IDENTIFICATION OF MICRORNA PRECURSORS VIA SVM

LIANG HUAI YANG,[1,2] WYNNE HSU,[1] MONG LI LEE,[1] LIMSOON WONG[1]

[1]*School of Computing, National University of Singapore*
{*yanglh, whsu, leeml, wongls*}*@comp.nus.edu.sg*

[2]*School of Electronics Engineering and Computer Science,*
Peking University, P.R. China

MiRNAs are short non-coding RNAs that regulate gene expression. While the first miRNAs were discovered using experimental methods, experimental miRNA identification remains technically challenging and incomplete. This calls for the development of computational approaches to complement experimental approaches to miRNA gene identification. We propose in this paper a *de novo* miRNA precursor prediction method. This method follows the "feature generation, feature selection, and feature integration" paradigm of constructing recognition models for genomics sequences. We generate and identified features based on information in both primary sequence and secondary structure, and use these features to construct SVM-based models for the recognition of miRNA precursors. Experimental results show that our method is effective, and can achieve good sensitivity and specificity.

1. Introduction

Traditionally, the "Central Dogma" has decreed that genetic information flows linearly from DNA to RNA to protein, and never in reverse. The role of RNA in the cell has been limited to its function as mRNA, tRNA, and rRNA. The discovery of a diverse array of transcripts that are not translated to proteins but rather function as RNAs has changed this view profoundly. Now, it is increasingly hard to have a comprehensive understanding of cellular processes without considering functional RNAs. Efficient identification of functional RNAs—non-coding RNAs (ncRNAs) as well as cis-acting elements—in genomic sequences is, therefore, one of the major goals of current bioinformatics.

1.1. *Background*

MicroRNAs (miRNAs) are the smallest functional non-coding RNAs of animals and plants. They have been called "the biological equivalent of dark matter, all around us but almost escaping without detection." The mature miRNAs are synthesized from a longer precursor (pre-miRNA) forming a long hairpin structure that contains the mature miRNA in either of its arms. All reported mature miRNAs are between 17 and 29 nucleotides (nt) in length and the majority of them are about 21-25 nt long and have been found in a wide range of eukaryotes, from Arabidopsis thaliana and Caenorhabditis elegans to mouse and human.[3] MicroRNAs play an important regulatory functions in eukaryotic gene expression through mRNA degradation or translation inhibition. The regulatory functions of miRNAs range

from cell proliferation, fat metabolism, neuronal patterning in nematodes, neurological diseases, modulation of hematopoietic lineage differentiation in mammals, development, cell death, cancer, and control of leaf and flower development in plants. An miRNA downregulates the translation of target mRNAs through base-pairing to these target mRNAs.[16,1] In animals, miRNAs tend to bind to the 3' untranslated region (3' UTR) of their target transcripts to repress translation. The pairing between miRNAs and their target mRNAs usually includes short bulges and/or mismatches. In contrast, in all known cases, plant miRNAs bind to the protein-coding region of their target mRNAs with three or fewer mismatches and induce target mRNA degradation[10] or repress mRNA translation.

1.2. *Related Works*

The experimental identification of miRNA is technically challenging and incomplete for two reasons. First, miRNAs tend to have highly constrained tissue- and time-specific expression patterns. Second, degradation products from mRNAs and other endogenous noncoding RNAs coexist with miRNAs and are sometimes dominant in small RNA molecule samples extracted from cells.

MicroRNAs and their associated proteins appear to be one of the more abundant ribonucleoprotein complexes in the cell. A single organism may have hundreds of distinct miRNAs, some of which are expressed in stage-, tissue- or cell type-specific patterns. Nonetheless, miRNAs whose expression is restricted to nonabundant cell types or specific environmental conditions could still be missed in cloning efforts. Thus, computational methods have been developed to complement experimental approaches to identify miRNA genes.

Many miRNAs have been predicted through various computational screens, such as comparative genomics, that can detect entirely new RNA families.[13,12] To date, over 1600 miRNAs have been identified in different organisms.[6]A variety of computational methods have been applied to several animal genomes, including Drosophila melanogaster, C. elegans and humans.[5,12,13] They use the following strategies:

(1) Homology searches for orthologs and paralogs of known miRNA genes. This strategy exploits the observation that some miRNAs are conserved across great evolutionary distances which indicates that their sequence is not arbitrary. Such sequence conservation in the mature miRNA and long hairpin structures in miRNA precursors facilitates genome-wide computational searches for miRNAs.
(2) Searching for a genomic cluster[15] in the vicinity of known miRNA genes. This strategy is important because some of the most rapidly evolving miRNA genes are present as tandem arrays within operon-like clusters, and the divergent sequences of these genes make them relatively difficult to spot if general approaches are used.
(3) Gene-finding approaches that do not depend on homology or proximity to known genes have also been developed and applied to entire genomes.[5,13,12,19] They typically start by identifying conserved genomic segments that both fall outside of predicted protein-coding regions and potentially could form stem loops and then scoring these candidate miRNA stem loops for the patterns of conservation and pairing that characterize known miRNAs genes.

MiRscan[12,15] and SRNALoop[5] have been systematically applied to nematode and vertebrate candidates, and miRseeker[13] has been systematically applied to insect candidates. Wang et al.[19] applied their method to plants. Dozens of new genes have been identified that were subsequently (or concurrently) experimentally verified. Other methods like profile-based detection of miRNA precursors[11] have also been proposed. In addition, several groups have developed computational methods to predict miRNA targets in Arabidopsis, Drosophila and humans.

1.3. *Paper Organization*

Notwithstanding its progress, *de novo* prediction is still a largely unsolved issue. Here, we follow the "feature generation, feature selection, feature integration" paradigm[14] of constructing recognition models for genomic sequences to develop a *de novo* method based on SVM for recognition of miRNA precursors. The paper is organized as follows: Section 2 details our methodology which includes the input data and feature generation. The data generation and experimental results are presented in Section 3 to demonstrate the effectiveness of our method and we conclude in Section 4.

2. Proposed Methodology

To predict new miRNAs by computational methods, we need to define sequence and structure properties that differentiate known miRNA sequences from random genomic sequence, and use these properties as constraints to screen intergenic regions/whole genome (introns excluding those protein encoding exons) in the target genome sequences for candidate miRNAs. Unlike protein coding genes, ncRNAs lack in their primary sequence common statistical signals that could be exploited for reliable detection algorithms. For miRNAs, different methods need to be contrived.

2.1. *Signals Used*

Computational gene-finding for protein-coding genes in both prokaryotic and eukaryotic genomes has been quite successful. These methods exploit genomic features such as long open-reading-frames and codon signatures. Many signal sensors have been designed to detect signals like splice sites, start and stop codons, branch points, promoters and terminators of transcription, polyadenylation sites, ribosomal binding sites, topoisomerase II binding sites, topoisomerase I cleavage sites, and various transcription factor binding sites and CpG islands.

However, it is not so easy for noncoding RNA (ncRNA) genes like miRNA. Usually only weakly-conserved promoter and terminator signals (and possibly other poorly known transcription binding sites) are present in ncRNA genes.[2] EST searches indicate that some human and mouse miRNAs are co-transcribed along with their upstream and downstream neighboring genes.[17] A recent study shows that microRNA genes are transcribed by RNA polymerase II.[9] This leads us to exploit some possible signals that might exist in the up-

stream and downstream of miRNA precursors. We distinguish the possible transcription of miRNA into two categories:

(1) Co-transcribed miRNAs: miRNAs located in the introns of annotated host genes. For this case, miRNAs share the same $\pm$1000 up/downstream of the host genes.
(2) Independently transcribed miRNAs: These miRNAs are not far away from the annotated genes. We further divide them into two categories: (a) clustered miRNAs: we use the -1000 upstream of the first miRNA precursor in the cluster and the +1000 downstream of the last miRNA precursor in the cluster; (b) non-clustered miRNAs: we use the $\pm$1000 up/downstream of the miRNAs precursor.

For the secondary category, it is observed that a prominent characteristic of animal miRNAs is that their genes are often organized in tandem, and are closely clustered on the genome.

Again the situation with miRNAs is more challenging. Far fewer miRNAs are available in the databases. MicroRNA sequences can be compared only at the nucleotide level—not as translated amino acids and miRNA sequences are quite short. As noted previously, the mature miRNA has only about 17-25nts and its precursor has about 100nts for animals. Consequently, distinguishing weakly conserved genes from random "hits" is more difficult when searching for miRNAs than for protein-coding genes. Moreover, even in cases where there are large RNA families, sequence conservation is often at the secondary-structure level, i.e., what is conserved are base pairing rather than the individual base sequence. Consequently, sequence alignment alone may fail to identify miRNAs that diverged too far apart in their primary sequence while retaining their base-paired structure.

To capture the information of secondary structure, we first fold the miRNA precursor using the Vienna RNA package RNAFold.[8] Next, to facilitate data processing, we encode the base-pairing by: A:U–"1", C:G–"2", G:C–"3", G:U–"4", U:A–"5", U:G–"6", Other-"0". An example cel-mir-1 miRNA precursor of C. elegans is shown in Figure 1. We ignore the loop part and mismatch starting part because of their large variations and low conservations.

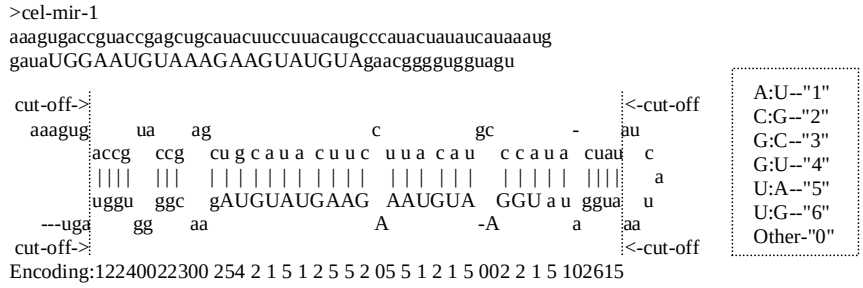

Figure 1. Encoding the secondary structure.

Figure 2 shows the conceptual view of one input sequence. Each input consists of four components: upstream sequence, the primary sequence of miRNA precursor, the encoding

sequence of the secondary structure of miRNA precursor, and the downstream sequence. Thus, the input contains the information of both primary sequence and secondary structure.

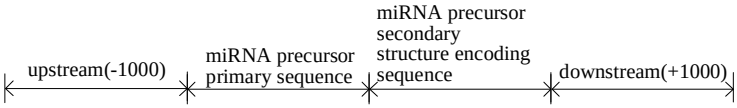

Figure 2. A conceptual view of input sequence.

2.2. *Feature Generation and Feature Selection*

To enable machine learning algorithms to learn from known miRNA sequences, we need to map the input sequence into a feature vector in the feature space. In this work, we follow the "feature generation, feature selection, feature integration".[14] In the "feature generation" process, we exploit the widely used so-called k-$gram$[14] frequency in our feature mapping.

Let Σ denote an alphabet, whose length is $|\Sigma| = L$. Let X be a sequence of letters from Σ. Given $1 \leq k < L$, a k-$gram$ is a k-length contiguous subsequence. We define our feature map as an indexed vector by all possible subsequences α of length-k from Σ^k. Formally, the feature map Φ_k: $X \rightarrow \mathbb{R}^{L^k}$ is defined as:

$$\Phi_k(X)=(\phi_\alpha(X))_{\alpha\in\Sigma^k}$$

where $\phi_\alpha(X)$ is the frequency count of α that occurs in X.

For our input data, the upstream, the primary sequence of the precursor and the downstream have the same alphabet $\sum$={A, C, G, U}. Given k=6, each sequence is coded into a vector with $\sum_{k=1}^{6} 4^k = 1364$ elements. The encoding sequence of the secondary structure of the precursor has an alphabet {1, 2, 3, 4, 5, 6}. We ignore the mismatch code "0". Let $k = 5$, the latter sequence is coded into a vector with $\sum_{k=1}^{5} 6^k = 1554$ elements. Hence, an input sequence will be mapped into a feature vector which will have a total of $3 * 1364 + 1554 = 5646$ elements. We use a suffix tree to accelerate the generation of features. Each depth-k node of the suffix tree stores a count of the number of leaf nodes it leads to.

The feature dimensionality is very large even for a small k. Most learning algorithms suffer from the "curse of dimensionality"— these methods typically require an exponential increase in the number of training samples with respect to an increase in the dimensionality of the samples in order to uncover and learn the relationship of the various dimensions to the nature of the samples. Hence, the selection of relevant informative features among the large collection of candidate features is necessary for machine learning tasks faced with high dimensional data. In the "feature selection" process, we use a correlation-based feature selection method based on the concept of entropy.[20]

2.3. *Support Vector Machines*

Support Vector Machines (SVMs) are a class of supervised learning algorithms first introduced by Vapnik.[18] Given a set of labelled training vectors (positive and negative input examples), an SVM learns a linear decision boundary to discriminate between the two classes. The result is a linear classification rule that can be used to classify new test examples. SVMs have exhibited excellent generalization performance (accuracy on test sets) in practice and have strong theoretical motivation in statistical learning theory.

In our application, we integrate the features selected previously into a model for classifying a candidate sequence as a miRNA precursor or as "other". This "feature integration" process is a typical application for SVMs.

3. Experiments: Classification of miRNA Precursors

In this section, we first describe how to generate the required data set for training and testing. Then we show the prediction result of the trained SVM.

3.1. *Data Generation*

All miRNA genes and precursors (Version 6; April2005) are downloaded from the microRNA Registry[6] which has 1650 precursors. Genome sequences for Caenorhabditis elegans and Caenorhabditis briggsae are available from WormBase at `ftp://ftp.wormbase.org`. Drosophila melanogaster and Drosophila pseudoobscura genome release 4.1 are obtained from FlyBase at `ftp://flybase.net/genomes`. Genomes and the corresponding annotation files of Homo sapiens, Mus musculus, Rattus norvegicus, and Gallus gallus are acquired from Ensembl at `http://www.ensembl.org/Download/`.

3.1.1. *Generating Positive Examples*

Animal miRNAs are often closely clustered together. We call two miRNAs on the same strand as "adjacent" if the number of nucleotides between the end of one miRNA and the start of the other is less than 1000 nts. If miRNAs mr_1, mr_2, ..., mr_k satisfy $(mr_{i+1}.\text{start} - mr_i.\text{end}) < 1000$nts for $i = 1$, ..., $k-1$, we say they form a miRNA cluster. The procedure of generating positive examples is as follows:

(1) For each species considered, we merge the adjacent miRNAs in the same strand to form clusters;
(2) According to the GFF annotations,
   - For miRNAs located in the introns of CDS, we obtain the –1000 upstream and +1000 downstream of the CDS, along with the miRNA precursor to form one input sequence;
   - For each independently transcribed miRNA, we extract the $\pm$1000 upstream/downstream of the miRNA or miRNA cluster, along with the miRNA precursor to form one input sequence;

3.1.2. *Generating Negative Examples*

It is an inherently difficult problem in bioinformatics to get negative examples. However, knowing that only a very very small fraction of non-annotated sequences correspond to "coding" sequences for miRNAs, we can generate negative examples of miRNA genes from inter-genic regions for learning. We make this assumption realizing that our negative examples might be somewhat contaminated with currently unknown miRNA genes. Hence, to alleviate the problem, we filter the negative examples in an iterative manner after making the initial predictions, i.e., we remove strongly predicted genes and re-train in order to purify our training examples.

Since all the miRNA precursors form a stem-loop secondary structure and each arm of the stem may contain the miRNA, we also require these negative examples to be as similar as possible to the true miRNA precursors. Otherwise, it will be trivial for the learning algorithm to detect these fake outliers. Specifically, when generating the negative examples, two conditions must be satisfied. First, they form a stem-loop. We use RNAfold[8] for folding the selected sequence using the C-libraries of the Vienna RNA package version 1.4. Second, the matching part of the stem is at least 15 nt long (currently the smallest miRNA is 17nt).

The procedure of generating negatives is as follows:

(1) With the help of GFF annotation file, we sort each CDS of the same strand according to its (start, end) position, and form the inter-genic regions;
(2) For each inter-genic region, we slide along the sequence and use a normal distribution $N(\mu, \sigma)$ to simulate the length of the precursor, where the μ, σ are estimated from the known miRNA precursors of the species in question. For instance, $\mu = 98$, $\sigma = 6.3$ for C. elegans.

During the generation of a sequence for stem-loops of a certain length, we may find two or more stem-loops on the same strand that has a large percentage of overlap. To avoid excessive overlap, when sliding along the intergenic region, we make a hop of about 50 nt by using a normal distribution $N(50, 20)$ with a large variation.

3.2. *Experimental Results*

We obtain a binary classification SVM on training sets by using the support vector machine library LIBSVM.[4] The input data for the SVM are scaled to $[-1, 1]$. We choose a radial basis function (RBF) kernel. All the experiments were performed in a PC with 1G RAM.

We present the results of three sets of experiments: training the SVM with one of three species D. melanogaster (dme), C. briggsae (cbr), and Mus musculus (mmu) separately and then use the resulting SVMs to predict other species. Due to memory restriction, we are not able to include a large number of negatives in the training set for feature selection. In the experiments, we only include 4000 negatives for feature selection.

Note that the choice of the negatives is an art since different combinations of negatives can lead to different selected feature sets. Hence, we test different combinations and keep those with good testing performance. For example, one data set may consists of 220 mmu

Table 1. Characteristics of training data for feature selection.

species	# of positives	# of negatives	# of features by CFS	# of features by CBFS
mmu	220	4000mmu	177	72
dme	78	4000dme	95	55
cbr	82	2000cbr+2000cel	134	55

Table 2. Experimental Results

Trained species	Test Species	Sensitivity(TP/(TP+FN)%)	Specificity(TN/(TN+FP)%)
dme(120dme150dps,39, 2, 2^{-1})	dps	62/(62+10)=86%	39666/(39666+2996)=93%
cbr(80cbr0cel, 44, 32, 2^{-9})	cel	88/(88+27)=76.52%	76661/(76661+3418)=95.73%
mmu(600mmu150hsa, 62, 8, 2^{-3})	rno	172/(172+13)=92.97%	77370/(77370+4842)=94.11%
mmu(0mmu350hsa, 62, 512, 2^{-7})	hsa	258/(258+63)=80.37%	69792/(69792+6518)=91.46%
mmu(600mmu450hsa, 62,32,2^{-3})	gga	110/(110+12)=90.16%	75069/(75069+4338)=94.54%
mmu(600mmu450hsa, 62, 32, 2^{-3})	ptr	57/(57+10)=85.08%	75203/(75203+3451)=95.61%

positives and 4000 mmu negatives; another data set may consists of 220 mmu positives, 2000 mmu negatives and 2000 hsa negatives. We also use the recursive feature selection method, i.e., we first obtain a feature set from a data set and then form a new data set by projecting the original data against this feature set. This method can put more instances into consideration. However, this method does not necessarily lead to better performance since the feature selection in the first step may be biased. In our experiments, we try two feature selection methods: CFS[7] and CBFS[20] for each combination. In general, the selected features are different for different data sets. The prominent property for all these feature sets is that they primarily consist of features from the encoded secondary structure. Some simple combinations of the negatives for feature selection are listed in Table 1.

Given one species, our purpose is to see if we can find a model to predict the miRNA precursors of another species. For this reason, during the training stage, we only use the positives of one species for training and hold out all the positives of the other species for testing. However, we use some negatives of the target species randomly chosen by assuming that most of the intergenic regions do not contain miRNA precursors. For the first experiment, we use all the known positives (78) of D. melanogaster (dme) and 4000 negatives to perform feature selection using CBFS. Among the 55 selected features, we choose the top 39 to train SVM models—we refer to it as dmeSVM. Among these models, we choose the one with larger area under the ROC curve (AUC). In general, we can get many models with equal AUC. Here, we report the model with 120 dme negatives and 150 dps negatives which has a sensitivity of 86% and a specificity of 93%. We optimize the parameters γ=0.5 and C=2. The prediction results of a species for its related species are given in Table 2. In the first column, the selected model is presented as *species(negative data combination, number of features used, C value, γ value)*.

To see the relationship between these miRNA precursors $\mathcal{T}$ in the training set and the miRNA precursors $\mathcal{P}$ to be predicted in the testing set, we implemented a Needleman-Wunsch-based similarity computing algorithm with match score $= 1$, mismatch $= -1$, and gap penalty $= 1$, and the similarity is computed by the ratio of identities over the whole alignment length denoted as $sim(s, \mathcal{P})$, where $s \in \mathcal{T}$ and $sim(s, \mathcal{P}) =$

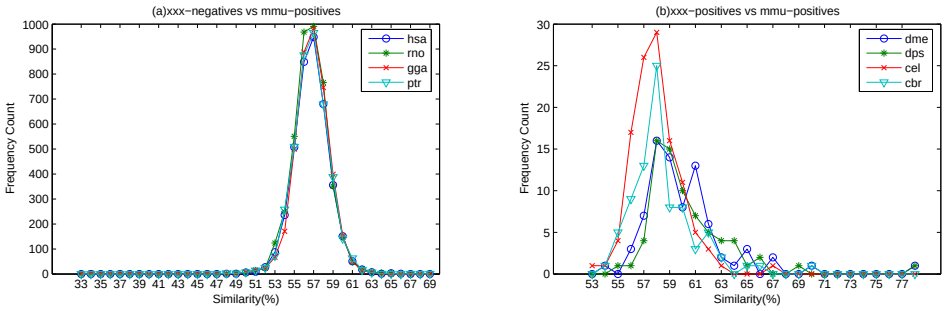

Figure 3. Similarity Histograms against the mmu Positives.

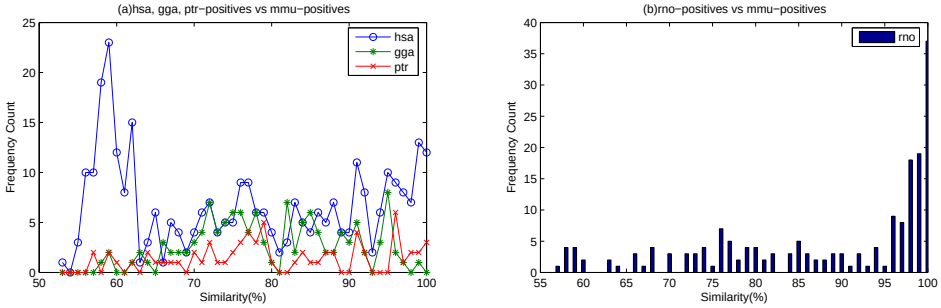

Figure 4. Similarity Histograms of hsa, gga and rno Positives against the mmu Positives.

$max\{sim(s,p) \mid p \in \mathcal{P}\}$. By sampling the negatives with a rate $\frac{1}{20}$ and taking the whole set of positives, we build the histograms of similarities to mmu of both negatives and positives of species other than mmu (Fig. 3). For the histogram of negatives of all species vs mmu positives (Fig. 3(a)), we know that the distribution is an approximate normal with their center around 56-58. This trend is also observed in the histogram of other positives of remote species against mmu-positives Fig. 3(b) which centers around 56-59. Only the later has a little bit longer tail. We show the similarity histogram of its related species in Fig. 4. The comparisons between other species are similar. For human(hsa), there are about 102 miRNA precursors with similarity around $53-62$ %. Based on these observations, we can see that SVM's performance is not solely dependent on the primary sequence similarity in some sense. This point is reflected in the selected features.

We also check some false positives by looking at their conservations in their related species. We find that some false positives reach 88% identity in conservation. This indicates that the false positive may be a true positive.

4. Conclusion

In this work, we have described a SVM-based method to predict miRNA precursors. Based on the current number of candidates generated, the method performs well for related

species. Future research directions include examining the selected features for biological explanations, investigate the performance for predicting unrelated species, and locating mature miRNA in its precursor.

References

1. J. E. Abrahante, A. L. Daul, M. Li, M. L. Volk, J. M. Tennessen, E. A. Miller, A .E. Rougvie. The Caenorhabditis elegans hunchback-like gene lin-57/hbl-1 controls developmental time and is regulated by microRNAs. Dev Cell 2003, 4:625-637.
2. J. Barciszewski, V. A. Erdmann(Ed.). Noncoding RNAs: Molecular Biology and Molecular Medicine. Kluwer Academic, 2004. Ch3 33-48: P. Schattner. Computational Gene-Finding for Noncoding RNAs.
3. D. P. Bartel. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell, 116:281-297, 2004.
4. C. C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines (2001).
5. Y. Grad, J. Aach, G. D. Hayes, B. J. Reinhart, G. M. Church, G. Ruvkun, J. Kim. Computational and experimental identification of C. elegans microRNAs. Mol Cell 2003, 11:1253-1263.
6. S. Griffiths-Jones. The microRNA Registry. Nucleic Acids Res, 32 Database issue:D109-11, 2004.
7. M. A. Hall. Correlation-based feature selection for machine learning. PhD thesis, The University of Waikato, 1999.
8. I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. Monatshefte fir Chemie, 125:167-188, 1994.
9. Y. Lee, M. Kim, J. Han, K. H. Yeom, S. Lee, S. H. Baek, V. N. Kim. MicroRNA genes are transcribed by RNA polymerase II. EMBO J. 23(20):4051-60, Epub 2004.
10. C. Llave, K. D. Kasschau, M. A. Rector, J. C. Carrington. Endogenous and silencing-associated small RNAs in plants. Plant Cell 2002, 14:1605-1619.
11. M. Legendre, A. Lambert, D. Gautheret. Profile-based detection of microRNA precursors in animal genomes. Bioinformatics 21:841-845, 2005.
12. L. P. Lim, N. C. Lau , E. G. Weinstein, A. Abdelhakim, S. Yekta, M. W. Rhoades, C. B. Burge, D. P. Bartel. The microRNAs of Caenorhabditis elegans. Genes Dev 2003, 17:991-1008.
13. E. C. Lai, P. Tomancak, R. W. Williams, G. M. Rubin. Computational identification of Drosophila microRNA genes. Genome Biol, 4:R42, 2003.
14. H. Liu and L. Wong. Data mining tools for biological sequences. J Bioinform Comput Biol., 1(1):139-67, 2003.
15. U. Ohler, S. Yekta, L. P. Lim, D. P. Bartel, C. B. Burge. Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. RNA, 10(9):1309-22,2004.
16. A. E. Pasquinelli, G. Ruvkun. Control of developmental timing by microRNAs and their targets. Annu Rev Cell Dev Biol 2002, 18:495-513.
17. N. R. Smalheiser. EST analyses predict the existence of a population of chimeric microRNA precursor-mRNA transcripts expressed in normal human and mouse tissues. Genome Biol 2003, 4:403.
18. V.N. Vapnik. Statistical Learning Theory. Springer, 1998.
19. X. J. Wang, J. L. Reyes, N. H. Chua, T. Gaasterland. Prediction and identification of Arabidopsis thaliana microRNAs and their mRNA targets. Genome Biol. 5(9):R65. Epub 2004.
20. L. Yu, H. Liu. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. ICML, pp:856-863,2003.