Sequence analysis

A k-mer scheme to predict piRNAs and characterize locust piRNAs

Yi Zhang1,2, Xianhui Wang1 and Le Kang1,*

1State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101 and
2Department of Mathematics, Hebei University of Science and Technology/Hebei Laboratory of Pharmaceutical Molecular Chemistry, Shijiazhuang, Hebei 050018, China

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Identifying piwi-interacting RNAs (piRNAs) of non-model organisms is a difficult and unsolved problem because piRNAs lack conservative secondary structure motifs and sequence homology in different species.

Results: In this article, a k-mer scheme is proposed to identify piRNA sequences, relying on the training sets from non-piRNA and piRNA sequences of five model species sequenced: rat, mouse, human, fruit fly and nematode. Compared with the existing static scheme based on the position-specific base usage, our novel dynamic algorithm performs much better with a precision of over 90% and a sensitivity of over 60%, and the precision is verified by 5-fold cross-validation in these species. To test its validity, we use the algorithm to identify piRNAs of the migratory locust based on 603 607 deep-sequenced small RNA sequences. Totally, 87 536 piRNAs of the locust are predicted, and 4426 of them matched with existing locust transposons. The transcriptional difference between solitary and gregarious locusts was described. We also revisit the position-specific base usage of piRNAs and find the conservation in the end sequences of piRNAs.

Availability: The web server for implementing the algorithm and the source code are freely available to the academic community at http://ls9.168.30/piRNA/index.php.

Contact: kang@ioz.ac.cn

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on August 19, 2010; revised on December 23, 2010; accepted on January 5, 2011

1 INTRODUCTION

Non-coding RNAs (ncRNAs) are functional RNA molecules that are not translated into proteins, including highly abundant and functionally important RNAs such as transfer RNA (tRNA) and ribosomal RNA (rRNA), as well as other RNAs such as snoRNAs, microRNAs, siRNAs and piRNAs and the long ncRNAs. Among them, the ones of typically 20–30 nt in length are called small RNA. Piwi-interacting RNA (piRNA) is the largest class of small RNA molecules expressed in animal cells, especially in germ cells, and 25–32 nt long in general (Aravin et al., 2006; Girard et al., 2006; Grivina et al., 2006). piRNAs form RNA–protein complexes through interactions with PIWI proteins, and has no clear secondary structure motifs (Kandhavelu et al., 2009), and its length is slightly longer than miRNA. Compared with miRNAs, piRNA is lack of primary sequence conservation, and the presence of a 5′ uridine is common in both vertebrates and invertebrates. piRNAs in the nematode have a 5′ monophosphate and a 3′ modification that block either 2′ or 3′ oxygen (Ruby et al., 2006), and are confirmed to exist in fruit fly (Yin and Lin, 2007; Vagin et al., 2006), zebrafish (Houring et al., 2007), mice (Kurino and Mourelatos, 2007; Watanabe et al., 2006) and rats (Houring et al., 2007). PIWI/ARGONAUTE (also known as PAZ-PIWI domain or PPD) protein family is evolutionarily conserved owing to its functional significance in stem cell self-renewal and germline development (Vagin et al., 2006).

piRNA derives from the post-transcriptional amplification ‘Ping-Pong Model’, and it may be involved in germ cell formation, germline stem cell maintenance, spermiogenesis and oogenesis (Brennecke et al., 2007; Cox et al., 1998; Thomson and Lin, 2009). Therefore, available piRNA data mainly come from model species with complete genome sequences. A general approach to detecting piRNA is based on the combination of immunoprecipitation and deep sequencing in model and sequenced organisms (Yin and Lin, 2007). However, the lowly expressed or issue-specific piRNAs might be missed using this method. In addition, some of piRNAs are not produced by ‘Ping-Pong Model’ (Das et al., 2008; Robine et al., 2009).

Thus, computational methods may provide an alternative approach to detect piRNAs, which can summarize general properties from known piRNAs and then train them to predict novel piRNAs. Betel et al. (2007) first use the position-specific usage of 10 upstream bases and 10 downstream bases of 5′ UTR to construct a vector with 21 × 4 components, by which they characterized and identified mouse piRNAs with a precision of 61–72%. They also found that mouse piRNAs have some position-special properties, such as G or A at +1 position, A at +4 position and a slight underrepresentation of G at –1 position. However, their method has limitations in predicting piRNAs from the organism without genome information (Lakshmhi and Agrawal, 2008). Meanwhile, this method cannot efficiently detect those piRNAs derived from 3′ UTR of mRNA, which are not produced by ‘Ping-Pong Model’ (Das et al., 2008; Robine et al., 2009). Furthermore, piRNA sequences are quite divergent among different species (Lakshmhi and Agrawal, 2008).
vectors provide a novel approach to distinguish piRNA from non-piRNA. Of string usages between piRNA and non-piRNA sequences, the 1364 D
are used. To characterize piRNA sequences, we use all the 1–5 nt strings, methods are urgently demanded. The
256–mer strings, 1024
mer strings, 64
2–tuples or
k
mer statistics giving
discrete probability distributions of a number of possible k-mer combinations are used. To characterize piRNA sequences, we use all the 1–5 nt strings, including 4 k-mer statistics: A, G, C and T. 16 k-mer strings, 64 1-mer strings, 256 1-mer strings, 1024 5-mer strings, and totally 1364 strings. A bio-sequence can be characterized by a vector consisting of frequencies of the 1364 k-mer (k = 1, 2, 3, 4, 5) strings. Because there are significant differences of string usages between piRNA and non-piRNA sequences, the 1364 D vectors provide a novel approach to distinguish piRNA from non-piRNA.

2.2 Construction of training set
Constructing training set to computationally detect piRNA with Fisher discriminant algorithm (Fisher, 1936), we use two groups of samples: a positive group consisting of true piRNA sequences from five model species and a negative group of non-piRNA sequences. The positive dataset consists of known piRNA sequences of five species: rat, mouse, human, fruit fly and nematode: piRNAs from the first three species are downloaded from NONCODE version 2.0 (Liu et al., 2005), and piRNAs from the last two species are obtained from NCBI (nematode: gi222139841 ~ gi222139820; fruit fly: gi15362817 ~ gi15361675). In total, we obtain 173,090 positive samples, including 32,046 human piRNAs, 72,747 mouse piRNAs, 66,758 rat piRNAs, 552 Caenorhabditis elegans piRNAs and 987 Drosophila piRNAs.

The negative samples are derived from NONCODE version 2.0 (Liu et al., 2005). NONCODE is a database of a wide variety of ncRNA classes (small and long ncRNAs) from 861 organisms covering all kingdoms of life (276 eukaryotes, 557 prokaryotes, and 28 archaea). For each of the 34,675 non-piRNA sequences in the negative samples, we shuffled it 10,000 times to destroy any potential functional structures, then randomly selected start points and generated no more than 5 subsequences with a length of 18–32 nt. Since there are 9257/207,765 = 4.46% ncRNAs shorter than 25 nt in ncRNA database NONCODE version 2.0 (Liu et al., 2005), we randomly produced 8678/193,321 = 4.49% sequences shorter than 25 nt to make the length distribution of negative samples similar to that of a real database.

In detail, the random processes generating 158,646 random sequences as negatives from the 34,675 non-piRNA ncRNA by the following method. For each of the 34,675 non-piRNA sequences, we shuffled it 10,000 times to destroy any potential functional structures, then randomly selected start points and generated no more than 5 subsequences with a length of 18–32 nt. There are 34,675 non-piRNA non-coding RNA sequences, and it should be noted that most of them are much longer than positive sequences. At first, the 34,675 non-piRNA ncRNA were selected as negative samples. Then, to make the number of negative samples close to that of the positive samples, we generated 158,646 random sequences as negative samples from the 34,675 non-piRNA ncRNA sequences by the following method. For each of the 34,675 non-piRNA sequences, we shuffled it 10,000 times to destroy any potential functional structures, then randomly selected start points and generated no more than 5 subsequences with a length of 18–32 nt. Since there are 9257/207,765 = 4.46% ncRNAs shorter than 25 nt in ncRNA database NONCODE version 2.0 (Liu et al., 2005), we randomly produced 8678/193,321 = 4.49% sequences shorter than 25 nt to make the length distribution of negative samples similar to that of a real database.

In detail, the random processes generating 158,646 negative samples cover three steps. Firstly, we divided each sequence into 40 nt-long non-overlap blocks, and chose no more than five blocks as random candidate blocks. Secondly, the length distribution was confined to 18–32 nt, which has little effect on the result because we only use the frequency of strings. Finally, the negative sequences can start at every possible position in a selected block.

2.3 Improved Fisher Algorithm in a 1364 D space
The Fisher discriminant algorithm uses a training set formed by these two groups of samples to obtain a discriminant vector w and threshold \( y_0 \). The Fisher linear discriminant equation in this case represents a super-plane in the 1364 D space, described by a vector \( w \), which is extremely simple in the two-class case. Let Group 1 (denoted by \( g_1 \)) and \( \tilde{g}_1 \) non-piRNA samples and \( \tilde{x}_1 = (\tilde{x}_{1,1}, \ldots, \tilde{x}_{1,1364}) \) the 1364 D vector defined above of the \( \tilde{g}_1 \) non-piRNA sequences, and it should be noted that most of them are much longer than positive sequences. At first, the 34,675 non-piRNA ncRNA were selected as negative samples. Then, to make the number of negative samples close to that of the positive samples, we generated 158,646 random sequences as negative samples from the 34,675 non-piRNA ncRNA sequences by the following method. For each of the 34,675 non-piRNA sequences, we shuffled it 10,000 times to destroy any potential functional structures, then randomly selected start points and generated no more than 5 subsequences with a length of 18–32 nt. Since there are 9257/207,765 = 4.46% ncRNAs shorter than 25 nt in ncRNA database NONCODE version 2.0 (Liu et al., 2005), we randomly produced 8678/193,321 = 4.49% sequences shorter than 25 nt to make the length distribution of negative samples similar to that of a real database.

In detail, the random processes generating 158,646 negative samples cover three steps. Firstly, we divided each sequence into 40 nt-long non-overlap blocks, and chose no more than five blocks as random candidate blocks. Secondly, the length distribution was confined to 18–32 nt, which has little effect on the result because we only use the frequency of strings. Finally, the negative sequences can start at every possible position in a selected block.

The Fisher linear discriminant equation in this case represents a super-plane in the 1364 D space, described by a vector w, which is extremely simple in the two-class case. Let Group 1 (denoted by \( g_1 \)) correspond to piRNA samples, Group 2 (denoted by \( g_2 \)) non-piRNA samples and \( \tilde{x}_1 = (\tilde{x}_{1,1}, \ldots, \tilde{x}_{1,1364}) \) the 1364 D vector defined above of the \( \tilde{g}_1 \) non-piRNA sequences, and it should be noted that most of them are much longer than positive sequences. At first, the 34,675 non-piRNA ncRNA were selected as negative samples. Then, to make the number of negative samples close to that of the positive samples, we generated 158,646 random sequences as negative samples from the 34,675 non-piRNA ncRNA sequences by the following method. For each of the 34,675 non-piRNA sequences, we shuffled it 10,000 times to destroy any potential functional structures, then randomly selected start points and generated no more than 5 subsequences with a length of 18–32 nt. Since there are 9257/207,765 = 4.46% ncRNAs shorter than 25 nt in ncRNA database NONCODE version 2.0 (Liu et al., 2005), we randomly produced 8678/193,321 = 4.49% sequences shorter than 25 nt to make the length distribution of negative samples similar to that of a real database.

In detail, the random processes generating 158,646 negative samples cover three steps. Firstly, we divided each sequence into 40 nt-long non-overlap blocks, and chose no more than five blocks as random candidate blocks. Secondly, the length distribution was confined to 18–32 nt, which has little effect on the result because we only use the frequency of strings. Finally, the negative sequences can start at every possible position in a selected block.
The size distribution of all known piRNAs largely varied ranging from 18 nt to 32 nt, and mainly distributed in 28, 29, 30 and 31 nt which cover 72.32% known piRNAs (Supplementary Material S1). For example, for the ratio of absolute value of string frequency difference to the sum of all string frequencies in piRNAs and non-piRNAs, we define the string frequency relative difference as significant strings whose usage are different between piRNAs and non-piRNAs. For the two variables, we may improve the Fisher discriminant algorithm. In detail, we change the discriminant formula into the new one shown below.

\[ f(x) = \tilde{w} \cdot \tilde{x} + \tilde{b} \]

and once the Fisher vector \( w \) is obtained, the decision of piRNA/non-piRNA is performed by the criterion of \( f(x) > \tilde{f}_0 \) or \( f(x) < \tilde{f}_0 \), respectively. Obviously, the cutoff value is \( \tilde{w} \cdot \tilde{x} + \tilde{b} \).

### 3 RESULTS AND DISCUSSION

#### 3.1 Different string usage of piRNA and non-piRNA

piRNA and non-piRNA sequences have significant differences in string usage. First, for each sequence (piRNA or non-piRNA), we calculate the frequencies of all the 1364 k-mer (\( k = 1, 2, 3, 4, 5 \)) strings, and construct a 1364 D vector to characterize the sequence. Then, we use rank sum test to determine which string usage is significantly different between piRNAs and non-piRNAs. With a significance level of \( 10^{-300} \), we found that the usage of 1337 strings (Supplementary Material S1) is significantly different between piRNAs and non-piRNAs. Therefore, the k-mer string scheme can spot the difference between piRNAs and non-piRNAs, and the difference can be visualized by comparing the frequencies of each string in piRNAs and non-piRNAs (Fig. 1). To identify the most significant strings whose usage are different between piRNAs and non-piRNAs, we define the string frequency relative difference as the ratio of absolute value of string frequency difference to the sum of string frequency in piRNAs and non-piRNAs. For example, for string ‘TGCTG’, its string frequency relative difference is

\[ \frac{f_{\text{piRNA}}(\text{TGCTG}) - f_{\text{nonpiRNA}}(\text{TGCTG})}{f_{\text{piRNA}}(\text{TGCTG}) + f_{\text{nonpiRNA}}(\text{TGCTG})} \]

where \( f_{\text{piRNA}}(\text{TGCTG}) \) is the frequency of string TGCTG appeared in piRNAs. There are 32 strings with string frequency relative difference larger than 0.7 (Supplementary Material S2), while only the string ‘TGCTG’ with a higher frequency in piRNAs than in non-piRNAs, perhaps because ‘TGCTG’ is the first 5 bases of many piRNAs. The left 31 strings all have low expression in piRNAs, but their biological significance requires further exploration.

#### 3.2 Position-specific base usage of piRNA

The size distribution of all known piRNAs largely varied ranging from 18 nt to 32 nt, and mainly distributed in 28, 29, 30 and 31 nt which cover 72.32% known piRNAs (Supplementary Material Fig. S1). With the comparison of piRNAs and non-piRNAs, we revisited the position-specific properties in detail. Then, we calculated the frequencies of four bases in each position, and identified conserved position-specific bases at the beginning and the end of piRNAs (Fig. 2A), besides G or A at +1 position, A at +4 position and a slight underrepresentation of G at −1 position, especially in the 30, 31 and the 32 position. Furthermore, we detected position-specific base usages by using rank sum test. In the analysis, we only considered the beginning 21 base positions of piRNA to cover all possible piRNA sequences. Setting significant level to be \( 10^{-100} \), we found that 21 position-specific base usages are significantly different between piRNA and non-piRNA. They are a1, g1, c1, t1, g2, t2, g3, c3, t3, a4, t4, a5, g5, t5, a6, c6, t6, t7, a10, c10 and c12. The difference can be visualized by comparing the frequencies of these position-specific bases in piRNAs and non-piRNAs (Fig. 2B).
Table 1. Definitions of precision and sensitivity of prediction

<table>
<thead>
<tr>
<th>Actual positives</th>
<th>Predicted positives</th>
<th>Predicted negatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td></td>
<td>FN</td>
</tr>
<tr>
<td>Actual negatives</td>
<td>FP</td>
<td>TN</td>
</tr>
<tr>
<td>sn</td>
<td></td>
<td>sp</td>
</tr>
</tbody>
</table>

TP: True positives; FP: False positives; TN: True negatives; sn: Sensitivity; sp: Precision.

3.3 Cross-validation tests

Prediction precision and sensitivity are widely used to evaluate the performance of an algorithm. Sensitivity is the ratio of number of true positive samples to that of actual positive samples, and the precision is the ratio of number of true positive samples to those of predicted positive samples. Their definitions are listed in Table 1. We performed five cross-validation tests in five species: rat, mouse, human, fruit fly and nematode. In order to evaluate the prediction precision and sensitivity of current algorithm in predicting piRNA for a new species, we used the piRNAs of four species as training set and the piRNAs of another species as test set. Each time we use 50000 pairs of piRNA and non-piRNA sequences derived from four species as drill set to predict piRNAs of another species. An improved Fisher formula

\[ f(x) = w^T \cdot x - \hat{m}_2 + t \cdot N_{\text{std}} \]

can promote the prediction precision over 90%. To compare current algorithm with that proposed by Betel et al. (2007), 36373 mouse piRNAs were taken as training positive set to predict the remaining 36374 piRNAs, when \( t = 0 \), precision is 68.41%, and sensitivity is 99.31%; setting \( t = 2 \), precision is 95.53%, and sensitivity is 72.47%. However, the prediction precision of Betel et al. (2007) is only 61%, suggesting that our algorithm may still be useful for the species with full genome information.

3.4 The method validity tests and locust piRNA prediction

Wei et al. (2009) reported the small RNA transcriptome of the migratory locust (Locusta migratoria) from gregarious and solitary phase libraries containing 603 607 sequences and a subset of small RNA in a peak at 25–29 nt. These data provide a valuable source to test the validity of new method and to identify piRNAs of the migratory locust. With the improved Fisher Algorithm, using 120000 piRNAs derived from the five model species mentioned above and 120000 non-piRNAs as drill set, we identified 87536 locust piRNAs with length larger than 24 nt (Supplementary Material), including 12386 gregarious-specific piRNAs, 69151 solitary-specific piRNAs and 5999 piRNAs for both two phases. The analysis of prediction sensitivity showed that the sensitivity decreases as \( t \) increases (Fig. 3A). Especially, when \( t = 2 \), the sensitivities of most species (except for fruit fly) are 60–70%, indicating that the 87536 predicted piRNAs are only a fraction of all locust piRNAs. Therefore, we estimated the total number of locust piRNAs is about 130000, which is less than that of Drosophila’s piRNAs (Lakshmi and Agrawal, 2008). After analyzing the usage of the 1337 strings in the 87536 predicted locust piRNAs, we found that the usage of the strings in locust piRNAs is consistent with that of the five model organisms (Fig. 4). The principle of the Fisher discriminant algorithm improved is to detect the different orientations of string usage between positive and negative samples, and then determines the sequences with amplified positive orientations as piRNAs.
The 15 strings that are differently used in two phases with a significance level of 10\(^{-100}\)

<table>
<thead>
<tr>
<th>k-mer</th>
<th>Base strings</th>
</tr>
</thead>
<tbody>
<tr>
<td>1mer</td>
<td>C</td>
</tr>
<tr>
<td>2mer</td>
<td>CC</td>
</tr>
<tr>
<td>3mer</td>
<td>ACC, GCC, CCA, CCC, CCT, CTC</td>
</tr>
<tr>
<td>4mer</td>
<td>CACC, CCGT, CCCC, CGGT, CGTT</td>
</tr>
<tr>
<td>5mer</td>
<td>CTGCA, CTCGT, TCCGA, TTGCT, TTGTA</td>
</tr>
</tbody>
</table>

comparing the advantage of the two methods through the relation of precision/sensitivity versus \(t\) values, the ‘dynamic’ method outperforms the ‘static’ method constructing an 84 D vector (21 positions \(\times\) 4 bases, to include all piRNA sequences) in identifying piRNAs (Fig. 3B). The ‘dynamic’ method can identify 1337 strings in piRNAs and non-piRNAs with a significant level of 10\(^{-100}\), while ‘static’ method can only identify 21 position-specific base usages with a significance level of 10\(^{-10}\).

3.5 The difference of piRNAs between the solitary and gregarious locusts

We found that 15 strings are differently used between two phase locusts, and the expression of piRNAs in solitary locust is much higher than in gregarious ones. Differences between solitary and gregarious locusts are contributed to gene expression and regulation level modulated by piRNA (Wei et al., 2009), because they have the same genome sequence. In the 87 536 locust predicted piRNAs, there are 12 386 gregarious-specific piRNAs and 69 151 solitary-specific piRNAs. Fifteen strings in gregarious and solitary-specific piRNAs display significantly different utilization rate with a significance level of 10\(^{-100}\) (Table 2).

This difference of string utilization rate can be visualized by comparing the average frequencies of the 15 strings in solitary and gregarious locusts (Fig. 5A). The most significant difference is the high content of C in the gregarious locust piRNAs. Based on the figure presented the framework of this article, which showed the robustness of our method in detecting piRNAs. There are 5999 piRNAs shared by two phase locusts, and the piRNAs in solitary locusts have more reads than in gregarious locusts. Of total, 3912 of piRNAs have more reads in solitary locusts, 1435 piRNAs have equal reads in two phases and only 652 piRNAs have more reads in gregarious than in solitary locusts (Supplementary Material Fig. S2). These highly expressed piRNAs may play an important role in maintaining strong propagation of the solitary locusts. We calculated ratios of piRNA reads in solitary to gregarious locusts, and found that the ratios of 84 piRNA reads are above 30 (Supplementary Material). The 84 piRNAs are ideal candidates for further piRNA interference in investigating piRNA modulation mechanism of phase transition in locusts.

3.6 Match the 87 536 predicted piRNAs with transposons

There are 4426 of 87 536 locust piRNAs matched with locust transposons from the locust transcriptome data (Kang et al., unpublished data). Not all transposons are transcribed in the locust transcriptome data, so we only get 6635 locust transposons. When the locust piRNAs are compared with the transposons, 4426 matches are found and over half transposons are hit (Supplementary Material). As expected, most of them have the largest values of \(F(x) = w^x\) among all 603 607 candidate sequences (Fig. 6). In fact, the figure presented the framework of this article, which showed the distributions of projective points for 120 000 pairs of drill sequences, 603 607 deep-sequenced candidate small RNA sequences, 87 536 predicted locust piRNAs, and 4426 locust piRNAs matched with locust transposons.

4 CONCLUSION

In this article, we implemented a \(k\)-mer algorithm to predict piRNAs. Compared with previous approaches, the new method does not require a reference genome and gives a much better performance on piRNA prediction. We also improved the Fisher algorithm by setting different cutoffs and elevating the precision rate. The basic work is to obtain the Fisher vector \(w\), the mean value \(\bar{m}_2\) and the SD \(N_{\text{std}}\) of the negative samples. Then, a sequence represented by a 1364 D vector \(x\) can be regarded as a piRNA if its \(w^x\) is larger than \(\bar{m}_2 + 2 \times N_{\text{std}}\). Using this new scheme, we obtained 87 536 putative piRNAs from the locust, which would be very helpful in studying the phase transition mechanism of insects, especially hemimetamorphosis insects. Moreover, the 84 locust piRNAs, which have the largest ratio of solitary to gregarious reads, and the 4426 locust piRNAs matched with existing transposons may provide excellent candidates for further experiments.
The authors are grateful to the associate editor and anonymous reviewers for comments and helping to improve the earlier version. The authors thank Drs Zhao, F.Q., Wu J.Y. and Zhu E.L. for their critical discussion, and Zhang X.X. for her polishing of the manuscript.

Funding: National Basic Research Program of China (2006CB10000-1 to L.K.); National Natural Science Foundation of China (3087022 to L.K.); and Hebei University of Science and Technology Foundation (Q200951, XL200902 to Y.Z.); Beijing Institutes of Life Science Foundation (2010-BioCAS-0304 to X.W.).

Conflict of Interest: none declared.

REFERENCES


