Mature miRNA identification via the use of a Naive Bayes classifier

Gkirtzou Katerina, Tsakalides Panagiotis and Poirazi Panayiota

Abstract— **MicroRNAs (miRNAs) are small single stranded RNAs, on average 22 nt long, generated from endogenous hairpin–shaped transcripts with post–transcriptional activity. Although many computational methods are currently available for identifying miRNA genes in the genomes of various species, very few algorithms can accurately predict the functional part of the miRNA gene, namely the mature miRNA. We introduce a computational method that uses a Naive Bayes classifier to identify mature miRNA candidates based on sequence and secondary structure information of the miRNA precursor. Specifically, for each mature miRNA, we generate a set of negative examples of equal length on the respective precursor(s). The true and negative sets are then used to estimate probability distributions for sequence composition and secondary structure on each position along the RNA. The distance between these distributions is estimated using the symmetric Kullback-Leibler method. The positions at which the two distributions differ significantly and consistently over a 10-fold cross-validation procedure are used as features for training the Naive Bayes classifier. A total of 120 classifiers were trained with true positive and negative examples from human and mouse using a 10-fold cross-validation procedure. A performance of 76% sensitivity and 65% specificity was achieved using a consensus averaging. Our findings suggest that position specific sequence and structure information combined with a simple Bayes classifier achieve a good performance on the challenging task of mature miRNA identification.**

I. Introduction

MIcroRNAs (miRNAs) are small, non–coding RNAs that play an important role in regulating the expression of numerous genes across several species [1]. As regulatory molecules, they influence the output of many protein–coding genes by targeting mRNAs for cleavage or translational repression [2].

Although miRNAs are functionally similar to short interfering RNAs (siRNAs), they are unique in terms of their bio-genesis. Most of the miRNA genes are transcribed by RNA Polymerase II. The primary transcripts of miRNAs (pri–miRNAs) are then processed into hairpin intermediates (precursor miRNAs or pre–miRNAs) by the microprocessor complex (the enzyme Drosha and the binding protein DGCR8/Pasha). The pre-miRNAs are then exported to the cytoplasm by RanGTP and Exportin–5. In the cytoplasm, the pre-miRNAs are processed by Dicer into short RNA duplexes termed miRNA duplexes. The mature miRNA from the miRNA duplexes then binds to an Argonaute protein, forming the miRNP complex. The miRNAs base–pair with their mRNA targets, leading either to mRNA cleavage, if there is sufficient complementarity between miRNA and the target mRNA, or to translational repression [3].

Several computational methods have been developed and are currently used in parallel with experimental techniques in order to facilitate the discovery of new miRNAs. Most computational methods focus on the discovery of either novel miRNA genes in the genomes of various species or possible mRNA targets of the known miRNAs. On the contrary, few attempts have been made to computationally predict the functional part of the miRNA precursor, namely the mature miRNA. A number of studies ([4], [5], [6]) combine miRNA gene prediction with the identification of a possible start position for the mature. To our knowledge, only one study ([7]) focuses exclusively on mature miRNA prediction, utilizing thermodynamic and structural information.

In this work, we introduce a computational method that uses a Naive Bayes classifier to identify mature miRNA candidates based on sequence and secondary structure information of the miRNA precursor.

II. Method

Naive Bayes is a simple probabilistic classification method that assumes independence among the features of its input patterns. In this case, features are location-specific sequence and structure information derived from experimentally verified miRNA precursors. To better learn those features which are specific to mature miRNAs, we trained the classifier to discriminate between true mature samples and a set of negative examples.

A. Negative class

Given that known miRNA precursors do not produce multiple non–overlapping mature miRNAs from the same arm of the foldback precursor [8], we generated a set of negative examples in the following way: for each true mature miRNA, we use a same-size sliding window and select all possible 'negative' matures which can be created by sliding 1 base pair towards either direction from the mature. These procedure results in a very large negative set, where each true mature has a variable number of respective 'negatives', depending on the length and number of precursors. To avoid overfitting the classifier to the negative data, we only use a randomly selected subset of 10 negative examples for each true mature.

This work was supported by ICS–FORTH

K. Gkirtzou is with the Department of Computer Science, University of Crete and the Institute of Computer Science (ICS), FORTH , Heraklion, Greece gkirtzou@csd.uoc.gr

P. Tsakalides is with the Department of Computer Science, University of Crete and the Institute of Computer Science (ICS), FORTH , Heraklion, Greece tsakalid@csd.uoc.gr

P.Poirazi is with the Institute of Morecular Biology and Biotechnology (IMBB), FORTH, Heraklio, Greece poirazi@imbb.forth.gr

B. Feature Selection

The miRNA precursors form irregular hairpin structures, containing various mismatches, internal loops and bulges. In our method, a mature miRNA is represented as a sequence of positions along the respective precursor(s), where each position contains sequence information (A, C, U, G) or structural information (match or mismatch). This location-specific information is used to select a set of features, namely those positions on the precursor that contain discriminatory information between true matures and negative samples. The discriminatory power of each position is estimated using the symmetric Kullback–Leibler divergence between the distributions of positive and negative data.

The Kullback–Leibler divergence (K–L divergence) is a measure of the difference between two probability distributions [9]. For probability distributions P and Q of a discrete random variable the K–L divergence of Q from P is defined to be:

$$D_{KL}(P||Q) = \sum_i P(i) \log_2 \frac{P(i)}{Q(i)}$$

Unfortunately, the KL divergence is not a true metric since it is not symmetric. In order to overcome this problem we used the symmetric Kullback–Leibler divergence which is defined as:

$$\frac{1}{2}(D_{KL}(P||Q) + D_{KL}(Q||P))$$

which is symmetric and nonnegative [10]. This metric is commonly used for feature selection in classification problems, where P and Q are the conditional Probability Mass Functions (PMFs) of a feature in the two different classes.

III. Results

We have evaluated our method using a dataset of experimentally verified human and mouse miRNAs from miRBase (version 10.0 , [11], [12], [13]). The human dataset consists of 533 precursors and 722 mature miRNAs, while the mouse dataset consists of 442 precursors and 579 mature miRNAs. We used 500 of the human precursors with their 692 respective mature and 347 of the mouse precursors with their 440 respective matures to train and validate our classifiers utilizing a leave-10-out cross validation procedure.

For each of the mature miRNAs in the training set, a negative set was generated as described in section II-A. A total of 150 classifiers were trained and the classification performance was assessed using consensus averaging. It is worth mentioning that in order to have a realistic estimation of the accuracy of the classifiers, the validation sets consisted of all potential mature miRNAs of size 22nt that could be produced by the precursors, whose mature miRNAs were left out from the training phase in the cross validation procedure.

Tables I, II and III show the top scoring classifiers, based on Matthews Correlation Coefficient (MCC), for different input features. We use three types of classifiers, each utilizing location-specific information about the sequence (table I), the structure (table II), and both sequence and structure (table III) of the training examples. Each table shows the sensitivity, specificity and Matthews Correlation Coefficient (MCC) [14] achieved with different numbers of such features and with different sizes of flanking regions around the mature miRNA. The positions along the precursor which served as input features were selected based on the K–L divergence metric and they were located either within the mature miRNA, or inside a flanking region around it.

TABLE I

CLASSIFIER TRAINED WITH FEATURES OF SEQUENCE INFORMATION.

Classifier's Description	Sensitivity	Specificity	MCC
Combination of 12 Features, 0nt flanking region	67.1%	55.10%	0.0850
Combination of 16 Features, 5nt flanking region	76.04%	53.34%	0.1074
Combination of 31 Features, 7nt flanking region	75.96%	53.20%	0.1071
Combination of 19 Features, 10nt flanking region	79.15%	47.01%	0.0960
Combination of 35 Features, 12nt flanking region	74.30%	51.33%	0.0945

TABLE II

CLASSIFIER TRAINED WITH FEATURES OF STRUCTURE INFORMATION.

Classifier's Description	Sensitivity	Specificity	MCC
Combination of 10 Features, 0nt flanking region	65.70%	54.30%	0.0730
Combination of 26 Features, 5nt flanking region	76.34%	52.64%	0.1056
Combination of 23 Features, 7nt flanking region	77.85%	54.29%	0.1186
Combination of 39 Features, 10nt flanking region	81.01%	56.63%	0.1373
Combination of 38 Features. 12nt flanking region	79.89%	55.51%	0.1300

We found that as the size of the flanking region increased, the sensitivity of the classifiers tended to improve, while the specificity remained relatively unaffected, independently of the type of features used in the classifier. This is not the case , though, for the classifiers with flanking region of size 12nt with features either sequence or structure alone (tables I and II respectively), where the extra features probably add more noise than useful information.

Moreover, the classifiers utilizing features with combined information for both sequence and structure achieved an overall better performance -in terms of improved specificity and MCC- than the ones using sequence or structure information alone. This is very important to have a high specificity score in this task, since the number of negative examples is higher than the number of positive ones, as it is also reflected in the MCC. Finally, all classifiers achieved a much higher sensitivity than specificity score, most likely because of the very high similarity between negative and positive examples.

IV. Conclusions

In this paper, we present a computational approach that identifies mature miRNAs based on the secondary structure and sequence of the precursor. We have used experimentaly

TABLE III

CLASSIFIER TRAINED WITH FEATURES OF SEQUENSE AND STRUCTURE INFORMATION.

Classifier's Description	Sensitivity	Specificity	MCC
Combination of 20 Features, 0nt flanking region	68.50%	62.50%	0.1250
Combination of 29 Features, 5nt flanking region	71.32%	65.34%	0.1394
Combination of 36 Features, 7nt flanking region	74.26%	66.46%	0.1562
Combination of 42 Features, 10nt flanking region	76.50%	65.61%	0.1606
Combination of 39 Features, 12nt flanking region	77.81%	64.14%	0.1590

verified miRNAs to train and evaluate the performance of a Naive Bayes Classifier in terms of Sensitivity and Specificity.

Most of the methods that have been made to computationally predict the functional part of the miRNA precursor calculate their performance in terms of true positive rate only, ignoring the false positive rate. It is a matter of semantics and a great challenge what is consider to be a negative example, but the major issue in such a task is to minimize the false positive rate.

In conclusion, our findings suggest that position specific sequence and structure information combined with a simple Bayes classifier achieve a good performance on this challenging task.

REFERENCES

[1] L. He and G. J. Hannon, "Micrornas: small rnas with a big role in gene regulation," *Nature Genetics*, vol. 5, pp. 522–532, 2004.

[2] D. P.Bartel, "Micrornas: Genomics,biogenesis, mechanism, and function," *Cell*, vol. 116, pp. 281–297, 2004.

[3] X. Liu, K. Fortin, and Z. Mourelatos, "Micrornas: Biogenesis and molecular functions," *Brain Pathology*, vol. 18, no. 1, pp. 113–121, 2008.

[4] J.-W. Nam, K.-R. Shin, J. Han, Y. Lee, V. N. Kim, and B.-T. Zhang, "Human microrna prediction through a probabilistic co-learning model of sequence and structure," *Nucleic Acids Research*, vol. 33, no. 11, pp. 3570–3581, 2005.

[5] M. Yousef, M. Nebozhyn, H. Shatkay, S. Kanterakis, L. C. Showe, and M. K. Showe, "Combining multi-species genomic data for microrna identification using a naive bayes classifier," *Bioinformatics*, vol. 22, no. 11, pp. 1325–1334, 2006.

[6] Y. Sheng, P. G. Engstrom, and B. Lenhard, "Mammalian microrna prediction through a support vector machine model of sequence and structure," *PloS ONE*, vol. 2, no. 9, 2007.

[7] M. Tao, "Thermodynamic and structural consensus principle predicts mature mirna location and structure, categorizes conserved interspecies mirna subgroups and hints new possible mechanisms of mirna maturization," Control and Dynamical Systems, California Institute of Technology, Tech. Rep., 2007. [Online]. Available: http://www.citebase.org/abstract?id=oai:arXiv.org:0710.4181

[8] V. Ambros, B. Bartel, D. P. Bartel, C. B. Burge, J. C.Carrington, X. Chen, G. Dreyfuss, S. R. Eddy, S. Griffiths-Jones, M. Marshall, M. Matzke, G. Ruvkun, and T. Tuschl, "Rna a uniform system for microrna annotation," *RNA*, vol. 9, pp. 277–279, 2003.

[9] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[10] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, vol. 186, no. 1007, pp. 453–461, 1946.

[11] G. S. Jones, "The microrna registry," *Nucleic Acids Res*, vol. 32, no. Database issue, pp. D109–D111, Jan 2004.

[12] S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright, "mirbase: microrna sequences, targets and gene nomenclature." *Nucleic Acids Res*, vol. 34, no. Database issue, January 2006. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/16381832

[13] S. Griffiths-Jones, H. K. K. Saini, S. v. V. Dongen, and A. J. J. Enright, "mirbase: tools for microrna genomics." *Nucleic Acids Res*, November 2007. [Online]. Available: http://dx.doi.org/10.1093/nar/gkm952

[14] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 18, no. 5, pp. 412–424, 2000.