BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 6 2011, pages 771–776
doi:10.1093/bioinformatics/btr016

Sequence analysis

Advance Access publication January 11, 2011

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

Yi Zhang[1,2], Xianhui Wang[1] and Le Kang[1,*]

[1]State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101 and
[2]Department of Mathematics, Hebei University of Science and Technology/Hebei Laboratory of Pharmaceutic 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 of piRNAs. Therefore, the method we developed can be used to identify piRNAs of non-model organisms without complete genome sequences.

Availability: The web server for implementing the algorithm and the software code are freely available to the academic community at http://59.79.168.90/piRNA/index.php.

Contact: lkang@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; Grivna *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 (Houwing *et al.*, 2007), mice (Kirino and Mourelatos, 2007; Watanabe *et al.*, 2006) and rats (Houwing *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′ U 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 (Lakshmi 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 (Lakshmi and Agrawal,

*To whom correspondence should be addressed.

2008; Seto *et al.*, 2007). Most general methods, such as BLAST and MEME, are inappropriate for piRNA prediction. For example, we cannot find any homologous piRNAs between fruit fly and other species with BLAST; and not any conserved motifs are found in piRNAs with MEME. Therefore, more efficient computational methods are urgently demanded. The *k*-mer scheme is widely used to characterize biosequences (Burge *et al.*, 1992; Gutierrez, 1993; Karlin and Ladunga,1994), because patterns using *k*-mers have been found to be species or taxon special (Karlin *et al.*, 1994; Karlin and Mrazek, 1997; Madera *et al.*, 2010).

It is important to note that, the Solexa small RNA data may also include miRNA, piRNA, as well as some short fractions of snoRNA, snRNA, long ncRNA and un-annotated mRNAs, which have similar lengths to piRNA. In fact, NONCODE version 2.0 (Liu *et al.*, 2005) has 9257/20 7765 = 4.46% ncRNAs shorter than 25 nt, suggesting that most ncRNA may produce a sequence fraction with similar length to piRNA. Moreover, possible background noise also exists, including the background information introduced by Solexa, the degraded RNA fragments in sample preparation and the noise caused by the random match of data with genome. Studies of Wei *et al.* (2009) demonstrated that there are about 30–40% short sequences in 603 607 possible candidates of locust small RNAs are unannotated. On the other words, there are about 200 000 unannoted small RNAs in which there would be a huge number of piRNAs and other short sequences. Obviously, these remaining sequences could cover real piRNAs, the fractions of long ncRNAs, unannotated mRNAs and noise produced by Solexa. Therefore, piRNAs in Solexa small RNA data cannot be identified merely based on their lengths. In order to predict piRNAs, an efficient algorithm is required to distinguish the real piRNA from the sequences within similar range of length. Here, we used all the 1364 1–5 nt strings and an improved Fisher Discriminant algorithm to characterize piRNA sequences in five model species: rat, mouse, human, fruit fly and nematode. The novel algorithm reached a prediction precision of over 90% and a sensitivity of over 60% in these species. We applied this algorithm to the deep-sequenced small RNA data (Wei *et al.*, 2009) of the migratory locust (*Locusta migratoria*), which is an important agricultural pest and model species for physiology, neuroscience and behavior. This method successfully identified 87 536 piRNAs of the locust with the same precision and sensitivity as the five model species. Therefore, the method proposed in this study can be used to predict piRNAs for both model and non-model organisms.

2 METHODS

2.1 *k*-mer string

In bioinformatics, *k*-mers usually refer to specific *k*-tuples or *k*-grams of nucleic acid or amino acid sequences that can be used to identify certain regions within biomolecules like DNA or proteins. Either *k*-mer strings as such can be used for finding regions of interest 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 1mer strings: A, G, C and T, 16 *2*mer strings, 64 *3*mer strings, 256 *4*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: gi222138841 ∼ 222138290; fruit fly: gi157362817 ∼ 157361675). 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 (eukaryotes, eubacteria, archea and viruses). Data are from three sources: (i) manual extracts from literature, (ii) automatically filtered and manually confirmed GenBank sequences and (iii) experimental data from Chen's laboratory (Giulia *et al.*, 2009). In detail, it includes 'miRNA', 'piRNA', 'mlRNA', 'snoRNA', 'snRNA', 'tmRNA', 'SRP RNA', 'gRNA', 'sbRNA', 'snlRNA', etc. Thus, it is qualified as the source for the ncRNA study.

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) correspond to piRNA samples, Group 2 (denoted by G_2) non-piRNA samples and $x_k^g = (x_{k1}^g, x_{k2}^g, \ldots, x_{k1364}^g)$ the 1364 D vector defined above of the k-th sample in group g (g = 1, 2), where $k = 1, 2, \ldots, n_g$ (n_1, n_2 are the numbers of samples in G_1 and G_2, respectively). We calculate the average vector m_g for each group: $m_g = \frac{1}{n_g}\sum_{k=1}^{n_g} x_k^g, g = 1, 2$. Denoting by S_w the sum of the covariance matrices of two groups, we have $S_w = \sum_{g=1}^{2}\sum_{k=1}^{n_g}(x_k^g - m_g)(x_k^g - m_g)^\tau, g = 1, 2$. The Fisher vector w is simply determined by the following equation: $w = S_w^{-1}(m_1 - m_2)$, where S_w^{-1} is the inverse of the matrix S_w. Thus, for any 1364 D vector $x_k^g = (x_{k1}^g, x_{k2}^g, \ldots, x_{k1364}^g)$, $k = 1, 2, \ldots, n_g$. its projective point is $y_k^g = w^\tau x_k^g$. Notice that w is not unique in the sense that w multiplied by a constant is still acceptable. Without loss of generality, we choose such w satisfying $||w|| = 1$. Based on the data in the training set, an appropriate threshold y_0 is determined to make the piRNA/non-piRNA decision. The threshold y_0 is determined by the formula: $y_0 = \frac{1}{2}(\frac{n_1\widetilde{m_1} + n_2\widetilde{m_2}}{n_1 + n_2} + \frac{1}{2}(\widetilde{m_1} + \widetilde{m_2}))$,

where $\widetilde{m}_g = \frac{1}{n_g}\sum_{k=1}^{n_g} y_k^g, g=1,2$. Once the Fisher vector w and the threshold y_0 are obtained, the decision of piRNA/non-piRNA in the test set is simply performed by the criterion of $f(x)>0/f(x)<0$, where $f(x)=w^\tau x - y_0$. To improve the Fisher algorithm, we introduce the 'cutoff', which in theoretical physics means the maximal (or minimal) value of energy, momentum or length, so that the objects with even smaller (or larger) values than these physical quantities are ignored. A popular method in increasing precision is to set higher cutoff values (Candolfi *et al.*, 1993; Huang *et al.*, 2006). The $\widetilde{m}_2$ is the mean value of the projective points [i.e. $f(x)=w^\tau x$] of non-piRNAs. Here, we set N_{std} to be the SD of the projective points of non-piRNAs. With the two variables, we may improve the Fisher discriminant algorithm. In detail, we change the discriminant formula into the new one shown below.

$$f(x)=w^\tau x-\widetilde{m}_2-t\times N_{\text{std}},$$

and once the Fisher vector w is obtained, the decision of piRNA/non-piRNA is performed by the criterion of $f(x)>0$ and $f(x)<0$, respectively. Obviously, the cutoff value is $\widetilde{m}_2+t\times N_{\text{std}}$.

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

$$\frac{|f_{\text{piRNA}}(TGCTG)-f_{\text{nonpiRNA}}(TGCTG)|}{(f_{\text{piRNA}}(TGCTG)+f_{\text{nonpiRNA}}(TGCTG))},$$

where $f_{\text{piRNA}}(TGCTG)$ 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.

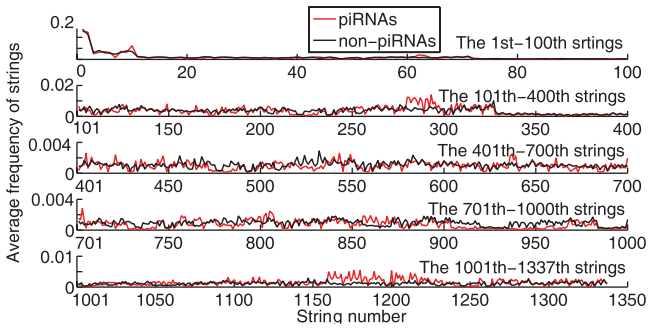

Fig. 1. Average frequencies of 1337 strings in piRNAs and non-piRNAs. The 1337 strings are used differently between piRNAs and non-piRNAs, and the difference is visualized by comparing the average frequencies of the 1337 strings in two groups of samples. Here, the red and black lines represent piRNAs and non-piRNAs, respectively.

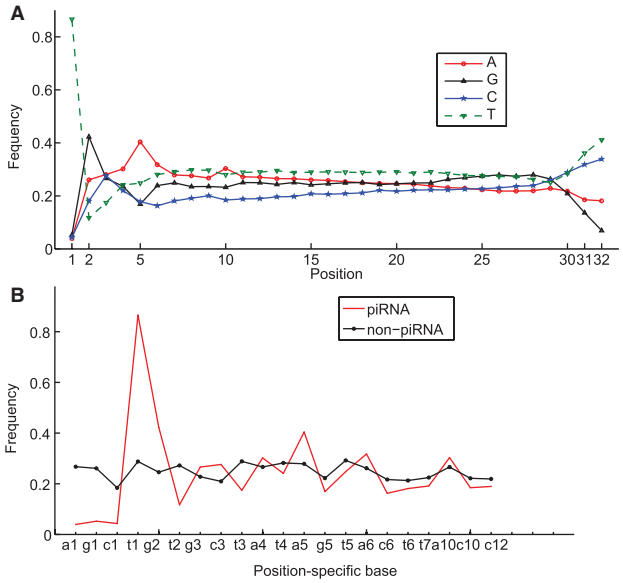

Fig. 2. (**A**) The frequencies of 32×4 position-specific bases in piRNA. Conservative base usages are found in the first 10 and the last three positions of piRNAs. (**B**) With a significance level of 10^{-100}, the usage of 21 position-specific bases is different between piRNAs and non-piRNAs. The first 10 positions, except for the 8th and 9th positions, all have conserved base usage.

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

	Predicted positives	Predicted negatives
Actual positives	TP	FN
Actual negatives	FP	TN
sn	$\mathrm{sn}=\frac{\mathrm{TP}}{\mathrm{TP+FN}}$	
sp	$\mathrm{sp}=\frac{\mathrm{TP}}{\mathrm{TP+FP}}$	

TP, True positives; FP, False positives; TN, Ture 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 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 50 000 pairs of piRNA and non-piRNA sequences derived from four species as drill set to predict piRNAs of another species. An improved Fisher formula:$f(x)=w^{\tau}x-\widetilde{m_2}-t\times N_{\mathrm{std}}$ can promote the prediction precision. As t augments from 0 to 3.4, the precisions of five cross-validation tests are significantly increased. When $t=2$, the precisions for all species are above 90% (Fig. 3A). In this study, we used the $\widetilde{m_2}+2\times N_{\mathrm{std}}$ as piRNA cutoff value, to ensure the piRNA prediction precision over 90%. To compare current algorithm with that proposed by Betel *et al.* (2007), 36 373 mouse piRNAs were taken as training positive set to predict the remaining 36 374 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 120 000 piRNAs derived from the five model species mentioned above and 120 000 non-piRNAs as drill set, we identified 87 536 locust piRNAs with length larger than 24 nt (Supplementary Material), including 12 386 gregarious-specific piRNAs, 69 151 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 87 536 predicted piRNAs are only a fraction of all locust piRNAs. Therefore, we estimated the total number of locust piRNAs is about 130 000, which is less than that of *Drosophila*'s piRNAs (Lakshmi and Agrawal, 2008). After analyzing the usage

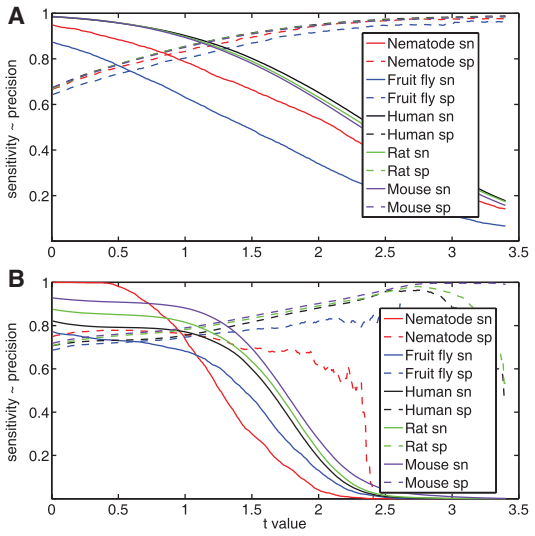

Fig. 3. The relationship of t to precision and sensitivity of 5-fold cross-validation tests. Here, 'sn' and 'sp' denotes the prediction sensitivity and precision, respectively. Nematode: *C.elegans*; fruit fly: *D.melanogaster*; rat: *R.norvegicus*; human: *H.sapiens*; mouse: *M.musculus*. (**A**) The dynamic algorithm based on string usage has an increasing precision and a decreasing sensitivity with t increasing. When $t=2$, all precisions are above 90% and most sensitivities (except for fruit fly) are 60–70%. (**B**) The static algorithm based on the position-specific base usage has an increasing precision and a decreasing sensitivity with t increasing. When $t=2.5$, most precisions reach 90%, but sensitivities are only about 10%.

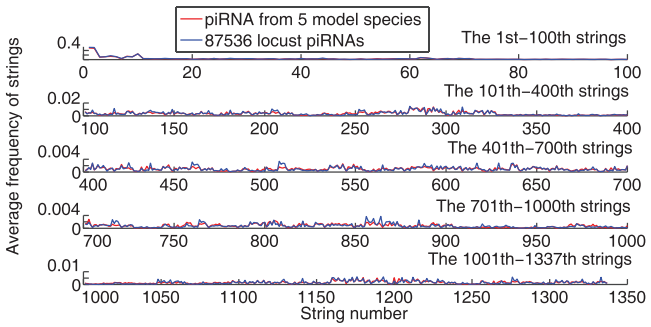

Fig. 4. Average frequencies of 1337 strings in piRNAs of the five model organisms and 87 536 locust piRNAs. The two groups of string usages are similar because the two curves are very close. Comparing with the significant difference between piRNAs and non-piRNAs as shown in the Figure 1, the algorithm firstly detects the different orientations of string usage between positive and negative samples, and then determines the sequences with amplified positive orientations as piRNAs.

of the 1337 strings in the 87 536 predicted locust piRNAs, we found that the usage of the strings in locust piRNAs are 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 the sequences with amplified positive orientations will be predicted as piRNAs. The method using k-mer frequency 'dynamic' we proposed is different from the method using position-specific base frequency 'static' method (Betel *et al.*, 2007). When

Table 2. The 15 strings that are differently used in two phases with a significance level of 10^{-100}

k-mer	Base strings
1mer	C
2mer	CC
3mer	ACC,GCC,CCA,CCC,CCT,CTC
4mer	CACC,CCTT
5mer	CTGCA,CTCTG,TCCGA,TTGCT,TTGTA

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^{-300}, while 'static' method can only identify 21 position-specific base usages with a significance level of 10^{-100}.

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 87 536 predicted piRNAs in the locust, the distribution patterns (Fig. 5B) of piRNA number versus the length and transcriptional profiling of solitary and gregarious locusts are consistent with the results reported by Wei *et al.* (2009). This result further confirms 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 piRNAs 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 $F(x)=w^{\tau}x$ 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.

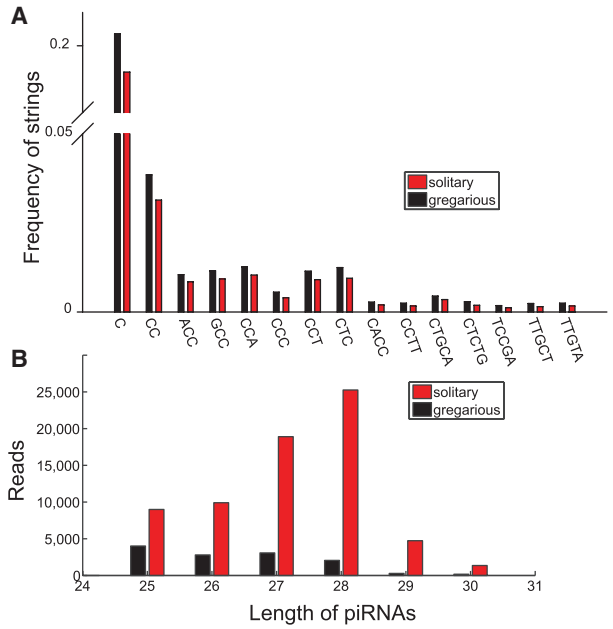

Fig. 5. The difference between solitary and gregarious locust piRNAs. (**A**) 15 strings are differently used between two phases with a significance level of 10^{-100}, and this difference is visualized by comparing the frequencies of the 15 strings in two phases. The most significant difference is the high content of C in gregarious locust piRNAs. (**B**) The distribution of piRNA reads versus length in solitary and gregarious locust. The solitary locusts always have more reads in each length than gregarious ones.

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 $\widetilde{m_2}$ and the SD N_{std} of the negative samples. Then, a sequence represented by a 1364 D vector x can be regarded as a piRNA if its $w^{\tau}x$ is larger than $\widetilde{m_2}+2\times N_{\text{std}}$. 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

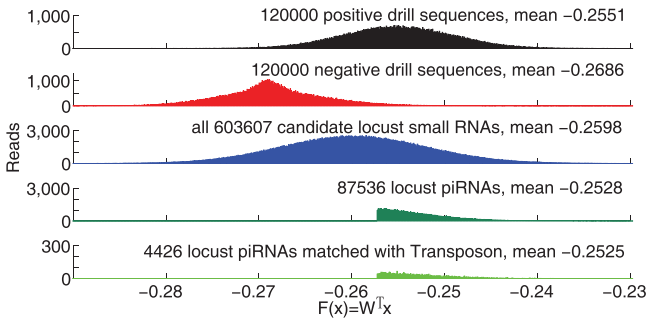

Fig. 6. The distribution of projective points representing small RNAs A small RNA sequence is first characterized by a 1364 D vector, and then further mapped to a projective point by Fisher formula $F(x)=w^{\tau}x$. This figure shows the distributions of projective points of 120 000 pairs of drill sequences (piRNAs and non-piRNAs), 603 607 deep-sequenced candidate small RNA sequences, 87 536 predicted locust piRNAs and 4426 locust piRNAs matched with the locust transposons. From the top down, the framework of this article is presented.

for studying phase transition via locust transcriptome and RNAi experiments. On the other hand, the results provided important cues understanding the molecular mechanism of fecundity difference between solitary and gregarious locusts. Most notably different from other methods in literatures, the novel prediction approach is based on the general property of string usage in piRNAs, which is extracted from all available piRNAs. We believe that this method can be widely used to predict piRNAs from both model and non-model organisms.

ACKNOWLEDGEMENTS

The authors are grateful to the associate editor and anonymous referees 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 (2006CB102000-1 to L.K.); National Natural Science Foundation of China (30830022 to L.K., 10926054 to Y.Z.); China Postdoctoral Science Foundation (20090460519 to Y.Z.); and Hebei University of Science and Technology Foundation (QD200951, XL200902 to Y.Z.); Beijing Institutes of Life Science Foundation (2010-Biols-CAS-0304 to X.W.).

Conflict of Interest: none declared.

REFERENCES

Aravin,A. *et al*. (2006) A novel class of small RNAs bind to MILI protein in mouse testes. *Nature*, **442**, 203–207.

Betel,D. *et al*. (2007) Computational analysis of mouse piRNA sequence and biogenesis. *PLoS Comput. Biol.*, **3**, 2219–2227.

Brennecke,J. *et al*. (2007) Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell*, **128**, 1089–1103.

Burge,C. *et al*. (1992) Over- and under-representation of short oligonucleotides in DNA sequences. *Proc. Natl Acad. Sci. USA*, **89**, 1358–1362.

Candolfi, E. *et al*. (1993) Determination of a new cut-off value for the diagnosis of congenital toxoplasmosis by detection of specific IgM in an enzyme immunoassay. *Enferm. Infecc. Microbiol. Clin.*, **12**, 396–398.

Cox,D.N. *et al*. (1998) A novel class of evolutionarily conserved genes defined by piwi are essential for stem cell self-renewal. *Genes Dev.*, **12**, 3715–3727.

Das,P.P. *et al*. (2008) Piwi and piRNAs Act upstream of an Endogenous siRNA Pathway to Suppress Tc3 Transposon Mobility in the Caenorhabditis elegans Germline. *Mol. Cell*, **31**, 79–90.

Fisher,R.A. (1936) The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, **7**, 179–188.

Girard,A. *et al*. (2006) A germline specific class of small RNAs binds mammalian Piwi proteins. *Nature*, **442**, 199–202.

Giulia,S. *et al*. (2009) An Ariadnes thread to the identification and annotation of noncoding RNAs in eukaryotes. *Brief Bioinform.*, **10**, 475–489.

Grivna,S.T. *et al*. (2006) A novel class of small RNAs in mouse spermatogenic cells. *Genes Dev.*, **20**, 1709–1714.

Gutierrez,G. *et al*. (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.

Houwing,S. *et al*. (2007) A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. *Cell*, **129**, 69–82.

Huang,Y.W. *et al*. (2006) Higher cut-off index value of immunoglobulin M antibody to hepatitis B core antigen in Taiwanese patients with hepatitis B. *J. Gastroenterol. Hepatol.*, **21**, 859–862.

Kandhavelu,M. *et al*. (2009) Existence of snoRNA, microRNA, piRNA characteristics in a novel non-coding RNA: x-ncRNA and its biological implication in Homo sapiens. *J. Bioinformatics Seq. Anal.*, **1**, 31–40.

Karlin, S. and Mrazek, J. (1997) Compositional differences within and between eukaryotic genomes. *Proc. Natl Acad. Sci. USA*, **94**, 10227–10232.

Karlin,S. *et al*. (1994) Heterogeneity of genomes: measures and values. *Proc. Natl Acad. Sci. USA*, **91**, 12837–12841.

Karlin,S. and Ladunga, I. (1994) Comparisons of eukaryotic genomic sequences. *Proc. Natl Acad. Sci. USA*, **91**, 12832–12836.

Kirino, Y. and Mourelatos, Z. (2007) Mouse Piwi-interacting RNAs are 2[prime]-O-methylated at their 3[prime] termini. *Nat. Struct. Mol. Biol.*, **14**, 347–348.

Lakshmi, S.S. and Agrawal, S. (2008) piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. *Nucleic Acids Res.*, **36**, 173–177.

Liu,C. *et al*. (2005) NONCODE: an integrated knowledge database of non-coding RNAs. *Nucleic Acids Res.*, **33**, D112–D115.

Madera,M. *et al*. (2010) Improving protein secondary structure prediction using a simple k-mer model. *Bioinformatics*, **26**, 596–602.

Robine,N. *et al*. (2009) A broadly conserved pathway generates 3′UTR-directed primary piRNAs. *Curr. Biol.*, **19**, 2066–2076.

Ruby,J.G. *et al*. (2006) Large-scale sequencing reveals 21U-RNAs and additional MicroRNAs and Endogenous siRNAs in C. elegans. *Cell*, **127**, 1193–1207.

Seto,A.G. *et al*. (2007) The coming of age for Piwi proteins. *Mol. Cell*, **26**, 603–609.

Thomson,T. and Lin,H.F. (2009) The Biogenesis and function of PIWI proteins and piRNAs: progress and Prospect. *Annu. Rev. Cell Dev. Biol.*, **25**, 355–376.

Vagin,V.V. *et al*. (2006) A distinct small RNA pathway silences selfish genetic elements in the germline. *Science*, **313**, 320–324.

Watanabe,T. *et al*. (2006) Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. *Genes Dev.*, **20**, 1732–1743.

Wei,Y.Y. *et al*. (2009) Characterization and comparative profiling of the small RNA transcriptomes in two phases of locust. *Genome Biol.*, **10**, R6.

Yin,H. and Lin,H.F. (2007) An epigenetic activation role of Piwi and a Piwi-associated piRNA in Drosophila melanogaster. *Nature*, **450**, 304–309.