Computational Approaches in Detecting Non- Coding RNA

Chunyu Wang[1], Leyi Wei[2], Maozu Guo[1,]* and Quan Zou[2,]*

[1]*School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China;* [2]*School of Information Science and Technology, Xiamen University, Xiamen 361005, China*

Abstract: The important role of non coding RNAs (ncRNAs) in the cell has made their identification a critical issue in the biological research. However, traditional approaches such as PT-PCR and Northern Blot are costly. With recent progress in bioinformatics and computational prediction technology, the discovery of ncRNAs has become realistically possible. This paper aims to introduce major computational approaches in the identification of ncRNAs, including homologous search, *de novo* prediction and mining in deep sequencing data. Furthermore, related software tools have been compared and reviewed along with a discussion on future improvements.

Received on: April 26, 2013- Revised on: July 18, 2013- Accepted on: July 18, 2013

Keywords: Non-coding RNA, Machine learning, Bioinformatics, lncRNA, microRNA, Deep sequencing.

INTRODUCTION

In the early biological researches, scientists mainly focused on the prokaryotes, which are dominated (85%-90%) by protein-coding genes [1] and it is publicly considered that the cellular activities are implemented by the proteins which are transcribed from those coding genes. But in the evolution of species, the relative proportions of the coding genes are gradually reducing, whereas the variety of cellular functions increasing. It is estimated that 98% of mammalian genomic output may be non-coding RNAs (ncRNAs), while the remaining 2% encodes the proteins [2]. However, at present, we have incomplete knowledge of those non-coding regions containing both non-coding genes and genomic elements which may regulate gene expression [3]. As a result, we are currently more interested in the non-coding regions which could lead to the better understanding of biological processes. They may be involved in the gene expression control, cancer and aging [61].

As stated in the central dogma of molecular biology, gene sequences (DNA) are transcribed into RNA according to the law of chemistry and physics. Some RNAs, including messenger RNA (mRNA), are called coding RNAs, since they are translated into protein and the others are called non-coding RNAs, because of their function as RNA molecules rather than coding protein. Non-coding RNAs are involved in translation, splicing, gene regulation, chromatin remodeling, gene modification, degradation and other functions [61]. They are also closely associated with cancer and other complex diseases [4]. Many kinds of functional ncRNAs have been discovered with biological experiments and computational methods. In the literature, ncRNAs are divided into several categories [5]. Some ncRNAs are named according to their functions, like microRNAs (miRNA), package RNAs (pRNA) or transfer RNAs (tRNA), etc. Others are named by their cellular localizations, such as piRNAs (interact with piwi protein) and rasiRNAs (repeat associated small interfering RNAs). There are still many other ncRNAs unclassified.

Non-coding RNAs are recognized only in biological experiments with technologies such as full-length complementary DNA cloning and genomic tiling arrays in the transcriptomes of organisms. Although these technologies can suit long ncRNA (lncRNA) genes in an efficient way, they are costly always requiring enough RNA samples, and are therefore limited. To overcome this shortage, researchers have developed computational biology approaches to discover ncRNAs [6] and are incorporating these computational approaches in experimental methods [7].

Although these computational methods and software tools have their characteristics, a unified framework for identifying all ncRNAs still needs to be discovered due to the diversity of ncRNAs, missing common sequence features and the lack of post-transcriptional processing information. Firstly, there are many kinds of ncRNAs in the species. For example, tRNAs and rRNAs involve in protein production; miRNAs control gene expression; snoRNAs modify post-transcription of other RNA molecules [8]. Different functions are induced by diverse ncRNA structures and there are also variations in ncRNA length. Secondly, they are different from protein-coding genes which have a lot of common conserved features of primary sequence, including splice cites, promoters, terminator and binding motifs etc. There are few common primary features among the non-coding RNAs, so it is difficult to identify all ncRNAs from primary sequence [9]. Therefore, although common primary features may exist in certain ncRNA families, researchers could not simply apply these features for identifying all

*Address correspondence to these authors at the Computer Science Department, 422#Siming Road, Xiamen, China, (Zip, 361005); Tel: +86-13656009020; Fax: +86-592-2580033; E-mail: zouquan@xmu.edu.cn and 319 box, 92#Xidazhi Road, Harbin, China, (Zip, 150001); Tel: +86-13654646103; Fax: +86-451-86402407; E-mail: maozuguo@hit.edu.cn

1875-5488/13 $58.00+.00

ncRNAs. Finally, many ncRNAs primarily transcribed from non-coding genes will go through the post-transcriptional processing to reach maturity. The modifications are intended to change their structures, which are closely related to their functions. So far, we are still unable to explain the modifications with the current knowledge and predict them with software programs. Consequently, we try to introduce most currently existing approaches about the identification of ncRNAs.

METHODS USED IN ncRNA IDENTIFICATION

A lot of computational approaches for detecting ncRNA genes have been designed and reported, but as the variations of ncRNAs, most of these methods are developed for specific ncRNAs or specific ncRNA family. In general, these methods can be divided into two classes.

1) Methods based on homology information. These methods always require homology information and a good quality of alignment among sequences. Only those ncRNAs, which are homologous to known ncRNA family, can be discovered with these methods. Novel computational methods need to be developed for identifying novel ncRNAs.

2) Methods based on common features in ncRNA genes. These methods are called "*de novo*" approaches, which do not require homology information and sequence alignment except known sequence and structural features derived easily from the genome. In addition, the machine learning method known as Random Forest or Support Vector Machine is often used to predict ncRNA genes based on features. In fact, ncRNA gene-finding method based on nucleotide composition like (G+C)% has made some success in some specific species genomes [10]. However, other investigations have indicated that programs based compositions alone are not sufficient to identify ncRNA genes effectively [11]. As a consequence, when using the *de novo* methods for identifying ncRNAs, we can derive features, including sequence and structural features, from sequences and select the appropriate classifier to model these features and train the model to achieve high accuracy of ncRNA prediction [12].

METHODS FOR HOMOLOGY-BASED ncRNA IDENTIFICATION

It is commonly believed that most ncRNAs are less conserved in sequence [8]. Although there are few common features in ncRNA sequences, it may be different for some special ncRNA families. So common sequence and structural characteristics are used as homology information to detect these ncRNAs. Homology search is to detect all homologous genes in the target sequences, given one or more ncRNA which could represent for a specific ncRNA family. Consequently, we mainly focus on their sequence homology and structure homology, which are based on their features respectively.

There are some software tools based on sequence homology, such as BLAST [13], FASTA [14], S Search [15] and BLAT [16] etc. BLAST (Basic Local Alignment Search Tool) employs a measure of sequence similarities between input sequences and known ncRNAs. A score derived from BLAST approximately quantifies this similarity. However, ncRNAs rarely preserve high degree of the similarity. Furthermore, it relies too heavily on individual sequences rather than focusing on the common features of the ncRNA family. Consequently, a family of homologous sequences is aligned to find the positions conserved than others. And then BLAST is used for finding these common positions in the alignments for target sequence when looking for the additional ncRNA family member. BLAT, which is the abbreviation of "BLAST-like alignment tool", is similar in many ways to BLAST. When multiple sequences as inputs are aligned to a large sequence database, BLAT performs at higher speed than BLAST. In addition, BLAT has also high sensitivity and specificity for ncRNA detection [12]. In the research of sequence homology, a probabilistic model, named Hidden Markov Model (HMM), which models the features of the homologous sequences, is also used to predict ncRNAs. It builds a model representing the consensus sequence for the family, not the sequence of any particular member [17].

It is naturally hard to identify ncRNAs effectively on sequence level when the level of sequence homology is low. Secondary structure of sequence is more conserved than sequence in the long evolutionary time [18]. Consequently, structure homology is also used to detect ncRNA genes. Programs INFERNAL [18], Rsearch [19] and FastR [20] are all based on structure homology. FastR package is applied to search homologous structure of ncRNAs in large genomic sequence [20]. INFERNAL and Rsearch allow for searching a sequence database for homologous ncRNAs, which are given and structured [15]. Taken tRNA as an example, it is publicly believed that tRNA has a classical "cloverleaf" structure. When Rsearch and BLAST or FASTA are both used to identify tRNA, prediction accuracy from Rsearch is better than that from BLAST or FASTA.

At present, there are many approaches based on a combination of sequence and structure homology to identify ncRNA. For example, when we annotate miRNA genes, known mature miRNAs are mapped into the predicted sequences with BLAST. After that, the mapped sequences are in high level of sequence homology with known miRNA. Then a model is built for the mapped sequences with their structure homology information, including secondary structures, pairwise sequence alignment and structural alignment. Finally, we get a measure of similarity of sequences to annotate true miRNAs, as shown in Fig. (**1**).

Among the current prediction tools based on homology information, most are designed using both sequence homology and structure homology information, such as ERPIN [21] and miRAlign [22]. ERPIN and BLAST are both used to detect new miRNAs. As a consequence, ERPIN increases the number of new miRNA candidates by 17% compared to a BLAST search. The result means that the programs using a combination of sequence and structure homology can get higher accuracy to identify conserved structure ncRNAs than those programs based only on sequence homology. But ERPIN package is limited to identify miRNAs when there are not sufficient known miRNA samples. On the contrary, miRAlign is applicable to identify novel miRNAs with few

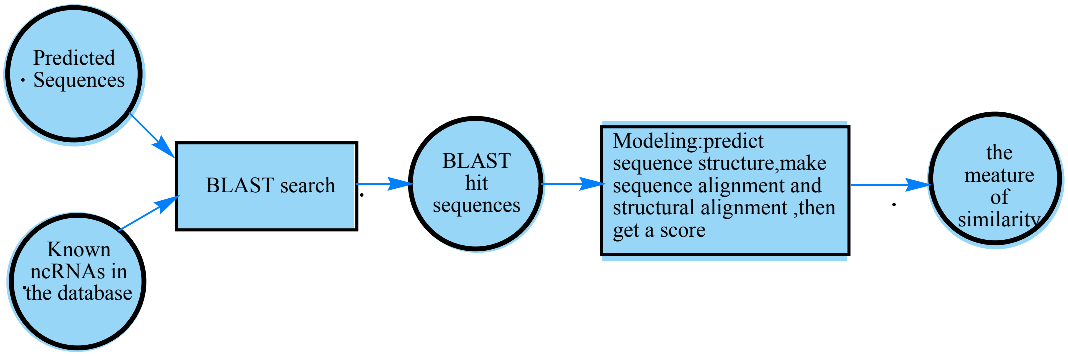
Predicted Sequences
Known ncRNAs in the database
BLAST search
BLAST hit sequences
Modeling:predict sequence structure,make sequence alignment and structural alignment ,then get a score
the meature of similarity

Fig. (1). Overview of approaches based on a combination of sequence and structure homology.

known miRNA samples. And in order to investigate the ability of miRAlign to identify miRNA in different species, researchers compared miRAlign with BLAST and ERPIN. Consequently, miRAlign achieved higher specificity and sensitivity compared to that exhibited by BLAST and ERPIN searches [18]. As the sequences and structures are deeply researched in miRNA, HMM (Hidden Markov Model) is used for the miRNA precursors [23] and targets [24] identification. For some ncRNAs, when the secondary structure is conserved, such as tRNA and H/ACA box snoRNA, context-sensitive HMM is used for identification [25]. And the non-coding RNA database RFAM is built based on HMM [26].

METHODS FOR *de novo* ncRNA IDENTIFICATION

We can not discover novel ncRNA families with homology-based methods which rely mostly on homology information. Thus, de novo approaches are developed to solve this problem using features derived from the sequences and structures of known ncRNA genes.

Methods Based on Sequence Features

In the earlier studies, nucleotide composition was used as sequence features to identify ncRNA in some nucleotide compositional bias species. For example in an AT-rich extreme hyperthermophile, ncRNA genes with a stable secondary structure might be found by calculating GC content**,** which is intended to stabilize their structures in the high temperature environment [8]. However, a single feature is not sufficient to identify ncRNA effectively. As a result, other sequence features have been employed to combine with nucleotide composition to detect ncRNA, including di- and tri-nucleotide frequencies, known RNA motifs and folding energy etc.

Currently, many programs based on sequence features have been developed, such as CRITICA [27], CST miner [28] and EST scan [29]. Researchers utilize these three programs to identify ncRNA from the 102801 FANTOM sequences respectively and find that CRITICA shows the highest degree of concordance which is up to 94.8% with the other two programs. And its concordance reveals the individual prediction accuracy of each program [30]. Furthermore, the machine learning algorithms are added into ncRNA identification. For example, CONC [31] takes sequence features as input and then uses SVM (Support Vector Machine) to train these features. It has high specificity and sensitivity for ncRNA annotation. However, CONC is slow to the large datasets and spends much computing resources. Compared to CONC, we run CPC [32] on two datasets including one non-coding RNA dataset and one protein-coding dataset respectively and record its result in (Table **1**). What we get from the result is that CPC has higher accuracy and consumes lower time and space than CONC.

Methods Based on Structure Features

It is publicly known that RNA molecule is a single strand, and usually folds into secondary structure, which is more conserved than primary sequence in long distant evolution. Thus, we investigate approaches for incorporating secondary structure into identification of novel ncRNAs. Actually, the minimum folding energy (MFE) approach is extensively used to predict secondary structure of the target sequences [33]. For example, the programs RNAfold [34], Mfold [35] and Afold [36] all based on this approach have successfully been applied for novel ncRNA identification. To achieve high sensitivity and specificity, an alternative approach, Sfold also incorporates a probabilistic model in the prediction [37]. In addition, searching novel H/ACA snoRNA in the yeast or other eukaryote genomes, the approach based on MFE structure could also provide good prediction [38].

Table 1. Evaluation of Accuracy and CPU Time of CPC and CONC on Two Datasets

Dataset	Dataset type	Dataset size	Accuracy		Time(min)	
			CPC	CONC	CPC	CONC
Rfam	Non-coding	9020	87.57%	85.36%	1053	13594
Embl cds	Coding	8949	93.24%	92.93%	5073	60647

However, secondary structure alone is generally not efficient enough for the detection of ncRNAs [39]. For a given sequence, it might fold into different secondary structures but these structures intend to have similar MFE. Thus, other structure features are extensively discovered and applied to distinguish ncRNA from the target sequences, such as thermodynamic stability and shannon entropy etc. For example, in a data set, containing real ncRNAs and their di-shuffled sequences, the di-shuffled sequences are intended to have higher MFE and Shannon entropy than the real ncRNAs [40]. In addition, a program MiPred [41], employs a combination of features, containing local contiguous structure-sequence composition, MFE and P-value of randomization test, uses a novel machine-learning technique based on random forest algorithm to identify putative miRNA precursors and seems to provide high sensitivity and specificity. Furthermore, PlantMiRNAPred [42] can classify plant miRNA precursors efficiently by SVM together with feature and sample selection strategies. It selects a variety of features from both primary sequence and secondary structure and provides a viable method for discovering novel plant pre-miRNAs. However, all these methods and software mentioned above suit for the long DNA sequences, such as EST or genome data. When dealing with deep sequencing or the next generation sequencing data, it needs more mapping or assembling strategies.

Methods Based on Deep Sequencing Technology

With the development of the next generation sequencing technologies, it has been implemented for small ncRNAs discovery, particularly for miRNAs. Massively next generation sequencing technologies (also named deep sequencing) are currently in widespread use, including 454, Solexa and SOLiD. Compared to conventional sequencing technologies, deep sequencing technologies accelerate biological research and significantly reduce the cost. Here we present currently available tools for miRNA identification with the deep sequencing technologies, including miRDeep [43], CID-miRNA [44], MiRank [45], miRCat (identify plant miRNAs) [46], mirTool [47], and miRanalyzer [48].

MiRNA discovery with deep technologies is generally divided into two steps. The first step is called filtering. The sequence reads derived from deep sequencing are mapped to the whole genome. The reads that map to tRNA or sRNA, etc are discarded and then the remaining reads are mapped to known miRNA database again. The sequence reads that map to the known miRNA database are passed and recognized as miRNA candidates. The other step is called modeling. The miRNA candidates are simply modeled by some algorithms. For example, in the core algorithm of miRDeep, potential miRNA candidates are modeled for the combined compatibility of energetic stability, positions and frequencies of reads with Dicer processing [33]. A number of features contribute to the final score derived from the model. miRDeep could discover not only known and novel miRNAs but also provide a statistical evaluation of false positive rate and sensitivity, which most machine learning algorithm could not provide. The flow is shown in Fig. (**2**).

However, compared to other tools, miRDeep relies on the characteristic pattern of high expression, thus it is limited for the miRNA at low level of expression. In this situation, it needs researchers to explore other means to indentify novel miRNA in the low expression sequences. For example, the conservation pattern of structure can be used to discover miRNA precursors [49]. Firstly, taking the sequence reads to map the whole genome, we can remove those reads that do not map to genome then fold remaining reads with RNA by Vienna package [25]. The novel hairpins produced by Vienna are filtered, while those single –loop hairpins with mature-miRNA in one side of hairpin are passed as possible hairpins. Secondly, these possible hairpins are refolded by the Vienna package and filtered again with Ambros criteria [50]. Finally, real mature-miRNA can be discovered in the remaining hairpins.

MiRanalyzer, which is similar to miRDeep, could search known miRNA in the miRNA database and discover novel miRNA, particularly those undiscovered miRNA family. The core algorithm of MiRanalyzer is a sensitive machine learning method using random forest algorithm. And the feature selection technologies are also used in the MiRanalyzer. Subsequently, the prediction of new miRNAs using MiRanalyzer could reach high sensitivity and keep a low level of false positive rate [38]. Similar to miRDeep and MiRanalyzer, MirTool also can predict known and novel miRNAs. Furthermore, it could provide detailed information for the known miRNAs, such as miRNA/miRNA* and absolute/relative reads count [37]. Here is another program called miRank using random walk-based ranking algorithm. The miRank method has some remarking properties. For example, it does not rely on cross-species conservation so that it can identify species-specific miRNAs. In addition, it does not require a number of miRNA samples but could reach a high discover accuracy. Hence, miRank is a useful tool for the miRNA identification [35]. Besides using deep sequenc-

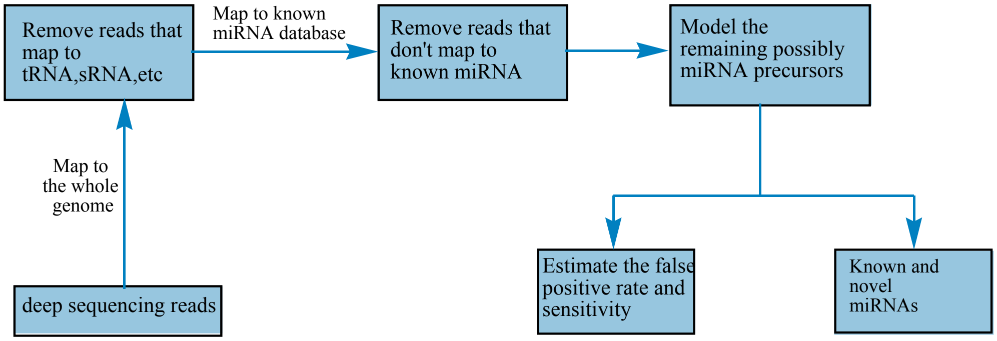
Remove reads that map to tRNA,sRNA,etc
Map to known miRNA database
Remove reads that don't map to known miRNA
Model the remaining possibly miRNA precursors
Map to the whole genome
deep sequencing reads
Estimate the false positive rate and sensitivity
Known and novel miRNAs

Fig. (2). Flow chart of miRNA discovery using deep sequencing technology.

ing technologies to detect miRNAs, other small ncRNA molecules such as snoRNA, piRNA, endo-siRNA, are also identified by deep sequencing technologies. For example, we can apply SnoSeeker which can identify snoRNAs from deep sequencing data [51]. (Table **2**) lists the main software tools for the ncRNA discovery.

Indeed, the main disadvantage of the popular software tools is that they are designed specially for just one kind of ncRNAs. For example, tRNAscanSE is used for detecting tRNA; snoSeeker is to look for snoRNA; miRDeep, miRCat, mirTool, miRanalyzer and MIReNA are all designed for mining microRNA. So there are working repeat and confusion for the conflict result when we annotate new sequencing data. Moreover, they are unfair to compare for ncRNA annotation.

Two methods, i.e., CSHMM and MIReNA, are employed for comparison. CSHMM is the machine learning based method used either to analyze individual sequences or scan potential pre-miRNAs from human genome-scale data. MIReNA is the method based on a genome-wide search algorithm for pre-miRNAs search. Both methods are used to search for known pre-miRNAs from the Chr 19. Results show that 70 and 74 true positives are correctly predicted by CSHMM, and MIReNA, respectively. Regarding methods to detect pre-miRNAs from human genome-scale data, maintaining high specificity is even of greater importance. MIReNA correctly predicts the highest number of true positives, but also produces 10,626 pre-miRNA candidates, while CSHMM predicts 18,258 false true results. This large number of pre-miRNA candidates is perhaps due to the low specificity of MIReNA. Similarly, the low specificity results in CSHMM as a poor choice for the identification of pre-miRNAs. Actually, MIReNA is capable of performing in different modes when handling different data types (e.g., the genome data, and the deep-sequencing data). It is believed that MIReNA can achieve better performance when evaluated on a more comprehensive data.

In the general, *de novo* methods use machine learning algorithms to train the features from sequence, structure and deep sequencing data to identify ncRNAs. With the development of bioinformatics, more and more features are derived and used in the ncRNA discovery. However, in most cases, these features are redundant. In order to reduce the redundancy, features selection technologies are created and applied to *de novo* methods. Now the main feature selection technologies constantly used are Filter, Wrapper and Embedded technologies [52]. On the other hand, different machine learning methods employed in the discovery are intended to lead to different efficiency and accuracy. Compared to those classifiers, such as SVM and Bayesian, an integrated machine learning model called incRNA is developed and can significantly improve the results in the ncRNA identification [53].

lncRNA AND lncRNA IDENTIFICATION

Besides small non-coding RNAs, there are long non-coding RNAs (lncRNA), which are longer than 200nt. They can be categorized into long intronic non-coding RNA and intergenic non-coding RNA. They are considered to regulate gene expression through changes in chromatin state, implicate in cancer pathogenesis and correlate with clinical features [58]. With the increasing amount of lncRNAs, identification and function research for lncRNA is called lncRNome [59].

Since the lack of conservation among lncRNA primary sequences, detecting lncRNAs from genomes relies on expression analysis that makes comprehensive characterization of lncRNome difficult. The latest GENCODE has specially collected lncRNAs. First, the transcriptome data were annotated and the protein coding sequences were filtered. Then short sequences, which were shorter than 200nt were removed and the remaining ones were viewed as lncRNAs [56]. Since the lack of experimental transcriptome data, computational prediction of lncRNAs is necessary and meaningful. Machine learning method based on SVM was employed for detecting polycomb-associated lncRNAs [57]. It can distinguish lncRNAs from transcription noise. However, features extraction for lncRNA is a challenging task since lncRNAs are largely unstructured. So regulation elements such as enhancers or promoters are always utilized. Although lncRNAs do not have common secondary structures, structure features can be used for distinguishing lncRNAs with other small ncRNA precursors [60].

DISCUSSIONS

Although many ncRNA families have been discovered by variety of identification tools, there is currently no unified prediction tool which could detect all kinds of ncRNA. For the specific ncRNA research, it will lead us to develop different programs. In fact, most current approaches and tools are complementary. In order to improve specificity and sensitivity as well as reduce false-positive, we are interested at how to combine these methods and tools effectively. This process seems a little complicated, because it requires us to evaluate different combinational methods, which represent another direction of ncRNA identification.

Table 2. The Main Software Tools of ncRNA Discovery

Homology-based ncRNA identification methods	BLAST, Blat, INFERNAL, FASTA, SSEARCH, Rsearch, FastR, ERPIN, miRAlign	
De novo-based ncRNA identification methods	Sequence features-based methods	CRITICA, CSTminer, ESTscan, CONC, CPC
	Structure features-based methods	RNAfold, Mfold, Afold, MiPred
	Deep sequencing-based methods	miRDeep, CID-miRNA, MiRank, miRCat, mirTool, snoSeeker, MiRanalyzer

In the field of ncRNA identification based on homology information, the selection of window size of alignment sequences would be a problem to limit us to use sequence homology methods, because the fixed alignment programs typically assume a window size to reduce computational requirements. In addition, the window size not suitable for the sequence alignments might reduce the prediction accuracy. When level of sequence homology is relatively low, alternative methods based on structure homology are applicable to detect new ncRNAs. Structural alignment approaches are incorporated into the structure homology research. And how to improve speed of structural alignments and their accuracy becomes another area of active research.

In *de novo* methods, most of them are based on the features derived from sequence and structure. By utilizing these features, kinds of classifiers have been applied to the research. At present, how to combine these features and select a proper classifier represent another direction in the field of ncRNA identification. With the quick development of next generation sequencing technologies, massive sequencing data provides a great deal of power to the ncRNA research.

Computation detecting methods mentioned above are mostly designed for the short non-coding RNAs, such as miRNAs, tRNAs, siRNAs, piRNAs, etc. When dealing with long non-coding RNAs (lncRNA), the computation methods always can not work well. To our knowledge, RT-PCR or CHIP-SEQ is the main detecting method for lncRNA [54, 55]. More research ought to be done on the lncRNA and the computational detecting methods are required for decreasing the molecular biology experiments cost.

In conclusion, ncRNA research is still at its infancy. To get more knowledge about complex ncRNA world, we still have to explore other novel methods, either biological experiments or computational methods. However, since experimental technology has not yet been developed, we should still focus on exploring ncRNA world by computational tools. And novel insights do not only help us to increase our knowledge about RNA world, but also help us to improve computational tools for further identification. Besides detection methods, more computation methods and tools need to be researched deeply, including identification alternative splicing or SNP in ncRNA. They are both interesting and important for the function research of ncRNA.

CONFLICT OF INTEREST

The author(s) confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

The work was supported by the Natural Science Foundation of China under Grant No. 60932008, 61172098, 61001013, 61102136 and 61271346, and the Specialized Research Fund for the Doctoral Program of Higher Education of China under Grant No. 20112302110040.

REFERENCES

[1] Mattick, J.S.; Makunin, I.V. Non-coding RNA. *Hum Mol Genet.*, **2006**, 15 (suppl 1), R17-R29.

[2] Hu, L.L.; Huang, Y.; Wang, Q.C.; Zou, Q.; Jiang, Y. A benmark comparison of ab initio microRNA identification methods and software. *Genet. Mol. Res.*, **2012**, *11*(4), 4525-4538.

[3] Wei, L.Y.; Huang, Y.; Qu, Y.Y.; Jiang, Y.; Zou, Q. Computational analysis of miRNA target identification. *Curr. Bioinform.*, **2012**, *7*(4), 512-525.

[4] Liang, R.Q.; David, J.B.; Wang, E. Epigenetic Control of MicroRNA Expression and Aging. *Curr. Genomics.*, **2009**, *10*(3), 184-193.

[5] Kapranov, P.; Willingham, A.T.; Gingeras, T.R. Genome-wide transcription and the implications for genomic organization. *Nat. Rev. Genet.*, **2007**, *8*(6), 413–423.

[6] Eddy, S.R. Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.*, **2007**, *2*(12), 919–929.

[7] Huang, Y.; Zou, Q.; Tang, S.M.; Wang, L.G.; Shen X.J. Computational identification and characteristics of novel microRNAs from the silkworm (Bombyx mori L.). *Mol. Biol .Rep.*, **2010**, *37*(7), 3171-3176.

[8] Wang X.J.; Reyes, J.L.; Chua, N.H.; Gaasterland, T. Prediction and identification of Arabidopsis thaliana microRNA genes and their mRNA targets. *Genome. Biol*., **2004**, 5:R65.

[9] Wang, E. MicroRNA Regulation and its Biological Significance in Personalized Medicine and Aging. *Curr. Genomics.*, **2009**, *10*(3), 143.

[10] Livny, J.; Waldor, M.K. Identification of small RNAs in diverse bacterial species. *Curr. Opin .Microbiol.*, **2007**, *10*(2), 96-1001.

[11] Klein, R.J.; Ziva, M.; Sean R.E. Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proc. Natl. Acad. Sci. USA.*, **2002**, *99*(11), 7542-7547.

[12] Xue, C.H.; Li, F.; He, T.; Liu, G.P.; Li, Y.D.; Zhang, X.G. Classification of Real and Pseudo MicroRNA Precursors Using Local Structure-Sequence Features And Support Vector Machine. *BMC Bioinformatics.*, **2005**, *6*, 310.

[13] Jana, H.; Danielle, D.J.; Manja, M.; Dominic, R.; Hakim, T.; Andrea, T.; Bernd, S.; Peter, F.S. Non-coding RNA annotation of the genome of Trichoplax adhaerens. *Nucleic Acids Res.*, **2009**, *37*(5), 1602-1615.

[14] Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.*, **1990**, *215*(3), 403-10.

[15] Pearson, W.R. Flexible sequence similarity searching with the FASTA3 program package. *Method Mol. Biol.*, **2000**, *132*, 185-219.

[16] Pearson, W.R. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomic.*, **1991**, *11*(3), 635-650.

[17] Kent, W.J. Blat: the BLAST-like alignment tool. *Genome Res.*, **2002**, *12*(4), 656-664.

[18] Wang, X.Q.D.; Crutchley, J.L.; Dostie, J. Shaping the Genome with Non-Coding RNAs. *Curr. Genomics.*, **2011**, *12*(5), 307-321.

[19] Eddy, S.R. A memory efficient dynamic programming algorithm for optimal structural alignment of a sequence to an RNA secondary structure. *BMC Bionformatic.*, **2002**, *3*, 18.

[20] Klein, R.J.; Eddy, S.R. RSEARCH: finding homologs of single structured RNA sequences. *BMC Bioinformatics.*, **2003**, *4*, 44.

[21] Zhang. S,; Haas, B.; Eskin, E.; Bafna, V. Searching Genomes for Noncoding RNA Using FastR. *IEEE/ACM Trans Comput .Biol .Bioinform.*, **2005**, *2*(4), 366-379.

[22] Gautheret, D.; Lambert, A. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *J.Mol.Biol.*, **2001**, *313*(5), 1003-1011.

[23] Wang, X.W.; Zhang, J.; Li, F.; Gu, J.; He, T.; Zhang, X.G.; Li, Y.D. MicroRNA identification based on sequence and structure alignment. *Bioinformatics.*, **2005**, *21*(18), 3610-3614.

[24] Sumeet, A.; Candida, V.; Alok, B.; Ashwin S. Prediction of novel precursor miRNAs using a context-sensitive hidden Markov model (CSHMM). *BMC Bioinformatics.*, **2010**, 11, S29.

[25] Fu, H.Y.; Xue, D.Y.; Zhang, X.D.; Yang, P.Y. Assessing potential miRNA targets based on a Markov model. *Genet. Mol. Res.*, **2009**, *8*(3), 848-860.

[26] Eddy, S. R. Computational genomics of noncoding RNA genes, *Cell.*, **2002**, *109*(2), 137-140.

[27] Sam, G.J.; Alex, B.; Mhairi. M.; Ajay K.; Eddy, S.R. Rfam: an RNA family database, *Nucleic Acids Res.*, **2003**, *31*(1), 439-441.

[28] Badger, J.H.; Olsen, G.J. CRITICA: Coding region identification tool invoking comparative analysis. *Mol. Bio. Evol.*, **1999**, *16*(4), 512-524.

[29] Tiziana, C.; Alessandro, C.; Giorgio, G.; Sabino, L.; Flavio, M.;

Graziano, P. CSTminer: a web tool for the identification of coding and noncoding conserved sequence tags through cross-species genome comparison. *Nucleic Acids Res.*, **2004**, 32 (suppl 2), 624-627.

[30] Lotta, C.; Iseli, C.; Jongeneel, C.V.; Bucher, P. Modeling sequencing errors by combining Hidden Markov models. *Bioinformatics.*, **2003**, 19 (suppl 2), 103-112.

[31] Zou, Q.; Lin, C.; Liu, X.Y.; Han, Y.P.; Li, W.B.; Guo. M.Z. Novel representation of RNA secondary structure used to improve prediction algorithms. *Genet. Mol. Res.*, **2011**, *10*(3), 1986-1998.

[32] Liu, J.; Gough, J.; Rost, B. Distinguishing protein-coding from non-coding RNAs through support vector machines. *PLoS. Genet*, **2006**, *2*(4), e29.

[33] Kong, L.; Zhang, Y.; Ye, Z.Q.; Liu, X.Q.; Zhao, S.Q.; Wei L.Q.; Gao G. CPC: assess the protein-coding potential of the transcripts using sequence features and support vector machine. *Nucleic Acids Res.*, **2007**, 35 (suppl 2), 345-349.

[34] Mathews, D.H.; Turner, D.H. Prediction of RNA secondary structure by free energy minimization. *Curr .Opin. Struct .Biol.*, **2006**, *16*(3), 270-278.

[35] Hofacker, I.L. Vienna RNA secondary structure server. *Nucleic Acids Res*, **2003**, *31*(13), 3429-3431.

[36] Zuker, M.; Stiegler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **1981**, *9*(1), 133-148.

[37] Ogurtsov, A.Y.; Shabalina, S.A.; Kondrashov, A.S.; Roytberg, M.A. Analysis of internal loops within the RNA secondary structure in almost quadratic time. *Bioinformatics.*, **2006**, *22*(11), 1317-1324.

[38] Chan, C.Y.; Lawrence, C.E.; Ding, Y. Structure clustering features on the Sfold web server. *Bioinformatics*, **2005**, *21*(20), 3926-3928.

[39] Sverker, E.; Paul P.G.; Anthony M.P.; Nichael D.H.; Dacid P.; Vincent M. A search for H/ACA snoRNAs in yeast using MFE secondary structure prediction. *Bioinformatics.*, **2003**, *19*(7), 865-873.

[40] Rivas, E.; Eddy, R. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics.*, **2000**, *16*(7), 583-605.

[41] Tran, T.T.; Zhou F.; Marshburn, S.; Stead, M.; Kushner, S.R.; Xu Y. De Novo computational prediction of non-coding RNA genes in prokaryotic genomes. *Bioinformatics.*, **2009**, *25*(22), 2897-2905.

[42] Jiang, P.; Wu, H.N.; Wang, W.K.; Ma, W.; Sun, X.; Lu, Z.H. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res.*, **2007**, 35 (suppl 2), W339-344.

[43] Xuan, P.; Guo, M.Z.; Liu, X.Y.; Huang, Y.C.; Li, W.B.; Huang, Y.F. PlantMiRNAPred: efficient classification of real and pseudo plant pre-miRNAs. *Bioinformatics.*, **2011**, *27*(10), 1368-1376.

[44] Marc, R.F.; Wei, C.; Catherine, A.; Jonas, M.; Ralf, E.; Signe K.; Nikolaus R. Discovering microRNAs from deep sequencing data using miRDeep. *Nat .Biotechnol.*, **2008**, *26*(4), 407-415.

[45] Sonika, T.; Candida, V.; Vipin, G.; Rohit, B.; Sachin, M.; Ashwin, S.; Alok, B. CID-miRNA: a web server for prediction of novel miRNA precursors in human genome. *Biochem. Biophys. Res. Commun.*, **2008**, *372*(4), 831-834.

[46] Xu, Y.; Zhou, X.; Zhang, W. MicroRNA prediction with a novel ranking algorithm based on random walks. *Bioinformatics*, **2008**, *24*(13), i50-i58.

[47] Simon M.; Frank S.; Tamas D.; Dan M.L.; David J.S.; Vincent M. A toolkit for analyzing large-scale plant small RNA datasets. *Bioinformatics.*, **2008**, *24*(19), 2252-2253.

[48] Zhu, E.L.; Zhao, F.Q.; Xu, G.; Hou, H.B.; Zhou, L.L.; Li, X.K.; Sun, Z.S.; Wu, J.Y. mirTools: microRNA profiling and discovery based on high-throughput sequencing. *Nucleic Acids Res.*, **2010**, 38 (suppl 2), W392-W397.

[49] Michael, H.; Martin, S.; David, L.; Juan, M.F.P.; Ana, M.A. miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res.*, **2009**, 37 (suppl 2), 68-76.

[50] Chad, J.C.; Jeffrey, G..R.; Preethi, H.G. Expression profliling of microRNAs by deep sequencing. *Briefings in Bioinformatics.*, **2009**, *10*(5), 490-497.

[51] Victor, A.; Bonnie, B.; David, P.B.; Christopher, B.B.; James, C.C.; Chen, X.M.; Gideon D.; Sean, R.E.; Sam, G.J.; Mhairi, M.; Marjori, M.; Gary, R.; Thomas, T. A uniform system for microRNA annotation. *RNA.*, **2003**, *9*(3), 277-279.

[52] Yang, J.H.; Zhang, X.C.; Huang, Z.P.; Zhou, H.; Huang, M.B.; Zhang, S.; Chen, Y.Q.; Qu, L.H. snoSeeker: an advanced computational package for screening of guide and orphan snoRNA genes in the human genome. *Nucl Acids Res*., **2006**, *34*(18), 5112-5123.

[53] Yvan, S.; Inaki, I.; Pedro, L. A review of feature selection techniques in bioinformatics. *Bioinformatics*., **2007**, *23*(19), 2507-2517.

[54] Lu, Z.J.; Kevin, Y.Y.; Wang, G.L.; Chong, S.; Hillier, L.D.W.; Ekta, K.; Ashish, A.; Raymond, A.; Joel, R.; Chao, C.; Masaomi, Kato.; David, M.M.; Frank, S.; Michael, S.; Robert, H.W.; Valerie, R.; Mark, B.G. Prediction and characterization of noncoding RNAs in C.elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data. *Genome Res.*, **2011**, *21*(2), 276-285.

[55] Karen, M.C.; Yue, W.; Rui, L.; Kelli, D.M.; Howard, Y.C.; Robert, B.W. Detection of Long Non-Coding RNA in Archival Tissue: Correlation with Polycomb Protein Expression in Primary and Metastatic Breast Carcinoma. *PLoS. ONE*, **2012**, *7*(10), e47998.

[56] David, M.; Alexander, E.L.; Yuri, I.W.; Svetlana, A.S.; Igor, B.R.; Eugene, V.K. The Vast, Conserved Mammalian lincRNome. *Plo.Comp. Biol.*, **2013**, *9*(2), e1002917.

[57] Thomas, D.; Rory, J.; Giovanni, B.; Andrea, T.; Sarah, D.; Hagen, T.; Gregory, G.; Davidn M.; Angelika, M.; David, G.K.; Julien, L.; Lavanya, V.; Ruan, X.A.; Ruan, Y.J.; Timo, L.; Piero, C.; James, B.B.; Leonard, L.; Jose, M.G.; Mark, T.; Carrie, A.D.; Ramin, S.; Thomas, R.G.; Tim, J.H.; Cedric, N.; Jennifer, H.; Roderic, G. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res.*, **2012**, *22*(9), 1775-1789.

[58] Galina, V.G.; Boris, L.Z.; Igor, B.R.; Computational Prediction of Polycomb-Associated Long Non-Coding RNAs. *PLoS. ONE.*, **2012**, *7*(9), e44878.

[59] Susan, B.; Karen, M.M. Computational Identification and Functional Predictions of Long Noncoding RNA in Zea mays. *PLoS. ONE.*, **2012**, *7*(8), e43047.

[60] Jasmina, P.; Peter, L.O.; Gerton, L.; Chris, P.P. Genomic and Transcriptional Co-Localization of Protein-Coding and Long Non-Coding RNA Pairs in the Developing Brain. *PLoS .Genetics.*, **2009**, *5*(8), e100617.

[61] Li, Z.; Liu, M.; Zhang, L.; Zhang, W.X.; Gao, G.; Zhu, Z.Y.; Wei, L.P.; Fan Q.C.; Long, M.Y. Detection of intergenic non-coding RNAs expressed in the main developmental stages in Drosophila melanogaster. *Nucleic Acids Res.*, **2009**, *37*(13), 4308-4314.