Eur. J. Entomol. **110**(1): 13–20, 2013
http://www.eje.cz/pdfs/110/1/13
ISSN 1210-5759 (print), 1802-8829 (online)

Bioinformatics analysis on structural features of microRNA precursors in insects

JISHENG LI[1, 2], WEI FAN[1], ZHENGYING YOU[1] and BOXIONG ZHONG[1]*

[1]College of Animal Sciences, Zijingang Campus, Zhejiang University, Hangzhou 310058, P.R. China
[2]Institute of Sericulture, Chengde Medical University, Chengde 067000, P.R. China

Key words. Structural features, bioinformatics, insects, microRNA precursors

Abstract. To date, thousands of microRNAs (miRNAs) and their precursors (pre-miRNAs) have been identified in insects and their nucleotide sequences deposited in the miRBase database. In the present work, we have systematically analyzed, utilizing bioinformatics tools, the featural differences between human and insect pre-miRNAs, as well as differences across 24 insect species. Results showed that the nucleotide composition, sequence length, nucleotides preference and secondary structure features between human and insects were different. Subsequently, with the aid of three available SVM-based prediction programs, pre-miRNA sequences were evaluated and given corresponding scores. Thus it was found that of 2633 sequences from the 24 chosen insect species, 2229 (84.7%) were successfully recognized by the Mirident classifier, higher than Triplet-SVM (72.5%) and PMirP (72.6%). In contrast, four species, including the domesticated silkworm, *Bombyx mori* L., the fruit fly, *Drosophila melanogaster* Meigen, the honeybee, *Apis mellifera* L. and the red flour beetle, *Tribolium castaneum* (Herbst), were found to be largely responsible for the poor performance of some sequence matching. Compared with other species, *B. mori* especially showed the worst performance with the lowest average MFE index (0.73). Collectively these results pave the way for understanding specificity and diversity of miRNA precursors in insects, and lay the foundation for the further development of more suitable algorisms for insects.

INTRODUCTION

MicroRNAs (miRNAs) are a large class of endogenous and small non-coding RNAs approximately 22 nucleotides (nt) in length that regulate gene expression at a post-transcriptional level and play various fundamental roles in multiple biological processes, including cell differentiation, proliferation and apoptosis as well as disease processes (Ambros, 2004; Bartel, 2004; Bushati & Cohen, 2007; Wang & Li, 2007). According to recent studies, mature miRNAs are originally transcribed from a long primary miRNA (pri-miRNA) and processed into a 60–70 nt miRNA precursor (pre-miRNA) with the aid of two different enzymes, RNA polymerase II and RNase III Drosha, respectively (Lee et al., 2003, 2004). Since their initial discovery in the nematode, *Caenorhabditis elegans*, the study of miRNAs has become a rapidly growing field in the life sciences (Lee et al., 1993). Compared with miRBase 16.0 including 142 species, the latest version miRBase18.0 has grown to 18226 miRNA gene loci in 168 species and 21643 distinct mature miRNA sequences (Kozomara & Griffiths-Jones, 2011), directly attributable to the development of deep sequencing technology.

Although the founding members of miRNAs, such as lin-4 and let-7, were identified by a genetic screening approach, computational approaches still play a critical role in the identification of novel miRNAs (Griffiths-Jones, 2004; Jones-Rhoades & Bartel, 2004; Li et al., 2010; Dong et al., 2012). Previous studies reported that miRNA genes were conserved in the primary sequences and secondary structures (Gesellchen & Boutros, 2004; Nam et al., 2005; Wang et al., 2005, 2007). Thus, for most computational approaches attempting to identify miRNAs, one of the critical discoveries has been the finding that all pre-miRNAs have a stem-loop hairpin in their secondary structure, as predicted by RNAfold or Mfold (Mathews et al., 1999; Hofacker, 2003; Zuker, 2003; Unver et al., 2009). To date, more and more computational approaches have been developed and widely applied to predict miRNAs from nematodes (Lim et al., 2003b), flies (Lai et al., 2003), humans (Lim et al., 2003a) and plants (Jones-Rhoades & Bartel, 2004). However, some such studies are limited when there are no known close homologies between compared sequences or enough information about species genomes (Bentwich et al., 2005). Three programs, Triplet-SVM, PMirP and the latest Mirident classifier, which can be used to identify pre-miRNAs without utilizing comparative genomics information, are all based on support vector machine (SVM) and reported to be of superior performance in predicting pre-miRNAs from humans (Xue et al., 2005; Zhao et al., 2010; Liu et al., 2012). Triplet-SVM is based on a set of novel structure features of stem-loops, while PMirP utilizes various features to distinguish real pre-miRNAs (Xue et al., 2005; Zhao et al., 2010). Unlike them, Mirident classifier applies the software Teiresias to recognize sequence-structure motifs (ss-motifs) of different length in data sets (Liu et al., 2012). Undoubtedly, all the programs can be used to distinguish real pre-miRNAs from pseudo ones, on the other hand, this function can also be

* Corresponding author; e-mail: bxzhong@zju.edu.cn

utilized to illustrate the features differences of pre-miRNAs.

In our study, utilizing information from the miRBase18.0 database, the featural differences of human and insect pre-miRNA were systematically analyzed and compared, including composition and preference of nucleotides and secondary structure characteristics. At the same time, three programs were separately applied to predict each pre-miRNA dataset. The differences across 24 species were also compared with the intention of illustrating the diversity among them and providing more information to bioinformaticians for developing more efficient tools for predicting insect pre-miRNAs. Moreover, by comparing the performance of three programs, the study aims to help those who are engaged in miRNA research find more suitable research tools to study insects.

MATERIAL AND METHODS

Data sets

Firstly, all the miRNA precursors were downloaded from the miRNA registry database release 18.0 (November 2011) (http://www.mirbase.org/ftp.shtml). Human miRNA hairpins contain 1527 sequences, while there are 2633 sequences available for 24 insect species. In order to reduce bias caused by redundant sequences, all the sequences were gathered together and redundancies filtered (sequence identity >90%) using CD-HIT software which was originally written by Weizhong Li (http://www.bioinformatics.org/cd-hit/).

Three types of prediction software

Two of the important processes, shared by three types of software, are constructing the training models and adopting Support Vector Machine (SVM) to classify pre-miRNAs versus non-pre-miRNA hairpins. The ab initio program Triplet-SVM was kindly provided by Chenhai Xue of Tsinghua University (China). The PMirP classifier package was obtained on the web (http://ccst.jlu.edu.cn/ci/bioinformatics/MiRNA), which may be run directly on Windows with a C++ compiler. The Mirident classifier package, written by Xiuqin Liu of the Chinese Academy of Sciences, was executed with Python program support (http://www.regulatoryrna.org/pub/Mirident/). Different versions of a third-party software libsvm were downloaded and installed on the Linux system according to the manufacturer's instructions (Chang & Lin, 2011).

Statistical analysis

RNAfold (http://www.tbi.univie.ac.at/~ivo/RNA/windoze/) was used to predict the secondary structure and minimum free energy (MFE) of each sequence (Hofacker et al., 1994).

In order to compare the hairpin features between human and insects, numbers of nucleotides (A, U, C or G), MFE, length of sequence, G + C content, and A + U content were calculated and exported into an Excel file for each sequence. The Student's *t*-test was applied in order to compare nucleotide difference between human and insects. At the same time, the nucleotide frequency at each position on the pre-miRNA sequences was counted separately for analyzing nucleotides preference. According to a previous report (Zhang et al., 2006), (adjust MFE) AMFE and MFE index (MFEI) were calculated by the following equations:

$$AMFE = (MFE/\text{sequence length}) \times 100,$$

$$MFEI = AMFE/(100 \times Ratio_{G-C}),$$

$Ratio_{G-C}$ stands for the content of (G + C).

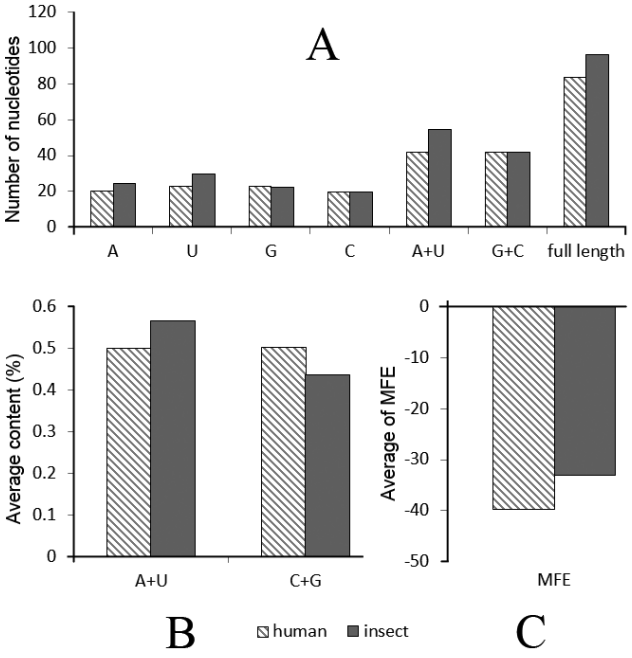

Fig. 1. Nucleotide composition and MFEs comparison of microRNA precursors (pre-miRNAs) between humans and insects.

In Table 2, the averages of MFEI for the 24 insect species were calculated separately, and then, using 0.85 as an index, numbers of pre-miRNAs with MFEI greater than 0.85 in each species were counted. Subsequently, their percentages in each species were also calculated, termed as $PPM_{0.85}$ (percentage of pre-miRNAs with MFEI higher than 0.85).

RESULTS AND DISCUSSION

Features comparison of pre-miRNA precursors between human and insects

Based on the sequences available in miRBase, stem-loop features between human and insects were compared according to nucleotide formation, contents of A-U or C-G, full length, nucleotide preference and secondary structure. After trimming high similar stem-loop sequences in human and insects separately, 1428 and 1638 non-redundant sequences were retrieved, respectively. Fig. 1A shows that the averages of two nucleotides (A or U) are different between human and insects, while the left two nucleotides G and C seem largely the same. Moreover, two datasets contained more (A+U) nucleotides than C-G pairs. The content of U is higher than others, especially in insects, which contains around 31% U compared to human (26.6%). A previous study revealed that more than 28% of the nucleotides in miRNA precursors were U, a fact which could be used to distinguish miRNAs from other RNAs (Zhang et al., 2006). A higher A-U pair content makes the pre-miRNA less stable and easier to be processed into mature miRNA. Although the sum and content of A-U in insects is obviously higher than its counterpart, two pairs (A-U and C-G) in human seem to show no obvious difference (Fig. 1B, C). Specifically, the average content of A-U in insects (56.5%) is clearly higher than for C-G (43.5%), while these are

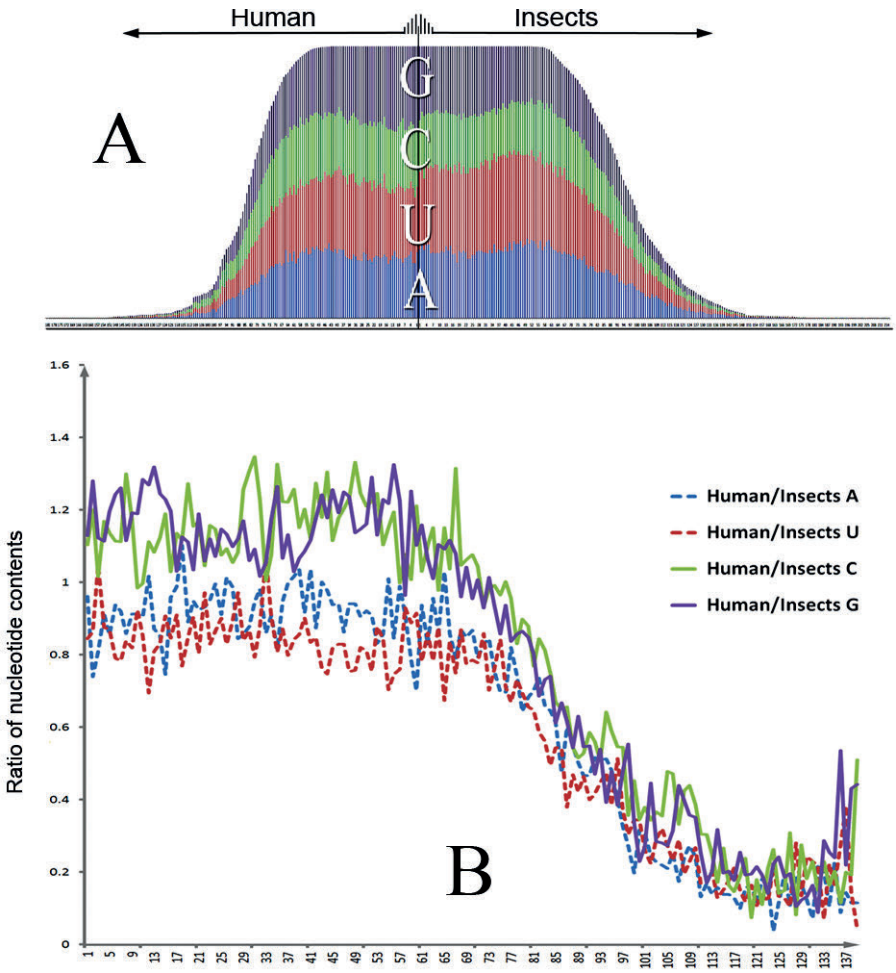

Fig. 2. Nucleotide distribution and ratios of pre-miRNAs between humans and insects. A – Contents of four nucleotides from two datasets were calculated and scattered on the X-axis according to their positions on the individual sequence. The length of waves represents sequence length, while the width represents individual contents of nucleotides on each position. From the head-head figure, the left region represents 1428 non-redundant pre-miRNAs of humans, while the right one represents 1638 non-redundant precursor sequences of insects. The X-axis represents the order of nucleotide on their pre-miRNA sequence. B – The lines drew in four colours represents the ratio of nucleotide content (RNC) for the same nucleotide between human and insects.

slightly different or even opposite in human (50.1% A-U vs 49.9% G-C). The subsequent *t*-test also illustrated that the content of the A-U pair was significantly higher in insects than for the C-G pair, compared to no difference in humans ($P < 0.001$).

Full length and minimum free energy (MFE) are two important features used to distinguish miRNA precursors. Some miRNA predictors often take them into account when distinguishing real and pseudo sequences (Xue et al., 2005; Jiang et al., 2007; Zhao et al., 2010; Liu et al., 2012). There are also differences between two datasets in both full length and MFEs (Fig. 1A, C). The average of full length in insects is longer than human, but the MFE in insects (–33.07 kcal/mol) is less negative than that in humans (–39.83 kcal/mol) (Fig. 1C). Generally, as more bonding interactions are possible in longer molecules, increasing sequence length is responsible for the more negative MFE (Seffens & Digby, 1999). However, this is not surprising as the C-G pair content in insects is lower than that in humans (Fig. 1B). Thermodynamic data illustrate that an A-U pair contributes –0.9 kcal/mol, while a C-G pair contributes –2.9 kcal/mol (Freier et al., 1986).

Thereafter, contents of four nucleotides at each position on the individual sequences from two datasets were calculated and scattered on the X-axis according to their positions on the individual sequence (Fig. 2A). From the head-to-head figure, it is seen that both share the similar colour waves of nucleotides. This result indicates that, although the width of each wave shows slightly differently between insects and humans, they have similar nucleotides preference. To better understand the differences between them, the ratio of nucleotide content (RNC) for the identical nucleotides between humans and insects was used to observe the diversity of the two datasets. For example, human/insects (A) represented the RNC of adenine between humans and insects. If the value of RNC for a nucleotide in a given position is greater than 1.0, it means that the content of this nucleotide in humans is higher than that in insects. Due to their different full lengths, 1–139 nucleotides shared by two datasets were selected for further analysis. As shown in Fig. 2B, most

TABLE 1. Performance evaluation across 24 insect species.

Insects 24 species	pre-miRNA Total/Non-redundant	Accuracy (prediction sequences)			Full length	Negative MFE
		Triplet-SVM	PMirP classifier	Mirident classifier		
Bombyx mori	487/441	0.46(483)	0.4(471)	0.69(487)	100.17	31.16
Acyrthosiphon pisum	123/101	0.85(118)	0.77(123)	0.94(123)	65.54	24.65
Aedes aegypti	101/72	0.86(97)	0.94(96)	0.92(101)	94.24	35.31
Anopheles gambiae	67/61	0.91(64)	0.95(63)	0.94(67)	95.22	37.47
Apis mellifera	174/168	0.61(170)	0.61(170)	0.81(174)	104.75	35.94
Culex quinquefasciatus	72/43	0.89(70)	0.89(70)	0.89(72)	89.36	34.25
Drosophila melanogaster	240/158	0.71(235)	0.78(226)	0.8(240)	94.58	33
Drosophila ananassae	76/36	0.89(72)	0.92(72)	0.92(76)	84.61	32.01
Drosophila erecta	81/22	0.82(78)	0.87(78)	0.91(81)	84.9	31.52
Drosophila grimshawi	82/25	0.8(82)	0.93(82)	0.94(82)	83.82	30.39
Drosophila mojavensis	71/21	0.88(69)	0.88(69)	0.94(71)	84.46	31.35
Drosophila persimilis	75/12	0.88(73)	0.9(73)	0.96(75)	85.12	32.16
Drosophila pseudoobscura	211/137	0.79(208)	0.83(197)	0.85(211)	97.22	35.54
Drosophila sechellia	78/4	0.92(72)	0.93(73)	0.92(78)	84.81	31.93
Drosophila simulans	136/53	0.79(131)	0.82(127)	0.9(136)	93.33	32.7
Drosophila virilis	74/12	0.86(72)	0.93(72)	0.95(74)	84.42	31.11
Drosophila willistoni	77/34	0.85(74)	0.89(74)	0.95(77)	85.47	30.75
Drosophila yakuba	80/1	0.89(76)	0.89(76)	0.94(80)	85.2	31.59
Heliconius melpomene	2/2	0.5(2)	1(2)	1(2)	79.5	37.2
Locusta migratoria	7/1	0.86(7)	1(7)	0.86(7)	60	25.44
Nasonia giraulti	32/0	0.9(31)	0.97(31)	0.94(32)	83.5	34.06
Nasonia longicornis	28/1	0.93(28)	0.96(27)	0.93(28)	84.64	34.02
Nasonia vitripennis	53/39	0.94(51)	1(46)	0.87(53)	93.58	39.42
Tribolium castaneum	206/194	0.77(204)	0.75(204)	0.79(206)	94.73	33.18
Total	2633	0.74(2567)	0.76(2529)	0.85(2633)		
#Total	1536	0.85(1475)	0.89(1458)	0.92(1526)		
Non-redundant sequences	1638	0.67(1582)	0.67(1556)	0.80(1638)		
#Non-redundant sequences	841	0.77(812)	0.79(801)	0.88(841)		
Redundancies	995	0.86(985)	0.90(973)	0.92(995)		
MiRNAs conserved sequences	1597	0.87(1555)	0.89(1541)	0.95(1597)		

Notes: Twenty-four insect species were used to evaluate the performance of three nucleotide analysis programs. Accuracy is the degree of closeness of a measured quantity to its actual value (number in bracket). Non-redundant sequences stand for the output of the above total of 2633 sequences in insects after filtering the redundancies (sequence identity >90%) using CD-HIT software. miRNAs conserved sequences represent stem-loop sequences whose mature miRNAs were found to be conserved in at least two insect species. The items with "#" stand for the corresponding sequences after removing four insect species, such as *Bombyx mori*, *Drosophila melanogaster*, *Apis mellifera* and *Tribolium castaneum*.

RNCs of C and G from the forward 56 nucleotides are >> 1.0, contrary to A or U. But after around 76 nucleotides, all nucleotide contents in humans seem smaller than those in insects. This result means that the RNC of pre-miRNAs between humans and insects are profoundly different, and this difference might well be caused by insect species diversity.

Global comparison and evaluation

Although three published programs had made great strides in human hairpin prediction, there was evidence that different features used for pre-miRNA detection could largely influence the performance of an algorithm (Jiang et al., 2007; Liu et al., 2012). We assumed that difference of prediction performance could be served as an index to directly reflect the difference of stem-loop structure. On the other hand, it was also necessary to evaluate the feasibility and performance of an algorithm before employing it for different cases.

As shown in Table 1, Mirident classifier shows the best performance, achieving an overall accuracy of 84.7% for 2633 sequences tests. There are 1909 out of 2567 pre-miRNAs correctly recognized by Triplet-SVM and 1913 out of 2529 sequences successfully detected by PMirP classifier, which give accuracies of 74.4% and 75.6%, respectively. In order to decrease the bias caused by the redundancies, non-redundant sequences were employed to evaluate the performance of three programs. Result showed that 1315 out of 1638 sequences were correctly recognized by Mirident classifier, which still kept the first place with an accuracy rate of 80.3% higher than that in Triplet-SVM (67.0%) and PMirP classifier (66.8%). And of those successfully recognized precursors, only 835 pre-

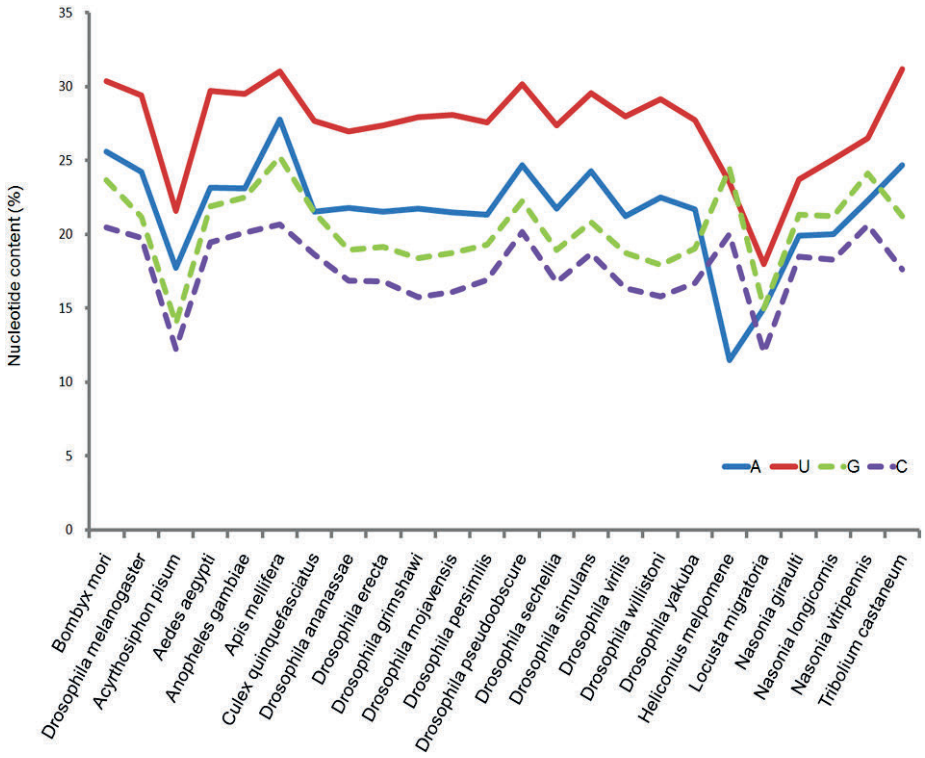

Fig. 3. Change tendency of four nucleotides in 24 insect species. The X-axis represents 24 species, the Y-axis, nucleotide contents.

miRNAs were jointly detected by the three programs, accounting for about 51% prediction.

From global evaluation of the three programs, all of them showed an obvious decline compared to the 2633 sequences. One possible reason for this was the fact that 995 highly similar sequences of the 24 insect species discarded by the CD-hit procedure possessed a large share of correctly detectable sequences. This being so, we also predicted these sequences using the three programs. Our results showed that 914 out of the 995 sequences were successfully recognized by Mirident classifier with the accuracy rate of 91.9%, still higher than that using Triplet-SVM (86.2%) and PMirP classifier (89.8%). These data confirmed our assumption and further demonstrated that these highly similar sequences shared similar features with human stem loop structures (Table 1). In addition, some pre-miRNA sequences detected using Triplet-SVM and PMirP processing were deleted before prediction, because of undesirable multiple loops according to individual designing principle. Although concerned with sequence structures, this data pretreatment cause problems with subsequent results processing compared to the Mirident classifier system. Nevertheless, compared to previous tests on human precursor sequences, Triplet SVM and PMirP classifier produced mis-diagnoses in insects. Given that the test sequences were composed of 24 insect species, a plausible interpretation is that two of them primarily relied on the features of human pre-miRNAs such as base-pair, full length and MFE, etc. Unlike them, Mirident classifier seems to be more suitable for predicting the pre-miRNA of insects. According to the analysis of nucleotides preference, especially in Fig. 2A, a similar tendency in human and insects pre-miRNAs is consistent with their initial designing conception on ss-motif of variable length, without considering sequence-structure features of fixed size (Liu et al., 2012).

In contrast, highly similar precursor sequences did not mean that their microRNAs were also conserved; this is because lots of microRNAs were almost identical while their precursors differed. For more objective and accurate evaluating of the three programs, 1597 stem-loop sequences whose mature miRNAs were conserved in at least two insect species were selected and predicted using the three softwares. The results as shown in Table 1 revealed of 1597 sequences, 1521 were successfully recognized by Mirident classifier, with the accuracy rate 95.2%, still higher than that in triplet SVM (87.9%) and PMirP classifer (89.7%). As compared with the above performances in 995 high similar precursor sequences, two datasets showed similar trends.

Collectively, the global performance comparison illustrated that different precursor features between human and insect miRNAs must influence prediction performance. Perhaps, species differences in insects were the main reason leading to the bad prediction performance obtained. To illustrate this further studies were still required in order to compare and analyze the diversity across the 24 insect species.

Comparison and analysis across 24 species

As shown in Table 1, the accuracy rates for the 24 insect species were obtained and summarized according to three prediction algorithms. Most recognized rates reached >> 80%, except in the case of several species

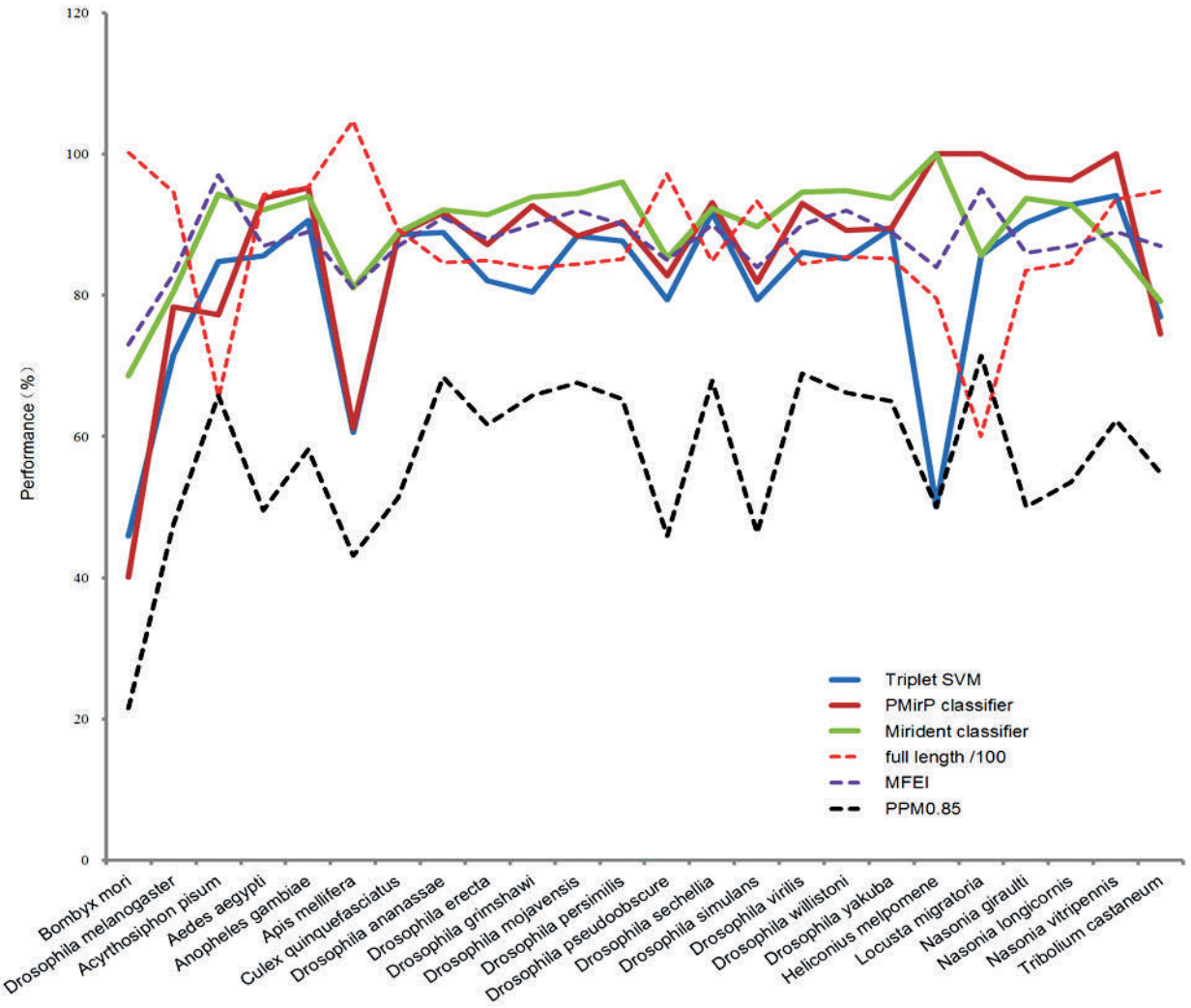

Fig. 4. Performance comparisons of three predictors through testing 24 insect species. The full lines represent the accuracy rates of three predictors, while the dotted line stands for percentage of full length, MFEI and percentage of pre-miRNAs with MFEI higher than 0.85 ($PPM_{0.85}$). The X-axis contains 24 insect species, which come from miRBase 18.0.

including *B. mori*, *D. melanogaster*, *A. mellifera* and *T. castaneum* which are largely responsible for the bad performance noted. This is especially so for 961 non-redundant pre-miRNA precursors from these four species, representing around 59% of 1638 non-redundant sequences. After removing these from the global set, the accuracy rates of the three test programs were individually raised to higher levels, which were closer to previous human tests. Our results demonstrated that many non-redundant stem-loop sequences in these four species might well vary from those of other species in the sample of 24 species. Consequently, nucleotides formations, full length and MFE for 24 species were calculated in order to analyze such potential differences. Fig. 3 illustrates that four nucleotides in the 24 species have almost the same change tendency. Nearly all the U nucleotides in these species are in excess of other nucleotides, with the exception of the butterfly, *Heliconius melpomene* (L.) and the locust, *Locusta migratoria* (L.). The contents of the four nucleotides reached a minimum value in the aphid, *Acyrthosiphon pisum* (Harris) and *L. migratoria*. Furthermore, the negative MFEs also fell to the lowest values noted (Table 1). This demonstrates that stem-loop structure feature of the two species is probably very different from that of the other species in the collection. Moreover and surprisingly, compared with full sequence length, the accuracy rates of the three programs most likely expressed opposite change tendencies (Fig. 4). This phenomenon indicates that full sequence length might affect the performance of the three chosen predictors.

Zhang and co-workers (Zhang et al., 2006) developed a new term named *minimum free energy index* (MFEI) to detect different types of RNA in plants. Their results showed that when the MFEI was >> 0.85, the sequence was most likely to be real miRNA. Using this concept, we employed MFEI to disclose the differences across the miRNAs of the 24 insect species tested. As shown in Table 2, most species had average MFEIs >> 0.85, except for five species, i.e. *B. mori*, *D. melanogaster*, *D. simulans* (Sturtevant) *A. mellifera*, and *H. melpomene*. Interestingly, three species – *B. mori*, *D. melanogaster* and *A. mellifera* not only had lower MFEIs, but also accounted for three quarters of the bad performance of species noted. Accordingly, more than half of pre-miRNAs in 17

TABLE 2. Nucleotide characteristics of 24 insect species.

Species	A	U	G	C	(A+U) %	(G+C) %	Full length	MFE	MFEI
Bombyx mori	25.62±8.69	30.39±10.19	23.67±6.85	20.49±6.94	55.91	44.09	100.17±21.61	–31.16±10.23	0.73
Acyrthosiphon pisum	17.72±3.94	21.57±4.26	13.98±3.54	12.28±3.93	59.93	40.07	65.54±8.95	–24.65±6.26	0.97
Aedes aegypti	23.16±5.85	29.73±7.29	21.9±6.08	19.45±6.4	56.12	43.88	94.24±20.82	–35.31±8.15	0.87
Anopheles gambiae	23.12±5.22	29.49±5.73	22.49±4.46	20.12±6.3	55.25	44.75	95.22±15.56	–37.47±7.48	0.89
Apis mellifera	27.78±9.32	31.05±9.51	25.25±7.14	20.68±6.48	56.15	43.85	104.75±18.64	–35.94±9.84	0.81
Culex quinquefasciatus	21.54±5.37	27.67±6.06	21.51±4.79	18.64±5.57	55.07	44.93	89.36±15.63	–34.25±7.51	0.87
Drosophila melanogaster	24.22±7.7	29.43±7.45	21.18±7.22	19.75±8.14	56.72	43.28	94.58±22.98	–33±10.99	0.83
Drosophila ananassae	21.8±4.87	26.97±4.27	18.95±4.3	16.88±4.24	57.65	42.35	84.61±11.26	–32.01±5.83	0.91
Drosophila erecta	21.52±4.93	27.38±4.37	19.16±4.21	16.84±3.76	57.60	42.40	84.9±11.41	–31.52±6.58	0.88
Drosophila grimshawi	21.73±5.31	27.93±4.48	18.38±4.09	15.78±4.04	59.25	40.75	83.82±11.65	–30.39±5.7	0.90
Drosophila mojavensis	21.51±5.11	28.1±4.72	18.73±4.32	16.13±4.19	58.73	41.27	84.46±11.95	–31.35±5.7	0.92
Drosophila persimilis	21.35±4.35	27.57±4.64	19.29±4.08	16.91±4.13	57.47	42.53	85.12±10.89	–32.16±5.5	0.90
Drosophila pseudoobscura	24.68±6.43	30.16±6.41	22.23±6.13	20.15±6.46	56.41	43.59	97.22±15.07	–35.54±10.37	0.85
Drosophila sechellia	21.73±5.01	27.36±4.39	18.94±4.21	16.78±3.87	57.88	42.12	84.81±11.75	–31.93±6.52	0.90
Drosophila simulans	24.27±6.03	29.54±6	20.85±5.42	18.68±5.6	57.66	42.34	93.33±16.15	–32.7±7.67	0.84
Drosophila virilis	21.26±5.27	28±4.63	18.77±4.31	16.39±4.33	58.35	41.65	84.42±12.1	–31.11±6.33	0.90
Drosophila willistoni	22.53±4.91	29.14±5.8	17.96±3.74	15.83±3.78	60.47	39.53	85.47±12.48	–30.75±5.95	0.92
Drosophila yakuba	21.7±4.9	27.75±4.49	19.05±4.11	16.7±3.8	58.04	41.96	85.2±11.11	–31.59±6.23	0.89
Heliconius melpomene	11.5±2.12	23.5±0.71	24.5±3.54	20±2.83	44.03	55.97	79.5±2.12	–37.2±1.13	0.84
Locusta migratoria	15±1.73	18±3.21	15±1.15	12±3.06	55.00	45.00	60±1.63	–25.44±2.37	0.95
Nasonia giraulti	19.91±3.81	23.75±4.08	21.34±4.15	18.5±3.15	52.29	47.71	83.5±8.56	–34.06±6.48	0.86
Nasonia longicornis	20±4.35	25.11±3.98	21.25±3.52	18.29±3.85	53.28	46.72	84.64±8.54	–34.02±4.79	0.87
Nasonia vitripennis	22.3±4.7	26.53±4.64	24.11±4.38	20.64±4.22	52.18	47.82	93.58±9.79	–39.42±5.71	0.89
Tribolium castaneum	24.69±8.03	31.19±9.33	21.23±6.82	17.62±5.34	59.00	41.00	94.73±21.5	–33.18±9.68	0.87

Note: Table 2 shows the average number of each term such as four nucleotides, C-G pair, A-U pair, full sequence length, minimum free energy (MFE) and minimum free energy index (MFEI). The following "± number" = individual standard deviation (STDV).

species had MFEIs >> 0.85, while only 21.6% miRNA precursors did so in *B. mori* (Fig. 4). Considering the curves shown in Fig. 4, $PPM_{0.85}$ almost has the same change tendency with performances of Triplet-SVM and PMirP classifier. The result illustrates that the values of MFEIs indeed influence the performance of the two programs. More importantly, the over-reliance on C-G content, full sequence length and MFE might be the main reason leading to the poor accuracy rates of two algorisms in some insect species. Lastly, compared with their report (Zhang et al., 2006), more than 90% of pre-miRNAs had an MFEI >> 0.85, but in our study, 71.4% for *L. migratoria* was the highest ratio observed (Fig. 4). This lack of uniformity might arise due to differences between plants and insects. Thus, MFEI might be a good, but not absolute parameter, for detecting pre-miRNAs; however, the average for total MFEIs could better serve as an important index to describe pre-miRNA features of a given species.

Despite stem-loop hairpin structure not being a unique feature of miRNA, it is nevertheless still the most important factor enabling definition of pre-miRNAs (Ambros et al., 2003). Full sequence length and MFE reflect the most immediate features of a pre-miRNA, while nucleotides preference leads to diversity and specificity among species. Based on the close relation between MFEI and the performances of two programs, Triplet-SVM and PMirP classifier, the lower MFEI appeared to be a key factor leading to bad prediction performance in several insect species tested. Perhaps the training models of the two programs placed over-reliance on the features of human pre-miRNAs. A more suitable and effective training model based for insect is presently under consideration and improvement. In the case of *B. mori* especially, this becomes more interesting because of the poor performance (Fig. 4), lowest average MFEI (0.73) and smallest PPM0.85 (21.6%) according to the above analyses. On the other hand, it cannot be ignored that potential error might be caused by false positive pre-miRNAs in miRBase. As is well known, many such errors arise as a result of computational methodology without extra experimental confirmation and validation (Griffiths-Jones, 2004).

CONCLUSION

Taken as a whole, developments of deep sequencing and bioinformatics have greatly boosted research into miRNAs. Butas found by us, one important step is clearly the selection of s more suitable tools to distinguish real miRNAs from pseudo ones. Given the performance of Mirident classifier in predicting pre-miRNAs, general sequence features such as ss-motifs should be taken as important factors to construct new tools or algorithms for pre-miRNA prediction. Furthermore, based on the features understanding of known pre-miRNA sequences of insects, a better and more specific training model is still

required for future studies. For those insect species whose genomes are presently unknown, such bioinformatics algorithms may play important roles in discovering many more useful miRNAs.

ACKNOWLEDGEMENTS. We thank C. Xue at MOE Key Laboratory of Bioinformatics at Tsinghua University for advice, F. Gong and X. Liu at National Center for Mathematics and Interdisciplinary Sciences of the Chinese Academy of Sciences for their technical assistance, and H.D. Loxdale for his editorial help in improving the text. This work was supported by the National Basic Research Program of China (Grant No. 2012CB114601) and the National Nature Science Foundation of China (Grant No. 30972142).

REFERENCES

AMBROS V. 2004: The functions of animal microRNAs. — *Nature* **431**: 350–355.

AMBROS V., BARTEL B., BARTEL D.P., BURGE C.B., CARRINGTON J.C., CHEN X., DREYFUSS G., EDDY S.R., GRIFFITHS-JONES S., MARSHALL M., MATZKE M., RUVKUN G. & TUSCHL T. 2003: A uniform system for microRNA annotation. — *RNA* **9**: 277–279.

BARTEL D.P. 2004: MicroRNAs: genomics, biogenesis, mechanism, and function. — *Cell* **116**: 281–297.

BENTWICH I., AVNIEL A., KAROV Y., AHARONOV R., GILAD S., BARAD O., BARZILAI A., EINAT P., EINAV U., MEIRI E., SHARON E., SPECTOR Y. & BENTWICH Z. 2005: Identification of hundreds of conserved and nonconserved human microRNAs. — *Nat. Genet.* **37**: 766–770.

BUSHATI N. & COHEN S.M. 2007: MicroRNA functions. — *Annu. Rev. Cell Dev. Biol.* **23**: 175–205.

CHANG C.C. & LIN C.J. 2011: LIBSVM: a library for support vector machines. — *ACM Trans. Intell. Syst. Technol.* **2**: 27.

DONG Q.H., HAN J., YU H.P., WANG C., ZHAO M.Z., LIU H., GE A.J. & FANG J.G. 2012: Computational identification of microRNAs in strawberry expressed sequence tags and validation of their precise sequences by miR-RACE. — *J. Heredity* **103**: 268–277.

FREIER S.M., KIERZEK R., JAEGER J.A., SUGIMOTO N., CARUTHERS M.H., NEILSON T. & TURNER D.H. 1986: Improved free-energy parameters for predictions of RNA duplex stability. — *Proc. Natl. Acad. Sci. U.S.A*. **83**: 9373–9377.

GESELLCHEN V. & BOUTROS M. 2004: Managing the genome: microRNAs in *Drosophila*. — *Differentiation* **72**: 74–80.

GRIFFITHS-JONES S. 2004: The microRNA registry. — *Nucl. Acids Res*. **32**: D109–111.

HOFACKER I.L. 2003: Vienna RNA secondary structure server. — *Nucl. Acids Res*. **31**: 3429–3431.

HOFACKER I.L., FONTANA W., STADLER P.F., BONHOEFFER L.S., TACKER M. & SCHUSTER P. 1994: Fast folding and comparison of RNA secondary structures. — *Mh. Chem. / Chem. Month.* **125**: 167–188.

JIANG P., WU H., WANG W., MA W., SUN X. & LU Z. 2007: MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. — *Nucl. Acids Res.* **35**: W339–W344.

JONES-RHOADES M.W. & BARTEL D.P. 2004: Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. — *Mol. Cell* **14**: 787–799.

KOZOMARA A. & GRIFFITHS-JONES S. 2011: miRBase: integrating microRNA annotation and deep-sequencing data. — *Nucl. Acids Res*. **39**: D152–157.

LAI E.C., TOMANCAK P., WILLIAMS R.W. & RUBIN G.M. 2003: Computational identification of *Drosophila* microRNA genes. — *Genome Biol.* **4**: R42.

LEE R.C., FEINBAUM R.L. & AMBROS V. 1993: The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. — *Cell* **75**: 843–854.

LEE Y., AHN C., HAN J., CHOI H., KIM J., YIM J., LEE J., PROVOST P., RADMARK O., KIM S. & KIM V.N. 2003: The nuclear RNase III Drosha initiates microRNA processing. — *Nature* **425**: 415–419.

LEE Y., KIM M., HAN J., YEOM K.H., LEE S., BAEK S.H. & KIM V.N. 2004: MicroRNA genes are transcribed by RNA polymerase II. — *EMBO J.* **23**: 4051–4060.

LI L., XU J., YANG D., TAN X. & WANG H. 2010: Computational approaches for microRNA studies: a review. — *Mamm. Genome* **21**: 1–12.

LIM L.P., GLASNER M.E., YEKTA S., BURGE C.B. & BARTEL D.P. 2003a: Vertebrate microRNA genes. — *Science* **299**: 1540.

LIM L.P., LAU N.C., WEINSTEIN E.G., ABDELHAKIM A., YEKTA S., RHOADES M.W., BURGE C.B. & BARTEL D.P. 2003b: The microRNAs of *Caenorhabditis elegans*. — *Genes Dev.* **17**: 991–1008.

LIU X., HE S., SKOGERBO G., GONG F. & CHEN R. 2012: Integrated sequence-structure motifs suffice to identify microRNA precursors. *PLoS One* **7**: e32797.

MATHEWS D.H., SABINA J., ZUKER M. & TURNER D.H. 1999: Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. — *J. Mol. Biol.* **288**: 911–940.

NAM J.W., SHIN K.R., HAN J., LEE Y., KIM V.N. & ZHANG B.T. 2005: Human microRNA prediction through a probabilistic co-learning model of sequence and structure. — *Nucl. Acids Res.* **33**: 3570–3581.

SEFFENS W. & DIGBY D. 1999: mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. — *Nucl. Acids Res.* **27**: 1578–1584.

UNVER T., NAMUTH-COVERT D.M. & BUDAK H. 2009: Review of current methodological approaches for characterizing microRNAs in plants. — *Int. J. Plant Genomics* **2009**: 262463.

WANG Q.L. & LI Z.H. 2007: The functions of microRNAs in plants. — *Front Biosci.* **12**: 3975–3982.

WANG X., ZHANG J., LI F., GU J., HE T., ZHANG X. & LI Y. 2005: MicroRNA identification based on sequence and structure alignment. — *Bioinformatics* **21**: 3610–3614.

WANG Y., STRICKER H.M., GOU D. & LIU L. 2007: MicroRNA: past and present. — *Front Biosci.* **12**: 2316–2329.

XUE C., LI F., HE T., LIU G.P., LI Y. & ZHANG X. 2005: Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. — *BMC Bioinform.* **6**: 310.

ZHANG B., PAN X., COX S., COBB G. & ANDERSON T. 2006: Evidence that miRNAs are different from other RNAs. — *Cell. Mol. Life Sci.* **63**: 246–254.

ZHAO D., WANG Y., LUO D., SHI X., WANG L., XU D., YU J. & LIANG Y. 2010: PMirP: A pre-microRNA prediction method based on structure-sequence hybrid features. — *Artificial Intell. Medicine* **49**: 127–132.

ZUKER M. 2003: Mfold web server for nucleic acid folding and hybridization prediction. — *Nucl. Acids Res.* **31**: 3406–3415.

Received July 4, 2012; revised and accepted August 19, 2012