NTNU
Innovation and Creativity

A Classifier for Microprocessor Processing Site Prediction in Human MicroRNAs

Snorre Andreas Helvik

Master of Science in Computer Science
Submission date: July 2006
Supervisor: Arne Halaas, IDI
Co-supervisor: Pål Sætrom, Interagon AS

Norwegian University of Science and Technology
Department of Computer and Information Science

Problem Description

Predicting microRNA genes by predicting the Microprocessor processing site in the microRNA primary transcript

MicroRNAs are a newly discovered class of small regulatory RNA genes with important functions in animal development, stress resistance, fat metabolism, and brain morphogenesis and microRNAs are also implicated in cancer as both oncogenes and tumor suppressors. Several hundred microRNAs are currently characterized in the human genome, but the current estimate of around 1000 microRNAs indicate that several genes remain to be discovered.

MicroRNAs are recognized by a characteristic hairpin structure and the first step in microRNA processing is the cleavage of this hairpin from the microRNA primary transcript. The cleavage site defines one end of the mature microRNA and predicting this site is therefore an important step in microRNA gene prediction. This project will focus on developing methods for predicting this processing site and test whether such a processing site predictor can be used to predict microRNA genes in a set of hairpin candidates.

Assignment given: 05. February 2006
Supervisor: Arne Halaas, IDI

Abstract

MircoRNAs are ∼22nts long non-coding RNA sequences that play a central role in gene regulation. As the microRNAs are temporary and not necessarily expressed when RNA from tissue samples are sequenced, bioinformatics is an important part of microRNA discovery. Most of the computational microRNA discovery approaches are based on conservation between human and other species. Recent results, however, estimate that there exists around 350 microRNAs unique to human. It is therefore a need for methods that use characteristics in the primary microRNA transcript to predict microRNA candidates. The main problem with such methods is, however, that many of the characteristics in the primary microRNA transcript are correlated with the location where the Microprocessor complex cleaves the primary microRNA into the precursor, which is unknown until the candidate is experimentally verified.

This work presents a method based on support vector machines (SVM) for Microprocessor processing site prediction in human microRNAs. The SVM correctly predicts the processing site for 43% of the known human microRNAs and shows a great performance distinguishing random hairpins and microRNAs.

The processing site SVM is useful for microRNA discovery in two ways. One, the predicted processing sites can be used to build an SVM with more distinct features and, thus, increase the accuracy of the microRNA gene predictions. Two, it generates information that can be used to predict microRNA candidates directly, such as the score differences between the candidate's potential and predicted processing sites. Preliminary results show that an SVM that uses the predictions from the processing site SVM and trains explicitly to separate microRNAs and random hairpins performs better than current prediction-based approaches. This illustrates the potential gain of using the processing site predictions in microRNA gene prediction.

Preface

As we learn more about human biological functions, even more important questions are raised. Many of these question are not only important for understanding an extremely complex system, but also quite facinating. The microRNA genes were discovered and accepted as a separate gene class only five years ago, and have the past few years shown to have great biological importance in areas such as development and cancer. The functions of the microRNAs are still widely unknown and a hot topic.

This thesis work was done at Interagon AS; a company that uses search solutions to help the pharmaceutical and biotechnology industry overcome bioinformatic challenges. A special thanks to my supervisor Pål Sætrom for showing a great interest in this work and providing a superior guidance.

The result of the work is here presented in the form of an extended scientific article, with focus on the results and how informatics was applied to obtain the results. We will first give an introduction about some of the basic biology and related work that has influenced this work. Further, the results are presented, and include how the methods were developed and how they performed. The results also describe related analyses that may help our understanding of the topic. Next, a discussion investigates some of the results combined with results from other recent studies. Then, the work is summarized and the natural next steps are pointed out in the Summary and Conclusion. Finally, the Materials and Methods describe the material sources and methods used in the work and give detailed information necessary to recreate the work.

Basic knowledge of molecular biology is required to understand this paper. Information beyond the introduction section can be found in the references.

Trondheim, July 2nd 2006

Snorre Helvik

Contents

1 Introduction

1.1 MicroRNAs play a central role in gene regulation

Traditionally, genes were thought to only encode proteins. That is, it was though that the gene sequence was transcribed from DNA into RNA, which then was translated into a protein [1]. Although some genes were known to not encode proteins, such as tRNAs and rRNA, these non-protein-coding RNA genes were thought to be few compared to the protein-coding genes. In 2001, a large class of non-coding RNA genes, called microRNAs were discovered [2, 3]. MircoRNAs are approximately 22 nucleotides long non-coding RNA sequences that regulate the expression of genes at the posttranscriptional level [4], and have been found to play key roles in a wide variety of biological processes, including cell fate specification, cell death, proliferation, and fat storage [5]. As the microRNAs' importance was discovered relatively recently, their function in gene regulation is still widely unknown, thus, identifying them is of great interest.

The microRNA primary transcript; that is, the microRNA sequence in the RNA transcribed from the DNA, has a characteristic secondary structure, called a hairpin structure. In a hairpin structure there are two ∼35nts long stems joined together by base-pairs and closed by a loop in the end; see figure 1.

This hairpin structure is shown to play an important role in the processing of the microRNAs from the primary microRNA structure to the single-stranded mature ∼22nts long mature microRNA [6, 7]. This process consists of four steps; see figure 2. First, the microRNA is cleaved from the primary transcript (pri-microRNA) to a separate hairpin, called the microRNA precursor (pre-microRNA). This process is done by the Microprocessor protein complex, and defines one of the mature microRNA ends [6, 8]. Second, the pre-miRNA is transported from the nucleus to the cytoplasm by Ran-GTP and the export receptor Exportin-5 [9, 4]. Third, the loop is cut from the microRNA precursor by the RNase III endonuclease enzyme Dicer [6]. This defines the second end of the mature microRNA, leaving only the mature microRNA joined with the other stem, called the microRNA*, by near-perfect base-pairing. Fourth, the two stems are unwound by the helicase to create two separate ∼22nts long sequences [10]. One of these strands is incorporated into the protein complex that effect microRNA function and is called the mature microRNA. This protein complex is often referred to as the RNA induced silencing complex (RISC). The mature microRNA functions by guiding the protein complex to mRNAs that are partially complementary to the microRNA. The protein complex then either causes mRNA cleavage, translational suppression, or polyadenylation, depending on the degree of sequence complementary. Collectively, microRNAs target thousands (30%) of protein coding genes [11].

The hairpin structure is not only characteristic for all known microRNAs, it is also

Figure 1: The hairpin structure is characterised by a relatively long stem of base-pairs, closed by a terminal loop in the end.

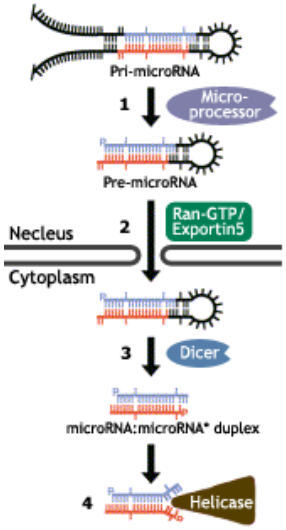


Figure 2: The microRNA is processed in four steps: 1) the hairpin is first cleaved from the primary transcript; 2) the precursor is transported to the cytoplasm; 3) the loop is cut from the hairpin; and 4) the two stems are unwound. The figure originates from [4].

a requirement for the microRNA annotation. Mainly, a microRNA candidate must fulfill two criteria to be annotated as a microRNA [12]. First, the mature microRNA must be expressed and detected, either by northern blot analysis, cloning, micro-array or other experimental methods. Second, the mature microRNA must be located in the stem of a hairpin structure without large internal loops or bulges. Alternatively, microRNA candidates that are homologs to experimentally verified microRNAs are accepted as microRNAs without experimentally evidence.

To date, 462 human microRNAs have been identified and registered in version 8.1 of the microRNA registry [13]. The first estimates suggested that there were only a few hundred human microRNAs [14, 15]. However, several recent studies estimate that this number could be as high as one thousand [16, 17, 18, 19]. Thus, microRNA discovery remains an important aspect to understand the regulation mechanisms.

1.2 MicroRNA discovery

1.2.1 Bioinformatics play an important role in microRNA discovery

The microRNA discovery methods can be divided into two groups: experiment-driven and bioinformatics-driven. Experiment-driven microRNA discovery methods

use the first of the two mentioned annotation criteria as the starting point. That is, small RNAs from tissue samples are sequenced and analysed, and then bioinformatic tools are used to verify the structural requirements of the selected microRNA candidates. For a review of experimental-driven methods, see [20]. Bioinformatics-driven methods use the second annotation criteria as the starting point. A set of microRNA candidates are computationally predicted based on common characteristics, and then verified through experimental techniques to demonstrate their expression. The bioinformatics-driven methods are reviewed in the two following sections.

Initially, the vast majority of the microRNAs were discovered with experiment-driven methods such as massive cloning and sequencing [2, 21, 3, 22]. Although these approaches were quite successful, they have three major limitations: 1) many microRNAs are tissue-specific; that is, microRNAs expressed in one tissue might not be expressed in other types of tissues [22]. 2) the tissues are often dominated by a few microRNAs. Thus, even with a considerable amount of RNA as raw material, the cloned products run a great risk of being dominated by a few highly expressed microRNAs [21]. 3) the expression of microRNAs is temporary and regulated by other factors such as environment and development [3]; that is, microRNAs expressed in a specific tissue might not be expressed in the same type of tissue at a latter stage. In other words, many microRNAs are only expressed in certain situations and for a short period of time, and cannot be identified by cloning or sequencing. Therefore, the need for computational approaches to overcome these problems has been recognized the past few years.

After the first wave of microRNA identification, the bioinformatics-driven methods became more popular. There was now sufficient information about the microRNAs to find common microRNA characteristics, which in turn could be used to identify novel microRNAs. Other studies point out that computational approaches are suited to identify microRNAs due to their characteristic secondary structure and the evolutionary conservation of microRNAs [21], which not only is a strong microRNA characteristic, but also a part of the requirements for microRNA annotation.

1.2.2 The principles of bioinformatics-driven microRNA discovery

The bioinformatics-driven methods can roughly be divided into two categories: conservation-based methods and prediction-based methods.

The conservation-based methods exploit the assumption that the biologically important parts of the genome often stay conserved through evolution, and that many of the known human microRNAs are conserved in other species [17, 14, 3, 3]. There are basically two ways to identify conserved microRNAs: to use known microRNAs to identify homologs, or to search for new sequences that are conserved between two or more species and verify whether these are part of hairpin structures. See [23, 20] for

reviews of conservation-based microRNA discovery approaches. The conservation-based methods have been widely used with successful results, but there is one major drawback for these methods. They are per definition restricted to only find conserved microRNAs, which means that the microRNAs that are unique to a specie or have diverged too far from their homologs will stay undiscovered. Recent work has identified several non-conserved novel microRNAs [24, 25], and estimate that there exist around 350 non-conserved microRNAs in the human genome [24].

The prediction-based methods use machine learning to distinguish microRNAs and random hairpins. The general principle consists of four steps. First, two datasets are gathered: one set of positive examples of known microRNAs, and one set of negative examples, usually random hairpin structures believed not to be microRNAs. Second, the positive and negative examples are examined and the features believed to be different between the positive and negative examples are selected. Third, a machine-learning algorithm is constructed based on the selected features and trained with the examples, such that the algorithm is optimized to best distinguish the positive examples from the negative. Fourth, the trained algorithm is used to classify unknown examples, that is, to predict whether a given sequence is positive or negative. An important strength with prediction-based approaches is that they not only consider how a microRNA should look like, but also how they should not look like. This gives a richer ability to determine which features are the most important to distinguish microRNAs and random hairpins. Details about how support vector machines, the machine learning algorithm used in this work, function can be found in section 5.2.5 in Materials and Methods.

1.2.3 MicroRNA identification by prediction-based approaches

There are four prediction-based microRNA discovery approaches, all published recently. Although all four approaches are independent of conservation, only one of them verifies non-conserved microRNAs [24]. Two of the approaches identify novel microRNAs that are homologs with known microRNAs [26, 18], while the fourth approach does not predict any microRNA candidates, but only describes the methods and the features used in the approach [27].

Bentwich et al. [24] developed PalGrade, which is the only published prediction-based approach to predict and verify non-conserved human microRNAs. They constructed a vector-based classifier, and optimized the vectors by finding the combination of features that best distinguished the microRNAs and a background set of 10,000 random hairpins found in non-protein-coding genomic regions. The features used in their analysis were structural features such as hairpin length, loop length, stability score, free energy per nucleotide, number of matching base-pairs and bulge size, and sequence features such as sequence repetitiveness, regular and inverted internal repeats and free energy per nucleotide composition. A total number of $\sim$ 11,000,000 random hairpins were extracted from the entire human genome

and scored with PalGrade. The 5,300 of the 434,239 hairpins that had a minimum threshold score were selected for experimentally investigation. 89 of these were validated as microRNAs, of which 53 are not conserved beyond primates.

ProMir [26] is a first order hidden markov model (HMM) based on the relationship between nucleotide information and secondary structure. Four states are created to describe the structure of a position in the paired hairpin; base-paired, mismatched, inserted or deleted. As the model uses two separate states for insertion and deletion, the HMM only model one of the stems at the time. Therefore, the 5' microRNAs and the 3' microRNAs were processed separately. Additionally, two hidden states describe whether the position is a part of the mature region or not. The total probability for an unknown sequence is a multiplication of the transition probability and emission probability for each position. The transition probability is the probability of going from a prior state to the current state, and the emission probability is the probability for observing the two given nucleotides in that particular state. Both the transition and emission probabilities were trained on a positive set of 136 human microRNAs and a negative set of 1,000 random hairpins. The results, based on a 5-fold cross-validation, gave a sensitivity of 73% and specificity of 96%. The processing site predictions gave an average distance of 2.66 nucleotides between the predicted and real processing site. ProMir does not directly depend on conservation. However, as the homologs have both similar mature sequence and similar structure, ProMir captures the conservation features better than the general microRNA features. They experimentally confirmed 9 of 23 predicted microRNA gene candidates. Some of these did not demonstrate a clear sequence similarity to the known microRNAs genes.

Sewer et al. [18] created an SVM based on over 25 distinct features, calculated from four different regions of the candidate hairpin. The regions are 1) the entire hairpin structure; 2) the longest symmetrical region in the hairpin; 3) the longest region with a total asymmetry not larger than four; and 4) all possible windows of size 20 that could be placed along the hairpin. The model is trained using 178 known microRNAs as positive examples and 5,395 hairpin structures isolated from tRNA, rRNA and mRNA genes as negative examples, and gives a sensitivity of 71% at a 97% specificity. For predicting novel microRNAs they focus on the genomic regions around already known microRNAs, to exploit that microRNAs often are found in clusters. 68 of their 260 top-scoring predictions were experimentally validated by searching a large database of cloned small RNAs from human, mouse and rat. Although the approach is not based on any conservation, their validated microRNAs are homologs with microRNAs in at least one other specie. The reason for this might be because they used hairpins from genomic regions that are already known to contain other conserved microRNAs, and that they considered a candidate as verified if it was expressed in any specie, not only the specie the candidate was extracted from.

Xue et al. [27] created an SVM based on a combination of structure and nucleotide

content, which they call triplet elements. Each nucleotide is part of a local structure according to the predicted secondary structure. That is, either the nucleotide is part of a base-pair or it is a mismatch/insertion. A triplet element is the nucleotide structure for three continuos nucleotides combined with the nucleotide type of the nucleotide in the middle of these three, yielding 32 different combinations of triplets. For example, if all thee of the nucleotides in the subsequence "AGC" are involved in base-pairs, the triplet element will be "G(((", where "(" symbolizes a base-pair. The hairpin structure's feature vector is the relative number of appearances of each of these 32 triplets along the stem of the candidate hairpin. The SVM is trained on a set of 163 known human microRNAs as positive examples and 168 random hairpins extracted from the protein coding regions as negative examples. The classificator uses a separate test set of 30 microRNAs and 1,000 random hairpins to estimate the SVM's performance, which gives a sensitivity of 93.3% and a specificity of 88.1%. However, no attempt was done to predict microRNA genes in genomic hairpins, mainly because such a low specificity would give too many false positives.

1.2.4 Many features correlate with Microprocessor

The position where Microprocessor cleaves the primary microRNA into the microRNA precursor, called the Microprocessor processing site, is the start of the mature region. That is, Microprocessor defines the mature microRNA region and is, thus, of great biological importance. Lab-experiments have shown that Microprocessor's processing is exact. That is, Microprocessor creates a single product in most cases, but may occasionally use two different processing sites [28]. It is therefore likely that there are certain characteristics that describe where Microprocessor will cleave the primary microRNA.

A recent study [29] investigating known microRNA transcripts found that both structural and sequential features are conserved in the microRNA precursor but not random hairpins. That is, described in terms of the distance from the processing site, the occurrences of bases, bulges and mismatches are not random, but are more prevalent at specific positions in the microRNA precursor. Similar observations have also been reported by several other studies [30, 31, 14]. Sætrom et al. also found that structural and sequential features were conserved at some positions in the flanking region, and that there was a significantly different base composition in the precursor and flanking region. Several studies show that the flanking region plays an important role in the Microprocessor process [6, 8, 28]. For example, altering the base-pairing in the beginning of the flanking region can reduce the expression of the microRNA [8, 32]. Sætrom et al. found that the base-pairing in the flanking region is conserved up to approximately 13 bases, but not beyond this point.

How Microprocessor recognizes the processing site is not yet fully understood. Initial studies suggested that a relatively large loop is required for efficient Micro-

processor processing and that Microprocessor cleaved the stem by measuring two helical turns from the loop [8, 33]. However, a recent study suggests that the loop is unessential and that the 5' end of the processing site is mainly determined by a distance of approximately 13 bases from the end of the hairpin double-stranded structure [28], which is consistent with the results from Sætrom et al.

These results not only suggest that many features are related to the Microprocessor processing site, but also that there are characteristics in the primary microRNA transcript that describe where Microprocessor will cleave the primary transcript.

1.3 The processing site can be predicted for 43% of the human microRNAs

Several microRNA characteristics have been discovered and used in prediction-based microRNA discovery. These characteristics include both structural features such as hairpin length, base pairing and bulge symmetry, and sequential features such as nucleotide content in the primary structure, often in combination with the secondary structure. However, there is still one major weakness with the prediction-based microRNA discovery approaches. Research suggest that there are distinct features that describe the location of the Microprocessor processing site in the primary microRNA transcript. Thus, seen from a microRNA discovery point of view, if we know the location of the processing site we can create distinct features to distinguish microRNAs from other hairpin structures. Yet, all the current prediction-based approaches fail to include this information, because unless the microRNA candidate already has been experimentally verified they have no information of where the Microprocessor processing site is.

To solve this, the approaches have based the features on the entire hairpin (PalGrade and ProMir), different regions of the hairpin that are independent of the processing site (Sewer et al.), or modeled each position in the hairpin independently (Xue et al.). The only approach predicting the mature region (ProMir) does not use this information in the microRNA gene prediction but as a parallel step to the gene prediction.

We consider this as one of the main problems with prediction-based microRNA discovery. The predictions are only based on information found in the initial primary transcript, although the features are shown to greatly depend on how the microRNAs are being processed. It is therefore of great interest to develop a program to predict the Microprocessor processing site for a given candidate microRNA. Not only will such a program help to predict microRNA genes more accurately, but as the processing sites define the mature microRNA region, the processing sites have an important biological function. Consequently, finding these processing sites can help our understanding of the microRNA process.

This work reports a method based on support vector machines (SVM) that predicts the Microprocessor processing sites in microRNA candidates. The SVM correctly predicts the processing site for 43% of the known human microRNAs and gives an average distance of only one nucleotide between the predicted processing site and the real processing site for the human microRNAs. Further, the SVM shows a great performance in distinguishing microRNA genes and genomic hairpins, which indicates the advantage of processing site prediction in microRNA discovery. The results also show that a much higher performance is gained by creating a new SVM trained explicitly to distinguish the predicted processing sites for microRNAs and random hairpins. In general, the Microprocessor processing site predictor is useful for microRNA discovery in two ways. One, the processing site can be predicted for any given hairpin structure, which in turn can be used to build the machine learning algorithm using more distinct features and, thus, increase the accuracy of the predictions. Two, it generates information that can be used to distinguish microRNA genes and genomic hairpins directly. This is not only information about how likely the predicted processing site is, but also about the differences between the predicted and the other potential processing sites. This information can wither be used to predict microRNA genes directly, or it can be used as new features in a gene prediction classifier.

2 Results

2.1 An SVM to predict Microprocessor processing sites in human microRNAs

2.1.1 Processing the microRNAs into positive and negative examples

A machine learning approach for binary classification is always trained using two distinct training sets: one set of positive examples, and one set of negative examples. In this case, one example is a microRNA with a given processing site, further referred to as a processing site.

The entire set of human microRNAs was downloaded from version 8.0 of the microRNA registry [13], 332 microRNAs in total. All potential processing sites in these 332 microRNAs were extracted and used to train and test the SVM (see section 5.2.1 in Materials and Methods for details). A processing site is considered as positive when the processing site is the same as the annotated one. To be more exact; the processing sites are strictly defined by the 5' end, which means the positive processing site is the processing site where the 5' end is the same as the 5' end of the annotated microRNA precursor.

The predicted secondary structures of 110nts long sequence windows centered on the microRNA hairpins were used to find the microRNAs' potential processing sites (see

section 5.2.4 in Materials and Methods for details). As the microRNAs' annotated secondary structures are based on sequences in a range of 70-110 nucleotides, our predicted secondary structures may vary from the annotated ones. A verification step was therefore necessary to insure the predicted secondary structures can be correctly processed by our software. The verification step consists of two tests. Test 1) there must be a closed loop within minimum precursor of 50nts. Several small loops within the minimum precursor are considered the same loop. Test 2) there must exist at least one base-pair along the stem that give a valid precursor length, that is, a precursor length between 50-80nts.

All the 332 microRNAs passed the verification step. Five of the microRNAs, however, failed to give the real processing site. Two of these, hsa-mir-448 and hsa-mir-451, had a precursor length outside the range defined in the premisses, a length of 82 and 42 respectively. The other three, hsa-mir-25, hsa-mir-198 and hsa-mir-453, had a predicted secondary structure that resulted in a precursor length longer than 80 for the real 5' end of the processing site. The number of false processing sites found for these three were 3, 13 and 10, respectively. The false processing sites for these five microRNAs were included in the training set, but, unless otherwise specified, not included when evaluating the classifier's performance on individual microRNAs. Additionally, 22 real processing sites did not have a 3' end identical with the annotated 3' end. Only three of these had a predicted secondary structure with the 3' end more than four nucleotides away from the annotated 3' end. These were hsa-mir-24-2, hsa-mir-346 and hsa-mir-431. Due to consistency, these were included anyway.

All in all, the training set consists of 327 positive processing sites from 327 microRNAs, and 4,681 negative processing sites from 332 microRNAs.

2.1.2 Both primary and secondary features describe a processing site

The work by Sætrom et al. [29] found that several characteristics, both structural and sequential, are conserved in primary microRNAs, both the precursor and the flanking region. Structural features such as position of bulges, internal loops and base-paring are more prevalent at specific regions and positions. Similarly, sequence features such as base distribution is significantly different in different parts of the primary microRNAs, and some nucleotides appear more frequently at specific positions in the precursor and surrounding regions. As a minimum, the processing site classifier should include features covering the discussed characteristics.

The features used in the Microprocessor processing site classifier are listed in table 1, where feature 3 to 11 cover the characteristics described by Sætrom et al. In addition, some common features not already covered is used. These are precursor length, loop size and distance from the mature region to the loop. The precursor length of 60-70nts is a known characteristic among microRNAs [6] and was therefore

Table 1: The features used in the Microprocessor processing site predictor include both structural and sequential features to create extensive profiles of the processing sites.

Id	Feature description
1	Precursor length and loop size
2	Distance from the 5' end to loop
3	Occurrence of each base separately the first 24 positions of the precursor
4	Base pair information the first 24 positions of the precursor
5	Total number of each base separately the first 24 positions of the precursor
6	Total number of base pairs the first 24 positions of the precursor
7	Occurrence of each base separately the first 50 positions of the flank
8	Base pair information the first 48 positions of the flank
9	Total number of each base separately the first 50 positions of the flank
10	Total number of base pairs the first 15 positions of the flank
11	Total number of base pairs the first 48 positions of the flank

included. The loop size is the same for each potential processing site for the same microRNA, but should be included due to the potential relationship between loop size and other features, especially precursor length. The precursor length and loop size were considered as the same feature. The distance from the mature region to the loop is important because only one microRNA has a mature region overlapping with the loop. Note that this microRNA (hsa-mir-451) is radically different from the other known human microRNAs, as its annotated precursor is only 42nts long, and is therefore considered as an outlier in the analysis. As this classifier is created to predict processing sites in unknown microRNAs, however, we generally do not know the length or position of the mature region. Dicer's processing of the mature region from the precursor occurs at a later stage than the Microprocessor's processing of the primary transcript. Thus, the final length of the mature microRNA likely has no influence on determining the Microprocessor complex's processing site. The distance from the 5' end to the loop is therefore used instead of the distance from the mature region to the loop.

All features listed in table 1 were used in the analysis. Feature 9 and 10, however, showed to decrease the performance and were therefore excluded in the final version of the SVM, see section 2.5.1 for details. For details about the features and how they were calculated, see section 5.2.3 in Materials and Methods.

2.2 Performance of the Microprocessor processing site SVM

2.2.1 The SVM distinguishes between real and false processing sites

The machine learning method used in this study is the gist implementation of support vector machine (SVM) (`http://benzer.ubic.ca/gist/`), using a gaussian

radial basis kernel function, see section 5.2.5 in Materials and Methods. SVMs have been widely applied to the prediction and classification of important biological signals such as promoters, translation initiation sites, splice sites and proteins [27], and have recently been used in several studies predicting [27, 18, 34] and analysing [29] microRNAs and siRNAs.

To test the performance of the SVM and give an estimate on unknown data, a 10-fold cross-validation experiment was set up, where all potential processing sites for the same microRNAs were put into the same fold to prevent bias in the prediction estimates on unknown data. The known microRNAs fall into families with similar mature sequence and secondary structure [35], and a regular 10-fold cross-validation that randomly assigns microRNAs or processing sites to folds may therefore give a biased estimate. In other words, by assigning microRNAs or processing sites randomly into folds it is most likely that very similar processing sites end up in different folds, which in turn means that the processing site in the test set has partly been seen by the SVM during training. To avoid this, all members of the same family were placed into the same fold. All performance estimates are based on this 10-fold cross-validation procedure unless otherwise specified. See section 2.3 for details about this bias analysis.

The average score for the real processing sites is -0.0839, with a standard deviation of 0.439, while the average score for the false processing sites is -0,968, with a standard deviation of 0.592. It is a large difference in the average score for real and false processing sites, meaning the SVM is able to distinguish real and false processing sites. It is also interesting to see that the standard deviation is lower for the real than false processing sites, which could mean that the real processing sites are scored more stably. The ROC-curve in figure 3 shows that the SVM can truly distinguish between real and false processing sites. For example, a sensitivity of 50.0% yields a specificity as high as 93.8%, which gives a positive prediction rate of 36.1%. Similarly, a sensitivity of 80.0% yields a specificity of 79.9% and a positive prediction rate of 21.8%. The positive prediction rate for the 10, 20, 50 and 100 highest scored processing sites is 50.0% in all cases.

It is a very important factor that several processing sites belong to the same microRNA, which is the motivation for building the processing site SVM in the first place. Although this is unknown for the SVM itself, this may, and probably will, lead to a correlation in the score between the real and false processing sites for the same microRNA. The results in this section only describe the overal performance of the SVM, i.e. the performance independent of which microRNA the different processing sites belong too.

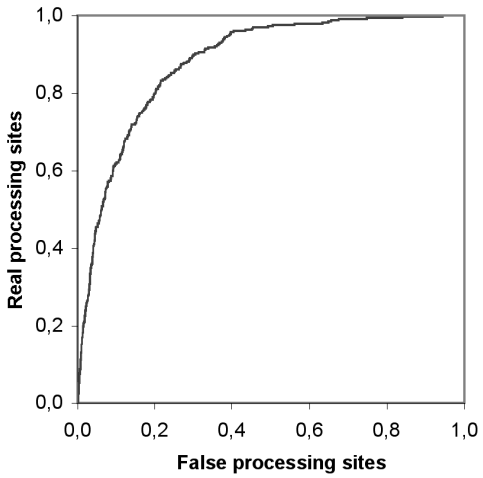


Figure 3: The SVM distinguishes between real and false processing sites. The figure shows the ROC-curve from the 10-fold experiments, where the x-axis shows the relative number of false processing sites and the y-axis shows the relative number of real processing sites.

2.2.2 The SVM predicts correct processing sites for over 43% of the known human microRNAs

The SVM has a good performance at distinguishing between real and false processing sites. However, the intuitive usage of the SVM is to predict which of the microRNA's potential processing sites is the real one; that is, to predict the processing site for any given candidate microRNA primary transcript. To take a closer look at this performance, all processing sites were grouped by their respective microRNA, such that each microRNA is evaluated separately. We designate a prediction to be correct when the real processing site received the highest score, and false otherwise. The prediction rate is the percentage of microRNAs with correct prediction.

The SVM predicted the real processing site for 141 of the 327 microRNAs, yielding a prediction rate as high as 43.1%. Although the prediction rate gives a better indication of the SVM's performance than a simple cutoff value, it does not describe the overall performance for the microRNAs. Therefore, two addition factors describing how well the SVM performed were calculated. First, the distance in nucleotides from the 5' end of the predicted processing site to the 5' end of the real processing site were calculated. The average distance for all the 327 microRNAs is 1.13 nucleotides, which means that the predicted processing site averagely lies within one nucleotide from the real processing site. Figure 4 shows the distribution of the distances between the predicted and real processing sites. Second, the average number of false processing sites that received a higher score than the real was 1.66. Figure 5 shows the overall distribution of the number of processing sites with a higher score than the real processing site.

Both these figures show that even for the microRNAs with false predictions, the

real processing sites are nearby in both distance and score. There is in particular a strong relationship between the position of the real processing site and the predicted processing site, that is, the processing sites close to the real one gets in general a high score. For example, 88.1% of the microRNAs' real processing sites were within a distance of two nucleotides from the predicted processing site.

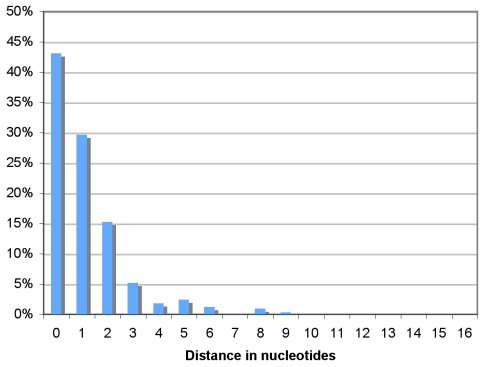


Figure 4: Nearly 90% of the predicted processing sites are within a distance of two nucleotides from the real processing site. Each bar shows the percentage of microRNAs with that particular distance between real and highest scored processing site.

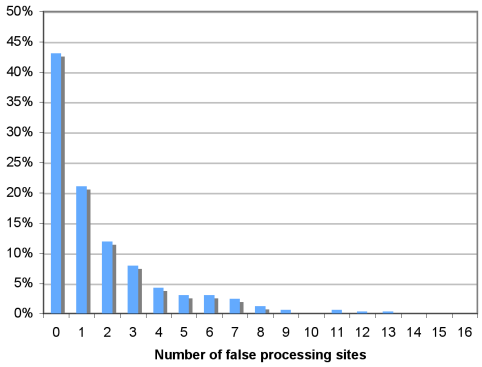


Figure 5: The number of false processing sites with higher score than the real is not random, that is, the real processing site is more frequently found among the top scoring sites. Each bar shows the percentage of microRNAs where that particular number of false processing sites received a higher score than the real site.

2.2.3 Newly discovered microRNAs differ from the previously known

In the beginning of May 2006, while this work was in progress a new version of the Sanger microRNA database was released, containing 130 new human microRNAs. These were used as a separate independent test set.

Of the 130 microRNAs, two, hsa-mir-614 and hsa-mir-639, did not pass the validation step, failing to find any base pairs in the stem (test 2). Further, the algorithms failed to find the real processing site for six microRNAs. Three of these, hsa-mir-553, hsa-mir-563 and hsa-mir-626, were too short, having precursor lengths of 46, 45, and 48 nucleotides. The other three, hsa-mir-591, hsa-mir-611, and hsa-mir-650, found 6, 5, and 2 false processing sites, but not the real site. In total, 122 of the 130 new microRNAs were successfully processed and tested. This set is in the following referred to as the independent test.

The performance for the independent test is shown in table 2. The prediction rate for the independent test is 35.3% lower than the estimate given by the original SVM. Further, the average distance between the predicted and real processing site is increased by 111.4% and the average number of false processing sites with higher

score than the real is increased by 128.1%. This is far from what the estimates suggested. The ROC curve for the independent test is drawn in figure 6, as well as the ROC curve for the original SVM as a reference, confirming the performance difference between the estimate and test set. Detailed predictions of the independent test set can be found in appendix A.

Table 2: The performance of the SVM trained on the original set of 332 microRNAs is much better than on an independent test of 130 new microRNAs and the SVM trained on the original dataset plus the 130 new microRNAs. The table shows the prediction rate, the average distance between predicted and real processing site, and average false processing sites with higher score than the real site for the SVMs and test.

	Pred. rate	**Distance**	**Num false**
Original SVM	43.1%	1.128	1.664
Independent test set	27.9%	2.385	3.795
Retrained SVM (452 microRNAs)	33.4%	1.668	2.281

However, published studies examining these new microRNAs state that they are fundamentally different from the rest of the known microRNAs [20]. These 130 new microRNAs were discovered with an experimental-based approach by sequencing approximately 274,000 cDNA tags [25]. That is, they were first expressed, and then computationally confirmed to be in a hairpin structure. But although they meet the two criteria in the microRNA annotation [12], they are suspicious due to low support among the tags and because only six were differentially expressed in a cell line where Dicer expression was knocked down. Because of this, it has been raised a question about whether these new microRNAs could be erroneously annotated [20].

To investigate the properties of these new microRNAs, the SVM was retrained on the original 332 plus the new 130. Table 2 shows the performance of the retrained SVM, yielding a 22.5% lower prediction rate than the original SVM. Also the average distance between the predicted and the real processing site and the average number of false processing sites with higher score than the real increased. With an increase of 47.6% and 37.4% respectively for there two averages, the SVM shows a great difficulty at predicting the newly released microRNAs, both when tested on the original SVM and trained together with the existing microRNAs.

Figure 6 also includes the ROC-curve for the SVM retrained with the dataset including the new microRNAs. The performance of the retrained SVM is better than the independent test with the 130 new microRNAs. However, the new microRNAs are only 28% of the the total training set, and still decrease the SVM's performance by 22.5% based on the prediction rates. It is therefore not surprising that these 130 new microRNAs also score a lot poorer than the original estimate. These results indicate that the new microRNAs indeed differ from the already known microRNAs. Whether this is due to bias in the set of known microRNAs or whether these new microRNAs might have been misclassified due to the method of discovery remains

Figure 6: The performance of both the independent test with the 130 new microRNAs and the SVM trained with the original dataset plus the 130 new microRNA is lower than the original SVM's estimate.

a question.

2.3 Investigating the bias due to the microRNA family partitioning

2.3.1 A random fold distribution gives a biased estimate

The known microRNAs fall into families with similar mature sequence and secondary structure [35]. Currently 9 of the microRNAs have no family, and the remaining 323 are divided into 157 different families, where the largest family, mir-515, has 41 members.

A common technique to estimate a classifier's performance on unseen data is to use a standard 10-fold cross-validation, where the dataset is randomly divided into ten evenly distributed folds with approximately the same size and the same number of positive examples. Nine of the folds will be used as training while the tenth is used as a test set. This process is repeated ten times such that each fold has been tested once on a classifier trained on data other than the test set. This gives an accurate estimate of the classifier's performance on unknown data [36]. However, this process assumes there is no correlation between the examples in the different folds. In our case, where the microRNAs are divided into families according to certain characteristics, this assumption is most likely wrong. The standard 10-fold cross-validation can therefore give a too optimistic estimate.

To study this closer, as well as try to estimate this bias, we compared the performance of the SVM trained using two different fold distributions. One, a random

distribution of the microRNA, further referred to as the standard distribution, and two, a distribution where all members of the entire family were put in the same fold, referred to as the family distribution. The prediction rate estimate for the standard distribution is 50.2%, compared to the 43.1% for the family distribution, see table 3. Despite that the standard distribution and family distribution use the same datasets and SVM, the standard distribution gives an estimate 16.3% higher than the family distribution. Similarly, the performance for the average distance between predicted and real processing site and average false processing sites with higher score than the real site is 16.5% and 19.3% lower, respectively. The ROC curves for both distributions are drawn in figure 7, confirming that the standard distribution gives a higher performance estimate.

Table 3: A fold distribution where the entire families are put into the same fold gives a more realistic performance estimate than a standard fold distribution, and is a better SVM than an SVM using only one microRNA from each family. The table shows the prediction rate, the average distance between predicted and real processing site, and average false processing sites with higher score than the real site for the SVMs and test.

	Pred. rate	**Distance**	**Num false**
Normal SVM, standard distribution	50.2%	0.942	1.343
Normal SVM, std. dist, famsize 1 only	41.4%	1.207	1.721
Normal SVM, std. dist, famsize 2 only	55.9%	0.882	1.397
Normal SVM, std. dist, famsize 3 only	53.2%	0.702	0.872
Normal SVM, family distribution	43.1%	1.128	1.664
Family SVM	38.7%	1.276	1.816
Family SVM test set	43.3%	1.110	1.634

To investigate this difference in more detail, the microRNAs from the standard distribution were grouped together by the number of members in their families; that is, the family size. To be more specific, as five microRNAs were missing in the dataset and several microRNAs in the same family were put into the same folds, each microRNA was grouped by the number of family members that were present in other folds than it's own fold. For example, if two microRNAs are in the same family and both are put into the same fold, these will both be considered as family size 1 microRNAs since they are tested on an SVM not trained on any microRNAs from that family. Figure 8 shows the estimated prediction rate for each family size, with the total prediction rate of 50.2% drawn as a reference line.

When only considering the microRNAs with a family size of 1, the prediction rate is 41.4%, which is considerably lower than the overall estimate and indicates that the families contribute to a positive bias. As a comparison, the prediction rates on all the different family sizes except family size 4 are higher than the prediction rate for family size 1. Especially family size 2 and 3 give us a good indication of how large the bias is, as these contain both a fair number of microRNAs and a fair number of different families. The prediction rate for the group of family size

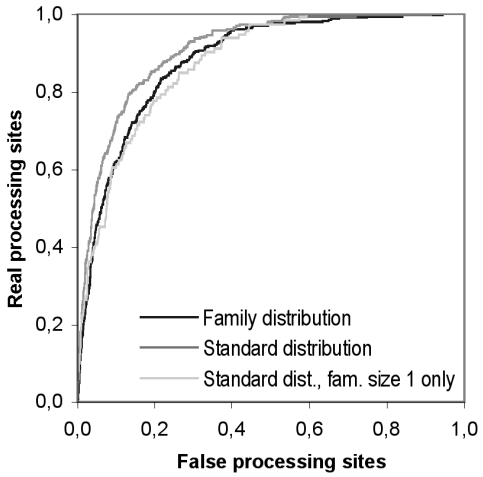


Figure 7: A standard fold distribution of the microRNA gives bias, resulting in a overestimated accuracy on unseen data. The figure shows the ROC curves describing the 10-fold estimate when the microRNAs in the same family are put into the same folds (family distribution), and when the microRNAs are randomly put into folds (standard distribution). Additionally, the figure shows the ROC curve when only microRNAs with family size 1 in the standard distribution are considered.

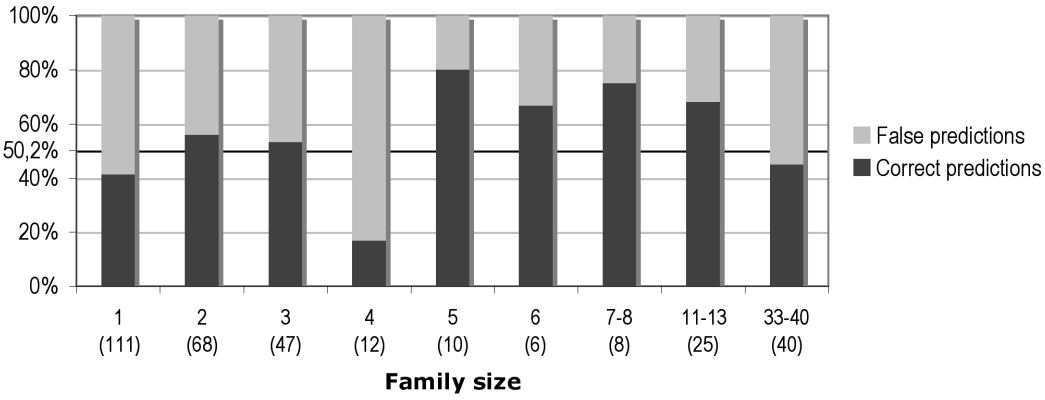


Figure 8: Dividing the microRNAs in folds randomly gives bias in the prediction estimate, as microRNA with large family size tend to get higher prediction rate than those with family size 1.

2 and 3 microRNAs is 32.2% higher than for the family size 1 microRNAs. Similarly, the average distance between predicted and real processing site and average false processing sites with higher score than the real site is 33.0% and 31.3% lower, respectively. In other words, the SVM performs approximately 30% better on microRNAs in the same family as microRNAs used for training compared to unknown microRNA families, which is important to consider in a performance estimate of unknown data.

An estimate based on the assumption that unknown microRNAs are not in family with the already known will remove the positive bias and give a much more realistic performance estimate on unknown microRNAs. As there might be undiscovered microRNAs that are in family with known microRNAs, however, the estimate based on such an assumption might be an underestimate.

The prediction rate estimate for the family size 1 microRNAs is as mentioned 41.4%, which is 3.9% lower than the family distribution. The average distance between the predicted and real processing site and the average number of false processing site with higher score than the real is 7.0% and 3.4% higher, respectively. The ROC curve for the family size 1 microRNAs is drawn in figure 7, giving a ROC area 0.6% less than the ROC area for the family distribution. Although this is not a significantly lower performance this could indicate that there is a larger correlation between the known families than correlation between the microRNAs with no other family members. The estimate given by the family distribution involves the performance of the entire set, while the estimate given by the family size 1 microRNAs only involves one third. As the results indicate that the family size 1 microRNAs might not represent the entire set of microRNA families, the family distribution estimate is therefore the best estimate on unknown data for this SVM.

2.3.2 The SVM performs better when trained on all microRNAs rather than on one per family

Distributing entire families into the same fold removes bias in the performance estimate, as shown in the previous section. Another way of dealing with the bias is to train the SVM on only one member from each family. In additional to removing bias in the estimate, this will also remove bias in the prediction of unknown data. That is, since the microRNAs in the same family have very similar characteristics, the characteristics for the large-family microRNAs might be weighted heavier by the SVM than single-family microRNAs, and therefore reduce the SVM's performance on new families. On the other hand, training on only one microRNA from each family will reduce the size of the training set, which in turn may lead to a poorer performance.

The SVM trained using a dataset with only one microRNA from each family is here

referred to as the family SVM, and the SVM trained on the entire set of microRNAs for normal SVM. The prediction rate for the normal SVM is the estimate given by the family distribution, and should therefore be directly comparable with the estimates from the family SVM. There are in total 166 families (version 8.1 of the microRNA registry). However, three of the five microRNAs that missed the real processing site have a family size of 1 (hsa-mir-198, hsa-mir-451 and hsa-mir-448) and only 163 of the families are therefore processed, yielding 163 real processing sites and 2,292 false processing sites. A list a of the microRNAs included in the family SVM can be found in table B-1 in the appendix.

As shown in table 3, the prediction rate for the family SVM is 38.7%, which is a performance 10.4% lower than the normal SVM. Similarly, the family SVM has an average distance from the predicted and the real processing site of 1.28 nucleotides and an average number of false processing sites with a higher score than the real of 1.82 processing sites, which is 13.1% and 9.2% higher than the normal SVM. The ROC curves for the normal and family SVM are drawn in figure 9, showing a slightly better performance for normal SVM.

As the family SVM only uses 163 microRNAs as training set, the remaining 164 were used as a separate test set. Although this test set has not been used to train the SVM it is important to note that this test set is not independent, as it completely consists of microRNAs that are in families with exactly one microRNA already used to train the SVM. In other words, a performance above the family SVM estimate is expected. To be more precise, the expected performance for the test set is the same as the performance for the family size 2 microRNAs of the normal SVM using a standard distribution, as also this set of microRNAs have been tested on an SVM trained on exactly one microRNA in the same family.

The performance rate for this test is 43.3%, while the average distance from the predicted and the real processing site and average number of false processing sites with a higher score than the real is 1.11 and 1.63, respectively. Compared to the family size 2 microRNAs this is a 17%-26% lower performance. In fact, the performance of the test is even only 0%-2% better than the normal SVM using the family distribution. The performance of an SVM trained on the entire set of human microRNAs is therefore better than the performance of an SVM trained on only one human microRNA from each family.

2.4 The processing site SVM can be used for microRNA gene prediction

2.4.1 Extracting six million hairpins from the human genome

To fully understand the regulation mechanism microRNAs are a part of, it is also very important to discover the unknown microRNA genes, which has been a hot

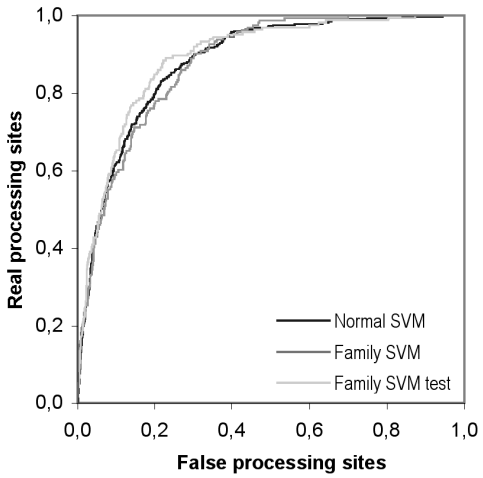


Figure 9: An SVM trained on the entire set of human microRNA gives a slightly higher performance than an SVM trained on only one microRNA from each family. Also the family SVM test gives poorer performance than expected.

topic the past 4-5 years. Although this SVM was built and trained to predict Microprocessor processing sites, it is interesting to see how well it performs at distinguishing microRNA genes and random hairpins. To investigate this, we set up an experiment to test whether one can use the score of the best processing site of a given sequence to separate microRNAs from random hairpins. In other words, the predicted processing site, i.e. the processing site with the highest score, for a microRNA represents the microRNA itself. Likewise, the predicted processing site for each of the random hairpins represents the hairpins.

An initial set of hairpin candidates were extracted from the entire human genome using a program called ScorePin [23]. Using a threshold score of 110, the program found a set of 8,556,723 hairpin candidates. This set includes 326 of the 332 known microRNAs; only hsa-mir-98, hsa-mir-198, hsa-mir-134, hsa-mir-384, hsa-mir-425, and hsa-mir-484 were missed. At this early stage it is important to maintain as high sensitivity as possible, as latter steps including more variables than only the structure will be able to distinguish microRNA hairpins from random hairpins more precisely [23]. Although this threshold does not give the optimal trade off between sensitivity and specificity, it gives a sensitivity as high as 98% and a reasonable number of random hairpins, relative to previous published results [24, 26], and was therefore chosen to gather the initial set of hairpin candidates.

ScorePin is developed with special emphasis to extract hairpins from the entire genome in a resonable time. To do this, ScorePin uses an edit distance-based algorithm instead of a traditional secondary structure tool, scoring each hairpin candidate by how well it may be folded as a hairpin structure. Although the program showed a great performance, it does not guarantee that the predicted secondary structure for the hairpin candidate actually is a hairpin. It is therefore useful to

run the random hairpin candidates through a verification, not only to eliminate falsely predicted hairpins, but also to verify that the sequences have predicted secondary structures that can be processed by the program. The verification step for the random hairpins is therefore the same as the verification step for the microRNAs, as described in section 2.1.1. 20.5% of the random hairpins were removed in this verification step; 2.1% at test 1 and 18.5% at test 2. As a comparison, 0% of the microRNAs were removed in this step. Note, however, that only the top predicted secondary structure for each candidate is used in this filtering step, so some true hairpins might be lost due to alternatively predicted secondary structures.

In total 6,798,341 random hairpins were processed, resulting in 75,124,493 potential processing sites.

2.4.2 The SVM distinguishes random hairpins and microRNA genes

The set of random hairpins was scored by the SVM trained on the entire set of known human microRNAs. The score for a random hairpin was, as mentioned, the score of the highest scoring processing site for that random hairpin. Similarly, the score of the predicted processing sites were the scores for the microRNAs. As the score for a microRNA is independent for the real processing site, the five microRNAs that did not give the real processing site were also included in this experiment. This is consistent with an approach to discover new microRNA genes, in which we do not know whether or not the correct processing sites have been included.

Figure 10 shows the ROC curve describing the performance of the SVM at distinguishing microRNAs and random hairpins. For a desired sensitivity, the ROC-graph shows that the SVM can separate random hairpins from microRNAs with high specificity. For example, a sensitivity of 10.1% gives a specificity 99.9751%. In other words, a total number of 1,667 random hairpins have as high score as 10.1% of the microRNAs. These random hairpins should be considered as microRNA candidates. In fact, 322 of these random hairpins are known microRNAs. As ScorePin uses a fairly small window size of 14 nucleotides, several high-scoring predicted hairpins could in reality be the same hairpin. This is also the case for the known microRNAs in the initial set of random hairpins. These 322 known microRNAs are in reality only 285 different microRNAs, as 37 is represented twice. Using an estimate of a total of 1,000 human microRNAs, there are 668 microRNAs that are unseen by the SVM. As ScorePin has a sensitivity of 98.2%, it is expected that 655 of these 668 unseen microRNAs are included in the initial set of random hairpins. Based on the 10-fold cross-validation estimate for unknown microRNAs, 10.1% (66) of these 655 unknown microRNAs are expected to be present among the 1,345 novel microRNA candidates, that is, 4.9% of these 1,345 are expected to be novel microRNA genes.

Further, if we increase the sensitivity to 30.0% this gives a specificity of 99.738%,

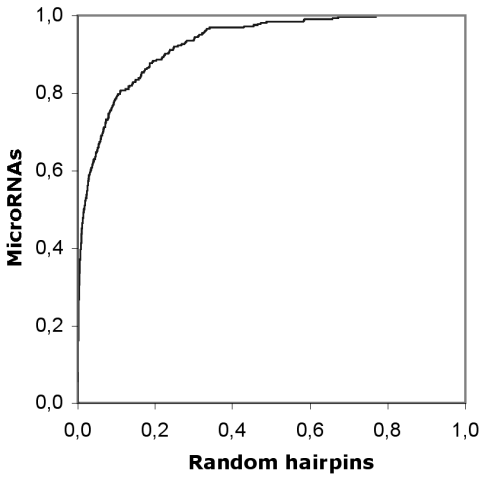


Figure 10: The SVM distinguishes very well between microRNA and random hairpins, independent of which processing sites belong to which microRNAs.

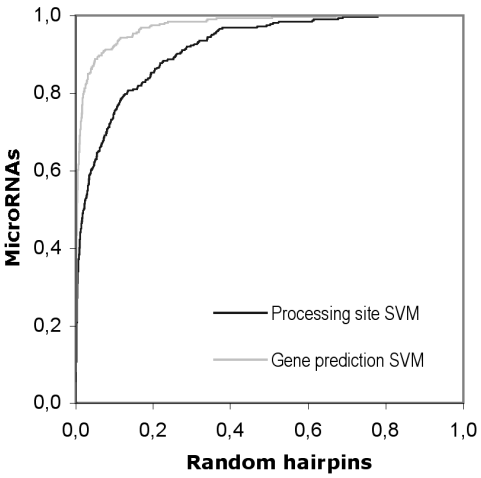


Figure 11: For gene prediction, an SVM trained to distinguish microRNAs and random hairpins gives a much better performance than the original SVM trained to predict the processing site.

which means 17,427 microRNA candidates, of which 330 are known microRNAs. Using the same estimates, the number of expected novel microRNAs among these 17,177 novel microRNA candidates is 197, or 1.1%. The highest percentage of expected novel microRNAs among the microRNA candidates is 30.8%, and is found at a sensitivity of 0.612% and specificity of 99.99893%, yielding 72 microRNA candidates, of which 59 are known microRNAs. In other words, there are 13 novel microRNA candidates, of which 4 are expected to be microRNAs.

A comparison of the microRNA scores and hairpin scores shows that there are some major differences, especially in the score for the highest scoring processing site and the standard deviation of the scores for a candidate's processing sites. The difference in the score of the predicted processing site is reflected in the ROC curve. The average score for the predicted processing site is 0.0422 and -0.454 for the microRNAs and hairpins. Thus, microRNAs are on average scored higher than random hairpins. The standard deviation of the scores from the candidates' processing sites shows a striking difference between the microRNAs and random hairpins. For the microRNAs, the standard deviation is much higher (0.518) than for the random hairpins (0.249). This indicates that some of the microRNAs' processing sites receive very good scores, while others very bad scores. To take a closer look at this, the difference between the score for the predicted processing site and the average score for all of the candidate' processing sites was calculated. This average difference is nearly 60% lower for the random hairpins (0.375) than the microRNAs (0.910). These values give valuable information about how likely the predicted processing site is compared to the rest of the potential processing

sites, and in turn how likely the candidate is to be processed by the Microprocessor in the first place. In other words, by using the SVM to predict a candidate's processing site, one additionally gains valuable information that can be used to separate microRNAs from random genomic hairpins in a latter step.

2.4.3 An SVM trained to distinguish microRNAs and random hairpins gives a much better performance

The approach to identify a set of microRNA candidates described in the previous section uses only the score of the predicted processing site as a separation criteria, and can therefore be seen as a parallell step to the processing site prediction. However, an SVM that includes the predicted processing sites alone and trains explicitly to distinguish microRNAs and random hairpins is expected to give a much higher performance, as random hairpins now are a part of the training.

To investigate this, a new SVM, called the gene prediction SVM, was created by using the same features as the processing site SVM. However, the new SVM was trained with the predicted processing sites in the known human microRNAs as positive examples and the predicted processing sites in 3,000 random hairpins as negative examples. A test set of 17,000 random hairpins representative for the entire genome was used to test the performance of the new SVM. Figure 11 shows the performance of the new SVM compared to the processing site SVM on the same test set. As we can see from the figure, there is a drastic difference in performance. This illustrates the importance of first predicting the processing site as a separate step and then create a new SVM based on this information. An SVM including other features that are shown to be characteristic for the microRNAs is expected to give an even higher performance. These could be features that distinguish microRNAs from hairpins but are independent of the processing site and therefore not included in the processing site SVM, such as free energy, symmetry of hairpin and stability of the secondary structure.

As described in the previous section, the processing site SVM gave a sensitivity of 30.0% at 99.738% specificity. Based on the two SVMs' performance on the test set, we can estimate the gene prediction SVM's performance on the entire genome. At this specificity the gene prediction SVM gives a 86.7% better sensitivity than the processing site SVM on the same test set. Converted to the entire genome, a specificity of 99.738% means approximately 17,177 novel microRNA candidates. The sensitivity for the processing site SVM at this specificity was as mentioned 30.0%, and with an increase in the sensitivity of 86.7% the estimated sensitivity for the gene prediction SVM is 56.0% at this specificity. In other words, over half of the unknown microRNAs are expected to be among this set of $\sim$17,000 microRNA candidates. Based on the estimate of 1,000 human microRNAs in total, 367 novel microRNAs are expected in the set of 17,177 top-scoring random hairpins predicted by the gene prediction SVM.

2.5 Analysing which features are the most important in the SVM

2.5.1 The precursor and loop size, and base occurrences are most important for the processing site SVM

It is interesting to see which features are most important for the processing site prediction. First, this gives a good idea of what features characterise the real microRNAs, and second, it might give an indication of features that are important for the Microprocessor processing itself. Thus, the features described in table 1 were removed one at the time and a new SVM was trained with the remaining features. The performance is measured by the following four variables: 1) the ROC area given by the ROC curve of real and false processing sites independent of which microRNA the processing sites belong to; 2) the prediction rate; 3) the average distance from the predicted to the real processing site, called average distance for short; and 4) the average number of false processing sites found before the real, called average false for short. The average distance and average false are based on one calculation from each of the 327 microRNAs, and thus, the importance is best described by a statistical significance test. However, none of these variables are normal distributed and both have a very low distribution range. It it therefore very difficult to calculate any good statistical significance. The variables' relative changes give, however, a good indication of the features' importance.

Table 4 shows the results as the percentage change compared to the SVM based on all features. As we can see from the table, three features, 1, 3 and 7, gave a much lower performance when removed. Especially feature 1, which is the precursor and loop size, had a drastic impact on all four variables. It is, for instance, very interesting to see that this feature impacts the ROC area, average distance and average false over twice as much as the other features. These results show that the precursor and loop size is a very important characteristic for the microRNAs. Further, feature 3 and 7, which are occurrences of the bases at each position in the stems (3) and at the flank (7), show a lower performance in nearly all variables. This indicates that also these two features should be considered important for predicting the processing site. In addition, feature 8, the base-pair information for the 48 first bases at the flank, gives a lower performance, although not as low as the three others.

Five features, 2, 4, 5, 6, and 11, gave a small performance change. When removed, almost all of them gave a higher average distance and average false, which indicate that the features are indeed helping the SVM, although not dramatically. Removing features 4, 5, or 6, though, results in a higher prediction rate, which indicates that these features may create bias instead. To investigate this more carefully, all of these three features were removed and a new SVM was trained. The result (see table 4) showed a lower performance for this SVM, which indicates that these three features also help the SVM.

Table 4: Some features are more important for predicting the processing site than others. Each feature was separately removed from the SVM. For each feature removed the SVM was re-trained to measure the feature's importance for processing site prediction, and used to classify a test set to measure the feature's importance in distinguishing microRNAs and random hairpins. The performances are given as the percentage change compared to the SVM based on all features.

Feature	**Processing site prediction**				**Gene pred**
removed	**ROC area**	**Pred. rate**	**Avg dist.**	**Avg false**	**ROC area**
1	-5.3%	-13.4%	25.2%	26.6%	2.5%/-4.5%
2	-1.5%	0.0%	2.0%	7.2%	0.6%
3	-0.8%	-7.9%	-0.5%	10.2%	1.0%
4	-0.5%	1.6%	2.3%	3.0%	0.9%
5	-0.4%	1.6%	1.3%	4.3%	-2.2%
6	-0.3%	1.6%	0.5%	-1.8%	-0.9%
7	-2.5%	-15.0%	6.0%	12.3%	3.3%
8	-0.8%	-3.9%	4.0%	10.0%	-6,5%
9	1.0%	7.9%	-3.0%	-3.2%	-0.8%
10	-0.2%	3.9%	-0.7%	-0.4%	-5.3%
11	0.3%	-1.6%	0.5%	0.5%	-0.8%
4, 5 and 6	-1.6%	-3.9%	8.2%	13.4%	-
9 and 10	0.9%	11.0%	-7.8%	-2.7%	-

Finally, the last two features, 9 and 10, gave a better performance when removed, which suggests that these features should not be included in the processing site SVM at all. A new SVM without these two features were created, showing a much better performance (see table 4). This version of the SVM was therefore used as the final version of the SVM.

In summary, the precursor and loop size, as well as the base occurrences in the entire primary microRNA are the most important features for processing site prediction. Further, the total number of each of the bases in the flank and the total number of base-pairs at the first 15nts in the flank give bias and lower performance of the SVM.

2.5.2 Loop size and base-pair information in the flank are most important for gene prediction

The previous section looked at the features' importance for processing site prediction. It is also of great interest to determine which features are important for distinguishing microRNAs and random hairpins. As the features originally were selected to predict processing sites, this analysis is not complete. That is, it will only look at the features used in this study and not all potential features for predicting microRNAs. However, it will indicate which features that are important for microRNAs in general, and give a hint at which direction one should look when

gathering features to distinguish microRNAs and random hairpins.

A test set of 20,000 random hairpins representative for the human genome was tested on each of the SVMs created in the previous section. As the random hairpins do not have any real processing sites, the only performance variable we can use is the ROC area given by the ROC curve of microRNAs and random hairpins. The score for the mircoRNAs and random hairpins are based on the predicted processing sites for each of the microRNAs and random hairpins. Table 4 shows each SVM's performance change in percentage of the performance of the SVM based on all features. As the performance now is measured by only one processing site from each hairpin, the precursor length and loop size should be considered as two different features and were therefore tested separately. Removing the precursor length gives a performance 2.5% higher than when the precursor length was included. Removing the loop size gives a performance 4.5% lower.

Three features; 1(II), 8 and 10, gave a performance lower than 3% when removed. Feature 1(II) is the loop size. Features 8 and 10 is the base-pair information in the flanking region and the total number of base-pairs in the first 15 positions of the flank which are consistent with other studies [29].

Only one feature gave a performance increase higher than 3% when removed, which is the base occurrences in the flanking region (feature 7). In contrast, feature 7 was one of the most important features when predicting the processing site. The reason for this might be the large number of vector columns the features gives. With 50 positions at each stem, each describing four different nucleotide types, this yields 400 columns, which easily could create bias in such a relatively small test set. That is, even though the base occurrences are important for fine-tuning the Microprocessor processing site, other features are more important for separating random hairpins from microRNAs. Random hairpins can for example have the correct base-occurrences, but miss important secondary structure elements that are essential for microRNAs.

The rest of the features only give a small change in performance. That is, as the decision is made from only one performance variable, there might be relative large bias and a performance change under 3% should therefore be considered as a small change.

In summary, the loop size and the base-pair information in the flanking region, especially the first 15 bases, are the most important for distinguishing microRNAs and random hairpins. Further, the base occurrences in the flanking region create bias and should therefore be excluded. Note, however, that although these features are important when the SVM is used to distinguish microRNAs and random hairpins, it does not directly indicate that these features are the most important features for microRNA discovery in general. That is, in this case the SVM distinguishes microRNAs and random hairpins in parallell to predicting the best processing site for the candidates. An SVM trained to distinguish microRNAs and random hairpins

given only the best processing site from each candidate would give a much better indication of the features' importance in microRNA discovery.

2.6 The SVM also predicts processing sites in other species

The SVM has so far only been trained and tested on human microRNAs, but it is also of great interest to see how it performs on other species, both closely and distantly related to human. 1,976 microRNAs from 15 species were used as a test. Of these, 118 have not yet been mapped to a position in the genome and can therefore not be classified by the SVM. Additionally, 15 microRNAs were too short, 26 too long, and 14 did not find the real processing site due to a predicted secondary structure that was different from the annotated secondary structure. All these microRNAs passed the verification step. In total 1,803 of the microRNAs were processed and classified by the SVM. The results for each specie are listed in table 5.

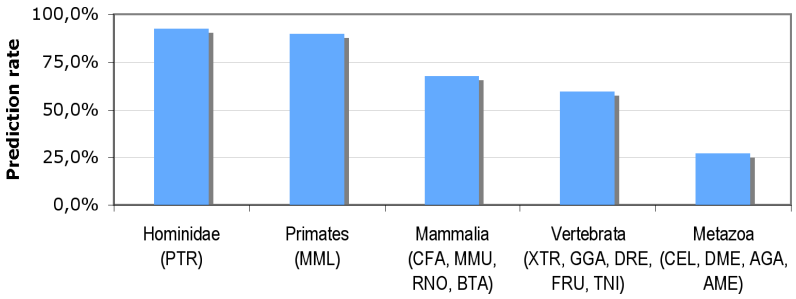


Figure 12: The prediction rate for other species decreases with evolutionary distance from humans. The figure shows the total prediction rate for the different evolutionary steps from human.

First of all, the prediction rate for 11 of these 15 species is much better than the estimate for human microRNAs. However, this is not very surprising as many of these microRNAs are homologs or in family with the human microRNAs, meaning they got a very simliar mature sequence and structure. For a more in-depth study, the predictions should be independent of conservation. That is, the microRNAs should be classified on an SVM strictly trained on microRNAs in other families. These results give, however, still a good impression of the performance on other species.

To take a closer look on the relationship between diversity and prediction rate, the species were grouped according to the diversity; see figure 12. The figure indicates a higher prediction rate the closer the species are to human. We find a best prediction rate in chimpanzee (hominidae) which has a genome highly conserved with the human genome. We find the lowest prediction rate, which is even lower than the estimate for unknown human microRNAs, in the non-vertebrate species (metazoa). Some microRNA characteristics, such as the base-paring in the flanking region differ

Table 5: The SVM predicts processing sites for other species, but gives a lower performance the more evolutionary distant a specie is to human. The table shows how many of the microRNAs found in the microRNA registry were removed due to an unknown location in the genome, a too short precursor length, a too large precursor length, and because the real processing site was not extracted. The SVM prediction rate describes the percentage of processed microRNAs that were correct predicted, while the total prediction rate describes the percentage of microRNAs with known location that were correct predicted, which means the microRNAs that did not make it through the processing step are considered as a false prediction.

Specie	**Sym**	**Num of miRNA**	**Unknown location**	**Too short**	**Too large**	**Real not found**	**Processed miRNA**	**Total pred rate**	**SVM pred rate**
Anopheles gambiae	AGA	38	1	1	0	0	36	29.7%	30.6%
Apis mellifera	AME	25	1	0	0	0	24	37.5%	37.5%
Bos taurus	BTA	33	11	0	0	0	22	77.3%	77.3%
Caenorhabditis elegans	CEL	114	1	2	3	1	107	21.2%	22.4%
Canis familiaris	CFA	6	0	0	1	0	5	50.0%	60.0%
Danio rerio	DRE	372	62	3	2	0	305	54.8%	55.7%
Drosophila melanogaster	DME	78	1	0	3	1	73	27.3%	28.8%
Fugu rubripes	FRU	130	0	0	0	0	130	55.4%	55.4%
Gallus gallus	GGA	144	0	0	0	0	144	75.7%	75.7%
Macaca mulatta	MML	71	9	0	0	3	59	85.5%	89.8%
Mus musculus	MMU	340	10	6	15	6	303	59.1%	64.4%
Pan troglodytes	PTR	83	16	0	0	0	67	92.5%	92.5%
Rattus norvegicus	RNO	234	6	2	1	2	223	69.7%	71.3%
Tetraodon nigroviridis	TNI	131	0	0	1	1	129	55.0%	55.8%
Xenopus tropicalis	XTR	177	0	1	0	0	176	58.8%	59.1%

Table 6: The gene prediction SVM gives a better performance than the other published prediction-based approaches. The table shows the sensitivity for both the processing site SVM and the gene prediction SVM at the same specificity as the other approaches.

		Sensitivity		
Approach	**Specificity**	**Approach**	**Proc site SVM**	**Gene pred SVM**
Sewer et al.	97%	71%	59%	89%
ProMir	96%	73%	62%	90%
Xue et al.	88.1%	93.3%	80.7%	96.5%

in metazoa compared to human [29]. The microRNAs in the non-vertebrate species are processed by a Microprocessor that might have different characteristics from the Microprocessor processing human microRNAs; for example, the Microprocessor is larger (650kDa) in humans than in Drosophila melanogaster (500kDa) [9].

3 Discussion

3.1 The gene prediction SVM performs better than other prediction-based MicroRNA discovery approaches

To get an indication how the processing site SVM and gene prediction SVM perform compared to other methods, we compared our results with the three other prediction-based microRNA discovery approaches that estimate the sensitivity and specificity. The sensitivity was chosen as the comparison variable; that is, the sensitivity of the other approaches is compared to the SVMs' sensitivity at the same specificity. As all approaches, including ours, use different sets of hairpins and different number of microRNAs in the training set, there might be some bias in the comparisons. We consider this bias as relatively small, as all approaches used test sets that are fairly representable for the entire genome.

Table 6 shows the SVMs' performance compared to the three other published approaches. The performance of the processing site SVM for gene prediction is much lower than the other approaches. As the processing site SVM was only trained to distinguish small differences between real and false processing sites and have not seen any random hairpins during training, the lower performance is expected. The gene prediction SVM, however, score much better than current approaches. Despite the excellent performance, the gene prediction SVM was only meant as a comparison to the processing site SVM on microRNA discovery and to illustrate the potential gain of using the predictions of the processing site SVM in microRNA gene prediction. The features used in the gene prediction SVM are therefore the same as the ones used in the processing site SVM analyses. We expect even better predictions when including other and more general microRNA characteristics, such

as bulge symmetry and hairpin stability, and the statistics from the initial processing site analysis. Nevertheless, these results illustrate the benefits of predicting the potential Microprocessor processing site in hairpin candidates before predicting whether the hairpin candidates are microRNAs.

3.2 Investigating the base-pair conservation in the flanking region

A recent study by Han et al. [28] suggests that the microRNA primary transcripts are cleaved ∼13nts (at the 5' stem) from the junction where the double-stranded hairpin structures (dsRNA) splits into two single-stranded sequences (ssRNA). These results are consistent with Sætrom et al.'s work [29] that found that the base-pair structure is conserved the first 13 bases of the flanking region, but not beyond this point.

However, our results do not indicate a particular strong importance of this feature. In fact, the results indicated the opposite; feature 10, the total number of base-pairs the first 15nts of the flanking region, creates bias in the processing site SVM and was therefore removed from the final version. The reason for this bias may be that feature 10 only describes the base-pair conservation at the beginning of the flanking region, and not the relationship between the beginning and the other parts. This can therefore create bias by for example favoring short hairpins over long hairpins.

A feature describing the distance between the junction and the processing site would be an optimal replacement for feature 10. However, this turned out to be very difficult due to the large bias in the secondary structure predictions. It has for example been shown that 8 out of 10 experimentally investigated microRNA secondary structures were different from the predicted secondary structure, especially in the areas around the loop and flanking region [31]. The bias in secondary structure prediction tools increases exponentially as a function of the sequence length. As a relatively large flanking region is required to find the dsRNA-ssRNA junction, the task become very difficult. Besides, it is a difficult task to define computationally where the dsRNA-ssRNA junction is.

As an alternative solution, feature 10 was replaced by two new features: 1) the total number of base-pairs the first 11 bases of the flanking region, counted from the first position in the 3' stem; and 2) the total number of base-pairs the 11 next bases of the flanking region, that is, at position 12 inclusive to 22 inclusive, relative to the 3' stem. It is expected that these two features will describe the conservation in the flanking region better than the original feature 10, as they not only describe the number of base-pairs up to the expected dsRNA-ssRNA junction, but also the number of base-pairs the same number of positions after the junction.

However, the results show only a small performance improvement. The ROC area

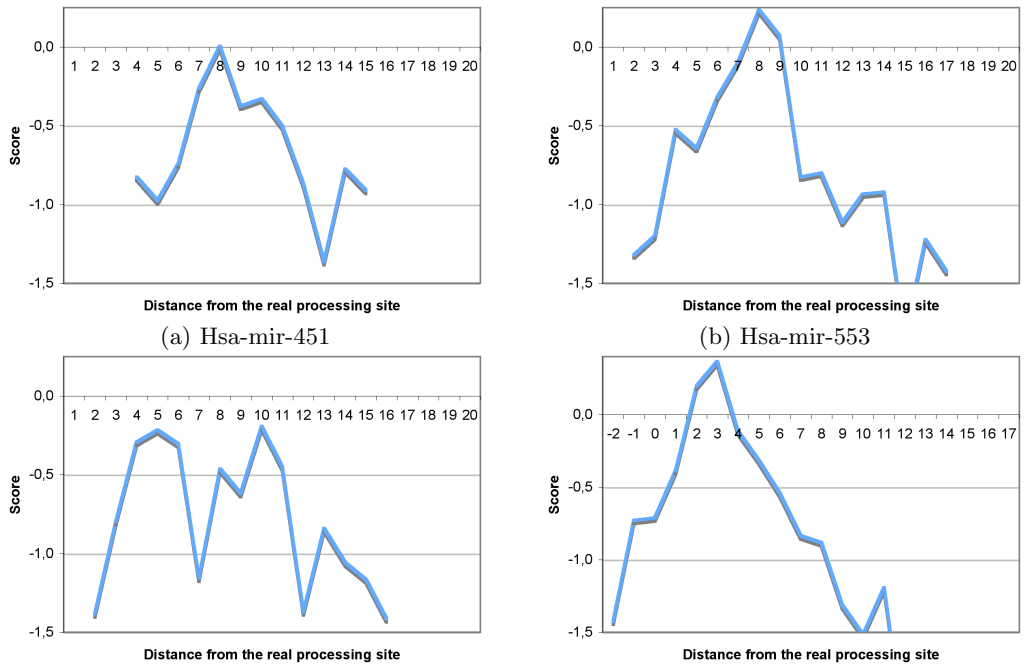
and prediction rate were increased by 0.1% and 0.8%, while the average distance between the predicted and real processing site and the average number of false processing sites with higher score than the real were decreased by 3.2% and 5.2%, respectively. This indicates that the processing site SVM does not find any strong relation between the processing site and the number of base-pairing in the beginning of the flank. One explanation could be that the base-pairing information in this region is already covered by feature 8. It could be very interesting to see a more in-depth study of the flanking region.

3.3 Examining four short microRNAs that may have falsely annotated processing sites

The results from Han et al. also suggests that the microRNA primary transcripts have two processing sites. The processing site ∼13nts from the dsRNA-ssRNA juction is called the productive processing site. Additionally, the primary microRNAs can occasionally be cleaved at a distance of ∼13nts from the terminal loop, called the abortive processing site. In the abortive processing the loop is mistaken as the dsRNA-ssRNA junction, which occurs more frequently when there are small internal loops near the productive processing site and lack of the same near the abortive processing site. In other words, the abortive processing site is more likely to be used when the precursor stems are joined together by a perfect or close to perfect base-pairing.

Four human microRNAs have a precursor length shorter than 50 nucleotides, and were considered as outliers in our study. Three of these (hsa-mir-553, hsa-mir-563 and hsa-mir-626) are newly predicted in a set of microRNA that are shown to be fundamentally different from the previous known microRNAs [20]. The fourth (hsa-mir-541) is radically different from the other microRNAs as the mature region continues through the entire predicted terminal loop. The question whether these four microRNAs are correctly annotated should therefore be raised. Figure 13 shows a detailed analysis of three of these four microRNAs, where each microRNA's score is plotted as a function of the processing sites' distance from the real processing site. The SVM predicts a singel processing site region for hsa-mir-451 and hsa-mir-533, with the predicted processing site a distance of 8nts from the annotated processing site (figure 13a and 13b). The SVM did not find such a singel region for hsa-mir-563 figure 13c). Hsa-mir-626 was not analysed due to an insufficient number of potential processing sites.

As both hsa-mir-451 and hsa-mir-553 have a perfectly base-paired stem this could indicate that the abortive processing of these two microRNAs has been detected and annotated instead of the productive processing site, which the SVM predicts 8nts downstream from the currently annotated site. Also a fifth short hairpin, hsa-mir-617, differ from the rest. Although it has a precursor length of 52, the mature region overlaps with the loop, which is unusual. Similar to hsa-mir-451 and



(a) Hsa-mir-451 (b) Hsa-mir-553

(c) Hsa-mir-563 (d) Hsa-mir-617

Figure 13: The SVM predicts a single processing site region for three of the four analysed short microRNAs, including a predicted processing site with a relatively high score. The figures show the score as a function of the processing sites' distance from the annotated site.

hsa-mir-553, the SVM predicts a singel processing site region with the predicted processing site a distance of 3nts downstream from the annotated processing site; see figure 13d.

In summary, the results from the processing site SVM combined with recent research indicate that three short microRNAs (hsa-mir-451, hsa-mir-553, and hsa-617) might have a falsely annotated processing sites. That is, it might be the abortive processing that cleaved the microRNAs when their expression were detected. Further investigation of these microRNAs is encouraged. The forth examined short microRNA (hsa-mir-663) also differ from the previously known microRNAs. However, the processing site SVM does not clearly indicate that the abortive processing site has been annotated for this microRNA.

4 Summary & Conslusion

This work describes an approach to predict the Microprocessor processing site in any given microRNA candidate. Eleven different features from the entire primary microRNA transcript were calculated. These features were related to the Micropro-

cessor processing site and include both structural features such as percursor length, loop size, distance from the 5' end to the loop and base-pair information, and sequence features such as occurrences of the different nucleotides at each position in the primary microRNA. An SVM was then trained with the known human microRNAs' real processing site as positive examples and all other possible processing sites in the known human microRNAs as negative examples.

The SVM correctly predicts the annotated processing site for 43% of the known human microRNAs, based on a performance estimate from 10-fold cross-validation. To avoid a positive biased estimate due to closely related microRNAs, both real and false processing sites from all microRNAs in the same family were placed in the same fold. The predicted processing site is on average within a distance of one nucleotide from the real processing site, which means that for many of the cases where the SVM falsely predict a processing site, it is close enough to the real site, to be used in a latter step of gene prediction.

The SVM was also used to distinguish microRNAs and random hairpins extracted from the entire human genome. Although the SVM was trained to only distinguish real and false processing sites in human microRNAs and has not seen any random hairpins during training, it also performs well for microRNA gene prediction. Further, the results show that an SVM trained explicitly to distinguish microRNAs and random hairpins can give 87% higher sensitivity at the same specificity. An SVM that additionally includes other features known to be characteristic for the microRNAs is expected to give an even better performance.

To get an indication of the most important features in the processing site SVM and, thus, the Microroprocessing process, each of the eleven features were removed separately from the SVM. This revealed that the most important features are the precursor length and loop size, and tre nucleotide occurrences on each position in the precursor and in the flanking region. Similarly, a representative test set of random hairpins were then classified with each of the different SVMs with one feature removed. This was done to get an idea of the different features' importance at separating random hairpins from true microRNAs. The results indicate that the loop size, base-pair information in the flanking region and total number of base-pairs the first 15nts of the flank are the most important features distinguishing microRNAs from random hairpins.

MicroRNAs from 15 other species were also classified by the SVM. The 15 species were selected from all levels of evolutionary diversity from human, and included 1,976 microRNAs in total. The results showed that the SVM performed very well on all species but the non-vertebrates and that the prediction rate decreases with evolutionary distance from humans. Although these microRNAs were not used to train the SVM, many of the microRNAs are homologs to human microRNAs and are therefore not representable for the unknown microRNAs in their respective species. For a more in-depht study, the microRNAs' conservation with microRNAs

used for training should be considered.

In general, the processing site SVM is useful for microRNA discovery in two ways. One, it predicts the processing site for any given hairpin structure. This predicted processing site can be used to calculate more distinct features, which in turn will increase the accuracy of the microRNA predictions. Two, it gives information that can be used to distinguish microRNAs genes and genomic hairpins directly. This is not only information about how likely the predicted processing site is, but also information about the difference between the predicted and the other potential processing sites. This information can be used to extract novel microRNA candidates directly, or it can be used as new features in microRNA gene prediction. Preliminary results show that an SVM that uses the predictions from the processing site SVM and trains explicitly to separate microRNAs and random hairpins performs better than all the current prediction-based approaches.

A natural next step will therefore be to train a new machine learning algorithm based on the features used in this work, the information generated by the processing site SVM, and other general features that have shown to distinguish microRNAs and random hairpins. The algorithm should be trained explicitly to distinguish microRNAs and random hairpin structures, and aim to extract a small set of novel microRNA candidates for experimental verification.

This work is the first of its kind and will hopefully increase the accuracy of computationally microRNA gene discovery, as well as contribute to a better understanding of the microRNA processing.

5 Materials and Methods

5.1 Materials

The microRNA registry [13] is a publicly available database of all annotated microRNA sequences. The set of known human microRNAs used to train and estimate the performance of the SVM was downloaded from version 8.0 of February 2006 of the microRNA registry (`ftp://ftp.sanger.ac.uk/pub/mirbase/sequences/8.0/genomes/hsa.gff`). The unseen independent test set with 130 new human microRNAs was downloaded from version 8.1 of May 2006. Version 8.1 was also used for downloading microRNAs sequences for all other species than human.

The partitioning of known microRNAs into families was downloaded from version 8.1 of the microRNA registry, and is consistent with version 7.0 of the RNA family database Rfam [35] (`ftp://ftp.sanger.ac.uk/pub/mirbase/sequences/8.1/miFam.dat.gz`).

The random hairpins used for microRNA gene prediction were extracted from the

```
1  def FindProcessingSites():
2    loopstart = FindLoopPosition5prime()
3    foreach position p from loopstart to 0:
4      if p base-pairs with a 3prime nucleotide:
5        lastBP5 = p
6        lastBP3 = the position base-paring with p
7      pos3End = (the distance from p to lastBP5) + lastBP3 + 2
8      if the distance from p to pos3End < 50:
9        continue
10     elif the distance from p to pos3End > 80:
11       break
12     processing_sites.add([p, pos3End])
13   return processing_sites
```

Listing 1: Pseudo code for finding all potential processing sites in a sequence.

entire human genome. The DNA sequences are based on NCBI build 35 of November 2005, and were downloaded from the Ensembl FTP site (`ftp://ftp.ensembl.org/pub/release-37/homo_sapiens_37_35j/data/fasta/dna/`). The DNA sequences consist of the 24 chromosomes, the mitochondrial chromosome, and the nonchromosomal sequences, i.e. DNA that has not yet been assembled to a chromosome.

5.2 Methods

5.2.1 Extracting and processing the processing sites

Four principles to define the concept of a processing site were used in all analyses and experiments: 1) the microRNA's processing site is strictly defined by the 5' end of the processing site; 2) all processing sites give a two nucleotides overhang on the 3' end, that is, the 3' end has two additional bases relative to the 5' end; 3) the distance between the 5' end and the 3' end must be between 50 and 80 nucleotides; that is, the allowed precursor length is between 50 and 80 nucleotides; and 4) the positive processing site is the processing site where the 5' end is the same as the real 5' end as found in the microRNA registry [13].

The algorithm that extracts all possible processing sites from a microRNA candidate is based on the three first principles. Listing 1 shows the pseudo code for this algorithm. The fourth principle is used in the latter step where the processing sites are divided into the positive and negative examples.

5.2.2 Estimating the classifier's performance

All performance estimates were based on 10-fold cross-validation. A standard fold distribution, which was used in the family SVM and in one version of the normal

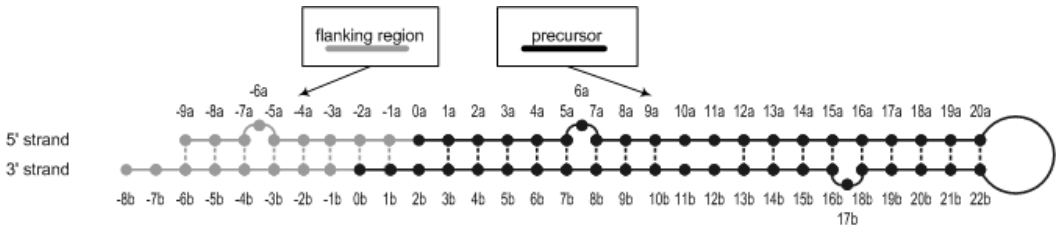


Figure 14: A position refers to a unique nucleotide. The index number refers to the distance from the nucleotide to the processing site. The postfix 'a' and 'b' refers to the 5' and 3' stem, respectively. The position reference system is therefore applicable to any given primary microRNA.

SVM, assigns each microRNA to a random fold that is not already full. This means that all processing sites from the same microRNA were located in the same fold. As there are five microRNAs missing the real processing site and the different microRNAs have different number of false processing sites, this could lead to slightly uneven folds. This was, however, not considered to give a significant bias, and was necessary in order to measure the performance of each microRNA candidate and not only the overall performance of the set of real processing sites compared to the false processing sites. The final performance estimate for the SVM was, however, based on a family distribution of the dataset. The family distribution assigns all microRNAs from the same family into the same fold. To ensure as even folds as possible, the families were first sorted by their size. The largest family not yet assigned to a fold was put into the currently smallest fold. This process was repeated until all families had been assigned to a fold. As the largest family consists of 40 microRNAs this gave one fold that was larger than the others. However, results showed that this did not give a noticeable bias.

5.2.3 Calculating the features

All features used in the analysis, except the loop size, are specific to the particular processing site candidate they are describing. Figure 14 shows how each nucleotide in any given primary microRNA with a defined processing site is referred to with a unique position index, consisting of a number and a postfix letter. The number refer to the distance from the processing site, where the nucleotides in the microRNA precursor is referred to with a positive distance, while the nucleotides in the flanking region is referred to with a negative distance. The distance for the first nucleotide of the stem is defined as 0 and the first nucleotide of the flank is defined as 1. Due to the two nucleotides overhang at the processing site, position 1 in the 5' stem will not necessary match position 1 in the 3' stem. In other words, the nucleotides in the 5' stem and 3' stem are indexed separately. To avoid confusion, the positions in the 5' stem is given the postfix 'a' while the positions in the 3' stem the postfix 'b'; for example, position '4a' refers to the 4th nucleotide in the 5' precursor stem.

Table 7: The table shows a detailed description of the features used in the analysis, where each feature is relative to the specific processing site they describe.

Id	Index	Feature description
1	1	Precursor length; distance from pos 0a inclusive to 0b inclusive
	2	Loop size
2	3	Distance from position 0a inclusive to the loop exclusive
3	4 - 99	Occurrence of each base separately at pos 0a to 23a, both incl
	100 - 195	Occurrence of each base separately at pos 23b to 0b, both incl
4	196 - 219	Base-pair information for and inclusive pos -2 to 21, relative to the 5' stem
5	220 - 223	Total number of each base separately from pos 0a to 23a, both incl
	224 - 227	Total number of each base separately from pos 23b to 0b, both incl
6	228	Total number of base-pairs from pos -2 to 21 relative to 5' side, both inclusive
7	229 - 428	Occurrence each base separately at pos -50a to -1a, both incl
	429 - 628	Occurrence each base separately at pos -1b to -50b, both incl
8	629 - 676	Base-pair information for and inclusive pos -50 to -3 relative to 5' side
9	677 - 680	Total number of each base separately from pos -50a to -1a, both inclusive
	681 - 684	Total number of each base separately from pos -1b to -50b, both inclusive
10	685	Total number of base-pairs from pos -17 to -3 relative to the 5' stem, both inclusive
11	686	Total number of base-pairs from pos -50 to -3 relative to the 5' stem, both inclusive

```
1  def FindLoop():
2    loop_start, loop_end = −1
3    flag_loop = 0
4    foreach position p in the minimum precursor:
5      bp_p = the position base−pairing with p
6      if flag_loop != 1:
7        if bp_p and bp_p > p and bp_p is inside the minimum precursor:
8          loop_start = p + 1
9        elif bp_p and bp_p < p and loop_start > −1:
10         loop_end = p − 1
11         flag_loop = 1
12     else:
13       if bp_p and bp_p > p and bp_p is inside the minimum precursor:
14         endloop1 = the closest position < p that is base−pairing
15         startloop1 = the position base−pairing with endloop1
16         loop_start = (the closest position < startloop1 that base−
               pairs) + 1
17         startloop2 = p
18         endloop2 = the position base−pairing with startloop2
19         loop_end = (the first position > endloop2 that is base−
               pairing) − 1
20         continue at position p = loop_end
21   return [loop_start, loop_end]
```

Listing 2: Pseudo code for finding the loop in a hairpin structure.

Table 7 shows a detailed description of the features used in the analysis. All features were designed for precursors with flanking regions of 50nts. As the maximum precursor length is 80nts, the maximum sequence length we consider is 180nts.

The size of the loop is calculated by an algorithm that detects the beginning and end of the loop. The algorithm has one premiss; that there must be a closed loop within the 50nts of the minimum precursor. Listing 2 shows the pseudo code for the algorithm that detects the loop. The algorithm works in two steps. The algorithm first finds the first closed loop inside the minimum precursor and then searches for base-pairs that create new closed loops inside the minimum precursor. If any additional loops are found, these loops are considered as the same loop and the start and end indexes are extended accordingly.

The occurrence of bases at one specific position results in four columns with a binary value for each of the four nucleotide types, according to whether that nucleotide type occurs at that position or not. For example, occurrences of bases for the sequence ACGG will be [1,0,0,0, 0,1,0,0, 0,0,1,0, 0,0,1,0]. The total number of bases in a subsequence is the number of occurrences of each of the four nucleotide types in that subsequence. The example sequence above yields a total number of bases equal to [1,1,2,0].

The base-pair information for each position is calculated strictly from the position

indexes, independent of the degree of asymmetri in the stem. That is, the base-pair value for a position relative to the 5' side is the average number of nucleotides base-pairing in that position on the 5' stem and in that position plus two on the 3' stem. The base-pair value BP_x at a position x is:

$$BP_x = \begin{bmatrix} 0 & \text{if neither pos } xa \text{ nor } (x+2)b \text{ is base-paring with the other stem} \\ 0.5 & \text{if either pos } xa \text{ or } (x+2)b \text{ is base-paring with the other stem} \\ 1 & \text{if both pos } xa \text{ and } (x+2)b \text{ is base-paring with the other stem} \end{bmatrix}$$

Note that $BP_x = 1$ does not necessarily mean that the two nucleotides are base-pairing with each other, but only that both are base-paring with a nucleotide in the opposite stem. The total number of base-pairs for an area is the sum of all the BP_x for each position x in that area.

5.2.4 The classifier uses two different secondary structure predictions

As mentioned, to be able to include features from the 50nts flank, a sequence length of 180nts is necessary. As the features include base-pair information in the flank, a secondary structure prediction of the entire 180nts sequences is also required.

However, the secondary structure predictions of the 180nts sequences often differ from the secondary structure predictions of the hairpin structures alone, which typically are 90-110nts long. As the algorithm that extracts potential processing sites from a sequence is entirely based on the secondary structure prediction, the difference between the structure of the 180nts sequence and the structure of the 90-110nts annotated sequence could therefore impair the precision of the algorithm. Also the features involving base-pair information will be impacted by such a difference in the secondary structure. A subsequence of the 110 nucleotides in the center of the original 180nts sequence was therefore folded into a secondary structure, and is referred to as the 110nts folding.

20.5% of the random hairpins failed the verification step described in section 2.1.1 using the 110nts foldings, against 35.3% when the 180nts foldings were used. All the known microRNAs passed the verification step when both the 110nts and the 180nts foldings were used. However, this verification step should not be considered as a step to eliminate bad hairpins and distinguish random hairpins from microRNAs, but to insure that there are at least one potential processing site present and that all features can be calculated. Further, the SVM showed a better performance when the features involving the precursor's secondary structure, features 1, 2, 4 and 6, were calculated based on the 110nts foldings rather than the 180nts foldings (results not shown).

Based on these results, the processing site detection and the features involving the precursor's secondary structure were based on the 110nts foldings, despite that this could give an inconsistent set of features for some of the candidates. All

secondary structures were predicted using version 1.4 of RNAfold [37, 38]. The default parameter setting was used, and only the top predicted secondary structure was considered.

5.2.5 SVMs combined with kernels give non-linear classification

The machine learning method used in this work is the gist implementation of the support vector machine (SVM) (`http://benzer.ubic.ca/gist/`). SVM is a linear machine learning method that can be used for classification and regression [39]. The SVM uses dot product to find the similarity between two instances. Combined with a separate weight for each instance, the SVM classifies an unknown instance according to the similarity between the unknown instance and the entire set of known instances. For a training set X with N instances, let instance x_i in X have the label l_i and weight α_i. The score $F(y)$ for an unknown instance y is given by

$$F(y) = \text{sgn}(\sum_{i=1}^{N} l_i \alpha_i \|x_i, y\|)$$

The training of the SVM consists of optimizing the weights α_i for $0 < i \leq N$. This is done by maximizing

$$\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} l_i l_j \alpha_i \alpha_j \|x_i, x_j\|$$

such that

$$0 \leq \alpha_i \leq C_i \text{ and } \sum_{i=1}^{N} l_i \alpha_i = 0$$

where C_i is the penalty parameter for misclassifying instance x_i.

For data that is not possible to separate linearly, the SVM can be combined with a kernel to enable non-linearity [40, 39]. That is, the SVM now uses a kernel function instead of the dot product to find the similarity between two instances. This kernel function maps the non-linear data into a higher dimensional space where the data is more likely to be linearly separable, and returns a score of the similarity between the two instances in this higher space. An example of such a kernel function is the gaussian radial basis function which maps the data to the Hilbert space of infinite dimensions [40]. For two instances x and y, the similarity score $K(x, y)$ between them is given by

$$K(x, y) = \exp(-\frac{\|x, y\|^2}{2w^2}))$$

where w is the width of the kernel. The width can be optimised or chosen in some heuristic manner; for example as the median of the distance from each positive training point to the nearest negative training point.

5.2.6 Extracting genomic hairpin structures using ScorePin

The set of random hairpins used for the microRNA gene prediction was extracted from the entire human genome, using a program called ScorePin [23]. In general, ScorePin is based on the edit distance algorithm and uses dynamic programming and several states to calculate the structural similarity between the given sequence and a hairpin structure with a perfect base-paired stem. ScorePin uses in other words the predicted base-pairing in the stem alone as scoring criteria.

ScorePin is designed to quickly find hairpin structures in large amount of data, typically entire genomes, and differ from the traditional RNA folding tools such as Vienna [38] because of the low runtime. That is, instead of finding the most likely secondary structures for a given sequence, ScorePin predicts how likely the sequence is to be folded like a hairpin structure. By doing so, ScorePin reduces the runtime from $O(n^3)$, which is the runtime for the RNA folding tools, to $O(n^2)$, where n is proportional to the sequence length. Adding additional constraints such as maximun hairpin asymmetry reduces the runtime further to $O(n)$, which in practice means ScorePin uses one day to extract hairpin structures from the entire genome, compared to six months for the traditional RNA folding tools. As this work is strictly time limited, ScorePin was used instead of a traditional RNA folding tool.

ScorePin assigns a score to each position in the input sequence based on the similarity score where that position is in the middle of the hairpin loop. In this way, ScorePin enables the user to extract a set of hairpin structures according to a preferred sensitivity or specificity. For our approach, a high sensitivity was crucial, and a threshold of 110 was chosen, which gives a sensitivity of 98.2%. A window size of 14 was chosen, which means that all high-scoring positions within each window size of 14 position are considered as the same hairpin.

References

[1] Benjamin Lewin. *Genes VII*. Oxford University Press Inc., 2000.

[2] Nelson C. Lau, Lee P. Lim, Earl G. Weinstein, and David P. Bartel. An abundant class of tiny rnas with probable regulatory roles in caenorhabditis elegans. *Science*, 294(5543):858–862, October 2001.

[3] Mariana Lagos-Quintana, Reinhard Rauhut, Winfried Lendeckel, and Thomas Tuschl. Identification of novel genes coding for small expressed rnas. *Science*, 294(5543):853–858, October 2001.

[4] David P. Bartel. Microrнas: Genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–297, January 2004.

[5] Victor Ambros. The functions of animal microrнas. *Nature*, 431(7006):350–355, September 2004.

[6] Yoontae Lee, Chiyoung Ahn, Jinju Han, Hyounjeong Choi, Jaekwang Kim, Jeongbin Yim, Junho Lee, Patrick Provost, Olof Rådmark, Sunyoung Kim, and V. Narry Kim. The nuclear rnase iii drosha initiates microrna processing. *Nature*, 425(6956):415–419, September 2003.

[7] Chang-Zheng Chen, Ling Li, Harvey F. Lodish, and David P. Bartel. Microrнas modulate hematopoietic lineage differentiation. *Science*, 303(5654):83–86, January 2004.

[8] Yan Zeng and Bryan R. Cullen. Sequence requirements for micro rna processing and function in human cells. *RNA*, 9(1):112–123, January 2003.

[9] V. Narry Kim. Microrna biogenesis: coordinated cropping and dicing. *Nat Rev Mol Cell Biol*, 6(5):376–385, May 2005.

[10] Dianne S. Schwarz, György Hutvágner, Tingting Du, Zuoshang Xu, Neil Aronin, and Phillip D. Zamore. Asymmetry in the assembly of the rnai enzyme complex. *Cell*, 115(2):199–208, October 2003.

[11] Nikolaus Rajewsky. Microrna target predictions in animals. *Nature*, 38:S8–S13, June 2006.

[12] Victor Ambros, Bonnie Bartel, David P. Bartel, Christopher B. Burge, James C. Carrington, Xuemei Chen, Gideon Dreyfuss, Sean R. Eddy, Sam Griffiths-Jones, Mhairi Marshall, Marjori Matzke, Gary Ruvkun, and Thomas Tuschl. A uniform system for microrna annotation. *RNA*, 9(3):277–279, March 2003.

[13] Sam Griffiths-Jones. The microrna registry. *Nucleic Acids Res.*, 1(32):D109–D111, January 2004.

[14] Lee P. Lim, Nelson C. Lau, Earl G. Weinstein, Aliaa Abdelhakim, Soraya Yekta, Matthew W. Rhoades, Christopher B. Burge, and David P. Bartel. The microrna of caenorhabditis elegans. *Genes Dev.*, 17(8):991–1008, April 2003.

[15] Benjamin P. Lewis, I hung Shih, Matthew W. Jones-Rhoades, David P. Bartel, and Christopher B. Burge. Prediction of mammalian microrna targets. *Cell*, 115(7):787–798, December 2003.

[16] Isaac Bentwich, Amir Avniel, Yael Karov, Ranit Aharonov, Shlomit Gilad, Omer Barad, Adi Barzilai, Paz Einat, Uri Einav, Eti Meiri, Eilon Sharon, Yael Spector, and Zvi Bentwich. Identification of hundreds of conserved and nonconserved human microrna. *Nature*, 37(7):766–770, July 2005.

[17] Eugene Berezikov, Victor Guryev, José van de Belt, Erno Wienholds, Ronald H. A. Plasterk, and Edwin Cuppen. Phylogenetic shadowing and computational identification of human microrna genes. *Cell*, 120(1):21–24, January 2005.

[18] Alain Sewer, Nicodème Paul, Pablo Landgraf, Alexei Aravin, Sébastien Pfeffer, Michael J. Brownstein, Thomas Tuschl, Erik van Nimwegen, and Mihaela Zavolan. Identification of clustered microrna using an ab initio prediction method. *BMC Bioinformatics*, 6(267), November 2005.

[19] Xiaohui Xie, Jun Lu, E. J. Kulbokas, Todd R. Golub, Vamsi Mootha, Kerstin Lindblad-Toh, Eric S. Lander, and Manolis Kellis. Systematic discovery of regulatory motifs in human promoters and 3' utrs by comparison of several mammals. *Nature*, 434(7031):338–345, March 2005.

[20] Eugene Berezikov, Edwin Cuppen, and Ronald H. A. Plasterk. Approaches to microrna discovery. *Nature*, 38(1):2–7, June 2006.

[21] Eric C. Lai, Pavel Tomancak, Robert W. Williams, and Gerald M. Rubin. Computational identification of drosophila microrna genes. *Genome Biol.*, 4(7):R42, June 2003.

[22] Mariana Lagos-Quintana, Reinhard Rauhut, Abdullah Yalcin, Jutta Meyer, Winfried Lendeckel, and Thomas Tuschl. Identification of tissue-specific microrna from mouse. *Current Biology*, 12(9):735–739, April 2002.

[23] Snorre Helvik. Computational identification of hairpin structures using edit distance-like approach. Master's thesis, NTNU, December 2005.

[24] Isaac Bentwich. Prediction and validation of microrna and their targets. *FEBS*, 579(26):5904–5910, October 2005.

[25] Jordan M. Cummins, Yiping He, Rebecca J. Leary, Ray Pagliarini, Luis A. Diaz Jr., Tobias Sjoblom, Omer Barad, Zvi Bentwich, Anna E. Szafranska, Emmanuel Labourier, Christopher K. Raymond, Brian S. Roberts, Hartmut

Juhl, Kenneth W. Kinzler, Bert Vogelstein, and Victor E. Velculescu. The colorectal micrornaome. *PNAS*, 103(10):3687–3692, March 2006.

[26] Jin-Wu Nam, Ki-Roo Shin, Jinju Han, Yoontae Lee, V. Narry Kim, and Byoung-Tak Zhang. Human microrna prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Res.*, 33(11):3570–3581, June 2005.

[27] Chenghai Xue, Fei Li, Tao He, Guo-Ping Liu, Yanda Li, and Xuegong Zhang. Classification of real and pseudo microrna precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, 6(310), December 2005.

[28] Jinju Han, Yoontae Lee, Kyu-Hyeon Yeom, Jin-Wu Nam, Inha Heo, Je-Keun Rhee, Sun Young Sohn, Yunje Cho, Byoung-Tak Zhang, and V. Narry Kim. Molecular basis for the recognition of primary microrna by the drosha-dgcr8 complex. *Cell*, 125(5):887–901, June 2006.

[29] Pål Sætrom, Ola Snøve Jr., Magnar Nedland, Thomas B. Grünfeld, Yun Lin, Michael Bass, and Jude R. Canon. Conserved microrna characteristics in mammals. *Oligonucleotides*, 16(2):115–144, June 2006.

[30] Anastasia Khvorova, Angela Reynolds, and Sumedha D. Jayasena. Functional sirna and mirna exhibit strand bias. *Cell*, 115(2):209–216, October 2003.

[31] Jacek Krol, Krzysztof Sobczak, Urszula Wilczynska, Maria Drath, Anna Jasinska, Danuta Kaczynska, and Wlodzimierz J. Krzyzosiak. Structural features of microrna (mirna) precursors and their relevance to mirna biogenesis and small interfering rna/short hairpin rna design. *J Biol Chem*, 279(40):42230–42239, October 2004.

[32] Hongxia Zhou, Xu Gang Xia, and Zuoshang Xu. An rna polymerase ii construct synthesizes short-hairpin rna with a quantitative indicator and mediates highly efficient rnai. *Nucleic Acids Res.*, 33(6):e62, April 2005.

[33] Yan Zeng, Rui Yi, and Bryan R. Cullen. Recognition and cleavage of primary microrna precursors by the nuclear processing enzyme drosha. *EMBO J.*, 24(1):138–148, January 2005.

[34] Reiji Teramoto, Mikio Aoki, Toru Kimura, and Masaharu Kanaoka. Prediction of sirna functionality using generalized string kernel and support vector machine. *FEBS*, 579(13):2878–2882, May 2005.

[35] Sam Griffiths-Jones, Alex Bateman, Mhairi Marshall, Ajay Khanna, and Sean R. Eddy. Rfam: an rna family database. *Nucleic Acids Res.*, 1(31):439–441, January 2003.

[36] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International*

Joint Conference on Artificial Intelligence, pages 1137–1143. Morgan Kaufmann Publishers, Montreal, Canada, 1995.

[37] Ivo L. Hofacker, Walter Fontana, Peter F. Stadler, L. Sebastian Bonhoeffer, Manfred Tacker, , and Peter Schuster. Fast folding and comparison of rna secondary structures. *Momathefte für Chemie*, 125(1):167–188, 1994.

[38] Ivo L. Hofacker. Vienna rna secondary structure server. *Nucleic Acids Research*, 31(13):3429–3431, July 2003.

[39] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[40] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

A Prediction results for the newly discovered microRNAs

Table A-1: The table shows the prediction results for the 130 newly discovered microRNAs [25]. Two of these (hsa-mir-614 and hsa-mir-639) failed the verification step and could therefore not be predicted at all. Six of the remaining 128 (hsa-mir-553, hsa-mir-563, hsa-mir-591, hsa-mir-611, hsa-mir-626, and hsa-mir-650) did not give the real processing site. The table shows the score assigned to the predicted processing site and the real processing site. Further, the table shows the distance from the predicted processing site to the real site (Dist) and the number of false processing sites that got a higher score than the real (# false). Finally, the table shows the average and standard deviation of the processing sites' scores.

	Predicted		**Real**			**#**	**Avg**	
MicroRNA	**Site**	**Score**	**Site**	**Score**	**Dist**	**false**	**score**	**Stdev**
hsa-mir-33b	60,120	-0.04	60,120	-0.04	0	0	-0.75	0.50
hsa-mir-411	62,118	0.39	62,118	0.39	0	0	-0.93	0.65
hsa-mir-421	59,122	0.05	61,120	-0.42	2	2	-0.84	0.57
hsa-mir-449b	61,120	0.34	60,121	-0.22	-1	1	-0.81	0.57
hsa-mir-532	60,121	-0.32	61,120	-0.59	1	1	-1.01	0.37
hsa-mir-548a-1	61,119	-0.01	61,119	-0.01	0	0	-0.87	0.41
hsa-mir-548a-2	61,120	0.19	61,120	0.19	0	0	-0.87	0.53
hsa-mir-548a-3	61,119	1.21	61,119	1.21	0	0	-0.80	0.69
hsa-mir-548b	61,119	0.44	61,119	0.44	0	0	-0.85	0.59
hsa-mir-548c	61,119	0.72	61,119	0.72	0	0	-0.74	0.62
hsa-mir-548d-1	61,119	0.45	61,119	0.45	0	0	-0.90	0.71
hsa-mir-548d-2	61,119	0.60	61,119	0.60	0	0	-0.81	0.55
hsa-mir-549	59,121	0.21	63,117	-0.56	4	4	-0.87	0.64
hsa-mir-550-1	58,122	-0.02	60,120	-0.15	2	2	-0.81	0.48
hsa-mir-550-2	59,121	0.16	60,120	-0.03	1	1	-0.77	0.48
hsa-mir-551a	60,120	-0.10	60,120	-0.10	0	0	-0.89	0.44
hsa-mir-551b	59,121	-0.16	60,120	-0.20	1	1	-1.14	0.57
hsa-mir-552	60,121	0.41	62,119	-1.21	2	2	-0.94	0.56
hsa-mir-553	59,121	0.24	-	-	-	-	-0.81	0.57
hsa-mir-554	60,121	0.09	58,123	-0.10	-2	2	-0.85	0.64
hsa-mir-555	61,120	0.23	55,126	-1.46	-6	6	-1.01	0.61
hsa-mir-556	61,119	0.06	60,120	-0.22	-1	1	-0.98	0.58
hsa-mir-557	59,122	0.35	64,117	-0.76	5	5	-0.87	0.53
hsa-mir-558	56,124	0.44	63,118	-0.76	7	7	-0.71	0.54
hsa-mir-559	59,122	-0.05	59,122	-0.05	0	0	-0.78	0.49
hsa-mir-560	59,121	0.60	55,125	-0.79	-4	4	-0.67	0.49
hsa-mir-561	62,119	0.29	62,119	0.29	0	0	-0.74	0.62
hsa-mir-562	60,121	0.06	59,122	-0.61	-1	1	-0.94	0.59
hsa-mir-563	58,124	-0.19	-	-	-	-	-0.78	0.45

Continues next page

Table A-1 - continued from previous page

	Predicted		Real			#	Avg	
MicroRNA	**Site**	**Score**	**Site**	**Score**	**Dist**	**false**	**score**	**Stdev**
hsa-mir-564	59,122	-0.05	59,122	-0.05	0	0	-0.46	0.26
hsa-mir-565	53,128	-0.55	55,126	-0.86	2	2	-0.81	0.16
hsa-mir-566	63,117	-0.23	64,116	-0.68	1	1	-0.79	0.38
hsa-mir-567	61,119	-0.27	61,119	-0.27	0	0	-0.94	0.39
hsa-mir-568	64,117	0.04	58,123	-0.94	-6	6	-0.96	0.63
hsa-mir-569	59,121	0.39	58,122	-0.90	-1	1	-1.07	0.55
hsa-mir-570	61,120	0.20	61,120	0.20	0	0	-0.62	0.46
hsa-mir-571	61,120	-0.07	57,124	-1.05	-4	4	-0.81	0.53
hsa-mir-572	60,120	0.00	64,116	-1.04	4	4	-0.72	0.38
hsa-mir-573	60,119	0.04	56,125	-0.52	-4	4	-0.99	0.56
hsa-mir-574	57,123	-0.19	62,118	-0.48	5	5	-0.67	0.35
hsa-mir-575	59,120	0.18	54,127	-1.10	-5	5	-0.71	0.56
hsa-mir-576	61,119	0.09	59,121	-0.11	-2	2	-0.96	0.72
hsa-mir-577	60,121	0.41	62,119	-0.52	2	2	-0.87	0.68
hsa-mir-578	63,117	0.05	61,119	-0.12	-2	2	-1.12	0.74
hsa-mir-579	63,117	-0.09	61,119	-0.11	-2	2	-0.90	0.61
hsa-mir-580	60,121	0.25	60,121	0.25	0	0	-0.95	0.64
hsa-mir-581	60,121	-0.18	60,121	-0.18	0	0	-0.92	0.51
hsa-mir-582	58,122	0.13	60,120	0.03	2	2	-0.86	0.67
hsa-mir-583	63,117	-0.21	60,120	-0.52	-3	3	-1.08	0.50
hsa-mir-584	59,121	0.60	59,121	0.60	0	0	-0.89	0.74
hsa-mir-585	60,120	-0.02	64,116	-1.32	4	4	-0.95	0.60
hsa-mir-586	59,122	0.02	62,119	-0.57	3	3	-0.75	0.42
hsa-mir-587	60,120	0.45	55,125	-1.30	-5	5	-0.91	0.63
hsa-mir-588	61,119	0.07	61,119	0.07	0	0	-0.79	0.51
hsa-mir-589	60,121	0.47	60,121	0.47	0	0	-0.91	0.60
hsa-mir-590	59,122	0.18	60,121	-0.19	1	1	-1.00	0.70
hsa-mir-591	53,108	-0.48	-	-	-	-	-0.96	0.39
hsa-mir-592	60,120	0.11	60,120	0.11	0	0	-0.92	0.69
hsa-mir-593	63,117	0.06	54,126	-0.91	-9	9	-1.14	0.66
hsa-mir-594	61,119	0.10	59,123	-1.01	-2	2	-1.06	0.63
hsa-mir-595	62,124	-0.01	57,124	-0.86	-5	5	-0.98	0.48
hsa-mir-596	61,120	0.36	63,118	-0.73	2	2	-0.92	0.56
hsa-mir-597	58,122	0.17	59,121	-0.39	1	1	-0.83	0.58
hsa-mir-598	58,123	0.30	61,120	0.05	3	3	-0.96	0.56
hsa-mir-599	59,123	0.07	62,118	-0.70	3	3	-0.84	0.48
hsa-mir-600	61,119	-0.13	63,117	-1.13	2	2	-0.84	0.44
hsa-mir-601	59,124	-0.18	62,119	-0.20	3	3	-0.85	0.44
hsa-mir-602	60,121	0.27	61,120	0.06	1	1	-0.83	0.56
hsa-mir-603	61,120	0.80	61,120	0.80	0	0	-0.89	0.65

Continues next page

Table A-1 - continued from previous page

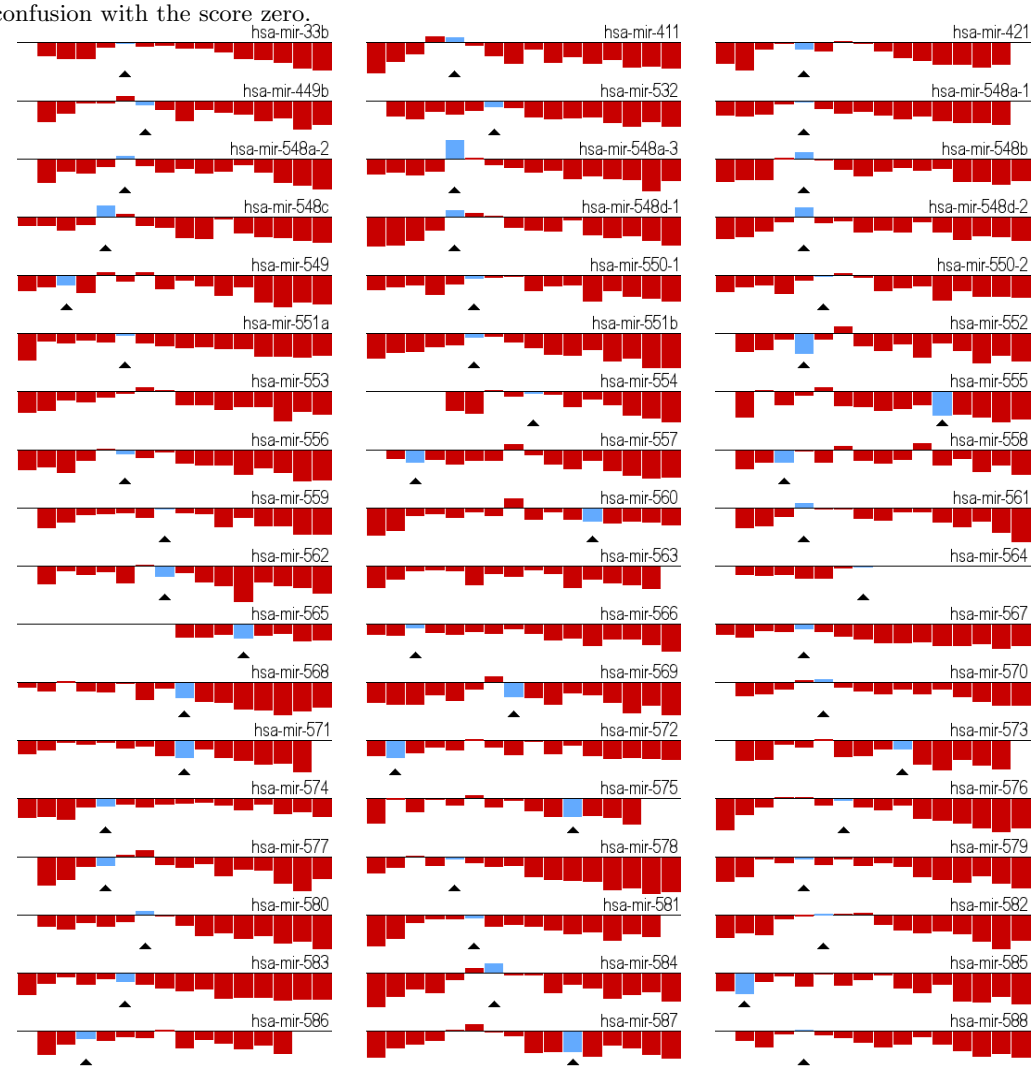
	Predicted		Real			#	Avg	
MicroRNA	**Site**	**Score**	**Site**	**Score**	**Dist**	**false**	**score**	**Stdev**
hsa-mir-604	64,117	0.02	56,124	-1.06	-8	8	-1.00	0.52
hsa-mir-605	61,120	0.29	62,119	0.04	1	1	-0.72	0.56
hsa-mir-606	61,119	0.22	52,128	-1.52	-9	9	-0.86	0.56
hsa-mir-607	60,121	0.21	60,121	0.21	0	0	-0.75	0.41
hsa-mir-608	53,112	-0.81	55,110	-0.95	2	2	-1.01	0.09
hsa-mir-609	57,123	0.40	58,122	-0.39	1	1	-0.85	0.70
hsa-mir-610	60,121	-0.22	60,121	-0.22	0	0	-0.75	0.33
hsa-mir-611	66,129	-0.67	-	-	-	-	-0.92	0.18
hsa-mir-612	61,120	0.29	59,122	0.20	-2	2	-0.65	0.58
hsa-mir-613	63,120	-0.15	57,123	-0.69	-6	6	-1.04	0.52
hsa-mir-614	-	-	-	-	-	-	-	-
hsa-mir-615	63,117	-0.08	59,121	-0.77	-4	4	-0.75	0.40
hsa-mir-616	58,122	-0.04	59,121	-0.16	1	1	-1.01	0.63
hsa-mir-617	60,120	0.36	63,117	-0.71	3	3	-0.86	0.73
hsa-mir-618	62,119	0.45	57,123	-0.69	-5	5	-0.82	0.70
hsa-mir-619	61,120	-0.42	61,120	-0.42	0	0	-1.13	0.55
hsa-mir-620	62,118	0.12	53,127	-1.58	-9	9	-0.97	0.57
hsa-mir-621	62,118	-0.04	61,119	-1.03	-1	1	-0.71	0.46
hsa-mir-622	63,117	0.25	58,122	-0.06	-5	5	-0.89	0.55
hsa-mir-623	61,123	0.19	56,125	-0.93	-5	5	-0.86	0.53
hsa-mir-624	57,123	0.01	61,119	-0.40	4	4	-0.86	0.52
hsa-mir-625	61,120	0.07	61,120	0.07	0	0	-0.69	0.47
hsa-mir-626	51,109	-0.78	-	-	-	-	-1.00	0.16
hsa-mir-627	61,119	0.12	60,120	-0.02	-1	1	-1.02	0.57
hsa-mir-628	60,121	0.43	62,119	-0.54	2	2	-1.03	0.61
hsa-mir-629	61,120	0.15	60,121	-0.02	-1	1	-0.51	0.40
hsa-mir-630	60,119	0.26	58,122	-0.42	-2	2	-0.82	0.50
hsa-mir-631	58,119	-0.24	64,116	-0.68	6	6	-0.80	0.39
hsa-mir-632	60,120	-0.25	59,121	-1.09	-1	1	-0.98	0.35
hsa-mir-633	60,121	0.06	61,120	-0.02	1	1	-0.63	0.46
hsa-mir-634	62,119	-0.32	63,118	-0.71	1	1	-0.82	0.32
hsa-mir-635	63,119	-0.09	53,127	-1.57	-10	10	-1.14	0.61
hsa-mir-636	60,121	-0.03	59,122	-0.91	-1	1	-0.76	0.39
hsa-mir-637	60,119	-0.24	53,127	-1.19	-7	7	-0.72	0.29
hsa-mir-638	56,126	-0.11	51,130	-1.43	-5	5	-0.84	0.39
hsa-mir-639	-	-	-	-	-	-	-	-
hsa-mir-640	62,119	-0.20	54,127	-1.40	-8	8	-0.88	0.40
hsa-mir-641	61,119	0.04	55,125	-0.65	-6	6	-0.89	0.59
hsa-mir-642	62,119	0.42	62,119	0.42	0	0	-0.66	0.54
hsa-mir-643	60,121	0.28	60,121	0.28	0	0	-0.83	0.75

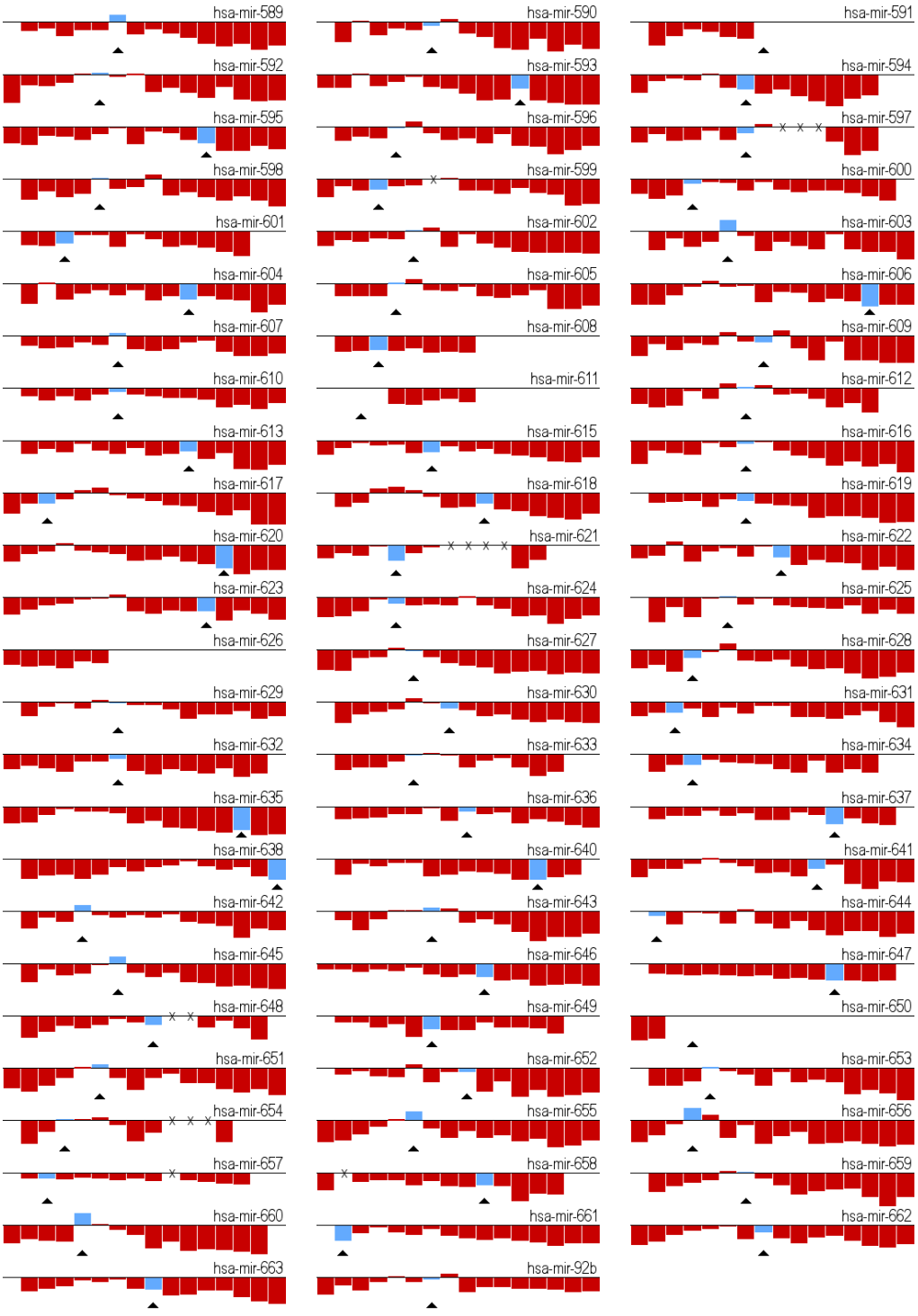
Continues next page

Table A-1 - continued from previous page

	Predicted		Real			#	Avg	
MicroRNA	**Site**	**Score**	**Site**	**Score**	**Dist**	**false**	**score**	**Stdev**
hsa-mir-644	60,121	0.14	65,116	-0.27	5	5	-0.90	0.59
hsa-mir-645	60,121	0.50	60,121	0.50	0	0	-0.91	0.59
hsa-mir-646	61,121	-0.19	58,123	-0.69	-3	3	-0.86	0.44
hsa-mir-647	66,117	-0.65	56,127	-1.14	-10	10	-0.90	0.18
hsa-mir-648	60,120	-0.16	58,122	-0.55	-2	2	-0.76	0.42
hsa-mir-649	57,126	-0.37	59,122	-0.94	2	2	-0.77	0.29
hsa-mir-650	50,101	-1.56	-	-	-	-	-1.62	0.07
hsa-mir-651	60,120	0.23	60,120	0.23	0	0	-0.99	0.59
hsa-mir-652	61,120	0.27	58,123	-0.24	-3	3	-0.94	0.72
hsa-mir-653	62,119	0.04	62,119	0.04	0	0	-0.99	0.61
hsa-mir-654	61,120	0.19	63,118	0.07	2	2	-0.67	0.68
hsa-mir-655	60,120	0.62	60,120	0.62	0	0	-0.97	0.63
hsa-mir-656	62,118	0.85	62,118	0.85	0	0	-0.99	0.76
hsa-mir-657	61,120	-0.29	63,118	-0.35	2	2	-0.46	0.14
hsa-mir-658	62,118	-0.46	55,125	-0.88	-7	7	-0.95	0.42
hsa-mir-659	61,120	0.14	60,121	0.06	-1	1	-0.98	0.64
hsa-mir-660	61,119	0.84	61,119	0.84	0	0	-1.07	0.77
hsa-mir-661	63,118	-0.10	65,116	-1.07	2	2	-0.83	0.37
hsa-mir-662	60,120	-0.03	58,122	-0.43	-2	2	-0.87	0.43
hsa-mir-663	59,121	-0.10	57,123	-0.80	-2	2	-0.89	0.49
hsa-mir-92b	59,122	0.24	60,121	-0.10	1	1	-0.59	0.40

Figure A-1: The figures show the score for each processing site in each of the 130 newly discovered microRNAs. Two of these (hsa-mir-614 and hsa-mir-639) failed the verification step and could therefore not be processed. The triangle at the bottom of each figure marks the real processing site in the microRNA. The columns give the scores for the different processing sites, sorted ascending according to the precursor length. That is, the first column represents the processing site that gives a precursor length of 50-51nts, and the last column a precursor length of 79-80nts. Gaps in precursor lengths due to bulges and internal loops are filled with a 'x' on the x-axis for avoid confusion with the score zero.




hsa-mir-589
hsa-mir-590
hsa-mir-591
hsa-mir-592
hsa-mir-593
hsa-mir-594
hsa-mir-595
hsa-mir-596
hsa-mir-597
hsa-mir-598
hsa-mir-599
hsa-mir-600
hsa-mir-601
hsa-mir-602
hsa-mir-603
hsa-mir-604
hsa-mir-605
hsa-mir-606
hsa-mir-607
hsa-mir-608
hsa-mir-609
hsa-mir-610
hsa-mir-611
hsa-mir-612
hsa-mir-613
hsa-mir-615
hsa-mir-616
hsa-mir-617
hsa-mir-618
hsa-mir-619
hsa-mir-620
hsa-mir-621
hsa-mir-622
hsa-mir-623
hsa-mir-624
hsa-mir-625
hsa-mir-626
hsa-mir-627
hsa-mir-628
hsa-mir-629
hsa-mir-630
hsa-mir-631
hsa-mir-632
hsa-mir-633
hsa-mir-634
hsa-mir-635
hsa-mir-636
hsa-mir-637
hsa-mir-638
hsa-mir-640
hsa-mir-641
hsa-mir-642
hsa-mir-643
hsa-mir-644
hsa-mir-645
hsa-mir-646
hsa-mir-647
hsa-mir-648
hsa-mir-649
hsa-mir-650
hsa-mir-651
hsa-mir-652
hsa-mir-653
hsa-mir-654
hsa-mir-655
hsa-mir-656
hsa-mir-657
hsa-mir-658
hsa-mir-659
hsa-mir-660
hsa-mir-661
hsa-mir-662
hsa-mir-663
hsa-mir-92b

B The microRNAs used in the family SVM

Table B-1: The family SVM was trained on only one microRNA from each family. The table shows which microRNA was used from each of the families and the size of the family, that is, the total number of human microRNAs in the family. Nine microRNAs are not assigned to a family yet (version 8.1 of the microRNA registry) and were therefore considered as separate families with family size 1. Three of the microRNAs with family size 1 (hsa-mir-198, hsa-mir-451, and hsa-mir-448) were excluded because the real processing sites were not found. This gives a total of 163 families used in the family SVM.

Family ID	**Used microRNA**	**Family size**
MIPF0000001	hsa-mir-17	8
MIPF0000002	hsa-let-7a-1	12
MIPF0000005	hsa-mir-30a	6
MIPF0000006	hsa-mir-15a	5
MIPF0000007	hsa-mir-181a-2	6
MIPF0000009	hsa-mir-29a	4
MIPF0000011	hsa-mir-19a	3
MIPF0000013	hsa-mir-92-1	4
MIPF0000014	hsa-mir-9-1	3
MIPF0000017	hsa-mir-125b-1	3
MIPF0000018	hsa-mir-154	16
MIPF0000019	hsa-mir-200b	5
MIPF0000020	hsa-mir-520e	40
MIPF0000021	hsa-mir-124a-1	3
MIPF0000022	hsa-mir-7-1	3
MIPF0000024	hsa-mir-103-2	3
MIPF0000025	hsa-mir-99a	3
MIPF0000026	hsa-mir-218-1	2
MIPF0000027	hsa-mir-23a	2
MIPF0000028	hsa-mir-135a-1	3
MIPF0000029	hsa-mir-133a-1	3
MIPF0000031	hsa-mir-196a-1	3
MIPF0000033	hsa-mir-10a	2
MIPF0000034	hsa-mir-130a	3
MIPF0000036	hsa-mir-27a	2
MIPF0000038	hsa-mir-1-2	3
MIPF0000039	hsa-mir-34a	3
MIPF0000040	hsa-mir-199a-1	3
MIPF0000041	hsa-mir-24-1	2
MIPF0000042	hsa-mir-204	2
MIPF0000043	hsa-mir-26a-1	3
MIPF0000044	hsa-mir-219-1	2

Continues next page

Table B-1 - continued from previous page

Family ID	Used microRNA	Family size
MIPF0000046	hsa-mir-101-1	2
MIPF0000048	hsa-mir-128a	2
MIPF0000050	hsa-mir-153-1	2
MIPF0000051	hsa-mir-221	2
MIPF0000053	hsa-mir-22	1
MIPF0000054	hsa-mir-216	1
MIPF0000055	hsa-mir-194-1	2
MIPF0000056	hsa-mir-148a	3
MIPF0000057	hsa-mir-28	2
MIPF0000058	hsa-mir-205	1
MIPF0000059	hsa-mir-184	1
MIPF0000060	hsa-mir-21	1
MIPF0000061	hsa-mir-365-1	2
MIPF0000062	hsa-mir-214	1
MIPF0000063	hsa-mir-192	2
MIPF0000064	hsa-mir-31	1
MIPF0000065	hsa-mir-212	2
MIPF0000066	hsa-mir-183	1
MIPF0000067	hsa-mir-223	1
MIPF0000068	hsa-mir-371	2
MIPF0000069	hsa-mir-32	1
MIPF0000070	hsa-mir-33	1
MIPF0000071	hsa-mir-302a	4
MIPF0000072	hsa-mir-96	1
MIPF0000073	hsa-mir-129-1	2
MIPF0000074	hsa-mir-105-1	2
MIPF0000075	hsa-mir-138-2	2
MIPF0000076	hsa-mir-190	1
MIPF0000077	hsa-mir-217	1
MIPF0000078	hsa-mir-187	1
MIPF0000079	hsa-mir-145	1
MIPF0000080	hsa-mir-127	1
MIPF0000082	hsa-mir-193a	2
MIPF0000084	hsa-mir-142	1
MIPF0000085	hsa-mir-140	1
MIPF0000086	hsa-mir-210	1
MIPF0000088	hsa-mir-224	1
MIPF0000091	hsa-mir-368	4
MIPF0000093	hsa-mir-144	1
MIPF0000094	hsa-mir-143	1
MIPF0000095	hsa-mir-122a	1

Continues next page

Table B-1 - continued from previous page

Family ID	Used microRNA	Family size
MIPF0000097	hsa-mir-338	1
MIPF0000098	hsa-mir-95	3
MIPF0000099	hsa-mir-136	1
MIPF0000103	hsa-mir-146a	2
MIPF0000105	hsa-mir-147	1
MIPF0000106	hsa-mir-137	1
MIPF0000108	hsa-mir-203	1
MIPF0000109	hsa-mir-186	1
MIPF0000110	hsa-mir-329-1	3
MIPF0000111	hsa-mir-489	1
MIPF0000112	hsa-mir-134	1
MIPF0000113	hsa-mir-188	3
MIPF0000114	hsa-mir-375	1
MIPF0000115	hsa-mir-126	1
MIPF0000116	hsa-mir-182	1
MIPF0000117	hsa-mir-139	1
MIPF0000118	hsa-mir-220	1
MIPF0000121	hsa-mir-202	1
MIPF0000123	hsa-mir-197	1
MIPF0000126	hsa-mir-379	3
MIPF0000128	hsa-mir-450-1	2
MIPF0000129	hsa-mir-455	1
MIPF0000130	hsa-mir-509	2
MIPF0000133	hsa-mir-449	1
MIPF0000137	hsa-mir-383	1
MIPF0000138	hsa-mir-363	1
MIPF0000139	hsa-mir-500	3
MIPF0000142	hsa-mir-431	1
MIPF0000143	hsa-mir-326	1
MIPF0000147	hsa-mir-325	1
MIPF0000157	hsa-mir-155	1
MIPF0000159	hsa-mir-296	1
MIPF0000162	hsa-mir-367	1
MIPF0000163	hsa-mir-320	1
MIPF0000164	hsa-mir-424	1
MIPF0000165	hsa-mir-324	1
MIPF0000167	hsa-mir-370	1
MIPF0000168	hsa-mir-378	1
MIPF0000172	hsa-mir-361	1
MIPF0000173	hsa-mir-499	1
MIPF0000176	hsa-mir-506	3

Continues next page

Table B-1 - continued from previous page

Family ID	Used microRNA	Family size
MIPF0000177	hsa-mir-433	1
MIPF0000178	hsa-mir-208	1
MIPF0000180	hsa-mir-483	1
MIPF0000183	hsa-mir-503	1
MIPF0000185	hsa-mir-542	1
MIPF0000186	hsa-mir-299	1
MIPF0000188	hsa-mir-346	1
MIPF0000189	hsa-mir-345	1
MIPF0000190	hsa-mir-342	1
MIPF0000191	hsa-mir-340	1
MIPF0000192	hsa-mir-412	1
MIPF0000193	hsa-mir-339	1
MIPF0000194	hsa-mir-191	1
MIPF0000195	hsa-mir-337	1
MIPF0000196	hsa-mir-335	1
MIPF0000197	hsa-mir-150	1
MIPF0000199	hsa-mir-331	1
MIPF0000200	hsa-mir-330	1
MIPF0000201	hsa-mir-485	1
MIPF0000202	hsa-mir-185	1
MIPF0000203	hsa-mir-328	1
MIPF0000209	hsa-mir-362	1
MIPF0000211	hsa-mir-432	1
MIPF0000217	hsa-mir-505	1
MIPF0000219	hsa-mir-484	1
MIPF0000220	hsa-mir-486	1
MIPF0000229	hsa-mir-490	1
MIPF0000230	hsa-mir-493	1
MIPF0000231	hsa-mir-497	1
MIPF0000242	hsa-mir-425	1
MIPF0000274	hsa-mir-149	1
MIPF0000287	hsa-mir-452	1
MIPF0000288	hsa-mir-374	1
MIPF0000289	hsa-mir-384	1
MIPF0000291	hsa-mir-512-1	2
MIPF0000301	hsa-mir-511-1	2
MIPF0000314	hsa-mir-513-1	2
MIPF0000318	hsa-mir-488	1
MIPF0000319	hsa-mir-491	1
MIPF0000329	hsa-mir-423	1
-	hsa-mir-373	1

Continues next page

Table B-1 - continued from previous page

Family ID	**Used microRNA**	**Family size**
-	hsa-mir-422a	1
-	hsa-mir-492	1
-	hsa-mir-498	1
-	hsa-mir-504	1
-	hsa-mir-514-1	1
-	hsa-mir-514-2	1
-	hsa-mir-514-3	1
-	hsa-mir-544	1