applied sciences

MDPI

Article

Support Vector Machine Classifier for Accurate Identification of piRNA

Taoying Li *, Mingyue Gao, Runyu Song, Qian Yin and Yan Chen

School of Maritime Economics and Management, Dalian Maritime University, Dalian 116026, China; mygao@dlmu.edu.cn (M.G.); songrunyv@163.com (R.S.); YinQian_Qian@126.com (Q.Y.); chenyan@dlmu.edu.cn (Y.C.)

* Correspondence: litaoying@dlmu.edu.cn; Tel.: +86-155-6680-2152

Received: 26 September 2018; Accepted: 6 November 2018; Published: 9 November 2018

Featured Application: A support vector machine was used to achieve the best jackknife and the 5-fold cross-validation outcomes for identifying piRNAs (Piwi-interacting RNA) by combining these multiple features.

Abstract: Piwi-interacting RNA (piRNA) is a newly identified class of small non-coding RNAs. It can combine with PIWI proteins to regulate the transcriptional gene silencing process, heterochromatin modifications, and to maintain germline and stem cell function in animals. To better understand the function of piRNA, it is imperative to improve the accuracy of identifying piRNAs. In this study, the sequence information included the single nucleotide composition, and 16 dinucleotides compositions, six physicochemical properties in RNA, the position specificities of nucleotides both in N-terminal and C-terminal, and the proportions of the similar peptide sequence of both N-terminal and C-terminal in positive and negative samples, which were used to construct the feature vector. Then, the F-Score was applied to choose an optimal single type of features. By combining these selected features, we achieved the best results on the jackknife and the 5-fold cross-validation running 10 times based on the support vector machine algorithm. Moreover, we further evaluated the stability and robustness of our new method.

Keywords: Piwi-interacting RNA; sequence information; feature extraction; feature selection; machine learning

1. Introduction

Piwi-interacting RNA (PiRNA), a newly identified class of small non-coding RNA of which the length is 26–33 nt, can combine with PIWI proteins to regulate a transcription gene silencing process, heterochromatin modifications and to maintain germline and stem cell function in animals [1–4]. However, high-throughput sequencing indicates that tens of thousands of different piRNAs produced in various animals cannot recognize transposons [5]. Therefore, the function of piRNA needs to be further investigated. Experimental verification of piRNA targets and the piRNA-targeting rules are quite difficult to prove [6]. Crosslinking immunoprecipitation (CLIP) analyses of PIWIs suggest that they associate with diverse mRNAs. However, because diverse piRNAs engage with many mRNAs, it is hard to infer the target of a given piRNA from these CLIP analyses [6–8]. Therefore, additional approaches are required to distinguish piRNA sites in vivo.

In recent years, several computing biology tools have been proposed to identify piRNAs. The first model to identify piRNAs was piRNApredictor, firstly developed by Zhang et al. [9]. After three years, Wang et al. [10] proposed the second model for predicting piRNAs based on the transposon interaction and SVM (Support Vector Machine). Recently, Luo et al. [11] applied the sequence

information and physicochemical features of piRNAs and non-piRNAs to construct the model to predict piRNAs. In addition, Li et al. [12] used a powerful ensemble approach, which achieved a substantial improvement. According to the most attractive work by Lin et al. [13], 2L-piRNA (two-layer ensemble classifier for identifying piRNAs) can be used to identify piRNAs and their function types. 2L-piRNA yielded an accuracy of 86.1% for identifying piRNAs and non-piRNAs, and achieved an accuracy of 77.6% for identifying piRNAs with the function of instructing target mRNA deadenylation and piRNAs without the function of instructing target mRNA deadenylation.

Aiming to improve the prediction accuracy, we developed a novel predictor, 2L-piRNAPred (2-layer integrated program for identifying piRNAs in the first layer and determining if piRNAs have the function of instructing target mRNA deadenylation in the second layer.), by considering single nucleotide composition (SNC), the 16 dinucleotides compositions (DNC), six physicochemical properties in RNA, the position specificities of nucleotides both in N-terminal and C-terminal, and the proportions of similar peptide sequences of both N-terminal and C-terminal in the positive and negative samples. Consequently, F-Score was utilized to select the most efficient unique type of features. Furthermore, all the optimized feature vectors were combined to build our prediction model based on a support vector machine classifier. Both the jackknife test and 5-fold cross-validation running 10 times were implemented to test the stability and robustness of the model. In addition, the major comparison with the previous work based on 2L-piRNA showed that our model, 2L-piRNAPred, is superior both in sensitivity and specificity for the first layer and the second layer.

2. Methods

2.1. Datasets

We applied the same dataset as in [13]. Firstly, the piRNA sequences were originally taken from piRBASE [14] and non-piRNA sequences were obtained from [15]. Then, the CD-HIT with a cutoff threshold of 0.8 was employed to remove high-similarity sequences. Thirdly, we randomly selected the same number of negative samples as that of the positive samples to avoid the high false negative rate caused by the imbalanced dataset. Finally, there were 709 piRNA samples having the function of instructing target mRNA deadenylation (denoted as S^+_{inst}), 709 piRNA samples without this function (denoted as $S^+_{non-inst}$), and 1418 non-piRNA samples (denoted as S^-). The benchmark positive dataset was the union of S^+_{inst} and $S^+_{non-inst}$. Therefore, the training datasets in this study can be formulated as:

$$\begin{cases} S = S^+ \cup S^- \\ S^+ = S^+_{inst} \cup S^+_{non-inst} \end{cases} \quad (1)$$

2.2. Sequence Information

2.2.1. Pse-Nucleotide Composition

The concept of the pseudo amino acid composition or Chou's PseAAC (Pseudo Amino Acid Composition) was proposed in 2001 and has been rapidly applied in all fields of computational biology [16]. To learn the detailing introduction of Chou's PseAAC and its recent development and applications, we can see the comprehensive review in [17]. In this work, SNC, DNC and tri-nucleotides composition (TNC) were employed to extract sequence information, which were formulated as follows, respectively:

$$NC(i) = \frac{Total\ number\ of\ nucleotide\ (i)}{The\ length\ of\ sequence} \times 100 \quad (2)$$

where $i \in \{A,\ C,\ G,\ U\}$ and the length of the sequence is the number of nucleotides in this

$$DNC(j) = \frac{Total\ number\ of\ dinucleotide(j)}{The\ length\ of\ sequence - 1} \times 100,\ j = 1,\ 2,\ \ldots L,\ 16 \quad (3)$$

where $j \in \{AA, AC, \ldots L, GT, TT\}$ sequences;

$$TNC(k) = \frac{Total\ number\ of\ dinucleotide(k)}{The\ length\ of\ sequence - 2} \times 100,\ k = 1,\ 2,\ \ldots, 64 \tag{4}$$

where $k\ s \in \{AAA,\ AAC,\ \ldots,\ UUG,\ UUU\}$ equences.

2.2.2. Split-Position-Specific Matrix

It has been indicated in [13] that N- and C-terminal segments are critical to piRNAs. In this study, we considered the amino acids distribution both in N- and C-terminals. Moreover, the number of selected nucleotides of N- and C-terminal segments decreased successively, which were set to 15, 13, 11, 9, 7 and 5 nucleotides, respectively. Then, the bi-profile Bayes (BPB) features were used to characterize the probability distributions of each nucleotide at each position. As reported in many studies [18–21], the BPB feature vector was formulated as follows:

$$P = (x_1,\ x_2, \ldots,\ x_n,\ x_{n+1},\ \ldots, x_{2n}) \tag{5}$$

where P is the posterior probability vector; $x_1, x_2, \ldots, x_n$ represent the posterior probability of each nucleic acid at each position in positive peptide sequence datasets, respectively; and $x_{n+1}, \ldots, x_{2n}$ represent the posterior probability of each nucleic acid at each position in negative datasets, respectively.

2.2.3. Six RNA Dimer's Physicochemical Properties

Six RNA dimer's physicochemical properties, including rise, roll, shift, slide, tilt, and twist, have been used in [13], and have shown decently the prediction performance. The normalized values of six physicochemical properties for 16 dimers were derived in [13], and we listed them in Table 1 for convenience. We mapped the RNA sequence to the following vector according to the physicochemical properties:

$$V = (f_{AA} \cdot Rise_{AA},\ f_{AA} \cdot Roll_{AA},\ \cdots,\ f_{UU} \cdot Rise_{UU},\ f_{UU} \cdot Roll_{UU}) \tag{6}$$

Then, we obtained a physicochemical property vector with 96 dimensions (96-D).

Table 1. The Normalized Values of Rise, Roll, Shift, Slide, Tilt, and Twist for the 16 Dinucleotides in RNA.

	Physicochemical Property					
Dimer	**Rise**	**Roll**	**Shift**	**Slide**	**Tilt**	**Twist**
AA	−0.862	−0.689	−1.163	1.386	−1.896	−0.27
AC	−0.149	−1.698	1.545	0.51	0.555	0.347
AG	0.565	0	−0.813	0.127	0.096	−0.888
AU	−0.149	−0.643	−0.988	0.894	1.015	0.965
CA	−1.931	0.643	0.497	0.346	0.862	−0.27
CC	0.802	0.092	−0.551	−0.1407	−0.211	0.347
CG	0.565	1.652	2.156	−2.009	−0.823	−2.741
CU	0.565	0	−0.813	0.127	0.096	−0.888
GA	1.515	0.413	0.147	−0.969	1.321	0.347
GC	−0.386	−1.102	0.147	0.729	−0.67	2.201
GG	0.802	1.652	−0.551	−1.407	−0.211	0.347
GU	−0.149	−1.698	1.545	0.51	0.555	0.347
UA	0.089	1.01	−0.639	0.401	−0.977	0.347
UC	1.515	0.413	0.147	−0.969	1.321	0.347
UG	−1.931	0.643	0.497	0.346	0.862	−0.27
UU	−0.862	−0.689	−1.163	1.386	−1.896	−0.27

2.2.4. Feature Optimization

When four types of features were incorporated to train the model, the dimension of the hybrid features vector was 240. Additionally, the initial combined features may contain redundant and noisy information. This might exert the negative effect on model training. In this work, the importance of each feature was ranked by a feature selection tool known as F-score. The F-score of the j-th feature was defined as:

$$\mathrm{F-score}(j) = \frac{\left(\overline{x}_j^{(+)} - \overline{x}_j\right)^2 + \left(\overline{x}_j^{(-)} - \overline{x}_j\right)^2}{\frac{1}{m^+-1}\sum_{k=1}^{m^+}\left(\overline{x}_{k,j}^{(+)} - \overline{x}_j^{(+)}\right)^2 + \frac{1}{m^--1}\sum_{k=1}^{m^-}\left(\overline{x}_{k,j}^{(-)} - \overline{x}_j^{(-)}\right)^2} \quad (7)$$

where $\overline{x}_j$, $\overline{x}_j^{(+)}$ and $\overline{x}_j^{(-)}$ are the average values of the j-th feature in whole, positive and negative datasets, respectively, m^+ denotes the number of positive data, m^- denotes the number of negative data, $\overline{x}_{k,j}^{(+)}$ denotes the j-th feature of the k-th positive instance and $\overline{x}_{k,j}^{(-)}$ denotes the j-th feature of the k-th negative instance. The greater F-score indicates that the feature is more different between positive and negative samples, and is useful to classification [22].

2.3. SVM Implementation and Parameter Selection

SVM is a set of related supervised learning methods used for classification and regression based on the statistical learning theory, and has been illustrated to be powerful in many areas of bioinformatics [23–27]. As in other works [28–31], SVM was trained and tested by using the LIBSVM package [30] to build the model and implement the prediction. The radial basis function kernel was used in our SVM model. For different input features, the penalty parameter C and kernel parameter γ were optimized using SVMcg in the LIBSVM package based on a 5-fold cross-validation test. The optimal parameters C = 8 and γ = 0.044194 were set for the detection of piRNAs and non-piRNAs, while C = 0.35355 and γ = 1.4142 were assigned for the distinguishing samples with the function of instructing target mRNA deadenylation.

2.4. Model Construction and Evaluation

The performance of 2L-piRNAPred was evaluated using four measurements derived based on the symbols introduced by Chou in predicting signal peptides. Particularly, its advantages have been analyzed and endorsed by a series of studies published very recently [28–30]. The four measurements were given as follows:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \quad (8)$$

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

$$Sp = \frac{TN}{FP + TN} \quad (10)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (11)$$

where TP represents the number of true positives, TN represents the number of true negatives, FP represents the number of false positives and FN represents the number of false negatives.

3. Results and Discussion

The test process of our model piRNAPred is summarized in Algorithm 1.

Algorithm 1: The Predictive Algorithm for piRNAPred.

Input: $X = \{x_i\}_{i=1}^{N}$ is a set of samples and the number of categories is c.
Output: The prediction label of each sample.
For each X_i, $i = 1, 2, \ldots, N$ do
Take X_i as the test sample, and the others as the training dataset.
Extract features.
Predict the category.
end

3.1. Window Size Optimization for Bi-Profile Bayes

For the number of selected nucleotides, both N- and C-terminals affect the prediction performance, so we tried different window sizes to find the optimal prediction performance. The detailed results for the different window sizes are listed in Table S1 (Supplementary Materials). The highest *Acc* 82.81% was achieved at a window size of 15. Considering the minimum length in the whole training dataset was 24, we selected the maximum window size of 15. If we chose the window size that was much greater, there would be more repeated nucleic acids in the selected N-terminal and C-terminal segments.

3.2. Feature Selection for the First Layer

We first picked out the most contributing characteristics with a step of 2 for the other three types of characteristics, except for SNC with a step of 1. We decided which characteristics were retained based on the average results of 5-fold cross-validation test running 5 times for both the first and the second layers. To avoid wordiness, we only described the process of feature selection for the first layer. For the characteristic of SNC, the best prediction performance was achieved at all the four features (Table S2, Supplementary Materials). For the DNC, we sorted the 16 values according to the F-scores, and then selected 2, 4, . . . , 14, 16 features step by step. As listed in Table S3, the best performance achieved at 8 di-nucleotides selected (i.e., CG, UG, GA, AG, UA, CC, UU, and AU), with an *Sn* of 86.55%, an *Sp* of 82.05%, an *Acc* of 84.29%, and an *MCC* of 0.6870. The same processes were made for TNC and physicochemical properties (See Tables S4 and S5 (Supplementary Materials) for detail description) and the best performance parameters for these features are listed in Table 2. As we can see from Table 2, the DNC features achieved the best performance with an *Acc* of 84.29% and an *MCC* of 0.6870; on the contrary, the SNC feature achieved the worst prediction results with an *Acc* of 62.96% and an *MCC* of 0.2791. At last, all the selected characteristic vectors were combined together to get the higher results with an *Acc* of 88.95% and an *MCC* of 0.7791. Both the *Sn* and *Sp* were satisfactory.

Table 2. The best prediction achieved by single type of features for the first layer.

Cross-Validation	Features	Dimension	*Sn* (%)	*Sp* (%)	*Acc* (%)	*MCC*
5-fold	BPB	15	88.4	77.1	82.8	0.660
	SNC	4	45.4	80.62	63.0	0.279
	DNC	8	86.6	82.1	84.3	0.687
	TNC	56	85.3	82.6	83.9	0.678
	PP	84	84.5	79.1	81.8	0.636
	BPB + SNC + DNC + TNC + PP	-	90.4	87.5	89.0	0.779
Jackknife	BPB + SNC + DNC + TNC + PP	-	90.4	87.9	89.2	0.784

3.3. Feature Selection for the Second Layer

The same features selection processes for the second layer were performed as those for the first layer (see Tables S6–S10, Supplementary Materials for the detail). As can be seen from Table 3, using the unique type feature of TNC, we achieved the best *Acc* of 77.7% and *MCC* of 0.554, and the worst performance again was achieved by SNC with an *Acc* of 71.4% and an *MCC* of 0.427. When we combined all the features, the *Acc* slightly increased to 78.7% and *MCC* slightly increased to 0.573.

However, the *Sp* was 77.3%, less than 77.60% achieved by using TNC only. From the analysis, the prediction results were not as satisfactory as those for the first layer. Therefore, we had to add another type of features to further improve the prediction performance. Inspired by the *k* nearest neighbor (KNN) classification algorithm, we used KNN to embody the distribution of neighbor sequences [32]. Unlike in the samples investigated above, the lengths of samples in S^{+}_{inst} and $S^{+}_{non-inst}$ were not the same. To deal with this problem, the similar strategy was adopted to consider 10 nucleotides in N-terminal and C-terminal separately. The algorithm of using KNN to extract features is described in detail in [33]. The KNN features (k = 10) of N-terminal were firstly added to the original combination of BPB(15) + SNC(4) + DNC(10) + TNC(48) + PP(24), which increased the *Sp* to 79.3%, and *Sn* to 80.1%. Then, the KNN features (k = 10) of C-terminal were further added, which increased the *Sp* to 83.6%, and *Sn* to 84.3%.

Table 3. The best prediction achieved by single type of features for the second layer.

Cross-Validation	Features	Dimension	*Sn* (%)	*Sp* (%)	*Acc* (%)	*MCC*
5-fold	BPB	15	73.0	73.0	72.9	0.459
	SNC	4	69.2	73.6	71.4	0.427
	DNC	10	75.2	73.6	74.3	0.488
	TNC	48	77.8	77.6	77.7	0.554
	PP	24	74.7	74.2	74.3	0.489
	BPB + SNC + DNC + TNC + PP	101	80.0	77.3	78.7	0.573
	BPB + SNC + DNC + TNC + PP + KNN (N-terminal)	111	80.1	79.3	79.8	0.598
	BPB + SNC + DNC+ PP + KNN (N- and C-terminals)	121	84.3	83.6	84.0	0.68
Jackknife	BPB + SNC + DNC + PP + KNN (N- and C-terminals)	121	85.1	83.2	84.1	0.683

3.4. Performance of 2L-piRNAPred

One way to prove the superiority of the new model is to compare its prediction performance with that obtained by other existing methods. The compared results are listed in Table 4. For the first layer, our model achieved the best values of *Sn* and *Sp* among the four methods. While for the second layer, our model also achieved the best *Sn* and *Sp* values if our model piRNAPred was compared with the first two-layer model named 2L-piRNA. We noted that the increase in *Sp* value was more significant than that in *Sn* value both for the first and the second layers. Next, we further implemented the jackknife validation test. From the last line in Tables 2 and 3, we showed that our model piRNAPred obtained the *MCC* value of 0.784 for the first layer and the *MCC* value of 0.683 for the second layer. To further rank the classification methods for the first layer, a Friedman or a Friedman Aligned Ranks test (number of datasets: <20) with the Holm post-hoc test [34] was performed using [R]. Among all predictors, piRNAPred was significantly better than other approaches according to the MCC measurement. These comparisons illustrate the stability and robustness of our model piRNAPred. The reason why the methodology works well is the comprehensive features we extracted. The features reflected the global and local information of the samples in the dataset. Moreover, the running time for predicting one sample was about one second.

Table 4. The best prediction achieved by single type of features for the first layer.

Methodology	*Sn* (%)	*Sp* (%)	*Acc* (%)	*MCC*
First Layer				
piRNAPred	90.4	87.5	89	0.779
2L-piRNA	88.3	83.9	86.1	0.723
Accurate piRNA prediction	83.1	82.1	82.6	0.651
GA-WE	90.6	78.3	84.4	0.694
Second Layer				
piRNAPred	84.3	83.6	84	0.68
2L-piRNA	79.1	76	77.6	0.552

3.5. Comparison for Different Classifiers

In order to test whether the prediction performance of piRNAPred can be further improved, we tried different classifiers on the jackknife test. The prediction performances of Random Forest (RF), *KNN* and Ensemble for Boosting are listed in Table 5. It is illustrated that both for the first layer and the second layer, SVM achieved the best performance. Therefore, our predictor adopted SVM as the final classifier.

Table 5. The performances of different classifiers on the jackknife test.

Method	*Sn* (%)	*Sp* (%)	*Acc* (%)	*MCC*
First Layer				
SVM	90.4	87.5	89	0.779
RF	85.8	88.4	87.1	0.743
KNN	88.7	83.6	86.1	0.724
Ensemble	89.9	87.0	88.5	0.770
Second Layer				
SVM	84.3	83.6	84.0	0.680
RF	72.9	72.8	72.9	0.457
KNN	73.3	69.7	71.5	0.431
Ensemble	75.9	73.6	74.8	0.495

4. Conclusions

In this work, we proposed a computational method for identifying piRNAs with the function of instructing target mRNA deadenylation and piRNAs without the function of instructing target mRNA deadenylation. According to the average outcomes of 5-fold cross-validation test for running 10 times, the combination of BPB(15) + SNC(4) + DNC(8) + TNC(56) + PP(84) achieved the best *Sn*, *Sp*, *Acc*, and *MCC* values for the first layer. While for the second layer, it was a bit complex. The original combination BPB(15) + SNC(4) + DNC(10) + TNC(48) + PP(24) must contain KNN (k = 10) features in N-terminal and C-terminal to get satisfactory *Sn*, *Sp*, *Acc*, and *MCC* average results of 5-fold cross-validation test for 10 times. Moreover, the comparison between the jackknife and 5-fold cross-validation outcomes shows the robustness of 2L-piRNAPred. It should be pointed that the predictor was not tested on an independent dataset, and the prediction performance might cause a certain overfitting.

Supplementary Materials: The following are available online at http://www.mdpi.com/2076-3417/8/11/2204/s1, Table S1: The average 5-fold cross-validation results using split-BPB on different window sizes both on N- and C- terminal for the first layer, Table S2: Features selection process for SNC by F-scores for the first layer, Table S3: Features selection process for DNC by F-scores for the first layer, Table S4: Features selection process for TNC by F-scores for the first layer, Table S5: Features selection process for physicochemical properties by F-scores for the first layer, Table S6: The average 5-fold cross-validation results using split-BPB on different window sizes both on N- and C- terminal for the second layer, Table S7: Features selection process for SNC by F-scores for the second layer, Table S8: Features selection process for DNC by F-scores for the second layer, Table S9: Features selection process for TNC by F-scores for the second layer, Table S10: Features selection process for physicochemical properties by F-scores for the second layer.

Author Contributions: T.L. conceived and designed the experiments; M.G. and Y.C. implemented SVM and created the webserver; T.L. performed the analysis and wrote the paper; R.S. and Q.Y. revised the paper. All authors read and approved the final manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (grant No. 71271034), the National Social Science Foundation of China (grant No. 15CGL031), the Fundamental Research Funds for the Central Universities (grant No. 3132016306, 3132018160, 3132018227), the Program for Dalian High-Level Talent Innovation Support (grant No. 2015R063), the National Natural Science Foundation of Liaoning Province (grant No. 20180550307 and 20180550223), and the National Scholarship Fund of China for Studying Abroad.

Acknowledgments: We are deeply grateful to the anonymous referees and to the editors for their suggestion and help in significantly improving the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Aravin, A.; Gaidatzis, D.; Pfeffer, S.; Lagos-Quintana, M.; Landgraf, P.; Iovino, N.; Morris, P.; Brownstein, M.J.; Kuramochi-Miyagawa, S.; Nakano, T.; et al. A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* **2006**, *442*, 203–207. [CrossRef] [PubMed]
2. Grivna, S.T.; Beyret, E.; Wang, Z.; Lin, H. A novel class of small RNAs in mouse spermatogenic cells. *Gene Dev.* **2006**, *20*, 1709–1714. [CrossRef] [PubMed]
3. Grivna, S.T.; Pyhtila, B.; Lin, H. MIWI associates with translational machinery and PIWI–interacting RNAs (piRNAs) in regulating spermatogenesis. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 13415–13420. [CrossRef] [PubMed]
4. Goh, W.S.; Falciatori, I.; Tam, O.H.; Burgess, R.; Meikar, O.; Kotaja, N.; Hammell, M.; Hannon, G.J. piRNA–directed cleavage of meiotic transcripts regulates spermatogenesis. *Gene Dev.* **2015**, *29*, 1032–1044. [CrossRef] [PubMed]
5. Gong, S.H. Identification and verification of potential piRNAs from domesticated yak testis. *Reproduction* **2018**, *155*, 117–127. [CrossRef] [PubMed]
6. Zhang, D.; Tu, S.; Stubna, M.; Wu, W.S.; Huang, W.C.; Weng, Z.; Lee, H.C. The piRNA targeting rules and the resistance to piRNA silencing in endogenous genes. *Science* **2018**, *359*, 587–592. [CrossRef] [PubMed]
7. Svendsen, J.M.; Montgomery, T.A. piRNA Rules of Engagement. *Dev. Cell* **2018**, *4*, 657–658. [CrossRef] [PubMed]
8. Wu, W.S.; Huang, W.C.; Brown, J.S.; Zhang, D.; Song, X.; Chen, H.; Tu, S.; Weng, Z.; Lee, H.C. pirScan: A webserver to predict piRNA targeting sites and to avoid transgene silencing in *C. elegans*. *Nucleic Acids Res.* **2018**, *46*, W43–W48. [CrossRef] [PubMed]
9. Zhang, Y.; Wang, X.; Kang, L. A *k*–mer scheme to predict piRNAs and characterize locust piRNAs. *Bioinformatics* **2011**, *27*, 771–776. [CrossRef] [PubMed]
10. Wang, K.; Liang, C.; Liu, J.; Xiao, H.; Huang, S.; Xu, J.; Li, F. Prediction of piRNAs using transposon interaction and a support vector machine. *BMC Bioinform.* **2014**, *15*, 419. [CrossRef] [PubMed]
11. Luo, L.; Li, D.; Zhang, W.; Tu, S.; Zhu, X.; Tian, G. Accurate prediction of transposon–derived piRNAs by integrating various sequential and physicochemical features. *PLoS ONE* **2016**, *11*, e0153268. [CrossRef] [PubMed]
12. Li, D.; Luo, L.; Zhang, W.; Liu, F.; Luo, F. A genetic algorithm–based weighted ensemble method for predicting transposon–derived piRNAs. *BMC Bioinform.* **2016**, *17*, 329. [CrossRef] [PubMed]
13. Liu, B.; Yang, F.; Chou, K.C. 2L–piRNA: A Two–Layer Ensemble Classifier for Identifying Piwi–Interacting RNAs and Their Function. *Mol. Ther. Nucleic Acids* **2017**, *16*, 267–277. [CrossRef] [PubMed]
14. Zhang, P.; Si, X.; Skogerbø, G.; Wang, J.; Cui, D.; Li, Y.; Sun, X.; Liu, L.; Sun, B.; Chen, R.; et al. piRBase: A web resource assisting piRNA functional study. *Database* **2014**, *2014*, 110. [CrossRef] [PubMed]
15. Bu, D.; Yu, K.; Sun, S.; Xie, C.; Skogerbø, G.; Miao, R.; Xiao, H.; Liao, Q.; Luo, H.; Zhao, G.; et al. NONCODE v3. 0: Integrative annotation of long noncoding RNAs. *Nucleic Acids Res.* **2012**, *40*, D210–D215. [CrossRef] [PubMed]
16. Brett, T.; Anthony, K. Computational phosphorylation site prediction in plants using random forests and organism-specific instance weights. *Bioinformatics* **2013**, *29*, 686–694.
17. López, Y.; Dehzangi, A.; Lal, S.P.; Taherzadeh, G.; Michaelson, J.; Sattar, A.; Tsunoda, T.; Sharma, A. SucStruct: Prediction of succinylated lysine residues by using structural properties of amino acids. *Anal. Biochem.* **2017**, *527*, 24–32. [CrossRef] [PubMed]
18. Liu, B.; Liu, F.; Wang, X.; Chen, J.; Fang, L.; Chou, K.C. Pse–in–One: A web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.* **2015**, *43*, W65–W71. [CrossRef] [PubMed]
19. Shao, J.; Xu, D.; Tsai, S.N.; Wang, Y.; Ngai, S.M. Computational identification of protein methylation sites through bi–profile Bayes feature extraction. *PLoS ONE* **2009**, *4*, e4920. [CrossRef] [PubMed]
20. Song, J.; Tan, H.; Shen, H.; Mahmood, K.; Boyd, S.E.; Webb, G.I.; Akutsu, T.; Whisstock, J.C. Cascleave: Towards more accurate prediction of caspase substrate cleavage sites. *Bioinformatics* **2010**, *26*, 752–760. [CrossRef] [PubMed]
21. Jia, C.; Liu, T.; Chang, A.K.; Zhai, Y. Prediction of mitochondrial proteins of malaria parasite using bi–profile Bayes feature extraction. *Biochimie* **2011**, *93*, 778–782. [CrossRef] [PubMed]

22. Jia, C.Z.; Zuo, Y.; Zou, Q. O–GlcNAcPRED–II: An integrated classification algorithm for identifying O–GlcNAcylation sites based on fuzzy undersampling and a K–means PCA oversampling technique. *Bioinformatics* **2018**, *34*, 2029–2036. [CrossRef] [PubMed]
23. Senawi, A.; Wei, H.L.; Billings, S.A. A new maximum relevance–minimum multicollinearity (MRmMC) method for feature selection and ranking. *Pattern Recognit.* **2017**, *67*, 47–61. [CrossRef]
24. Chen, L.; Chen, W.; Cheng, Q.; Wu, Y.; Krishnan, S.; Zou, Q. LibD3C: Ensemble Classifiers with a Clustering and Dynamic Selection Strategy. *Neurocomputing* **2014**, *123*, 424–435.
25. Li, S.; Li, D.; Zeng, X.X.; Wu, Y.F.; Li, G.; Zou, Q. nDNA–prot: Identification of DNA–binding Proteins Based on Unbalanced Classification. *BMC Bioinform.* **2014**, *15*, 298.
26. Li, D.; Ju, Y.; Zou, Q. Protein Folds Prediction with Hierarchical Structured SVM. *Curr. Proteom.* **2016**, *13*, 79–85. [CrossRef]
27. Wei, L.Y.; Tang, J.J.; Zou, Q. Local–DPP: An Improved DNA–binding Protein Prediction Method by Exploring Local Evolutionary Information. *Inf. Sci.* **2017**, *384*, 135–144. [CrossRef]
28. Chang, C.C.; Lin, C.J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 27. [CrossRef]
29. Meher, P.K.; Sahu, T.K.; Saini, V.; Rao, A.R. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. *Sci. Rep.* **2017**, *7*, 42362. [CrossRef] [PubMed]
30. Farman, A.; Maqsood, H. Unb-DPC: Identify mycobacterial membrane protein types by incorporating un-biased dipeptide composition into Chou's general PseAAC. *J. Theor. Biol.* **2015**, *384*, 78–83.
31. Rahimi, M.; Bakhtiarizadeh, M.R.; Mohammadi-Sangcheshmeh, A. OOgenesis_Pred: A sequence-based method for predicting oogenesis proteins by six different modes of Chou's pseudo amino acid composition. *J. Theor. Biol.* **2017**, *415*, 13–19. [CrossRef] [PubMed]
32. Chen, X.; Qiu, J.D.; Shi, S.P.; Suo, S.B.; Huang, S.Y.; Liang, R.P. Incorporating key position and amino acid residue features to identify general and species–specific Ubiquitin conjugation sites. *Bioinformatics* **2013**, *29*, 1614–1622. [CrossRef] [PubMed]
33. Jia, C.Z.; Zhang, J.J.; Gu, W.Z. RNA–MethylPred: A high–accuracy predictor to identify N6-methyladenosine in RNA. *Anal. Biochem.* **2016**, *510*, 72–75. [CrossRef] [PubMed]
34. Rodríguez-Fdez, I.; Canosa, A.; Mucientes, M.; Bugarín, A. STAC: A web platform for the comparison of algorithms using statistical tests. In Proceedings of the 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Istanbul, Turkey, 2–5 August 2015; pp. 1–8.