武漢大學
Wuhan University

Predicting small RNAs in bacteria via sequence learning ensemble method

Wen Zhang, Jingwen Shi, Guifeng Tang, Wenjian Wu, Xiang Yue, Dingfang Li

Presenter：Xiang Yue
（Supervised by Prof. Wen Zhang）

Biomedical Big Data Mining Lab (BBDM-Lab)

Wuhan University, P.R China

OUTLINE

WHU

1 Background

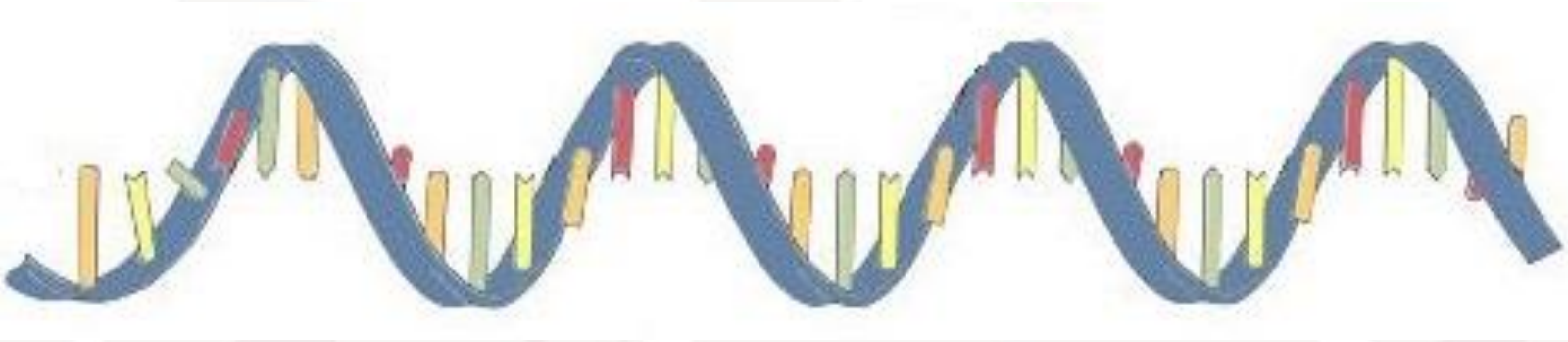
2 Method

3 Results

4 Conclusion

1 Background

1 Background



A G C U

Adenine Guanine Cytosine Uracil

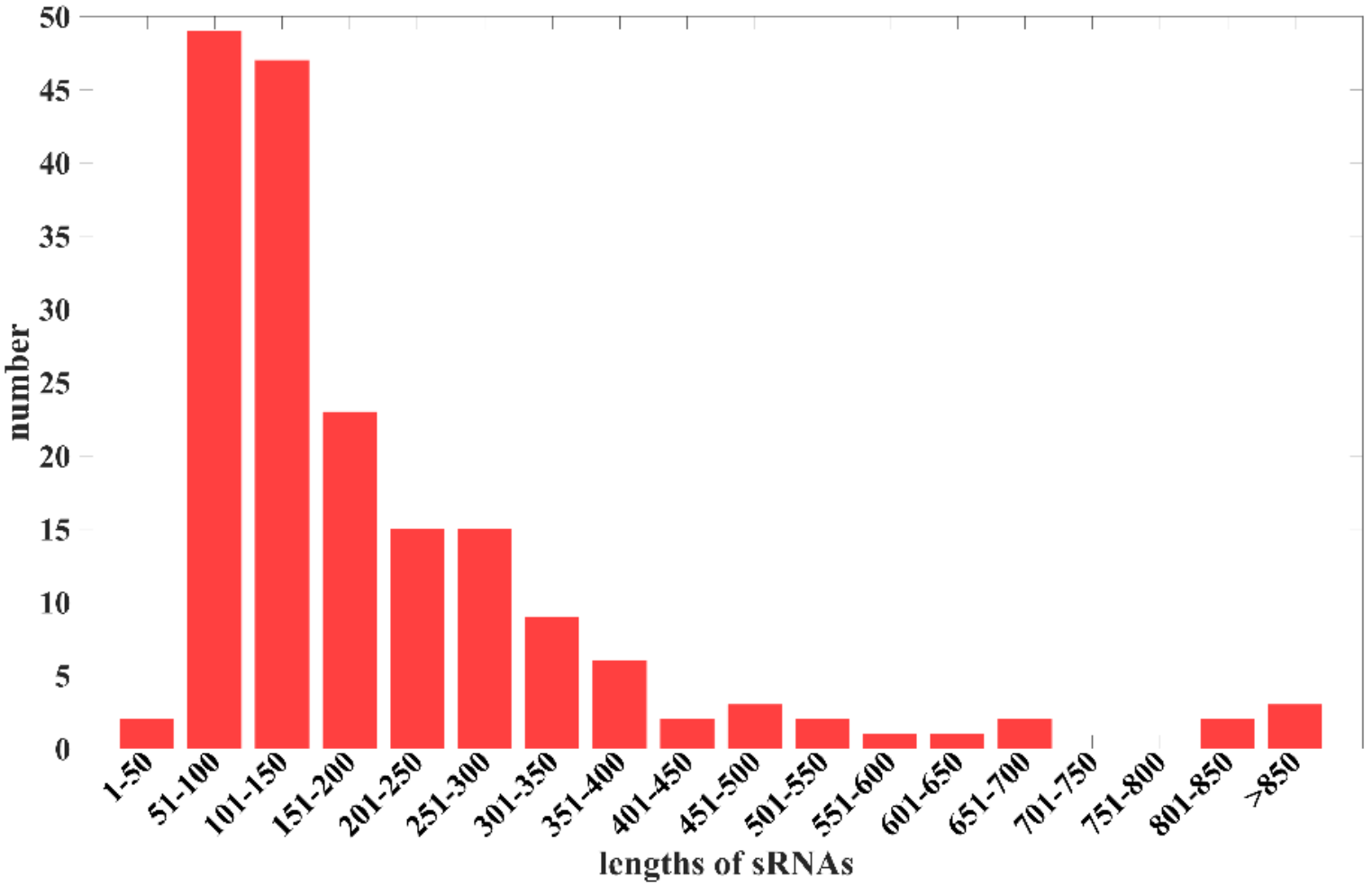
RNA sequence

1 Background

WHU

What's sRNA ?

- Small non-coding RNAs (sRNAs) exist in bacteria.
- Acting as functional RNAs
- Samples:
  GTTACAGGACGACCTGTAAAC
  GCTATTCTCACCGGGGACGGC
  CCC
- typical size : 50-500 nucleotides



Topic

- sRNAs play important roles in various physiological processes, including growth, development, cell proliferation, differentiation, metabolic reactions and carbon metabolism
- The identification of sRNAs is the prerequisite for understanding biological mechanisms
- The prediction of sRNAs is an important task and is a kind of supervised binary classification problem

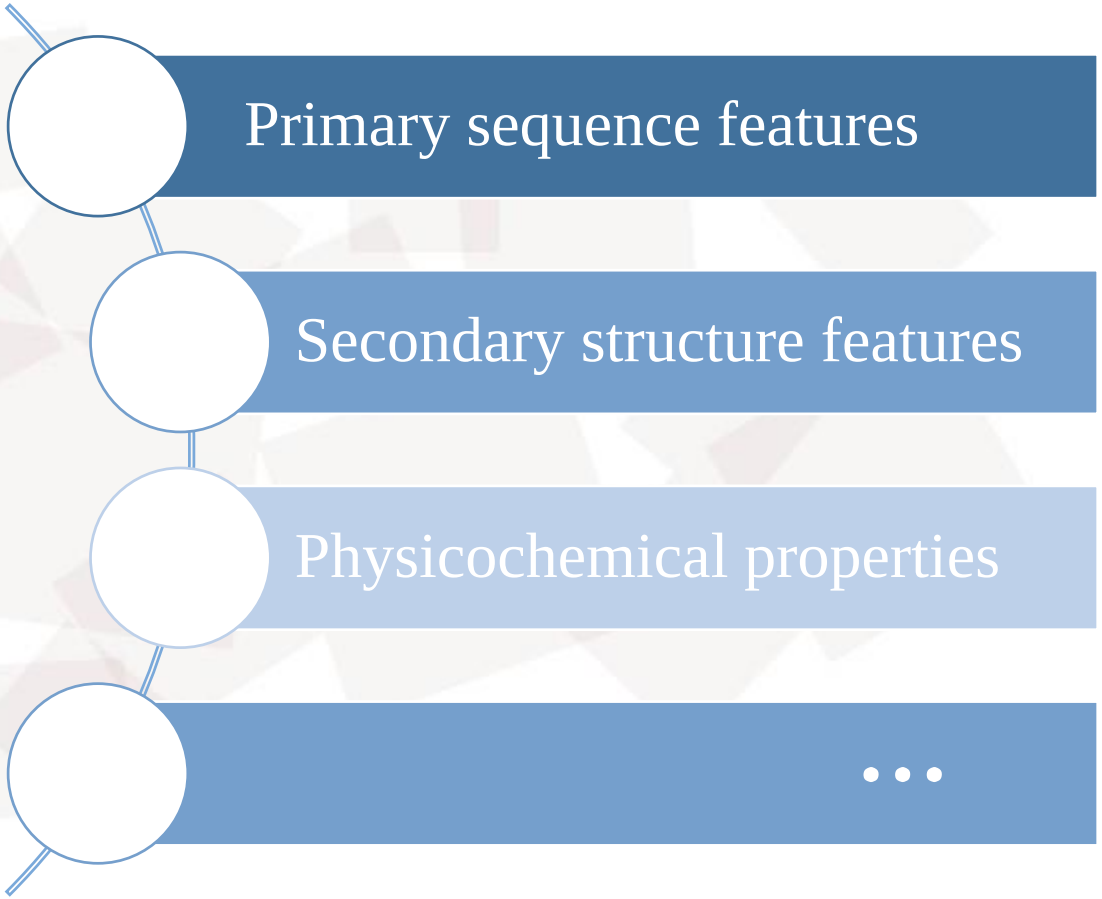
Keys

- Feature extraction
- Model construction
- Ensemble strategy

WHU

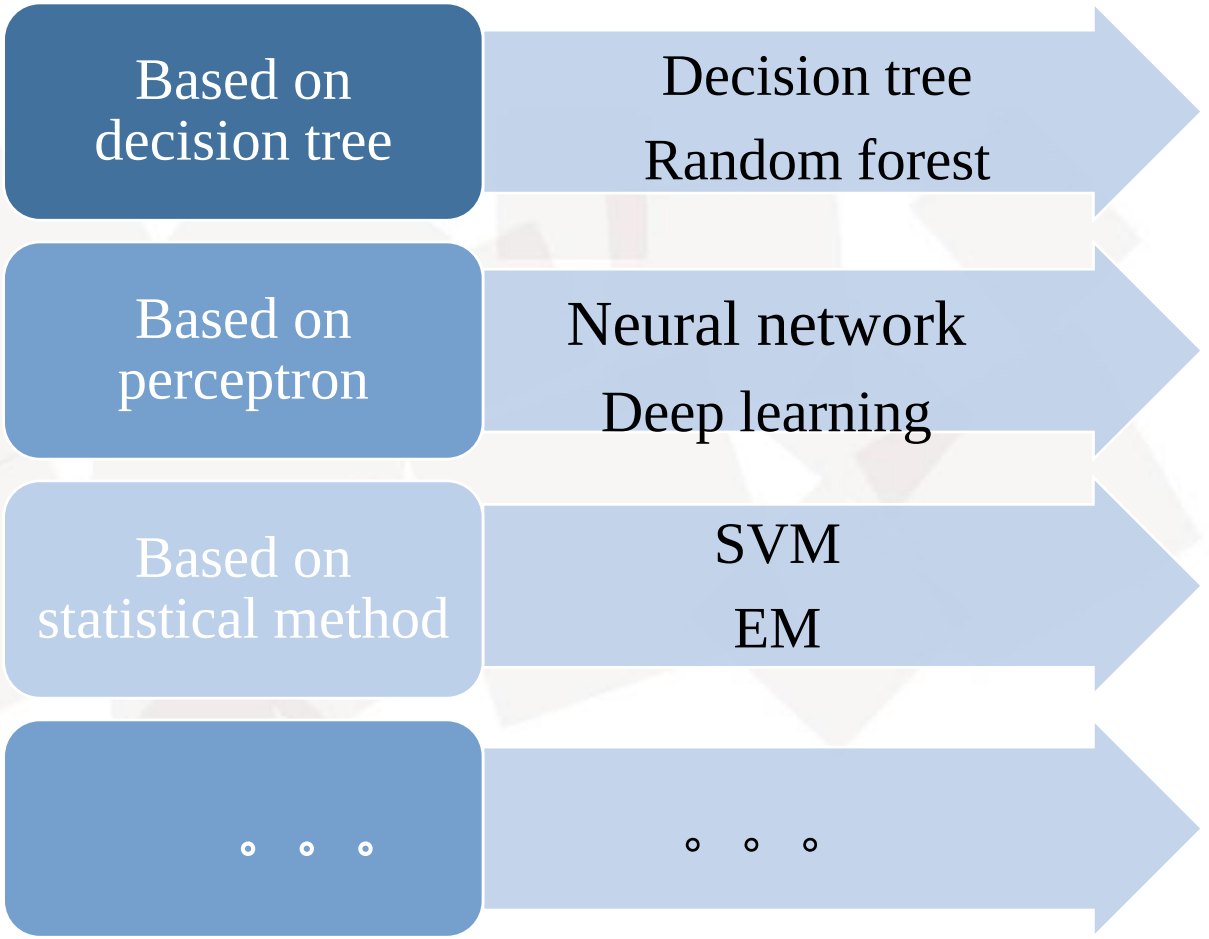
Feature extraction

Diverse features bring diverse information



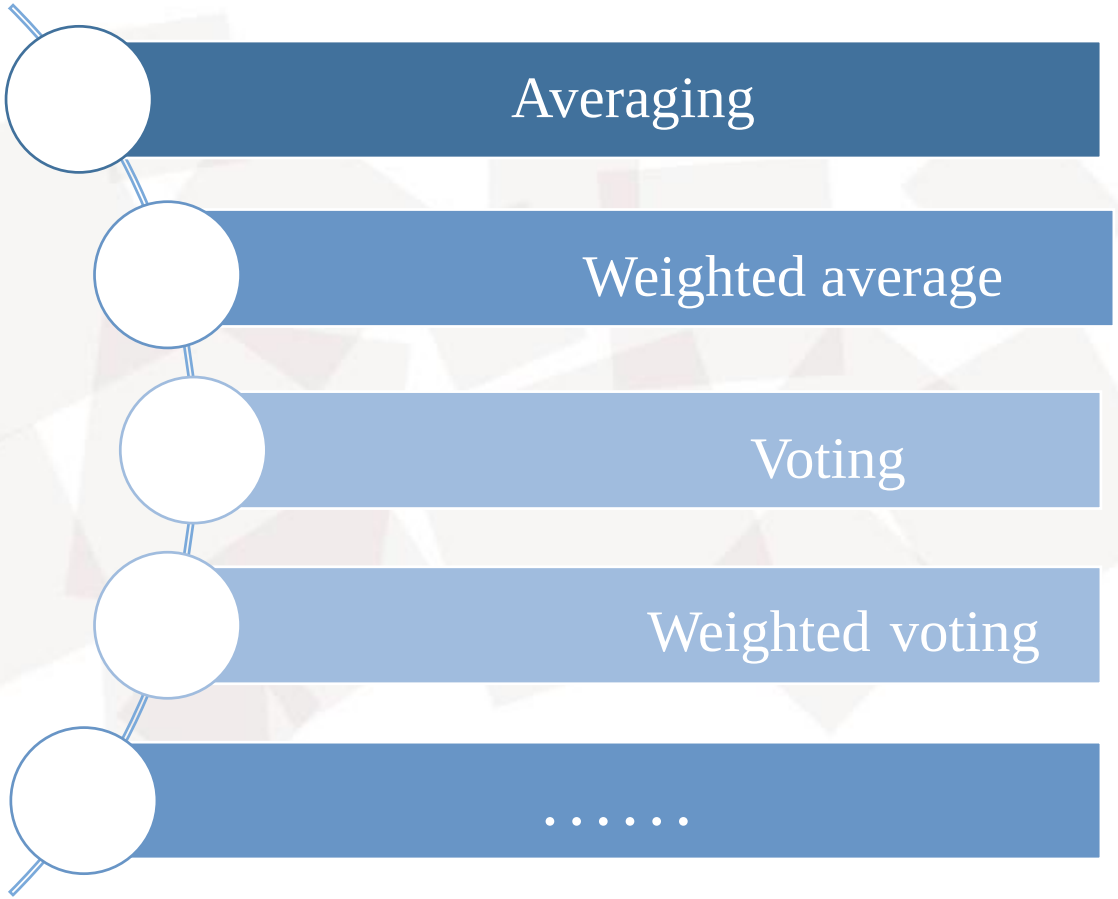
WHU

Model construction



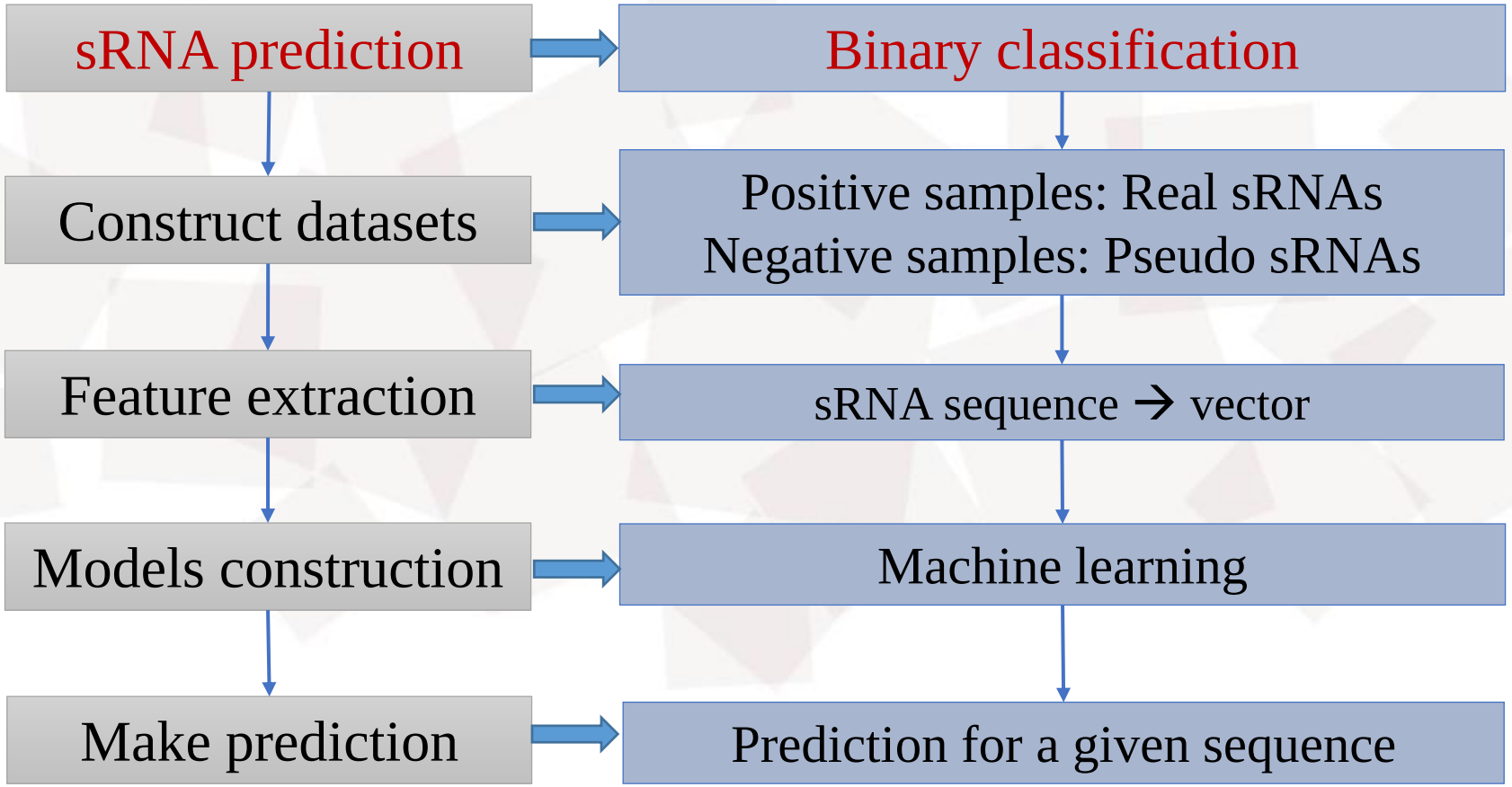
WHU

Ensemble strategy



2 Method

2 Method


sRNA prediction
Binary classification
Construct datasets
Positive samples: Real sRNAs
Negative samples: Pseudo sRNAs
Feature extraction
sRNA sequence → vector
Models construction
Machine learning
Make prediction
Prediction for a given sequence

2 Method

2.1 Datasets

Table1. Datasets

Species	Datasets	N(P): N(N)	Positive instances	Negative instances
SLT2	Balanced	1:1	182	182
	Imbalanced	1:2	182	364
		1:3	182	566
		1:4	182	728
		1:5	182	910

- NCBI

2.2 Feature extraction

Table 2. sRNA sequence-derived features

Features	Index	Dimensions
Spectrum Profile	F1~F5	4、16、64、256、1024
Mismatch profile	F6~F8	64、256、1024
Reverse compliment k-mer	F9~F13	4、16、64、256、1024
Pseudo nucleotide composition features	F14~F17	Concerned with the sequence length

- Each sequence-derived feature can be used to construct a individual feature-based prediction model by machine learning methods. Here, seventeen classifiers can be obtained.

2.3 Models

2.3.1. Individual sequence-derived feature-based model by machine learning method

$$\text{TAGG}\ldots\text{ACAT} \rightarrow x=(x_1, x_2, \ldots, x_d)$$

$$y = f(x), \quad x \in R^d; y \in [0,1].$$

- The function f is obtained by machine learning, such as support vector machine, random forest, deep belief network, neural network, and so on.

2.3 Models

2.3.2. The Sequence Learning Ensemble Method (SLEM)

- Considering the set of classifiers:$\{f_1, f_2,..., f_n\}$, the ensemble model is defined as:

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

$$\sum_{i=1}^{N} w_i = 1,\ w_i \geq 0$$

- Here, we adopt genetic algorithm(GA) to search the optimal weights ($w_1, w_2, \ldots, w_n$)

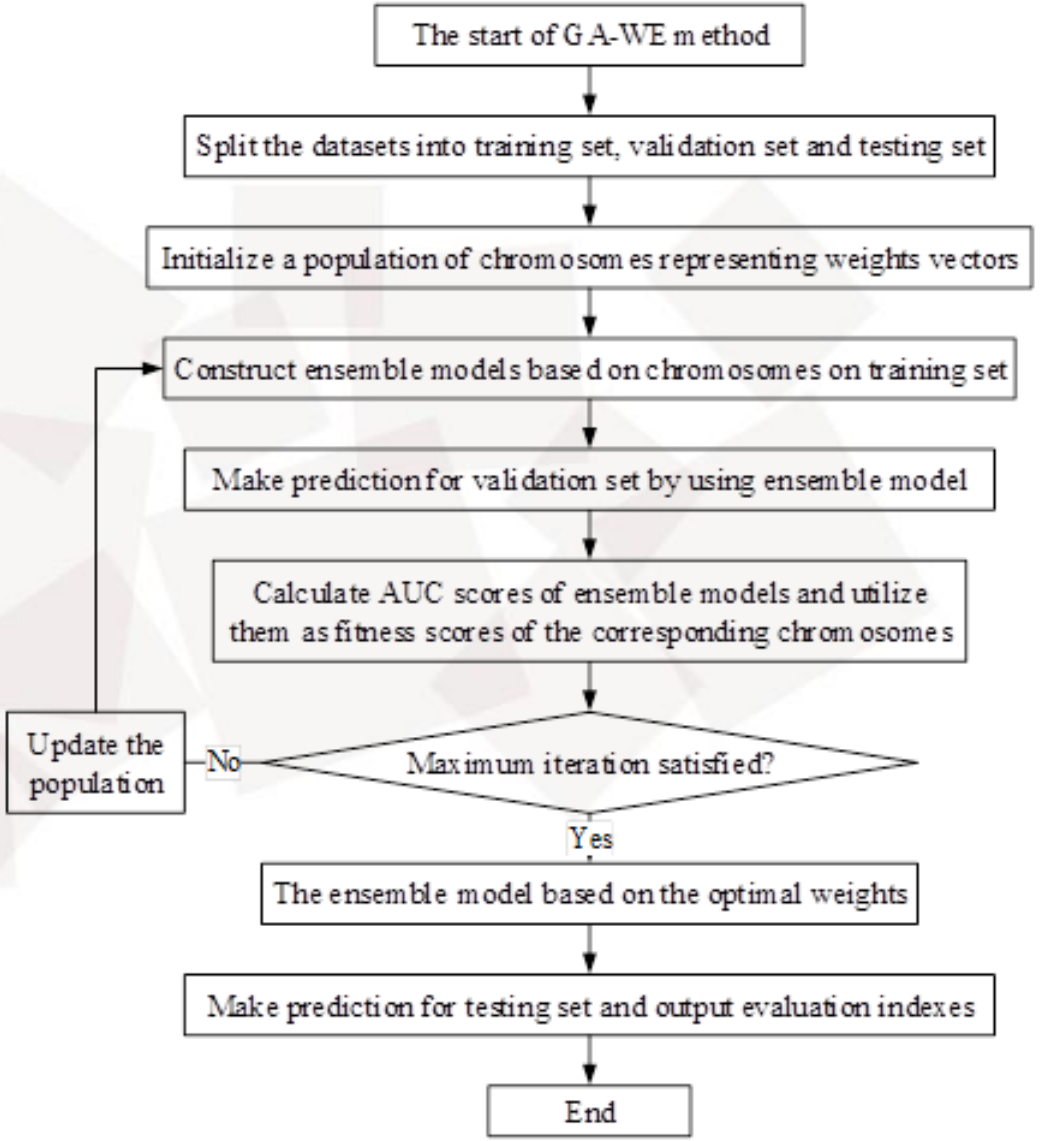
Optimal weights

2 Method

2.3 Models

SLEM:

- 5-fold cross validation (5-CV) is adopted
- The prediction models are constructed on the train sets, and the weights are optimized on the validation set via GA. Finally, the prediction is made on the testing set.



3 Results

3 Results

Table 3. The performance of individual feature-based models constructed by RF on benchmark SLT2 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

3 Results

Table 4. the performance of SLEM on the balanced and imbalanced datasets

Datasets	N(P): N(N)	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

3 Results

Table 5. performance measures of different methods on balanced and imbalanced SLT2

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

4 Conclusion

4 Conclusion

WHU

- The sequence learning ensemble method(SLEM) can automatically determine the importance of different information resources and produce high-accuracy performances
- Compared with other state-of-the-art methods, the SLEM can lead to better performances. Therefore, the SLEM has a great potential for sRNA prediction

Q & A

Thanks !