	0.445	0.996
	1:5	Carter's method	0.636	0.835	0.011	1.000
		Barman's method	0.943	0.875	0.884	0.865
		SLEM	0.958	0.898	0.405	0.997

IV. CONCLUSION

Bacterial small non-coding RNAs play important regulatory roles in controlling various physiological processes. Predicting sRNAs is important for understanding the biological mechanism of bacteria. We design a computational method for the sRNA prediction by combining sRNA sequence-derived features. This method is evaluated on the benchmark SLT2 datasets, and achieves high-accuracy performances. Compared with other state-of-the-art sRNA prediction methods, our method can produce better performances. In conclusion, the proposed method could efficiently predict sRNAs in bacteria. The source codes, datasets and supplementary are available in http://www.bioinfotech.cn/BIBM2017/SLEM.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (61772381, 61572368) and The Fundamental Research Funds for the Central Universities (2042017kf0219).

REFERENCES

[1] B. Tjaden, "Prediction of small, noncoding RNAs in bacteria using heterogeneous data," *Journal of mathematical biology,* vol. 56, no. 1-2, pp. 183-200, 2008.

[2] L. S. Waters, and G. Storz, "Regulatory RNAs in bacteria," *Cell,* vol. 136, no. 4, pp. 615-628, 2009.

[3] S. Gottesman, and G. Storz, "Bacterial small RNA regulators: versatile roles and rapidly evolving variations," *Cold Spring Harbor perspectives in biology,* vol. 3, no. 12, pp. a003798, 2011.

[4] S. Altuvia, "Identification of bacterial small non-coding RNAs: experimental approaches," *Current opinion in microbiology,* vol. 10, no. 3, pp. 257-261, 2007.

[5] E. Rivas, R. J. Klein, T. A. Jones, and S. R. Eddy, "Computational identification of noncoding RNAs in E. coli by comparative genomics," *Current biology,* vol. 11, no. 17, pp. 1369-1373, 2001.

[6] M. Zuker, "Mfold web server for nucleic acid folding and hybridization prediction," *Nucleic acids research,* vol. 31, no. 13, pp. 3406-3415, 2003.

[7] B. Tjaden, S. S. Goodwin, J. A. Opdyke, M. Guillier, D. X. Fu, S. Gottesman, and G. Storz, "Target prediction for small, noncoding RNAs in bacteria," *Nucleic acids research,* vol. 34, no. 9, pp. 2791-2802, 2006.

[8] J. Arnedo, R. Romero-Zaliz, I. Zwir, and C. Del Val, "A multiobjective method for robust identification of bacterial small non-coding RNAs," *Bioinformatics,* vol. 30, no. 20, pp. 2875-2882, 2014.

[9] R. J. Carter, I. Dubchak, and S. R. Holbrook, "A computational approach to identify genes for functional RNAs in genomic sequences," *Nucleic acids research,* vol. 29, no. 19, pp. 3928-3938, 2001.

[10] R. K. Barman, A. Mukhopadhyay, and S. Das, "An improved method for identification of small non-coding RNAs in bacteria using support vector machine," *Scientific Reports,* vol. 7, 2017.

[11] L. Luo, D. Li, W. Zhang, S. Tu, X. Zhu, and G. Tian, "Accurate prediction of transposon-derived piRNAs by integrating various sequential and physicochemical features," *PloS one,* vol. 11, no. 4, pp. e0153268, 2016.

[12] D. Li, L. Luo, W. Zhang, F. Liu, and F. Luo, "A genetic algorithm-based weighted ensemble method for predicting transposon-derived piRNAs," *BMC Bioinformatics,* vol. 17, no. 1, pp. 329, 2016.

[13] W. Zhang, X. Zhu, Y. Fu, J. Tsuji, and Z. Weng, "The prediction of human splicing branchpoints by multi-label learning." *IEEE International Conference on Bioinformatics and Biomedicine IEEE,* 2016, pp. 254-259.

[14] W. Zhang, Y. Niu, H. Zou, L. Luo, Q. Liu, and W. Wu, "Accurate prediction of immunogenic T-cell epitopes from epitope sequences using the genetic algorithm-based ensemble learning," *PloS one,* vol. 10, no. 5, pp. e0128194, 2015.

[15] W. Zhang, J. Liu, Y. Xiong, M. Ke, and K. Zhang, "Predicting immunogenic T-cell epitopes by combining various sequence-derived features." *bioinformatics and biomedicine,* 2013, pp. 4-9.

[16] W. Zhang, Y. Niu, Y. Xiong, M. Zhao, R. Yu, and J. Liu, "Computational Prediction of Conformational B-Cell Epitopes from Antigen Primary Structures by Ensemble Learning," *PLOS ONE,* vol. 7, no. 8, 2012.

[17] W. Zhang, J. Liu, M. Zhao, and Q. Li, "Predicting linear B-cell epitopes by using sequence-derived structural and physicochemical features," *International Journal of Data Mining and Bioinformatics,* vol. 6, no. 5, pp. 557-569, 2012.

[18] W. Zhang, J. Liu, and Y. Niu, "Quantitative prediction of MHC-II binding affinity using particle swarm optimization," *Artificial Intelligence in Medicine,* vol. 50, no. 2, pp. 127-132, 2010.

[19] W. Zhang, J. Liu, and Y. Niu, "Quantitative prediction of MHC-II peptide binding affinity using relevance vector machine," *Applied Intelligence,* vol. 31, no. 2, pp. 180-187, 2009.

[20] G. Padalon-Brauch, R. Hershberg, M. Elgrably-Weiss, K. Baruch, I. Rosenshine, H. Margalit, and S. Altuvia, "Small RNAs encoded within genetic islands of Salmonella typhimurium show host-induced expression and role in virulence," *Nucleic acids research,* vol. 36, no. 6, pp. 1913-1927, 2008.

[21] 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, no. suppl_1, pp. D121-D124, 2005.

[22] P. M. Rice, I. Longden, and A. J. Bleasby, "EMBOSS: The European Molecular Biology Open Software Suite," *Trends in Genetics,* vol. 16, no. 6, pp. 276-277, 2000.

[23] B. Liu, F. Liu, L. Fang, X. Wang, and K.-C. Chou, "repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects," *Bioinformatics,* vol. 31, no. 8, pp. 1307-1309, 2014.

[24] M. Srinivas, and L. M. Patnaik, "Adaptive probabilities of crossover and mutation in genetic algorithms," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. 24, no. 4, pp. 656-667, 1994.