SCIENCE CHINA
Life Sciences

• REVIEWS •

May 2010 Vol.53 No.5: 548–562
doi: 10.1007/s11427-010-0101-9

Computational RNomics: Structure identification and functional prediction of non-coding RNAs *in silico*

ZHENG LingLing & QU LiangHu*

Key Laboratory of Gene Engineering of the Ministry of Education, State Key Laboratory for Biocontrol, Sun Yat-sen University, Guangzhou 510275, China

Received March 20, 2009; accepted June 28, 2009

The eukaryotic genome contains varying numbers of non-coding RNA (ncRNA) genes. "Computational RNomics" takes a multidisciplinary approach, like information science, to resolve the structure and function of ncRNAs. Here, we review the main issues in "Computational RNomics" of data storage and management, ncRNA gene identification and characterization, ncRNA target identification and functional prediction, and we summarize the main methods and current content of "computational RNomics".

non-coding RNA, eukaryote, computational RNomics

Citation: Zheng L L, Qu L H. Computational RNomics: Structure identification and functional prediction of non-coding RNAs *in silico*. Sci China Life Sci, 2010, 53: 548−562, doi: 10.1007/s11427-010-0101-9

The Human Genome Project revealed that less than 2% of the high-mammal genome codes are used for proteins, whereas the other 98% is non-protein coding DNA. The vast majority of regions are not "junk DNA" but contain varying numbers of transcriptional control elements and non-coding RNA (ncRNA) genes. After 4 years of work, the ENCODE project proposed by the National Human Genome Research Institute has completed a pilot project for identifying all of the functional elements in a limited (~1%) region of the human genome, providing convincing evidence that most of the human genome is transcribed and contains a large number of ncRNA genes [1]. In the second phase, the ENCODE project will seek to annotate all types of sequence-based functional elements [2]. The Functional Annotation of the Mammalian Genome Project has started the FANTOM4 project, which focuses on building a regulation network of transcriptional regulation elements and functional RNAs in mammals [3]. There is an "RNA world" hidden in higher organism cells. RNomics has become a new frontier in the post-genomic era [4].

As a key component of "RNomics", which aims at revealing the structures and functions of ncRNAs on a genome-wide scale, "computational RNomics" takes a multidisciplinary approach, comparable to information science, for resolving the structures and functions of ncRNAs, mainly focusing on 3 problems:

(1) ncRNA data storage and management. Numerous ncRNAs and their related molecules have been detected by experimental or computational methods. In particular, deep sequencing technology has resulted in the generation of a massive amount of data. The first step of RNomics research is to effectively manage and store such sequence data. "Computational RNomics" uses databases to accomplish the storage and categorization of data and the mining of internal relationships, thus facilitating usage of the information by researchers.

(2) ncRNA gene identification and characterization. There has been uncertainty regarding the exact number and

*Corresponding author (email: lssqlh@mail.sysu.edu.cn)

categories of ncRNAs within any organism. The classification of these ncRNA genes into various kinds of species and their annotation is one of the main aims of RNomics. "Computational RNomics" develops algorithms for the prediction and annotation of ncRNAs on genome scale.

(3) ncRNA target identification and functional prediction. The final purpose of identifying ncRNA genes is to reveal their function in the cell and to clarify their biological purpose. ncRNAs serve regulatory function by either directly or indirectly interacting with other molecules. "Computational RNomics" develops a series of algorithms to find the target molecules that bind to ncRNAs, predicting their function and establishing the regulation network.

In this paper, we classify current methods and software and summarize the main problems associated with "computational RNomics".

1 Data resources for ncRNA

NcRNA databases are typically divided into 3 categories: general ncRNA databases, specialized ncRNA databases, and ncRNA databases which utilize deep-sequencing results. Some of the frequently used ncRNA databases and their websites are shown in Table 1, from which users can choose an appropriate database according to their own needs.

1.1 General ncRNA database

This kind of database collects all types of ncRNAs, including long ncRNAs and short ncRNAs, and stores them in accordance with certain criteria. Data generally originates from literature mining, experimental results and computer predictions. Such databases include the following:

RNAdb. This database contains information about sequences and the annotation of mammalian ncRNA. The data is from a comprehensive literature review, the Functional Annotation of Mouse (FANTOM) and the H-Invitational databases, the human chromosome 7 annotation project, the miRNA registry, and public sequence databases at NCBI, Ensembl and UCSC [5]. Users search for their sequence of interest in a variety of ways, including casually browsing the collection, searching queries by keywords or applying filters across nominated fields. Users also use BLAST tools to locate regions of similarity between sequences of interest and those stored in the database. Entire datasets and user-specified queries are available for download in FASTA and XML format for local viewing or for conversion into the BED format for directly loading into the UCSC Genome Browser.

NONCODE. This database collects almost all types of ncRNAs, except tRNAs and rRNAs, and includes a novel classification system based on the cellular process and function of a given ncRNA to integrate existing classification systems [6]. The current release of *NONCODE* (v2.0) contains 212527 sequences from 861 organisms, including eukaryotes, eubacteria, archaebacteria, viruses and viroids. ncRNA data are from 3 sources: (i) manual extracts from the relevant literature, (ii) automatically filtered and manually confirmed GenBank sequences, and (iii) biological experimental data from Chen's laboratory [6]. Users browse the database by class of species or ncRNA and search ncRNAs by their GenBank ID or other keywords. It also supports BLAST searches as well as the conversion of data into the BED format for UCSC genome browsing. The entire dataset is available for downloading in FASTA format for local viewing.

Rfam. This database is a collection of RNA sequence families represented by multiple sequence alignments and covariance models (CMs) [7]. Users browse the database by RNA family and search for a sequence of interest against the Rfam model library using BLAST. In addition, files in *Rfam* are downloadable by anonymous ftp (ftp: //selab. janelia.org/pub/Rfam/), and the alignments are in the Stockholm format, which is parsable by the HMMER package.

1.2 Specialized ncRNA database

This kind of database stores specialized types of ncRNAs, e.g. microRNAs or snoRNAs. Other databases that contain

Table 1 Data resources for ncRNA

Database	Website	Feature	Reference
RNAdb	http: //jsm-research.imb.uq.edu.au/rnadb/	Mammalian ncRNA database	[5]
NONCODE	http: //www.noncode.org/	Database contain Eukaryotes, Bacteria, Archea, Viruses data	[6]
Rfam	http: //rfam.sanger.ac.uk/	Collection of RNA families	[7]
fRNAdb	http: //www.ncrna.org/frnadb/	A comprehensive non-coding sequence database	[8]
miRBase	http: //www.mirbase.org/	MicroRNA database	[9,10]
snoRNABase	http: //www-snorna.biotoul.fr/index.php	Human snoRNA database	[11]
microrna.org	http: //www.microrna.org/microrna/home.do	Resource for predicted microRNA targets and expression	[12]
Tarbase	http: //diana.cslab.ece.ntua.gr/tarbase/	Database of experimentally supported microRNA targets	[13]
deepBase	http: //deepbase.sysu.edu.cn/	Database for deeply annotating and mining deep sequencing data	[14]

resources for the target sequences of microRNA have been available, and the microRNA expression data has been integrated, allowing users to look up the expression profile of the ncRNA of interest in different tissues. We briefly describe the principal features of a few databases:

miRbase. A searchable database of both hairpin and mature miRNA sequences that includes annotation as well as information about the miRNA target. The database supports the prediction of miRNA targets across species. Users search and browse by names, keywords, references, and annotations, and all data in the *miRbase* is available for download.

microRNA.org. The database contains resources for microRNA target prediction and expression profiles. Users search and view the expression profile of their microRNA of interest in different tissues and cell lines of normal and disease origin and download the results by heat map or bar graph. Target information predicted by the *miRanda* algorithm is also available for download.

snoRNABase. A special resource for human C/D box and H/ACA modification guide snoRNAs. Users search by the name of the snoRNA, HGNC approved symbol and GenBank accession number for their snoRNA of interest. They also find guide RNAs for rRNAs and snRNAs. The database is linked to the UCSC Genome Browser for users' convenience.

1.3 ncRNA database that applies deep-sequencing results

Deep-sequencing sequencing technology facilitates the sequencing of the transcriptome and the high-throughput generation of massive data sequences. Therefore, it is essential to effectively store and manage this data. Our laboratory has integrated data from current transcriptome sources to develop a database, *deepBase,* with high-throughput and the deep annotation of ncRNAs to facilitate the comprehensive annotation and discovery of small RNAs.

deepBase. Figure 1 shows the basic framework for *deepBase*. Yang *et al*. [14] collected deep sequencing data from 230 small RNA libraries from diverse tissues and cell lines of 7 organisms: human, mouse, chicken, *Ciona intestinalis*, *Drosophila melanogaster*, *Caenhorhabdites elegans* and *Arabidopsis thaliana.* The RNA libraries contain ~14.6 million unique reads which perfectly map onto more than 284 million genomic loci. Many known and novel ncRNAs were identified, and the mapped reads have been grouped into about 1.2 million RNA clusters. The results are downloadable in FASTA and BED format for local viewing and mining. DeepBase provides a user-friendly interface, many convenient mining and search tools and substantial statistical information from the database, including the distribution of read length, the distribution of nucleotides at the 5′ end, and ncRNA expression profiles in different libraries. Users conveniently exchange information with such other websites as UCSC.

2 ncRNA gene finding algorithm

The development of a variety of model organism genome projects provides an unprecedented opportunity for ncRNA discovery and identification at the entire genome scale. This requires individuals to adopt computational methods for developing a suitable algorithm for the large-scale, high-throughput identification of ncRNAs. A challenge for complete genome ncRNA identification is that ncRNAs do not possess such clear characteristics for coding sequences as Open Reading Frames, start and stop codons, or special splice sites. Therefore, conventional gene prediction algorithms are unsuitable. It is necessary to develop new prediction algorithms according to the characteristics of ncRNA genes. "Computational RNomics" has developed a number of ncRNA gene prediction algorithms. According to their characteristics, they are divided into 3 main categories: structure-based prediction, sequence signal-based prediction and deep-sequencing results-based prediction.

2.1 Structure-based prediction

Early prediction algorithms were based on the assumption that the ncRNA sequence folded into a more stable secondary structure than a simple nucleotide. Therefore, sequences that have a stable secondary structure in the genome are considered to be ncRNAs. There are many software packages for RNA secondary structure prediction, e.g. *Mfold* [15,16], *RNAfold* [17], *Sfold* [18–20], which are used in ncRNA identification. However, subsequent studies have found that this feature is not a sufficient criterion for ncRNA prediction because many random sequences also fold into a lower free energy state [21]. A conserved secondary structure was therefore added as a new parameter. A consensus secondary structure is found by multiple sequence alignment. Algorithms that are based on structural features mainly use a combination of the 2 above parameters.

In 2001, Eddy *et al*. [22] developed *QRNA*, which is a comparative sequence analysis algorithm for structural RNA gene identification. Based on covariance variance theory, they built 3 probabilistic "pair-grammars": a pair of stochastic, context-free, grammar modeling alignments constrained by structural RNA evolution, a pair of hidden Markov models that models alignments constrained by coding sequence evolution, and a pair of hidden Markov models modeling a null hypothesis of position-independent evolution [22]. When a pairwise sequence alignment is entered, it is classified by the algorithm as coding, RNA, or null class according to the posterior probability of each class. When using QRNA to test the *Escherichia coli* genome by comparisons to the related genome of *Salmonella*

thyphi, the results indicated that the sensitivity of the program is largely dependent upon the availability of appropriate comparative sequence data [22]. Using multiple comparative genome alignment results will improve the sensitivity.

QRNA has been used to identify ncRNAs in bacteria and yeast but is not suitable for screens of large genomes. This is because of its limited pairwise alignments, such that its reliability is low, especially if the evolutionary distance between the 2 sequences lies outside the optimal range [23]. Several software packages have been developed that also detect a limit to the prediction of ncRNAs. Faced with these problems, Stadler *et al*. [23] developed *RNAz* in 2004, which combines comparative sequence analysis and structure prediction. The algorithm consists of 2 basic components: (i) a measure for RNA secondary structure conservation based on computing a consensus secondary structure, and (ii) a measure for thermodynamic stability by a *z* score that is normalized with respect to both sequence length and base composition but is calculated without sampling from shuffled sequences. *RNAz* analyzes 2 and more sequence alignment results. *RNAz* is used to screen alignments in the conserved noncoding elements in the upstream regions of orthologous genes from human, mouse, rat, *Fugu*, and zebrafish and has found all known ncRNAs and *cis*-acting elements as well as many other novel conserved RNA secondary structures.

With the gradual completion of the genome project, an increasing number of species' genome sequences have been acquired, which has provided more data for the construction of evolutionary relationships to find ncRNA through comparative genomics. The Haussler laboratory developed *Evofold* in 2006, which is based on phylogenetic, stochastic, context-free grammars for identifying functional RNAs encoded in the human genome [24]. An eight-way, genome-wide alignment of the human, chimpanzee, mouse, rat, dog, chicken, zebrafish, and puffer-fish genomes was used as input for deeply conserved functional RNA identification.

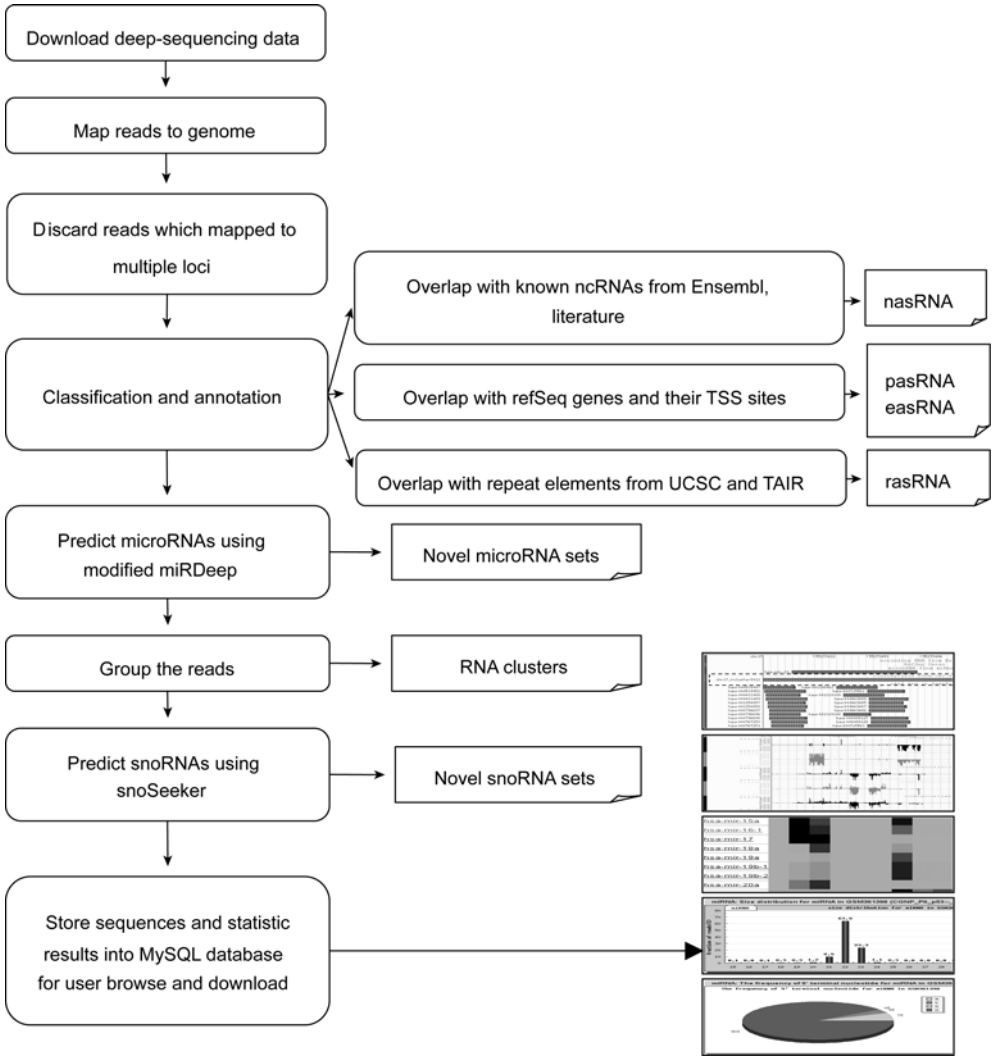

Figure 1 The basic framework for deepBase.

Finally, a large number of known functional RNAs and various types of known genetic recoding elements were identified, and the authors also found new candidate selenocysteine insertion sites, RNA editing hairpins, RNAs involved in transcript auto regulation, and many folds which form singletons or small functional RNA families of completely unknown function [24].

Table 2 lists several structure-based ncRNA prediction software packages. These packages primarily use structural stability and structure conservation across species as parameters. They achieved a stable recognition effect for both long and short ncRNA, with the method serving as a general ncRNA identification technique. However, the predicted results may be elements that are conserved in both mRNA and ncRNA. For instance, the mRNA 5′ and 3′ UTR regions that bind regulatory proteins contain abundant conserved local secondary structures [25]. In addition, this method cannot find ncRNAs that do not possess a conserved secondary structure. The method cannot determine the particular class to which the predicted ncRNA belongs. Therefore, programs based on secondary structure prediction might lead to significant false positive and false negative discoveries [26] and cannot predict particular classes of ncRNA.

2.2 Sequence signal-based prediction method

Structure-based prediction methods are generally used but are not suitable for the identification of specialized ncRNAs. Considering the large number and complexity of ncRNA types, finding a commonality among them is in itself a daunting task. The understanding of ncRNA remains in its infancy. Therefore, programs have been developed for the specialized identification and annotation of ncRNAs using specific categories of known ncRNA sequence analysis to establish models for the identification of novel ncRNAs on the genome scale.

With increased understanding of ncRNA, thousands of small ncRNAs have been identified, including microRNAs, snoRNAs, siRNAs, and piRNAs. In addition, hundreds of thousands of long ncRNAs have been identified through full-length cDNA cloning and genomic tilling array technology. The sequence features of long and short ncRNAs show significant differences and require different approaches for their identification. Here, we describe the methods proposed for ncRNA prediction for both long and short ncRNAs.

2.2.1 Long ncRNA prediction algorithm

A great number of transcript data is available as a result of experimentation (cDNA or EST data), and both protein-coding and non-coding sequences are contained in such data. Protein-coding sequences possess a clear open reading frame (ORF), a certain range of length, and ORFs are conserved between different species. Therefore, sequences are recognized as coding or non-coding according to their characteristics. For example, some programs use ORF length as a criterion. The FANTOM consortium originally used 300 nt (100 codons) as a cutoff, and transcripts that were below this value would be considered non-coding sequences [29]. However, this criterion could pose problems. For example, certain long ncRNAs possess ORF characters. *H19*, *Xist*, *Mirg*, *Gtl2*, and *KcnqOT1* all have putative ORFs>100 codons [30]. By contrast, other short coding sequences are regarded as ncRNAs, e.g. the tal (tarsal-less) gene, which was initially classified as a ncRNA [31]. Therefore, more characteristics are integrated, e.g. the conserved sequence domain, specific receptor/donor site, amino acid composition, and polyadenylation signals. There are many programs which discriminate coding RNAs from ncRNAs, such as *ESTScan* [32], *CSTminer* [33], and *CRITICA* [34]. Mattick *et al*. [35] applied 10 such software packages to the 102801 FANTOM mouse cDNA sequences to discriminate mRNAs from ncRNAs, and after comparison, *CRITICA,* which combined comparative analysis with statistical analysis of coding sequences, showed the highest degree of concordance with the other methods and made the fewest recognizable errors.

Some algorithms use machine-learning methods to distinguish between coding potential and putative ncRNAs. Given confirmed ncRNA and mRNA sequences as input, the method first extracts the characters and uses machine-learning or statistical methods to train and test new

Table 2 Structure-based ncRNA prediction software

Software	Website	Web-based	Downloadable	Reference
Mfold	http: //mfold.bioinfo.rpi.edu/	√	√	[15,16]
RNAfold	http: //rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi	√	√	[17]
QRNA	ftp: //selab.janelia.org/pub/software/qrna/		√	[22]
RNAz	http: //www.tbi.univie.ac.at/~wash/RNAz/	√	√	[23]
Randfold	http: //bioinformatics.psb.ugent.be/software/details/Randfold		√	[27]
RNAstrand	http: //www.bioinf.uni-leipzig.de/Software/RNAstrand/		√	[28]

sequences. Many methods use support vector machines (SVM) to distinguish coding and non-coding RNAs. *CONC* classifies transcripts according to features, including peptide length, amino acid composition, predicted secondary structure content, predicted percentage of exposed residues, compositional entropy, number of homologs from database searches, alignment entropy and nucleotide frequencies [36]. The authors used eukaryotic proteins and known ncRNAs as a true positive and true negative training set to train the SVM model. Using 102801 pieces of mouse cDNA data for prediction, over 14000 ncRNAs were identified. In addition, cross-validation suggested that the specificity and sensitivity of the software were respectively 97% and 98%. The total number of *CONC* input extract features was 180, while a subsequent software package, *CPC,* that was developed by Peking University and was also based on an SVM model but used 6 extract features, achieved a comparable performance with *CONC* in less time [37].

2.2.2 *Short ncRNA prediction algorithm*

The general length of a short ncRNA is no more than 200 nt, including microRNAs, snoRNAs, siRNAs, and piRNAs, and each type has special characteristics. Algorithms used to predict these ncRNAs are mainly based on the following parameters: special length, position in the genome, structure, free energy, sequence pattern, GC content, and sequence and structure homology information.

Searching ncRNAs at the genome scale according to structure produces many results, but many sequences which fold into a stem-loop structure are not true ncRNAs. As real ncRNAs fold into a secondary structure with a lower free energy than randomly generated stem-loop structures and miRNA genes are more conserved in their secondary structure than in their primary structure, the free energy character and conservation is considered. Wang *et al*. [38] from Tsinghua University developed *MiRAlign*, which is based on both sequence and structure alignment, to detect miRNAs in animals. They identified 59 new miRNA genes in the genome of *Anopheles gambiae*. *MiPred* extracted sequence features of the secondary structure of the minimum free energy and used machine-learning methods to distinguish between real and random miRNA precursor stem-loop structures, obtaining 95.09% sensitivity and 98.21% specificity [39]. *MiRFinder* compared 2 genomes and used SVM to conduct genome-wide microRNA precursor prediction, achieving more than a 99% accuracy rate [40].

SnoRNA prediction has proved to be complex, because snoRNA contains 2 types of modifications to the target RNA molecules—box C/D snoRNA and box H/ACA snoRNA. Early snoRNA prediction programs were *Snoscan* [41] and *snoGPS* [42], which mainly searched for guide snoRNA, e.g. the target for rRNAs or snRNAs. Subsequently, Yang *et al*. [43], in our laboratory, have developed *snoSeeker*, which is carried out at the genome scale to search for various types of snoRNAs, including orphan snoRNAs that do not have target molecules. The algorithm also predicts the snoRNA target. There are 2 programs in the *snoSeeker* package, *CDseeker* and *ACAseeker*. The algorithm first removes repeat sequences in the genome. Based on comparative genomics, it identifies homologous sequences as candidates for snoRNA and uses a series of features to build statistical models. Using this method to screen the human genome, 266 known snoRNAs were found, as well as 54 new snoRNAs that include 26 orphan snoRNAs. Another program, *snoReport* [44], used secondary structure prediction information and machine learning methods to predict both types of snoRNAs. The program does not rely on alignment or homology information but searches for new snoRNAs at the genome scale and verifies the snoRNAs predicted by other methods.

2.2.3 *Summary of the sequence signal-based prediction method*

Several software packages based on sequence signals for predicting both long and short ncRNAs are listed in Table 3. It is feasible to use this method to predict special types of ncRNAs, but there are some points which must be ad-

Table 3 Sequence signal-based ncRNA prediction software

Software	Website	Web-based	Downloadable	Reference
ESTScan	http: //www.ch.embnet.org/software/ESTScan2.html	√	√	[32]
CSTminer	http: //t.caspur.it/CSTminer/	√		[33]
CRITICA	http: //rdpwww.life.uiuc.edu/	√	√	[34]
CPC	http: //cpc.cbi.pku.edu.cn/	√	√	[37]
MiRAlign	http: //web.archive.org/web/20080803072157/bioinfo.au.tsinghua.edu.cn/miralign/	√		[38]
Snoscan	http: //lowelab.ucsc.edu/snoscan/	√	√	[41]
snoGPS	http: //lowelab.ucsc.edu/snoGPS/	√	√	[42]
snoSeeker	http: //genelab.zsu.edu.cn/snoseeker/	√	√	[43]
snoReport	http: //www.bioinf.uni-leipzig.de/Software/snoReport/		√	[44]

dressed. First, it is difficult to unequivocally annotate an RNA as protein coding or noncoding [30] because some are bifunctional and serve as a regulatory RNA and are also translated into protein [45]. In addition, one decisive factor in using transcription data is the completeness of the input data, which calls for the full-length status of sequences. However, there is a wide range of experimental variables that make this requirement difficult to attain, such as incomplete reverse transcription, internal priming of pre-mRNAs and genomic DNA contamination [26]. These variables are likely to generate spurious or truncated transcripts that would be difficult to annotate [46]. Second, the homologous sequence analysis method only finds new members of known ncRNAs, not new classes or non-conservative ncRNAs. Due to restrictions on experimental methods, there are certain difficulties inherent in the prediction and verification of some low or tissue-specific expression ncRNAs.

2.3 Deep-sequencing results-based prediction method

Deep-sequencing technology generates thousands of sequence fragments at one time, which are used by various programs to map these fragments against the reference genome for small ncRNA identification. Figure 2 shows the pipeline for microRNA identification using deep-sequencing results [9,10]. Large-scale reads are acquired using deep-sequencing technology. After a preliminary treatment in which these reads are mapped onto the reference genome, the reads that are mapped to multiple loci are filtered. Then, other known RNAs are filtered, e.g. mRNAs, tRNAs, snoRNAs, and snRNAs. According to such characteristics as secondary structure or free energy information to extract possible microRNA precursors, these sequences are next aligned with the known microRNAs to find novel microRNAs. *miRDeep* [47] uses Illumina/Solexa sequencing results according to the locations and the abundance of reads

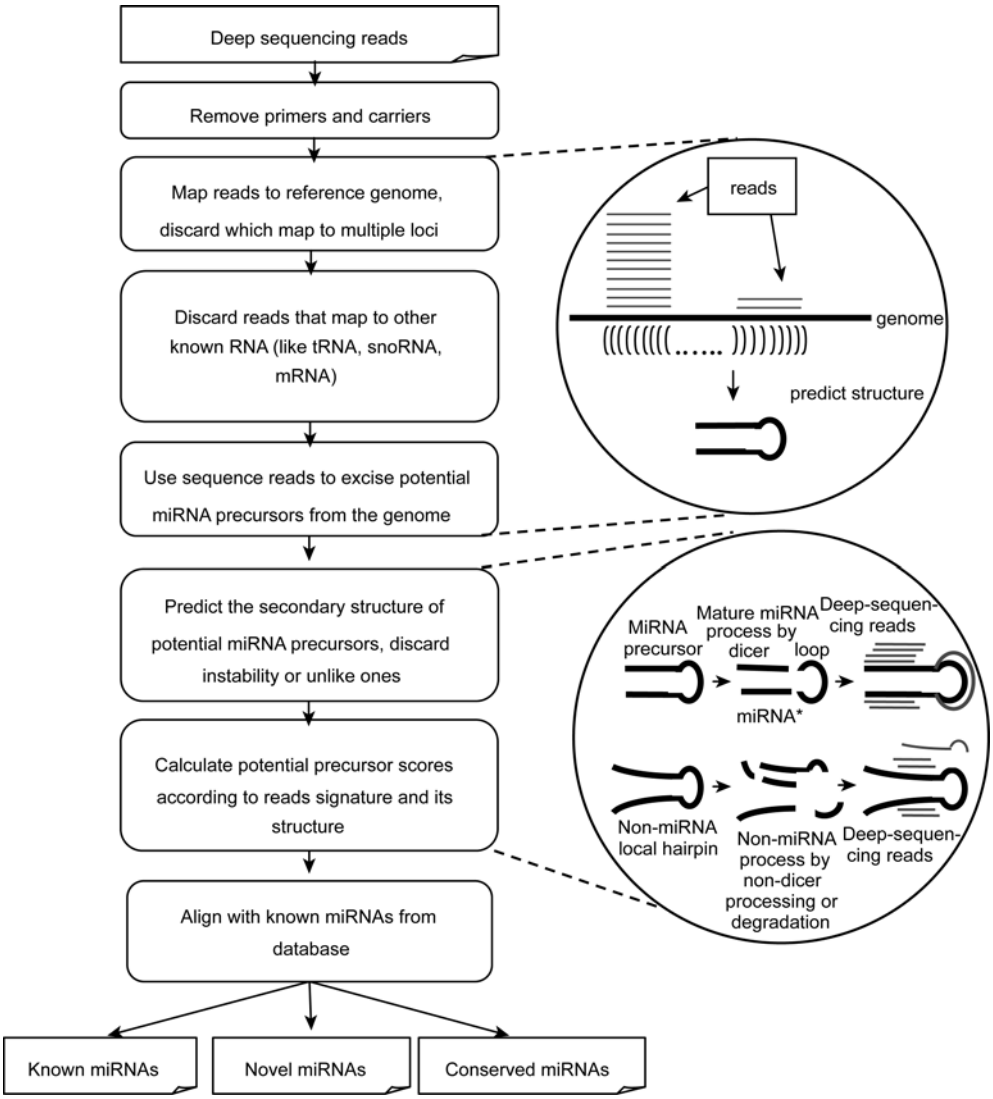

Figure 2 Scheme of the algorithm using deep-sequencing results to detect miRNAs.

that match the microRNA precursor sequence to establish a probabilistic model for novel microRNA prediction. Data from *Caenorhabditis elegans,* human and dog transcriptome sequencing were used for analysis, and ~230 previously unannotated miRNAs were found, some of which were validated by using Northern blot analysis**.**

Researchers on the FANTOM4 project team have used deep-sequencing technology results and have obtained several significant results. Mattin *et al*. [48] analyzed human, chicken and *Drosophila* large-scale sequencing data, and discovered a type of tiny RNAs with an 18 nt conserved sequence pattern in the −60 to +120 region of the gene transcription start site. They called these RNAs transcription initiation RNAs (tiRNAs) and suggested that they may be a general feature of transcription in metazoa and possibly all eukaryotes [48]. Another research result was obtained by Carninci *et al*. [49], which used Cap Analysis Gene Expression (CAGE) data from FANTOM3 to find 6%–30% of the RNA transcription start sites in humans and mice that contain repetitive elements. They identified 23000 candidate regulatory regions derived from retrotransposons, together with more than 2000 examples of bidirectional transcription, and they concluded that retrotransposon transcription has a key influence upon the transcriptional output of the mammalian genome [49].

Several software packages that use deep-sequencing results to predict ncRNAs are listed in Table 4. Deep-sequencing technology provides a reliable data resource for ncRNA prediction. Using this data in combination with the biological characteristics and processing mechanisms of ncRNAs could have an efficient and accurate effect on ncRNA gene identification.

2.4 Summary of the ncRNA gene-finding algorithm

Using ncRNA identification programs, new ncRNA types may eventually be discovered, and their function in the regulation of gene expression will be investigated. While there is no one universal tool that enables the accurate identification of all ncRNAs, the majority of the computer methods have been developed for specific ncRNA prediction. For known ncRNA types, specific methods according to the character of the RNA categories or a combination of several approaches are used to predict a higher sensitivity. However, because a large number of ncRNA classes remain unknown, it is necessary to develop a common universal prediction approach. In addition, with the development of experimental techniques, there will be many new categories added to the ncRNA family. Finding these ncRNA features and establishing appropriate algorithms for genome-wide identification will be a future direction of development. At present, the role of small RNAs and their function has been gradually understood. The development of deep-sequencing technology which provides a reliable transcript analysis resource has enabled considerable advances in the identification of small RNAs. By contrast, long ncRNA research is somewhat slow and needs further exploration. Many tissue- or organism-specific ncRNAs remain to be discovered. Therefore, the development of more accurate and faster recognition algorithms remains a critical direction for future development.

3 ncRNA target prediction algorithm

A major step in ncRNA gene annotation is the identification of molecules with which ncRNA interacts and the clarification of the effects it has on regulating the course of the entire pathway. Many ncRNAs directly interact with other molecules. In addition, the silencing mechanism is determined by the degree of complementarity between the ncRNA and its target molecule. When nearly perfect complementarity is achieved, e.g. plant siRNAs or miRNAs, target molecules are cleaved. However, most animal miRNAs are only partly complementary with their target molecules, and this mainly represses mRNA translation. The interaction between the miRNA and its target also depends on the location and local secondary structure of the target molecule. Therefore, the discovery of ncRNA target molecules is the basis and a key step for revealing the ncRNA function.

microRNAs, due to their critical biological functions and target specific combination mechanisms, have become the main object for target prediction algorithms. microRNA biological functions include the control of cell proliferation, cell death, lipid metabolism, and the control of leaf and flower development in plants. It is predicted that more than 30% of human genes are regulated by microRNAs. microRNAs regulate gene expression primarily by binding to the 3′ UTRs of their target mRNAs for either cleavage or

Table 4 Deep-sequencing, results-based prediction software

Software	Website	Web-based	Downloadable	Reference
snoSeeker	http: //deepbase.sysu.edu.cn/	√		[14]
miRDeep	http: //www.mdc-berlin.de/en/research/research_teams/systems_biology_of_gene_regulatory_elements/projects/miRDeep/		√	[47]
MiRanalyzer	http: //web.bioinformatics.cicbiogune.es/microRNA/	√		[50]

translational repression, thereby affecting cell proliferation, apoptosis, differentiation, metabolism, growth, metastasis and other biological processes. Therefore, it is an essential step to find the miRNA target in revealing the microRNA function. However, due to the diversity of interaction styles and characteristics of the target secondary structure, identification is problematic. The binding mechanisms in plants and animals are quite different. Plant microRNA prediction is rather straightforward because plant microRNAs usually bind their targets with nearly perfect complementarity. Consequently, most researchers have focused their attention on the problem of animal microRNA target identification. Dozens of prediction software packages are available and sum up many interaction rules. Based on the different principles of target identity, the algorithms are divided into 3 categories: generalized principle-based strategies, breaking generalized rules strategies, and integrative principle strategies.

3.1 Generalized principle-based strategies

Most microRNA target identification programs are based on 3 basic principles: (i) the base complementarity between the miRNA 5′ seed region with the target site in the mRNA 3′ UTR, (ii) the thermodynamic stability of the predicted miRNA-mRNA duplex, and (iii) the degree of conservation of the target sites of related species. Figure 3 shows the basic process of microRNA target identification using generalized principles.

In 2003, Enright *et al*. [51] developed a microRNA target prediction algorithm, *miRanda*, which is based chiefly on the 3 general principles: sequence complementarity, the thermo-dynamic stability of RNA-RNA duplexes, and the conservation of target sites in related genomes. They used 73 available *D. melanogaster* miRNAs as probes and scanned the 3′ UTRs of 9805 *D. melanogaster* genes for possible complementarity matches, eventually finding 535

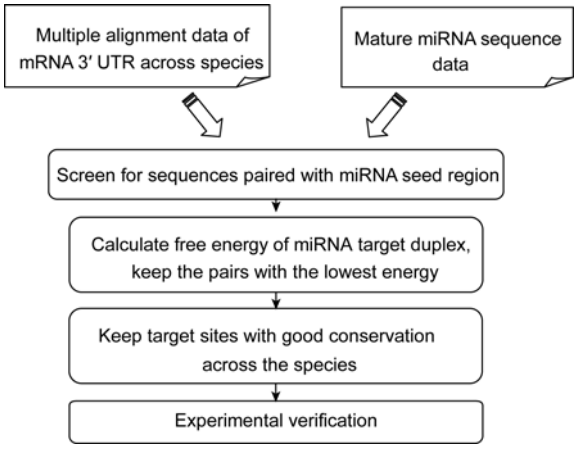

Figure 3 The basic process of microRNA target identification using general principals.

potential target genes. Due to the strong conservation of miRNA sequences between *D. melanogaster* and *A. gambiae* [52], the same procedure was used for a subset of targets that were also conserved in the 3′ UTRs of *A. Gambia* genes, and 150 potential targets were found, among which 40% exhibited more than 60% conservation identity to the corresponding *D. melanogaster* site. In addition, they constructed randomized microRNA data to perform control calculations and used a simple estimate of the rate of false positives (R^-/R^+), where R^- represents the number of targets predicted for randomized miRNAs, and R^+ is the number of targets predicted for the actual miRNAs, to obtain a 35% false positive rate. They further analyzed the predicted target molecules according to their biological process and molecular function derived from the Gene Ontology [53] database to elucidate the miRNA function, and they concluded that miRNAs target the control of gene activity at multiple levels, specifically transcription, translation and protein degradation [51].

Lewis *et al*. [54] developed the *TargetScan* program to predict vertebrate miRNA targets. They provide not only an accurate calculation for false-positives but also experimental evidence for the predicted target. The approach also considers complements of the RNA duplex and conservation of target sites between different species. In the calculation of the process of matching complements, they suggested that the 5′ portion of the miRNA, particularly nucleotides 2–8, appears to be the most critical sites for target recognition, and they referred to this segment as the "miRNA seed" [54]. Therefore, they initially sought for segments of mRNA 3′ UTRs that were ideally matched with the 7-nucleotide length "seed" region and extended each match with additional complementary base pairs as far as possible in each direction until mismatches were encountered. For the remaining 3′ portion of the miRNA, they optimized the base-pairing of the 35 bases of the UTR immediately preceding the 5′ of each seed match and used the RNAfold program to extend each seed match to a longer "target site". They searched the 3′ UTRs of orthologous human, mouse, and rat data derived from the Ensembl database and found 451 putative miRNA: target interactions, with an average of 5.7 targets per miRNA. They used this data as their signal, while another group of sequences, which were predicted through randomly permuted miRNAs, were used as noise. On average, every shuffled miRNA had 1.8 targets. Therefore, the signal: noise ratio was 3.2:1 (5.7/1.8), for a false-positive rate of about 31% (1.8/5.7). In addition, 11 of the 15 predicted targets were experimentally validated. They analyzed the molecular function by gene ontology (GO) and found that the predicted mammalian miRNA target genes were significantly involved in the "transcription" and "regulation of transcription" GO process categories.

In 2005, Lewis *et al*. [55] refined the *TargetScan* algorithm and developed *TargetScanS*. The algorithm differs by starting with whole-genome alignments, thereby require-

ing the conserved seed matches to be at conserved positions within the UTRs [55]. The definition of a seed region also changed. Previously defined as the 7-nt segment of the miRNA 5′ portion, the improved algorithm emphasizes the pairing to a 6-nt miRNA seed flanked by either an m8 match or a t1A anchor. They used the algorithm to predict 4 vertebrate genome 3′ UTRs, discovered about 13000 interactions and concluded that more than 30% of human genes are regulated by microRNAs.

Krek *et al*. [56] developed the *PicTar* program, which uses probability methods to predict common targets of microRNAs by calculating the probabilities for mRNA sequences to function as a binding site for microRNAs. The software is divided into 2 parts; the first part identifies single microRNA target sites given a single microRNA and calculates the probabilities for an mRNA sequence to be a target site. The second part, for a given RNA sequence (typically a 3′ UTR), calculates the maximum likelihood score that a fixed set of microRNAs binds the sequence. The *PicTar* algorithm follows the conclusion by Doench *et al*. [57] that microRNAs competitively bind to the target gene. Using genome-wide alignments of 8 vertebrate genomes to test the algorithm, the result specifically recovered known microRNA targets. The false-positive rate of the algorithm has been estimated to be around 30%. In addition, Krek *et al*. suggested that vertebrate microRNAs each target roughly 200 transcripts.

Researchers used *PicTar* in multiple species for miRNA target identification. In 2005, Grün *et al*. [58] used this program to exploit cross-species comparisons to show that there are generally 54 targeted genes per microRNA above noise in *Drosophila melanogaster*. In 2006, Lall *et al*. [59] used a new version of *PicTar* to predict microRNA targets in *C. elegans*. Using sequence alignments of 3 nematodes, they predicted that miRNAs regulate at least 10% of *C. elegans* genes, and a number of nematode miRNAs seem to regulate biological processes by targeting functionally related genes. Chen *et al*. [60] used the program to analyze microRNA targets in the human genome and showed that negative selection in humans is stronger for computationally predicted conserved miRNA binding sites than for other conserved sequence motifs in 3′ UTRs. For nonconserved miRNA binding sites, they estimated that 30%–50% of these are functional when the mRNA and miRNA are endogenously coexpressed, which indicates that polymorphisms at predicted miRNA binding sites are likely to be deleterious, and are thus candidates for causal variants of human disease.

Table 5 lists several software packages using the general principle-based identification strategy, together with their supported organisms. The algorithms using general principles reliably predict, but some researchers think that these features are not always sufficient for miRNA target prediction determination. Andrew *et al*. [61] combined computational and experimental approaches to suggest 5 other characteristics of the site context that boost site efficacy: (i) AU-rich nucleotide composition near the site, (ii) proximity to sites for coexpressed miRNAs, (iii) proximity to residues pairing to miRNA nucleotides 13–16, (iv) positioning within the 3′ UTR at least 15 nt from the stop codon, and (v) positioning away from the center of long UTRs [61]. Therefore, many programs have made some improvements and breakthroughs concerning the general principles.

3.2 Breaking general rules strategies

Although most target miRNA binding mechanisms follow the 3 generalized principles, many other features also are factors in miRNA target gene identification. For example, the number of target sites, AU content near target sites, and expression levels of miRNA and mRNA are used as certain standards for miRNA target determination. Many researchers found that there are some exceptions to these generalized principles and conclude that target selection mechanisms vary from species to species [65], creating a number of new criteria. These criteria serve, in addition to the general principles, with some possibly superseding the conventional principles.

3.2.1 3′ UTR rule

Most prediction algorithms follow a general rule that animal miRNAs target the 3′ UTR region of mRNAs. Therefore, many researchers have searched in this region and have found a large number of real targets. However, it has been

Table 5 Software packages using general principal-based identification strategy

Software	Supported organism (s)	Website	Web-based	Downloadable	Reference
miRanda	Vertebrate	http: //www.microRNA.org/	√	√	[51]
TargetScan *TargetScanS*	Vertebrate	http: //www.targetscan.org/	√		[54] [55]
PicTar	Vertebrate, fly, worm	http: //pictar.mdc-berlin.de/	√		[56]
DIANA-microT	Mammals	http: //diana.cslab.ece.ntua.gr/microT/	√		[62]
RNAhybrid	Mammals	http: //bibiserv.techfak.uni-bielefeld.de/rnahybrid/	√	√	[63]
miTarget2	Human, mouse, rat, chicken, dog	http: //mirdb.org/miRDB/		√	[64]

reported that miRNAs can bind promoter regions [66], coding regions [67] and 5′ UTR regions of mRNA [68].

Forman *et al*. [69] developed a computational algorithm to find conserved sequence motifs within coding regions and analyzed a multiple-sequence alignment of 17 genomes. They first searched for conserved motifs of 8-bp length in coding regions and assigned the motifs a Sequence Level Conservation Score (SLCS) that represents the degree of the motif at the sequence level over the amino acid level. By matching these SLCS motifs against known miRNAs, they found that high-scoring sequence motifs matched the miRNA target sites. They experimentally demonstrated that the let-7 miRNA directly targets the miRNA-processing enzyme Dicer within its coding sequence, thus establishing a mechanism for an miRNA/Dicer autoregulatory negative feedback loop [69]. They used *RNAcofold* to computationally fold the secondary structure of let-7 and its seed matches in the coding regions, 3′ UTRs and 5′ UTRs. Through a comparison, they revealed that the results for the 3′ UTR show a preference for looping at base pairs 10–12, followed by a preference for binding at positions 13–16, whereas the results for the coding regions showed a statistically significant preference for binding in positions 13–15 and 17–19 and no preference for looping, suggesting different mechanisms of miRNA targeting in coding regions and 3′ UTRs.

3.2.2 "Seed" region rule

The rule of seed pairing is a fundamental criterion used in miRNA target prediction. In an early study, the complete sequence complementary to the 6-nt sequence of the miRNA 5′ region was found to be significant for target recognition [70]. Through mutation analysis using human, *Drosophila*, and *Arabidopsis thaliana* cells, it was concluded that mismatches could occur [57,71,72]. The next study further improved the rules. The 2–7 position of the miRNA 5′ end was considered to be the "seed" region, and the matching extent with this region determined the recognition of the target molecule [55]. Most miRNA target prediction algorithms follow the "seed" region rules and identify the actual target molecules. In addition, high-throughput transcript and proteomics research has shown that in miRNA overexpression or inactivation of human cells, a large number of genes are enriched in the regions that match the "seed" sequences [73–75]. Although numerous studies have shown that the "seed" region is a vital factor in miRNA target identification, other characteristics also serve to specify target molecules. For example, earlier studies found that let-7 miRNA has many non-seed binding target sites on the lin-41 and lin-14 genes [18,76]. Therefore, the seed match rule is not always sufficient for repression.

Zhang *et al*. [77], using vascular endothelial growth factor (VEGF) as a model, examined the effect of the central loop region in the miRNA: MRE duplex in miRNA- mediated gene regulation. In the 2.0 version of their prediction program, *FindTar*, the central loop law was involved such that 3 elements were considered: loop position, loop size and loop priority. The results showed that central loops in miRNA: MRE duplexes affect efficiency in miRNA target identification. They suggested the addition of a loop score that combines both the location and the size as new criteria for predicting MREs. Their cognate miRNAs significantly decreased the false positive rates and increased the specificity of MRE prediction.

3.2.3 Target site conservation rule

The conservation of target sites between species is an essential rule that is a factor in miRNA target prediction algorithms. However, researchers have demonstrated that many of the nonconserved target sites, which outnumber the conserved sites 10 to 1, also mediate repression [78]. Using this rule, nonconserved target sites may be missed. Therefore, a number of algorithms have been developed that do not rely on the conservation of target sites and that provide satisfactory functions.

In 2006, Miranda *et al*. [79] changed the ideas of their predecessors. They developed a microRNA target gene prediction software package, *RNA22* which is a pattern-based approach that does not rely upon cross-species conservation and is therefore utilized for the analysis of any single genome. Another unique feature of the algorithm is that it does not proceed from the known microRNAs to the search for their binding targets but starts from the sequence of interest, determines whether it might be a microRNA target, and finally determines the microRNAs which regulate it [80]. They used several species, and their experimental results challenge many of the previous conclusions about microRNA target prediction. They suggested that the number of microRNA precursors in mammalian genomes may possibly number in the tens of thousands. They also suggested that in addition to 3′ UTRs, numerous binding sites may exist in the 5′ UTRs and CDSs of mRNA, which has since been proved to be the case [69,81]. The software also predicted microRNA precursors, and the results suggested that the number of precursors is more than that currently estimated. Several software packages that use the principle strategies to predict microRNA targets are listed in Table 6.

3.3 Integrative principle strategies

More features have been integrated into microRNA target prediction algorithms, including the secondary structure of target molecules, the expression levels of miRNAs and target genes, and other information about the target molecule, e.g. gene function, the role of the pathway, and

Table 6 Software packages that use breaking conventional principle strategies to predict microRNA targets

Software	Supported organism (s)	Website	Web-based	Downloadable	Reference
FindTar	Human, mouse, rat, chimpanzee, cow	http: //bio.sz.tsinghua.edu.cn/findtar/	√		[77]
RNA22	Mammals	http: //cbcsrv.watson.ibm.com/rna22.html	√	√	[79]
microTar	Worm, fly, mouse	http: //tiger.dbs.nus.edu.sg/microtar/		√	[82]

associated diseases.

The generalized principles have been introduced for the analysis of secondary structure by predicting the structure of the miRNA-mRNA duplex, calculating their free energy, and choosing the more stable structure. However, the actual situation is more complicated than expected. Studies have shown that although a given molecule might constitute a good match and have a stable secondary structure with the miRNA, because the sequence itself may constitute a stem loop structure, the regulation by miRNA may be abrogated [83]. Therefore, the secondary structure upstream and downstream of the binding site region is used as one of the factors considered by an miRNA target prediction algorithm.

Long *et al*. [84] analyzed the effect of the target secondary structure on the efficacy of repression by miRNAs. They used *Sfold* software to predict the mRNA secondary structure and to establish a two-step hybridization reaction model of the interaction between an miRNA and a target. They used the model to analyze the interaction of let-7 with various mutant forms of the *Caenorhabditis elegans* lin-41 3' UTR region and also tested miRNA-target interactions in *C. elegans* and *Drosophila melanogaster*. The results suggested a fundamental effect of target structure on target recognition and established a structure-based framework for the genome-wide identification of animal miRNA targets [84].

PITA software, developed by Michael *et al*. [83] provides a parameter-free model for microRNA-target interaction by computing the difference between the free energy gained from the formation of the microRNA-target duplex and the energetic cost of unpairing the target to make it accessible to the microRNA [83]. For each given microRNA, the authors calculated the target score to represent its ability to be combined with the predicted target molecules. The experimental results revealed that the mutation of target sites would reduce target accessibility, which is a critical factor in microRNA function.

Several software packages that use integrative principle strategies to predict miRNA targets are listed in Table 7. The ultimate aim of the target prediction was to reveal the function of microRNAs. Therefore, new methods for the integration of more information about target molecules, including their expression profile, functional categories and the pathways in which they are involved are needed.

3.4 Summary of ncRNA target prediction

Determining the target molecules for ncRNAs is crucial for revealing their functions and the specific associated mechanism. Although a wide range of ncRNAs have been found in various species, few functions have been fully elucidated, indicating the disconnection between gene identification and functional prediction. At present, microRNA target prediction is the focus of much research. Most micro RNA target prediction algorithms combine the complementarity between the miRNA 5′ seed region and the target site in the mRNA 3′ UTR, the thermodynamic stability of the miRNA- mRNA duplex, and the conservation of target sites across species. Although these characteristics are essential for the combination of the miRNAs and their target molecules, the relationship between these features and how they affect the miRNA and target combination remains unclear [26]. Other features are considered, with the searching scale expanded to include other regions, such as the coding regions or the 5′ UTRs of mRNA. The miRNA target prediction algorithm has a high specificity and sensitivity for known miRNAs. However, various programs do not yield the same prediction for unknown target genes, and the false-positive rate is high [86]. Achieving consistent predictions requires a more in-depth study of the interaction mechanism of miRNAs and target molecules, establishing more credible parameters and thus improving the prediction accuracy rate. With the development of new experimental techniques, the validated miRNA target data may continue to accumulate, providing a workable training set for recognition algorithms, thereby continually improving and optimizing prediction algorithms.

4 Opportunities and challenges

Such next generation sequencing technology as Roche/454, Illumina/Solexa, and ABI/SOLiD, substantial amounts of sequencing data will be added to databases, which may provide a driving force for ncRNA research in the post- genomic era as well as new opportunities and challenges.

The development of new experimental techniques has expanded the field of research in "computational RNomics". In addition to model organisms, a large number of non-model organisms are utilized for genetic and developmental studies. They are also employed to design various studies

Table 7 Software packages that use the Combination Principle Strategies to predict miRNA targets

Software	Supported organism(s)	Website	Web-based	Downloadable	Reference
PITA	Fly, worm, mouse, human	http: //genie.weizmann.ac.il/pubs/mir07/mir07_prediction.html	√	√	[83]
GenMir++	Human	www.psi.toronto.edu/genmir		√	[85]

for the transcriptome of normal cells or tissue samples with pathological analysis, revealing the genetic characteristics or pathological mechanism with which they are associated. Comparative genomics and comparative transcriptomics studies will be at different levels and on a larger scale to reveal the function of ncRNAs and the genetic information regulated by them in the biological evolution process. In addition, the results of the different sequencing technologies combined with genomics and proteomics data may serve to establish an overall regulatory network. Ultimately, integrating the full range of computational methods to solve complex biological problems is the goal of "computational RNomics".

Genomic and transcriptome sequence data are exponentially expanding, posing new requirements for computer hardware (such as the continued expansion of computing and greater storage capacity). How to effectively enhance the capacity of the database interaction, reduce costs for maintenance, and improve the transmission efficiency and protection of information security represent major challenges for "computational RNomics".

RNomics constitutes a major frontier in the post-genomics era. The combination of "computational RNomics" and "experimental RNomics" may facilitate the exploration of ncRNAs to be more extensive and in-depth, thereby revealing the full expression of genetic information and its regulatory network in life and the evolution of the "RNA World".

The authors would like to thank Liao JianYou, Guan Daogang, Chen HuaCheng, and Li JunHao for their valuable assistance in collecting relative information. We are also grateful to Dr. Yang JianHua, Dr. Shao Peng for their constructive suggestions. This work was supported by the National Natural Science Foundation of China (*Grant Nos. 30830066 and 30771151*), *and the National Basic Research Program* (*Grant No. 2005CB724600*).

1 Birney E, Stamatoyannopoulos J A, Dutta A, *et al*. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature, 2007, 447: 799–816
2 Rosenbloom K R, Dreszer T R, Pheasant M, *et al*. ENCODE whole-genome data in the UCSC Genome Browser. Nucl Acids Res, 2010, 38: D620–625
3 Maeda N, Kasukawa T, Oyama R, *et al*. Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs. PLoS Genet, 2006, 2: e62
4 Qu L H. RNomics: The new frontier in the post-genomic era. Sci China C Life Sci, 2009, 52: 193–194
5 Pang K C, Stephen S, Engstrom P G, *et al*. RNAdb--a comprehensive mammalian noncoding RNA database. Nucl Acids Res, 2005, 33: D125–130
6 Liu C, Bai B, Skogerbo G, *et al*. NONCODE: An integrated knowledge database of non-coding RNAs. Nucl Acids Res, 2005, 33: D112–115
7 Gardner P P, Daub J, Tate J G, *et al*. Rfam: Updates to the RNA families database. Nucl Acids Res, 2009, 37: D136–140
8 Kin T, Yamada K, Terai G, *et al*. fRNAdb: A platform for mining/annotating functional RNA candidates from non-coding RNA sequences. Nucl Acids Res, 2007, 35: D145–148
9 Griffiths-Jones S, Saini H K, van Dongen S, *et al*. miRBase: Tools for microRNA genomics. Nucl Acids Res, 2008, 36: D154–158
10 Griffiths-Jones S, Grocock R J, van Dongen S, et al. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res, 2006, 34: D140–144
11 Lestrade L, Weber M J. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. Nucleic Acids Res, 2006, 34: D158–162
12 Betel D, Wilson M, Gabow A, *et al*. The microRNA.org resource: Targets and expression. Nucleic Acids Res, 2008, 36: D149–153
13 Papadopoulos G L, Reczko M, Simossis V A, *et al*. The database of experimentally supported targets: A functional update of TarBase. Nucleic Acids Res, 2009, 37: D155–158
14 Yang J, Shao P, Zhou H, *et al*. deepBase: A database for deeply annotating and mining deep sequencing data. Nucleic Acids Res, 2010, 38: D123–130
15 Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res, 2003, 31: 3406–3415
16 Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucl Acids Res, 1981, 9: 133–148
17 Hofacker I L. Vienna RNA secondary structure server. Nucleic Acids Res, 2003, 31: 3429–3431
18 Chan C Y, Lawrence C E, Ding Y. Structure clustering features on the Sfold web server. Bioinformatics, 2005, 21: 3926–3928
19 Ding Y, Chan C Y, Lawrence C E. Sfold web server for statistical folding and rational design of nucleic acids. Nucl Acids Res, 2004, 32: W135–141
20 Ding Y, Lawrence C E. A statistical sampling algorithm for RNA secondary structure prediction. Nucl Acids Res, 2003, 31: 7280–7301
21 Rivas E, Eddy S R. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. Bioinformatics, 2000, 16: 583–605
22 Rivas E, Eddy S R. Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics, 2001, 2: 8
23 Washietl S, Hofacker I L, Stadler P F. Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci USA, 2005, 102: 2454–2459
24 Pedersen J S, Bejerano G, Siepel A, *et al*. Identification and classification of conserved RNA secondary structures in the human genome. PLoS Comput Biol, 2006, 2: e33
25 Meyer I M, Miklos I. Statistical evidence for conserved, local secondary structure in the coding regions of eukaryotic mRNAs and pre-mRNAs. Nucl Acids Res, 2005, 33: 6338–6348
26 Solda G, Makunin I V, Sezerman O U, *et al*. An Ariadne's thread to the identification and annotation of noncoding RNAs in eukaryotes. Brief Bioinform, 2009, 10: 475–489
27 Bonnet E, Wuyts J, Rouze P, *et al*. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. Bioinformatics, 2004, 20: 2911– 2917

28 Reiche K, Stadler P. RNAstrand: Reading direction of structured RNAs in multiple sequence alignments. Algori Mol Biol, 2007, 2: 6
29 Frith M C, Forrest A R, Nourbakhsh E, *et al*. The abundance of short proteins in the mammalian proteome. PLoS Genet, 2006, 2: e52
30 Dinger M E, Pang K C, Mercer T R, *et al*. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. PLoS Comput Biol, 2008, 4: e1000176
31 Tupy J L, Bailey A M, Dailey G, *et al*. Identification of putative noncoding polyadenylated transcripts in *Drosophila melanogaster*. Proc Natl Acad Sci USA, 2005, 102: 5495–5500
32 Lottaz C, Iseli C, Jongeneel C V, *et al*. Modeling sequencing errors by combining Hidden Markov models. Bioinformatics, 2003, 19: ii103–112
33 Castrignanò T, Canali A, Grillo G, *et al*. CSTminer: A web tool for the identification of coding and noncoding conserved sequence tags through cross-species genome comparison. Nucleic Acids Res, 2004, 32: W624–627
34 Badger J H, Olsen G J. CRITICA: Coding region identification tool invoking comparative analysis. Mol Biol Evol, 1999, 16: 512–524
35 Frith M C, Bailey T L, Kasukawa T, *et al*. Discrimination of non-protein-coding transcripts from protein-coding mRNA. RNA Biol, 2006, 3: 40–48
36 Liu J, Gough J, Rost B. Distinguishing protein-coding from non-coding RNAs through support vector machines. PLoS Genet, 2006, 2: e29
37 Kong L, Zhang Y, Ye Z, *et al*. CPC: Assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res, 2007, 35: W345–349
38 Wang X, Zhang J, Li F, *et al*. microRNA identification based on sequence and structure alignment. Bioinformatics, 2005, 21: 3610–3614
39 Jiang P, Wu H, Wang W, *et al*. MiPred: Classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic Acids Res, 2007, 35: W339–344
40 Huang T, Fan B, Rothschild M F, *et al*. MiRFinder: An improved approach and software implementation for genome-wide fast microRNA precursor scans. BMC Bioinformatics, 2007, 8: 341
41 Lowe T M, Eddy S R. A computational screen for methylation guide snoRNAs in yeast. Science, 1999, 283: 1168–1171
42 Schattner P, Barberan-Soler S, Lowe T M. A computational screen for mammalian pseudouridylation guide H/ACA RNAs. RNA, 2006, 12: 15–25
43 Yang J, Zhang X, Huang Z, *et al*. snoSeeker: An advanced computational package for screening of guide and orphan snoRNA genes in the human genome. Nucl Acids Res, 2006, 34: 5112–5123
44 Hertel J, Hofacker I L, Stadler P F. SnoReport: Computational identification of snoRNAs with unknown targets. Bioinformatics, 2008, 24: 158–164
45 Lanz R B, McKenna N J, Onate S A, *et al*. A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex. Cell, 1999, 97: 17–27
46 Ravasi T, Suzuki H, Pang K C, *et al*. Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. Genome Res, 2006, 16: 11–19
47 Friedlander M R, Chen W, Adamidi C, *et al*. Discovering microRNAs from deep sequencing data using miRDeep. Nat Biotech, 2008, 26: 407–415
48 Taft R J, Glazov E A, Cloonan N, *et al*. Tiny RNAs associated with transcription start sites in animals. Nat Genet, 2009, 41: 572–578
49 Faulkner G J, Kimura Y, Daub C O, *et al*. The regulated retrotransposon transcriptome of mammalian cells. Nat Genet, 2009, 41: 563–571
50 Hackenberg M, Sturm M, Langenberger D, *et al*. miRanalyzer: A microRNA detection and analysis tool for next-generation sequencing experiments. Nucleic Acids Res, 2009, 37: W68–76
51 Enright A J, John B, Gaul U, *et al*. microRNA targets in *Drosophila*. Genome Biol, 2003, 5: R1
52 Aravin A A, Lagos-Quintana M, Yalcin A, *et al*. The small RNA profile during *Drosophila melanogaster* development. Dev Cell, 2003, 5: 337–350
53 Ashburner M, Ball C A, Blake J A, *et al*. Gene ontology: Tool for the unification of biology. Nat Genet, 2000, 25: 25–29
54 Lewis B P, Shih I, Jones-Rhoades M W, *et al*. Prediction of mammalian microRNA targets. Cell, 2003, 115: 787–798
55 Lewis B P, Burge C B, Bartel D P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell, 2005, 120: 15–20
56 Krek A, Grun D, Poy M N, *et al*. Combinatorial microRNA target predictions. Nat Genet, 2005, 37: 495–500
57 Doench J G, Sharp P A. Specificity of microRNA target selection in translational repression. Genes Dev, 2004, 18: 504–511
58 Grün D, Wang Y, Langenberger D, *et al*. microRNA target predictions across seven *Drosophila* species and comparison to mammalian targets. PLoS Comput Biol, 2005, 1: e13
59 Lall S, Grün D, Krek A, *et al*. A genome-wide map of conserved microRNA targets in *C. elegans*. Curr Biol, 2006, 16: 460–471
60 Chen K, Rajewsky N. Natural selection on human microRNA binding sites inferred from SNP data. Nat Genet, 2006, 38: 1452–1456
61 Grimson A, Farh K K, Johnston W K, *et al*. microRNA targeting specificity in mammals: Determinants beyond seed pairing. Mol Cell, 2007, 27: 91–105
62 Maragkakis M, Reczko M, Simossis V A, *et al*. DIANA-microT web server: Elucidating microRNA functions through target prediction. Nucl Acids Res, 2009, 37: W273–276
63 Rehmsmeier M, Steffen P, Höchsmann M, *et al*. Fast and effective prediction of microRNA/target duplexes. RNA, 2004, 10: 1507–1517
64 Wang X, El Naqa I M. Prediction of both conserved and nonconserved microRNA targets in animals. Bioinformatics, 2008, 24: 325–332
65 Watanabe Y, Tomita M, Kanai A. Computational methods for microRNA target prediction. Meth Enzymol 2007, 427: 65–86
66 Place R F, Li L, Pookot D, *et al*. microRNA-373 induces expression of genes with complementary promoter sequences. Proc Natl Acad Sci USA, 2008, 105: 1608–1613
67 Duursma A M, Kedde M, Schrier M, *et al*. miR-148 targets human DNMT3b protein coding region. RNA, 2008, 14: 872–877
68 Lee I, Ajay S S, Yook J I, *et al*. New class of microRNA targets containing simultaneous 5′-UTR and 3′-UTR interaction sites. Genome Res, 2009, 19: 1175–1183
69 Forman J J, Legesse-Miller A, Coller H A. A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence. Proc Natl Acad Sci USA, 2008, 105: 14879–14884
70 Lai E C. Micro RNAs are complementary to 3′ UTR sequence motifs that mediate negative post-transcriptional regulation. Nat Genet, 2002, 30: 363–364
71 Mallory A C, Reinhart B J, Jones-Rhoades M W, *et al*. microRNA control of PHABULOSA in leaf development: importance of pairing to the microRNA 5' region. EMBO J, 2004, 23: 3356–3364
72 Brennecke J, Stark A, Russell R B, *et al*. Principles of microRNA-target recognition. PLoS Biol, 2005, 3: e85
73 Lim L P, Lau N C, Garrett-Engele P, *et al*. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature, 2005, 433: 769–773
74 Selbach M, Schwanhäusser B, Thierfelder N, *et al*. Widespread changes in protein synthesis induced by microRNAs. Nature, 2008, 455: 58–63
75 Baek D, Villén J, Shin C, *et al*. The impact of microRNAs on protein output. Nature, 2008, 455: 64–71
76 Ha I, Wightman B, Ruvkun G. A bulged lin-4/lin-14 RNA duplex is sufficient for *Caenorhabditis elegans* lin-14 temporal gradient forma-

tion. Genes Dev, 1996, 10: 3041–3050
77 Ye W, Lv Q, Wong C A, *et al*. The effect of central loops in miRNA: MRE duplexes on the efficiency of miRNA-mediated gene regulation. PLoS ONE, 2008, 3: e1719
78 Farh K K, Grimson A, Jan C, *et al*. The widespread impact of mammalian microRNAs on mRNA repression and evolution. Science, 2005, 310: 1817–1821
79 Miranda K C, Huynh T, Tay Y, *et al*. A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. Cell, 2006, 126: 1203–1217
80 Xia W, Cao G, Shao N. Progress in miRNA target prediction and identification. Sci China C Life Sci, 2009, 52: 1123–1130
81 Tay Y, Zhang J, Thomson A M, *et al*. microRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. Nature, 2008, 455: 1124–1128
82 Thadani R, Tammi M T. MicroTar: Predicting microRNA targets from RNA duplexes. BMC Bioinformatics, 2006, 7: S20
83 Kertesz M, Iovino N, Unnerstall U, *et al*. The role of site accessibility in microRNA target recognition. Nat Genet, 2007, 39: 1278–1284
84 Long D, Lee R, Williams P, *et al*. Potent effect of target structure on microRNA function. Nat Struct Mol Biol, 2007, 14: 287–294
85 Huang J C, Babak T, Corson T W, *et al*. Using expression profiling data to identify human microRNA targets. Nat Meth, 2007, 4: 1045–1049
86 Xu X. Same computational analysis, different miRNA target predictions. Nat Methods, 2007, 4: 191