BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 2 2007, pages 142–149
doi:10.1093/bioinformatics/btl570

Genome analysis

Reliable prediction of Drosha processing sites improves microRNA gene prediction

Snorre A. Helvik[1], Ola Snøve Jr[2,3] and Pål Sætrom[1,2,*]

[1]Department of Computer and Information Science, Norwegian University of Science and Technology, NO-7052 Trondheim, Norway, [2]Interagon AS, Laboratoriesenteret, NO-7006 Trondheim, Norway and [3]Department of Cancer Research and Molecular Medicine, Faculty of Medicine, Norwegian University of Science and Technology, N0-7489 Trondheim, Norway

Received on August 13, 2006; revised and accepted on November 8, 2006
Advance Access publication November 14, 2006
Associate Editor: Chris Stoeckert

ABSTRACT

Motivation: Mature microRNAs (miRNAs) are processed from long hairpin transcripts. Even though it is only the first of several steps, the initial Drosha processing defines the mature product and is characteristic for all miRNA genes. Methods that can separate between true and false processing sites are therefore essential to miRNA gene discovery.

Results: We present a classifier that predicts 5′ Drosha processing sites in hairpins that are candidate miRNAs. The classifier, called Microprocessor SVM, correctly predicts the processing site for 50% of known human 5′ miRNAs, and 90% of its predictions are within two nucleotides of the true site. Another classifier that is trained on the output from the Microprocessor SVM outperforms existing methods for prediction of unconserved miRNAs. Reanalysis of characteristics and supporting evidence for a set of newly annotated miRNAs shows that some miRNAs may be misannotated. This suggests that expressed hairpins should not be annotated as miRNAs until they are verified to be Drosha and Dicer substrates.

Availability: The classifiers are publicly available at https://demo1.interagon.com/miRNA/

Contact: paal.saetrom@interagon.com

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

1 INTRODUCTION

MicroRNAs (miRNAs) constitute a large class of non-protein-coding genes with the potential to downregulate protein-coding genes via sequence-specific target mechanisms (Bartel, 2004). Current estimates suggest that almost a thousand miRNAs exists in humans, and while most are well-conserved in other species, some are specific to mammals (Bentwich *et al*., 2005). MicroRNAs are crucial in animal development (Ambros, 2004) and seem to play an important role in genetic diseases, including cancer (Esquela-Kerscher and Slack, 2006).

In humans, miRNAs are primarily transcribed by RNA polymerase II (Lee *et al*., 2004) and undergo multistep processing before silencing their targets (Murchison and Hannon, 2004). First, primary transcripts (pri-miRNAs) are processed into precursors (pre-miRNAs) by the Microprocessor complex, which comprise Drosha and DGCR8 (Gregory *et al*., 2004; Denli *et al*., 2004). Second, another complex that consists of Exportin 5, a carrier protein, and RanGTP, transports pre-miRNAs from the nucleus and into the cytoplasm (Lund *et al*., 2004; Bohnsack *et al*., 2004) where they are released and processed by Dicer into short double-stranded RNAs with 2 nt 3′ overhangs (Bernstein *et al*., 2001; Hutvágner *et al*., 2001). These mature products are then unwound and one strand is incorporated into what could potentially be several variants of the RNA-induced silencing complex (RISC) (Schwarz *et al*., 2003; Mourelatos *et al*., 2002; Hutvágner and Zamore, 2002; Hammond *et al*., 2001). The RNA component guides RISC to messages with at least partial sequence complementarity to the guide strand (Jackson *et al*., 2003; Brennecke *et al*., 2005).

The Microprocessor step—where pri-miRNAs are processed into pre-miRNAs in the nucleus—leaves a 2 nt 3′ overhang (Lee *et al*., 2003). In the cytoplasm, Dicer cuts at a certain distance from the overhang created by the Microprocessor's cut (Vermeulen *et al*., 2005). The first processing step therefore determines the mature product. Several algorithms have been developed to find miRNA genes (Berezikov *et al*., 2006) and targets (Bentwich, 2005), but despite the importance of the problem, we are not aware of any attempts to predict Microprocessor processing sites. Some algorithms do, however, predict the mature miRNA sequence and thereby also the processing site as an integral part of the gene predictions (Lim *et al*., 2003; Lai *et al*., 2003; Xie *et al*., 2005; Nam *et al*., 2005).

Primary miRNAs share various sequence and structure characteristics that are likely to contribute towards efficient processing in the miRNA pathway (Ohler *et al*., 2004; Zeng and Cullen, 2003, 2004; Zeng *et al*., 2005; Krol *et al*., 2004). We previously reported that many features are conserved between known miRNAs (Sætrom *et al*., 2006). These commonalities include sequence and structure characteristics both upstream and downstream of the Microprocessor cut site. As for sequence characteristics, a significant presence or absence of certain bases at specific positions are conserved between miRNAs. The structural features include internal loops and bulges that commonly appear in specific positions. Importantly, the conservation of base pairing continues into the stem, but no

*To whom correspondence should be addressed.

conservation is found beyond the first 13 bases upstream of the annotated start of the pre-miRNA (Sætrom *et al*., 2006). A recent publication that studies the molecular basis for pri-miRNA recognition by the Microprocessor complex concludes that the cleavage site is determined mainly by the distance from the first unpaired nucleotides in the stem, thus confirming our results experimentally (Han *et al*., 2006).

Here, we describe a support vector machine classifier that can predict 5′ Microprocessor processing sites in human 5′ miRNAs with 50% accuracy. Importantly, if the predicted site is wrong, the actual site is within 2 nt of the predicted site in ∼90% of the cases. This Microprocessor SVM can be useful as a post-processor for existing tools that only predict whether hairpins are likely miRNAs. Furthermore, we use the Microprocessor processing site predictions to create a miRNA gene predictor that performs better than currently available methods for predicting unconserved miRNAs. Our Microprocessor SVM predicts the 5′ processing site and predictions of the 3′ processing site—and, as a consequence, 3′ miRNAs—will therefore be less accurate. Nevertheless, the miRNA gene predictor's performance is independent on whether the mature miRNA is from the 5′ or 3′ stem. By using the two classifiers to analyze 130 recently published miRNA sequences (Cummins *et al*., 2006), we show that several of these sequences do not share the characteristics of previously known miRNAs. Importantly, the lack of common characteristics as measured by our classifiers, correlates with a lack of evidence for the reported sequences being miRNAs. This correlation suggests that current databases may contain falsely annotated miRNAs.

2 MATERIALS AND METHODS

2.1 Sequences

We downloaded all 332 human miRNA sequences from miRBase (Griffiths-Jones, 2004) version 8.0 and all 130 new human miRNA sequences from miRBase version 8.1 to use as a dataset for 10-fold cross-validation and as an independent test set. Genomic sequences were from NCBI build 35 of November 2005 and were downloaded from the Ensemble FTP site (ftp://ftp.ensembl.org/pub/release-37/homo_sapiens_37_35j/data/fasta/dna/). Non-coding RNA (ncRNA) gene annotations where from Ensembl v37 of February 2006.

2.2 Predicting RNA secondary structure

We used RNAfold (Hofacker, 2003) version 1.4 with default parameters to predict RNA secondary structures.

2.3 Finding candidate processing sites

We defined all candidate processing sites by the 5′ processing site; the 3′ processing site was given by the 2 nt overhang relative to the 5′ site. For a miRNA, we defined the true processing site as the candidate site where the 5′ end was the same as the 5′ end found in miRBase. For miRNAs with mature sequences in the 3′ stem and no annotated ∗-sequence, we assumed the real 5′ processing site gave a 2 nt 3′ overhang relative to the annotated 3′ site.

To find the candidate processing sites for a particular miRNA sequence or predicted genomic hairpin, we used the predicted secondary structures of 110 nt long sequence windows centered on the hairpin. The candidate processing sites were then defined as all the 5′ sites that gave precursors from 50 to 80 nt long. To ensure that all candidates were folded as hairpins, we excluded all sequences that (1) did not have a hairpin loop within the minimum precursor sequence or (2) did not have at least 1 bp between nucleotides on opposite sides of the hairpin loop, such that the distance between the nucleotides was between 50 and 80 nt. Furthermore, we required that the nucleotide at the site forming the shortest candidate precursor base-paired with a nucleotide in the 3′ stem. Because of this requirement, the number of candidate processing sites varied between different hairpins.

Table 1. Features used by the Microprocessor SVM

ID	Explanation
1	Precursor length and loop size
2	Distance from the 5′ processing site to the loop start
3	Nucleotide occurrences at each position in the 24 nt regions of the precursor 5′ and 3′ arms
4	Base-pair information of each nucleotide for the 24 nt at the precursor base
5	Nucleotide frequencies in the two regions in 3
6	Total number of base pairs in the region in 4
7	Nucleotide occurrences at each position in the 50 nt 5′ and 3′ flanking regions
8	Base-pair information of each nucleotide for the 48 nt in the flanking region outside the precursor
9	Nucleotide frequencies in the two regions in 7
10	Total number of base pairs for the 15 nt immediately flanking the precursor
11	Total number of base pairs in the region in 8

Table 2. Additional features used by the miRNA SVM

ID	Explanation
12	Number of potential processing sites
13	Score for the best processing site
14	Average score for all potential processing sites
15	Standard deviation for all potential processing sites
16	Difference between features 13 and 14
17	Distance between the three top-scoring processing sites
18	Number of local maximums in the processing site score distribution

2.4 Feature vectors for Microprocessor SVM

All features, except the loop size, were specifically calculated for each processing site candidate (Table 1). The secondary structure features in a candidate precursor and its flanking regions were calculated based on the predicted secondary structures of 110 and 180 nt long sequence windows centered on the hairpin. Precursor length is the number of nucleotides from 5′ processing site to 3′ processing site and includes the 2 nt 3′ overhang. Loop size is the number of predicted unpaired bases in the hairpin loop. Position specific nucleotide occurrences (features 3 and 7) are encoded using four binary features such that A = [1,0,0,0], C = [0,1,0,0], G = [0,0,1,0] and U = [0,0,0,1]. Nucleotide frequencies (features 5 and 9) are four features counting the number of occurrences of each nucleotide. Position specific base-pair information (features 4 and 8) is the average number of nucleotides base-pairing at a specific position relative to the 5′ and 3′ processing sites. For a specific position x, this value BP_x is 0, 0.5 or 1 depending on whether none one or both of nucleotides x upstream of the 5′ processing site and $x-2$ downstream of the 3′ processing site are base-paired with a nucleotide in the opposite strand. Total number of base pairs (features 6, 10 and 11) is the sum of BP_x in the corresponding region. In total, the feature vector consisted of 686 variables.

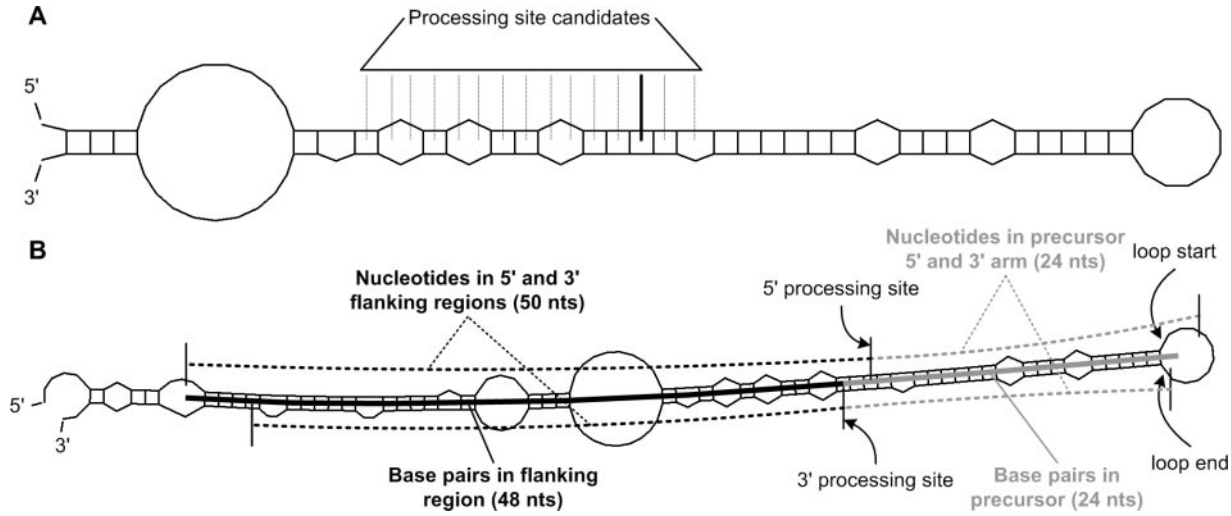

Fig. 1. The Microprocessor SVM calculates features relative to the putative 5′ processing site. (**A**) For a particular hairpin, the set of 5′ processing site candidates are the sites that result in precursor candidates of lengths between 50 and 80 nt. The figure shows hsa-mir-23a's predicted secondary tructure, and its processing site candidates; the true processing site, based on miRBase annotations, is in bold. (**B**) For a particular processing site, we compute the input features to the Microprocessor SVM from the regions immediately surrounding the site. More specifically, we compute features from the 50 nt in the 5′ and 3′ flanking regions, the 24 nt in the 5′ and 3′ arms of the putative precursor, the base-pair information for the 48 nt in the flanking region outside the precursor, and the base-pair information for the 24 nt part of the precursor.

2.5 Additional features for miRNA SVM

The seven additional features used by the miRNA SVM (Table 2) were all derived from the Microprocessor SVM predictions using basic statistics (features 12, 13, 14, 15 and 16). The distance between the three top-scoring processing sites (feature 17) is the total number of nucleotides between the three top-scoring sites. The number of local maximums in the processing site score distribution (feature 18) is the number of candidate sites with a score higher than −0.5 and a score higher than its closest neighboring sites.

2.6 Performance measures

In a classification problem, a prediction can either be a true or false positive or true or false negative prediction. From the counts of each case (TP, FP, TN and FN) sensitivity, specificity and positive predictive value are defined as:

$$\mathrm{Se} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{1}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \text{ and} \tag{2}$$

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}. \tag{3}$$

2.7 Classifier performance estimation

In 10-fold cross-validation, the data set is randomly divided into 10 equally sized folds and for each fold, a classifier is trained on the remaining 9 folds and tested on the 1 remaining fold. The test results in the 10 folds usually give a good estimate of an algorithm's ability to generalize to unseen data (Kohavi, 1995). Randomly dividing the set of known miRNAs, however, risk introducing bias as many groups of miRNAs have the same or similar sequences and structures. We therefore placed all similar miRNAs, as defined in miRBase version 8.1 (ftp://ftp.sanger.ac.uk/pub/mirbase/sequences/8.1/miFam.dat.gz), in the same fold. Similarly, for the Microprocessor SVM, we placed all candidate processing sites from the same miRNA in the same fold.

2.8 Extracting hairpins from the genome

We used a fast algorithm based on edit distance computations to extract hairpins from the human genome. The algorithm, called ScorePin, uses dynamic programming to calculate the structural similarity between the given sequence and a hairpin structure with a perfectly base-paired stem. See the online Supplementary Material for algorithm details.

We ran ScorePin on both strands of the complete human genome. All positions that scored below or equal to a given threshold were considered hairpin candidates, but we used only the best scoring candidate within a window of 14 nt. We used a threshold of 110 and this resulted in a total of 8 556 723 hairpin candidates, which included 98% of the miRNAs in miRBase release 8.1. The threshold was chosen based on a 10-fold cross-validation to maximize ScorePin's specificity and sensitivity. We then used RNAfold (Hofacker, 2003) to verify the hairpin structure and to ensure that the hairpin had candidate processing sites. This filtering reduced the number of hairpins to 6 798 341 without reducing the number of miRNAs. A total of 2004 (0.03%) of these had at least one of the strands overlapping with ncRNA annotations from Ensembl v37.

3 RESULTS AND DISCUSSION

We wanted to construct a classifier that automatically determines the processing site when presented with a hairpin candidate and to find out whether this property could be used to improve current miRNA gene prediction algorithms. SVMs are known to produce classifiers that generalize well to unseen data (Schölkopf, 1997) and have already been used for miRNA gene predictions (Xue *et al*., 2005; Sewer *et al*., 2005). We used the gist implementation of SVMs (Pavlidis *et al*., 2004) with a radial basis function kernel (Burges, 1998) and default parameters to construct a classifier that separates between true and false sites (Fig. 1A).

The SVM's feature vector included both sequence and structure properties (Fig. 1B) that were selected based on the results of our previous study (Sætrom *et al*., 2006). Table 1 lists the features; see Materials and Methods for details.

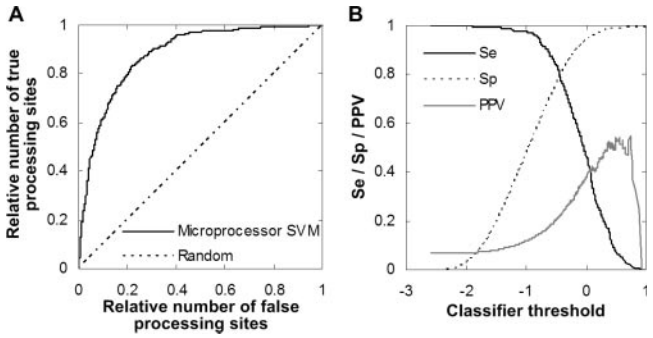

Fig. 2. The Microprocessor SVM distinguishes between real and false processing sites. (**A**) The ROC-curve shows, for varying thresholds, the relative number of true and false processing sites with a predicted score above the given threshold. Random predictions give a straight line from (0,0) to (1,1). (**B**) The Microprocessor SVM's sensitivity (*Se*), specificity (*Sp*), and positive predictive value (PPV) for varying thresholds.

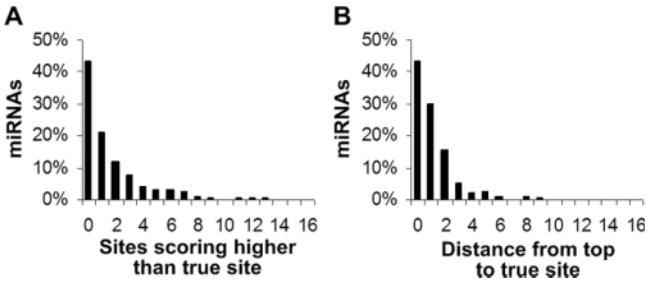

Fig. 3. The score depends on the probability of being a true site. (**A**) The number of positions with a higher score than the true site is usually low. (**B**) Almost 90% of all true cut sites are within two nucleotides of the highest scoring site.

3.1 The Microprocessor SVM accurately predicts annotated processing sites

We used a modified version of 10-fold cross validation on the miRNAs from miRBase (release 8.0) to test the Microprocessor SVM's predictive power (see Materials and Methods). For 327 of the 332 miRNAs in the database, the annotated 5′ processing site resulted in a predicted precursor between 50 and 80 nt. These 327 sites formed the positive dataset; the remaining sites formed the negative dataset.

A comparison of the average scores of true and false target sites on the test sets showed that the SVM could separate true from false processing sites (−0.084 for true sites versus −0.968 for false sites). As expected, true processing sites are more homogeneous than false processing sites, as the standard deviation of 0.439 for true sites is smaller than the corresponding 0.592 for false sites.

Sensitivity and specificity—the relative number of correctly predicted true and false processing sites (see Materials and Methods)—are intuitive and useful performance measures, but these measures depend on a predefined threshold that defines the minimum score a position must receive to be considered as a positive prediction. Figure 2A shows a receiver operating characteristic (ROC) curve (Hanley and McNeil, 1982) that displays the Microprocessor SVM's tradeoffs between sensitivity and specificity across all thresholds. Figure 2B shows the SVM's sensitivity and specificity for all thresholds, and also shows the corresponding positive predictive value. Thus, without taking into account that many processing sites are related as they are located in the same hairpin, the Microprocessor SVM can separate true from false sites; for example, ∼80% of the true sites can be picked up before 20% of the false sites are included in the results.

For a particular pri-miRNA, however, thresholds for defining positive predictions are not necessary, as the highest scoring position naturally becomes the predicted processing site. Of the 327 miRNAs in our dataset, the highest scoring site was the true site for 43.1%. We refer to this score as the prediction rate. Importantly, a site's score depends on the probability of the site being the miRNA's true site. Figure 3A shows that the number of false processing sites with scores above that of the true site is non-random and that >75% of the miRNAs have two or fewer false predictions above the true site. Furthermore, the true site is usually close to the highest scoring site. As shown in Figure 3B, ∼90% of predicted processing sites are either correct or 1 or 2 nt away from the true site. Thus, for most miRNAs the Microprocessor SVM either predicts the true processing site or a site that is very close to the true site.

Note that the above accuracy estimates depend on our initial assumption that for miRNAs from the 3′ stem, the real 5′ processing site always gave a 2 nt 3′ overhang relative to the annotated 3′ site. This is not generally true. Consequently, for the miRNAs where this simplified assumption does not hold, we have used the wrong 5′processing site in our experiments. This also means that even if the Microprocessor SVM predicted the correct processing site for these miRNAs, we will record the prediction as wrong. Indeed, the Microprocessor SVM has a lower prediction rate for miRNAs that are in the 3′ arm (35.1%) compared with miRNAs that are in the 5′ arm (50.3%). As we can be confident that the processing sites we inferred for the miRNAs from the 5′ stem are correct, it is likely that our initial prediction rate estimate of 43.1% is too low.

3.2 The precursor's flanking region is most important for processing site recognition

To construct a biological model for the recognition of processing sites by the Microprocessor complex, it is necessary to find which feature contributes most to the predictor's performance. Decoding black box classifiers like SVMs is difficult, but as an approximation, we did an analysis where we removed each of the 11 features from the Microprocessor SVM's feature vector, retrained the SVM, and measured the resulting 10-fold cross-validation performance. Figure 4 shows the resulting SVMs' performance relative to the original Microprocessor SVM with all features included.

The most important features for determining the processing site are the precursor length and loop size (feature 1). This is perhaps not surprising, as the variation in miRNA precursor length is small (Sætrom *et al*., 2006) (avg = 60.7 and SD = 4.9 for human miRNAs in miRBase 8.1). Except for the precursor length, the most important features are located outside the precursor. The second and third most important features are the nucleotide occurrences and base-pair information in the precursor's flanking region (features 7 and 8). Other important features are the nucleotide occurrences in the precursor and the length of the precursor 5′ arm (features 3 and 2). On the other end of the scale, the general base composition in the flank and the number of base pairs in the flank closest to the

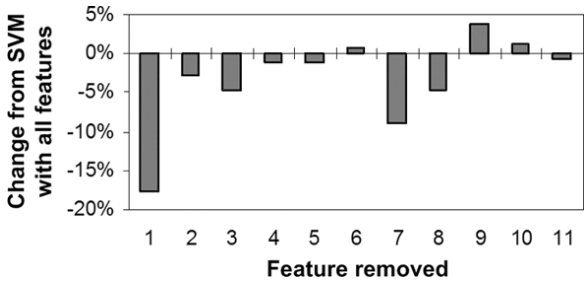

Fig. 4. Some features are more important for predicting the processing site than others. The graph shows how removing various features changes the SVM's performance. The reported performance change is the average relative change in ROC-area, prediction rate, average distance from the topscoring prediction to the true processing site and average number of false processing sites with a higher score than the correct site. A negative change in the two former measures indicates lower performance, whereas a negative change in the two latter measures indicates increased performance. Consequently, we reversed the sign of the two latter measures before computing the average. See Supplementary Figure 1 for changes in the individual measures.

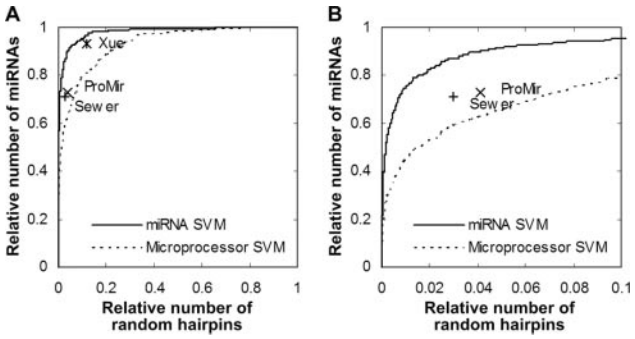

Fig. 5. Microprocessor processing site prediction improves miRNA gene prediction. (**A**) The dotted line shows the ROC-curve for the predictor that was trained on true versus false processing sites in pri-miRNAs, whereas the solid line shows ROC-curve for the predictor that was trained on miRNAs versus random hairpins. True positives are estimated based on cross-validation; false positives are based on the genomic predictions. The points for Sewer, ProMir, Xue are the sensitivities and specificities reported by Sewer *et al*. (2005), Nam *et al*. (2005), and Xue *et al*. (2005). (**B**) Detailed excerpt of (A).

precursor (features 9 and 10) are detrimental to the SVM's performance. Furthermore, the other features that summarize a region's general characteristics (features 5, 6 and 10) are relatively unimportant—possibly because these features vary little between close processing sites. Thus, the SVM likely uses the precursor length as a rough guide to find likely processing sites, and the position specific occurrences of nucleotides and base-pairs to fine-tune the predictions.

Whether particular nucleotides in the precursor's flanking region are single stranded or not is important for Microprocessor processing site recognition (Han *et al*., 2006) and is also a strong characteristic of mammalian miRNAs (Sætrom *et al*., 2006). These results confirm the importance of both a single stranded flanking region and certain nucleotide motifs in it (Sætrom *et al*., 2006) and suggest that these sequence motifs are important for processing site recognition. The motifs' full importance can, however, only be found by mutation studies that disrupt the sequence motifs but retain the base-pairing patterns.

3.3 Reliable processing site prediction improves miRNA gene prediction

Since recognition by Drosha is a hallmark of true pri-miRNA transcripts, we hypothesized that reliable predictions of processing sites could be used to distinguish miRNA genes from random hairpins. We used an edit distance-based algorithm to extract close to 6.8 million hairpins from the human genome (see Materials and Methods). The set included all known miRNAs, except mir-98, mir-198, mir-134, mir-384, mir-425 and mir-484, which translates to a sensitivity of 98%. The number of candidate hairpins is in accordance with previously published miRNA gene prediction algorithms (Bentwich *et al*., 2005; Nam *et al*., 2005), but the sensitivity is higher; for example, Bentwich *et al*. (2005) extracted ~10 million hairpins from the human genome, but missed 15% of the known miRNAs.

We then used the Microprocessor SVM to score the potential processing sites in all the genomic hairpins and used the highest scoring processing site for each hairpin as the hairpin's score. As shown by the dotted line in Figure 5, the Microprocessor SVM can predict miRNA genes with fairly good results. Sixty percent of the true miRNAs are picked up at a level where >95% of the initial hairpin set has been excluded.

The Microprocessor SVM shows good performance when separating miRNAs from random hairpins, despite that we did not train the SVM for this purpose. We therefore expected that an SVM explicitly trained to distinguish miRNAs and random hairpins based on the Microprocessor SVM's processing site predictions would have an even higher performance. This is because even though certain features may be unimportant for the Microprocessor SVM to predict the correct processing site, these features can be important to distinguish miRNAs from random hairpins. This is particularly true for the number of base-pairs in the precursor (feature 6), which Lim *et al*. (2003) reported to be the most important variable for predicting miRNAs in *Caenorhabditis elegans*. To train this miRNA SVM, we used a set of 3000 random genomic hairpins and the miRNAs from miRBase 8.0 to create a set of negative and positive feature vectors. More specifically, for a particular hairpin or miRNA we used the Microprocessor SVM's highest scoring processing site candidate to estimate the true processing site for that hairpin and used this site to create a feature vector that included the same features as for the Microprocessor SVM.

The value for the highest scoring processing site candidate indicates whether or not a given hairpin is a likely miRNA (Fig. 5). Similarly, as miRNAs have a higher standard deviation than random hairpins (0.518 and 0.249), the standard deviation of processing site scores for a particular hairpin may indicate whether the hairpin is a likely miRNAs. Consequently, we included several features derived from the Microprocessor SVM predictions in the miRNA SVM feature vectors (Table 2; see Materials and Methods for details).

Figure 5 shows that the miRNA SVM performs much better than the Microprocessor SVM alone (solid versus dotted line). These performance characteristics also hold when the two SVMs classify

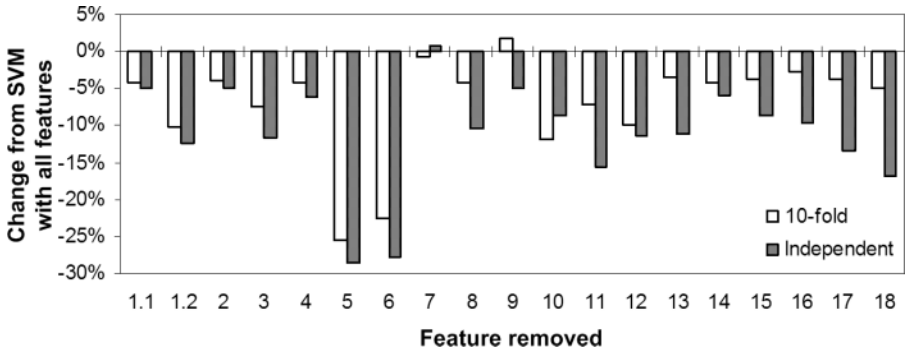

Fig. 6. All features except the nucleotide information in the precursor flanking regions are important for miRNA gene prediction. The graph shows how removing different features changes the SVM's performance in terms of ROC-area for specificities ≥99%. The performance changes are estimated based on 10-fold cross-validation and an independent test set of 17 000 random genomic hairpins. Features 1.1 and 1.2 are the precursor length and loop size.

the subset of genomic hairpins that overlap with other known ncRNAs (Supplementary Figure 2). Furthermore, the performance of the miRNA SVM does not depend on whether the miRNA is in the 5′ or 3′ stem, as there is no correlation between the miRNA SVM's predictions and whether the miRNAs is 5′ or 3′ (data not shown). The sensitivity of the algorithm has improved enough to pick up ∼90% of the true miRNAs if ∼5% false predictions can be tolerated. Of course, when the initial number of hairpins is in the millions, it quickly becomes impractical if the specificity of the predictions is decreased.

Nevertheless, compared to other existing methods that do not rely on sequence or structure conservation to predict miRNA genes, our method has a much better performance. At sensitivity levels comparable to those of Sewer *et al*. (2005) and ProMir (Nam *et al*., 2005), our method reduces the number of genomic hairpin candidates to 24 and 20%. Compared with Xue *et al*. (2005), the miRNA SVM reduces the number of hairpin candidates to 60%, but the sensitivity level reported by this method is impractical in a full genome screen, as the number of remaining hairpin candidates is almost half a million (483 490). This illustrates the difficulty of predicting miRNAs in whole genomes, but also highlights the merit of basing miRNA gene prediction on reliable processing site predictions.

3.4 MicroRNA gene prediction relies on overall differences between miRNAs and genomic hairpins

To evaluate the importance of the different features used by the miRNA SVM, we did a similar analysis as for the Microprocessor SVM; that is, we removed each of the different feature groups, retrained and tested the SVM using 10-fold cross-validation, and observed the SVM's performance change. We measured the ROC-area for specificities ≥99%, as we were primarily interested in the SVM's performance in the high-specificity region. Changes in the complete ROC-area were similar but of a smaller magnitude (data not shown). Figure 6 summarizes the results.

Compared to the Microprocessor SVM (Fig. 4), the miRNA SVM relies on a different set of features. The miRNA SVM's most important features by far are the nucleotide frequencies (5) and the total number of base pairs (6) in the putative precursor. These features had little influence on the Microprocessor SVM's performance (Fig. 4), but the miRNA SVM possibly uses these features to filter out hairpins that clearly are not miRNAs because of lack of base-pairing or an uncharacteristic nucleotide composition in the precursor. Similarly, the precursor length (1), nucleotide occurrences and base-pair information in the precursor's flanking region (7 and 8), and the length of the precursor 5′ arm (2) are less important. As these features are important for the Microprocessor SVM's predictions, they will be more homogeneous than other features in the miRNA SVM's input vector and therefore also less useful for distinguishing miRNAs from random hairpins. Most of the features do, however, contribute to the miRNA SVM's performance; the only exceptions are the nucleotide occurrences and nucleotide frequencies in the flanking regions (7 and 9). Thus, whereas the Microprocessor SVM mostly uses position specific features to pinpoint the likely processing site, the miRNA SVM uses overall features to separate miRNAs from random hairpins.

Note that by including random genomic hairpins in the Microprocessor SVM's negative training set, we could likely have trained a single SVM both to separate miRNAs from random hairpins and to find the true processing site in Microprocessor substrates. The features derived from the Microprocessor SVM's predictions would not, however, be available to this single SVM. As these features do contribute to the miRNA SVM's performance (Fig. 6), the single-SVM solution would likely have lower performance than our approach has.

3.5 Reanalysis suggests that some sequences in miRBase are not miRNAs

We also tested the Microprocessor and miRNA SVMs on a set of 130 new miRNAs from release 8.1 of miRBase, of which 122 passed our preprocessing filters. On this set, the Microprocessor SVM correctly predicted the processing site for 27.9% of the miRNAs. Similarly, the miRNA SVM had problems separating some of these new miRNAs from genomic hairpins (Supplementary Figure 3). These results seem to indicate that the SVMs had overfitted our initial training set and could not capture the characteristics of these new miRNAs, possibly because the known set of miRNAs is biased towards certain characteristics. Several

other results indicate, however, that some of these sequences may instead have been falsely annotated as miRNAs.

First, when we used a 10-fold cross-validation to retrain and test the Microprocessor SVM on the complete set of 449 miRNAs from miRBase 8.1, we got a prediction rate of 33.4%. Again, this rate is considerably lower than the prediction rate for the initial set of miRNAs. Thus, even when we add the new miRNAs to our training set, the SVM fails to classify these sequences as miRNAs.

Second, as was previously reported, these new miRNAs differ from the rest of the known miRNAs in several aspects (Berrzikov *et al*., 2006). All the new human miRNAs were discovered by sequencing small RNAs (Cummins *et al*., 2006), but (1) the number of tags for the new miRNAs was much lower than for the previously known miRNAs (2000 compared with 70 000), and (2) only 3 of the new miRNAs are differentially expressed in a Dicer-disrupted cell-line, as compared with 55 of the 97 known miRNAs.

Third, Berezikov *et al*. (2006) points out that two of the new miRNAs (hsa-mir-565 and hsa-mir-594) overlap with tRNA annotations. Neither sequence was represented by >1 clone nor differentially expressed in the Dicer-disrupted cell-line (Cummins *et al*., 2006). Furthermore, both sequences receive very low scores by the Microprocessor and miRNA SVM; the hairpins and annotated processing sites of hsa-mir-565 and hsa-mir-594 are scored 2nd and 8th (miRNA) and 29th and 20th (Microprocessor) lowest among the new miRNAs. We therefore believe that these two sequences were falsely annotated as miRNAs.

Fourth, the total body of evidence for these reported miRNAs truly being miRNAs (hairpin conservation, sequencing of miRNA*-strand, multiple observations of expression, genomic clustering with other miRNAs, strong homology to other miRNAs and multiple tags), correlates with the predictions from the Microprocessor and miRNA SVMs. More specifically, reported miRNAs that had much additional evidence tended to get scored higher by the two SVMs than reported miRNAs that had none or one additional supporting evidence. The correlations between Microprocessor and miRNA SVM scores and the total amount of supporting evidence were 0.35 and 0.34, which is significant ($P = 4 \times 10^{-5}$ and $P = 8 \times 10^{-5}$ with two-tailed Student, *t*-tests).

Granted, one cannot dismiss potential miRNAs as false based on lack of supporting evidence. Certain miRNAs are not conserved (Bentwich *et al*., 2005), have temporal or cell-specific expression patterns (Lu *et al*., 2005) that may complicate detection, are not clustered with other miRNAs (Altuvia *et al*., 2005), or have no homology to other miRNAs. But if we take the least biased approach and only look at the evidence from the sequencing, which is presence of multiple tags or the miRNA*-strand, the correlations between Microprocessor and SVM scores and total amount of evidence were even higher (0.39 and 0.36; $P = 4 \times 10^{-6}$ and $P = 3 \times 10^{-5}$). Similarly, when we removed the miRNAs with none or one additional supporting evidence and evaluated the SVMs' predictions on this set, we got results that were similar to those on the initial dataset (Microprocessor SVM prediction rate of 41.8%; see Supplementary Figure 4 for miRNA SVM ROC). Thus, though it is possible that the new sequences in miRBase 8.1 that were difficult to classify are miRNA genes with different characteristics than the majority of known miRNAs, we find the correlation between low prediction scores and lack of additional evidence suspect. We therefore propose that some of these sequences were falsely annotated as true miRNAs. Furthermore, we echo Berezikov *et al*.'s opinion that a single sequenced clone mapped to a non-conserved hairpin should only be referred to as a candidate miRNA (Berezikov *et al*., 2006) until further evidence can be found.

4 SUMMARY AND CONCLUSIONS

We have presented a support vector machine classifier that predicts Microprocessor processing sites with high accuracy. We have also shown that using the predictions from this classifier as input to a miRNA gene predictor gives more accurate predictions than currently available methods for finding unconserved miRNAs do. In addition to the classifier's obvious use in predicting miRNA genes, we believe that it will be useful for designing short hairpin expression constructs that mimic miRNAs.

Short hairpin RNAs (shRNAs) for sequence-specific silencing by RNA interference exploit the cellular machinery of miRNAs. Short hairpin RNAs are structurally similar to pri-miRNAs or pre-miRNAs (Brummelkamp *et al*., 2002; Zeng *et al*., 2002), but whereas the mature product from pre-miRNA-like shRNAs can be reliably predicted, the mature product from pri-miRNA-like shRNAs may be more difficult to predict. Short hairpin RNAs modeled as pri-miRNAs are often modified versions of a putative pri-miRNA, and these modifications may affect processing by the Microprocessor complex (Zeng and Cullen, 2003; Zhou *et al*., 2005). We believe that our classifiers will be useful in shRNA design as they can assess whether a construct resembles endogenous miRNAs and is a likely Microprocessor substrate, and predict the likely processing site.

Our results also indicate that miRNA annotation practices should be changed such that annotations make clearer distinctions between verified and candidate miRNAs. Current guidelines require both expression and biogenesis evidence for a sequence to be annotated as a miRNA (Ambros *et al*., 2003), but in practice, the biogenesis evidence is synonymous with the sequence being part of a hairpin structure. Short expressed sequences may, however, be part of predicted hairpins without being miRNAs. Sequences from tRNAs have previously been annotated as miRNAs and been removed from miRBase (Sætrom *et al*., 2006), and hsa-mir-565 and hsa-mir-594 in the current version of the registry may be part of tRNAs as well (Berezikov *et al*., 2006). Similarly, our results indicate that several newly sequenced small RNAs that were mapped to hairpins and annotated as miRNAs without additional evidence, lack features that may be essential for miRNA biogenesis.

Current knowledge of animal miRNAs indicate that they undergo two distinct processing steps, namely excision of the precursor from the primary transcript by the Microprocessor complex and cleavage of the precursor by Dicer. Current guidelines for miRNA annotation consider Dicer processing as sufficient evidence, but we believe that until a sequence can be shown to be both a Drosha and Dicer substrate, the sequence should be annotated as a miRNA candidate and not a verified miRNA. We expect that such a distinction will help future research by pointing out more and less reliable data in existing databases.

ACKNOWLEDGEMENTS

O.S. and P.S. received support by the bioinformatics platform of the Norwegian Research Council's Functional Genomics Program

(FUGE). Funding to pay the Open Access publication charges for this article was provided by the Norwegian University of Science and Technology.

Conflict of Interest: none declared.

REFERENCES

Altuvia,Y. *et al*. (2005) Clustering and conservation patterns of human microRNAs. *Nucleic Acids Res*., **33**, 2697–2708.

Ambros,V. (2004) The functions of animal microRNAs. *Nature*, **431**, 350–355.

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.

Bentwich,I. (2005) Prediction and validation of microRNAs and their targets. *FEBS Lett*., **579**, 5904–5910.

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

Berezikov,E. *et al*. (2006) Approaches to microRNA discovery. *Nat. Genet*., **38**, S2–S7.

Bernstein,E. *et al*. (2001) Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, **409**, 363–366.

Bohnsack,M.T. *et al*. (2004) Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. *RNA*, **10**, 185–191.

Brennecke,J. *et al*. (2005) Principles of microRNA-target recognition. *PLoS Biol*., **3**, e85.

Brummelkamp,T.R. *et al*. (2002) A system for stable expression of short interfering RNAs in mammalian cells. *Science*, **296**, 550–553.

Burges,C.J. (1998) A tutorial on support vector machines for pattern recognition. *J. Data Mining and Knowledge Discovery*, **2**, 121–167.

Cummins,J.M. *et al*. (2006) The colorectal microRNAome. *Proc. Natl Acad. Sci. USA*, **103**, 3687–3692.

Denli,A.M. *et al*. (2004) Processing of primary microRNAs by the Microprocessor complex. *Nature*, **432**, 231–235.

Esquela-Kerscher,A. and Slack,F.J. (2006) Oncomirs—microRNAs with a role in cancer. *Nat. Rev. Cancer*, **6**, 259–269.

Gregory,R.I. *et al*. (2004) The Microprocessor complex mediates the genesis of microRNAs. *Nature*, **432**, 235–240.

Griffiths-Jones,S. (2004) The microRNA registry. *Nucleic Acids Res*., **32**, D109–D111.

Hammond,S.M. *et al*. (2001) Argonaute2, a link between genetic and biochemical analyses of RNAi. *Science*, **293**, 1146–1150.

Han,J. *et al*. (2006) Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell*, **125**, 887–901.

Hanley,J. and McNeil,B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.

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

Hutvágner,G. and Zamore,P.D. (2002) A microRNA in a multiple-turnover RNAi enzyme complex. *Science*, **297**, 2056–2060.

Hutvágner,G. *et al*. (2001) A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science*, **293**, 834–838.

Jackson,A.L. *et al*. (2003) Expression profiling reveals off-target gene regulation by RNAi. *Nat. Biotechnol*., **21**, 635–637.

Kohavi,R. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th IJCAI,* San Francisco, CA, Morgan Kaufmann Publishers, pp. 1137–1143.

Krol,J. *et al*. (2004) Structural features of microRNA (miRNA) precursors and their relevance to miRNA biogenesis and small interfering RNA/short hairpin RNA design. *J. Biol. Chem*., **279**, 42230–42239.

Lai,E.C. *et al*. (2003) Computational identification of *Drosophila* microRNA genes. *Genome Biol*., **4**, R42.

Lee,Y. *et al*. (2003) The nuclear RNase III Drosha initiates microRNA processing. *Nature*, **425**, 415–419.

Lee,Y. *et al*. (2004) MicroRNA genes are transcribed by RNA polymerase II. *EMBO J*., **23**, 4051–4060.

Lim,L.P. *et al*. (2003) The microRNAs of *Caenorhabditis elegans*. *Genes Dev*., **17**, 991–1008.

Lu,J. *et al*. (2005) MicroRNA expression profiles classify human cancer. *Nature*, **435**, 834–838.

Lund,E. *et al*. (2004) Nuclear export of microRNA precursors. *Science*, **303**, 95–98.

Mourelatos,Z. *et al*. (2002) miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. *Genes Dev*., **16**, 720–728.

Murchison,E.P. and Hannon,G.J. (2004) miRNAs on the move: miRNA biogenesis and the RNAi machinery. *Curr. Opin. Cell Biol*., **16**, 223–229.

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.

Ohler,U. *et al*. (2004) Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. *RNA*, **10**, 1309–1322.

Pavlidis,P. *et al*. (2004) Support vector machine classification on the web. *Bioinformatics*, **20**, 586–587.

Sætrom,P. *et al*. (2006) Conserved microRNA characteristics in mammals. *Oligonucleotides*, **16**, 115–144.

Schölkopf,B. (1997) *Support Vector Learning*. Oldenbourg Verlag, Munich.

Schwarz,D.S. *et al*. (2003) Asymmetry in the assembly of the RNAi enzyme complex. *Cell*, **115**, 199–208.

Sewer,A. *et al*. (2005) Identification of clustered microRNAs using an ab initio prediction method. *BMC Bioinformatics*, **6**, 267.

Vermeulen,A. *et al*. (2005) The contributions of dsRNA structure to Dicer specificity and efficiency. *RNA*, **11**, 674–682.

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.

Xue,C. *et al*. (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, **6**, 310.

Zeng,Y. and Cullen,B.R. (2003) Sequence requirements for microRNA processing and function in human cells. *RNA*, **9**, 112–123.

Zeng,Y. and Cullen,B.R. (2004) Structural requirements for pre-microRNA binding and nuclear export by Exportin 5. *Nucleic Acids Res*., 4776–4785.

Zeng,Y. *et al*. (2002) Both natural and designed micro RNAs can inhibit the expression of cognate mRNA when expressed in human cells. *Mol. Cell*., **9**, 1327–1333.

Zeng,Y. *et al*. (2005) Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha. *EMBO J*., 138–148.

Zhou,H. *et al*. (2005) An RNA polymerase II construct synthesizes short-hairpin RNA with a quantitative indicator and mediates highly efficient RNAi. *Nucleic Acids Res*., **33**, e62.