GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Benchmark comparison of *ab initio* microRNA identification methods and software

L.L. Hu[1], Y. Huang[2], Q.C. Wang[1], Q. Zou[1] and Y. Jiang[1]

[1]School of Information Science and Technology, Xiamen University, Xiamen, Fujian, China
[2]Animal Science and Technology College, Henan University of Science and Technology, Luoyang, Henan, China

Corresponding author: Y. Jiang
E-mail: jiangyi@xmu.edu.cn

Genet. Mol. Res. 11 (4): 4525-4538 (2012)
Received January 10, 2012
Accepted June 7, 2012
Published October 17, 2012
DOI http://dx.doi.org/10.4238/2012.October.17.4

ABSTRACT. MicroRNAs (miRNAs) are short, non-coding RNA molecules that play an important role in the world of genes, especially in regulating the gene expression of target messenger RNAs through cleavage or translational repression of messenger RNA. *Ab initio* methods have become popular in computational miRNA detection. Most software tools are designed to distinguish miRNA precursors from pseudo-hairpins, but a few can mine miRNA from genome or expressed sequence tag sequences. We prepared novel testing datasets to measure and compare the performance of various software tools. Furthermore, we summarized the miRNA mining methods that study next-generation sequencing data for bioinformatics researchers who are analyzing these data. Because secondary structure is an important feature in the identification of miRNA, we analyzed the influence of various secondary structure prediction software tools on miRNA identification. MiPred was the most effective for classifying real-/pseudo-pre-miRNA sequences, and miRAbela performed relatively

better for mining miRNA precursors from genome or expressed sequence tag sequences. RNA-fold performed better than m-fold for extracting secondary structure features of miRNA precursors.

Key words: MicroRNAs; *Ab initio* methods; Pseudo-hairpins; Next-generation sequencing; Secondary structure

INTRODUCTION

MicroRNAs (miRNAs) are non-coding RNAs that play important roles in gene regulation by targeting mRNAs for cleavage or translational repression (Carrington and Ambros, 2003; Huang et al., 2007). In animals, miRNAs are initially transcribed as longer primary transcripts called pri-miRNAs and then processed by RNase III Drosha (Borchert et al., 2006) into 60- to 70-nt miRNA precursors (pre-miRNA) (Lee et al., 2003; Zeng et al., 2005). Pre-miRNA is transported from the nucleus to the cytoplasm by exportin-5 and cleaved into 21- to 25-nt mature miRNA. In plants, pri-miRNAs are cleaved into mature miRNA by Dicer-like 1 protein and transported from the nucleus to the cytoplasm by HASTY (Park et al., 2005; Li et al., 2007). The precursor of the miRNAs has the characteristics of stem-loop hairpin structures (Huang et al., 2010). miRNAs play an increasingly important role in the regulation and control of biological processes in organisms such as the larvae growth sequence (Moss et al., 1997; Reinhart et al., 2000) and cell proliferation (Brennecke et al., 2003; Huang et al., 2011a).

One of the most extensively developed methods for miRNA detection is the comparative approach, which depends primarily on sequence similarity to known pre-miRNAs. Lower sensitivity in detecting novel miRNAs is the main drawback; another drawback is the generation of false positives. To overcome these obstacles, researchers have increasingly turned to *ab initio* prediction methods. These methods predict miRNAs in a single genome without using conservation of structure or comparative sequence analysis. The number of non-conserved miRNAs is estimated to be relatively large (Bentwich et al., 2005), which enables the identification of completely novel miRNAs for which no close homologs are known. Unlike comparative genomics approaches, *ab initio* approaches discover species-specific miRNAs without known sequence homology, and many of these algorithms have recently been developed to detect novel pre-miRNAs for which no close homology is known. Bioinformatics play an important role in the research of miRNA; Figure 1 shows the prediction process for miRNA.

Many software tools have been released for the prediction of miRNAs. Summarization and comparison of the best-known software are carried out to help researchers choose effective software for miRNA prediction in experimental and public DNA data. The comparison presented herein provides a deeper understanding of various identification software tools to allow appropriate choices based on needed strengths, such as accuracy rate or specificity. The direction of experiments is taken into account simultaneously. We performed several experiments to compare various software and methods. The advantages and disadvantages of the software are presented based on aspects such as dataset accuracy rate. Our results are highly reliable because we obtained and compared data from a large number of experiments instead of relying on speculation and research.

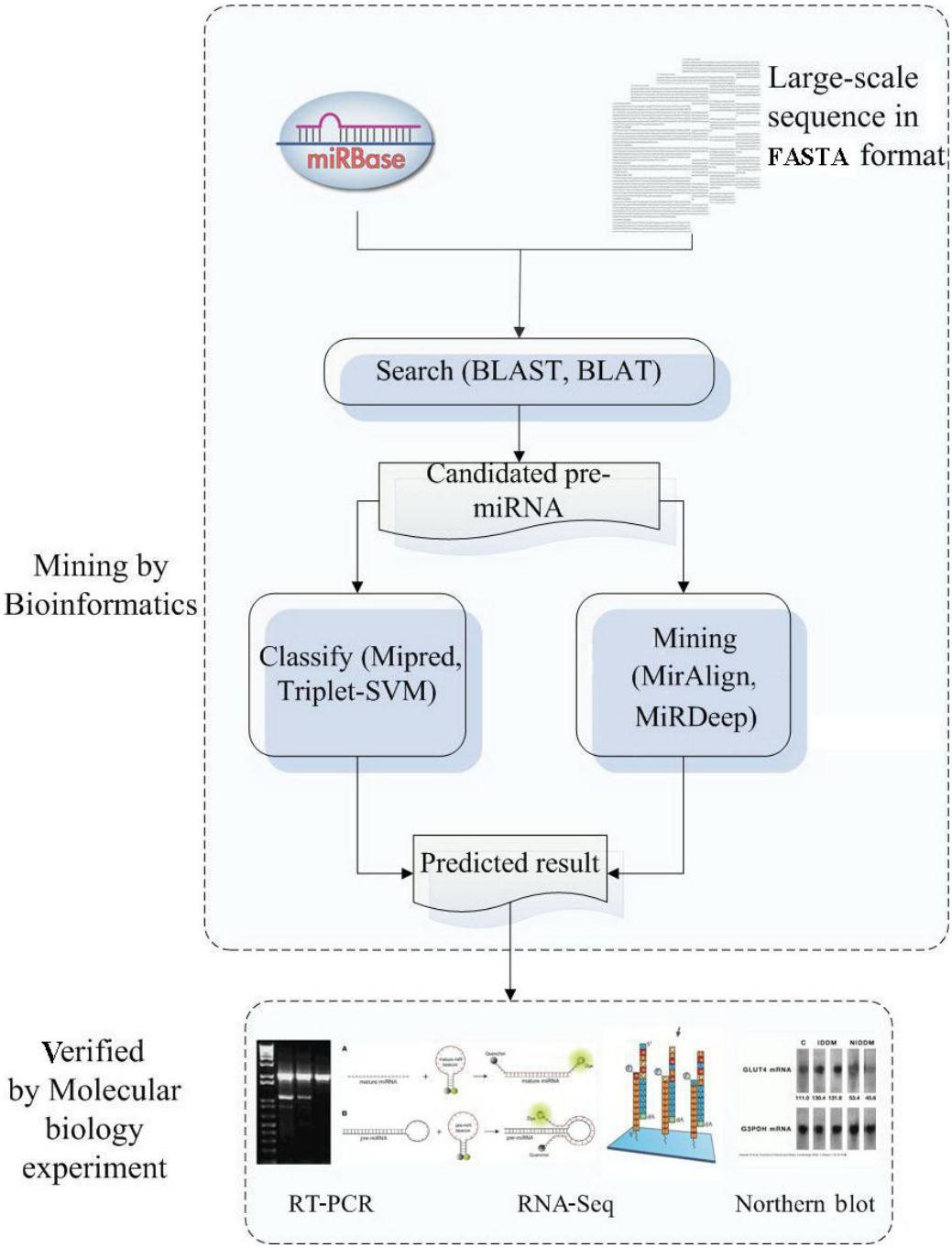

Figure 1. Predict process of miRNA.

MATERIAL AND METHODS

The highly species- and time-specific expression patterns of *ab initio* methods make them popular in detecting a mass of non-conserved miRNAs. Unlike comparative genomics approaches, *ab initio* approaches can discover species-specific miRNAs without known homology sequences (Huang et al., 2011b), which has spurred the development of multiple *ab initio* methods for miRNA prediction.

Herein, we introduce the main *ab initio* methods with the goal of analyzing the working accuracy of the algorithms instead of understanding their working mechanisms. We intend for users to be able to select a method for their experimental data according to the information we offer. Various methods are categorized and summarized according to their purposes and processes in miRNA identification.

Classifying miRNA precursors from pseudo-hairpins

miRNA precursor sequences can fold into a typical stem-loop structure considered to be the most important indicator of the maturation process; however, a large number of similar hairpins that are not pre-miRNAs can be found in many genome regions and are called pseudo-hairpins (Bentwich et al., 2005). Accurately distinguishing, classifying, and identifying miRNA precursors from pseudo-hairpins is difficult. Most of the current methods for the computational prediction of miRNAs make use of comparative genomic approaches to identify pre-miRNAs from candidate hairpins. Because distinguishing pre-miRNAs among hairpin secondary structures is extraordinarily difficult; the development of *ab initio* methods to distinguish pre-miRNAs from pseudo-pre-miRNA-like hairpin segments is crucial. Several of the main pre-miRNA software tools are summarized below.

MiPred

Pre-miRNAs can be distinguished from other hairpin sequences with MiPred. Random forest improves the accuracy, and 32 triplet structure-sequence features as well as minimal free energy (Hofacker, 2003) and P value of free energy are used to describe the sample. MiPred reached a sensitivity of 89.35% and a specificity of 93.21% on a test set (Jiang et al., 2007). Given a sequence, MiPred determines whether it is a pre-miRNA-like hairpin. If it is, the random forest classifier predicts and shows whether it is a pre-miRNA or a pseudo-hairpin. Additional information such as minimum free energy of the secondary structure and the P value of the randomization test are given.

microPred

microPred presents an effective classifier with appropriate machine learning techniques. The approach in microPred includes the introduction of more representative datasets, extraction of new biologically relevant features, feature selection, handling of class imbalance problems in datasets, and extensive classifier performance evaluation via systematic cross-validation methods (Batuwita et al., 2009).

Virgo

Virgo functions based on sequence and structure features. A support vector machine (SVM) trained on sequence-structure feature elements is used for efficient discrimination between miRNA precursor hairpins and pseudo-miRNA hairpins. The method is more efficient than that of other *ab initio* methods for predicting viral and mammalian miRNAs (Kumar et al., 2009). Virgo selects a model with maximum specificity (rather than sensitivity) and uses the Radial Bias Function kernel of the SVM to create a function corresponding to the hypersurface that optimally separates true and pseudo-miRNA hairpins. The method is fast enough for viral genome-wide predictions and can be helpful in the discovery of both novel non-conserved and virus-expressed miRNAs.

Triplet-SVM

Triplet-SVM classifies real- and pseudo-microRNA precursors using local structure-sequence features and SVM. An SVM classifier trained based on the triplet element features of a set of real-miRNA precursors and a set of pseudo-miRNA hairpins is used to analyze and predict the triplet elements of the query. Triplet-SVM runs directly on Linux with a Perl compiler. The SVM classifier that built on human data can correctly identify up to 90% of the pre-miRNAs (Xue et al., 2005) from other species without using any comparative genomics information, and 32 triplet structure sequence characteristics were used to describe the sample to construct Triplet-SVM. It reached a sensitivity of 93.3% and a specificity of 88.1% on a test set (Xue et al., 2005).

Mining miRNA precursors from genome or EST sequences

Ab initio methods that mine miRNA precursors from genome or ESTs do not depend entirely on known genetic sequence information but use it as training set to extract features and then make use of a machine-learning classification algorithm to identify candidate clips. Current research shows that the secondary structure of noncoding miRNA remains unaltered. For instance, the hairpin secondary structure of miRNA precursors and the distribution of nucleotides have certain rules (Zou et al., 2011) that can be discerned by studying known miRNA genes and their precursors; these rules can be used to classify unknown RNA and judge whether it is miRNA. Several major *ab initio* mining software tools for miRNAs are summarized below.

miRAlign

Most approaches detect novel miRNAs according to the conservation of whole pre-miRNA sequences or the nearly perfect match of mature parts of the sequence, but miRAlign is distinct from other search tools in 2 main ways (Wang et al., 2005): it can find distant homologs according to the relatively loose conservation of the mature sequence, and it considers additional properties of miRNA structure conservation. MiRAlign performed with higher sensitivity than and comparable specificity to those of other homolog searching methods, and 59 novel miRNA genes were detected (Wang et al., 2005).

miRAbela

Research has shown that miRNAs are occasionally found in clusters. Consequently, miRAbela focuses on genomic regions around already known miRNAs. It uses a growing set of mammalian sequences that have been cloned in the laboratory of Thomas Tuschl to evaluate the performance of the method. Cross-species comparisons are used in the methods to make conservative estimates of the number of novel miRNAs, and the species-specific identity and genome organization of miRNA loci can be studied because each genome is analyzed separately. miRAbela reached a sensitivity of 71% and a specificity of 97% on a training set (Sewer et al., 2005). Thirty-two possible pre-miRNAs were predicted in a test of virus miRNA in which 13 pre-miRNAs had been verified experimentally (Sewer et al., 2005).

miRPara

miRPara is a software tool that predicts the most probable mature miRNA-coding regions from genome-scale sequences in a species-specific manner. miRPara uses an SVM to train 3 models based on an initial set of 77 parameters (Wu et al., 2011), and these parameters are related to the physical properties of pro-miRNAs and their miRNAs. Given a genome sequence, miRPara locates miRNA-coding regions, and high-throughput screening experiments can use miRPara as a pre-screening step.

MIReNA

MIReNA can find miRNAs at the genome scale and from deep sequencing data without the machine learning that can confirm only pre-miRNAs that look like known pre-miRNAs. MIReNA explores a multidimensional space defined by only five parameters to identify pre-miRNA/miRNA pairs. In addition, the MIReNA algorithm can handle four kinds of data (known miRNAs, deep sequencing data, potential miRNAs occurring in long sequences, and putative pre-miRNAs containing potential miRNAs) (Mathelier and Carbone, 2010), and the first two kinds of data may be checked against full-genome sequences.

MiRscan

Together with molecular identification and validation methods, MiRscan can identify most of the miRNA genes in the nematode *Caenorhabditis elegans* (Lim et al., 2003). For the identification of miRNAs in nematodes, the *C. elegans* genome was first scanned for hairpin structures with sequences that were conserved in *C. briggsae*. Fifty published miRNA genes served as a training set for MiRscan. MiRscan was then used to assign scores to each of the hairpins and evaluate them for similarity to certain features of the training set. It indentified 30 novel miRNAs in *C. elegans* and 38 novel human miRNAs (Lim et al., 2003).

ProMiR II

ProMiR (http://bi.snu.ac.kr/ProMiR/) uses a hidden Markov model (Nam et al., 2005) to describe the real- and pseudo-pre-miRNA. ProMiR II is a general version of ProMiR that

searches for miRNA in stem-loop sequences. Low- or high-stringency prediction of conserved and non-conserved miRNA genes is allowed because, unlike ProMiR, ProMiR II adjusts several filtering criteria such as free-energy data, G/C ratio, conservation score, and entropy of candidate sequences (Nam et al., 2006). ProMiR II generates a list of nearby potential miRNAs according to score and filtering criteria. Additional services, such as the prediction of miRNA genes in long unrelated sequences such as viral genomes, are also provided. An MySQL database that structures the data improves the efficiency of data analysis.

BayesMiRNAfind

BayesMiRNAfind may be applicable to a wide variety of eukaryotes. It differs from other tools in two ways: 1) it generates a model automatically from the training data and identifies rules based on the miRNA gene structure and sequence, allowing prediction of non-conserved miRNAs, and the training data consist of sequence and structure information of known miRNAs from a variety of species. 2) It integrates data from multiple species for miRNA gene prediction and performs a comparative analysis over multiple species to reduce the false-positive rate. Using this classifier combined with a structure feature such as the length of species and the conservation of genome in humans and dolphins, the program predicted 533 possible pre-miRNAs, of which 135 are already known in the normal chain of the mouse genome (Yousef et al., 2006).

Mining miRNA from next-generation sequencing (NGS) data

A new generation of sequencing technologies (such as deep sequencing) has provided unprecedented opportunities for high-throughput detection of miRNAs and can detect many small RNAs with a high degree of reliability. NGS has made possible high-sensitivity discovery of tissue-specific and developmental stage-specific miRNAs and miRNAs expressed at low levels (Ruby et al., 2006; Friedlander et al., 2008). NGS data from Illumina/Solexa, ABI/SOLiD, and 454/Roche produce several sequence fragments in the 200- to 300-bp range and detect known and novel miRNAs with unprecedented sensitivity (Friedlander et al., 2008). Several major mining methods of miRNAs from NGS data are summarized below.

miRDeep

miRDeep uses a probabilistic algorithm to score features of miRNA candidates with Bayesian statistics. Its accuracy and robustness originate in published *C. elegans* data and data generated from deep-sequenced human and dog RNAs. The miRDeep algorithm excises the genomic DNA bracketing these alignments and computes their secondary structure after the sequencing reads are aligned to the genome. Plausible miRNA precursor sequences are identified and then scored for their likelihood to be real-miRNA precursors in the core part of the algorithm. miRDeep reported ~230 previously unannotated miRNAs, of which four novel *C. elegans* miRNAs were validated with Northern blot analysis (Friedlander et al., 2008). A scored list of known, novel precursors, mature miRNAs in the deep-sequencing sample, and estimates for the number of false positives was shown as the output.

miRanalyzer

miRanalyzer implements a variety of methods for integrated analysis of deep-sequencing experiments of small RNA molecules. The small RNA data obtained with NGS platforms such as Illumina or SOLiD are processed by miRanalyzer. miRanalyzer implements a highly accurate machine learning algorithm based on the random forest classifier and is trained on experimental data for the prediction of new miRNAs. The output website (http://web.bioinformatics.cicbiogune.es/microRNA/) shows tables of detailed information. miRanalyzer provided the number of predicted miRNAs and predicted new miRNAs numbers in the candidate miRNAs, reaching area under the curve value of 97.9% and recall value of up to 75% on unseen data (Hackenberg et al., 2009) for the prediction of new miRNAs. Resources for the *ab initio* approaches for miRNA prediction are shown in Table 1.

Table 1. Resources of the *ab initio* approaches for miRNA predictions.

Software	Website	References
MiPred	http://www.bioinf.seu.edu.cn/miRNA/	Jiang et al., 2007
microPred	http://web.comlab.ox.ac.uk/people/ManoharaRukshan.Batuwita/microPred.htm	Batuwita et al., 2009
Virgo	http://miracle.igib.res.in/virgo/	Kumar et al., 2009
Triplet-SVM	http://bioinfo.au.tsinghua.edu.cn/mirnasvm/	Xue et al., 2005
MiRAlign	http://bioinfo.au.tsinghua.edu.cn/miralign/	Wang et al., 2005
miRAbela	http://www.mirz.unibas.ch/cgi/pred_miRNA_genes.cgi	Sewer et al., 2005
miRPara	http://159.226.126.177/mirpara/cgi-bin/form.cgi	Wu et al., 2011
MIReNA	http://www.ihes.fr/~carbone/data8/	Mathelier and Carbone, 2010
MiRscan	http://genes.mit.edu/mirscan/	Lim et al., 2003
ProMiR II	http://cbit.snu.ac.kr/~ProMiR2/	Nam et al., 2006
BayesMiRNAfind	http://wotan.wistar.upenn.edu/miRNA/	Yousef et al., 2006
miRDeep	http://www.mdc-berlin.de/rajewsky/miRDeep	Friedlander et al., 2008
miRanalyzer	http://bioinfo2.ugr.es/miRanalyzer/miRanalyzer.php	Hackenberg et al., 2009

RESULTS

The comparative study used to evaluate the software falls into three categories: 1) comparison of various miRNA precursor classifiers, 2) comparison of mining performance, and 3) influence of various RNA fold-software.

Data preparation

The evaluation of platform accuracy was based mainly on a known test dataset. Our test set included positive and negative datasets, and these were used for the comparative study. The positive dataset consisted of 1700 positive miRNA sequences that were either experimentally supported or obtained from the literature; they were actual pre-miRNAs. The positive dataset helped identify the number of true positives and false negatives, which are explained below, to define a series of evaluation criteria such as sensitivity, specificity, and accuracy for each software tool. The negative dataset consisted of 1700 negative miRNA sequences with stem-loop structures similar to those of miRNAs but not reported as pre-miRNAs (Zuker, 2003). The negative dataset helped identify the number of true negatives and false negatives, also explained below, to define a series of evaluation criteria similar to those determined using the positive dataset.

Comparison of various miRNA precursor classifiers

The positive and negative datasets were used to test the software, and we compared miRNA precursor classifiers by analyzing their outputs. Some important standards defined to evaluate the software are as follows:

Number of true positives (NTP): the number of experimentally supported miRNA precursors predicted by a program. Number of false positives (NFP): the number of negatives predicted by a program. Number of true negatives (NTN): the number of negatives not predicted by a program. Number of false negatives (NFN): the number of experimentally supported miRNA precursors not predicted by a program.

Other standards used to evaluate the performance of the predictive software products were sensitivity, specificity, and accuracy. These standards are always defined based on the 4 above-mentioned standards and are calculated as follows:

$$\text{Accuracy} = (NTP + NTN) / (NTP + NTN + NFP + NFN) * 100$$
$$\text{Specificity} = NTN / (NTN + NFP) * 100$$
$$\text{Sensitivity} = NTP / (NTP + NFN) * 100$$

The comparison of the four classifiers based on real-/pseudo-miRNA precursors is shown in Table 2.

Table 2. Comparison of four classifiers based on real-/pseudo-miRNA precursor.

Software	Positive data		Negative data		Sensitivity (%)	Specificity (%)	Accuracy (%)
	NTP	NFN	NTN	NFP			
microPred	1591	109	260	1440	93.59	15.29	37.51
MiPred	73	17	62	28	81.11	68.89	75.00
Virgo	931	408	1080	506	69.52	68.10	68.75
Triplet-SVM	1007	283	442	307	78.06	59.01	71.41

NTP = number of true positives; NFN = number of false negatives; NTN = number of true negatives; NFP = number of false positives.

Of the four softwares for the classification of real- and pseudo-pre-miRNAs, microPred was more sensitive in identifying pseudo-precursor miRNAs (93.59%), whereas Virgo was less sensitive (69.52%). The specificity of MiPred was as high as 68.89% compared to the 15.29% specificity of microPred. Although microPred had the highest sensitivity but relatively low accuracy owing to the lower specificity, the sensitivity of both microPred and MiPred were relatively large but because of the high specificity of MiPred, its overall accuracy was higher. Hence, MiPred is more effective than the other tools for classifying miRNA precursors.

Comparison of mining performance

Given a sequence, the software tools can mine probable miRNA. The total number of miRNAs contained in the gene sequence was known in advance. We chose part of chromosome 14, *cds1* and *cds2*, as our test set. We then compared the performance of the various mining software tools according to the number of miRNAs mined from the sequence and the number of true miRNAs contained in the mining miRNA. The standard of performance consisted of precision and recall. We acquired useful information through the experimental

method and provide a detailed list of all softwares. The pre-miRNAs predicted by the various software tools are shown in Table 3.

We can get the precision by dividing the number of predicted putative pre-miRNAs by positives in the putative pre-miRNAs. The rate of accuracy can also be calculated: precision = number of positives in the putative pre-miRNAs / number of predicted putative pre-miRNAs. Recall can be calculated: recall = number of positives in the putative pre-miRNAs / number of known pre-miRNAs.

Table 3. Real pre-miRNAs predicted by different softwares.

Software	No. of predicted putative pre-miRNAs	No. of known pre-miRNAs	No. of positive in putative pre-miRNAs	Precision (%)	Recall (%)
MiRAlign	16	25	14	87.50	56.00
miRAbela	16	26	15	93.75	57.69
MIReNA	38	42	27	71.05	64.29

Of the software considered for mining miRNA, both miRAlign and miRAbela had high precision and similar recall, with miRAbela being relatively higher (93.75%). The recall of MIReNA was as high as 64.29% compared to the 56-58% recall of miRAlign and miRAbela. We concluded that miRAbela has better precision, whereas MIReNA performs better in recall. We then compared these two software tools from the perspective that they mined miRNA incorrectly in *cds*, as shown in Figure 2.

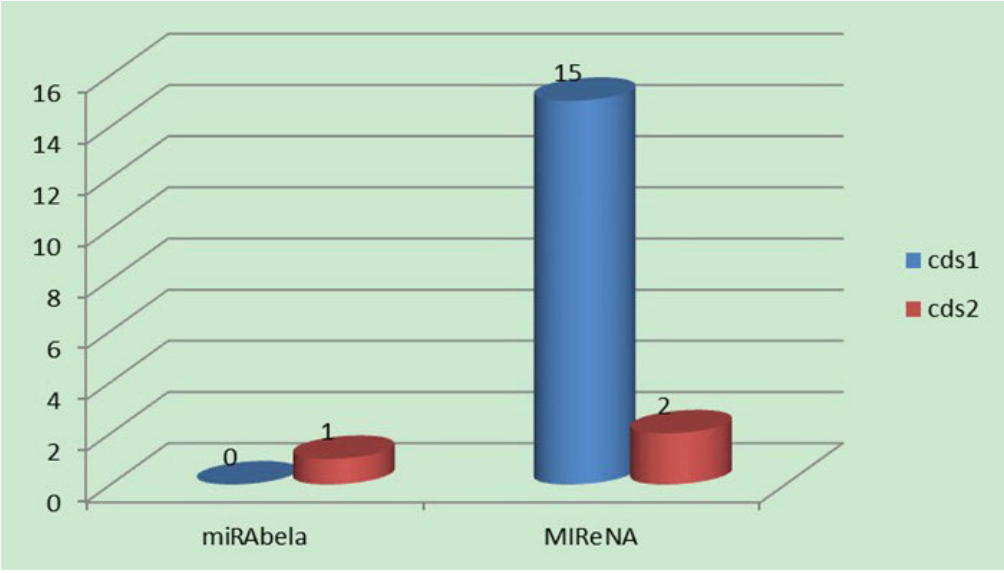

Figure 2. Pre-miRNAs predicted by different softwares.

miRAbela is superior to MIReNA and performed relatively well from this perspective. Other prediction software that mine miRNA precursors from genome or ESTs is summarized in Table 4.

Table 4. Some other prediction softwares were compared as follows.

Software	Online website	Local service	Online service	Others
miRPara	http://159.226.126.177/mirpara/cgi-bin/form.cgi	√	√	The result page cannot open after jumping from the home page
MiRscan	http://genes.mit.edu/mirscan/	√		Internal server error, web page cannot open
ProMiR II	http://cbit.snu.ac.kr/~ProMiR2/		√	Web page cannot open
BayesMiRNAfind	http://wotan.wistar.upenn.edu/miRNA/		√	Web page cannot open
miRDeep	http://www.mdc-berlin.de/rajewsky/miRDeep	√		Local service: the miRDeep package was developed to discover active known or novel miRNAs from deep sequencing data. Besides Perl, the Vienna package and Randfold application are required in the users's Linux box. Also needed is BLAST.
miRanalyzer	http://bioinfo2.ugr.es/miRanalyzer/miRanalyzer.php	√	√	Online service: the web server tool requires a simple input file containing a list of unique reads and its copy numbers. The users could choose to predict just new miRNA or just predict known miRNA in the input parameters optional. The output website shows some tables of detailed information. It provides the number of predicted miRNAs and predicts new miRNA numbers in the candidate miRNAs. Local service: besides the miRanalyzer, package should be installed, other packages like Weka and Vienna RNA package should be installed first. It is not convenient for the users to use the programs.

Influence of various RNA-fold softwares

The secondary structure of RNA (the base pair set for an RNA molecule) provides a scaffold for the tertiary structure (Zou et al., 2009). Yet the experimental determination of RNA structure remains difficult, and most researchers turn to computational methods. To date the most popular structure prediction algorithm is the minimum free-energy method for folding a single sequence. This algorithm has been implemented in two packages: m-fold (Zuker, 2003) and RNA-fold (Hofacker, 2003). RNA-fold (Sankoff et al., 1983; Zuker, 1989a,b) computes a single-minimum energy folding of an RNA sequence. The m-fold software was developed in the late 1980s (Zuker, 1989b). The m simply refers to multiple. In the prediction of a single sequence, both RNA-fold and m-fold use the dynamic programming method to calculate minimum free energy, so the forecast effect is approximate. Most of the prediction software above uses RNA-fold to fold the miRNA secondary structure.

To analyze the influence that the various secondary structure prediction software has on the recognition of miRNA, we choose Triplet-SVM as the analysis object from among the prediction software, as shown in Figure 3. The Triplet-SVM classifier runs directly on Linux with a Perl compiler, and this package requires a third-party software - namely RNA-fold and Libsvm. Therefore, we used RNA-fold and m-fold to fold the secondary structure of miRNA and then trained and tested the software with our test set. Our experiments showed that with the same training set and classification software, different secondary structure prediction software tools produce different effects in the predicting outcomes. With Triplet-SVM, RNA-fold yields better results.

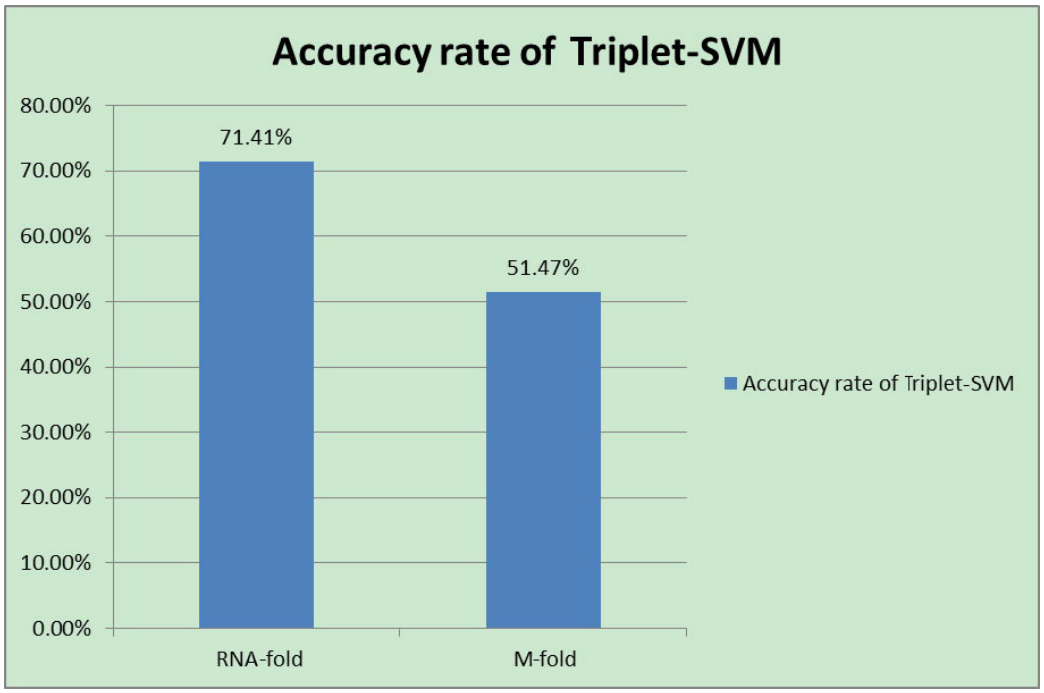

Figure 3. Prediction accuracy of Triplet-SVM that uses different RNA-fold softwares.

DISCUSSION

Many *ab initio* methods have been developed to predict miRNAs. Parts of the methods based on 3 characteristics of miRNA gene identification tools have been compared to understand their relative performance. Among the methods, MiPred performs best for classifying miRNA precursors from pseudo-hairpins, and miRAlign displayed superior performance in mining miRNA precursors from genome or ESTs. One class approach can be a good alternative, but for overall accuracy, improvements need to be incorporated for better performance.

Most prediction software tools currently predict novel miRNAs according to the structure and properties of a small number of known miRNAs, but little is known about miRNAs, and the existing methods have no general applicability, so the prediction effect is not one of confidence. To explore and reveal the mystery of miRNA further, more in-depth research is needed. In addition, the development of NGS technologies is central in the prediction and discovery of novel miRNAs.

ACKNOWLEDGMENTS

Research supported by the Natural Science Foundation of China (#61001013, #61102136, and #61001143) and the Natural Science Foundation of Fujian Province of China (#2011J05158).

REFERENCES

Batuwita R and Palade V (2009). microPred: effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics* 25: 989-995.

Bentwich I, Avniel A, Karov Y, Aharonov R, et al. (2005). Identification of hundreds of conserved and nonconserved human microRNAs. *Nat. Genet.* 37: 766-770.

Borchert GM, Lanier W and Davidson BL (2006). RNA polymerase III transcribes human microRNAs. *Nat. Struct. Mol. Biol.* 13: 1097-1101.

Brennecke J, Hipfner DR, Stark A, Russell RB, et al. (2003). bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. *Cell* 113: 25-36.

Carrington JC and Ambros V (2003). Role of microRNAs in plant and animal development. *Science* 301: 336-338.

Friedlander MR, Chen W, Adamidi C, Maaskola J, et al. (2008). Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.* 26: 407-415.

Hackenberg M, Sturm M, Langenberger D, Falcon-Perez JM, et al. (2009). miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res.* 37: W68-W76.

Hofacker IL (2003). Vienna RNA secondary structure server. *Nucleic Acids Res.* 31: 3429-3431.

Huang JC, Babak T, Corson TW, Chua G, et al. (2007). Using expression profiling data to identify human microRNA targets. *Nat. Methods* 4: 1045-1049.

Huang Y, Zou Q, Tang SM, Wang LG, et al. (2010). Computational identification and characteristics of novel microRNAs from the silkworm (*Bombyx mori* L.). *Mol. Biol. Rep.* 37: 3171-3176.

Huang Y, Shen XJ, Zou Q, Wang SP, et al. (2011a). Biological functions of microRNAs: a review. *J. Physiol. Biochem.* 67: 129-139.

Huang Y, Zou Q, Wang SP, Tang SM, et al. (2011b). The discovery approaches and detection methods of microRNAs. *Mol. Biol. Rep.* 38: 4125-4135.

Jiang P, Wu H, Wang W, Ma W, et al. (2007). MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res.* 35: W339-W344.

Kumar S, Ansari FA and Scaria V (2009). Prediction of viral microRNA precursors based on human microRNA precursor sequence and structural features. *Virol. J.* 6: 129.

Lee Y, Ahn C, Han J, Choi H, et al. (2003). The nuclear RNase III Drosha initiates microRNA processing. *Nature* 425: 415-419.

Li PW, Lu XY, Li CZ, Fang J, et al. (2007). Advances in the study of plant microRNAs. *Yi Chuan* 29: 283-288.

Lim LP, Lau NC, Weinstein EG, Abdelhakim A, et al. (2003). The microRNAs of *Caenorhabditis elegans*. *Genes Dev.* 17: 991-1008.

Mathelier A and Carbone A (2010). MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. *Bioinformatics* 26: 2226-2234.

Moss EG, Lee RC and Ambros V (1997). The cold shock domain protein LIN-28 controls developmental timing in *C. elegans* and is regulated by the lin-4 RNA. *Cell* 88: 637-646.

Nam JW, Shin KR, Han J, Lee Y, et al. (2005). Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Res.* 33: 3570-3581.

Nam JW, Kim J, Kim SK and Zhang BT (2006). ProMiR II: a web server for the probabilistic prediction of clustered, nonclustered, conserved and nonconserved microRNAs. *Nucleic Acids Res.* 34: W455-W458.

Park MY, Wu G, Gonzalez-Sulser A, Vaucheret H, et al. (2005). Nuclear processing and export of microRNAs in *Arabidopsis*. *Proc. Natl. Acad. Sci. U. S. A.* 102: 3691-3696.

Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, et al. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature* 403: 901-906.

Ruby JG, Jan C, Player C, Axtell MJ, et al. (2006). Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. *Cell* 127: 1193-1207.

Sankoff D, Kruskal JB, Mainville S and Cedergren RJ (1983). Fast Algorithms to Determine RNA Secondary Structures Containing Multiple Loops. In: Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison (Sankoff D and Kruskal JB, eds.). Chapter 3. Addison-Wesley, Reading, 93-120.

Sewer A, Paul N, Landgraf P, Aravin A, et al. (2005). Identification of clustered microRNAs using an *ab initio* prediction method. *BMC Bioinformatics* 6: 267.

Wang X, Zhang J, Li F, Gu J, et al. (2005). MicroRNA identification based on sequence and structure alignment. *Bioinformatics* 21: 3610-3614.

Wu Y, Wei B, Liu H, Li T, et al. (2011). MiRPara: a SVM-based software tool for prediction of most probable microRNA

coding regions in genome scale sequences. *BMC Bioinformatics* 12: 107.
Xue C, Li F, He T, Liu GP, et al. (2005). Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics* 6: 310.
Yousef M, Nebozhyn M, Shatkay H, Kanterakis S, et al. (2006). Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier. *Bioinformatics* 22: 1325-1334.
Zeng Y, Yi R and Cullen BR (2005). Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha. *EMBO J.* 24: 138-148.
Zou Q, Zhao T, Liu Y and Guo M (2009). Predicting RNA secondary structure based on the class information and Hopfield network. *Comput. Biol. Med.* 39: 206-214.
Zou Q, Lin C, Liu XY, Han YP, et al. (2011). Novel representation of RNA secondary structure used to improve prediction algorithms. *Genet. Mol. Res.* 10: 1986-1998.
Zuker M (1989a). Computer prediction of RNA structure. *Methods Enzymol.* 180: 262-288.
Zuker M (1989b). On finding all suboptimal foldings of an RNA molecule. *Science* 244: 48-52.
Zuker M (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31: 3406-3415.
Zuker M and Stiegler P (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* 9: 133-148.