

Pre-miRNA Features for Automated Classification

Andrei-Lucian Ioniță, Liviu Ciortuz

Faculty of Informatics,

“Al. I. Cuza” University of Iasi, Romania

E-mail: lucian.ionita@infoiasi.ro, ciortuz@infoiasi.ro

Abstract—We present a system for precursory microRNA classification that implements many types of features found in the literature: structural, thermodynamical, information-theoretical, and comparative. A total of 1485 features are computed and various tests are performed. We used Random Forests first as the classifier of choice and secondly in conjunction with various feature selection strategies in order to determine the most salient features and to increase the classification performance.

I. INTRODUCTION

MicroRNAs (miRNAs) are small non-coding RNA that have an extremely important role in the regulation of gene expression [9]. It has been estimated that there may be on the order of 1000 miRNAs in the human genome. MiRNAs may influence about two thirds of the mammalian genes and thus are probably involved in most biological processes. More interestingly, it has been found that there is a link between miRNA expression profiles and cancer [3]. This has led to a pronounced interest in miRNA study in the bio-medical community. Unfortunately it is difficult to experimentally ascertain whether a sequence is or is not a miRNA. A convenient alternative is to computationally predict regions of the genetic sequence that contain, with high probability, miRNAs. One method of formulating this task is using the classical machine learning framework: given a database of known true and false precursory miRNAs, one has to find a model that would classify, as accurately as possible, a given new sequence. Precursory miRNAs (pre-miRNAs) are long sequences (~80-120nt) that exhibit a hairpin-like structure, as a stem finished with a loop. This will be cleaved by the RNase III enzyme Dicer [10], resulting in an imperfect duplex (miRNA:miRNA*). Finally, one of those two strands in the duplex is incorporated into the RNA-Induced Silencing Complex (RISC). It is said that the pre-miRNA has matured into miRNA.

Aiming to create a system for accurate classification of pre-miRNAs, we have first realized a very comprehensive catalogue of features for such sequences. After we have analyzed most of the existing systems that extract information from pre-miRNAs in various ways, we have integrated many types of features in our system: primary structure features, secondary structure features, profile alignment score and edit distances, information-theoretical features, topological features etc, in total 1485 attributes. We have not included conservation or phylogeny data in our system because this raises the complexity of the approach, and it will be left for future work. Secondly, for the classification component of our system,

we have considered both Support Vector Machines (SVM) and Random Forests (RF). SVM enjoyed a wide success in many previous studies in bioinformatics and in particular pre-miRNA identification. RF is a well-rounded classifier and, remarkably, it can lead to different feature selection methods, easily derived from the initial algorithm.

In this work some feature ranking and feature selection strategies have been studied in conjunction with RF and it has been shown that approximately 98% of the very large feature pool could be eliminated without significantly reducing the classifier's performance.

The structure of this paper is it follows: this first section gave an overview of the subject and the work we have done, the second section will present the features, the classification and feature selection methods we used, the third section will show the results we obtained, and the last section will draw the conclusions and will sketch future work.

II. METHODS

Here we will discuss in detail the features included in our system and also the classifiers and the feature selection methods that we employed.

A. Features used

This first part of our system consists of extracting as much information as possible from real or pseudo miRNA sequences. To this end we have analyzed various systems created for identification of miRNA or similar. These include microPred [1], miPred [12], Droscha SVM [5], Triplet-SVM [18], MiRFinder [7], MiPred [8], Diana-microH [17], mirEncoding [19], yasMiR [13] [14], and others. Many pre-miRNA features have been used in more than one system, especially straightforward features like Minimum Free Energy or nucleotide counts, but others, more innovative features like the bio-chemical indices [16] have not been used extensively and almost never in combination with other innovative features. One *goal* of this research is to find sets of features that together have a strong discriminative power.

Tables I to V show an extensive list of the features we integrated in our system. They have been loosely grouped in primary and secondary structure features, energy related features, miscellaneous(which include information-theoretical and normalized) features and comparative features. Some of the most interesting or unintuitive features are described in the following paragraphs.

TABLE I
PRIMARY STRUCTURE FEATURES

Symbol(s)	Primary Structure Features
nA, nC, nG, nT	Number of each nucleotide
pA, pC, pG, pT	Frequencies of nucleotides
nAA, nAG, nCU etc	Number of each dinucleotide
pAA, pAG, pCU etc	Frequencies of dinucleotides
L	Sequence length
pcCG	CG content (pC + pG)

Primary structure features concern only those that can be simply deduced by looking at the nucleotide sequence. They are straightforward and consist of the number and frequency of nucleotides and dinucleotides, the sequence length and the CG-content of the sequence.

Secondary structure features are more complex. They are concerned with the folding of the RNA molecule. Here we usually consider only the Minimum Free Energy folding structure at the normal temperature of 37C, and the corresponding number of base pairs, the mean base-pair distance, the number of stems or bulges, the size of the hairpin loop etc. Other features take into account the base-pairing profile calculated by the Vienna RNA package [6]. This profile is an L by L matrix (L being the length of the sequence) representing the probabilities for each two nucleotides to form a base-pair. These probabilities are computed according to McCaskill's algorithm [11], which is based on thermodynamic principles. Features representing for instance the probabilities of each nucleotide type to be unpaired are calculated by using this profile.

The local structure features described by Xue et al [18] correspond to the number of contiguous nucleotide triplets, more precisely by taking into account the paired/unpaired status (represented by either parenthesis or a dot) of each nucleotide in the triplet, together with the base type in the middle position. To exemplify, such a feature is designated as "A.((" which indicates a triplet where the first and third nucleotides are paired, but the second (the Adenine) is not. The direction of pairing is considered unimportant, so only the symbol "((" is used for paired bases.

A probabilistic version of these features was used in yas-MiR [13], where each triplet pattern in the RNA molecule is weighted according to the probabilities of all possible secondary structures (of the given RNA sequence) in which the base pairs satisfying that triplet pattern may appear.

Another version of the local structure description was considered in MiRFinder [7], where five symbols were introduced ("=", ":", ".", "-" and "^") indicating the states of paired, unpaired, insertion, deletion and bulge, respectively. A feature is designated by two such symbols, and its value is the number of all base pairs satisfying the pattern. The information about the nucleotide type is discarded.

TABLE II
SECONDARY STRUCTURE FEATURES

Symbol(s)	Secondary Structure Features
pAnp, pCnp etc	Non-pairing probabilities for each nucleotide
A((. C.(. U... etc	Number of local continuous triplet structures as described in Triplet SVM[18]
pA((. pC.(. pU... etc	Probabilities of each continuous triplet structures, analogous to previously defined features
n(. n((. n... etc	Number of three-nucleotide pairing structures
p(. p((. p... etc	Probabilities of three-nucleotide pairing structures
pAleft, pAright etc	Frequency of each nucleotide in left, respectively right arms
loopen	Loop length
bulge	Bulge size
nbulge	Bulge
symdiff	Symmetric difference
tlen	Tail length
ntails	Number of tails
. = == :: := :. etc	Dinucleotide pairing local structures as described in [7]
nstems	Number of stems
njunc	Number of junctions
nend	Number of end points
nmid	Number of midpoints
F	Second (Fielder) Eigen Value of the Laplacian matrix of the tree-graph structure
nAUb, nGCb etc	Number of each base pairs
nAUb/L, nGCb/L etc	Number of each base pair normalized to sequence length
pAUb, pGCb etc	Probability of each base pair
nAU/nstems etc	Average number of each base pair per stem
BPs	Average number of base pairs per stem
BP	Number of base pairs
BP/L, P	Normalized base pair propensity
meanBPdist	Probabilistically, the mean base pair distance
d5loop	Distance from '5 end to start of loop
BTI ₀ , BTI ₁ , ..., BTI ₂₃	Biochemical topological indices
D	Average base pair distance
D/L	Normalized average base pair distance

In [16], Shu et al. introduced the idea of applying topological indices from the chemical graph theory to the pre-miRNA molecule. The pre-miRNA is represented as an "element-contact graph". Three graph representations of this kind are given, where the stems and loops are vertices, while adjacent elements are connected via edges. Different indices are computed from each graph, namely the Wiener, Balaban

TABLE III
ENERGY RELATED FEATURES

Symbol(s)	Energy Related Features
MFE	Minimum Free Energy
NMFE	Normalized Minimum Free Energy
MFEI ₁	Minimum Free Energy Index 1 (NMFE/pcCG)
MFEI ₂	Minimum Free Energy Index 2 (NMFE/nstems)
MFEI ₃	Minimum Free Energy Index 3 (NMFE/nloops)
MFEI ₄	Minimum Free Energy Index 4 (MFE/BP)
EFE	Ensemble Free Energy
NEFE	Normalized Ensemble Free Energy (EFE/L)
FREQ	Frequency of MFE structure
VI	Valley Index
dH	Structure Enthalpy
dH/L	Normalized Structure Enthalpy
Tm	Melting Energy
Tm/L	Normalized Melting Energy
DIV	Asta merge la cat 2 Structural Diversity
EDIV	Ensemble Diversity

TABLE IV
MISCELLANEOUS FEATURES

Symbol(s)	Miscellaneous Features
dS	Structure Entropy
dS/L	Normalized Structure Entropy
Q	Shannon Entropy
Q/L	Normalized Shannon Entropy
Z_MFE, Z_Q etc	Z-score for MFE, Q, D, P and F
P_MFE, P_Q etc	P-score for MFE, Q, D, P and F
ND	Nucleotide Descriptors

and Randić indices of first and second order. Moreover, the weighted versions of these indices are also used, totaling 24 features. These indices offer a method of quantifying the connectiveness of the graph.

A similar approach is used when computing the Fielder eigenvalue [1]. The RNA molecule is represented as a tree-graph structure, in which vertices represent loops while edges represent stems. The second eigenvalue of the Laplacian matrix associated to such a tree-graph is a measure of its compactness.

Among the *energy related features*, the MFE of the RNA structure is the most important and many variations of it were defined, for instance by normalizing the MFE by sequence length, number of stems etc. From a thermodynamical point of view, the RNA molecule exists in an assembly of structures that can be probabilistically modeled using a Boltzmann

TABLE V
COMPARATIVE FEATURES

Symbol(s)	Secondary Structure Features
PAS _{0..n}	Profile Alignment Scores
TED _{e0..n}	Tree Edit Distances for expanded notation
WTED _{e0..n}	Weighted Tree Edit Distances for expanded notation
TED _{c0..n}	Tree Edit Distances for coarse grained notation
WTED _{c0..n}	Weighted Tree Edit Distances for coarse grained notation
SED _{e0..n}	String Edit Distances for expanded notation
WSED _{e0..n}	Weighted String Edit Distances for expanded notation
SED _{c0..n}	String Edit Distances for coarse grained notation
WSED _{c0..n}	Weighted String Edit Distances for coarse grained notation
PED _{0..n}	Profile Edit Distances

distribution. This information is captured by the Ensemble Free Energy, the Ensemble Diversity and other related features.

Information-theoretical features measure the entropy of the base pairing profile (Structure entropy) and of the ensemble (Shannon entropy). These features offer a measure of the diversity of the possible structures of the RNA sequence.

So-called *normalized* versions of some features have been introduced in miPred [12]. This system uses the observation that pre-miRNAs have lower MFE than random sequences with the same dinucleotide frequencies. By using a dinucleotide shuffling algorithm, a large number of sequences are obtained starting from the given one, and then their MFEs are computed. A P-value is calculated, which is the fraction of the MFEs that are lower than that of the original sequence. The Z-value is an even more important measure which also takes into account the standard deviation of the samples obtained. This process of finding P- and Z-values has also been performed for the Shannon entropy, the average base-pair distance, the base pair propensity and the Fielder eigen value.

Nucleotide descriptors, used in [5] are a set of binary features that represent the information regarding nucleotide type and pairing for each of the 24 bases in the left, respectively right arms. Every nucleotide is represented by four features which describe its type (A C G or U) and 2 for its paired/unpaired status. These features were included in the miscellaneous category due to the fact that there is little probability that they would have a significant effect on classification accuracy. They have been included nonetheless for purposes of completion and comparison. Also in the *miscellaneous* category we include the information-theoretical features and *normalized* ones.

Comparative Features compute a measure of similarity between two RNA sequences. In yasMiR [13], McCaskill's profile alignment scores [11] were used with great success.

Every given pre-miRNA was aligned with a database of 100 random pre-miRNA-like sequences (“pivots”) and their similarity scores were computed. We extended this approach by introducing various other alignment scores that were computed with the Vienna RNA package, namely the tree edit distance, the string edit distance and the profile edit distance. Four representations are used in conjunction with string and tree edit distances, two of which are the weighted variants of the others. In total, there are 10 comparative features for each pivot in the database. The yasMiR database of 100 pivot sequences was used for doing comparisons with the analyzed sequences.

B. The Classifier

For this work, the classifier we have chosen is Random Forests (RF) [2]. A random forest is a collection of decision trees grown without pruning according to the CART methodology. Each decision tree is trained on a sample of the data and, at each node the best split is chosen from a number of randomly chosen features. A test instance is then classified by taking the majority vote among the trees in the RF. The generalization error of a RF depends largely on the strength of the individual trees and on the (lack of) correlation between them. The two types of randomness involved in the creation of a RF ensure low correlation, while the sound CART methodology ensures the classification strength.

The portion of samples that have not been used in training is called the out-of-bag (oob) sample and can be used to generate an unbiased estimate of the generalization error of the classifier. This is done by evaluating each example with the trees for which it was not part of the training sample and then choosing the class that had the most votes. The accuracy achieved with this method is the out-of-bag accuracy.

For this paper, the parameters for the RF are as follows: the number of features evaluated at each node, m_{try} , is \sqrt{m} , where m is the number of features, the number of trees N is set at 2500 and the fraction of in-bag samples is 0.66.

C. Feature Selection

Feature selection is an important topic in machine learning because it gives us information regarding the relevant variables, it decreases the dimensionality of the data, and sometimes it improves the classification performance. For this paper we have used a feature selection strategy called Recursive Feature Elimination [4]. The algorithm involves training a classifier, evaluating or ranking the features, eliminating a fraction of them and repeating this procedure until some criterion is met. In this work, we eliminated 10% of the features at each step. We have tested three feature evaluation methods in connection with RF. The first is a naive procedure which ranks each variable by the number of times it has been chosen as the best split in a node. The second method measures the change in average oob accuracy in the trees when, after training the classifier, a feature is randomly permuted between examples. The third is the change in overall RF oob accuracy when permuting the value of a feature.

III. RESULTS

This objective of this section is to evaluate the system defined in the previous sections and to compare its performance with that of other systems found in the literature. For comparisons, we chose two such systems, namely Triplet-SVM and yasMiR.

The dataset on which Triplet-SVM was trained consisted of 331 examples, of which 163 were positive. They have been randomly selected from the 193 positive pre-miRNA sequences from miRBase 5.0. The negative examples are pre-miRNA-like sequences randomly chosen from the CODING set, which itself is a subset of the NCBI RefSeq database [15]. The authors of Triplet-SVM also built four datasets for testing purposes: The TE-C set contains the 30 remaining human pre-miRNA from miRBase 5.0 and 1000 pseudo pre-miRNAs from the CODING set, excluding those examples selected for training. The UPDATED set contains the 39 pre-miRNAs discovered after the release of miRBase 5.0 and before the completion of Triplet-SVM. The CROSS-SPECIES set is composed of 581 pre-miRNAs from nonhuman species, found in miRBase 5.0. The CONSERVED-HAIRPIN set was created by scanning a part of chromosome 19 of the human genome and extracting 2444 hairpin structures, of which only 3 are real pre-miRNAs.

For comparison reasons, we have also included the results from 5-fold cross-validation accuracy on the TE-C dataset, since this was reported for both Triplet-SVM and yasMiR.

The second dataset used here is the miRBase set referred to in the yasMiR technical report [14]. The training set is composed of the 678 human pre-miRNAs from miRBase 11.0 along with 1256 sequences from CODING as negative examples. The testing set consists of 3651 pre-miRNAs from miRBase 12.0 and 7198 pseudo pre mi-RNAs from the CODING dataset. The positive examples were chosen using the methodology described by the authors of miPred [12] which ensured the retention of sufficiently dissimilar sequences.

The third dataset we used was the one created for testing miPred. The training set TR-H contains 200 randomly chosen pre-miRNAs from miRBase 8.2 and 400 pseudo pre-miRNAs from human RefSeq genes. Also, four test sets were created TE-H with 123 human pre-miRNAs and 246 randomly chosen pseudo hairpins, IE-NH containing 1918 non-human miRNA and 3836 randomly selected negative examples, IE-NC comprising of 12387 functional ncRNA from Rfam 7.0, and IE-M, which contains 31 mRNAs from the GenBank DNA database.

Table VI shows the comparative accuracy results we obtained at 5-fold cross-validation and on TE-C and on the test sets from Triplet-SVM and the above described miRBase 12.0. Table VII presents the results on the (re)created miPred dataset. The results listed in these two tables for Triplet SVM, miPred and yasMiR were taken from the respective papers.

Our system shows accuracies significantly above those of Triplet-SVM in all tests, which was expected. On the Triplet-SVM test sets, our results were at least on par with yasMiR’s, if not slightly superior, while on the miRBase dataset the

TABLE VI
COMPARATIVE RESULTS ON TRIPLET-SVM AND miRBASE 12.0
DATASETS

Test	Our system's accuracy (%)	yasMiR accuracy (%)	Triplet-SVM accuracy (%)
TE-C Cross Validation	96.99 (98.73 oob)	96.07	93.5
TE-C: Human pre-miRNA	100	100	93.3
TE-C: Pseudo pre-miRNA	96.4	96.2	88.1
UPDATED	97.4	94.9	92.3
CROSS-SPECIES	94.75	95.2	90.9
CONSERVED-HAIRPIN	93.08	94.24	89
miRBase 12.0	95.15	94.77	-

TABLE VII
COMPARATIVE RESULTS ON THE RECONSTRUCTED miPred DATASET

Test	Our system's accuracy (%)	yasMiR accuracy (%)	miPred accuracy (%)	Triplet-SVM accuracy (%)
TE-H	94.30	93.77	93.50	87.96
IE-NH	94.91	94.11	95.64	86.15
IE-NC	77.71	82.95	68.68	78.37
IE-M	96.77	100	87.09	0

results show that our system has a small advantage. On the miPred dataset our system was the best on TE-H (compared to yasMiR, miPred and Triplet-SVM) and it was second best on IE-NH and IE-M. It was significantly less good than yasMiR (the best) on IE-NC and IE-M.

For direct comparison between RF and SVM, we have trained an SVM classifier on the miRBase 11.0 dataset as used by yasMiR, using all 1485 features. We got a 94.58% test accuracy on the miRBase 12.0 dataset (also as used by yasMiR), which is lower than both our classifier and yasMiR.

We experimented with the feature selection strategy called Recursive Feature Elimination (RFE). At each iteration, a proportion of the weakest features – in this work 10% – are eliminated. The best feature subset is subsequently chosen by selecting the iteration which optimizes some criterion, e.g. the cross validation accuracy. Here we used as optimization criterion the RF out-of-bag accuracy estimator.

Figures 1, 2 and 3 correspond to the experiments that we performed using the three feature evaluation criteria described in Section II. The three curves represent the averaged oob accuracy, the overall RF oob accuracy and the test accuracy obtained on the miRBase 12.0 dataset. The subsets of features that were eventually selected in this experiment lead to test accuracies of 95.16%, 95.17% respectively 95.60% for the three methods. The first two results are not statistically signif-

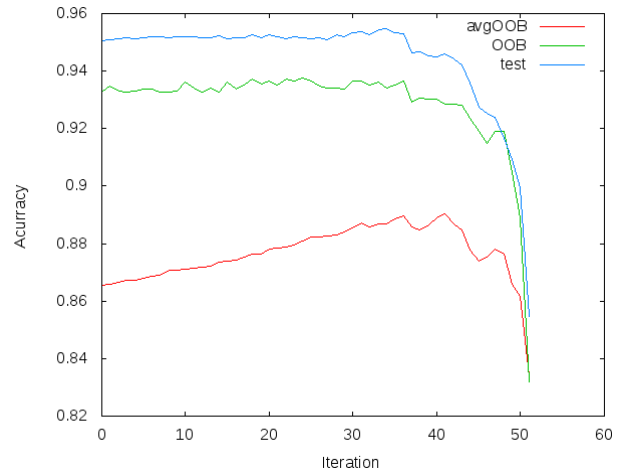


Fig. 1. Feature selection with feature counting.

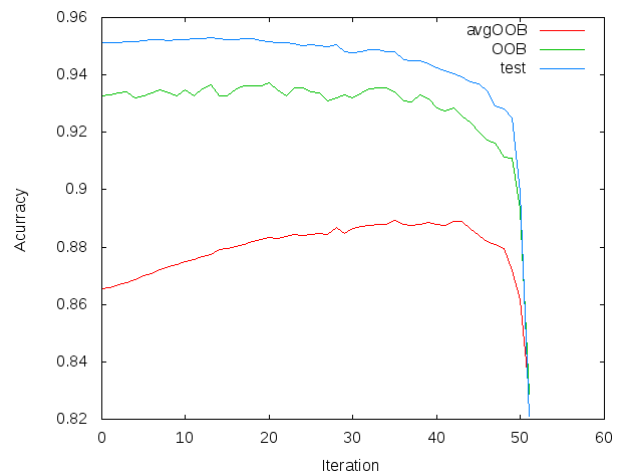


Fig. 2. Feature selection with average oob error estimation.

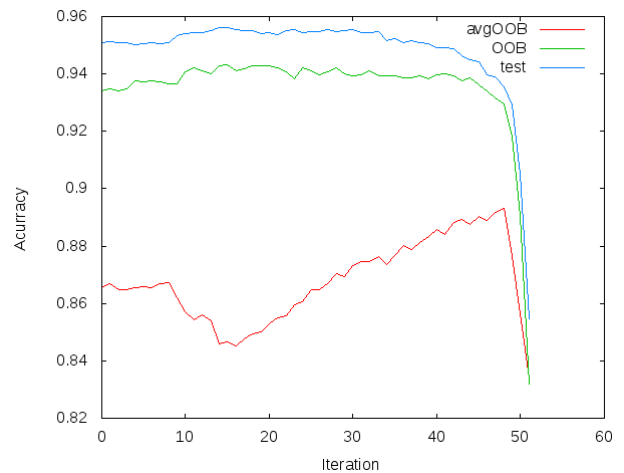


Fig. 3. Feature selection with forest oob error estimation.

icant and this is probably due to the enormous redundancy of information expressed in the feature set and the fact that these

TABLE VIII
COMPARATIVE RESULTS OF FEATURE SELECTION METHODS ON
TRIPLET-SVM DATASETS

Test	No FS	Naive FS	average oob FS	oob FS
TE-C: Human pre-miRNA	100	100	100	100
TE-C: Pseudo pre-miRNA	96.4	96.3	96.3	96.1
UPDATED	97.4	97.43	97.43	94.87
CROSS-SPECIES	94.75	95.35	94.33	96.12
CONSERVED-HAIRPIN	93.08	93.57	93.65	90.58

methods of feature evaluation hardly consider the interaction between trees. They can be considered as computing the feature relevance for each tree and returning the average. The oob method does take into account the interaction between trees and while it does not optimize individual oob accuracy, it results in higher overall oob scores. Also, the area under the accuracy curve for the third method is larger than for the other two methods, meaning that it consistently finds better subsets of features, according to the criterion.

It is interesting to note that with the naive feature counting method, we have found that 27 features are sufficient for satisfactory classification. This is shown by the sharp drops in oob accuracy (from 93.6% to 93.29%) and also for the test accuracy (from 95.30% to 94.64%) after one more iteration of feature removal. As for the third method, no clear steep drop is seen from an iteration with test accuracy or oob accuracy comparable to the best score, but with 27 features, the test score was 95.07%, so it was still somewhat acceptable. These features are listed here: Z_{MFE} , $p(((, A(((, BP/L, ==, nAU/L, p(., PAS_{91}, dH, dS, pAub, tlen, TED_{43}, TED_7, Z_P, pA(((, G(., n(((, PAS_{37}, FREQ, n..., P_{MFE}, BTI_{14}$ (Weighted Balaban Index), $C(., looplen, -=, \text{ and } pUA$. (See Section II for the description of these features.)

We also tested our feature selection techniques on the Triplet-SVM datasets. For each method, the RFE feature selection algorithm was run on the training dataset and the best model, according to oob accuracy, was chosen. The results are listed in Table VIII. They indicate that there is a clear potential to feature selection, but more research is needed in this area, since no method offers a clear advantage over all other methods on most datasets.

IV. CONCLUSIONS AND FUTURE WORK

In this paper we have presented a miRNA de novo detection system that works with a comprehensive set of features derived from the literature. The system can be used in connection with a couple of machine learning classification and feature selection techniques. Comparisons results with existing miRNA identification systems have been presented. Future work will include the use of better, multivariate feature selection, testing certain improvements of the random forests algorithm, and the

addition of other specialized features for instance phylogenetic ones.

ACKNOWLEDGMENTS

We acknowledge the importation in our system of certain features from the miPred, microPred, yasMiR, Triplet-SVM, MiRFinder, mirEncoding system which were freely available.

REFERENCES

- [1] Rukshan Batuwita and Vasile Palade. microPred: effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics*, 25(8):989–995, 2009.
- [2] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.
- [3] George A. Călin and Carlo M. Croce. MicroRNA-cancer connection: The beginning of a new tale. *Cancer Res*, 66(15):7390–7394, 2006.
- [4] Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and Lotfi A. Zadeh. *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
- [5] Snorre Helvik, Ola Jr. Snøve, and Pål Sætrom. Reliable prediction of Drosha processing sites improves microRNA gene prediction. *Bioinformatics*, 23(2):142–149, 2007.
- [6] Ivo L. Hofacker. The Vienna RNA secondary structure server. *Nucleic Acids Res*, 31(13):3429–3431, 2003.
- [7] Ting-Hua Huang, Bin Fan, Max F. Rothschild, Zhi-Liang Hu, Kui Li, and Shu-Hong Zhao. MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans. *BMC Bioinformatics*, 8(341), 2007.
- [8] Peng Jiang, Haonan Wu, Wenkai Wang, Wei Ma, Xiao Sun, and Zuhong Lu. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Research*, 2007.
- [9] R. C. Lee, R. L. Feinbaum, and V. Ambros. The *c. elegans* heterochronic gene *lin-4* encodes small rnas with antisense complementarity to *lin-14*. *Cell*, 75(5):843–854, December 1993.
- [10] Yoontae Lee, Chiyoun Ahn, Jinju Han, Hyounjeong Choi, Jaekwang Kim, Jeongbin Yim, Junho Lee, Patrick Provost, Olof Radmark, Sunyoung Kim, and V. Narry Kim. The nuclear rnaase iii drosha initiates microRNA processing. *Nature*, 425(6956):415–419, September 2003.
- [11] John S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structures. *Biopolymers*, 29:1105–1119, 1990.
- [12] Kwang Loong Stanley Ng and Santosh Mishra. De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics*, 23(11):1321–1330, 2007.
- [13] Daniel Pasailă, Irina Mohorianu, Andrei Sucilă, Ștefan Panțiru, and Liviu Ciortuz. Yet Another SVM for MiRNA Recognition: yasMiR. Technical Report TR 10-01, “ALi.Cuza” University of Iași, Faculty of Computer Science, 2010. URL: <http://www.infoiasi.ro/tr/tr.pl.cgi>.
- [14] Daniel Pasailă, Irina Mohorianu, and Liviu Ciortuz. Using base pairing probabilities for MiRNA recognition. In *SYNASC '08: Proceedings of the 2008 10th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, pages 519–525, 2008.
- [15] Kim D. Pruitt and Donna R. Maglott. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Research*, 29(1):137–140, 2001.
- [16] Wenjie Shu, Xiaochen Bo, Zhiqiang Zheng, and Shengqi Wang. A novel representation of RNA secondary structure based on element-contact graphs. *BMC Bioinformatics*, 9(1):188, 2008.
- [17] Karol Szafranski, Molly Megraw, Martin Reczko, and Artemis Hatzigeorgiou. Support vector machines for predicting microRNA hairpins. In Hamid Arabnia and Homayoun Valafar, editors, *Proceedings of the 2006 International Conference on Bioinformatics & Computational Biology, BIOCAMP'06*, pages 270–276. CSREA Press, 2006.
- [18] Chenghai Xue, Fei Li, Tao He, Guoping Liu, Yanda Li, and Xuegong Zhang. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, 6(310), 2005.
- [19] Yun Zheng, Wynne Hsu, Mong-Li Lee, and Limsoon Wong. Exploring essential attributes for detecting microRNA precursors from background sequences. In Mehmet M. Dalkilic, Sun Kim, and Jiong Yang, editors, *2006 VDMB Workshop on Data Mining in Bioinformatics*, volume 4316 of *Lecture Notes in Computer Science*, pages 131–145. Springer, 2006.