OMICS A Journal of Integrative Biology
Volume 17, Number 9, 2013
© Mary Ann Liebert, Inc.
DOI: 10.1089/omi.2013.0011

Effective Classification of MicroRNA Precursors Using Feature Mining and AdaBoost Algorithms

Ling Zhong,[1] Jason T. L. Wang,[1] Dongrong Wen,[1] Virginie Aris,[2]
Patricia Soteropoulos,[2] and Bruce A. Shapiro[3]

Abstract

MicroRNAs play important roles in most biological processes, including cell proliferation, tissue differentiation, and embryonic development, among others. They originate from precursor transcripts (pre-miRNAs), which contain phylogenetically conserved stem–loop structures. An important bioinformatics problem is to distinguish the pre-miRNAs from pseudo pre-miRNAs that have similar stem–loop structures. We present here a novel method for tackling this bioinformatics problem. Our method, named MirID, accepts an RNA sequence as input, and classifies the RNA sequence either as positive (i.e., a real pre-miRNA) or as negative (i.e., a pseudo pre-miRNA). MirID employs a feature mining algorithm for finding combinations of features suitable for building pre-miRNA classification models. These models are implemented using support vector machines, which are combined to construct a classifier ensemble. The accuracy of the classifier ensemble is further enhanced by the utilization of an AdaBoost algorithm. When compared with two closely related tools on twelve species analyzed with these tools, MirID outperforms the existing tools on the majority of the twelve species. MirID was also tested on nine additional species, and the results showed high accuracies on the nine species. The MirID web server is fully operational and freely accessible at http://bioinformatics.njit.edu/MirID/. Potential applications of this software in genomics and medicine are also discussed.

Introduction

MicroRNAs play a very important role in the transcriptional and post-transcriptional regulation of genes affecting protein levels (Bartel, 2004; Bindra et al., 2010). Lee et al. (1993) first reported that in *Caenorhabditis elegans*, lin-4 regulates the translation of lin-14 mRNA via an antisense RNA–RNA interaction. Since then, many functions of microRNAs have been discovered (Aukerman and Sakai, 2003; Brennecke et al., 2003; Bushati and Cohen, 2007; Johnston and Hobert, 2003; Mack, 2007). MicroRNAs (miRNAs) can have multiple mRNA targets as they bind to the targets with partial complementarities in animals. In addition, the mRNA targets can be regulated by multiple miRNAs. MicroRNAs are likely involved in regulation of all biological processes, and are also found circulating in blood (Mitchell et al., 2008; Schöler et al., 2011). Their expression has been correlated with the expression of oncogenes in cancer cells (Sampson et al., 2007; Zhu et al., 2008), cancer risk factors (Wang et al., 2007), and drug metabolism (Gomez and Ingelman-Sundberg, 2009; Pan et al., 2009; Takagi et al., 2008; Tsuchiya et al., 2006). They hold a great potential for pharmacogenomics applications, such as the tailoring of drugs to the specific cancers and monitoring the response to, and toxicity of, the drugs in individual patients.

MicroRNAs originate from precursor transcripts (pre-miRNAs), which contain phylogenetically conserved stem–loop structures (Lai et al., 2003). An important bioinformatics problem is to distinguish the pre-miRNAs from pseudo pre-miRNAs that have similar stem–loop structures. A common approach to tackling this bioinformatics problem is to transform the classification of real and pseudo pre-miRNAs to a feature selection problem.

Lim et al. (2003) reported some characteristic features in phylogenetically conserved stem–loop pre-miRNAs. Lai et al. (2003) considered hairpin structures predicted by mfold (Zuker, 2003) as well as the nucleotide divergence of pre-miRNAs. Xue et al. (2005) decomposed stem–loop hairpin structures into local structure–sequence features, and used these features in combination with a support vector machine to classify pre-miRNAs. Bentwich et al. (2005) proposed a

[1] Bioinformatics Program and Department of Computer Science, New Jersey Institute of Technology, Newark, New Jersey.
[2] Center for Applied Genomics, Public Health Research Institute, New Jersey Medical School, University of Medicine and Dentistry of New Jersey, Newark, New Jersey.
[3] Computational RNA Structure Group, Center for Cancer Research Nanobiology Program, National Cancer Institute, Frederick, Maryland.

scoring function for pre-miRNAs with thermodynamic stability and certain structural features, which capture the global properties of the hairpin structures in the pre-miRNAs. Ng and Mishra (2007) employed a Gaussian radial basis function kernel as a similarity measure for 29 global and intrinsic hairpin folding attributes, and characterized pre-miRNAs based on their dinucleotide subsequences, hairpin folding, nonlinear statistical thermodynamics, and topology. Huang et al. (2007) evaluated features valuable for pre-miRNA classification, such as the local secondary structure differences of the stem regions of real pre-miRNA and pseudo pre-miRNA hairpins, and established correlations between different types of mutations and the secondary structures of real pre-miRNAs. More recently, Zhao et al. (2010) considered structure–sequence features and the double helix structure with free nucleotides and base-pairing features. In general, the quality of selected features directly affects the classification accuracy achieved by a method.

In this article, we present a novel method, named MirID, for pre-miRNA classification. MirID accepts as input an RNA sequence, and predicts as output whether the RNA sequence is a pre-miRNA or not. MirID employs a feature mining algorithm for finding combinations of features suitable for building pre-miRNA classification models. These models are implemented using support vector machines (SVMs) (Cortes and Vapnik, 1995; Fan et al., 2005), which are combined to construct a classifier ensemble. The accuracy of the classifier ensemble is further enhanced by the utilization of an AdaBoost algorithm (Bindewald and Shapiro, 2006; Freund and Schapire, 1997; Schapire, 1999).

The study reported here expands upon our previous work (Zhong et al., 2012) where we outlined the algorithms utilized by MirID. This comprehensive study includes (1) a complete flowchart describing our feature mining algorithm; (2) a larger data set containing twenty-one species, as compared to the eleven species previously analyzed (Zhong et al., 2012), and new sequences; (3) new feature values mined from these new sequences and hence new (SVM) classification models obtained from the new data; (4) a thorough experimental study for evaluating the performance and behavior of the MirID algorithms; (5) a web server for online access as well as a downloadable tool for local use; and (6) discussion of potential applications of the software in genomics and medicine.

Materials and Methods

Datasets

Real pre-miRNAs and pseudo pre-miRNAs were collected from twenty-one species, some of which were studied previously while others have not been explored. This collection is comprehensive, covering a variety of species, from viruses to humans. Table 1 summarizes the datasets. The training set and testing set have roughly the same number of sequences. For example, refer to *Schmidtea mediterranea* in the table. We used 72 pre-miRNAs as positive training sequences and 73 pre-miRNAs as positive testing sequences. In addition, we used 201 pseudo pre-miRNAs as negative training sequences and 202 pseudo pre-miRNAs as negative testing sequences.

We downloaded the pre-miRNAs from miRBase (http://www.mirbase.org/) (Kozomara and Griffiths-Jones, 2011). These pre-miRNAs had between 60 and 120 nucleotides (nt). The pre-miRNAs were folded into secondary structures using RNAfold (Hofacker, 2003). As in Xue et al.,(2005), we collected the pseudo pre-miRNAs from GenBank (http://www.ncbi.nlm.nih.gov/genbank/) by selecting short sequences, having between 60 and 120 nt, from the protein-coding regions of the twenty-one species in Table 1. The pseudo pre-miRNAs were chosen in such a way that they had properties (stem–loop structures and free energies) similar to the real pre-miRNAs.

Table 1. Summary of Datasets

Species	*Real pre-miRNA*	*Pseudo pre-miRNA*
Arabidopsis thaliana	66, 67	923, 924
Caenorhabditis briggsae	66, 67	437, 438
Caenorhabditis elegans	84, 85	595, 596
Canis familiaris	161, 161	904, 905
Ciona intestinalis	160, 160	733, 734
Danio rerio	170, 170	1071, 1072
Drosophila melanogaster	81, 82	694, 694
Drosophila pseudoobscura	98, 99	495, 495
Epstein barr virus	12, 13	119, 119
Gallus gallus	241, 241	1186, 1186
Homo sapiens	504, 504	1999, 2000
Macaca mulatta	222, 223	1086, 1086
Medicago truncatula	111, 111	116, 116
Mus musculus	315, 315	2019, 2019
Oryza sativa	172, 172	522, 523
Physcomitrella patens	73, 74	703, 704
Populus trichocarpa	94, 95	809, 810
Pristionchus pacificus	60, 61	58, 58
Rattus norvegicus	193, 193	1238, 1238
Schmidtea mediterranea	72, 73	201, 202
Taeniopygia guttata	94, 95	483, 483

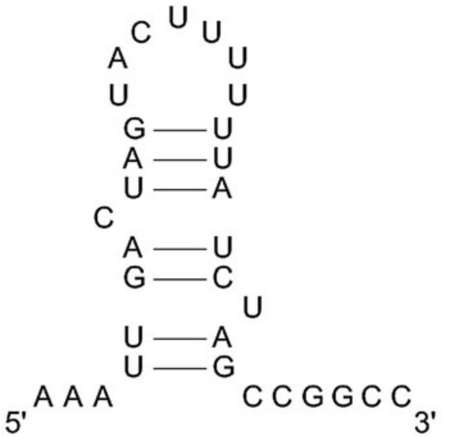
Feature selection

We selected 74 features from a (real or pseudo) pre-miRNA sequence and its secondary structure. These features included the hairpin loop size (Griesmer et al., 2011; Wang and Wu, 2006), the free energy of the secondary structure, its normalized free energy (Spirollari et al., 2009), GC content, the number of bulge loops (Sewer et al., 2005; Xue et al., 2005; Zheng et al., 2006), combined features such as the ratio between the number of base pairs and the sequence length (Zheng et al., 2006), the average size of symmetric and asymmetric internal loops (Sewer et al., 2005), and triplets of structure–sequence elements (Xue et al., 2005).

Triplets contain three consecutive structure elements, which are bases or base pairs (Liu et al., 2005), as well as the nucleotide in the middle of the elements. For example, Figure 1 shows the sequence and structure of a hypothetical pre-miRNA and its dot-bracket notation. Consider the first three dots (bases) and their corresponding nucleotides AAA in Figure 1. The middle nucleotide is A. Thus, the structure-sequence elements "A ..." constitute a triplet. As another example, consider the first three brackets (base pairs) and their corresponding nucleotides UUG in Figure 1. The middle nucleotide is U. Thus, the structure-sequence elements "U(((" constitute a triplet.

MirID algorithms

Central to MirID is a combinatorial feature mining algorithm, which uses a bagging approach (Breiman, 1996) to



```
5' A A A U U G A C U A G U A C U U U U U U A U C U A G C C G G C C 3'
   . . . ( ( ( ( . ( ( ( . . . . . . . ) ) ) ) ) . ) ) . . . . . .
```

FIG. 1. The sequence and structure of a hypothetical pre-miRNA and its dot-bracket notation.

randomly generate N combinations of features (feature sets) from the 74 features at hand. (The default of N is set to 100.) Each feature set contains between 1 and 150 features, randomly chosen with replacement from the 74 features. Each feature set is used to build a classification model, implemented using support vector machines (SVMs). The polynomial kernel in the LIBSVM toolkit (Fan et al., 2005) is used here for the SVM implementation, which achieves the best performance among all kernel functions available in the LIBSVM toolkit. Using the training and testing sequences in Table 1, the accuracy of each classification model is calculated. Classification models with accuracies less than a threshold t are eliminated. (The default of t is set to 0.8.) A classifier ensemble is constructed from the remaining classification models. This ensemble works by taking the majority vote from the individual classification models. The ensemble is continuously refined through an iterative procedure to get the best ensemble. The procedure increments t by a *step* value in each iteration. (The default of *step* is set to 0.005.)

One important function of the combinatorial feature mining algorithm is to perform *merge* and *split* operations on existing feature sets to generate new feature sets. Figure 2 illustrates how these operations work. In Figure 2, S_1 contains 7 features and S_2 contains 3 features. These two feature sets are merged to form S_3, which contains 10 features. Then a number is randomly generated (we assume this number is "6" in this example), and 6 features are randomly selected from S_3 and assigned to $S^{'}_1$, and the remaining 4 features are assigned to $S^{'}_2$. The new feature sets $S^{'}_1$ and $S^{'}_2$ are then used to build new classification models. Figure 3 presents details of our feature mining algorithm, whose output is the best classifier ensemble.

We next applied AdaBoost (Bindewald and Shapiro, 2006; Freund and Schapire, 1997; Schapire, 1999) to the classifier ensemble, which is treated as a weak classifier and is continuously refined into a strong classifier through a procedure with K iterations. (The default of K is set to 20.) The strong classifier is able to predict whether an input RNA sequence is a pre-miRNA or not.

Results

Performance analysis of the MirID method

We carried out a series of experiments to evaluate the proposed MirID method. All the experiments were performed on a 2 GHz Pentium 4 PC having a memory of 2G bytes. The operating system was Cygwin on Windows XP and the algorithms were implemented in Perl. To understand the effect of boosting, we also considered using the combinatorial feature mining algorithm alone to classify pre-miRNAs, and referred to it as the CFM method.

We first evaluated how the number of initial feature sets, N, affects the performance of CFM and MirID. As N increases, more feature sets are generated initially. Thus, the feature mining algorithm constructs a classifier ensemble using more diverse feature sets. Hence the accuracy of the ensemble

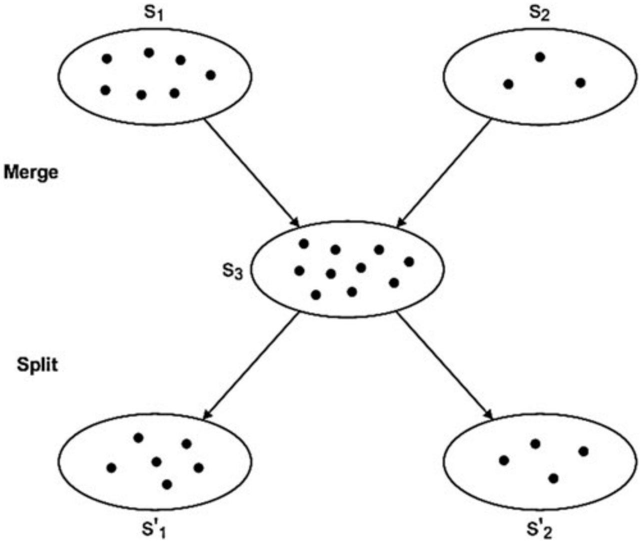


FIG. 2. Illustration of the merge and split operations on two feature sets.

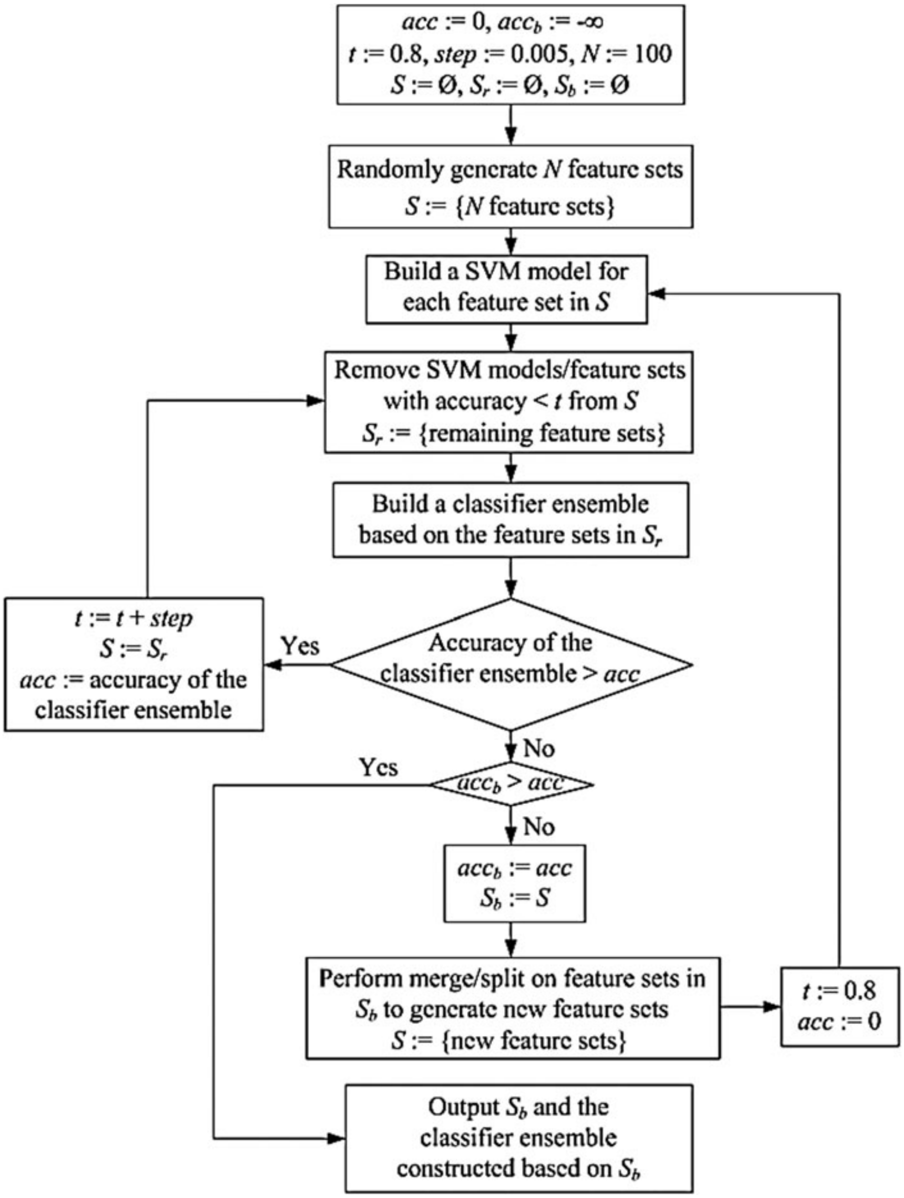

$acc := 0, acc_b := -\infty$
$t := 0.8, step := 0.005, N := 100$
$S := \emptyset, S_r := \emptyset, S_b := \emptyset$
Randomly generate N feature sets
$S :=$ {N feature sets}
Build a SVM model for each feature set in S
Remove SVM models/feature sets with accuracy < t from S
$S_r :=$ {remaining feature sets}
Build a classifier ensemble based on the feature sets in S_r
$t := t + step$
$S := S_r$
$acc :=$ accuracy of the classifier ensemble
Yes
Accuracy of the classifier ensemble > acc
No
Yes
$acc_b > acc$
No
$acc_b := acc$
$S_b := S$
Perform merge/split on feature sets in S_b to generate new feature sets
$S :=$ {new feature sets}
$t := 0.8$
$acc := 0$
Output S_b and the classifier ensemble constructed based on S_b

FIG. 3. Algorithm for combinatorial feature mining.

increases. On the other hand, as N increases, the inner loop in Figure 3 is run more times; as a consequence, the running time increases. MirID requires more time than CFM, due to the extra time spent in boosting. MirID in general is more accurate than CFM, indicating the benefit of including the AdaBoost algorithm.

We next evaluated how the threshold, t, used in the feature mining algorithm affects the performance of CFM and MirID. When t is very large (e.g., $t > 0.95$), the accuracies of the methods drop sharply. This happens because the accuracies of most SVM classification models are less than 0.95 (i.e., 95%), and hence these SVM classification models are eliminated from further consideration early in the feature mining algorithm, cf. Figure 3. When t approaches 1, it is likely that the set S_b returned by the feature mining algorithm is an empty set, and therefore the classifier ensemble constructed based on S_b is also empty, yielding an accuracy of 0. As t increases, fewer feature sets qualify and the set S_r is smaller. As a result, the inner loop in Figure 3 is executed fewer times, and hence the running time decreases.

Then we evaluated how the value, *step*, used to increment the threshold t in each iteration of the inner loop in Figure 3 affects the performance of CFM and MirID. With the default values of N and t used in this study, the feature mining algorithm is able to produce a classifier ensemble with high accuracy. The value of *step* has little impact on the accuracies of the proposed methods. However, as *step* increases, fewer iterations of the inner loop in Figure 3 are executed, and as a consequence, the running time decreases.

We also conducted experiments to test different numbers of iterations, K, in the AdaBoost algorithm. It was found that when K is sufficiently large (e.g., $K \geq 20$), the behavior of the AdaBoost algorithm becomes stable, with the accuracy

Table 2. Accuracies of Four Pre-miRNA Classification Methods Including TripletSVM, PMirP, CFM, and MirID on Twelve Species

Species	*TripletSVM*	*PMirP*	*CFM*	*MirID*
Arabidopsis thaliana	92	96	99	**100**
Caenorhabditis briggsae	96	97	98	**100**
Caenorhabditis elegans	86	86	97	**98**
Danio rerio	67	83	98	**99**
Drosophila melanogaster	92	96	97	**99**
Drosophila pseudoobscura	90	92	98	**100**
Epstein barr virus	**100**	80	98	**100**
Gallus gallus	85	**100**	96	96
Homo sapiens	93	**95**	93	**95**
Mus musculus	94	94	95	**97**
Oryza sativa	95	**100**	97	99
Rattus norvegicus	80	92	97	**98**

The unit of each number in the table is percentage (%).
The highest accuracy for each species is highlighted in boldface.

approaching 100%. On the other hand, when *K* is large, more running time will be needed.

Finally we compared CFM and MirID with two closely related methods, PMirP (Zhao et al., 2010) and TripletSVM (Xue et al., 2005). Table 2 shows the accuracies of these four methods on the twelve species available from the PMirP and TripletSVM web servers. The table shows that MirID is better than or as good as the existing tools on all the species except *Gallus gallus* and *Oryza sativa.* For *Gallus gallus* and *Oryza sativa,* PMirP achieves higher accuracies.

Web server

We have implemented MirID using Perl into a web server, accessible at http://bioinformatics.njit.edu/MirID. The web server accepts a testing sequence as input and classifies the testing sequence as a pre-miRNA or not. We pre-train our web server using the training sequences given in Table 1. In addition to the twelve species available from the PMirP and TripletSVM web servers (Xue et al., 2005; Zhao et al., 2010), we pre-train our web server using nine additional species (shown in Table 1 but not in Table 2). Our tool achieves high accuracies on these nine species, as shown in Table 3. (The PMirP and TripletSVM web servers were not pre-trained on these nine species, and hence we only show the results for CFM and MirID here.) MirID is more accurate than CFM, due to the AdaBoost algorithm.

Table 3. Accuracies of CFM and MirID on Nine Additional Species

Species	*CFM*	*MirID*
Canis familiaris	97	**100**
Ciona intestinalis	94	**100**
Macaca mulatta	**96**	**96**
Medicago truncatula	95	**100**
Physcomitrella patens	**100**	**100**
Populus trichocarpa	97	**99**
Pristionchus pacificus	96	**100**
Schmidtea mediterranea	95	**99**
Taeniopygia guttata	95	**99**

The unit of each number in the table is percentage (%).
The highest accuracy for each species is highlighted in boldface.

Table 4. Number of Feature Sets for Each Species in MirID

Species	*Number of feature sets*
Arabidopsis thaliana	1
Caenorhabditis briggsae	6
Caenorhabditis elegans	1
Canis familiaris	1
Ciona intestinalis	7
Danio rerio	11
Drosophila melanogaster	3
Drosophila pseudoobscura	4
Epstein barr virus	5
Gallus gallus	3
Homo sapiens	1
Macaca mulatta	1
Medicago truncatula	1
Mus musculus	3
Oryza sativa	3
Physcomitrella patens	1
Populus trichocarpa	1
Pristionchus pacificus	1
Rattus norvegicus	10
Schmidtea mediterranea	32
Taeniopygia guttata	5

Table 4 shows, for each species in Table 1, the number of feature sets produced by our feature mining algorithm. Table 5 shows the CPU time (in seconds) spent in pre-training the MirID web server. The training time depends on the number of feature sets, the number of features in each feature set, the number of iterations used by the feature mining algorithm, and the number of iterations used in the AdaBoost algorithm. Notice that this training is done once, and no more training is

Table 5. Training Time for Each Species in MirID

Species	*Training time (in seconds)*
Arabidopsis thaliana	80
Caenorhabditis briggsae	348
Caenorhabditis elegans	103
Canis familiaris	153
Ciona intestinalis	269
Danio rerio	1272
Drosophila melanogaster	199
Drosophila pseudoobscura	196
Epstein barr virus	113
Gallus gallus	274
Homo sapiens	1530
Macaca mulatta	243
Medicago truncatula	104
Mus musculus	786
Oryza sativa	214
Physcomitrella patens	90
Populus trichocarpa	138
Pristionchus pacificus	63
Rattus norvegicus	349
Schmidtea mediterranea	478
Taeniopygia guttata	156

needed on the testing data. It takes less than a second to classify an unlabeled testing sequence.

Discussion

The MirID method was shown experimentally to be better than two existing tools, PMirP and TripletSVM, for the majority of the twelve species analyzed with these two tools. For nine additional species, MirID also performed well. The good performance of the MirID method is due to its novel feature mining and AdaBoost algorithms.

Both the feature mining and AdaBoost algorithms contain user-specified parameters. As indicated by our experimental results in the performance analysis section, changing these parameter values may affect the running time and accuracy of our method. The MirID web server adopts the default parameter values as used in this study, to achieve good and stable performance. The server is able to process sequences of a variety of species, from viruses to humans. It does not include bacteria, however. While there are small regulatory RNAs in bacteria, bacteria do not have true miRNAs (Gottesman, 2005; Tjaden et al., 2006). Bacterial miRNA will be added to our server when such data is validated and becomes available in public databases.

Currently, the MirID web server is capable of classifying one testing sequence at a time, predicting whether the testing sequence is a pre-miRNA or not. When multiple testing sequences must be classified, we suggest that the user run the tool locally in a batch mode. Instructions for downloading the tool and running the tool locally can be obtained from http://bioinformatics.njit.edu/MirID-download.

MicroRNAs play important roles in most biological processes, including cell proliferation, tissue differentiation, and embryonic development, to name a few (Aukerman and Sakai, 2003; Brennecke et al., 2003; Bushati and Cohen, 2007; Johnston and Hobert, 2003; Tang et al., 2009; Xu et al., 2009). They interact with target mRNAs at specific sites to induce cleavage of the message or inhibit translation (John et al., 2004). They can have multiple mRNA targets as they bind to the targets with partial complementarities in animals. In addition, an mRNA target can be regulated by multiple miRNAs at different loci with different effects. This adds to the complexity of finding out the mRNA targets in genomes (John et al., 2004).

The total number of microRNA discovered continues growing every day. According to the latest miRBase release (version 19, August 2012), accessible at http://www.mirbase.org, there are 2019 unique mature human miRNAs up from 894 in the version 14. There seems to be a correlation between the tissue specificity of a human miRNA and the number of diseases with which the miRNA is associated (Lu et al., 2008). The fact that microRNAs are found circulating in blood (Mitchell et al., 2008; Schöler et al., 2011) holds great promise for the development of diagnostic tools that can be used in multiple ways, from noninvasive pregnancy diagnostic tests to cancer diagnostics and treatment. A tool like MirID for predicting pre-miRNAs will contribute to our basic understanding of the roles played by microRNA in regulating many biological processes, and their contribution to disease development and progression.

A potential application for the MirID tool is in the area of individualized genomic analysis. With the advent of high-throughput sequencing technologies, millions of short reads can now be generated from a library of nucleotide sequences. These technologies have catalyzed a new era of personalized medicine based on individualized genomic analysis (Anderson and Schrijver, 2010). Determining levels of known and novel microRNA from small RNA sequencing data is an important subject in this new era (An et al., 2013). With next-generation sequencing platforms, several prostate-expressed microRNAs related to prostate cancer have been identified (Martens-Uzunova et al., 2012; Ostling et al., 2011; Ribas et al., 2009; Wang et al., 2011; Watahiki et al., 2011). As a consequence, exploring microRNAs and their functions continues to be a highly active area of research. The MirID tool developed from this work can be used to assess aggregated RNA-seq reads for pre-miRNA secondary structure potential. The tool can be combined and integrated with other miRNA profiling tools (Friedlander et al., 2012; Hackenberg et al., 2011; Hendrix et al., 2010; Mathelier and Carbone, 2010) for applications to personalized medicine. We plan to investigate the integration of these tools and evaluate the effectiveness of the integrated pipeline system in the future.

Acknowledgments

We thank the reviewers for their thoughtful comments and constructive suggestions. This research was supported in part by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research. The current findings were presented in part in preliminary form in the Proceedings of the 2012 IEEE International Conference on Bioinformatics and Biomedicine (Zhong et al., 2012).

Author Disclosure Statement

The authors declare that no conflicting financial interests exist.

References

An J, Lai J, Lehman ML, and Nelson CC. (2013). miRDeep*: An integrated application tool for miRNA identification from RNA sequencing data. Nucleic Acids Res 41, 727–737.

Anderson MW, and Schrijver I. (2010). Next generation DNA sequencing and the future of genomic medicine. Genes 1, 38–69.

Aukerman NJ, and Sakai H. (2003). Regulation of flowering time and floral organ identity by a microRNA and its APETALA2-like target genes. Plant Cell 15, 2730–2741.

Bartel DP. (2004). MicroRNAs: Genomics, biogenesis, mechanism, and function. Cell 116, 281–297.

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

Bindewald E, and Shapiro BA. (2006). RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. RNA 12, 342–352.

Bindra RS, Wang JTL, and Bagga PS. (2010). Bioinformatics methods for studying microRNA and ARE-mediated regulation of post-transcriptional gene expression. Intl J Knowledge Disc Bioinformatics 1, 97–112.

Breiman L. (1996). Bagging predictors. Machine Learning 24, 123–140.

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

Bushati N, and Cohen SM. (2007). MicroRNA functions. Annu Rev Cell Dev Biol 23, 175–205.

Cortes C, and Vapnik V. (1995). Support-vector networks. Machine Learning 20, 273–297.

Fan R, Chen P, and Lin C. (2005). Working set selection using the second order information for training SVM. J Machine Learning Res 6, 1889–1918.

Freund Y, and Schapire RE. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. J Comput System Sci 55, 119–139.

Friedlander MR, Mackowiak SD, Li N, Chen W, and Rajewsky N. (2012). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res 40, 37–52.

Gomez A, and Ingelman-Sundberg M. (2009). Epigenetic and microRNA-dependent control of cytochrome P450 expression: A gap between DNA and protein. Pharmacogenomics 10, 1067–1076.

Gottesman S. (2005). Micros for microbes: Non-coding regulatory RNAs in bacteria. Trends Genet 21, 399–404.

Griesmer SJ, Cervantes-Cervantes M, Song Y, and Wang JTL. (2011). In silico prediction of noncoding RNAs using supervised learning and feature ranking methods. Intl J Bioinformat Res Applicat 7, 355–375.

Hackenberg M, Rodriguez-Ezpeleta N, and Aransay AM. (2011). miRanalyzer: An update on the detection and analysis of microRNAs in high-throughput sequencing experiments. Nucleic Acids Res 39, W132–W138.

Hendrix D, Levine M, and Shi W. (2010). miRTRAP, a computational method for the systematic identification of miRNAs from high throughput sequencing data. Genome Biol 11, R39.

Hofacker IL. (2003). Vienna RNA secondary structure server. Nucleic Acids Res 31, 3429–3431.

Huang TH, Fan B, Rothschild MF, Hu ZL, Li K, and Zhao SH. (2007). MiRFinder: An improved approach and software implementation for genome-wide fast microRNA precursor scans. BMC Bioinformat 8, 341.

John B, Enright AJ, Aravin A, Tuschl T, Sander C, and Marks DS. (2004). Human microRNA targets. PLoS Biol 2, e363.

Johnston RJ, and Hobert O. (2003). A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans*. Nature 426, 845–849.

Kozomara A, and Griffiths-Jones S. (2011). miRBase: Integrating microRNA annotation and deep sequencing data. Nucleic Acids Res 39, D152–D157.

Lai EC, Tomancak P, Williams RW, and Rubin GM. (2003). Computational identification of Drosophila microRNA genes. Genome Biol 4, R42.

Lee RC, Feinbaum RL, and Ambros V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75, 843–854.

Lim LP, Glasner ME, Yekta S, Burge CB, and Bartel DP. (2003). Vertebrate microRNA genes. Science 299, 1540.

Liu J, Wang JTL, Hu J, and Tian B. (2005). A method for aligning RNA secondary structures and its application to RNA motif detection. BMC Bioinformat 6, 89.

Lu M, Zhang Q, Deng M, et al. (2008). An analysis of human microRNA and disease associations. PLoS One 3, e3420.

Mack GS. (2007). MicroRNA gets down to business. Nat Biotechnol 25, 631–638.

Martens-Uzunova ES, Jalava SE, Dits NF, et al. (2012). Diagnostic and prognostic signatures from the small non-coding RNA transcriptome in prostate cancer. Oncogene 31, 978–991.

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

Mitchell PS, Parkin RK, Kroh EM, et al. (2008). Circulating microRNAs as stable blood-based markers for cancer detection. Proc Natl Acad Sci USA 105, 10513–10518.

Ng KL, and Mishra SK. (2007). De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. Bioinformatics 23, 1321–1330.

Ostling P, Leivonen SK, Aakula A, et al. (2011). Systematic analysis of microRNAs targeting the androgen receptor in prostate cancer cells. Cancer Res 71, 1956–1967.

Pan YZ, Gao W, and Yu AM. (2009). MicroRNAs regulate CYP3A4 expression via direct and indirect targeting. Drug Metab Dispos 37, 2112–2117.

Ribas J, Ni X, Haffner M, et al. (2009). miR-21: An androgen receptor-regulated microRNA that promotes hormone-dependent and hormone-independent prostate cancer growth. Cancer Res 69, 7165–7169.

Sampson VB, Rong NH, Han J, et al. (2007). MicroRNA let-7a down-regulates MYC and reverts MYC-induced growth in Burkitt lymphoma cells. Cancer Res 67, 9762–9770.

Schapire RE. (1999). A brief introduction to boosting. In: Proc Sixteenth Intl Joint Conf Artificial Intelligence, 1401–1406.

Schöler N, Langer C, and Kuchenbauer F. (2011). Circulating microRNAs as biomarkers—True blood? Genome Med 3, 72.

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

Spirollari J, Wang JTL, Zhang K, Bellofatto V, Park Y, and Shapiro BA. (2009). Predicting consensus structures for RNA alignments via pseudo-energy minimization. Bioinformat Biol Insights 3, 51–69.

Takagi S, Nakajima M, Mohri T, and Yokoi T. (2008). Post-transcriptional regulation of human pregnane X receptor by micro-RNA affects the expression of cytochrome P450 3A4. J Biol Chem 283, 9674–9680.

Tang YF, Zhang Y, Li XY, Li C, Tian W, and Liu L. (2009). Expression of miR-31, miR-125b-5p, and miR-326 in the adipogenic differentiation process of adipose-derived stem cells. OMICS 13, 331–336.

Tjaden B, Goodwin SS, Opdyke JA, et al. (2006). Target prediction for small, noncoding RNAs in bacteria. Nucleic Acids Res 34, 2791–2802.

Tsuchiya Y, Nakajima M, Takagi S, Taniya T, and Yokoi T. (2006). MicroRNA regulates the expression of human cytochrome P450 1B1. Cancer Res 66, 9090–9098.

Wang JTL, and Wu X. (2006). Kernel design for RNA classification using support vector machines. Intl J Data Mining Bioinformat 1, 57–76.

Wang T, Zhang X, Obijuru L, et al. (2007). A micro-RNA signature associated with race, tumor size, and target gene activity in human uterine leiomyomas. Genes Chromosomes Cancer 46, 336–347.

Wang WL, Chatterjee N, Chittur SV, Welsh J, and Tenniswood MP. (2011). Effects of 1alpha,25 dihydroxyvitamin D3 and testosterone on miRNA and mRNA expression in LNCaP cells. Mol Cancer 10, 58.

Watahiki A, Wang Y, Morris J, et al. (2011). MicroRNAs associated with metastatic prostate cancer. PLoS One 6, e24950.

Xu CF, Yu CH, and Li YM. (2009). Regulation of hepatic microRNA expression in response to ischemic preconditioning following ischemia/reperfusion injury in mice. OMICS 13, 513–520.

Xue C, Li F, He T, Liu GP, Li Y, and Zhang X. (2005). Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. BMC Bioinformat 6, 310.

Zhao D, Wang Y, Luo D, et al. (2010). PMirP: A pre-microRNA prediction method based on structure-sequence hybrid features. Artificial Intelligence Med 49, 127–132.

Zheng Y, Hsu W, Lee ML, and Wong LS. (2006). Exploring essential attributes for detecting microRNA precursors from background sequences. Lecture Notes Comput Science, Springer 4316, 131–145.

Zhong L, Wang JTL, Wen D, and Shapiro BA. (2012). Pre-miRNA classification via combinatorial feature mining and boosting. In: Proc 2012 IEEE Intl Conf Bioinformat Biomed 369–372.

Zhu H, Wu H, Liu X, et al. (2008). Role of MicroRNA miR-27a and miR-451 in the regulation of MDR1/P-glycoprotein expression in human cancer cells. Biochem Pharmacol 76, 582–588.

Zuker M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31, 3406–3415.

Address correspondence to:
Dr. Jason T. L. Wang
Bioinformatics Program and Department of Computer Science
New Jersey Institute of Technology
GITC Building, Room #4211
218 Central Avenue
Newark, New Jersey 07102

E-mail: wangj@njit.edu