BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 11 2006, pages 1325–1334
doi:10.1093/bioinformatics/btl094

Sequence analysis

Combining multi-species genomic data for microRNA identification using a Naïve Bayes classifier

Malik Yousef[1], Michael Nebozhyn[1], Hagit Shatkay[2], Stathis Kanterakis[1], Louise C. Showe[1] and Michael K. Showe[1,*]

[1]The Wistar Institute, Philadelphia, PA 19104, USA and [2]School of Computing, Queen's University, Kingston, Ontario, Canada

Received on December 1, 2005; revised on February 21, 2006; accepted on March 9, 2006
Advance Access publication March 16, 2006
Associate Editor: Keith A Crandall

ABSTRACT

Motivation: Most computational methodologies for microRNA gene prediction utilize techniques based on sequence conservation and/or structural similarity. In this study we describe a new technique, which is applicable across several species, for predicting miRNA genes. This technique is based on machine learning, using the Naïve Bayes classifier. It automatically generates a model from the training data, which consists of sequence and structure information of known miRNAs from a variety of species.

Results: Our study shows that the application of machine learning techniques, along with the integration of data from multiple species is a useful and general approach for miRNA gene prediction. Based on our experiments, we believe that this new technique is applicable to an extensive range of eukaryotes' genomes. Specific structure and sequence features are first used to identify miRNAs followed by a comparative analysis to decrease the number of false positives (FPs). The resulting algorithm exhibits higher specificity and similar sensitivity compared to currently used algorithms that rely on conserved genomic regions to decrease the rate of FPs.

Availability: The BayesMiRNAfind program is available at http://wotan.wistar.upenn.edu/miRNA

Contact: showe@wistar.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

INTRODUCTION

MicroRNAs (miRNAs) are single-stranded, non-coding RNAs averaging 21 nt in length. The mature miRNA is cleaved from a 70–110 nt 'hairpin' precursor with a double-stranded region containing one or more single-stranded loops. MiRNAs target messenger RNAs (mRNAs) for cleavage, repressing translation and causing nascent protein degradation (Bartel, 2004).

Several computational approaches have been implemented for miRNA gene prediction using methods based on sequence conservation and/or structural similarity (Lim *et al*., 2003a, b; Weber, 2005; Lai *et al*., 2003; Grad *et al*., 2003). Lim and others (Lim *et al*., 2003a, b; Weber, 2005) developed a program, for identification of miRNAs, called MiRscan with a 70% specificity at a sensitivity of 50%. MiRscan uses seven miRNA features with associated weights to build a computational tool, which assigns scores to hairpin candidates. The weights are estimated using statistics based on the previously known miRNAs from *Caenorhabditis elegans*. Grad *et al*. (2003) developed a computational method using sequence conservation and structural similarity to predict miRNAs in the *C.elegans* genome. Lai *et al*. (2003) used similar ideas to develop a different computational tool for the Drosophila genome, called miRseeker. These efforts have recently been reviewed by Bartel (2004). Others used homology searches for revealing paralog and ortholog miRNAs (Weber, 2005; Lagos-Quintana, 2001; Lau *et al*., 2001; Lee and Ambros, 2001; Pasquinelli *et al*., 2000). In addition, Wang *et al*. (2005) developed a method based on sequence and structure alignment for miRNA identification. The most recent published work of which we are aware that uses machine learning for miRNA discovery is by Nam *et al*. (2005). They constructed a highly specific probabilistic model (HMM) whose topology and states are handcrafted based on prior knowledge and assumptions, and whose exact probabilities are derived from the data.

In our study we present a machine learning approach based on the Naïve Bayes classifier for predicting miRNA genes. Our method differs from previous efforts in two ways: (1) we generate the model automatically and identify rules based on the miRNA gene structure and sequence allowing prediction of non-conserved miRNAs and (2) we use a comparative analysis over multiple species to reduce the false positive (FP) rate. This allows for a trade-off between sensitivity and specificity. Based on our experiments with multiple genomes we believe that our method is applicable to a wide variety of eukaryotes. The resulting algorithm demonstrates higher specificity and similar sensitivity compared to currently used algorithms, which use conserved genomic regions to reduce FPs (Lim *et al*., 2003a, b; Lai *et al*., 2003; Grad *et al*., 2003).

Like Nam *et al*. (2005), rather than relying on miRNAs homology between related species, we directly use features of the miRNA sequence and secondary structure. However in contrast to them, we train a Naïve Bayes classifier to identify miRNAs directly from the data. In our system prior knowledge is used for initial filtering of the data, but not for constructing the model. The Naive Bayes classifier is a standard model with no domain-specific assumptions (aside for the usual conditional independence assumptions inherent to the model). In addition, whereas Nam's model was trained and tested on a single type of data (136 Human miRNAs)

*To whom correspondence should be addressed.

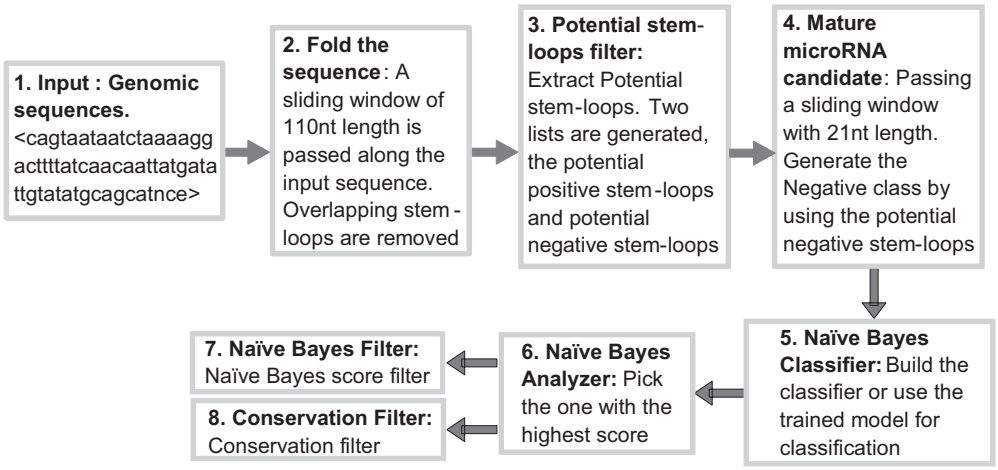

Fig. 1. The computation procedure components. This figure illustrates the pipeline of the computational procedure with eight (8) stages. Its main component is the Naïve Bayes algorithm.

with respect to a restricted set of negative examples, in our study, we trained and tested the model using a variety of miRNAs from multiple organisms. We note that the dataset used by Nam is strictly a subset of the data used in our work. We demonstrate here that the results obtained by training and testing, using arbitrarily selected numbers of negative samples are highly sensitive to the size of the negative set. We overcome this problem by using multiple sources of miRNA. The major novelty of the work presented here is the combination of data from multiple sources. By integrating data from multiple species, we stabilize the learning process and, more importantly, construct a model that is more likely to be applicable to a variety of genomes.

Our results, as presented in this paper, demonstrate the specificity of the algorithm for miRNA gene prediction and suggest that our approach is comparable to, and even exceeds the performance of other methods. The results also demonstrate the applicability of machine learning methods for identifying common features in DNA sequence.

MATERIALS AND METHODS

Reduction of the search space to stem–loops with general structural features and generation of the negative class

A machine learning training procedure for building a classifier typically requires positive and negative examples. In our case it is clear that the known miRNAs serve as the positive examples. However, it is more difficult to decide which are the best negative examples for the training stage. To produce high specificity in the selection of new candidate miRNAs, the negative examples should be highly similar to the miRNA themselves. Various techniques could be considered including (1) generating negative examples by permuting the original known miRNA sequences, (2) randomly selecting genomic sub-sequences from the whole genome where the relative frequency of miRNAs is very low, (3) using samples from the mRNA 3′-untranslated region (3′-UTR) region (Mignone *et al*., 2005). There is only one predicted miRNA sequence in the 159 000 reported 3′-UTR sequences in UTRdb at http://www.ba.itb.cnr.it/UTR/. Only 35% of mammalian miRNAs overlap annotated genes and 90% of these are intronic (Griffiths-Jones *et al*., 2006) In addition, of the 114 *C.elegans* miRNAs annotated in Rfam (http://microrna.sanger.ac.uk/cgi-bin/sequences/) none is located in a known 3′-UTR and only four are described as exonic (3′-UTRs are in exons) while 97 are listed as intergenic or intronic. We conclude that miRNAs are extremely rare in 3′-UTRs and thus the third choice for generating a filter that is later used to generate a negative class. These samples closely resemble miRNAs except in particular features that we select.

Since miRNAs are highly conserved between species, we draw our negative class sequences from highly conserved regions, as defined by homology using the BLAT sequence alignment (Kent, 2002; Kent *et al*., 2002). Beginning with these highly conserved sequences, we pass the sequences through the 110 base sliding window, which filters for the secondary structural features that define a typical miRNA (Fig. 1, Steps 1–4). Candidate sequences which fail only one of the four structural features (stem length; number of paired bases; hairpin length and free energy) are retained for the class of negative examples.

The computational procedure (BayesmiRNAfind)

Figure 1 illustrates the pipeline of the computational procedure. Its main component is the Naïve Bayes algorithm (Mitchell, 1997).

The input (First stage) is a genomic sequence of any length. It could be the whole genome or a sub-sequence. This is different from most existing methods, which start with only conserved genomic segments (Lim *et al*., 2003a, b; Weber, 2005; Lai *et al*., 2003). Use of pre-aligned genomic sequences (conserved segments) reduces the search space, at the cost of making the algorithm less general. Although most miRNAs are phylogenetically conserved, putative miRNAs that are not conserved between species will not be predicted by methods that use only conserved regions of the genome (Lim *et al*., 2003b; Lai *et al*., 2003; Grad *et al*., 2003; Wang *et al*., 2005).

MiRNAs are processed from a precursor that forms a stem–loop structure. In the second stage of our process, the mfold program (Mathews *et al*., 1999; Zuker, 2003) is used for the stem–loop predictions. A 110 nt sliding window is moved along the input sequence and mfold generates the secondary structure. Stem–loops that overlap are removed and the one candidate with the lowest free energy is kept for the next step. The parameters used for the prediction of potential stem–loops are based on the biogenesis criteria mentioned in Ambros *et al*. (2003) and on our own analyses as shown in Figures 3 and 4. We used two sequence sets to extract potential stem–loop criteria: the first consisted of the miRNAs from *Drosophila melanogaster*, *C.elegans*, *Caenorhabditis briggsae*, *Homo sapiens*, *Mus musculus* and

Table 1. Dataset of discovered miRNAs

Species name	#Mature
Drosophila melanogaster	79
Drosophila pseudoobscura	68
Caenorhabditis elegans	117
Caenorhabditis briggsae	78
Gallus gallus	101
Homo sapiens	207
Mus musculus	212
Rattus norvegicus	177
Zea Mays	40
Oryza sativa	134
Arabidopsis thaliana	114
Danio rerio	26
Epstein Barr virus	6
Total	1359

Rattus norvegicus as shown in Table 1. The second, a negative set, consists of random sequences from *Human* mRNA 3′UTR regions (Mignone *et al*., 2005). The first and the second stages of the computational procedure were applied to both sets to generate the secondary structure and folding, yielding 719 miRNA stem–loops and 190 739 non-miRNA stem–loops.

In the next (Third) stage, a filter is applied to reduce the number of potential stem–loops by discarding unlikely candidates. Four features are considered for building a filter that can distinguish potential stem–loops: (1) Stem length, (2) Folding free energy, (3) Base pairs and (4) Loop length (Fig. 2). For each of these features a statistical histogram plot was generated as shown in Figure 3. It is clear that none of them is a stand-alone criterion for potential stem–loop predictions. Therefore, a combination of these features has been used to generate a filter requiring stem–loops to satisfy the following criteria:

(1) 42–85 nt stem length (the number of nucleotide in the upper stem plus the lower stem arm).

(2) Atleast ∼25 kcal/mole of folding free energy.

(3) Loop length <26 nt.

(4) 16–45 base pairs (bp).

Stem–loops that satisfy all four criteria (Fig. 4) are retained for the next stage of analysis. Expressed sequences that fail to satisfy this filter form a pool from which negative examples are drawn for later use when training the classifier. Figure 4 shows that ∼80% of the miRNAs satisfy all four conditions and only ∼6% of the non-miRNAs. One could use stricter criteria that may decrease the sensitivity and increase the specificity, or alternatively use recomputed criteria based on organism-specific miRNA. For validation of these four criteria for true miRNA stem–loops, a different set of negative examples was used. Moreover, when we split the data using the first half to train and the second half to test, similar results were observed.

So far, we have shown the way we built the potential stem–loops to be used later for generating the negative class and for eliminating sequences that are unlikely to contain potential miRNA genes. We next describe the features that are extracted from the sequence and structure of the miRNA and the non-miRNA examples for training the Naïve Bayes classifier.

Definition of structural and sequence features

For the positive (miRNA) class, the 21 nt of the mature miRNA are mapped into its associated stem–loop [generated by the mfold program (Zuker, 2003)] and then the features are generated as described below. For the negative (non-miRNA) class, we use a 21 nt sliding window (Fourth stage) to represent the so-called mature miRNA candidate. Features are then extracted for each 21 nt window. Our main assumption is that each true hairpin precursor contains one mature miRNA located in one of its arms.

For a given 21 nt miRNA candidate, 62 secondary structure features are derived from the hairpin (stem–loop), and 12 sequence-based features (which we call 'words') are extracted from the candidate sequence. Next, the hairpin (stem–loop) is split into three parts: foot, mature and head as shown in Figure 10. For each of these parts the following features are extracted:

(1) The number of base pairs.

(2) The number of bulges.

(3) The number of loops

(4) The number of asymmetric loops.

(5) Eight additional features represent the number of bulges of lengths 1–7 and those with lengths >7.

(6) Another set of eight features represents the number of symmetric loops with lengths 1–7 and the eighth one representing those that have lengths >7.

(7) The distance of the start of the mature miRNA candidate from the first paired base of the foot and head part are two additional features that are extracted.

(8) Nucleotide sequence 'words' with lengths 4–9 are extracted from the candidate 21 nt sequence and from the reverse sequence. These 'motif' features are not fixed and influence the dimension of the vector space. The dimension of that vector is determined, at a later point, to be 62 plus the number of unique 'words' that are obtained. It could include thousands of features. Those features are a result of the criteria mentioned by Lim *et al*. (2003a, b), Weber (2005), Wang *et al*. (2005), Ambros *et al*. (2003) and recently summarized by Bartel (2004).

The weighted combination of these features is used for generating a model that describes the miRNA class. This model is then utilized for predicting novel miRNAs. Some of the generated features might be irrelevant to a particular model. However, since the machine learning algorithm is able to learn with noisy features and because a particular feature could be important to a specific genome or miRNA class, we decided to provide as much information as possible to the machine learning algorithm. This way, we can generate a model that may identify biologically relevant structures that were hitherto undiscovered. One could elect to reduce the number of features, using feature selection techniques (Sahami *et al*., 1996), and still obtain reasonable accuracy. An analysis of the most important features is shown in the Supplementary information. In this study all of the generated features were used for training the classifier.

Further performance analysis has been applied to evaluate the importance of the structural features and the sequence features separately; an accuracy increase of ∼10% has been observed when the two kinds of features are combined.

Stage 5, Naïve Bayes classifier

Naïve Bayes is a classification model obtained by applying a relatively simple method to a training dataset (Mitchell, 1997). A Naïve Bayes classifier calculates the probability that a given instance (example) belongs to a certain class. It makes the simplifying assumption that the features constituting the instance are conditionally independent given the class. In practice, Naïve Bayes often performs well, considering its simple structure and ease of implementation. Given an example X, described by its feature vector $(x_1, \ldots, x_n)$, we are looking for a class C that maximizes the likelihood: $P(X \mid C) = P(x_1, \ldots, x_n \mid C)$. The (Naïve Bayes) assumption of conditional independence among the features, given the class, allows us to express this conditional probability $P(X \mid C)$ as a product of simpler probabilities:$P(X \mid C) = \prod_{i=1}^{n} P(x_i \mid C)$.

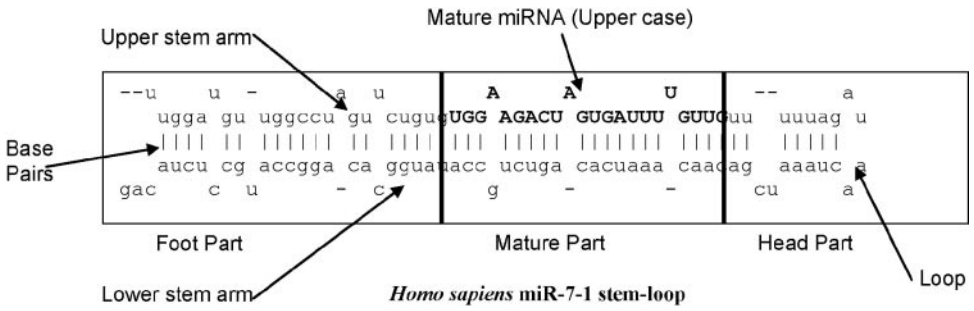

Fig. 2. Partition of stem–loop into three parts: foot, mature and head and features to determine potential stem–loops. For a given 21 nt miRNA candidate, different secondary structure features are derived from each part.

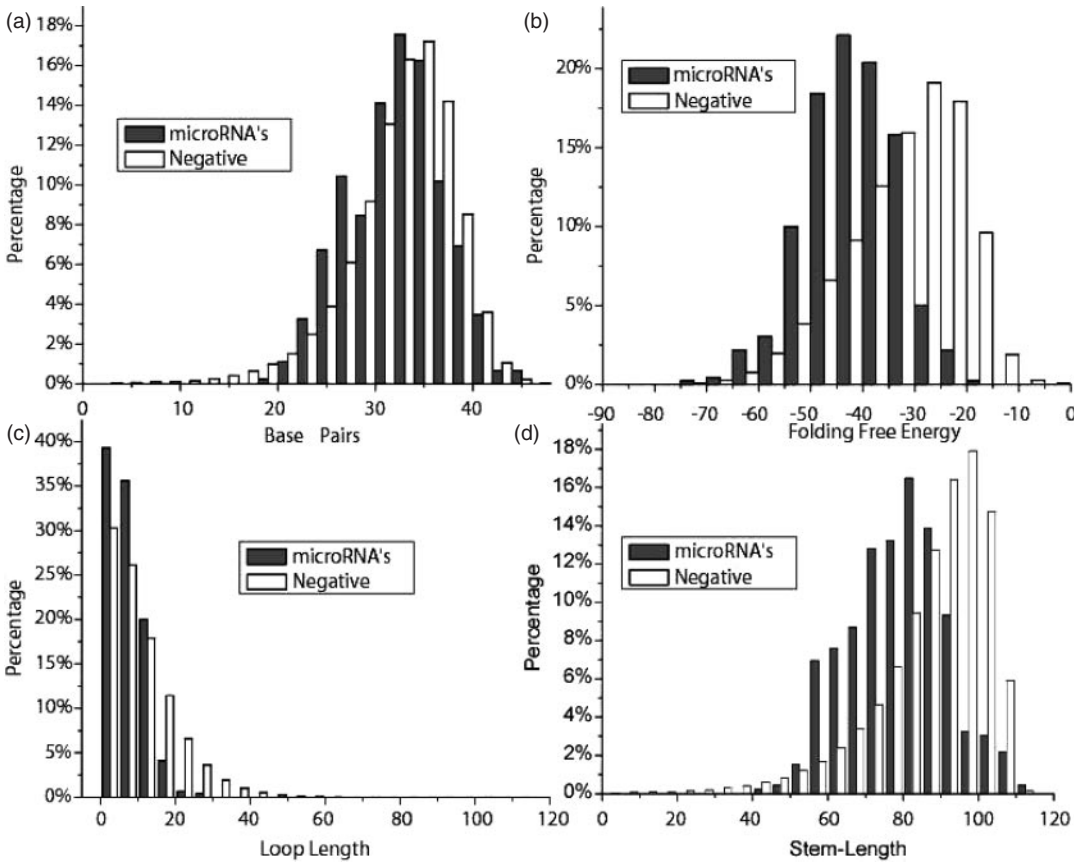

Fig. 3. Distribution of miRNAs and non-miRNAs for each feature of the potential stem–loop filter is shown in statistical histogram plots. Four features are considered for building a filter that can distinguish potential stem–loop: (**a**) number of base pairs, (**b**) folding free energy, (**c**) loop length and (**d**) stem length.

We used the Rainbow (McCallum, 1996, http://www.cs.cmu.edu/mccallum/bow) program to train the Naïve Bayes classifier. To combine the numeric features identified in the stem–loops and the sequence features ('words') in the miRNA candidate sequence, a dictionary of all the 'words' was generated and the frequency of each 'word' in the dictionary is used.

Training the classifier

A challenge in training the Naïve Bayes classifier (Fifth stage) is the imbalance between the number of positive and negative examples in the training data. An unbalanced distribution of the data could result in poor performance in the classification task. Previous studies in machine learning suggest methods that can be applied to deal with this kind of dataset (Japkowicz and Stephen, 2002).

Initially, we carefully chose the proportion of positive and negative examples to include in the training set and considered miRNAs derived from only one genome. This under-sampling approach was used to determine the appropriate proportions of positive and negative examples to include. In the next section, we demonstrate that combining all known miRNAs from a variety of species actually improves the performance of the classifier, making the performance more robust to changes in the proportion of positive and negative examples in the training set. We thus include all known miRNAs

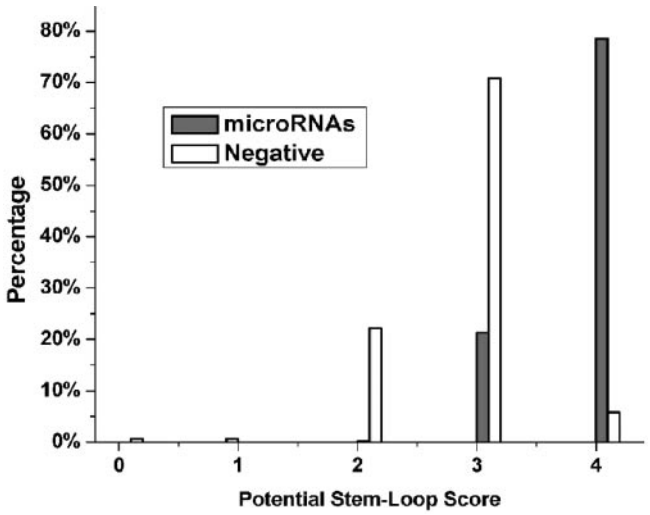

Fig. 4. The histogram shows the distribution of miRNAs and non-miRNAs for the combined four features characterizing potential stem-loops (stem length; folding free energy; loop length; base pairs) for potential stem–loop predictions. A combination of these features has been used to generate a filter requiring stem–loops to satisfy the following criteria: (1) stem length: 42–85 nt (the number of nt in the upper stem plus the lower stem arm); (2) folding free energy: at least ∼25 kcal/mol (3) loop length <26 nt and (4) 16–45 bp.

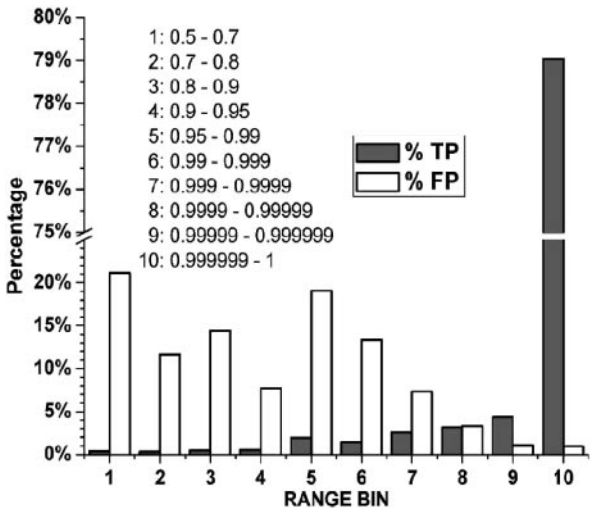

Fig. 5. Distribution of Naïve Bayes scores over the miRNAs class and the negative class from *mouse* sequences. This figure shows the distribution of the NBSs for the true positive (%TP) and the false positive (%FP) rates using *Mouse* sequences.

from several species, in the training set, and then apply our computational procedure to the various genomes, using the Naïve Bayes classification model that had been generated through the training. The output of the Naïve Bayes classifier (Fifth stage) is a pair of lists. The first list contains candidates that are assigned to the miRNA class, while the second contains candidates that are assigned to the non-miRNA class. The first list is subsequently refined by the next steps of the computational procedure.

Stage 6, Picking the best candidate from each stem–loop

A 21 nt sliding window representing the mature miRNA is moved along each stem–loop in the fourth stage of the computational procedure to generate the mature miRNA candidates. Each stem–loop generates a number of candidates. However, our initial assumption was that only one mature miRNA is associated with each stem–loop. Since different candidates that have been generated from one stem–loop may be predicted by the Naïve Bayes classifier to be a potential mature miRNA, we implemented an analyzer (Sixth stage) to pick the mature miRNA candidate with the highest score among all the candidates corresponding to one stem–loop. Further analysis is carried out on overlapping stem–loops that have been produced from the overlapping windows, which were generated at Stage 2 of the computational procedure (Fig. 1). From each family of overlapping stem–loops, only the one with the highest scoring mature miRNA is retained.

Stage 7, setting the Naïve Bayes score filter with the *Mouse* sequence

In the following sections we show that our computational procedure has high accuracy (high sensitivity and specificity) at finding known miRNA genes. However, we are still interested in further reducing the number of FP predictions, since the data to be examined at this point could reach millions of examples. Even with a small percentage of FPs, tens of thousands of predictions could be generated, making it difficult to validate these predictions in the laboratory. Thus, further analysis was applied to determine the appropriate threshold for eliminating FP predictions.

The Naïve Bayes classifier assigns a score to each mature miRNA candidate and classifies it into one of the two predefined classes: the miRNA and the non-miRNA. Figure 5 shows the distribution of the Naïve Bayes scores (NBSs) for the true positive (TP) and the FP using *Mouse* sequences (see sub-section 'miRNAs from *Mouse*' for information about the experiments). It is obvious that the 0.99999 (10e − 5) NBS is the appropriate threshold to reduce FP predictions. This value and others have been embedded, as additional filters, in our computational procedure at Stage 7.

Stage 8, conservation filter

The conservation of similar functional regions among evolutionarily related species can also be utilized to eliminate FP predictions. For each candidate precursor sequence (110 nt), a conservation measurement is obtained, with respect to a reference genomic sequence, using the BLAT program (Kent, 2002). Precursors which are highly conserved with respect to the reference genome (≥90%) are retained while those which are not conserved at this level are rejected. Figure 6 shows the distribution of this conservation score applied to all of the 224 known miRNAs from the *Mouse* genome against *Human*, *Chimpanzee* and *Fugu* genomes. The *Fugu* genome was also used by Lim *et al*. (2003b) to reduce the FP prediction.

It is obvious from Figure 6 that a conservation score >0.9 is appropriate to use as a filter in our computational procedure (Stage 8) when we use the *Human* or the *Chimpanzee* as the reference genome. One may also use the *Fugu* genome as a reference sequences for eliminating even more FP predictions, but it is expected that we would then begin to lose true potential candidate miRNAs. The choice of reference sequences and conservation score should vary with the new sequence being tested. We emphasize that the use of this filter is only to reduce the final number of predictions, and one would not use it if the outcome of Stage 6 is small enough for testing in the laboratory.

RESULTS

Evaluation of performance

To evaluate classification performance, several experiments have been executed as described in the following sections.

miRNAs from C.elegans In this experiment we aimed to identify miRNAs in the *C.elegans* genome. The 117 miRNAs from the

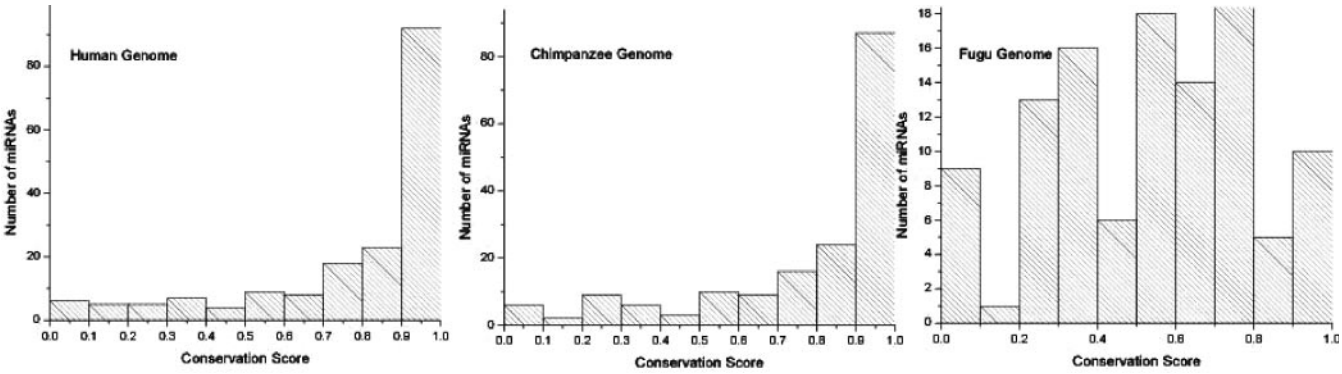

Fig. 6. Distribution of candidate *Mouse* miRNA sequence conservation scores in three species. This figure illustrates the distribution of the BLAT conservation score applied to all of the 224 known miRNAs from the *Mouse* genome against *Human, Chimpanzee* and *Fugu* genomes.

C.elegans genome were used as positive examples (Table 1) for training the Naïve Bayes classifier. For generating the negative examples, 300 sequences were randomly selected from the *C.elegans*/*C.briggsae* BLAT alignment [obtained from the UCSC genome Browser (http://genome.ucsc.edu/index.html) (Kent *et al*., 2002) on January 10, 2005 (version May 2003)]. The third and the fourth stages of the computational procedure (Fig. 1) were then applied to generate 434 955 negative examples.

The Naïve Bayes classifier was trained multiple times. In each training epoch, a set of 105 known miRNAs (90% of the positive data) were randomly selected from the 117 known *C.elegans* miRNAs and used as positive examples. We varied the number of negative examples across different sets of experiments, randomly choosing a set of 50, 75, 100, 125, 150, 175, 200, 300, 400, 500 or 1000 from the pool of negative examples (434, 955). The test was performed over the remaining 10% from the miRNA class and the remaining negative examples. The evaluation procedure was repeated 1000 times. The results are shown in Figure 7a. Using 150 samples from the negative data yielded ∼83% sensitivity and ∼96% specificity. Increasing the number of negative examples in the training set resulted in an increase in the classifier specificity and a decrease in its sensitivity. Using 300 examples from the negative class, the classifier reached ∼43% sensitivity and ∼99% specificity. It is clear that there is lack of stability in sensitivity as the number of negative examples grows. To verify that the stability does not change with respect to the number of positive examples, we altered the split of the training/testing portions, e.g. using 80% of the positive examples, with no significant difference.

Additional evaluation was implemented by 5-fold cross validation (repeated 100 times). The 117 known miRNAs served as the positive examples and 100, 200 or 1000 negative examples were selected from the negative pool. Receiver operating characteristic (ROC) (Metz, 1978) curves were generated and the area under the curves calculated. Results were 0.997, 0.992 and 0.960 respectively. The plot of the ROC curve for the 200 negative examples is shown in Figure 7b.

MiRNAs from Mouse The 224 known *Mouse* miRNAs downloaded from Rfam (Griffiths-Jones, 2004) were used as positive examples. We used 300 random sequences highly conserved between Mouse, Rat, Human, Dog and Chicken [using multiple alignment, May 2004, downloaded from USCS (Kent *et al*., 2002)]. By applying the third and the fourth stages of the computational procedure on those sequences, 239 674 stem–loops were generated to serve as a pool of negative examples for training the Naïve Bayes classifier.

The Naïve Bayes classifier was trained with 90% (201 miRNA) of the positive miRNA data and with 50, 75, 100, 125, 150, 175, 200, 300, 400, 500 or 1000 negative examples chosen randomly from the pool of 239 674 negative examples. The test was done with the remaining 10% from the miRNA class and the remaining negative examples. The evaluation procedure was repeated 1000 times, and the results are reported in Figure 8a. Using 150 samples from the negative data yielded ∼97% sensitivity and ∼91% specificity. As observed before, an increase in the number of negative examples in the training set leads to an increase in the classifier specificity and a decrease in its sensitivity. Again, it is obvious that there is lack of stability as the number of negative examples grows. For verifying the stability in the face of changes to the positive set, we again employed different splits of the training/testing portions, e.g. using 80% of the positive examples, with no significant change being observed.

Evaluation by 5-fold cross validation (repeated 100 times) was carried out using 224 known miRNA as the positive examples and 100, 200 or 1000 examples from the negative pool. ROC curves were generated and the area under the curves calculated. The results were 0.966, 0.980 and 0.965 respectively. The curve for the 200 negative examples is shown in Figure 8b.

Combining miRNAs from different species

We have so far shown the learning procedure applied to miRNA from a single species. The results demonstrated high specificity, but sensitivity was reduced with an increase in the number of negative examples. We now demonstrate that sensitivity benefits significantly in terms of both values and robustness, from combining multiple sets of miRNAs in the training data.

The currently known miRNAs from different species were downloaded from Rfam (Griffiths-Jones, 2004), (Table 1). 1359 mature miRNAs and 1420 precursor miRNAs served as positive examples. The difference between the number of mature and precursor miRNA stems from the fact that sequences for several mature miRNAs are found in more than a single precursor. The

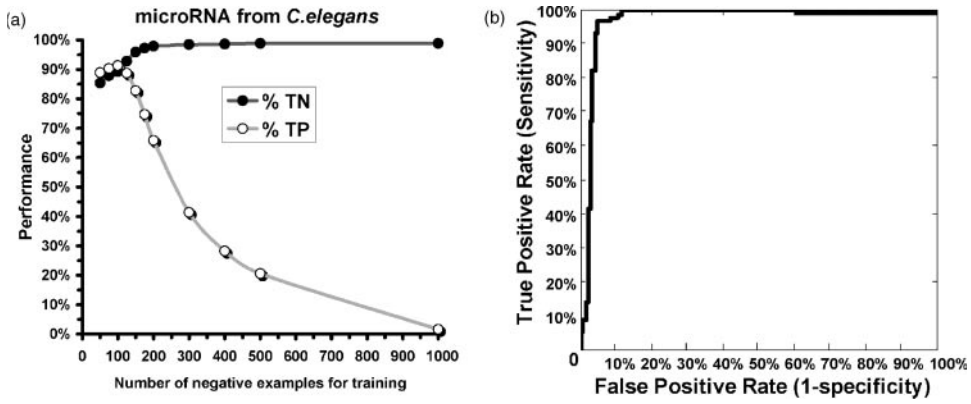

Fig. 7. Prediction performance as a function of size of the negative class for the *C.elegans* miRNAs. %TN is the true negative ratio and %TP is the true positive ratio. (**a**) The shown results are for the Naïve Bayes classifier that was trained multiple times. In each training epoch, a set of 105 known miRNAs (90% of the positive data) was randomly selected from the 117 known *C.elegans* miRNAs and used as positive examples. A varied number of negative examples across different sets of experiments, randomly choosing a set of either 50, 75, 100, 125, 150, 175, 200, 300, 400, 500 or 1000 from the pool of the negative examples (434 955) was used. The test was performed over the remaining 10% from the miRNA class and the remaining negative examples. The evaluation procedure was repeated 1000 times. (**b**) The ROC curve for 5-fold cross validation (repeated 100 times) using 200 negative examples; the area under the ROC curve is 0.992.

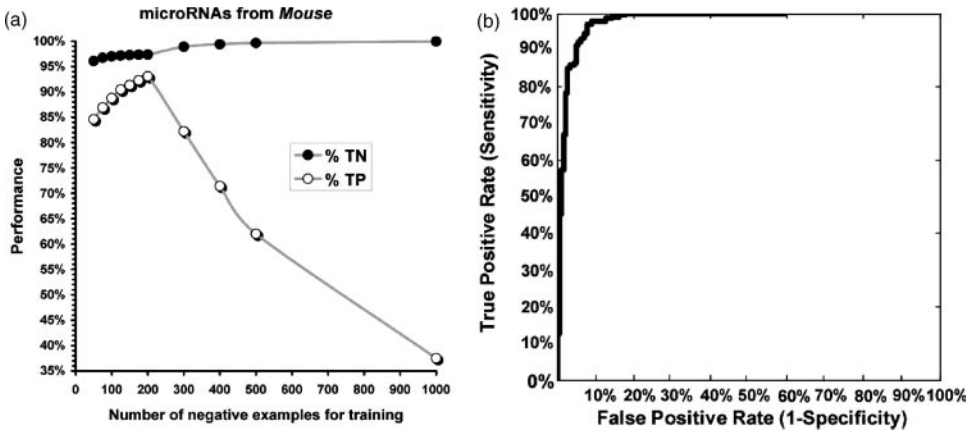

Fig. 8. Prediction performance as a function of size of the negative class for *mouse* miRNAs. %TN is the true negative ratio and %TP is the true positive ratio. (**a**) The shown results are for the Naïve Bayes classifier that was trained with 90% of the positive data *Mouse* miRNAs and with 50, 75, 100, 125, 150, 175, 200, 300, 400, 500 or 1000 negative examples chosen randomly from the pool of the negative examples (239, 674). The test was done with the remaining 10% from the miRNA class and the remaining negative examples. The evaluation procedure was repeated 1000 times. (**b**) The ROC curve for 5-fold cross validation (repeated 100 times) using 200 negative examples; the area under the ROC curve is 0.996.

10 redundant precursors were removed from the dataset. The pool of negative examples is the same as the one described in sub-section 'miRNAs from *Mouse*'.

The Naïve Bayes classifier was trained with 90% of the known miRNAs, and different numbers of negative examples that were randomly chosen from the pool of the negative examples (239 674). The test was performed with the remaining 10% from the miRNAs class and the remaining negative examples. The evaluation procedure was repeated 1000 times over 50, 75, 100, 125, 150, 175, 200, 300, 400, 500, 1000, 2000, 3000, 4000, 5000, 10 000, 20 000, 30 000, 40 000, 50 000 and 55 000 negative examples (Fig. 9a).

Once again we used 5-fold cross validation (repeated 100 times) for further evaluation with the 1420 known miRNA serving as the positive examples and 1000, 2000, 10 000, 20 000 or 30 000 negative examples selected from the negative example pool. The calculated areas under the ROC are 0.697, 0.897, 0.986, 0.982 and 0.978 respectively. The plot for the 10 000 negative examples is shown in Figure 9b.

The results shown in Figure 9 demonstrate that the classifier resulting from using all the miRNA sequences was more robust to changes in the size of the training set, than the ones trained using only the miRNAs from *C.elegans* or *Mouse*. Our explanation is that the distribution of the positive data used for training the classifier better represents the variety of miRNA classes. It is clear that when we provided more examples for training and for testing the classifier demonstrated better generalization. Using miRNA from multiple species for training allows our model to be applied to any genome, as it incorporates more information associated with different miRNA classes that may appear in different species.

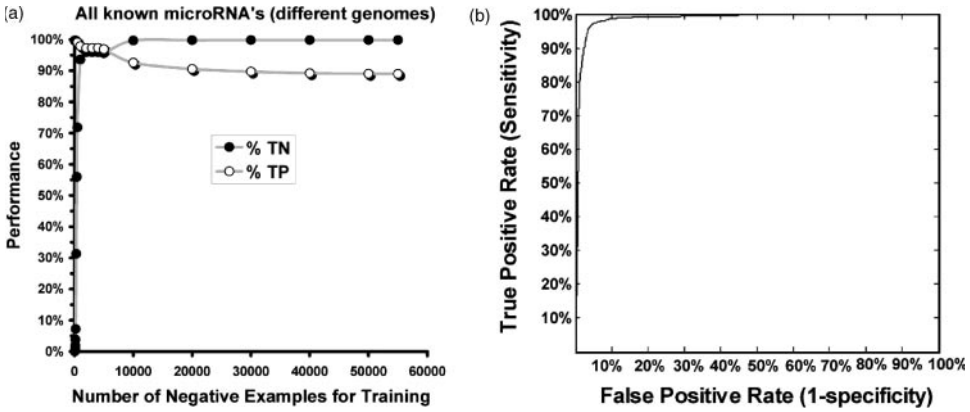

Fig. 9. Accuracy of prediction as a function of size of the negative class, for all miRNAs. %TN is the true negative ratio and %TP is the true positive ratio. (**a**) The shown results are for the Naïve Bayes classifier that was trained with 90% of the known miRNAs from different species, and different numbers of negative examples that were randomly chosen from the pool of the negative examples (239 674). The test was performed with the remaining 10% from the miRNAs class and the remaining negative examples. The evaluation procedure was repeated 1000 times, and the results for 50, 75, 100, 125, 150, 175, 200, 300, 400, 500, 1000, 2000, 3000, 4000, 5000, 10 000 20 000, 30 000, 40 000, 50 000 and 55 000 negative examples are shown. (**b**) The ROC curve for 5-fold cross validation using (repeated 100 times) 10 000 negative examples; the area under the ROC curve is 0.986.

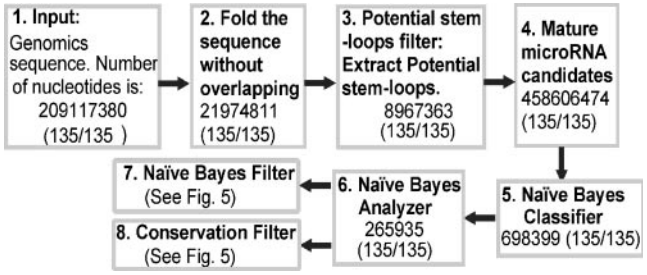

Fig. 10. Results from applying the computational procedure to the forward strand from the *Mouse*/*Human* genome. The classifier was trained using the 1420 known miRNAs from all species and 30 000 negative examples. The number of stem–loops retained at stage 2–6 is shown as well as the number of captured *Mouse* known miRNA genes (in parentheses).

Predicting miRNA genes in the *Mouse* genome

As input for this test, we used only the forward strand from the *Mouse*/*Human* BLAT alignments downloaded from UCSC [Mouse annotation: mm6, March 2005, NCBI Build 34 and Human annotation: hg17, May 2004, NCBI Build 35 (Kent *et al*., 2002)]. The computation was run on a parallel compute cluster with 100 nodes (http://core.pcbi.upenn.edu/tools/liniactools.html). The computation took 956 253 min of computer time (6.64 days elapsed time).

The classifier was trained using the 1420 known miRNAs from all species and 30 000 negative examples, as described above. The whole computational procedure was applied (Fig. 1) and the results are shown in Figures 10 and 11. Out of the 212 known mature miRNAs from the *Mouse* genome, 135 are on the forward DNA strand, and we kept track of those 135 miRNAs in our analysis, since we only analyzed the forward strand. For the conservation filter, we used the *Human* and *Fugu* genomes and 2909 miRNA precursors downloaded from Rfam (Release 7.0: June 2005) (Griffiths-Jones, 2004). Figure 10 shows the results as they are produced through the various stages of the Naïve Bayes analyzer pipeline. Stem–loop candidates were reduced to 265 935 from 21 974 811, a 100-fold reduction between stages 1 and 6, while retaining 100% of the 135 known miRNA genes.

The rest of the computational procedure, Stages 7 and 8 of Figure 10, is expanded in Figure 11, showing the application of a combination of conservation filters and species-specific feature rules. There are three main steps following the Naïve Bayes analyzer stage. For each step, we examined the values of the NBS filter. As the score increased, we reduced the percentage of TP predictions, as well as the number of predicted new miRNA genes to be tested.

In Step 1, we defined additional rules, that are based on the 135 captured miRNAs from the *Mouse* and built new rules as follows: (1) at least 60 nt stem length, (2) at least 25 bp, (3) at least ~26 kcal/mol of folding free energy, (4) at least 9 nt hairpin length and (5) at least a 0.9 NBS. These rules are now more specific to the *Mouse* genome, since the previous rules were extracted using multiple species. Applying these rules at the level of 10e − 7 of the NB score we detected 30% of the known miRNA genes and predicted 1466 miRNAs (of which 135 are already known).

In Step 2, before any additional *Mouse*-specific rules were applied, a conservation filter using the *Human* genome, as a reference (score 0.6 of conservation level), was applied reducing the number of candidates to 21 174 and finding 99 of the 135 *Mouse* miRNA genes. The subsequent application of the mouse-specific rules at a cutoff of 10e − 7 then resulted in only 379 candidates, which included 244 new genes not counting the 40 predicted known *Mouse* genes.

The left-most box in Figure 11 shows the results if conservation with all known miRNAs from different species is used as a filter (See Supplementary Information for the new predictions). There are 218 (83 new genes) predictions with 100% (135/135) prediction sensitivity. Our conclusion from these experiments is that the algorithm is able to identify the known miRNA with 100% sensitivity at the Naïve Bayes analyzer stage. However, owing to the large number of new predictions, we use additional filters to reduce the total number of predictions decreasing the sensitivity to different levels depending on the level of filtering.

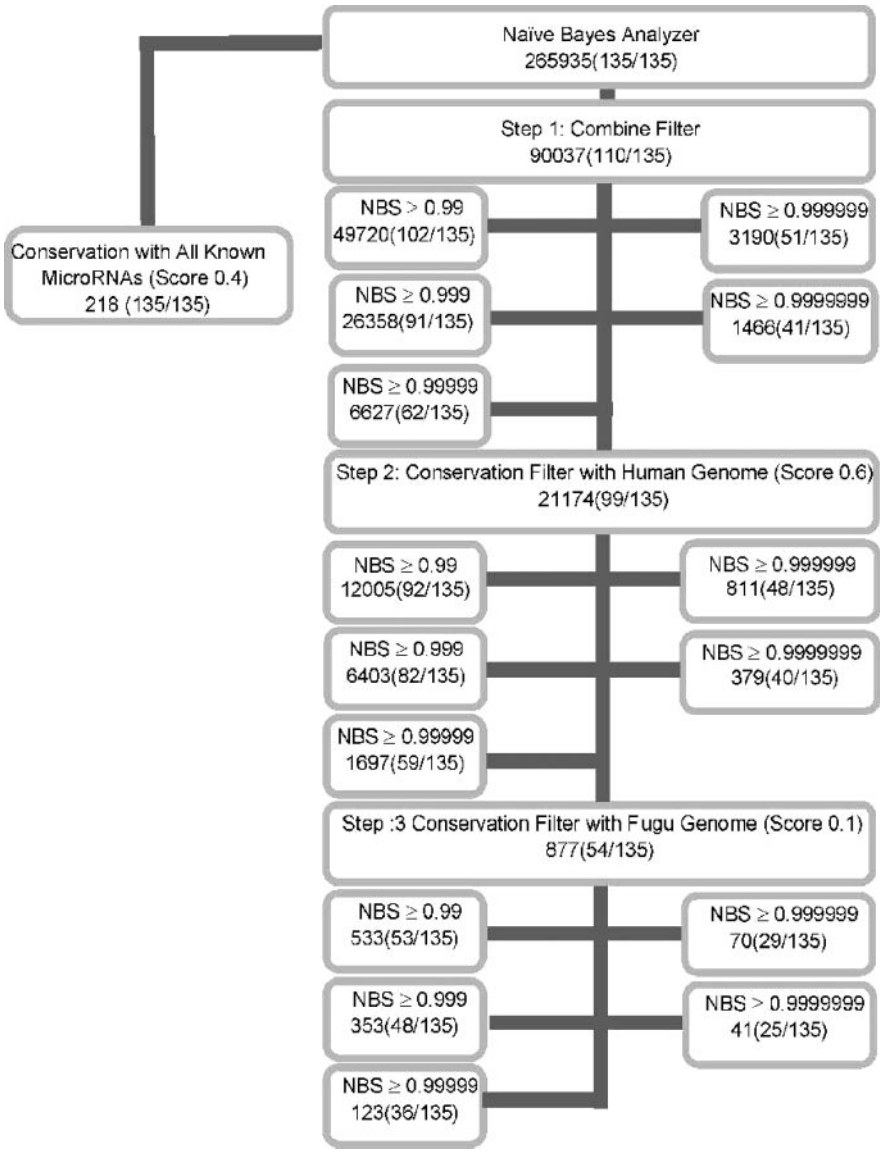

Fig. 11. Results of combining filters from the computational procedure (NBS = Naïve Bayes score). Three main steps following the Naïve Bayes analyzer stage are shown. For each step, the values of the NBS filter are examined. As the NBS increased, the percentage of TP predictions is reduced as well as the number of predicted new miRNA genes to be tested.

DISCUSSION

In this paper we present BayesmiRNAfind, a computational approach that predicts miRNAs based on their secondary structure and sequence. Our approach is more general than previously described algorithms, as it is not specific to a particular species. Instead, the program is based on using the sequence and structure of the known miRNAs for a specific genome as the input. Based on this data, the program uses machine learning to build a Naive Bayes classifier. It is likely that the computational procedure can be further improved by using other machine learning models such as SVM, or combinations of methods. The use of additional biological information to identify new miRNA genes could also be used to increase the sensitivity and the specificity of perditions. An example is the identification of conserved 8mer motifs in 3′-UTRs as potential miRNA binding sites and subsequent identification of new miRNA genes specific to those sites (Xie *et al*., 2005). The requirement that a putative miRNA contain a mature sequence complementary to a sequence in a known 3′-UTR might reduce false predictions enough to eliminate the need for a conservation filter.

As mentioned in the Introduction, the most recent work applying machine learning is by Nam *et al*. (2005). They use 5-fold cross validation on 1000 negative examples and 136 known miRNAs from Human with a threshold $P = 0.033$, reporting 73% sensitivity, 96% specificity and 0.936 area under the ROC curve. Our reported results demonstrate a higher sensitivity and specificity. For example, for the *C. elegans* genome, with 150 negative examples we obtained ∼83% sensitivity and ∼96% specificity. Using the *Mouse* genome and data with the same number of negative

examples we obtained ∼97% sensitivity and ∼91% specificity. Similarly, the area under the ROC curve exceeds previously reported results. It is clear that our method improves upon the state-of-the-art according to all measures used. The performance on Human is expected to be similar to that of Mouse.

In conclusion, our study shows that the application of a relatively simple machine learning technique while integrating data from multiple species can be a powerful approach for prediction of miRNA genes. We believe that this approach will enable the prediction of miRNA genes in organisms in which none has yet been identified so far.

ACKNOWLEDGEMENTS

The authors would like to thank Shere Billouin for preparing the manuscript. This project is funded in part by U01 CA85060 and the Pennsylvania Department of Health (PA DOH Commonwealth Universal Research Enhancement Program), and Tobacco Settlement grants ME01-740 and SAP 4100020718 (L.C.S.), NSF RCN 0090286 (M.K.S), M.N. is supported by NCI T32 CA09171 and caBIG subcontract 79522CBS10. H.S. is supported by NSERC Discovery Grant 298292-04.

Conflict of Interest: none declared.

REFERENCES

Ambros,V. *et al*. (2003) A uniform system for microRNA annotation. *RNA*, **9**, 277–279.
Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
Grad,Y. *et al*. (2003) Computational and experimental identification of *C.elegans* microRNAs. *Mol. Cell*, **11**, 1253–1263.
Griffiths-Jones,S. (2004) The microRNA Registry. *Nucleic Acids Res*., **32**, D109–D111.
Griffiths-Jones,S. *et al*. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*., **34**, D140–D144.
Japkowicz,N. and Stephen,S. (2002) The class imbalance problem: a systematic study. *Intell. Data Anal*., **6**, 429–450.
Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res*., **12**, 656–664.
Kent,W.J. *et al*. (2002) The human genome browser at UCSC. *Genome Res*., **12**, 996–1006.
Lagos-Quintana,M. *et al*. (2001) Identification of novel genes coding for small expressed RNAs. *Science*, **294**, 853–858.
Lai,E.C. *et al*. (2003) Computational identification of *Drosophila* microRNA genes. *Genome Biol*., **4**, R42.
Lau,N.C. *et al*. (2001) An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science*, **294**, 858–862.
Lee,R.C. and Ambros,V. (2001) An extensive class of small RNAs in *Caenorhabditis elegans*. *Science*, **294**, 862–864.
Lim,L.P. *et al*. (2003a) The microRNAs of *Caenorhabditis elegans*. *Genes Dev*., **17**, 991–1008.
Lim,L.P. *et al*. (2003b) Vertebrate microRNA genes. *Science*, **299**, 1540.
Mathews,D.H. *et al*. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol*., **288**, 911–940.
McCallum,A.K. (1996) *Bow: a toolkit for statistical language modeling, text retrieval, classification and clustering*. http://www.cs.cmu.edu/mccallum/bow.
Metz,C.E. (1978) Basic principles of ROC analysis. *Semin. Nucl. Med*., **8**, 283–298.
Mignone,F. *et al*. (2005) UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. *Nucleic Acids Res*., **33**, D141–D146.
Mitchell,T. (1997) *Machine Learning*. New York: McGraw Hill, Ch. 10.
Nam,J.-W. *et al*. (2005) Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Res*., **33**, 3570–3581.
Pasquinelli,A.E. *et al*. (2000) Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, **408**, 86–89.
Sahami,M. and Koller,D. (1996) Toward optimal feature selection. In *Proceedings of the Thirteenth International Conference on Machine Learning*. Morgan Kaufmann, San Francisco, CA, pp. 284–292.
Wang,X. *et al*. (2005) Sahami, M., Koller, D. (1996) Toward optimal feature selection. In Proceedings of the Thirteenth International Conference on Machine Learning. Morgan Kaufmann, San Francisco, CA, 284-292.. *Bioinformatics*, **21**, 3610–3614.
Weber,M.J. (2005) New human and mouse microRNA genes found by homology search. *FEBS Journal*, **272**, 59–73.
Xie,X. *et al*. (2005) Systematic discovery of regulatory motifs in human promoters and 3′UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*., **31**, 3406–3415.