J Inf Process Syst, Vol.13, No.6, pp.1527~1543, December 2017
https://doi.org/10.3745/JIPS.04.0051

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

miRNA Pattern Discovery from Sequence Alignment

Xiaohan Sun*,** and Junying Zhang*

Abstract

MiRNA is a biological short sequence, which plays a crucial role in almost all important biological process. MiRNA patterns are common sequence segments of multiple mature miRNA sequences, and they are of significance in identifying miRNAs due to the functional implication in miRNA patterns. In the proposed approach, the primary miRNA patterns are produced from sequence alignment, and they are then cut into short segment miRNA patterns. From the segment miRNA patterns, the candidate miRNA patterns are selected based on estimated probability, and from which, the potential miRNA patterns are further selected according to the classification performance between authentic and artificial miRNA sequences. Three parameters are suggested that bi-nucleotides are employed to compute the estimated probability of segment miRNA patterns, and top 1% segment miRNA patterns of length four in the order of estimated probabilities are selected as potential miRNA patterns.

Keywords

Deep Sequencing Data, miRNA, Pattern Discovery

1. Introduction

MicroRNAs (miRNAs) are short RNA sequences which regulate mRNA translation, play a crucial regulation role in nearly all important biological processes [1,2], and make fine-scale adjustments to protein outputs by regulating target mRNAs [3]. Identification of novel miRNAs are pivotal to understand the development mechanism of an organism and explore the pathogen of complex diseases. MiRNA computational identification methods have been improved unprecedentedly in sensitivity due to deep sequencing technologies, but many of them are still compromised by substantial false positives and low efficiency [4,5]. MiRNA identification methods based on high-throughput sequencing technologies regularly encompass the two steps: mapping the reads into genome or RNA database and analyze the stem-loop structure of a candidate sequence. The former is a heavy computational burden [5-7], and the latter is difficult to remove the pseudo-miRNAs with similar stem-loop structure [8]. A fast filtration of reads to remove spurious miRNA sequences will reduce the mapping time and false positives in the miRNA identification process.

Sequence patterns are a widely used feature for classification and they could be used to discover the specific sequences for the conservative knowledge implied in these patterns. MiRNA patterns are a set

Manuscript received January 6, 2017; second revision June 13, 2017; accepted June 21, 2017.
Corresponding Author: Junying Zhang (jyzhang@mail.xidian.edu.cn)
* School of Computer Science and Technology, Xidian University, Xi'an, China (sxhjpj@sina.com, jyzhang@mail.xidian.edu.cn)
**School of Network Security and Informationization, Weinan Normal University, Weinan, China (sxhjpj@sina.com)

of distinguishing subsequences of authentic miRNA sequences which appear frequently in the authentic miRNA sequences, yet infrequently in non-miRNA sequences [9], and they are the features selected by some miRNA identification methods for higher accurate classification, less computational time [10]. There has been many different supervised feature selection techniques for sequence analysis, such as content analysis focusing on the broad characteristics of sequences [11], signal analysis concentrating on identification of sequence motifs [11], automatic feature generation for sequences integrating feature construction and feature generation in a systematic way [12]. Due to the short length and little construction information, it is a challenging take to extract patterns from mature miRNA sequences.

In this paper, we propose an approach to discover miRNA patterns from sequence alignment and suggest three parameters to select the potential miRNA patterns. We first extract the common letters of multiple mature miRNA sequences as a primary miRNA pattern [13-15]. The primary miRNA patterns are then cut into short segments, from which we select the segments which appear in more authentic mature miRNA sequences than in non-miRNA ones as potential miRNA patterns. For balancing computation burden and classification performance, three parameters for selecting miRNA patterns are suggested that employing bi-nucleotides to compute the probability of a segment miRNA pattern, and selecting top 1% segment miRNA patterns of length four as potential miRNA patterns.

This paper is organized as follows: Section 2 introduces the materials and suggested approach to produce and select miRNA patterns. In Section 3, we analyze and discuss the classification performance, computation burden and significance of the potential miRNA patterns. Section 4 concludes this paper.

2. Materials and Methods

The proposed approach includes two stages: pattern discovery and pattern selection (Fig. 1).

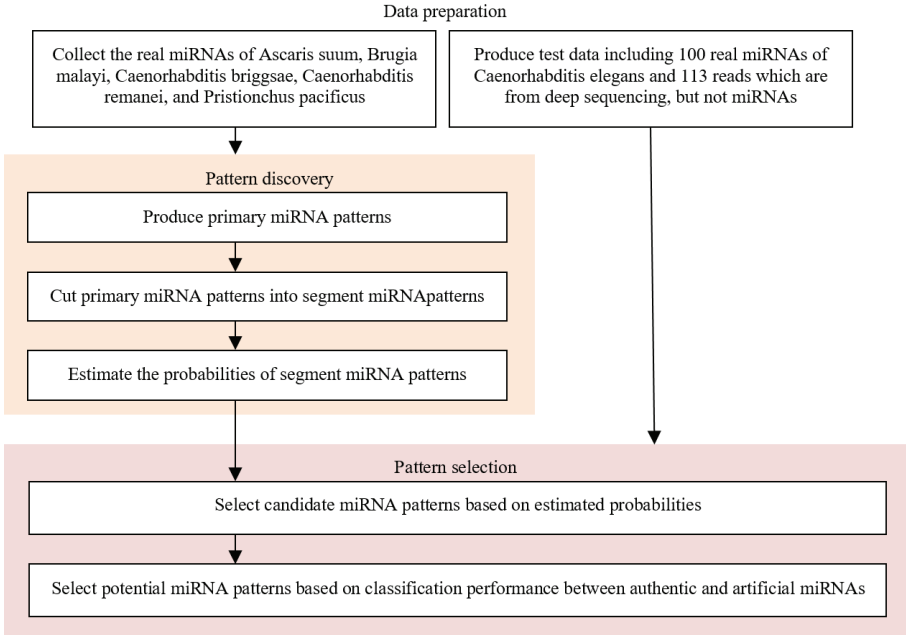

Fig. 1. Flowchart of the approach.

2.1 Data Preparation

Mature miRNA sequences of nematode (*Ascaris suum*, *Brugia malayi*, *Caenorhabditis briggsae*, *Caenorhabditis remanei*, and *Pristionchus pacificus*) are downloaded from miRBase (miRNA database, release 17) and used to produce patterns. The test data contains 100 randomly chosen authentic mature miRNAs of *Caenorhabditis elegans* and 113 artificial ones which are chosen randomly from the reads of *Caenorhabditis elegans* downloaded from miRDeep (a software to identify miRNAs from deep sequencing) [5].

2.2 Pattern Discovery

2.2.1 Production of primary miRNA patterns

Primary miRNA patterns are the common letters of multiple mature miRNA sequences. Here, Needleman-Wunsch (NW) alignment algorithm is employed to produce a primary miRNA pattern of any two miRNA sequences. We adopt NW algorithm, rather than some multiple sequences alignment methods, such as HandAlign [16], Phylo [17], and SINA [18], for the simplicity and global optimality of NW algorithm. The alignment results of two miRNA sequences are primary miRNA patterns which present in the forms that comprise a combination of literals (any one of bases A, C, G, and U) and wildcards (each denoted by "."). One such primary miRNA pattern is “AC..U”, all instances of which have their first and second positions occupied by bases A and C, the third and fourth positions by any two bases, and the fifth position by the base U.

2.2.2 Production of segment miRNA patterns

The liberals in a primary miRNA pattern are the common letters in the specific positions of the two mature miRNA sequences, but they might be non-conservative bases because the two mature miRNA sequences might happen to have the same letters in some positions. For extracting the true conservative bases, primary miRNA patterns are cut into short segments which contain m (m=2, 3, ..., 24) literals where 24 is the maximum of literals in a primary miRNA pattern. We define the short segments as segment miRNA patterns and pattern length as the number of literals in such a segment. We thus get 24 groups of segment miRNA patterns (23 groups of segment miRNA patterns whose length is from 2 to 24 and the group of primary miRNA patterns).

2.2.3 Probability estimation of segment miRNA patterns

We employ the method in reference [19] to compute estimated probability of a segment miRNA pattern. First, the frequency of x-nucleotides is computed based on mature miRNA sequences. X-nucleotides could be bi-nucleotides containing two literals, or tri-nucleotides containing three literals, etc. The frequencies of all bi-nucleotides are computed in the following forms:

"AA"
"A.A"
...

"U.....U"
And the frequencies of all tri-nucleotides are computed:
"AAA"
"A.AA"
...
"AA..A"
...
"UU.....U"

Second, the probability of a given segment miRNA pattern is estimated by Bayes' theorem and second order Markov chain.

Based on the frequencies of bi-nucleotides, the probability of a segment miRNA pattern "AU.CG" is computed as follows:

Pr(AU.CG)=Pr(AU/A])*Pr(U.C/U)*Pr(CG/C)=#(AU)/(#(AA)+#(AC)+#(AG)+#(AU))*
#(U.C)/(#(U.A)+#(U.C)+#(U.G)+#(U.U))*#(CG)/(#(CA)+#(CC)+#(CG)+#(CU))

Based on tri-nucleotides, its probability is computed:

Pr(AU.CG)=Pr(AU.C/AU])*Pr(U.CG/U.C)=#(AU.C)/(#(AU.A)+#(AU.C)+#(AU.G)+#(AU.U))*
#(U.CG)/(#(U.CA)+#(U.CC)+#(U.CG)+#(U.CU))

2.3 Pattern Selection

2.3.1 Selection of candidate miRNA patterns

For each group, the segment miRNA patterns are sorted in the order of the estimated probabilities, and those whose estimated probabilities above a given threshold are selected as candidate patterns. The threshold of each group is determined in the following steps:

First, we compute the negative base-10 logarithms of the estimated probabilities of all the segment miRNA patterns, and all the estimated probabilities constitute the threshold space.

Second, we count respectively authentic and artificial miRNA sequences which contain a segment miRNA pattern.

Third, we compute separately true positive (TP), false negative (FN), true negative (TN), false positive (FP) [20] for each group of segment miRNA patterns. In each group, TP, FP, TN and FN are firstly initialized to 0. When the negative logarithm of estimated probability of a segment pattern is less than a given threshold, we increase TP by one if the segment pattern is contained in more authentic miRNAs, otherwise increase FP by one. Similarly, we compute TN and FN. After traversing all the segment miRNA patterns in a group, true positive rate ($TPR = \frac{TP}{\mathrm{TP+FN}}$), and false positive rate ($FPR = \frac{FP}{FP+TN}$) of the threshold can be calculated. We walk through the whole threshold space of a group and compute TPR and FPR based on every threshold, then a ROC curve can be drawn for this group. In a ROC curve, the point which is nearest to the upper left corner ($\sqrt{(1-TPR)^2 + FPR^2}$) is selected as the final threshold, and the segment patterns whose negative logarithm is less than the final threshold are selected as candidate miRNA patterns of the group. Thus, we get 24 groups of candidate patterns.

2.3.2 Selection of potential miRNA patterns

The candidate miRNA patterns in the 24 groups are the patterns which contain most conservative knowledge in a specific length, from which we need to further select the group that has the highest classification performance between authentic and artificial miRNAs as the potential miRNA patterns.

We also use ROC curve to compare the classification performance of the 24 groups of candidate miRNA patterns. For each group, we firstly count candidate miRNA patterns contained in each sequence in the test data. We then divide the number of candidate miRNA patterns contained in a sequence by the length of the sequence to eliminate the deviation that longer sequences have higher possibility to contain more patterns. The all the quotients of a group of candidate miRNA patterns constitute the threshold space of classification performance, and we also compute the TP, FP, TN, FN, TPR, and FPR and draw the ROC curve of the group. We obtain the threshold which present the highest classification performance from the ROC curve of each group, then select the group with the highest classification performance as potential miRNA patterns from the 24 ROC curves.

3. Results and Discussion

The following three parameters should be selected prudently for balancing computation burden and classification performance.

3.1 X-nucleotides of Estimating Probabilities of Segment Patterns

The estimated probability of a segment miRNA pattern is computed according to the frequencies of the x-nucleotides. X-nucleotides could be bi-nucleotides, tri-nucleotides, etc. In fact, only bi-/tri-nucleotides can be used to estimate the probabilities because the frequencies of most x-nucleotides is 0 when x≥4. Take 4-nucleotides for example, more than half 4-nucleotides have a frequency of 0 which results in the probabilities of many segment miRNA patterns are also 0. As for bi-nucleotides and tri-nucleotides, the former is better because the number of bi-nucleotides (400) is far less the number of tri-nucleotides (20800) and the classification performance of the former is also a little better than the latter (Table 1).

3.2 Percentage of Candidate miRNA Patterns

The higher estimated probability a segment miRNA pattern has, the more conservative knowledge the segment miRNA pattern contains. The top segment miRNA patterns in the order of estimated probabilities in a group present higher classification performance between authentic and artificial miRNA sequences, and we need to determine the percentage of selected segment miRNA patterns. TPR, FPR and $\sqrt{(1-TPR)^2+FPR^2}$ of the top n% segment miRNA patterns in the groups whose length range is from four to seven are shown in Tables 2–5 and Figs. 2–5.

From the Figs. 2–5, we find the top 0.5%–60% segment miRNA patterns in all the groups are good and stable in classification performance. Taking computation burden into account, we suggest selecting top 1% segment miRNA patterns as potential miRNA patterns.

Table 1. Classification performance of miRNA patterns based on bi-/tri-nucleotides

Length of segment miRNA patterns	bi-nucleotides			tri-nucleotides		
	TPR	FPR	$\sqrt{(1-TPR)^2+FPR^2}$	TPR	FPR	$\sqrt{(1-TPR)^2+FPR^2}$
Full length	0.6	0.5	0.60	0.71	0.54	0.61
2	0.8	0.3	0.31	-	-	-
3	0.8	0.4	0.43	0.65	0.25	0.43
4	***0.8***	***0.3***	***0.36***	***0.68***	***0.24***	***0.40***
5	***0.7***	***0.2***	***0.34***	***0.69***	***0.29***	***0.42***
6	***0.7***	***0.2***	***0.35***	***0.61***	***0.22***	***0.45***
7	***0.8***	***0.2***	***0.30***	***0.66***	***0.21***	***0.40***
8	0.6	0.1	0.41	0.71	0.38	0.48
9	0.5	0.0	0.50	0.52	0.07	0.49
10	0.4	0.0	0.60	0.37	0.00	0.63
11	0.2	0.0	0.79	0.23	0.00	0.77
12	0.2	0.0	0.80	0.19	0.00	0.81
13	0.1	0.0	0.88	0.15	0.00	0.85
14	0.1	0.0	0.90	0.11	0.00	0.89
15	0.0	0.0	0.96	0.08	0.00	0.92
16	0.0	0.0	0.98	0.02	0.00	0.98
17	0.0	0.0	0.99	0.00	0.00	1.00
18	0.0	0.0	1.00	0.03	0.00	0.97
19	0.0	0.0	1.00	0.01	0.00	0.99
20	0.0	0.0	1.00	0.00	0.00	1.00
21	0.0	0.0	1.00	0.00	0.00	1.00
22	0.0	0.0	1.00	0.00	0.00	1.00
23	0.0	0.0	1.00	0.00	0.00	1.00

$\sqrt{(1-TPR)^2+FPR^2}$ in the table is the distance of the point nearest to the upper left corner in a ROC curve.
The rows shown in bold and italic present good differentiation performance.

Table 2. Classification performance of candidate patterns of length four

# of candidate patterns	Percentage of candidate patterns (%)	TPR	FPR	$\sqrt{(1-TPR)^2+FPR^2}$
5	0.011	0.35	0.09	0.66
18	0.04	0.5	0.21	0.54
178	0.4	0.66	0.21	0.40
333	**0.7**	**0.79**	**0.31**	**0.37**
3688	8.2	0.63	0.17	0.41
25509	56.6	0.63	0.14	0.39
45058	100	1	0.79	0.79

The percentage of candidate miRNA patterns is calculated as n/45058 where 45058 is the number of total segment miRNA patterns of length four.

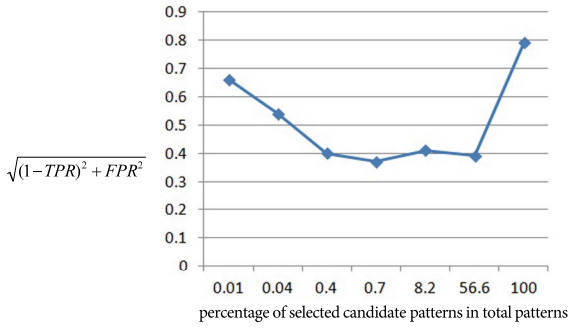

Fig. 2. Classification performance of candidate miRNA patterns of length four.

Table 3. Classification performance of candidate miRNA patterns of length five

# of candidate patterns	Percentage of candidate patterns (%)	TPR	FPR	$\sqrt{(1-TPR)^2+FPR^2}$
19	0.01	0.22	0.07	0.78
61	0.02	0.39	0.10	0.62
261	0.10	0.69	0.20	0.37
3166	1.24	0.84	0.30	0.34
17328	6.80	0.79	0.26	0.33
33849	13.29	0.74	0.19	0.32
59477	**23.36**	**0.80**	**0.22**	**0.30**
130704	**51.32**	**0.78**	**0.20**	**0.30**

The number of total segment miRNA patterns of length five is 254662.

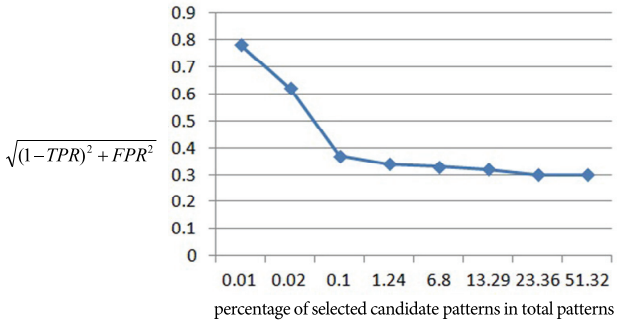

Fig. 3. Classification performance of candidate miRNA patterns of length five.

Table 4. Classification performance of candidate miRNA patterns of length six

# of candidate patterns	Percentage of candidate patterns (%)	TPR	FPR	$\sqrt{(1-TPR)^2+FPR^2}$
7	0.00001	0.22	0.07	0.78
386	0.06	0.43	0.06	0.57
5790	0.90	0.69	0.15	0.34
21004	3.27	0.76	0.25	0.35
41870	6.52	0.88	0.30	0.32
100079	15.59	0.84	0.22	0.27
361387	**56.30**	**0.86**	**0.20**	**0.24**

The number of total segment miRNA patterns of length six is 641908.

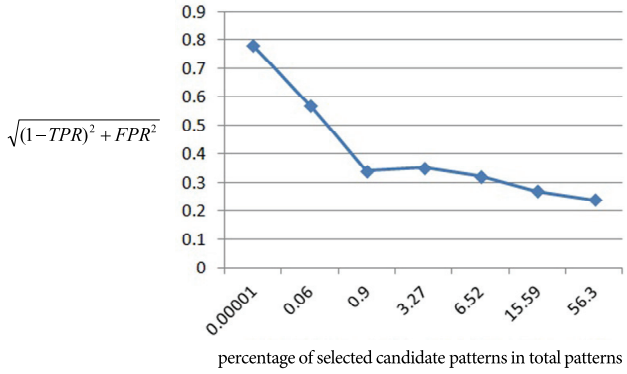

Fig. 4. Classification performance of candidate miRNA patterns of length six.

Table 5. Classification performance of candidate patterns of length seven

# of candidate patterns	Percentage of candidate patterns (%)	TPR	FPR	$\sqrt{(1-TPR)^2+FPR^2}$
4	0.00	0.01	0.00	0.99
754	0.09	0.31	0.05	0.69
3768	0.44	0.56	0.10	0.45
8854	1.05	0.72	0.23	0.36
18769	2.22	0.86	0.36	0.39
67484	7.97	0.79	0.21	0.30
457573	***54.03***	***0.80***	***0.18***	***0.27***

The number of total segment miRNA patterns of length seven is 846854.

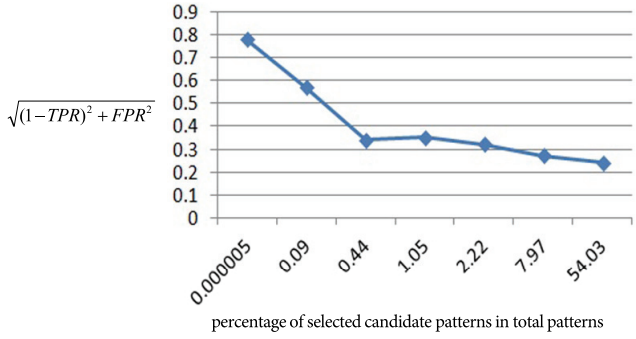

Fig. 5. Classification performance of candidate miRNA patterns of length seven.

3.3 Length of Potential miRNA Patterns

Figs. 2–5 also tell us that there is the difference of classification performance between the groups of different lengths. Fig. 6 shows the ROC curve of top approximately 50% segment miRNA patterns in the groups whose length range is from two to twelve and full length, and we find that the groups whose length range is between two and eight present better classification performance than others.

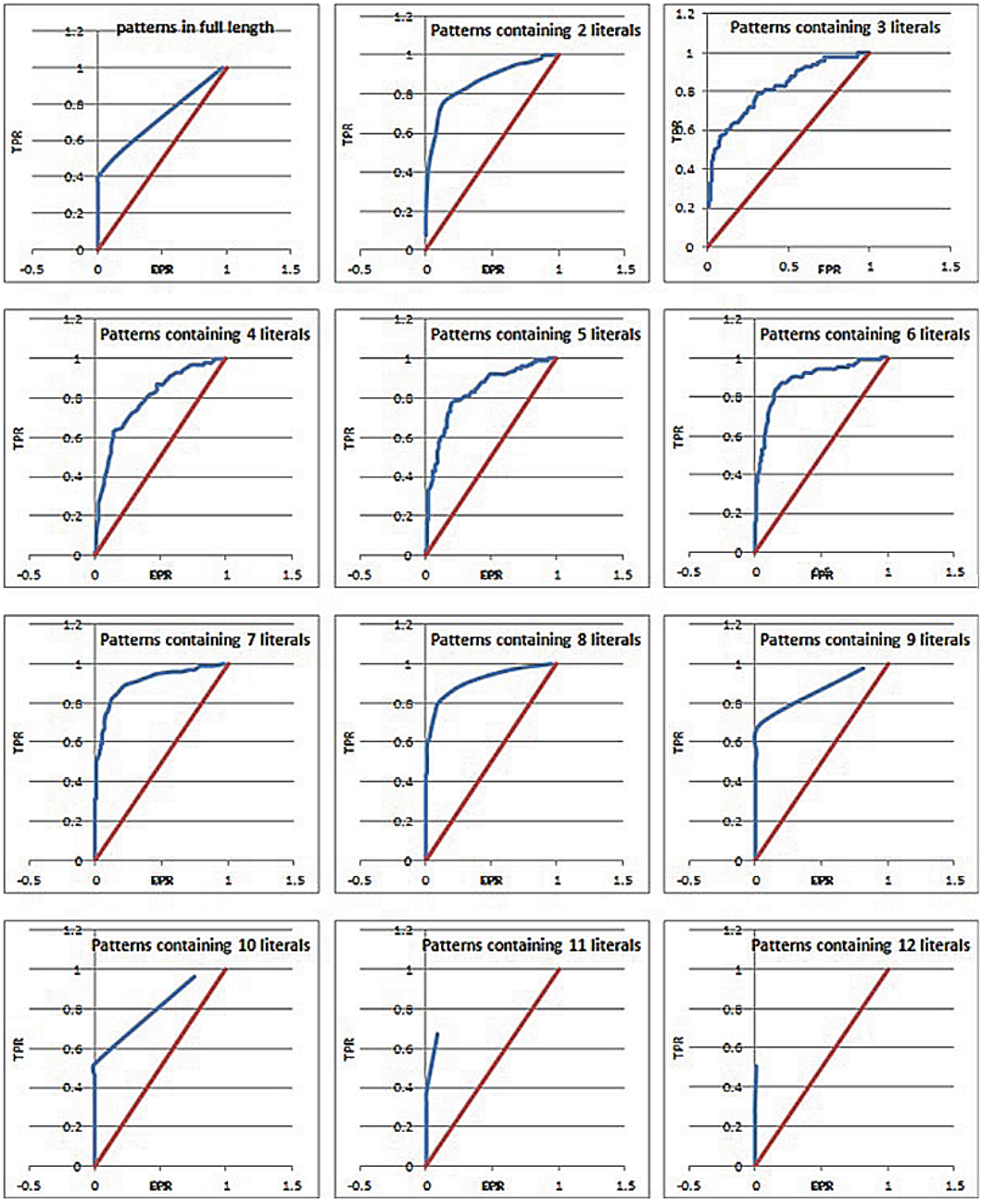

Fig. 6. ROC curves of top approximately 50% segment miRNA patterns in different groups. TPR and FPR are computed based on bi-nucleotides and the brown diagonal line is the random guess line. The figure does not show the ROC curves of the segment miRNA patterns whose length is larger than 12 because the ROC curves are either a straight line on y-axis or a point below the diagonal line.

We list the best TPR, FPR, and $\sqrt{(1-TPR)^2+FPR^2}$ of top approximately 50% segment miRNA patterns in Table 6 and find that the patterns in the groups whose length range is from five to nine are best in classification performance. Table 1 shows that the top 1% segment miRNA patterns in the

groups whose length range is from four to seven are best in classification performance. Although the classification performance of top approximately 50% segment miRNA patterns is a little better than that of top approximately 1% segment miRNA patterns, the number of top approximately 1% segment miRNA patterns (300–9000) is far less than the number of top approximately 50% segment miRNA patterns (4500–450000), therefore we select the top approximately 1% segment miRNA patterns in the groups of length range is from four to seven as potential miRNA patterns.

Table 6. Classification performance of top 50% of segment miRNA patterns

Length of segment patterns	TPR	FPR	The shortest distance to the upper left corner in ROC curve
Full length	0.4	0.0	0.60
2	0.8	0.1	0.28
3	0.8	0.3	0.37
4	0.6	0.1	0.40
5	**0.8**	**0.2**	**0.30**
6	**0.8**	**0.2**	**0.28**
7	**0.8**	**0.1**	**0.23**
8	**0.8**	**0.1**	**0.23**
9	**0.7**	**0.1**	**0.30**
10	0.5	0.0	0.50
11	0.7	0.1	0.34
12	0.5	0.0	0.49
13	0.5	0.0	0.54
14	0.4	0.0	0.57
15	0.4	0.0	0.65
16	0.3	0.0	0.71
17	0.2	0.0	0.81
18	0.2	0.0	0.85
19	0.1	0.0	0.91
20	0.1	0.0	0.91
21	0.1	0.0	0.92
22	0.0	0.0	0.98
23	0.0	0.0	1.00

The shortest distance to the upper left corner in ROC curve is computed by $\sqrt{(1-TPR)^2+FPR^2}$.

Moreover, Fig. 7 shows the classification performance of top 1% segment miRNA patterns in all the 24 groups based on bi-nucleotides and tri-nucleotides and the top 50% segment miRNA patterns in all the 24 groups based on bi-nucleotides. We find that the trend of the three curves is similar and the groups whose length range is from four to seven are good and stable. Therefore, taking the computation burden into account, we suggest computing estimated probabilities of segment miRNA patterns based on bi-nucleotides and selecting top 1% segment miRNA patterns of length four as potential miRNA patterns.

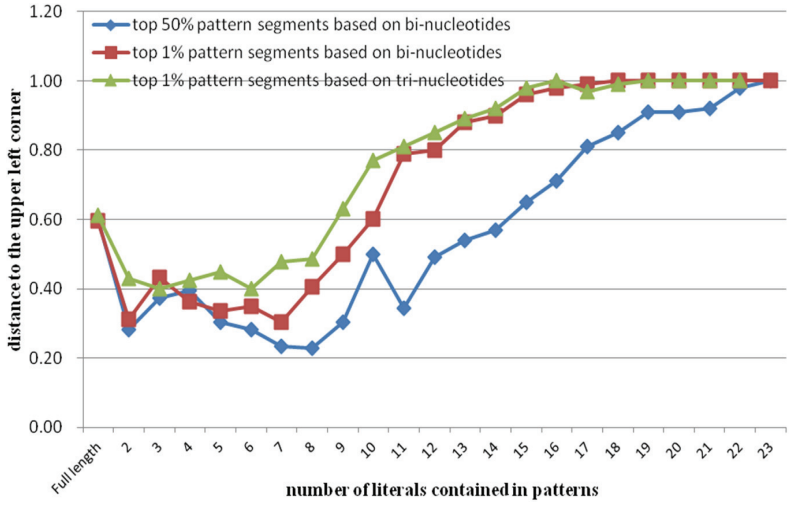

Fig. 7. The smallest value of $\sqrt{(1-TPR)^2+FPR^2}$ of the top 1% and 50% segment miRNA patterns.

3.4 Significance Analysis

3.4.1 Significance of potential miRNA patterns

The potential miRNA patterns are the set of subsequences which capture most conservative knowledge of mature miRNA sequences. There are 450 potential miRNA patterns, and we produce approximately 450 fake miRNA patterns randomly chosen from the 45058 segment miRNA patterns of length four.

Based on the test data, we use fake miRNA patterns to compute the $\sqrt{(1-TPR)^2+FPR^2}$. We repeat the steps above 10,000 times and compute the difference significance of $\sqrt{(1-TPR)^2+FPR^2}$ between potential miRNA patterns and fake ones. The results show that the values of $\sqrt{(1-TPR)^2+FPR^2}$ of fake miRNA patterns are significantly larger than that of potential miRNA patterns (p-value=0).

We also investigate the classification performance of the potential miRNA segments on other specifies. We download the mature miRNA sequences form miRBase (release 21) of *Drosophila melanogaster*, *Homo sapiens*, and *Mus musculus*, and produce the test data of the three species separately. The test data includes 100 randomly selected authentic mature miRNA sequences and 100 artificial miRNA sequences generated in the following two steps:

First, generate the 100 random numbers which subjects to a normal distribution (mean is 22 and variance is 3) as the lengths of the 100 artificial miRNA sequences;

Second, choose randomly one letter from the four literals (A, U, G, C) for each position in each artificial miRNA sequence.

We separately count the number of potential and fake miRNA patterns contained in the each sequence of the test data (Figs. 8–13). The number of potential miRNA patterns contained in authentic and artificial miRNAs is significantly different—the p-values of the three species are separately 5.248e-5 (*Mus musculus*), 0.007065 (*Homo sapiens*) and 6.801e-5 (*Drosophila melanogaster*). Simultaneously, the number of fake miRNA patterns contained in authentic and artificial miRNAs has no significant difference—the p-values are separately 0.8451 (*Mus musculus*), 0.08646 (*Homo sapiens*) and 0.08724 (*Drosophila melanogaster*).

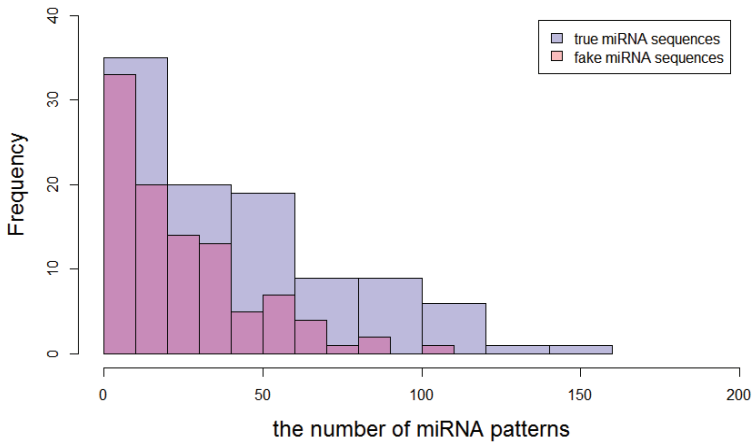

Fig. 8. Potential miRNA patterns contained in authentic and artificial miRNA sequences (*Mus musculus*).

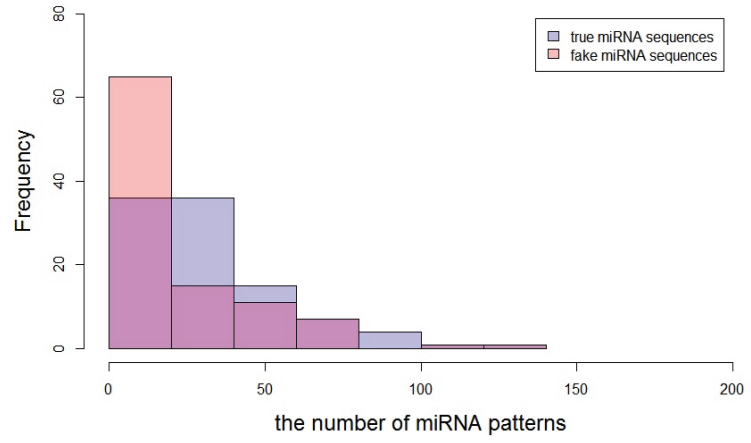

Fig. 9. Potential miRNA patterns contained in authentic and artificial miRNA sequences (*Homo sapiens*).

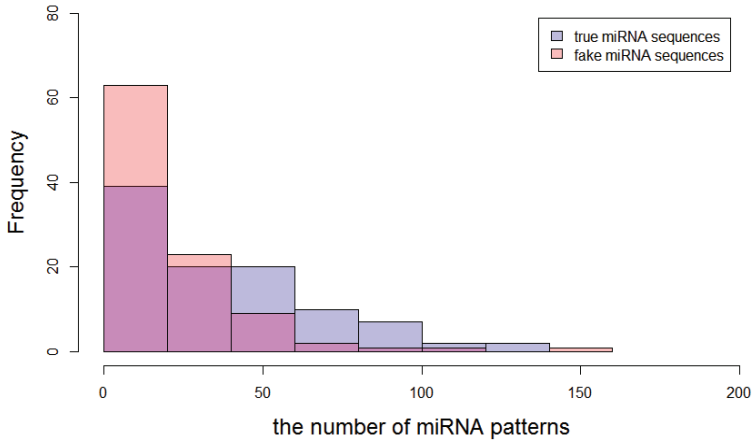

Fig. 10. Potential miRNA patterns contained in authentic and artificial miRNA sequences (*Drosophila melanogaster*).

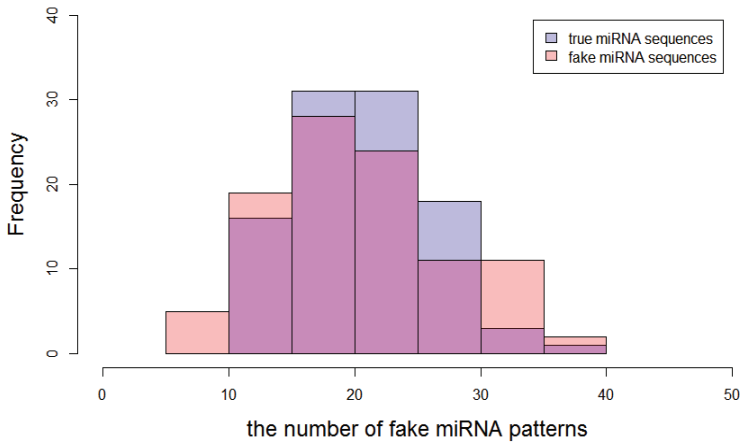

Fig. 11. Fake miRNA patterns contained in authentic and artificial miRNA sequences (*Mus musculus*).

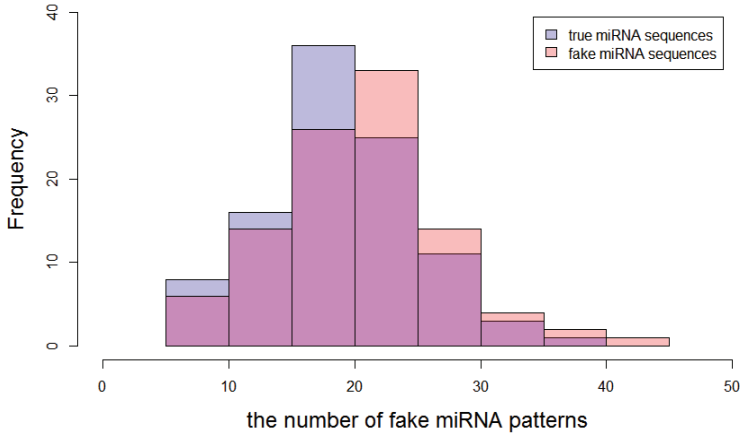

Fig. 12. Fake miRNA patterns contained in authentic and artificial miRNA sequences (*Homo sapiens*).

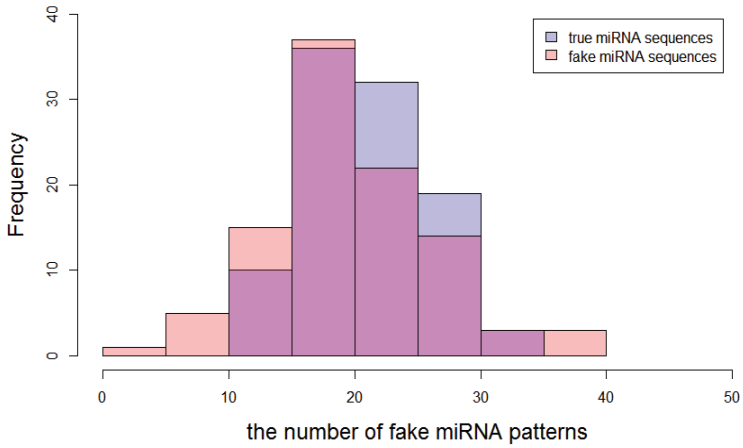

Fig. 13. Fake miRNA patterns contained in authentic and artificial miRNA sequences (*Drosophila melanogaster*).

Table 7. Comparison of Classification performance in three specifies

species	TP	FP	TN	FN	Sensitivity (%)	Specificity (%)	Precision (%)	Accuracy (%)
Homo sapiens	66	35	65	34	66	65	65.35	65.5
Mus musculus	69	53	47	31	69	47	56.56	58
Drosophila melanogaster	73	48	52	27	73	52	60.33	65.5

3.4.2 Classification performance of potential miRNA patterns on other test data

According to the numbers of potential and fake miRNA patterns included in the authentic and artificial miRNA sequences, we list the measures of TP, FP, TN, FN, sensitivity, specificity, precision and accuracy of the three specifies in Table 7 [21,22].

We also investigate the false positive artificial miRNA sequences and find there are separately 13, 11 and 8 artificial miRNA sequences correspond to authentic mature miRNAs (Table 8).

4. Conclusions

MiRNA pattern is a nucleotide sequence motif conjectured to have biological significance [15]. miRNA pattern discovery is a challenging task due to the small length and few structure information of mature miRNA sequences. The proposed miRNA pattern discovery approach is a fast way to produce the sequence patterns containing as much information as possible, reduce the output size and remove redundancy from the short mature miRNAs [23]. In addition, the discovered miRNA patterns are proved to implicate conservative knowledge of mature miRNA sequences and can be used to identify the authentic mature miRNAs.

Table 8. Artificial miRNA sequences corresponding to authentic miRNAs

Species	Artificial miRNA sequences	MiRNA ID
Drosophila melanogaster	auggaccauacacuuc	mdo-miR-7386m-3p
	agaguggaucgcuaucaagugcu	nve-miR-2046-5p
	agcucuguaugaggccgcuga	chi-miR-423-3p, hsa-miR-423-3p, mmu-miR-423-3p, bta-miR-423-3p, rno-miR-423-3p, mml-miR-423-3p, ptr-miR-423, eca-miR-423-3p, ssc-miR-423-3p, hsa-miR-3184-5p, ppy-miR-423-3p, tch-miR-423-3p, cgr-miR-423-3p
	auaugaaggguaugaaugcug	bmo-miR-3362
	aucgaugguuaaucaa	aly-miR4225
	ggucgagucggguucacca	osa-miR5077
	uacgggugugguuaaccucuga	mtr-miR5213-5p
	ucuggcaagguauaaaacuguca	gra-miR8743b

Table 8. (Continued)

Species	Artificial miRNA sequences	MiRNA ID
	agcgggcccggcugug	rno-miR-666-5p
	agguuguauggcgcggaaua	ppc-miR-8347-5p
	aucccuagucguccauguugagg	atr-miR8613
	cgggagauaugagcccuc	cel-miR-8200-3p
	cgggaugcugaaauguggguuuua	dme-miR-9372-5p
Mus musculus	cguaaguuggauacacucgc	str-miR-7880a-3p
	cucucuccgugcuauaaggagu	bdi-miR5174d-3p
	ucgacucguguaucgggaguaug	cel-miR-59-3p, crm-miR-59
	uggauacacucgcugaua	dre-miR-7148-5p
	uuccauguugagcgggcccg	bmo-miR-3378-5p
	uugccuuuuuccucccaug	hsa-miR-4423-5p
	aacuucugcgcagcuaaa	tca-miR-3818-3p
	agcauuggugcgcucugug	gga-miR-1737
	ccgccggucagaauacuug	hsa-miR-4465
	ccggucagaauacuugcacca	gra-miR8759
	cgcucugugcugcaaguugag	dre-miR-2196
	cuuauauugcgacucgucu	cte-miR-2686a-5p
	gcgugggauugugcacggc	lja-miR7527
Homo sapiens	guugcaagcgguug	gsa-miR-2b-5p
	Uaagcgugauuauggcguu	hsa-miR-122-3p, mmu-miR-122-3p, rno-miR-122-3p, hsa-miR-3591-5p, aca-miR-122-3p, ola-miR-122, gga-miR-122-3p, tgu-miR-122-3p, ssa-miR-122-2-3p
	ugaaccgcgucaau	zma-miR171j-5p
	ugccgccggucagaauacuugcacc	bmo-miR-3382-3p
	ugugcugcaaguugagcguguu	spu-miR-4849
	uucacuccgacaaagaacau	gga-miR-6693-3p

Search algorithm: BLASTN, cutoff: 10

Acknowledgement

This paper is supported by the Natural Science Foundation of China (Grant No. 61571341, 61201312, 91530113), the China Scholarship Council (No. 210508615092), Research Fund for the Doctoral Program of Higher Education of China (No. 20130203110017), Natural Science Foundation of Shaanxi Province (No. 2017JM6036), and the Research Projects of Weinan Normal University (No. 16YKP002).

References

[1] R. M. Marin, M. Sulc, and J. Vanicek, "Searching the coding region for microRNA targets," *RNA*, vol. 19, no. 4, pp. 467-474, 2013.

[2] S. T. Kalinowski, T. M. Andrews, M. J. Leonard, and M. Snodgrass, "Are Africans, Europeans, and Asians different 'races'? A guided-inquiry lab for introducing undergraduate students to genetic diversity and preparing them to study natural selection," *CBE Life Sciences Education*, vol. 11, no. 2, pp. 142-151, 2012.

[3] B. Liu, J. Li, and M. J. Cairns, "Identifying miRNAs, targets and functions," *Briefings in Bioinformatics*, vol. 15, no. 1, pp. 1-19, 2014.

[4] I. Bentwich, "Prediction and validation of microRNAs and their targets," *FEBS Letters*, vol. 579, no. 26, pp. 5904-5910, 2005.

[5] M. R. Friedlander, W. Chen, C. Adamidi, J. Maaskola, R. Einspanier, S. Knespel, and N. Rajewsky, "Discovering microRNAs from deep sequencing data using miRDeep," *Nature Biotechnology*, vol. 26, no. 4, pp. 407-415, 2008.

[6] V. Williamson, A. Kim, B. Xie, G. O. McMichael, Y. Gao, and V. Vladimirov, "Detecting miRNAs in deep-sequencing data: a software performance comparison and evaluation," *Briefings in Bioinformatics*, vol. 14, no. 1, pp. 36-45, 2013.

[7] M. R. Friedlander, S. D. Mackowiak, N. Li, W. Chen, and N. Rajewsky, "miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades," *Nucleic Acids Research*, vol. 40, no. 1, pp. 37-52, 2012.

[8] W. Shen, M. Chen, G. Wei, and Y. Li, "MicroRNA prediction using a fixed-order Markov model based on the secondary structure pattern," *PLoS One*, vol. 7, no. 10, article no. e48236, 2012.

[9] X. Ji, J. Bailey, and G. Dong, "Mining minimal distinguishing subsequence patterns with gap constraints," *Knowledge and Information Systems*, vol. 11, no. 3, pp. 259-286, 2007.

[10] G. Dong and J. Bailey, *Contrast Data Mining: Concepts, Algorithms, and Applications*. Boca Raton, FL: CRC Press, 2013.

[11] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507-2517, 2007.

[12] R. C. de Amorim, "Computational methods of feature selection," *Information Processing & Management*, vol. 45, no. 4, pp. 490-493, 2009.

[13] G. Nuel, L. Regad, J. Martin, and A. C. Camproux, "Exact distribution of a pattern in a set of random sequences generated by a Markov source: applications to biological data," *Algorithms for Molecular Biology*, vol. 5, article no. 15, 2010.

[14] R. Jackups and J. Liang, "Combinatorial analysis for sequence and spatial motif discovery in short sequence fragments," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 7, no. 3, pp. 524-536, 2010.

[15] H. Zheng, H. Wang, and F. Azuaje, "Improving pattern discovery and visualization of SAGE data through poisson-based self-adaptive neural networks," *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 4, pp. 459-469, 2008.

[16] O. Westesson, L. Barquist, and I. Holmes, "HandAlign: bayesian multiple sequence alignment, phylogeny and ancestral reconstruction," *Bioinformatics*, vol. 28, no. 8, pp. 1170-1171, 2012.

[17] A. Kawrykow, G. Roumanis, A. Kam, D. Kwak, C. Leung, C. Wu, et al., "Phylo: a citizen science approach for improving multiple sequence alignment," *PLoS One*, vol. 7, no. 3, article no. e31362, 2012.

[18] E. Pruesse, J. Peplies, and F. O. Glockner, "SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes," *Bioinformatics*, vol. 28, pp. 1823-1829, 2012.

[19] K. C. Miranda, T. Huynh, Y. Tay, Y. S. Ang, W. L. Tam, A. M. Thomson, B. Lim, and I. Rigoutsos, "A pattern-based method for the identification of microRNA binding sites and their corresponding heteroduplexes," *Cell*, vol. 126, no. 6, pp. 1203-1217, 2006.

[20] M. Hafner, P. Landgraf, J. Ludwig, A. Rice, T. Ojo, C. Lin, D. Holoch, C. Lim, and T. Tuschl, "Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing," *Methods*, vol. 44, no. 1, pp. 3-12, 2008.

[21] N. Lavrac, P. Flach, and B. Zupan, "Rule evaluation measures: a unifying view," in *Proceedings of 9th International Workshop on Inductive Logic Programming (ILP-99)*, Bled, Slovenia, 1999, pp. 174-185.

[22] L. Geng and H. J. Hamilton, "Interestingness measures for data mining: a survey," *ACM Computing Surveys (CSUR)*, vol. 38, no. 3, article no. 9, 2006.

[23] A. K. C. Wong, D. Zhuang, G. C. L. Li, and E. S. A. Lee, "Discovery of non-induced patterns from sequences," in *Proceedings of 5th IAPR International Conference on Pattern Recognition in Bioinformatics (PRIB 2010)*, Nijmegen, The Netherlands, 2010, pp. 149-160.

Xiaohan Sun https://orcid.org/0000-0002-7729-1438

She received the B.S. degree in information management from Shaanxi Economics and Trade College, Shaanxi, China, in 2001, the M.S. degree in computer application technology from Xidian University, Shaanxi, China, in 2009. Since 2011 she has been working towards a PhD degree in computer application technology from Xidian University. Her current research interests are microRNA and complex diseases.

Junying Zhang

She received her Ph.D. degree in Signal and Information Processing from Xidian University, Xi'an, China, in 1998. From 2001 to 2002, she was a visiting scholar at the Department of Electrical Engineering and Computer Science, The Catholic University of America, Washington, DC, USA, and in 2007, she was a visiting professor at the Department of Electrical Engineering and Computer Science, Virginia Polytechnic University, USA. She is currently a professor in the School of Computer Science and Technology, Xidian University, Xi'an, China. Her research interests focus on intelligent information processing, including machine learning and its application to cancer related bioinformatics, causative learning and pattern discovery.