Mature miRNA identification via the use of a Naive Bayes classifier

Katerina Gkirtzou, Panagiotis Tsakalides and Panayiota Poirazi

***Abstract*— MicroRNAs (miRNAs) are small single stranded RNAs, on average 22nt 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 metric. 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 15 classifiers were trained with true positive and negative examples from human and mouse. A performance of 76% sensitivity and 65% specificity was achieved using a consensus averaging over a 10-fold cross-validation procedure. 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 biogenesis. 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 each miRNA duplex 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 of the precursor RNA.

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

A. Naive Bayes Classifier

Naive Bayes is a simple probabilistic classifier which is based on the application of the Bayesian theorem with strong (naive) independence assumptions. According to the Bayesian classifier, a new sample $\mathbf{x}$ described by the feature vector $\mathbf{x} = \langle x_1, x_2, \ldots, x_n \rangle$ will be assigned to the class that minimizes the overall risk using the following formula:

$$\alpha(\mathbf{x}) = argmin_{\alpha_i \in A} \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|\mathbf{x})$$

where:

- $\omega_1, \ldots, \omega_c$ is a finite set of classes,
- $A = \{\alpha_1, \ldots, \alpha_c\}$ is a finite set of actions, where α_i means selecting class ω_i,
- $\lambda(\alpha_i|\omega_j)$ is the loss associated with deciding ω_i, when the true state of nature is ω_j and
- $P(\omega_j|\mathbf{x})$ is the posterior probability of ω_j being the true state of nature given $\mathbf{x}$.

The posterior probability $P(\omega_j|\mathbf{x})$ can be computed by the Bayes' formula (see section 2.9 of [8]):

This work was supported by FORTH–ICS and the EMBO Young Investigator Program

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

P. Poirazi is with the Institute of Molecular Biology and Biotechnology (IMBB), Foundation of Research and Technology, Hellas (FORTH), Heraklion, Greece poirazi@imbb.forth.gr

$$P(\omega_j|\mathbf{x}) = \frac{P(\mathbf{x}|\omega_j)P(\omega_j)}{P(\mathbf{x})},$$

where

- $P(\mathbf{x}|\omega_j)$ is the state–conditional probability for $\mathbf{x}$ conditioned on ω_j being the true class,
- $P(\omega_j)$ is the prior probability that nature is in state ω_j and
- $P(\mathbf{x}) = \sum_{j=1}^{c} P(\mathbf{x}|\omega_j)P(\omega_j)$ is the evidence for $\mathbf{x}$.

The Naive Bayes classifier is based on the simplifying assumption that the input features among samples of any given class are conditionally independent given the class [9]. In other words, given the class of a sample, the probability of observing the conjuction $x_1, x_2, \ldots, x_n$ is just the product of the probabilities for the individual features of this sample:

$$P(x_1, x_2, \ldots, x_n|c_j) = \prod_{i}^{n} P(x_i|c_j).$$

In our case, an observation for classification (i.e. a sample) is a mature miRNA candidate and the possible classes are two: the Positive class, which contains true mature miRNAs (denoted ω_1) and the Negative class, which contains false mature miRNAs (denoted ω_{-1}). Suppose we have a mature miRNA candidate with features $\mathbf{x} = \langle x_1, x_2, \ldots, x_n \rangle$ and we want to classify it to the class that minimizes the classification error. The simplest case is to consider that all errors have the same cost, so the loss function of interest is the zero–one loss function and the Bayes Decision Rule is converted to the following:

$$\textit{Decide } \omega_1 \textit{ if } P(\omega_1|\mathbf{x}) > P(\omega_{-1}|\mathbf{x})\textit{;}$$
$$\textit{otherwise decide } \omega_{-1}$$

(see section 2.3 of [8]). Since $P(\mathbf{x})$ is only a normalization factor, it can be omitted in order to minimize calculation time. Moreover, since there is no information about the probabilities of each class we can assume that the prior probability that nature is in state ω_j, $P(\omega_j)$, is 50% for both positive and negative data. This assumption prevents us from favoring a particular class. Under these assumptions, the Bayes Decision Rule is given by the following simplified formula:

$$\textit{Decide } \omega_1 \textit{ if } P(\mathbf{x}|\omega_1) > P(\mathbf{x}|\omega_{-1})\textit{;}$$
$$\textit{othewise decide } \omega_{-1}$$

In this work, we use the aforementioned formula to build multiple classifiers that are trained to discriminate between randomly selected sets of positive and negative miRNA samples as detailed below.

B. The Negative class

Given that known miRNA precursors do not produce multiple overlapping mature miRNAs from the same arm of the foldback precursor [10], we generate 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, excluding any hairpin loops. This 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 classifiers to the negative data, we only use a randomly selected subset of 10 negative examples for each true mature.

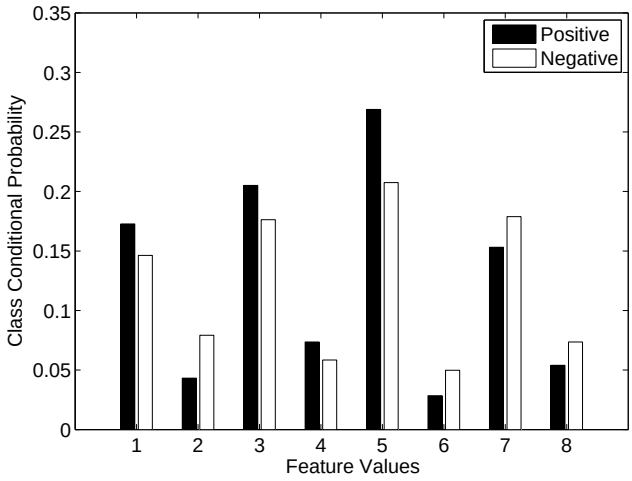

Fig. 1. Class conditional probabilities of a feature with combined sequence and structural information for position 0 within mature miRNA candidates of both the positive (black color) and the negative class (white color). The x axis shows the possible values of the feature, where: 1 → A match, 2 → A mismatch, 3 → C match, 4 → C mismatch, 5 → G match, 6 → G mismatch, 7 → U match and 8 → U mismatch. The y axis shows the class conditional probability for this feature.

C. Input Features

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, where each position contains sequence information (A, C, U, G) or structural information (match or mismatch), derived from the respective precursor(s). Apart from the features that lie in positions within the mature miRNA, we also consider features that lie within a flanking region of variable size (0, 5, 7, 10 or 12nt) that extends symmetrically along both sides of the mature sample. These features are also positions on the precursor RNA and contain sequence or structural information just like the features located within the mature miRNA. Since certain positions within the flanking region may be located outside the precursor, we use a special 'novalue' flag to indicate the lack of information at these positions and do not take them into account when estimating the Kullback–Leibler divergence between the two classes (see below).

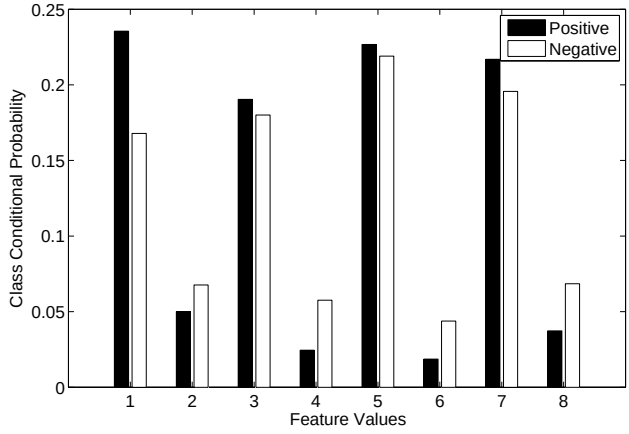

Fig. 2. Class conditional probabilities of a feature with combined sequence and structural information for position 3 within mature miRNA candidates of both the positive (black color) and the negative class (white color). The x axis shows the possible values of the feature, where: 1 → A match, 2 → A mismatch, 3 → C match, 4 → C mismatch, 5 → G match, 6 → G mismatch, 7 → U match and 8 → U mismatch, while the y axis shows the class conditional probability for this feature.

D. Feature Selection

The aforementioned 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 [11]. For Probability Mass Functions (PMFs) P and Q of a discrete random variable, the K–L divergence of Q from P is defined as:

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

Unfortunately, the KL divergence is not a true metric since it is not symmetric. To overcome this problem we used the symmetric and nonnegative Kullback–Leibler divergence [12], which is defined as:

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

and is commonly used in classification problems. Figures 1 and 2 show the class conditional probability distributions for two features with combined sequence and structure information for positions 0 and 3 within the mature miRNA candidates of the negative and positive data used to train our classifiers (see below). These features have the highest Kullback–Leibler divergence among all features that lie within the mature miRNA. As evident from both figures the two class distributions are very similar making discrimination a very challenging task.

III. RESULTS

We evaluated our method using a dataset of experimentally verified human and mouse miRNAs from miRBase (version 10.0, [13], [14], [15]). 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 and 347 of the mouse precursors -which generate 692 and 440 mature miRNAs, respectively- 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-B. A total of 15 classifiers were trained with different feature sets and the classification performance was assessed using consensus averaging over a 10–fold cross validation. To ensure a realistic estimation of classification accuracy, the validation sets consisted of true miRNA precursors, whose mature miRNAs were left out from the training sets in the cross validation procedure. Classification accuracy was estimated using a fixed 22nt size sliding window. All possible matures which could be created by sliding 1 base pair in both stem arms of the precursor, apart from the hairpin loop(s), were assigned to either class based on the averaged outcome of the 15 trained classifiers. It is important to note that classification accuracy was estimated based on exact match of the starting position of the predicted compared to the real mature miRNA. Even 1nt deviations were considered as negative examples.

TABLE I
BAYES CLASSIFIER TRAINED WITH FEATURES CONTAINING SEQUENCE INFORMATION.

Classifier's Description	**Sensitivity**	**Specificity**	**MCC**
Combination of 12 Features, 0nt flanking region	67.10%	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

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), or both the sequence and structure (Table III) of the training examples. Each table shows the sensitivity, specificity and Matthews Correlation Coefficient (MCC) [16] achieved with different numbers of such features and with different sizes of flanking regions around the

TABLE II

BAYES CLASSIFIER TRAINED WITH FEATURES CONTAINING 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

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.

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. This improvement seemed to reach a maximum for a flanking region of about 10nt. For classifiers with flanking regions of 12nt utilizing either sequence or structure information (Tables I and II respectively), the extra features did not further improve the accuracy, suggesting that they probably add more noise than useful information.

TABLE III

BAYES CLASSIFIER TRAINED WITH FEATURES CONTAINING BOTH SEQUENCE 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

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. Note that a high specificity score is particularly important in this task, since the number of negative examples is much larger than the number of positive ones. Finally, all classifiers achieved a much higher sensitivity than specificity score, most likely because of the very high similarity between negative and positive examples as well as the requirement for exact start position match between true and predicted miRNAs.

IV. CONCLUSIONS

In this work, we presented a computational approach that identifies mature miRNAs based on the secondary structure and sequence characteristics of the precursor. We used experimentally verified miRNAs to train and evaluate the performance of a Naive Bayes classifier in terms of sensitivity and specificity.

Unlike the method presented here, most of the computational tools that can be used to predict the functional part of the miRNA precursor estimate their performance accuracy in terms of true positive rate alone (sensitivity), ignoring the false positive rate ([4], [6], [7]). It is a matter of semantics as well as a great challenge to define a true negative example when it comes to mature miRNAs. However, a major issue in such a classification task is not only to maximize the identification of true positives but also to minimize the false positive rate. Our method tries to combine both of these criteria. Since minimizing the false positive rate is very difficult, we plan to develop and incorporate a number of filtering criteria that will help eliminate some of the false positive examples. Moreover, we plan to use our method to predict the mature miRNAs on six novel miRNA genes which were recently identified by Hidden Markov Models in our lab and were experimentally shown to produce short RNAs [17]. This combined computational and experimental approach will result in a more realistic evaluation of our method's performance accuracy.

In conclusion, 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.

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] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2nd Edition)*. Wiley-Interscience, November 2000.

[9] T. M. Mitchell, *Machine Learning*. McGraw-Hill International, 1997.

[10] 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.

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

[12] 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.

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

[14] 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

[15] 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

[16] 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.

[17] A. Oulas, A. Boutla, K. Gkirtzou, M. Reczko, K. Kalantidis, and P. Poirazi, "Prediction of novel microRNA genes for cancer associated genomic regions a combined computational and experimental approach," in preperation.