Genetics and Molecular Research

Detection of Piwi-interacting RNAs based on sequence features

Y.J. Liu[1], J.Y. Zhang[1], A.M. Li[2], Z.W. Liu[1], Y.Y. Zhang[1] and X.H. Sun[1]

[1]School of Computer Science and Technology, Xidian University, Xi'an, China
[2]School of Computer Science and Technology, Xi'an University of Technology, Xi'an, China

Corresponding author: J.Y. Zhang
E-mail: jyzhang@mail.xidian.edu.cn

Genet. Mol. Res. 15 (2): gmr.15028638
Received March 23, 2016
Accepted April 11, 2016
Published May 13, 2016
DOI http://dx.doi.org/10.4238/gmr.15028638

ABSTRACT. Piwi-interacting RNAs (piRNAs) are a class of small non-coding RNAs. Distinguishing piRNAs from other non-coding RNAs is important because of their important role in the physiological regulation of spermatogenesis, genome protection from transposons, and regulation of mRNAs and long non-coding RNAs. Few computational studies have addressed piRNAs detection, and both effectiveness and efficiency of piRNA detection tools require improvement. In this study, a piRNA detection method based on sequence features and a support vector machine was developed. Four types of features are proposed: weighted k-mer, weighted k-mer with wildcards, position-specific base, and piRNA length. The piRNA sequences from human, mouse, rat, and drosophila were respectively used in this experiment. Compared to existing algorithms, the proposed method provides a better balance between precision and sensitivity (both are approximately 90%), and although these values were slightly slower than those obtained using the piRNA annotation approach, the proposed method was four-fold faster than piRPred and 229-fold faster than piRNA predictor.

Key words: PiRNA; Detection; Weighted k-mer with wildcards; Weighted k-mer

INTRODUCTION

Piwi-interacting RNAs (piRNAs) are a class of newly discovered small non-coding RNAs (ncRNAs) that are approximately 24-33 nucleotides (nt) in length. They are primarily expressed in the germline and bind specifically to Piwi-class proteins rather than Ago-class Argonaute proteins. Similar to micro-RNAs, the 5' ends of piRNAs also exhibit a strong uracil bias (Aravin et al, 2006).

piRNAs play an important role in the physiological regulation of spermatogenesis, genome protection from transposons, and regulation of mRNAs and long non-coding RNAs (Watanabe et al., 2015). Only 17-20% of mammalian piRNA repeat sequences correspond to transposons and retrotransposons (Houwing et al., 2007). Limited computational studies have addressed piRNAs detection because of the lack of efficient secondary structure motifs and sequence homology in different species (Zhang et al., 2011). Furthermore, given the large size of ncRNAs, it is challenging to effectively and efficiently identify piRNAs (Reuter et al., 2011).

Finding features is a key point for the identification. Betel et al. (2007) found that mouse piRNAs have some position-specific properties (e.g. guanine or adenine at +1 position). They used these features to identify mouse piRNAs; however, the precision of this approach was only 61-72%.

The ultimate goal for identification is reaching an optimum balance between precision and sensitivity. Zhang et al. (2011) developed a piRNA predictor (piRNA predictor) based on the k-mer scheme and Fisher discriminant analysis. With a precision of over 90%, this approach is clearly superior to that proposed by Betel et al. However, the sensitivity of this method is not satisfactory (60-70%; except for fruit flies). Furthermore, the approach is both time- and space-consuming. Using 1364 features and Fisher discriminant analysis, the covariance matrices of piRNAs and non-piRNA groups must be estimated. Additionally, piRNAs are short; thus, most k-mer terms are 0. As a result, the feature vectors based on k-mer terms are sparse. For covariance matrix precision, a large-scale training set (173,090 positive samples and 193,321 negative samples) was employed.

Brayet et al. (2014) proposed a new method, named piRPred, which also provides unbalanced drosophila data (95% specificity but only 83% sensitivity). This method is based on multiple kernels and a support vector machine (SVM) classifier. Three kernels are used, and each is a square similarity matrix of size N x N, where N is the size of the training dataset. As the size of the training set increases, the building process requires increasing memory and time.

Wang et al. (2014) proposed a piRNA annotation approach (Piano), which shows a good balance between specificity and sensitivity; however, the approach is limited to only piRNA sequences that are aligned to transposons. Although the approach achieves a specificity of 95% and sensitivity of 96%, such piRNA sequences comprise only a small subset of all piRNAs. The percentage of piRNAs studied was only 9758/13,848 = 70.4% for *Drosophila melanogaster*, 7140/ 32,152 = 22.2% for human, 14,495/75,814 = 19.1% for mouse, and 14,195/66,758 = 21.2% for rat piRNAs.

It is important to establish an effective and efficient method for identifying piRNAs. Our study aimed to provide a better balance between precision and sensitivity as well as increase efficiency. New sequential features are proposed and an SVM is applied. Using a 5-fold cross-validation, our study provides approximately 90% precision and 90% sensitivity. Our method is slightly slower than Piano, but it is 4-fold faster than piRPred and 229-fold faster than piRNA predictor.

MATERIAL AND METHODS

Overview of methodology

piRNAs are distinguished from other small non-coding RNAs (ncRNAs) by extracting discriminate features to characterize the differences between piRNA and the ncRNA. A classifying model is then designed in the feature space to facilitate detection. The proposed method consists of two main aspects: feature extraction and classifying model construction (Figure 1). Feature extraction is more significant than classification because no discriminant features are extracted and no good detection performance can be obtained even when using the best classifying model. Thus, we focused on feature extraction in this approach while using the classification model as a standard conventional SVM.

In this study, piRNAs and non-piRNA sequences smaller than 50 nt were considered. Non-piRNA sequences larger than 50 nt were not considered because piRNAs are not larger than 50 nt. The sequences were transformed into feature vectors. Four types of features were considered: weighted k-mer scheme, weighted k-mer with wildcards, position-specific properties, and sequence length. An unmanageable number of features was originally obtained; therefore, this number was reduced to only 200 feature items by maximum relevance minimum redundancy (mRMR) (Peng et al., 2005). These features were used for constructing the SVM classifying model for detection. Here, the SVM classifier is trained using only a single species' piRNAs at a time. Four typical species are used (drosophila, human, mouse, and rat), and a user can choose one model according to the species information or build a new model using our approach.

In the classification process, any sequence smaller than 50 nt in the testing set was considered to be a piRNA candidate and thus was transformed into a feature vector. The trained model was used to classify the ncRNA sequences in the feature space into piRNAs or non-piRNAs.

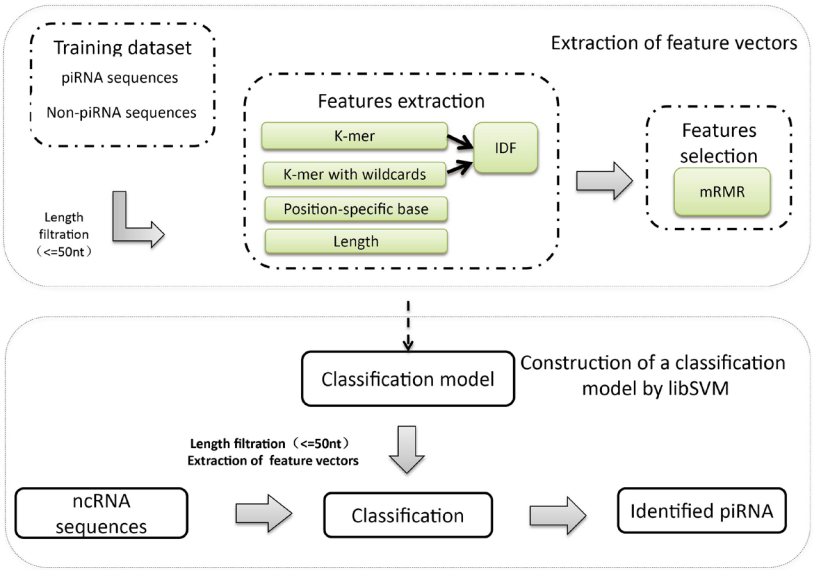


Figure 1. Framework illustration. The top panel details the feature vector extraction process, and the bottom panel illustrates the classification process.

Training dataset and length filter

Positive samples were downloaded from piRNA Bank (Sai Lakshmi and Agrawal, 2008) and divided into four datasets according to their species: 22,336 drosophila piRNAs, 23,439 human piRNAs, 39,986 mouse piRNAs, and 38,549 rat piRNAs. Non-piRNA sequences (negative samples) were obtained from Zhang et al. (2011). Because there are far less non-piRNA small ncRNAs in one species than piRNAs, only 34,675 real non-piRNA sequences from 861 organisms are contained in NONCODE (Bu et al., 2011), which was collected by Zhang et al. (2011). The remaining 158,646 sequences are well-designed and were generated by random processes according to real data, which was developed by Zhang et al. (2011). Furthermore, most non-piRNA sequences are considerably longer than piRNA sequences; therefore, if a sequence is too long, it is clearly not a piRNA and should be removed. In consideration of the maximum length of positive sets (38 nt) and possible sequencing errors, 50 nt was selected as the maximum tolerant length.

Feature extraction

K-mer term frequency (TF) (Burge et al., 1992; Gutiérrez et al., 1993; Karlin and Ladunga, 1994) typically refers to the specific k-grams frequency of nucleic acid or amino acid sequences and has been widely used to characterize biosequences. TFs can be used to identify certain regions of biomolecules such as DNA or proteins (Zhang et al., 2011). Furthermore, it has been determined that TFs are species- or taxon-specific (Karlin et al., 1994; Karlin and Mrázek, 1997; Madera et al., 2010).

Each piRNA is composed of four nucleobases: guanine (G), adenine (A), thymine (T), and cytosine (C). A sliding window of size k was used to scan each piRNA sequence from 5' to 3'. The frequency of each term was recorded using the moving window. In order to save storage space, k from 1 to 4 was used, so 1-4 nt terms were considered (including 4 1-mers, 16 2-mers, 64 3-mers, 256 4-mers, for a total of 340 terms).

Although TF has been widely used, a high TF value generally cannot reflect high discriminability. To illustrate this point, we describe an English email filtering problem. Suppose we have a collection of English emails (text documents) and wish to identify which email is an advertisement. Some terms are prevalent but cannot reflect true discriminability, e.g., "the" or "we". This indicates that a high TF value does not determine high discriminability. Thus, the aforementioned words are referred to as "stop words" in the fields of natural language processing. One solution is to filter these words out according to a specialist dictionary. Another solution is to introduce a new factor to weight the TF value; in this way, meaningless terms can be reduced. The factor is defined as inverse document frequency (or inverse collection frequency, short for IDF) (Jones, 1972), which favors terms concentrated in a few documents of a collection. A typical classical TF-IDF (Salton and Buckley, 1988) is defined below:

$$TFIDF(\mathrm{t},\mathrm{n},\mathrm{N}) = \mathrm{TF}(\mathrm{t}) \times \mathrm{IDF}(\mathrm{n},\mathrm{N}) = \mathrm{TF}(\mathrm{t}) \times \log\frac{N}{n}$$ (Salton and Buckley, 1988) (Equation 1)

where TF is the number of times term *t* occurs in a document, *N* is the total number of documents in collection, and *n* is the number of documents in which term *t* occurs.

As described by Salton and Buckley (1988), term discrimination considerations

suggest that the best terms for document content identification are those that can distinguish certain individual documents from the remainder of the collection. Hence, the IDF factor varies inversely with the number of documents *n* in which a term occurs in the whole collection of *N* documents, and a reasonable measure of term importance can be obtained by using the product of TF and IDF.

In bioinformatics, there is no specialist dictionary concerning piRNAs. Accordingly, the second solution was chosen for this study: in reference to the IDF theory, a factor was introduced to weight the k-mer term frequency to characterize piRNAs. This method is discussed below.

Weighted k-mer

We employed the k-mer term frequency as TF and a weighting factor to reflect k-mers' discriminability. However, classic IDF (Formula 1) may be problematic for piRNA detection; thus, specific issues must be addressed to improve piRNA detection.

Classic IDF uses the log ratio of N over n as the weight of a TF, where N is the total number of samples and n is the number of sequences in which the specific k-mer term occurs. By only considering the term distribution in the whole collection, classic IDF has an obvious drawback: it does not consider the term distributions of different classes in the whole collection; thus, variances in the term distributions of different classes are ignored. For example, a term always occurs in non-piRNAs but rarely occurs in piRNAs. Therefore, term discriminability is high, but the IDF calculated using Formula 1 is small. Another obvious case is that in which a term rarely appears in a whole collection, but its distribution in positive and negative sets is similar. In that case, the term is not suitable for classification, but its IDF is large.

By taking discriminability into consideration and referring to Fisher discriminant analysis (Belhumeur et al., 1997), it is hoped that there is a significant difference in the TF-IDF between the positive and negative sets and that the difference in each set is small. Therefore, for piRNA detection, a new IDF was designed in this study:

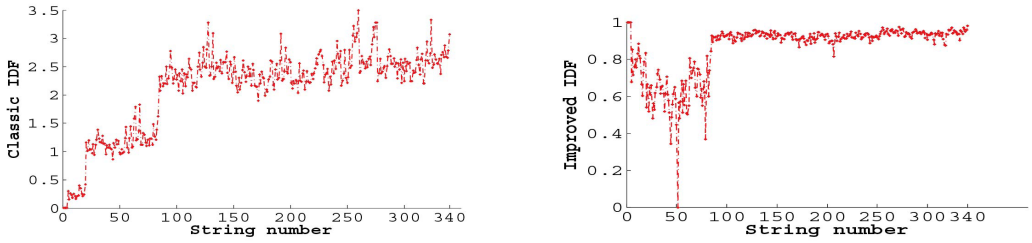
$$IDF(\mathrm{t}) = \log_2\left(\frac{|A - C|}{B + D + 1}\right) \qquad \text{(Equation 2)}$$

where A and B are the number of piRNA reads within which term t occurs and does not occur, respectively, and C and D are the number of non-piRNA reads within which term t occurs and does not occur, respectively. This information is listed in Table 1. As seen in Figure 2a and b, there was a clear difference between the classic IDF and the new IDF (normalized) of a piRNA.

Table 1. Terms distribution.

	Read number	
	Term t occurs	Term t does not occur
piRNA number	A	B
Non-piRNA number	C	D

It is also important to note that an unbalanced number of positive and negative samples will influence the IDF. To address this issue, positive and negative sample sizes were set to be the same as the training data in this study, and the samples were randomly selected from the dataset.



a b

Figure 2. a. Classic IDF of a piRNA. **b.** Weighted IDF (normalized) of a piRNA.

Weighted k-mer with wildcards

In order to enable flexibility as well as to avoid sequencing errors, a k-mer with wildcards was added as the second feature type. A term with a wildcard in the first or last place (e.g. "*A" and "A*") is analogous to a term without a wildcard (e.g. "A"); thus, only terms with wildcards in the middle were considered. As a result, 160 feature terms were selected: 16 3-mers with one wildcard (e.g. "A*C"), 16 4-mers with two wildcards (e.g. "A**C"), and 128 4-mers with one wildcard (e.g. "A*CG" and "AC*G"). Again, the IDF given in formula (2) was used as a factor to weight the k-mer with wildcards.

Position-specific base

The TF-IDF characterizes only TF frequency and does not consider specific positions; however, specific positions may be important for identifying piRNAs. In fact, Zhang et al. (2011) found 21 position-specific base usages that were significantly different between piRNA and non-piRNA. Thus, the position-specific properties were extracted as the third feature type.

Feature vectors were constructed by converting the starting 30-base sequence from both the 5' and 3' directions into 240-bit vectors (30 nt x 4 bases x 2 directions). Each nucleotide position was converted to a 4-bit vector representing the RNA base. In our method, 1 is the feature value if it is a specific base and 0 if it is not. Additionally, piRNAs are not uniform in length, so positions starting from the 5' and 3' directions are both considered, and the first 30 sequence positions are calculated to save storage space.

Length

There are numerous piRNAs, and their lengths generally range from 25-32 nt (Aravin et al., 2006; Girard et al., 2006; Seto et al., 2007; Siomi et al., 2011). MIWI-associated piRNAs are often 30-31 nt long (Girard et al., 2006), whereas MILI-associated piRNAs are 26-28 nt long (Seto et al., 2007). Both are much longer than siRNAs and micro-RNAs. In the analysis, the size distribution of the positive set varied greatly (18-38 nt), and the known piRNAs in

human, mouse, and rat were similar in length (Figure 3). Therefore, length was selected as a feature dimension.

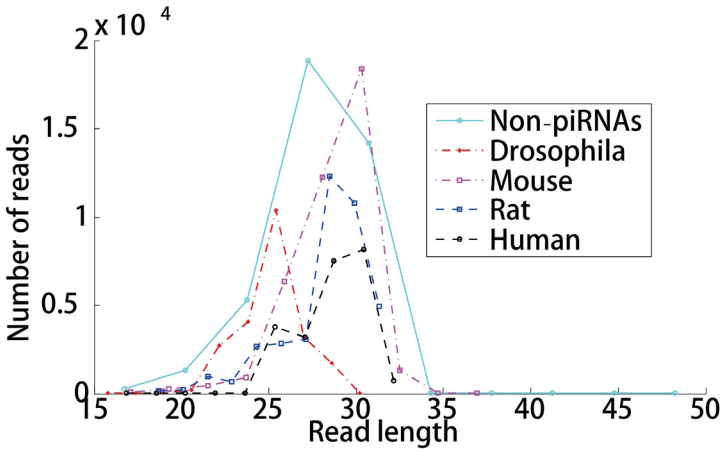


Figure 3. Length distributions in non-piRNA and in four species.

Feature selection and classifying model construction

As described above, one length feature, 340 k-mer feature terms, 160 k-mer feature terms with wildcards, and 240 position-specific base features were exacted. Overall, there were 741 features. However, it was unclear which features were closely related to classification. Therefore, mRMR was used for feature selection. Based on the results, the top 200 features for each species were retained and analyzed in section 3.2 "Selected features". The top 10 features are listed in Table 2.

Table 2. Top 10 features.

	Feature ranking									
	1	2	3	4	5	6	7	8	9	10
Drosophila	length	C*G	1T	-28A	-29A	29T	TC*G	29A	CTG	-29G
Human	C*G	1T	CTG	TG	CAG	length	TA	CG	G	C*GG
Rat	1T	CG	C*G	CTG	TG	CAG	length	TCG	CGA	CGT
Mouse	1T	CG	C*G	CTG	TG	CAG	length	1A	TCG	CGT

"1T" refers to T at the first position from 5' to 3'; '-28A' refers to A at the 28th position from 3' to 5'; and '*' represents wildcard.

The SVM is a powerful tool for binary classification. We choose libSVM 3.17 (Chang and Lin, 2011) to train the models using the standard radial basis function kernel. To optimize the SVM classifier, the C and gamma parameters were adjusted using the grid search strategy in libSVM. (In addition, grid.py is an optimizer of libSVM).

RESULTS AND DISCUSSION

Features analysis

Zhang et al. (2011) used the rank sum test to determine which string usage is

significantly different between piRNAs and non-piRNAs, and they found that using 1337 terms in 1-5mer (1364) was significant. The term frequency relative difference is:

$$fd_t = \frac{|f_{piRNA}(t) - f_{non-piRNA}(t)|}{f_{piRNA}(t) + f_{non-piRNA}(t)}$$ (Zhang et al., 2011) (Equation 3)

where t is a kmer term, and $f_{piRNA}(t)$ is the frequency of the term t as it appeared in piRNAs.

Among the k-mer terms, we evaluated whether there was a "stop word" in the sequence analysis. The same test was then used to determine which k-mer term was meaningless for identification. The frequencies of 340 k-mers in four species and non-piRNAs are shown in Figure 4. Based on the term frequency relative difference of the piRNAs and non-piRNA, three terms were determined to be meaningless for identification ("TTTG" with *fd* of 6.4134e-4; "GCTT" with *fd* of 4.7661e-4; "CAA" with *fd* of 3.2241e-4).

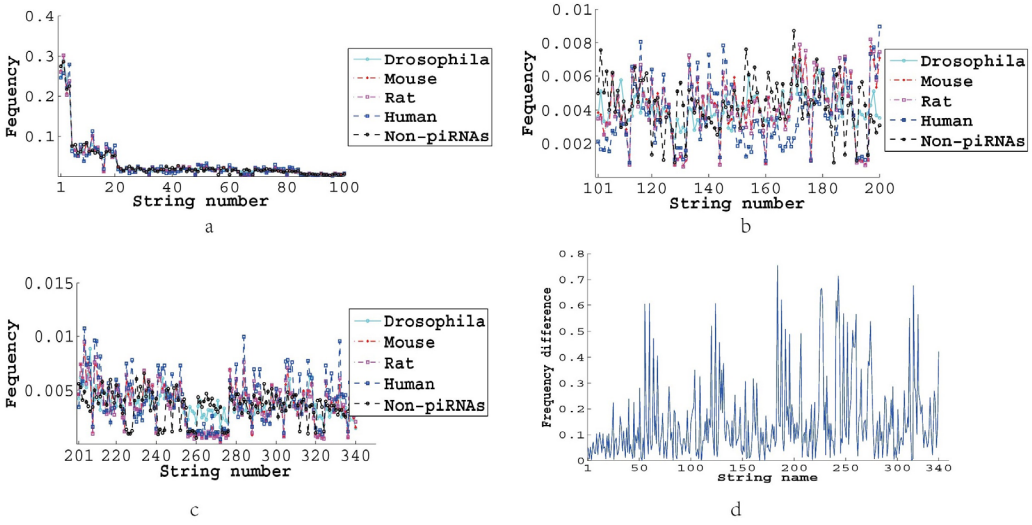


Figure 4. a. b. c. Average frequencies of 340 terms in the piRNAs of four species and non-piRNAs. **d.** Differences from comparing the frequencies of the piRNAs and non-piRNA.

For the piRNAs of four species and non-piRNA sequences, the frequencies of 160 k-mers with wildcards were calculated (Figure 5). By comparing non-piRNAs and piRNAs, the mouse's broken line was very similar to the rat's broken line, whereas non-piRNAs were significantly different from piRNAs. Hence, using k-mer with wildcards is a useful feature type.

The frequencies of four bases in the first 30 positions from 5' to 3' and from 3' to 5' were calculated for each species (Figure 6). The base distribution exhibited a strong bias for U at the 5' end, whereas an A at position 10 was a canonical feature of the piRNAs.

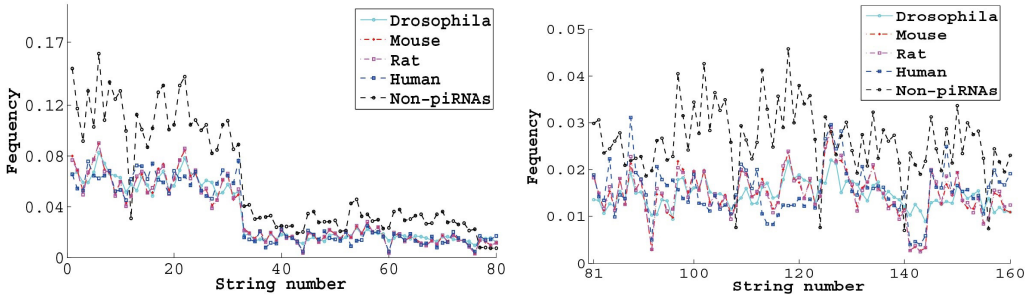


Figure 5. a. b. Average frequencies of 160 k-mers with wildcards in the piRNAs of four species and non-piRNAs.

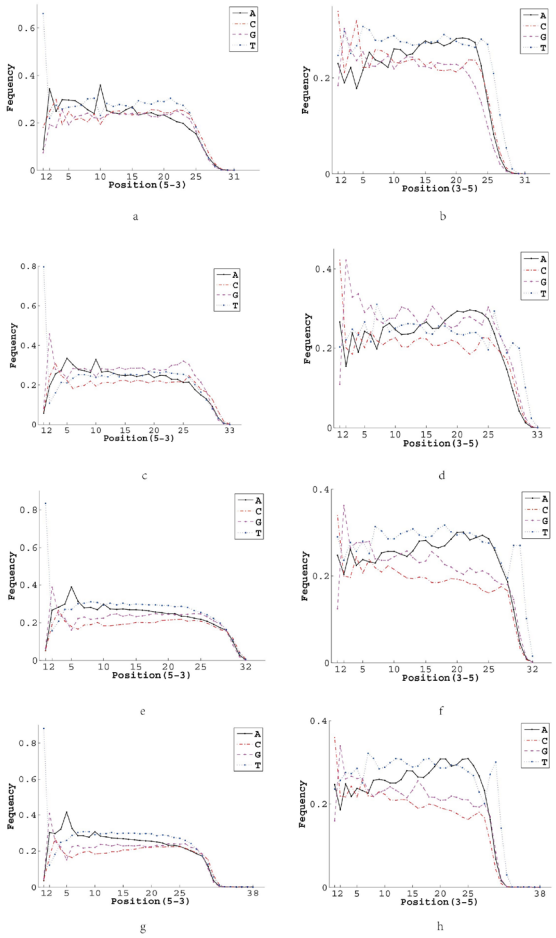


Figure 6. Frequencies of four bases in the first 30 position from 5' to 3' and from 3' to 5' in *Drosophila* (**a** and **b**), humans (**c** and **d**), rats (**e** and **f**), and mice (**g** and **h**).

Selected features

To improve efficiency, 200 features were selected by mRMR. In the feature rankings, length was sorted into the top six effect features for all four species.

The rankings of the four other feature types are shown in Figure 7. For the four species, k-mer was found to be the most useful feature type, which included more than half of the total feature items. The other types were effective to a certain extent. Of the 200 feature items, there were approximately 60 k-mers with wildcards and approximately 20 position-specific features.

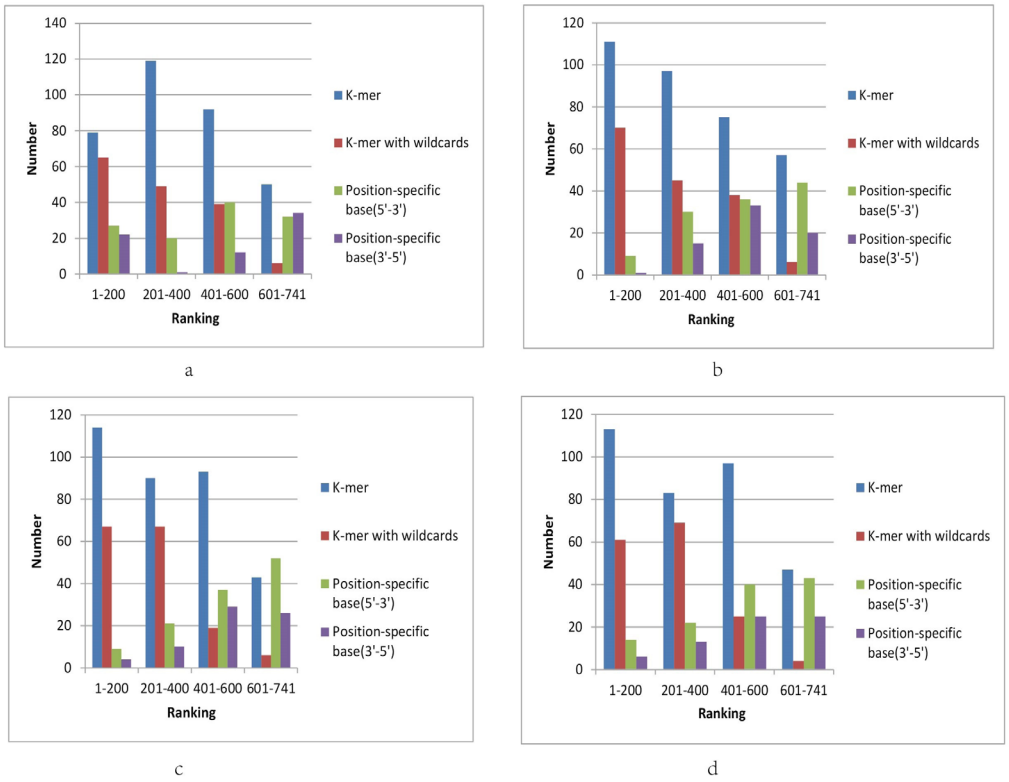


Figure 7. Distribution of feature rankings in *Drosophila* (**a**), humans (**b**), rats (**c**), and mice (**d**).

Cross-validation results

The piRNA sequences of the four species used in the experiment as the positive training set were as follows: 17,868 drosophila piRNAs, 18,751 human piRNAs, 31,989 mouse piRNAs, and 30,839 rat piRNAs. The remainder for each species was the validation set. Negative samples were randomly selected from the negative set, and the positive and negative sets were the same size. The precision and sensitivity of our method using a 5-fold cross-validation are shown in Table 3. All values were near 90%.

Table 3. Five-fold cross-validation results.

	Accuracy	Precision	Recall (Sensitivity)
	(TP + TN)/(TP + FN + FP + TN)	TP/(TP + FP)	TP/(TP + FN)
Drosophila	88.1%	86.7%	90.0%
Human	90.3%	90.1%	90.6%
Rat	88.6%	88.1%	89.3%
Mouse	89.5%	89.3%	89.7%

Comparison with existing tools

To compare the current algorithm's detection performance with those of existing tools, we investigated all algorithms on model species datasets. Four thousand stand-alone sequences were selected from single species at a time and 4000 negative samples were used. The experimental results were compared using different algorithms, and our algorithm showed the following advantages:

A better balance between precision and sensitivity

The detection performance was compared to those of the other tested algorithms (Table 4), and the results obtained using our method showed a better balance between precision and sensitivity. The piRNApredictor showed good precision but the lowest sensitivity. Furthermore, Piano is useful for recognizing piRNAs that target transposons, but it is not suitable for discovering other piRNAs. For example, among the 4000 drosophila piRNA sequences, 3240 piRNAs (81%) were aligned successfully and 2940 piRNAs (73.5%) were judged as positive samples. Additionally, piRPred produced higher precision than did our approach; however, it was much less sensitive.

Table 4. Comparison results.

		piRNApredictor	Piano	piRPred	Our method
Drosophila	Pre	95.0%	98.6%	94.3%	87.4%
	Se	40.1%	73.5%	83.6%	89.7%
Human	Pre	96.6%	94.9%	89.1%	89.0%
	Se	80.9%	24.6%	84.2%	86.1%

Pre = precision. Se = sensitivity.

Computation efficiency

The computation times of our method, piRNApredictor, Piano, and piRPred were measured for 500 positive and 500 negative sequences that were randomly selected from the drosophila set. All were run on a computer equipped with i5-4950 CPU and 8 GB RAM. Our method required 213 s for data processing, which was slightly slower than Piano (134 s) but was 4-fold faster than piRPred (954 s) and 229-fold faster than piRNApredictor (48,876 s).

Compared to piRNApredictor, our method included a significantly reduced feature number and sample size, and the SVM models of our method were all saved. Zhang et al. (2011) used 173,090 positive and 193,321 negative samples to build a classifier, and 1364 features were extracted. Here, the training set size was decreased, ranging from 50,000 to 80,000 samples. Additionally, only 200 features were selected from the 741 features obtained by mRMR.

Based on multiple kernels, piRPred requires considerably more time and memory than our method. Three kernels were used, and each kernel is a square similarity matrix of size N x N, where N is the size of the dataset. The matrix building process implemented in the R programming language requires much more memory and time than our method. In particular, the piRNA cluster feature of piRPred is extracted in a supervised manner and is very time-consuming.

Piano was the fastest among the tools tested in this study. This is likely because a number of sequences were filtered out when aligning to transposons in excess of three mismatches. Furthermore, Piano only considers 32 features and uses some mature tools. In the feature extraction process, the software SeqMap (Jiang and Wong, 2008) is used to map small RNAs to transposon sequences, and RNAplex (Tafer and Hofacker, 2008) is used to analyze piRNA-transposon interaction information.

Robustness to sequencing error

Our method uses a sliding-window to obtain k-mer and k-mer with wildcards features. K-mer features were found to be somewhat robust to indel sequencing errors (Li et al., 2014); however, the proposed K-mer wildcards must be more robust than K-mer since it introduces wildcards, allowing for additional non-indel nuclear change sequencing errors. This indicates that the proposed features are robust.

CONCLUSIONS

This approach proposes a method for detecting piRNAs based on sequence features. The proposed method provides a better balance between precision and sensitivity (both are approximately 90%) and is more efficient compared to some common existing approaches. Four feature types were employed, and a standard SVM was applied for detection. The features included weighted k-mer, weighted k-mer with wildcards, position-specific base, and piRNA length. Furthermore, this method is slightly slower than Piano but is 4-fold faster than piRPred and 229-fold faster than piRNApredictor.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Research supported by the Natural Science Foundation of China under Grants #61571341, #61201312, #91530113, the Research Fund for the Doctoral Program of Higher Education of China (#20130203110017), the Fundamental Research Funds for the Central Universities of China (#BDY171416 and #JB140306), the Natural Science Foundation of Shaanxi Province in China (#2015JM6275).

REFERENCES

Aravin A, Gaidatzis D, Pfeffer S, Lagos-Quintana M, et al. (2006). A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* 442: 203-207 10.1038/nature04916.

Belhumeur PN, Hespanha JP and Kriegman D (1997). Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *Pattern Analysis and Machine Intelligence*. *IEEE Transactions on* 19: 711-720 10.1109/34.598228.

Betel D, Sheridan R, Marks DS and Sander C (2007). Computational analysis of mouse piRNA sequence and biogenesis. *PLOS Comput. Biol.* 3: e222 http://dx.doi.org/10.1371/journal.pcbi.0030222.

Brayet J, Zehraoui F, Jeanson-Leh L, Israeli D, et al. (2014). Towards a piRNA prediction using multiple kernel fusion and support vector machine. *Bioinformatics* 30: i364-i370 http://dx.doi.org/10.1093/bioinformatics/btu441.

Bu D, Yu K, Sun S, Xie C, et al. (2011). NONCODE v3. 0: integrative annotation of long noncoding RNAs. *Nucleic Acids Res.* gkr1175: D210-D215. 10.1093/nar/gkr1175.

Burge C, Campbell AM and Karlin S (1992). Over- and under-representation of short oligonucleotides in DNA sequences. *Proc. Natl. Acad. Sci. USA* 89: 1358-1362 http://dx.doi.org/10.1073/pnas.89.4.1358.

Chang C-C and Lin C-J (2011). LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2: 1-27 http://dx.doi.org/10.1145/1961189.1961199.

Girard A, Sachidanandam R, Hannon GJ and Carmell MA (2006). A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* 442: 199-202 10.1038/nature04917.

Gutiérrez G, Oliver JL and Marín A (1993). Dinucleotides and G+C content in human genes: opposite behavior of GpG, GpC, and TpC at II-III codon positions and in introns. *J. Mol. Evol.* 37: 131-136 http://dx.doi.org/10.1007/BF02407348.

Houwing S, Kamminga LM, Berezikov E, Cronembold D, et al. (2007). A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. *Cell* 129: 69-82 http://dx.doi.org/10.1016/j.cell.2007.03.026.

Jiang H and Wong WH (2008). SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics* 24: 2395-2396 http://dx.doi.org/10.1093/bioinformatics/btn429.

Jones KS (1972). A statistical interpretation of term specificity and its application in retrieval. *J. Doc.* 28: 11-21. http://dx.doi.org/10.1108/eb026526

Karlin S and Ladunga I (1994). Comparisons of eukaryotic genomic sequences. *Proc. Natl. Acad. Sci. USA* 91: 12832-12836 http://dx.doi.org/10.1073/pnas.91.26.12832.

Karlin S and Mrázek J (1997). Compositional differences within and between eukaryotic genomes. *Proc. Natl. Acad. Sci. USA* 94: 10227-10232 http://dx.doi.org/10.1073/pnas.94.19.10227.

Karlin S, Ladunga I and Blaisdell BE (1994). Heterogeneity of genomes: measures and values. *Proc. Natl. Acad. Sci. USA* 91: 12837-12841 http://dx.doi.org/10.1073/pnas.91.26.12837.

Li A, Zhang J and Zhou Z (2014). PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinformatics* 15: 311 http://dx.doi.org/10.1186/1471-2105-15-311.

Madera M, Calmus R, Thiltgen G, Karplus K, et al. (2010). Improving protein secondary structure prediction using a simple k-mer model. *Bioinformatics* 26: 596-602 http://dx.doi.org/10.1093/bioinformatics/btq020.

Peng H, Long F and Ding C (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. Pattern Analysis and Machine Intelligence. *IEEE Transactions on Pattern Analysis Machine Intelligence* 27: 1226-1238 10.1109/TPAMI.2005.159.

Reuter M, Berninger P, Chuma S, Shah H, et al. (2011). Miwi catalysis is required for piRNA amplification-independent LINE1 transposon silencing. *Nature* 480: 264-267 http://dx.doi.org/10.1038/nature10672.

Sai Lakshmi S and Agrawal S (2008). piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. *Nucleic Acids Res.* 36 (Suppl 1): D173-D177 http://dx.doi.org/10.1093/nar/gkm696.

Salton G and Buckley C (1988). Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.* 24: 513-523 http://dx.doi.org/10.1016/0306-4573(88)90021-0.

Seto AG, Kingston RE and Lau NC (2007). The coming of age for Piwi proteins. *Mol. Cell* 26: 603-609 http://dx.doi.org/10.1016/j.molcel.2007.05.021.

Siomi MC, Sato K, Pezic D and Aravin AA (2011). PIWI-interacting small RNAs: the vanguard of genome defence. *Nat. Rev. Mol. Cell Biol.* 12: 246-258 http://dx.doi.org/10.1038/nrm3089.

Tafer H and Hofacker IL (2008). RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics* 24: 2657-2663 http://dx.doi.org/10.1093/bioinformatics/btn193.

Watanabe T, Cheng EC, Zhong M and Lin H (2015). Retrotransposons and pseudogenes regulate mRNAs and lncRNAs via the piRNA pathway in the germline. *Genome Res.* 25: 368-380 http://dx.doi.org/10.1101/gr.180802.114.

Wang K, Liang C, Liu J, Xiao H, et al. (2014). Prediction of piRNAs using transposon interaction and a support vector machine. *BMC Bioinformatics* 15: 419 http://dx.doi.org/10.1186/s12859-014-0419-6.

Zhang Y, Wang X and Kang L (2011). A k-mer scheme to predict piRNAs and characterize locust piRNAs. *Bioinformatics* 27: 771-776 http://dx.doi.org/10.1093/bioinformatics/btr016.