University of Crete
Computer Science Department

Mature miRNA identification via the use of a Naive Bayes Classifier

Gkirtzou Aikaterini
gkirtzou@csd.uoc.gr

Master of Science Thesis

Heraklion, March 2009

University of Crete
Computer Science Department

The undersigned hereby certify that they have read and recommended to the Faculty of Graduate Studies for acceptance a thesis entitled *"Mature miRNA identification via the use of a Naive Bayes Classifier"* by *Gkirtzou Aikaterini* in partial fulfillment of the requirements for the degree of *Master of Science*.

Author:

Gkirtzou Aikaterini

Committee:

Tsakalides Panagiotis, Supervisor
Associate Professor at CSD, UOC

Poirazi Panayiota, Advisor
Research Associate Professor at IMBB, FORTH

Tsamardinos Ioannis
Assistant Professor at CSD, UOC

Chairman of Graduate Studies Committee:

Trahanias Panos
Professor at CSD, UOC

March 2009

Contents

List of Figures

List of Tables

Abstract

MicroRNAs (miRNAs) are small single stranded RNAs, on average 22nt long, generated from endogenous hairpin–shaped transcripts with post transcriptional activity. Although many computational methods are currently available for identifying miRNA genes in the genomes of various species, very few algorithms can accurately predict the functional part of the miRNA gene, namely the mature miRNA. We introduce a computational method that uses a Naive Bayes classifier to identify mature miRNA candidates based on sequence and secondary structure information of the miRNA precursor. Specifically, for each mature miRNA, we generate a set of negative examples of equal length on the respective precursor(s). The true and negative sets are then used to estimate probability distributions for the sequence and secondary structure composition on each position along the mature or in flanking regions around it, as well as for the distances of the starting and ending position of the mature from the precursor's hairpin and ends. The divergence between these distributions is estimated using the symmetric Kullback-Leibler metric. The features at which the two distributions differ significantly and consistently over a 10-fold cross-validation procedure are used as features for training the Naive Bayes classifier. We used experimentally verified human and mouse miRNA data to train the classifier and a performance of $AUC \approx 0.88$ was achieved using a consensus averaging over a 10–fold cross–validation procedure. Moreover, we examined four strategies in order to provide the most accurate candidate mature, based on the ranking provided by our model. For each strategy, the confidence that the computational truth was $\pm 6nt$ away from the true mature was: *a*) 86.88% for the top scorer, *b*) 88.25% for the middle point of 4 top scorers, *c*) 89.34% for the mean value of 4 top scorers and *d*) 87.83% for the top scorer and its duplex. Our findings suggest that position specific sequence and structure information and the distance features

combined with a simple Bayes classifier achieve a good performance on the challenging task of mature miRNA identification.

Περίληψη

Τα microRNAs είναι μικρά μονόκλωνα μόρια RNAs, με μήκος 22 νουκλεοτιδίων κατά μέσο όρο, τα οποία παράγονται από ενδογενή μετάγραφα με μορφή 'φουρκέτας' και έχουν μέτα–μεταγραφική δραστηριότητα. Παρόλο που υπάρχουν πολλές υπολογιστικές μέθοδοι διαθέσιμες για την αναγνώριση microRNA γονιδίων στο γονιδίωμα πολλών οργανισμών, πολύ λίγοι αλγόριθμοι μπορούν με ακρίβεια να προβλέψουν το λειτουργικό μέρος ενός miRNA γονιδίου, γνωστό ως ώριμο miRNA. Στην εργασία αυτή προτείνουμε μια νέα υπολογιστική μέθοδος, η οποία χρησιμοποιεί έναν Naive Bayes classifier για να αναγνωρίζει υποψήφια ώριμα μόρια miRNA με βάση την ακολουθία και την δευτεροταγή δομή ενός πρώιμου miRNA (precursor miRNA). Συγκεκριμένα, για κάθε ώριμο miRNA, παράγουμε ένα σύνολο αρνητικών παραδειγμάτων ίσου μεγέθους, από τα αντίστοιχα πρώιμα μόρια miRNA. Τα δείγματα από τα πραγματικά και αρνητικά δεδομένα χρησιμοποιούνται κατόπιν για να εκτιμήσουμε τις κατανομές πιθανοτήτων των θέσεων που βρίσκονται είτε κατα μήκος του ώριμου μορίου, είτε σε περιοχες γύρω από αυτό, κρατώντας πληροφορίες για την ακολουθιά και την δευτεροταγή δομή της θέσης, καθώς και για την εκτίμηση των αποστάσεων της αρχικής και τελικής θέσης ενός ώριμου μορίου από τα όρια του κοντινότερου σχηματισμού 'φουρκέτας' του πρώιμου μορίου, καθώς και από τα άκρα του ίδιου του πρώιμου μορίου. Η απόκλιση μεταξύ αυτών των κατανομών υπολογίζεται από την συμμετρική απόκλιση των Kullback–Leibler. Τα χαρακτηριστικά των οποίων οι δύο κατανομές διαφέρουν σημαντικά χρησιμοποιούνται ως χαρακτηριστικά για την εκπαίδευση του Naive Bayes classifier. Χρησιμοποιούμε πειραματικά επιβεβαιωμένα miRNA δεδομένα από άνθρωπο και ποντίκι για την εκπαίδευση του μοντέλου μας και επιτυχγάνουμε μέση απόδοση $AUC \approx 0.88$ χρησιμοποιώντας 10–fold cross validation. Επιπλέον εξετάζουμε τέσσερις στρατηγικές για να παρέχουμε με μεγαλύτερη ακρίβεια ένα υποψήφιο ώριμο μόριο, βασισμένοι στην διάταξη των αποτελεσμάτων που παρέχει το μοντέλο μας. Για

κάθε μια στρατηγική, η βεβαιότητα ότι η υπολογιστική αλήθεια βρίσκεται $\pm 6nt$ μακριά από το πραγματικό ώριμο miRNA είναι: *α*) 86.88% για το υποψήφιο με την υψηλότερη επίδοση (top scorer), *β*) 88.25% για το υποψήφιο που σχηματίζεται από το μεσαίο στοιχείο του διαστήματος που ορίζουν οι τέσσερις υποψήφιοι με την υψηλότερη επίδοση, *ς*) 89.34% για το υποψήφιο που σχηματίζεται από τη μεσή τιμή του διαστήματος που ορίζουν οι τέσσερις υποψήφιοι με την υψηλότερη επίδοση, *δ*) 87.83% για τον υποψήφιο με την υψηλότερη επίδοση (top scorer), και την απένταντι αλληλουχία του όπως ορίζεται από την δευτεροταγή δομή του πρώιμο μορίου (duplex). Τα αποτελέσματά μας προτείνουν ότι η πληφορορίες ακολουθιάς και δευτεροταγής δομής που παρέχονται σε επίπεδο θέσεων, καθώς και οι χαρακτηριστικές αποστάσεις των ορίων του ώριμου μορίου miRNA σε συνδυασμό με έναν Naive Bayes classifier επιτυγχάνουν πολύ καλή απόδοση στο δύσκολο πρόβλημα της αναγνώρισης των ώριμων μορίων miRNA.

Acknowledgments

Αποφάσισα οι ευχαριστίες να γραφούν στα ελληνικά, καθώς μου είναι ευκολότερο να εκφραστώ στην μητρική μου γλώσσα. Κατάρχήν να ευχαριστήσω θερμά την επιβλέπουσα της μεταπτυχιακής μου εργασίας Ποϊράζη Παναγιώτα για την υποστήριξή της και την επιστημονική της καθοδήγηση σε όλη την διάρκεια της συνεργασία μας που ξεκίνησε πρίν από τρια χρόνια σχεδόν, όταν ήμουν ακόμα προπτυχιακή φοιτήτρια. Θα ήθελα ακόμα να ευχαριστήσω τα άλλα δυο μέλη της τριμέλης επιτροπής μου, κ.Τσακαλίδη Παναγιώτη και κ.Τσαμαρδίνο Ιωάννη, και τον ερευνητή Reckzo Martin για τις χρησιμες προτάσεις και σχόλια τους στα πλαίσια της μεταπτυχιακής μου εργασίας. Επιπλέον, θα ήθελα να ευχαριστήσω το τμήμα Επιστήμης Υπολογιστών του Πανεπιστημίου Κρήτης, καθώς και το Ινστιτούτο Πληροφορικής του Ιδρύματος Τεχνολογίας και Έρευνας για την οικονομική υποστήριξη τους κατά την διάρκεια των τελευταίων τεσσεράμιση χρόνων.

Πανώ από όλα , θα ήθελα να ευχαριστήσω από τα βάθη της καρδίας μου τους όλους τους φίλους μου που στάθηκαν δίπλα μου σε στιγμές ευχάριστές και δυσάρεστες, εύκολες και δύσκολες. Πιο συγκεκριμένα την Ελευθερία Τζαμαλή και τον Παναγιωτη Κουτσουράκη για την ηθική, αλλά και επιστημονική τους υποστήριξη. Μέσα από συζητήσεις μαζί τους από κοντά ή μέσω διαδικτύου με βοήθησαν να κατανοήσω αλλά και να ανακαλήψω πολλά πράγματα. Τον συγκάτοικό μου για ένα μήνα, Ηλία Γκρίνια, που μου χάρησε στιγμές χαράς στο δύσκολο κομμάτι την ολοκλήρωση της μεταπτυχιακής μου εργασίας. Σε όλους τους αγαπημένους μου φίλους που ήταν κοντά μου είτε σωματικά είτε ψυχικά για τις στιγμές που μου χάρησαν. Τέλος να ευχαριστήσω την οικογένεια μου για την υποστήριξη και την αγάπη τους όπου χωρίς αυτούς τίποτα δεν θα ήταν δυνατό.

Chapter 1

Introduction

MicroRNAs (miRNAs) are an abundant class of $\sim 22nt$ long endogenous non–protein–coding RNAs that regulate gene expression by binding to target sites on 3'UTRs of messenger RNAs (mRNAs). This binding results primarily in translational repression or mRNA degradation [54], although an enhancement of the target gene's expression has also been observed [65]. Mature miRNAs are derived from longer (70–100nt) precursors, the pre–miRNAs, which form a hairpin–like structure that contains one or two mature miRNAs in either or both of its arms. A large body of experimental findings indicates that the regulatory action of miRNAs is essential for most organisms as these tiny molecules play a central role in multiple processes, including development timing [41, 56], cell proliferation and differentiation [26, 66, 58, 16, 10], apoptosis [69, 7, 22], as well as in numerous diseases [28, 46, 38, 51, 27] and anti–viral defense [49, 24] (for more detailed description see also section 2.3).

An important step towards the understanding of miRNA–mediated regulation would be to assemble a complete catalog of miRNA genes, their products and their targets. Torwards this goal, experimental cloning efforts have successfully identified highly expressed miRNAs from various tissues and various organisms. However, cloning methods have a number of shortcomings, including high costs, while they are are highly biased towards miRNAs that are abundantly and/or ubiquitously expressed. On the other hand, computational prediction of miRNAs could become a powerful tool for finding tissue–specific or lowly expressed miRNAs. Several computational methods have been developed to facilitate the discovery of miRNAs (reviewed in [5]). Most of them focus on the discovery of either novel miRNA genes in the genomes of various species or possible mRNA targets of the known miRNAs. On the contrary, few attempts have been made to computationally predict the functional

part of the miRNA precursor, namely the mature miRNA. A number of studies ([50], [72], [61]) combine miRNA gene prediction with the identification of a possible start position for the mature. To our knowledge, only one study [64] focuses exclusively on mature miRNA prediction, utilizing thermodynamic and structural information.

1.1 Objectives

The purpose of this thesis is to build a Naive Bayes classifier capable of indentifying the mature miRNA(s) within a precursor miRNA with high accuracy. Torwards this goal, we consider biological features of miRNA precursors such as position specific sequence and structure information. We investigate numerous combinations of such features both within the mature as well as in regions around it. Features are selected according to their effect on classification performance in a two-class problem(true vs false mature), whereby all possible mature candidates that can be generated by sliding along the precursor are tested. The model's output is the predicted start position of the mature miRNA(s) for each precursor sequence.

1.2 Thesis Organization

This thesis is organized as follows: chapter 2 presents the biological properties of miRNAs, such as their biogenesis and functionality, and reviews other computational methods that have been developed to identify mature miRNAs. In chapter 3, we describe the methodology used to develop our Bayesian classifier, while in chapter 4 we analyze the training and evaluation of our model and contrast our findings with oter methods. Finally, in chapter 5 we conclude and propose some future work.

Chapter 2

Background Theory

MicroRNAs (miRNAs) are small 19–25 nucleotides long, single–stranded RNAs that are generated from endogenous hairpin shaped transcripts [35]. MicroRNAs function as regulatory molecules in post–transcriptional gene silencing by base pairing with target mRNAs, which leads to mRNA cleavage or translational repression, depending on the degree of complementarity between miRNA and its target transcript.

The first known miRNA, *lin-4*, was discovered in 1993 by Victor Ambros and his colleagues while studing the heterochronic gene *lin-14* in *C. elegans* [41]. Since the discovery of the first miRNA in 1993, thousands of miRNA genes have been identified from a wide range of eukaryotic organisms such as plants, mammals, fish, birds, worms and flies. Although it has been difficult to assign a specific function to miRNAs, important roles are emerging including the control of developmental timing, tumor suppression, cell differentiation and apoptosis.

In this chapter, we review the existing literature regarding the biogenesis (section 2.1), their mechanisms of action (section 2.2) and some of the known functions of miRNAs (section 2.3), as well as computational methods that focus on mature miRNA prediction (section 2.4).

2.1 Biogenesis

Although miRNAs are functionally similar to short interfering RNAs (siRNAs), they are unique in terms of their biogenesis. MicroRNA genes are transcribed into the pri–miRNAs, long double–stranded unstructured precursors, which sometimes can be several thousands bases long, with a 5' cap structure and a 3' Poly(A) tail [43]. It remains unclear which RNA polymerase is responsible for the transcription, although several observations have suggested that RNA polymerase II may be the key

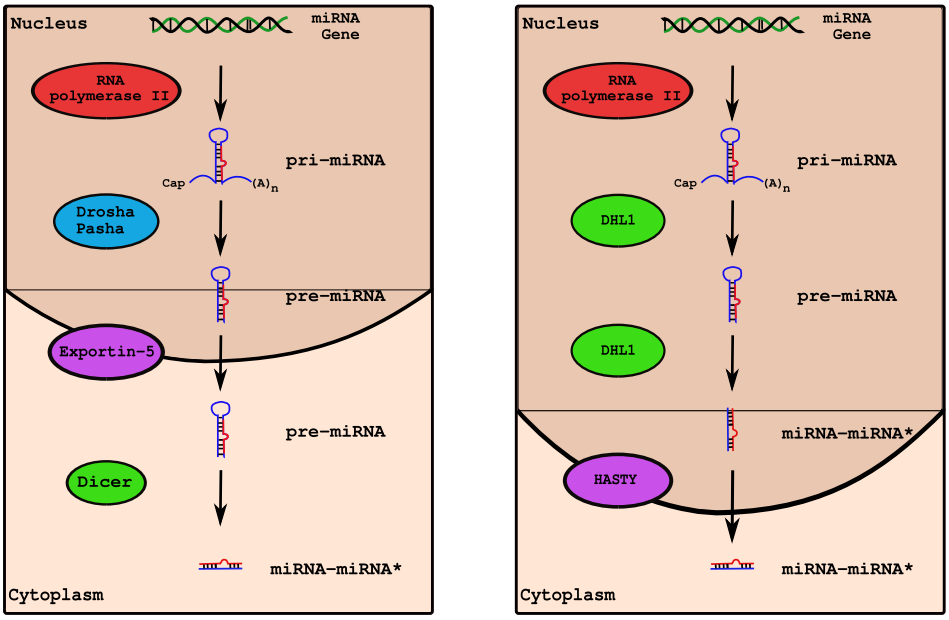
polymerase engaged in miRNA gene transcription [8, 43]. The most important of these findings are:

- The pri–miRNAs are transcribed as long molecules, which sometimes can be several thousands bases long, with a 5' cap structure and a 3' Poly(A) tail, which are unique properties of polymerase II gene transcripts [43].
- Stretches with more than four U's, which terminate the transcription of polymerase III, widely exist in pri–miRNAs sequences [57].

The primary transcript (pri–miRNA) is enzymatically processed in the nucleus by the Microprocessor complex into the precursor miRNA (pre–miRNA), a stem–loop of about 60–100 nt with a 2–nt 3' overhang. The Microprocessor complex in mammals consists of a specific ribonuclease of RNase III endonuclease family called Drosha which acts together with the cofactor called DGCR8 or Pasha. The latter is a double–stranded RNA binding protein that dimerizes with Drosha [39]. It is not very clear how the Microprocessor complex recognizes primary RNA substrates and selects its cleavage sites, since pri–miRNAs in animals don't seen to share any common sequence motifs. The cleavage site identification might result from the 3D structure of pre–miRNA. It was shown that in humans Drosha selectively cleaves RNA hairpin with a large terminal loop, greater than or equal to 10nt. It uses the distance information to decide where to cut: from the junction of the loop and the adjacent stem, Drosha cleaves approximately two helical RNA turns into the stem to produce the pre–miRNA [74].

Following the nuclear processing by Microprocessor in mammals, pre–miRNAs are transported to the cytoplasm by Exportin–5, a nucleus export factor, in a Ran–GTP dependent manner [33, 71]. Exportin–5 was originally known as a minor export factor for tRNAs, because it can transport tRNAs when the primary export factor, Exportin–t, is depleted or overloaded [23]. The binding of Exportin–5 to pre–miRNA is specific because a stem must be larger that 14 base pairs with a base-paired 5' end and a short 3' overhang in order for exportin–5 to bind efficiently [73].

Being exported from the nucleus, pre–miRNAs are subsequently processed into approximately 22 nucleotide miRNA duplexes by the cytoplasmic RNase III Dicer [6]. Dicer is a highly conserved protein that is found in almost all eukaryotic organisms. Some organisms contain multiple Dicer homologues, in which different Dicer isotypes are often assigned to take on distinct roles. For example, in *D.megalogaster*, Dicer–1 is required for pre–miRNA cleavages, whereas Dicer–2 is needed for siRNA generation [42].



(a) Biogenesis in mammals.

(b) Biogenesis in plants.

Figure 2.1: **Model for microRNA biogenesis.** Biogenesis of the mature miRNA is the result of a series of cleavage processes that begins with a primary miRNA. In mammals (a), pri-miRNA is processed in the nucleus into a precursor miRNA hairpin (60-100nt long) by Drosha; then the precursor is transported by Exportin-5 into the cytoplasm, where it is cleaved into the mature miRNA (∼22nt long) by Dicer. In plants (b), the pri-miRNA is processed in the nucleus into a precursor miRNA hairpin (60-100nt long) by Dicer; then the precursor is cleaved into the mature miRNA (∼22nt long) by Dicer also, and the duplex miRNA-miRNA* is transported by HASTY, a plant homologue of Exportin-5, into the cytoplasm.

Biogenesis in plants

The maturation of miRNAs in plants is very different from that in animals. First of all, plant miRNA precursors are quite diverse in structure, and their stem-loops are usually longer than in animals pre-miRNAs. Moreover, no Drosha homologue has been identified in plants so far. However, four Dicer homologues exist in *Arabidosis Thaliana*, and two of these Dicer proteins are likely to be localized in the nucleus. Dicer-like protein-1 (DCL1) possible performs both Drosha and Dicer-like activities for miRNA maturation [53]. Since DCL1 is a nucleus protein, this indicates that mature miRNAs might be generated in the nucleus in plants, unlike animals where the whole precursor is exported to the cytoplasm. While Exportin-5 transfers pre-miRNAs to the cytoplasm in animals, the Arabidopsis homolog of Exportin-5, HASTY, is proposed to export the miRNA-miRNA* duplex to the cytoplasm.

Figure 2.1 shows an overview of the microRNA biogenesis pathway for both mammals and plants. Overall, the biogenesis of a mature miRNA is the result of a series of cleavage processes that begin with a primary miRNA. Moreover, the regulatory mechanisms of miRNA maturation have different complexities between mammals and plants, despite their similarities.

2.2 MicroRNAs in action

After the pre-miRNA is processed into a miRNA-miRNA* duplex by Dicer, one of the RNA strands is incorporated into RISC for target recognition. RISC is composed of Dicer, Argonaute (AGO) and other non–specified proteins. AGO proteins bind to either miRNAs or siRNAs to create the core of the complex. Different Ago paralogs exist across species, and variants of the AGO protein within the same specie can have different functions. It is likely that the different AGO homologues along with the variable associating factors allow for different subtypes of RISC in order to provide a specific response to a particular siRNA and miRNA. RISC has many diverse functions in both siRNA and miRNA mechanisms. It acts as an effector complex in translational repression and mRNA cleavage [12, 14].

Translation repression

MiRNAs in animals mostly suppress translation of their target mRNAs due to an imperfect base–pairing within 3' untranslated regions (UTRs). By binding to the 3'UTR of the mRNA, the miRNA has the ability to inhibit translation by directly interfering with translation initiation factors or by disrupting poly(A) tail function (see figure 2.2).

mRNA cleavage

The distinction between translation repression and mRNA cleavage mediated by miRNAs relies primarily on the degree of complementarity between the miRNA and its target. In plants, miRNA regulation leads to mRNA cleavage due to the near perfect complementarity in base–pairing between miRNA and their target mRNA [44]. In mammals, miRNAs usually don't have perfect complementarity with their target thus leading to translational repression. There are a number of exceptions to the above rules including the mammalian *mir-196* which leads to cleavage of Hoxb8 mRNA instead of the expected translational repression [70] and the plant *miR-172* which acts as a translational repressor [11] . While perfect base–pairing is thought to be the critical feature of miRNA–mediated mRNA cleavage, it is not always sufficient in plants, suggesting the need for supplementary catalytic activity by RISC [11]. Regardless of the species, perfect match between the miRNA and the target mRNA is required for efficient cleavage, especially considering the precise

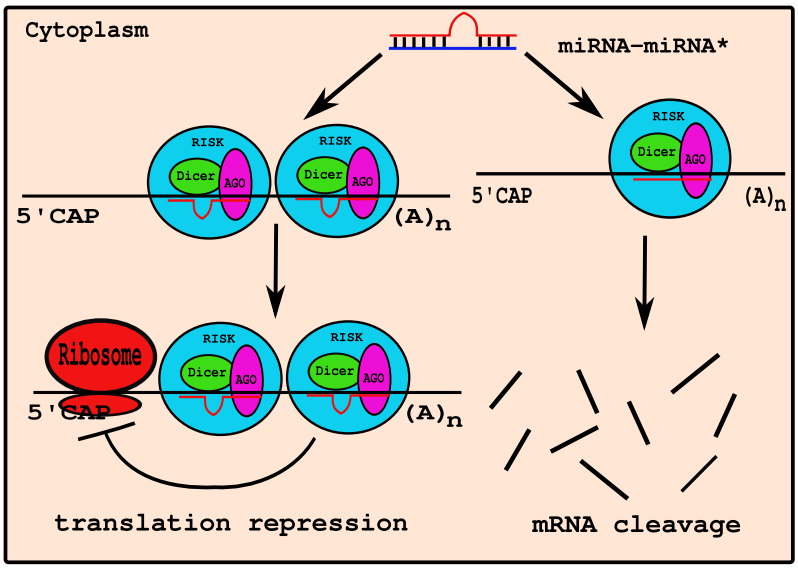


Figure 2.2: **MicroRNAs in action.** The mature miRNA (indicated by red), which forms one strand of the miRNA–miRNA* duplex , is incorporated into a large protein complex, termed RNA induced silencing complex (RISC), where it functions to guide RISC to target mRNA. Depending on the degree of complementarity between miRNA and its target transcript, miRNA either leads to translational repression of the target transcript (not perfect complementarity) or to mRNA cleavage (near perfect complementarity).

location of the cut between residues 10 and 11 of the miRNA [44] (see figure 2.2).

2.3 MicroRNA Functionality

MicroRNAs have been implicated in biological processes ranging from developmental timing to apoptosis. As biologists discover the role of microRNAs in even more processes, the importance of these tiny RNAs molecules will become more clear.

Developmental timing

The first biological process for which scientists discovered the effect of microRNAs was the developmental timing in *C. elegans*, where microRNA *lin-4* and *let-7* played a role [41, 56]. *Lin-4* and *let-7* bind to multiple conserveds sites in the 3'UTR of the *lin-14* and *lin-41* transcripts, respectively, through direct but imprecise base-pairing, thus inhibiting translation [41, 63]. In *C. elegans*, down-regulation of the LIN-14 protein at the end of the first larval stage initiates the second larval stage [41]. Lin-41, on the other hand, regulates the developmental transition from the last larval stage to the adult stage [63]. *Lin-4* and *let-7* also regulate two other genes, *lin-28* and *lin-57*, respectively [48, 1]. *Lin-28* encodes an RNA-binding pro-

tein that is important for neuronal differentiation of embryonic carcinoma cells [68], while *lin-57* encodes a protein responsible for the terminal differentiation of hypodermis in *C. elegans* [1]. Since the original *C. elegans* experiments, the regulatory abilities of *lin-4* and *let-7* have been extended to flies and mammals [59]. In mice, these miRNAs inhibit expression of *lin-41*, which is involved in key developmental events such as limb formation [59]. Studies of three miRNAs in *Drosophila*, *let-7*, *miR-125*(the *lin-4* homolog) and *miR-100*, not only show their up regulation during major points of development but also demonstrate the requirement for concurrent expression of a hormone in order to be functionally expressed [60]. Important mRNA targets responsible for developmental timing have also been found in *Arabidopsis thaliana*, suggesting that miRNA regulation in morphogenesis is a primitive mechanism [17, 3].

Cell proliferation and differentiation

Apart from guiding developmental timing, miRNAs have also been established as potent controllers of cell proliferation and differentiation. Although the division of cells is imperative for the growth of an organism, it can also be detrimental when occurring at inappropriate times. The latter is the hallmark of cancer, and several miRNAs have been shown to be up regulated in tumors (see paragraph 2.3 – Disease). Mutation studies in *Drosophila* show that disruption of miRNA processing causes stem cells to be locked between the G1 and S phases, thus halting division [26]. A neuron-specific miRNA, *miR-132*, is a target of the transcription factor, cAMP–response element binding protein. It regulates neuronal growth by decreasing the levels of a GTPase–activating protein [66]. Another brain–specific miRNA, *miR-134*, is expressed in the synapto–dendritic compartment of rat hippocampal neurons, where it is capable of down–regulating Limk–1, a protein responsible for spine development [58]. Regulatory roles of miRNAs are not limited to the brain. Adipose cell differentiation has been shown to be partially controlled by the expression of *miR-143* [16]. Also, *miR-1* and *miR-133* are important regulators of skeletal muscle proliferation and differentiation [10]. Since cell growth and differentiation are highly dynamic processes, it is no wonder that miRNA with the specific and fast–acting regulatory abilities play a vital role in shaping these processes.

Apoptosis

Apoptosis or programmed cell death, is an integral part of animal tissue development. Apoptosis is an evolutionarily conserved process that allows animals to remove cells that are useless or that are detrimental for survival. Once apoptosis is activated, caspase proteins cleave both the structural and functional elements of the cell. Therefore, cell death and survival depend largely on the control of active caspases in the cell. Because caspases are ubiquitous, it makes sense that miRNAs

would play a role in their regulation. Indeed, in the *Drosophila* eye, the absence of *miR-14* leads to an increase in the cell death effector, Drice, suggesting that *miR-14* is an inhibitor of apoptosis [69]. Likewise, the bantam gene encodes an miRNA that when over–expressed, it suppresses apoptosis in the *Drosophila* retina. One of the identified targets for *bantam* is the pro–apoptotic gene, *hid*, whose mRNA contains sequences that ase complementary to *bantam* [7]. It has been known for quite some time that viruses must prevent apoptosis in order to survive in the host cell. Recently, it has been discovered that the herpes simplex virus-1 inhibits apoptosis through a latency–associated miRNA (miR-*LAT*) that modulates TGF–β signaling [22]. By using miRNAs instead of proteins in the inhibition of apoptosis, viruses are able to survive as well as evade immune detection. As functional studies of miRNA continues, the list of targets involved in apoptosis will likely grow radically.

Disease

Although miRNAs have been established as being vital for animal development, they are also associated with diseases when their repressing activities are compromised. Some of miRNAs, including *miR-143* and *miR-145*, have been suggested to act as tumor suppressors. Thus, their down–regulation leads to tumorigenesis [28, 46]. The exact targets for these miRNAs have not been elucidated, but they are likely to be genes that regulate cell cycle. Another study found that *miR-15* and *miR-16* directly suppresses the BCL2 oncogene [13]. Furthermore, the previously discussed miRNA, *let-7*, has been linked to RAS, a potent activator of cell transformation [31]. Several different miRNA clusters have also been associated with the MYC oncogene [38, 51, 27]. Although the majority of miRNAs are down–regulated in cancers, some including *miR-21* are up–regulated due to their anti–apoptotic effects [9]. Whether *miR-21* has a direct role in cancer progression or is simply differentially modulated in tumors still needs to be clarified.

Due to their involvement in cancer, miRNAs may serve as important targets for therapeutic intervention. Indeed, experiments are already underway in model systems to inactivate miRNAs that may serve as oncogenes [29, 36, 19]. However, the new era of therapeutic targeting of miRNAs is not limited to cancer. A recent study of *miR-375*, a pancreatic-specific miRNA that regulates insulin secretion, suggests that miRNA therapies may also be applicable to diabetes [55]. As more miRNAs are linked to diseases, it is possible that this approach can be applied to virtually any organ system in the body.

Anti-viral defense

While microRNAs are implicated in diseases caused by malfunctions in the cellular machinery, they also play an important role in preventing diseases caused by viruses. Scientists studying plants first proposed that miRNAs may be able to in-

duce post transcriptional gene silencing of viral mRNAs [49, 24]. In plants, miRNAs have anti–viral capabilitiesy with short–lived effects because evolving viral factors eventually inactivate them [62]. In fact, many viruses have the ability to evade silencing by the host, but some viruses are better adapted for evading cellular machinery than others. In humans, for example, the adenovirus can block host miRNA biogenesis, thus squelching the very anti–viral miRNAs that are meant to stop adenovirus replication [45]. Also, tissue culture experiments show that the primate foamy virus type I (PFV-1) can escape silencing by *miR-32* with a silencing suppressor protein called Tas [40]. These observations suggest that host miRNA–mediated defence cannot always overcome viral attacks. However, these experiments do not account for the possibility of defence responses mounted by multiple miRNAs working together. A study of the hepatitis C virus demonstrates that the introduction of multiple siRNAs targeted to different areas of the viral genome prevents the virus from escaping siRNA–silencing [67].

2.4 Related Work

One of the prominent characteristics of miRNAs is that their expression is spatially and temporally regulated. Many miRNAs are highly expressed in certain organs or cell types and some are only expressed in certain stages during development. Considering this tight regulation and their small size, it is no wonder why miRNAs were not identified earlier, despite their wide spread occurrences in different species. From the very beginning, computational approaches have been extensively used in the research for novel miRNAs. Most computational methods focus either on the discovery of new miRNA genes in the genome of various species or the prediction of mRNA targets for the known miRNAs. On the contrary, few attempts have been made to computationally predict the functional part of the miRNA precursor, namely the mature miRNA. A number of studies ([50], [72], [61]) combine miRNA gene prediction with the identification of a possible start position for the mature. To our knowledge, only one study [64] focuses exclusively on mature miRNA prediction.

ProMiR – 2005

Nam *et al.* [50] proposed a probabilistic co–learning method based on paired hidden Markov Model (HMM), called ProMiR, to implement a general miRNA prediction method capable of to identifying close homologs as well as distant homologs. The method combines both sequential and structural characteristics of miRNA genes in a probabilistic framework, and simultaneously decides whether a miRNA gene and a region of mature miRNA are present by detecting the signals for the site cleaved by

Drosha. The accuracy of mature miRNA region prediction through was evaluated through 5–fold cross–validation with 136 known miRNAs. The measures used for assessment were the means of absolute distances and the square root of the mean of the squares. They also evaluated the predicton of the orientation of the mature region, with a mean accuracy of 72%.

BayesMiRNAfind – 2006

Yousef *et al.* [72] presented BayesmiRNAfind, a Naive Bayes classifier that predicts miRNAs based on their secondary structure and sequence. The major novelty of their work is the combination of data from multiple species, in order to create a more stable learning process. As far as the mature miRNA prediction is concerned they assume that only one mature miRNA is associated with each precursor and they are using the mature prediction for extracting the features of the classifier. Furthermore, they do not provide any evaluation performance associated with the mature prediction task.

mirCoS – 2007

Sheng *et al.* [61] proposed a computational method, called mirCoS, that uses three support vector machines (SVM) models sequentially to discover new miRNA candidates in mammalian genomes based on sequence, secondary structure and conservation. The first SVM uses features from sequence conservation of miRNA precursors, the second SVM uses features from the secondary structure of miRNA precursors and its conservation, while the last one focuses on mature miRNA prediction. For the third SVM the most discriminatory features measured the amount and conservation of base–pairing within the part of the predicted secondary structure corresponding to the miRNA. This is readily explained by the fact that mature miRNAs are always on the stems of hairpin structures, and the part of a stem that corresponds to a miRNA tends to have a high level of base pairing. They also estimated the method's performance using a repeated holdout scheme and obtained an average sensitivity of 85%.

Tao – 2007

Tao [64] proposed a method that focuses exclusively on mature miRNA prediction, utilizing thermodynamic and structural information of the precursor RNA. A simple K–NN model is employed for learning and predicting the mature miRNA, using features such as the distance of the mature from the ssRNA tail and loop and the length of the stem, where mean and standard deviation were calculated for each category. The method predicted as mature miRNA the candidate with the smallest score, using a window of 21 base pairs sliding along the precursor's sequence. Using 885 mature miRNA on 866 miRNA precursors as training samples, the algorithm predicted 79.4% of the 2780 test samples on 2722 miRNA precursors in 44 species. For human miRNAs, the prediction rate was 84.7% on 346 test mature

miRNA genes with the rest 127 mature miRNAs serving as training samples.

Let's briefly point out some disadvantages of the current methods used to predict the mature miRNA within a hairpin stem–loop. Most of the above methods hypothesize, direct or indirect, that the hairpin stem-loops contain one only mature miRNA ([50], [72]), a hypothesis that it is not true for a remarkable number of experimentally verified precursors. Moreover, they cannot exactly determine the mature miRNA regions [50] or they are not providing any information of their accuracy for predicting the exact starting position ([72], [61], [64]), an useful information for experimental biologists. Finally, most of these computational tools estimate their performance accuracy in terms of true positive rate alone (sensitivity), ignoring the false positive rate ([50], [61], [64]). It is a matter of semantics as well as a great challenge to define a true negative example when it comes to mature miRNAs. However, a major issue in such a classification task is not only to maximize the identification of true positives but also to minimize the false positive rate. The above bibliography review of methods used to predict the mature miRNA within a hairpin stem–loop makes it clear that there is room for improvement.

Chapter 3

Methodology

In this chapter, we describe in detail the methodology and datasets used in order to build a Naive Bayes classifier capabel of identifying mature miRNAs within a precursor sequence. Firstly, we briefly review the theory of the naive bayes classifier (section 3.1) and then we describe the parameters and hypotheses (section 3.2), the dataset (section 3.3) and the features (section 3.4 and 3.5) used for training our model.

3.1 The Naive Bayes Classifier

3.1.1 A first approach

The Naive Bayes classifier (NBC) is among the most popular classifiers used in the machine learning community and has a wide range of applications. Naive Bayes is a simple probabilistic classifier which is based on the application of the Bayesian theorem (equation (3.1)) and approximates a joint distribution with the product of the individual distributions. In simple terms, a NBC assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about $8\ cm$ in diameter. Even though these features may depend on the existence of the other features, a NBC considers all of them to independently contribute to the probability that this fruit is an apple.

In spite of their naive design and over-simplified assumptions, naive Bayes classifiers often work much better in many complex real–world situations than one might expect. Recently, careful analysis of the Bayesian classification problem has shown that there are some theoretical reasons for the unexpected efficacy of naive Bayes classifiers [75]. An advantage of the NBC is that it requires a relatively

Advantages of Naive Bayes

small amount of training data to estimate the parameters (means and variances of the variables for each class) necessary for classification, since independence of variables is assumed. This is very helpful for overcoming the curse of dimensionality, which requires scaling the data sets exponentially as the number of features increases. Another important advantage is the fact that NBC provides a *direct intuition* about the importance of the features used, especially in comparison to methods such as artificial neural networks or support vector machines, which work more like *black boxes*,that map features into a different, more complex space. Due to these advantages, NBC is very commonly used as the first approach in many classification problems.

3.1.2 The mathematical model

According to the Bayesian classifier, a new sample $\mathbf{x}$, which is described by the feature vector $\mathbf{x} = \langle x_1, x_2, \ldots, x_n \rangle$ and whose class label is unknown, will be assigned to the class ω_j among a finite set of possible classes $C = \{\omega_1, \ldots, \omega_c\}$, that minimizes the overall risk based on its features $\mathbf{x}$, according to the following formula:

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

where:

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

$P(\omega_j|\mathbf{x})$ is the posterior probability of class membership, meaning the probability that $\mathbf{x}$ belongs to ω_j and can be computed using the Bayes' formula (see section 2.9 of [15]):

$$P(\omega_j|\mathbf{x}) = \frac{P(\mathbf{x}|\omega_j)P(\omega_j)}{P(\mathbf{x})}, \tag{3.1}$$

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

The Naive Bayes classifier is based on the simplifying assumption that the input features among samples of any given class are conditionally independent given the class [47]. In other words, given the class of a sample, the probability of observing the conjunction $x_1, x_2, \ldots, x_n$ is just the product of the probabilities for the individual features of this sample:

$$P(\mathbf{x}|c_j) = P(x_1, x_2, \ldots, x_n|c_j) = \prod_i^n P(x_i|c_j).$$

Although the assumption that the predictor variables are independent is not always accurate, it does simplify the classification task dramatically, since it allows the class conditional densities $p(x_k|\omega_j)$ to be calculated separately for each variable k for each class ω_j, meaning it reduces a multidimensional task to a number of one–dimensional tasks. In effect, Naive Bayes reduces a high–dimensional density estimation task to an one–dimensional kernel density estimation.

3.2 Our model

In this thesis, we describe a method that uses a NBC for the identification of the mature miRNA(s) within a miRNA precursor. More specific, the observations for classification (i.e. the samples) are mature miRNA candidates that are produced from a miRNA precursor sequence by sliding a window of a specified size along the precursor. Each mature miRNA candidate is described by a set of features $\mathbf{x} = \langle x_1, x_2, \ldots, x_n \rangle$, that we consider to be independent, and it can be classified as a mature miRNA (*positive class* – denoted ω_1), or as non-mature miRNA (*negative class* – denoted ω_{-1}).

Ideally, we want to classify each sample to the class that minimizes the classification error, based on our training model. The simplest case is to consider that all misclassification errors have the same cost, using the zero–one loss function. Under these assumptions, the Bayes Decision Rule is converted to the following equation:

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

for some threshold $\lambda \in \mathbb{R}$ (see section 2.3 of [15]). Since $P(\mathbf{x})$ is only a normalization factor, it can be omitted in order to minimize calculation time, leaving the classification unchanged. Moreover, we assume that the prior probability is 50%

for both classes, which prevents us from favoring a particular class. Under these assumptions, the Bayes Decision Rule is given by the following simplified formula:

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

for some threshold $\lambda \in \mathbb{R}$.

3.3 Datasets

In section 3.2 we mentioned that each mature candidate can be classified as a mature miRNA (*positive class*), or as non-mature miRNA (*negative class*). This assumption formulates the mature identification problem into a two–class problem. As for any typical two–class classification problem, data samples from both classes are needed in order to train the classifier.

Positive dataset for training

The *positive class* is the main class of interest, i.e. the mature miRNAs. For the training procedure, we use as positive data the precursors of experimentally verified human and mouse microRNA downloaded from the miRBase Sequence Database (version 10.1, [32], [20], [21]). The human dataset consists of 533 precursors which produce 729 mature miRNAs, while the mouse dataset consists of 422 precursors which produce 530 mature miRNAs. We consider precursor and not just mature miRNA information, since some mature miRNAs come from more than one precursors [2]. Moreover, precursor information can provide more training examples depending on type of features used (see section 3.4).

Negative dataset for training

The definition of the *mature miRNA* is straight forward, but what is a non–mature miRNA? In order to answer this question and to create the *negative class* we consider two hypothesis:

1. As we already mentioned in section 3.2 the search area of the classifier will be a miRNA precursor sequence.

2. It has been observed that until now miRNA precursors do not produce multiple overlapping mature miRNAs from the same arm of the foldback precursor [2]. A hypothesis that holds in our positive dataset.

Based on the above constrains, we generate a set of negative examples from the precursor sequences in the following way: for each true mature miRNA, we use a same–size sliding window and select all possible “negative” matures which can be created by sliding 1 base pair towards either direction from the mature, excluding

any hairpin loops and the true mature. This procedure results in a very large negative set, where each true mature has a variable number of respective "negatives", depending on the length and number of precursors it comes from. In order to minimize computational time during the training procedure and at the same time have a good representation of the precursor for the areas where true mature miRNAs do not lay, we randomly select a subset of 10 negative examples for each true mature.

The evaluation dataset

Apart from the data used to train the classifier, we consider a final, blind dataset to evaluate the performance of our classifier. The dataset contains all new miRNA precursors of human and mouse, that were published under version 11 and 12 of the miRBase Sequence Database ([32], [20], [21]). The dataset consists of 155 human precursors, which produce 160 mature miRNAs, and 45 mouse precursors, which produce 48 mature miRNAs.

3.4 Features

The goal of this work is to produce a model that recognizes the mature miRNA(s) within each precursor sequence, as *Dicer* does in the real cell. Until now, the only information we share with *Dicer* is the sequence and the secondary structure of a miRNA precursor. MicroRNA precursors have a unique secondary structure forming irregular hairpin structures with various internal symmetric and non–symmetric loops, bulges and hairpins. Figure 3.1 presents a typical example of the secondary structure of a precursor miRNA. More specifically, it is the secondary structure of the human precursor *has-let-7a-1*, as it was produced by the RNAfold program [52].

Position oriented features

One would expect that the areas on the edges of the mature miRNA would have a common pattern that is recognized by *Dicer*. In order to evaluate whether this hypothesis holds, we consider the information of the sequence and the secondary structure of a precursor and we represent a mature miRNA as a sequence of positions. Each position is a single feature and contains sequence information (A, C, U, G) and/or structural information (match or mismatch), derived from each respective precursor(s). Notice that we simplify the secondary structure information that is provided by a typical secondary structure program, such as RNAfold, from hairpin, loops and bulges into match or mismatch, in order to represent the secondary structure into the position level without any loss of information. For example, figure 3.2 shows a feature found in position 2 [1] of the mature miRNA (indicated with red), which contains the information about its sequence (A) and its

[1] Position counting starts with zero.

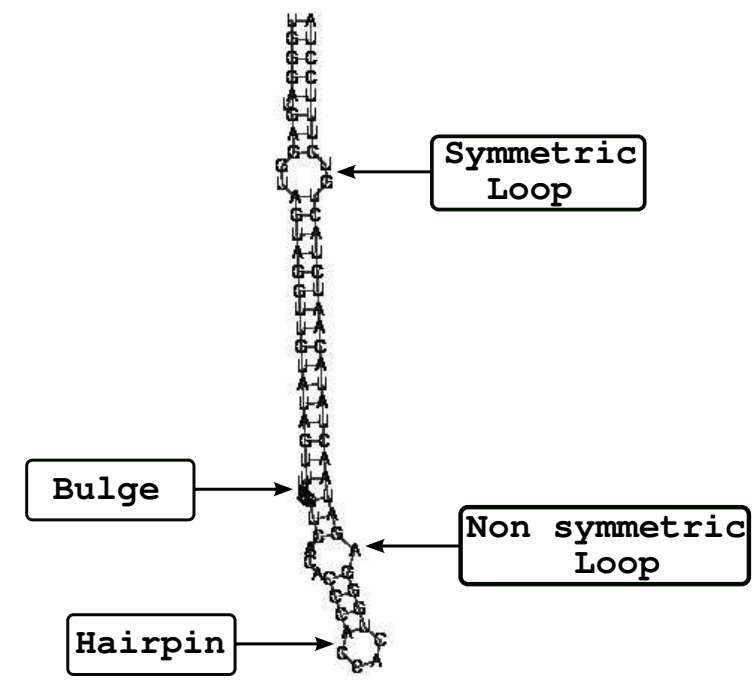


Figure 3.1: **The secondary structure of a miRNA precursor.** A typical example of the secondary structure of a precursor miRNA (the human precursor *hsa-let-7a-1*, as it was produced by the RNAfold program [52]), contains hairpins, bulges, symmetric and non–symmetric loops.

secondary structure (match).

The features that characterize the mature miRNA may lie in positions within the mature miRNA, but may also lie within a flanking region of variable size that extends symmetrically (or not) along both sides of the mature sample. It is possible that a feature that lies outside the mature sample, could also lie outside the precursor, depending on where the starting or ending position the mature miRNA lie within the precursor. In this case, the feature gets a special value indicating the lack of information. Figure 3.3 shows the areas where the position oriented features lie as a precursor miRNA folds in its secondary structure. With red color is indicated the mature miRNA, while green circles indicate the two flanking regions of specified length (i.e. 5nt) around the mature. Notice that the number of *position oriented* features depends on the length of the mature miRNA and the size of the flanking regions. More specifically there are $2 * N +$ mature length *position oriented* features that describe a mature miRNA, where N is the size of the flanking region.

Figure 3.4 shows the distributions of two *position oriented* features as they are

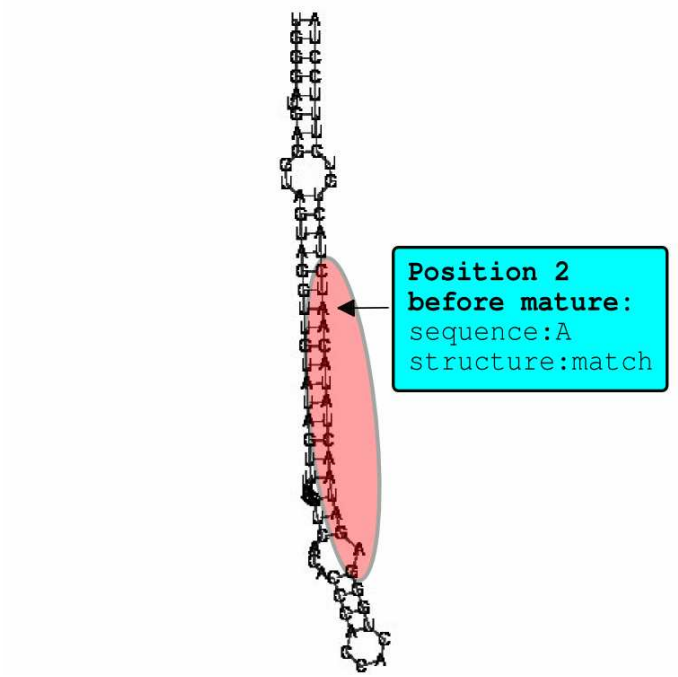


Figure 3.2: **An example of a position oriented feature.** The feature lies in position 2 within the mature region, indicated with red, and contains information about its sequence (A) and about its secondary structure (match).

calculated by the training data (see section 3.3). Both features are found in the flanking region before the actual mature miRNA, position 8 and 9 respectively, and have high divergence between their positive and negative data (see section 3.5 and table A.1).

Distance oriented features

Apart from the position oriented features, we also consider four additional features: the distances of the starting and ending position of a mature miRNA from the closest hairpin and the distance of the starting and ending position of a mature miRNA from the 5' or 3' precursor's end, depending whether the mature lies on the 5' or 3' stem respectively. Figure 3.5 shows the distance oriented features of a mature miRNA (indicated with red) which is found on the 5' stem of a miRNA precursor. In a similar way one can define the distance oriented features for a mature miRNA which is found on the 3' stem of a precursor. It should be noticed that these features have two distict distributions depending on stem the mature miRNA laye within the precursor sequence.

An example distribution of such a feature can be seen in figure 3.6. The feature shown is the starting position of a mature miRNA from the closest hairpin as is

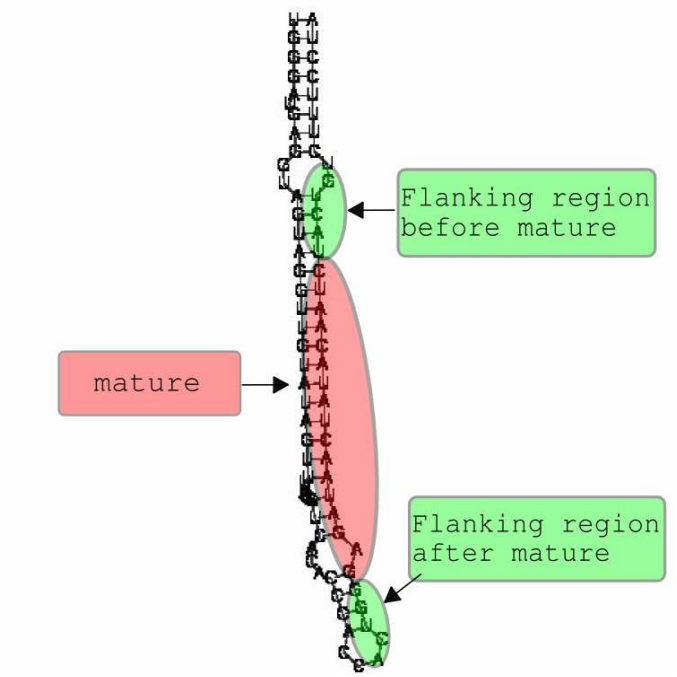


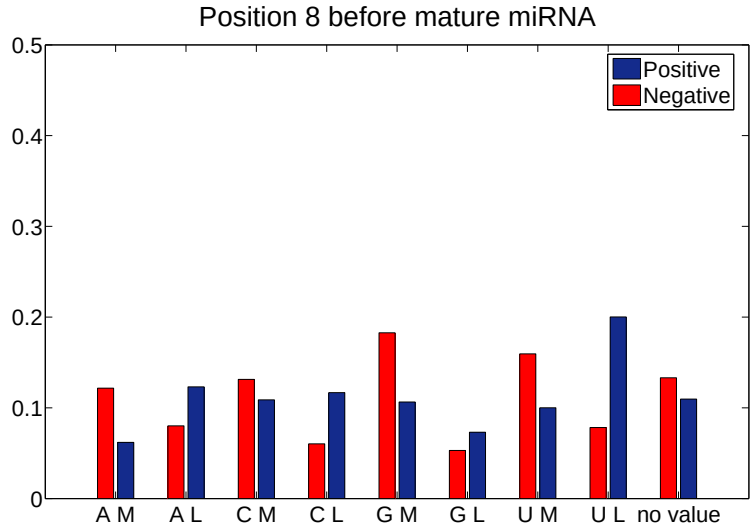
Figure 3.3: **The regions of position oriented features.** The regions where the position oriented features lie as the miRNA precursor folds in its secondary structure are the mature region, indicated with red, and the two flanking regions, indicated with green, that extends symmetrically around the mature region.

calculated by the training data (see section 3.3). The true mature miRNAs (positive data) tend to start in positions close to the hairpin, while the non-mature miRNAs (negative data) tend to form the uniform distribution, because of the way they were produced (see section 3.3).

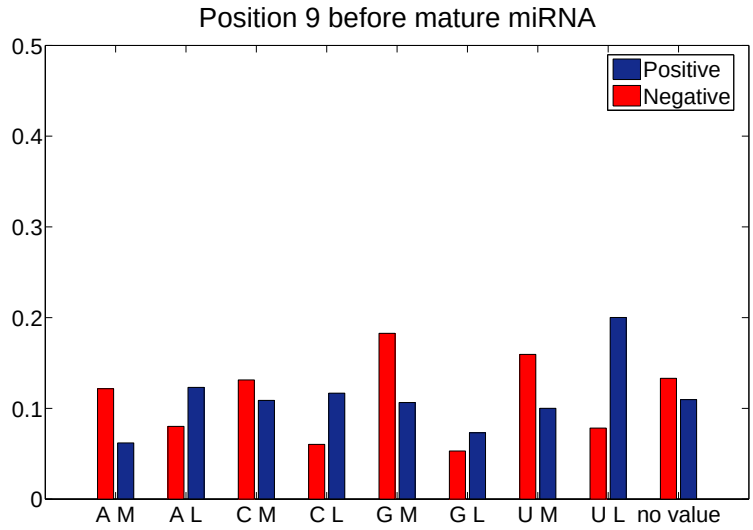
3.5 Feature Selection

As we already mentioned in section 3.4 there are two types of features used to describe a mature miRNA in our system, the *position oriented* and the *distance oriented* features. In order to select a set of features that contain discriminatory information between true matures and our negative samples, we rank our features using the symmetric Kullback–Leibler divergence metric (see below paragraph "Kullback–Leibler divergence") to measure the difference of the feature distributions for the positive and negative data.

Specifically, we follow the procedure below:



(a) Position 8 before the mature miRNA.



(b) Position 9 before the mature miRNA.

Figure 3.4: **Example distributions of two *position oriented* features.** The distributions are calculated based on the training data.

1. For each feature, either *distance oriented* or *position oriented*, we estimate the probability mass functions in both positive and negative data.

2. Using the symmetric K-L divergence, we estimate a score for each feature that measures how different the probability mass functions are between the two

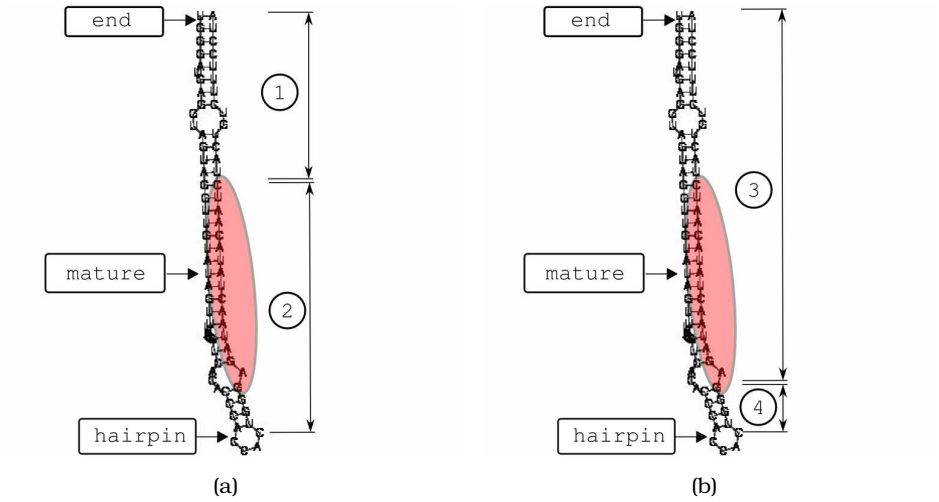


Figure 3.5: **The distance oriented features of a mature miRNA (indicated with red) which is found on the 5' stem of a precursor miRNA.** (a) The distances of the starting position of a mature miRNA from the 5' end of the precursor (Distance 1) and from the closest hairpin (Distance 2). (b) The distances of the ending position of a mature miRNA from the 5' end of the precursor (Distance 3) and from the closest hairpin (Distance 4).

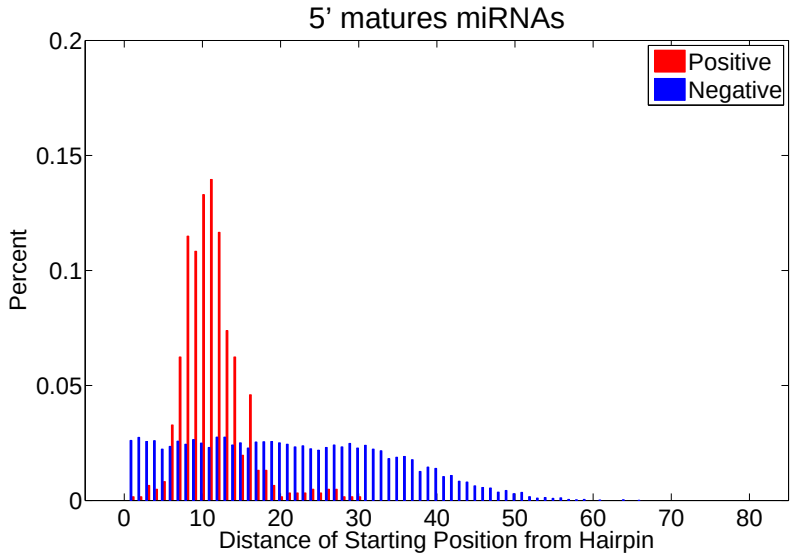
classes.

3. We rank the features according to the K-L provided score. Large distances are considered more informative.

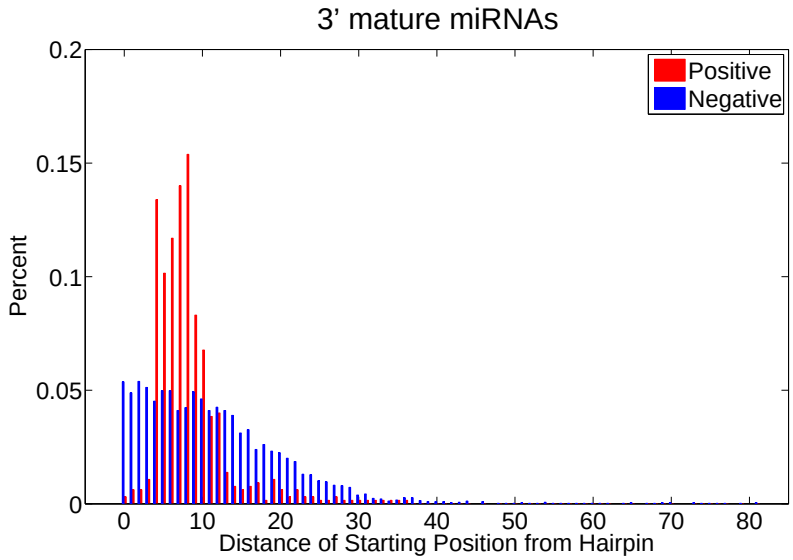
4. We then train the classifier using the top K features. Each feature is incopo-rated gradually into the classifier only if it helps increasing the performance of the classifier based on some evaluation metric. We vary both N, the size of flanking region, and K, the number of features used, until we find the optimal classifier.

Variable ranking advatages

The features selection method used in our model is a typical *variable ranking* method. *Variable ranking* method is a filter method, which is a preprocessing step, independent of the choice of the predictor. Still, under certain independence or orthogonality assumptions, it may be optimal with respect to a given predictor. For instance, using Fisher's criterion to rank variables in a classification problem where the covariance matrix is diagonal is optimum for Fisher's linear discriminant classifier [15]. Even when variable ranking is not optimal, it may be preferable to other variable subset selection methods because of its computational and statistical



(a) Distribution for 5' mature miRNAs.



(b) Distribution for 3' mature miRNAs.

Figure 3.6: **An example distribution of a *distance oriented feature*.** The feature is the starting position of a mature miRNA from the closest hairpin as is calculated by the training data.

scalability: Computationally, it is efficient since it requires only the computation of n scores and sorting the scores; Statistically, it is robust against overfitting because it introduces bias but it may have considerably less variance [25].

Kullback–Leibler divergence

Finally, the Kullback–Leibler divergence metric (K–L divergence) is a measure of the difference between two probability distributions [37]. For Probability Mass Functions (PMFs) P and Q of a discrete random variable, the K–L divergence of Q from P is defined as:

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

Two fundamental properties of $D_{KL}(P|Q)$ are:

- *non-negativity*: $D_{KL}(P|Q) \geq 0$ with equality if and only if $P = Q$.
- *asymmetry*: $D_{KL}(P|Q) \neq D_{KL}(Q|P)$.

Unfortunately, the property of asymmetry is the reason why K–L divergence is not a true distance metric. To overcome this problem we used the symmetric and nonnegative Kullback–Leibler divergence [30], which is defined as:

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

and is commonly used in classification problems.

Chapter 4

Results

In this chapter, we discuss the construction and fine tuning procedures for the Naive Bayes Classifier that was described in detail in chapter 3 (see sections 4.1, 4.2 and 4.3) and compare its performance with that of two existing tools the *BayesMiRNAfind* [72] and *ProMiR* [50] (see section 4.4).

4.1 Training the Naive Bayes Classifier

According to the hypotheses reported in section 3.2, there are a number of parameters that need to be tuned, in order to get the optimum Naive Bayes classifier. The main parameters are the size of the flanking regions, *N*, the number of features, *K*, used in the classifier and the type of information for the *position oriented* features.

A typical method for tuning the model's parameters is the m–fold cross–validation procedure [34], where the data are split into m subsets and a portion of them $\left(\frac{j}{m}\right)$ are used for training, while the remaining $\left(\frac{m-j}{m}\right)$ data are used for validation. This is repeated iteratively until all data are used for both training and validation. In this case, we use a 10–fold cross validation procedure and in order to ensure a realistic estimation of the classifier's performance, the validation sets consist of true miRNA precursors, instead of the mature miRNAs alone. It should be noted that the miRNA precursors used in the validation set correspond to the mature miRNAs that were left out from the training sets during cross–validation. Classification performance on the validation set is estimated using a sliding window of fixed size, whereby all possible mature candidates generated by sliding 1 base pair in both stem arms of the precursor apart from the hairpin loop(s), are assigned to one of the two classes. It is important to note that classification performance is estimated based on exact match of the starting position of the predicted compared to the real mature miRNA.

Even 1nt deviations are considered as negative examples. Based on the above convention, another parameter is introduced into the model, namely the size of the sliding window, *W*.

4.1.1 *Position oriented* features selection

Our first goal is to identify what type of *position oriented* features are more useful for the classification task at hand. In section 3.4, three categories of *position oriented* features were presented, depending on the type of information they contain:

1. Sequence Type, containing only sequence information (A, C, U, G).

2. Structure Type, containing only information of the secondary structure (match or mismatch).

3. Combined Type, containing information of both sequence and secondary structure.

The discriminatory power of each feature category is estimated via assessing the classification performance of Naive Bayes classifiers over 10–fold cross validation procedure. As mentioned above, for each Naive Bayes classifier three parameters need to be tuned: *a*) the size of the flanking region, N, which is assumed to lie within $N \in \{0, 5, 7, 10, 12\}$, *b*) the size of the scanning window, W, which is ssumed to be $W = 22nt$ and *c*) the number of *position oriented* features used into the classifier, K, which is assumed to lie within $K \in \{1, 2, \ldots, N + W\}$, since they can be located either within the mature miRNA, or inside the flanking regions around it. Moreover, the classification performance is based on Matthew's correlation coefficient [4], a measure of the quality of binary classifications. It is generally regarded as a balanced measure which can be used even if the classes are unbalanced. It returns a value between -1 and $+1$, where $+1$ represents a perfect prediction, 0 an average random prediction and -1 an inverse prediction.

Tables 4.1, 4.2 and 4.3 show the top scoring classifiers, based on Matthews Correlation Coefficient (MCC) calculated for threshold $\lambda = 1$ (see section 3.2), for the three categories of input features, each utilizing location-specific information about the sequence, the structure and both the sequence and structure of the training examples respectively. Each table shows the sensitivity, specificity and Matthews Correlation Coefficient (MCC) [4] achieved with different numbers of such features (*position oriented*) and with different sizes of flanking regions around the

Table 4.1: The Sequence–Based Naive Bayes Classifiers trained with emphposition oriented features containing only sequence information.

Classifier's Description	**Sensitivity**	**Specificity**	**MCC**
Combination of 12 Features, 0nt flanking region	67.10%	55.10%	0.0850
Combination of 16 Features, 5nt flanking region	76.04%	53.34%	0.1074
Combination of 31 Features, 7nt flanking region	75.96%	53.20%	0.1071
Combination of 19 Features, 10nt flanking region	79.15%	47.01%	0.0960
Combination of 35 Features, 12nt flanking region	74.30%	51.33%	0.0945

Table 4.2: The Structure–Based Naive Bayes Classifiers trained with *position oriented* features containing only structure information.

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

mature miRNA. Note that the positions along the precursor which served as input features were selected based on the K–L divergence metric (see section 3.5).

We found that as the size of the flanking region increased, the sensitivity of the classifiers tended to improve, while the specificity remained relatively unaffected, independently of the type of features used. This improvement seemed to reach a maximum for a flanking region of about 10nt. For classifiers with flanking regions of 12nt utilizing either sequence or structure information (Tables 4.1 and 4.2 respectively), the extra features did not further improve the accuracy, suggesting that

Table 4.3: The Combined Naive Bayes Classifiers trained with *position oriented* features containing both sequence and structure information.

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

they probably add more noise than useful information.

Moreover, the classifiers utilizing features with combined information for both sequence and structure achieved an overall better performance -in terms of improved specificity and MCC- than the ones using sequence or structure information alone. Note that a high specificity score is particularly important in this task, since the number of negative examples is much larger than the number of positive ones, suggesting that the *position oriented* features that utilize both sequence and secondary structure information have higher discriminatory power that the other two categories.

4.1.2 Tuning the parameters of the model

As we already mentioned in order to get the optimum Naive Bayes classifier, a number of parameters need to be tuned. In the previous subsection (4.1.1), we examined one of the parameters, the discriminatory power of different types of *position oriented* features, and showed that features which combined information of both sequence and secondary structure are more powerful. In this subsection we examine the rest of the parameters and present the best Naive Bayes classifier, based on our hypotheses (see section 3.2).

Area under the Curve

We trained a number of Naive Bayes classifiers using a 10–fold cross validation procedure with different parameters values. The classification performance was assessed using Area Under the Curve (AUC) of the average receiver operating characteristic (ROC) curve calculated using the threshold averaging algorithm in-

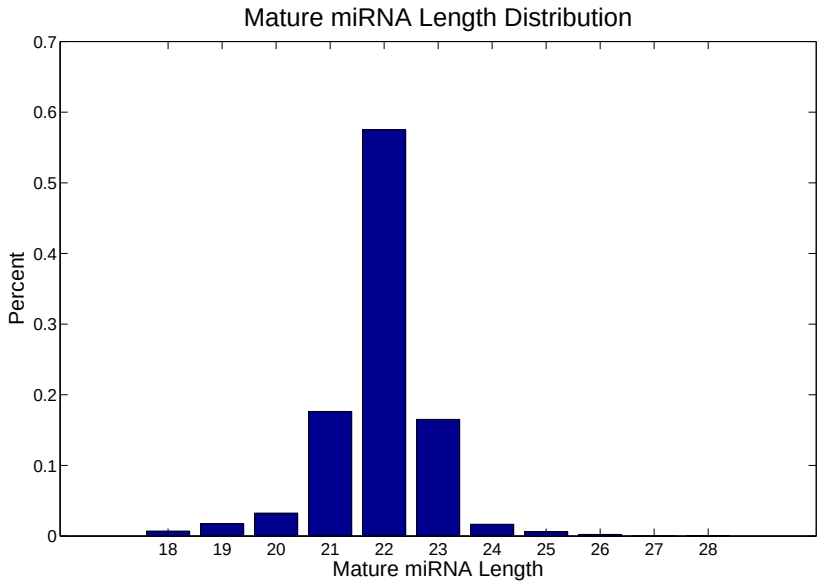


Figure 4.1: The length distribution of experimentally verified human and mouse mature miRNA's from the miRBase Sequence Database (version 10.1, [32], [20], [21]).

troduced by Fawcett [18]. We use AUC as a classification performance instead of the MCC, which was used in the previous section (4.1.1), because it is not limited by a specific λ threshold and is insensitive to both skewed class distributions and unequal classification error costs. Finally, the AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance [18], which is more intuitive than the MCC.

Parameters' Values

A set of different values for the parameters in question were selected based on the results of the previous set of experiments (subsection 4.1.1). Regarding the size of the flanking region, we tested values within the $N \in \{0, 3, 5, 7, 9\}$ set, since *position oriented* features derived from langer flanking regions didn't improve the classification accuracy. Regarding the scanning window, W, values $\{18, 20, 22, 24\}$, were investigated. Note that 18 is the size of the smallest mature miRNA in our training data, and 22 is the average size (see figure 4.1). Regarding the number of *position oriented* features used to train the classifier, K, we followed an incremental approach where $K \in \{1, 2, \ldots, N + W\}$. Features were added as described in 3.5 until no more improvement could be achieved. Four *distance oriented* features

Table 4.4: The AUC of the average ROC curve, over the 10–fold cross validation, of the naive bayes classifiers with *distance oriented* features for every scanning window value. In the array below with HS and HE are representing the distances of the starting and ending position of the mature from the hairpin respectively, and with ES and EE the distances of starting and ending position of the mature miRNA from the ends of the precursor respectively.

***Distance oriented* Features**	**Window 18nt**	**Window 20nt**	**Window 22nt**	**Window 24nt**
HS	0.8181	0.8155	0.8128	0.8147
HS-HE	0.7794	0.7914	0.8099	0.8100
HS-HE-ES	0.7621	0.7803	0.7787	0.7866
HS-HE-ES-EE	0.7587	0.7808	0.7875	0.7839

denoting the distance of the starting and ending position of the mature from the hairpin and the precursor, respectevely, were also examined.

Classifiation using *distance oriented* features

As we mentioned in section 3.5, we are using a *variable ranking* method to select the order of introducing our features into the classifier. The features with the highest Kullback–Leibler divergence, the metric for ranking the features, were the *distance oriented* features (see table A.1). Tables 4.4 shows the performance of the Naive Bayes classifiers that were trained using *distance oriented* features alone for every value of the scanning window W. We found that the most powerful feature based on Kullback–Leibler divergence (see table A.1), is the distance of the starting position of a mature miRNA from the closest hairpin (HS). Using this feature alone results a classification performance of approximately 0.81 AUC in the 10–fold cross validation procedure. Adding the rest of the *distance oriented* features unfortunately decreased the AUC below 0.80, suggesting that they probably add more noise than useful information.

Thus, in the next set of experiments we consider the combination of HS with *position oriented* features containing both sequence and secondary structure information. As shown (see tables A.2, A.3, A.4, A.5 and A.6) adding *position oriented* features into the classifier futher improves the classification performance . *Position oriented* features were inserted according to the Kullback–Leibler score, as long as their respective positions lied within the mature or flanking regions of the classifier. Table 4.5 shows the best Naive Bayes classifiers using HS and *position oriented* features for every combination of flanking region and scanning window. For each classifier the table presents its performance in terms of AUC over the average ROC

curve and the number of *position oriented* features used.

The best Naive Bayes classifier

The classifier with the highest performance on the cross–validation task was the one trained with 37 *position orieted* features and the distance of the starting position of the mature miRNA from the hairpin (HS feature). The optimal flanking region was $N = 9nt$ and the optimal scanning window $W = 22nt$. The highest performance achieved was $AUC \approx 0.88$ over the 10–fold cross validation, although the rest of the classifiers in table 4.5 have similar classification performances. To get a better estimate of the classifier's performance we tested our model against a blind dataset (see paragraph "The evaluation dataset" section 3.3). Figure 4.2 shows the ROC curves of the best classifier both in the cross validation (green line) and the evaluation dataset (black line). For each point in the ROC curve of the cross validation procedure (green line) we also provide the standard deviation for both false and true positive rate (red and blue line respectively). The AUC in the average ROC curve is ~ 0.88, while in the blind evaluation dataset is ~ 0.80. Even though the AUC decreases in the evaluation dataset, it remains sufficiently high.

4.2 Finding the best mature candidate

The purpose of this thesis is to create a model that would predict the mature miRNA(s) that is(are) produced by a precursor miRNA. The Naive Bayes classifiers will classify the candidates, that are created by shifting a scanning window of a specified size 1nt at a time along the precursor stems, into mature or non–mature. Based on the classification performance even our best classifier ($AUC \approx 0.88$ over the cross validation) will classify a number of candidates as mature for each precursor depending on a selected score threshold λ.

For example table 4.6 shows the mature candidates of the precursor *hsa-mir-576* using score threshold $\lambda = 1$. We select a score threshold $\lambda = 1$, which based on the average ROC curve of the cross validation (see figure 4.2), has an average Sensitivity of $86.95\% \pm 0.0348$ and an average Specificity of $73.27\% \pm 0.0120$. For each mature candidate, which is represented by its starting position within the precursor (column "Position"), the table shows the Bayesian score (column "Bayesian score") and the distance from the closest true mature miRNA of the precursor (column "Distance from Truth"). The candidates per stem are sorted based on the Bayesian score and it should be noted that consecutive candidate positions have close ranking positions. The precursor *hsa-mir-576* actually produces two mature miRNAs, one in position 15 on the 5' stem and one in position 54 on 3' the stem, but our best

Table 4.5: The AUC of the average ROC curve, over the 10–fold cross validation, of the best naive bayes classifiers for every combination of flanking region and scanning window.

Flanking Region	**Window 18nt**	**Window 20nt**	**Window 22nt**	**Window 24nt**
0nt	17 Position Features 0.8629	17 Position Features 0.8615	18 Position Features 0.8621	18 Position Features 0.8624
3nt	19 Position Features 0.8671	20 Position Features 0.8658	23 Position Features 0.8675	23 Position Features 0.8661
5nt	19 Position Features 0.8597	20 Position Features 0.8614	27 Position Features 0.8662	21 Position Features 0.8642
7nt	5 Position Features 0.8592	24 Position Features 0.8630	31 Position Features 0.8716	25 Position Features 0.8696
9nt	16 Position Features 0.8599	34 Position Features 0.8673	37 Position Features 0.8771	35 Position Features 0.8704

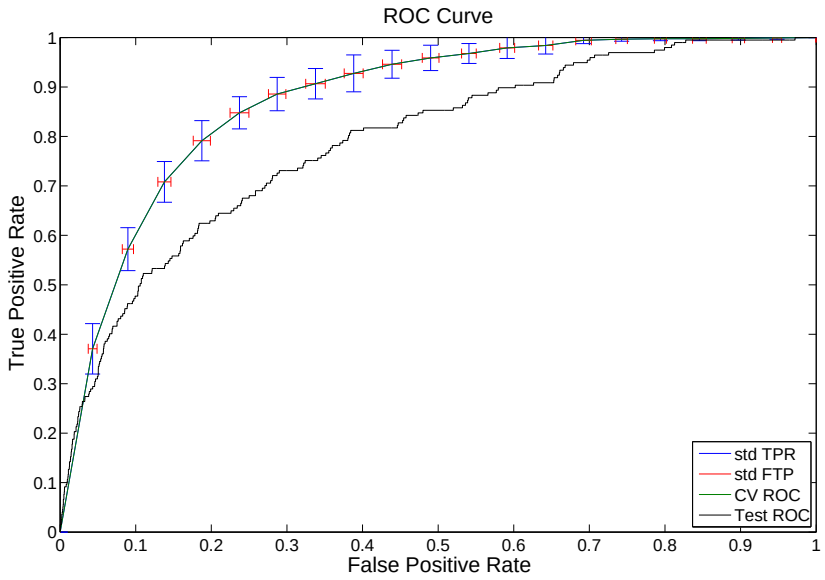


Figure 4.2: The ROC curve of the *Best Naive Bayes Classifier*. The green line represents the average ROC curve over the 10–fold cross validation, with blue the standard deviation of the true positive rate (TPR) and with red the standard deviation of the false positive rate (FPR), while the black line represents the ROC curve over the final blind dataset. The average AUC over the cross validation is 0.8771, while the AUC of the blind, test dataset is 0.791.

classifier provides 5 mature candidates on 5' stem and 13 mature candidates on 3' stem using score threshold $\lambda = 1$, respectively.

As we already mentioned it is known that experimentally miRNA precursors do not produce multiple overlapping mature miRNAs from the same arm of the fold-back precursor [2], and providing more than one position per stem as candidates may not be so useful for a biologist. In order to overcome this problem, we next try to provide one mature candidate per stem by using our best Naive Bayes classifier as a ranker and combining the highest scoring candidates to produce one computational candidate/truth. The evaluation of the methods of computational truth will be in terms of distance from the true mature miRNAs and our goal is to find the candidate with the smallest possible distance provided with the highest confidence. Note that distance in this case corresponds to the difference of the start position between the true mature and the predicted candidates.

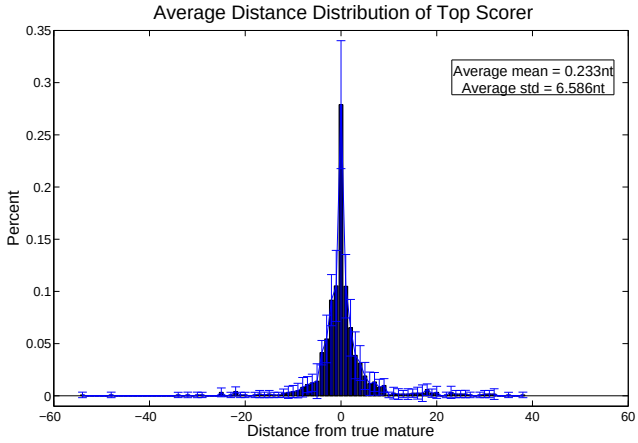
Table 4.6: Mature candidates of precursor *hsa-mir-576* from our best naive bayes classifier with score threshold $\lambda = 1$. The candidates are sorted by Bayesian score per stem and for each of them the table shows its Bayesian score and the distance from the closest true mature miRNA.

5'Stem – True mature:15			**3'Stem – True mature:54**		
Position	**Bayesian Score**	**Distance from Truth**	**Position**	**Bayesian Score**	**Distance from Truth**
16	24.61	1	53	145.07	-1
15	17.68	0	52	96.22	-2
14	12.01	-1	54	95.30	0
17	8.24	2	51	46.66	-3
18	5.68	3	50	25.80	-4
			60	11.91	6
			63	11.72	9
			65	11.54	11
			55	7.54	1
			67	5.18	13
			66	5.13	12
			68	2.30	14
			56	1.32	2

4.2.1 Finding the computational truth

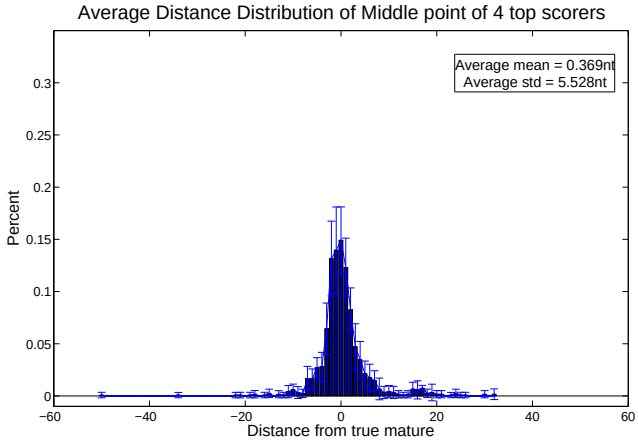
Average Candidate

The first idea is to provide the top scorer per stem as the computational truth with a score threshold $\lambda = 1$. Figure 4.3 shows the average distance distribution of the top scorers from the true mature miRNAs for the stems that produce mature miRNAs over the 10–fold cross validation. The average mean of the distribution is $0.2337nt$, while the average standard deviation is $6.586nt$. It should be noted that the 86.88% of the computational truth was $\pm 6nt$ away from truth. We also examined as computational truth the middle point of the positions' space defined by n top scorers and the mean value of n top scorers, where $n \in \{2, \ldots, 6\}$ (see tables A.7 and A.8). The best results for both computational solutions were obtained for $n = 4$. Figures 4.4 and 4.5 show the average distance distributions of the computational truth from the true mature miRNAs, only for the stems that produce mature miRNAs, over the 10-fold cross validation for $n = 4$. The computational truth in figure 4.4 is the middle point of 4 top scorers, while in figure 4.5 is the mean value of 4 top scorers, using in both cases a score threshold $\lambda = 1$. The average mean of the average distance distribution for the middle point of the 4 top scorers is



Distance from truth	Percent
0	27.89%
±1	48.91%
±2	64.59%
±3	73.92%
±4	81.18%
±5	84.48%
±6	86.88%
±7	89.28%

Figure 4.3: The average distance distribution over the 10–fold cross validation (left) and the percent for each distance away from the truth (right), when the computational truth is the *top scorer* using score threshold $\lambda = 1$.



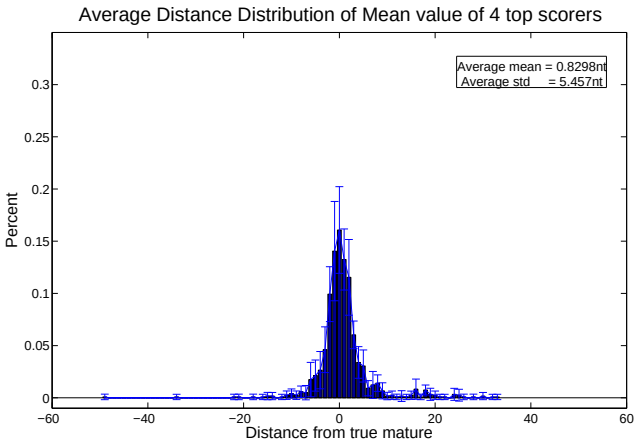
Distance from truth	Percent
0	14.92%
±1	41.18%
±2	62.60%
±3	73.74%
±4	80.00%
±5	84.88%
±6	88.25%
±7	91.39%

Figure 4.4: The average distance distribution over the 10–fold cross validation (left) and the percent for each distance away from the truth (right), when the computational truth is the *middle point* of the range defined by the 4 top scorers using score threshold $\lambda = 1$.

$0.3694nt$ with average standard deviation $5.5208nt$, while the average mean for the mean value of 4 top scorers is $0.8298nt$ with average standard deviation $5.4579nt$. Finally, the 88.25% of the middle point computational candidates and the 89.34% of the mean value computational candidates were within $\pm 6nt$ distance from truth.

The main problem with this approach is that even for precursors which produce, a single mature miRNA, our model will also provide a mature candidate for the

miRNA duplex

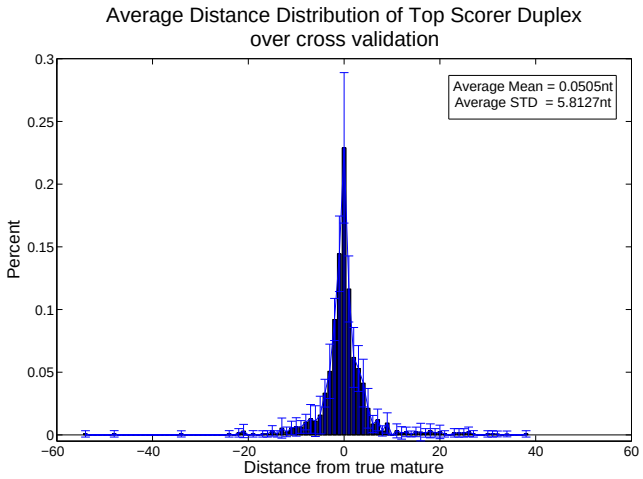


Distance from truth	Percent
0	16.07%
±1	43.37%
±2	64.84%
±3	75.48%
±4	81.51%
±5	86.68%
±6	89.34%
±7	91.06%

Figure 4.5: The average distance distribution over the 10–fold cross validation (left) and the percent for each distance away from the truth (right), when the computational truth is the *mean value* of the 4 top scorers using score threshold $\lambda = 1$.

opposite stem. For example, precursor *hsa-mir-140* produces only one mature miRNA in position 22 on 5'stem, but our model proposes two mature candidates one in position 23 on 5'stem and one in position 61 on 3'stem, if we use as computational truth the top scorer per stem. The second candidate appears because approximately half of our training data produce two mature miRNAs and the model learns to identify candidates in both stems. The first candidate in this specific example is only 1nt away from truth, while the second one is 39nt away. The second position based on the known information is a false positive, but it may be biologically significant if the two candidates correspond to the miRNA-miRNA* duplex (see section 2.1). Our next goal is thus to provide instead of two mature candidates, one double stranded candidate which is more likely to correspond to the miRNA–miRNA* duplex. Based on the observation that the miRNA-miRNA* duplex has approximately 2nt overhang in the 3' end, our model will first identify the top scoring mature over both stem and will then provide its miRNA* as the miRNA from the opposite stem which starts 2nt away from the matching position of the mature candidate's ending position.

Figure 4.6 shows the average distance distribution over the cross validation assuming as truth the top scorer of the precursor and its miRNA*. The distance is measured from the true mature, irrespectively of whether it corresponds to the predected miRNA or its miRNA* candidate. If the precursor produces two matures, both distances are calculated. The distribution of the miRNA–miRNA* duplex of the top scorer has average mean $0.0505nt$ and average standard deviation $5.8127nt$ over



Distance from truth	Percent
0	22.89%
±1	48.97%
±2	64.35%
±3	74.71%
±4	82.17%
±5	85.87%
±6	87.83%
±7	90.30%

Figure 4.6: The average distance distribution o (left) and the percent per distance away from the truth (right) of the top scorer miRNA–miRNA* duplex over the 10–fold cross validation.

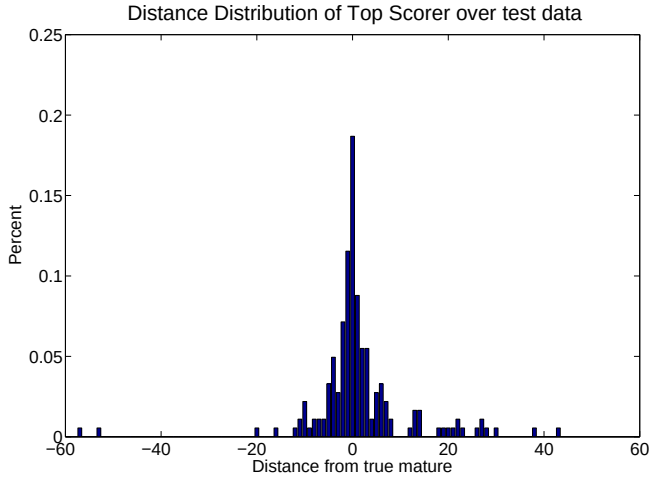
the cross validation. Moreover, the 87.83% of the candidates are $\pm 6nt$ away from the truth.

Overall, the best strategies for calculating the computational truth are the top scorer per stem if we have the prior knowledge which stem produces the mature miRNA, otherwise the top scorer per precursor and its duplex.

4.2.2 Evaluate best strategies in test dataset

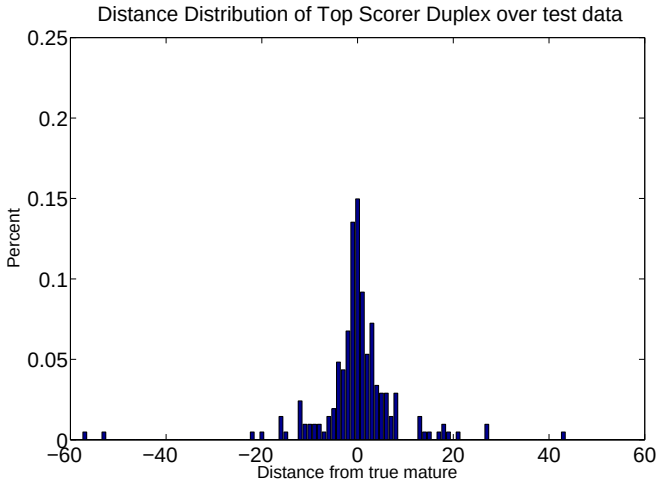
In this subsection we evaluate the performance of the best strategies of finding the computational truth over a blind test dataset (see section 3.3). As we methioned above the best strategies for calculating the computational truth are the top scorer per stem if we have the prior knowledge which stem produces the mature miRNA, otherwise the top scorer per precursor and its duplex. Figure 4.7 shows the distance distributions the two best strategies of calculating the computational truth as mentioned above. The 76.37% of the top scorers candidates per stem were found in $\pm 6nt$ away from the true mature miRNA, while the 78.74% of the top scorers with their duplexes as candidates lay within the same distance, over the blind dataset. Both strategies keep their percent of candidates for the distance of $\pm 6nt$ in high levels.

We also evaluate their performance of these strategies by splitting the test dataset into human and mouse set, in order to evaluate if there is a difference



Distance from truth	Percent
0	18.68%
±1	39.01%
±2	51.61%
±3	59.89%
±4	65.93%
±5	71.98%
±6	76.37%
±7	79.67%

(a) Distance distribution of top scorer per stem.



Distance from truth	Percent
0	14.98%
±1	37.68%
±2	49.76%
±3	61.35%
±4	69.57%
±5	74.40%
±6	78.74%
±7	80.68%

(b) Distance distribution of top scorer and its duplex.

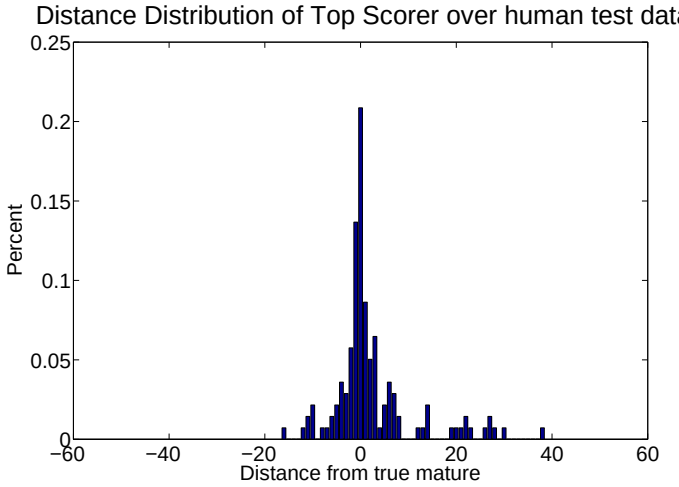
Figure 4.7: The distance distributions (left) and the percent per distance away from the truth (right) of the computational truth of our Best Naive Bayes classifier over the test dataset, for the most accurate strategies, top scorer per stem (see figure 4.7(a)), and top scorer per precursor and its duplex (see figure 4.7(b)).

that is drawn from the species. Figure 4.8 shows the distance distributions over the human and mouse test dataset, when the computational truth is the top scorer per stem. In both organisms the candidates that lay within $\pm 6nt$ away from the true mature miRNA have high percent, the 76.98% for the human dataset (see figure 4.8(a)) and the 74.42% for the mouse dataset (see figure 4.8(b)). Although the two distributions seems similar, the Kolmogorov-Smirnov Test shaw that the datasets come from different distributions (p-value ≈ 0.0352). On the other hand, figure 4.9 show the distance distributions over the human and mouse dataset, when the computational truth is the top scorer per precursor and its duplex. In both organisms the candidates that lay within $\pm 6nt$ away from the true mature miRNA have high percent, the 81.13% for the human dataset (see figure 4.9(a)) and the 70.83% for the mouse dataset (see figure 4.9(b)). The Kolmogorov–Smirnov Test in this cases shaw that the datasets come from the same distribution (p-value ≈ 0.3310). This evaluation shows that the top scorer per precursor with its duplex is more strong strategy for finding the optimal mature candidate within a miRNA precursor sequence.

4.3 Problem Complexity

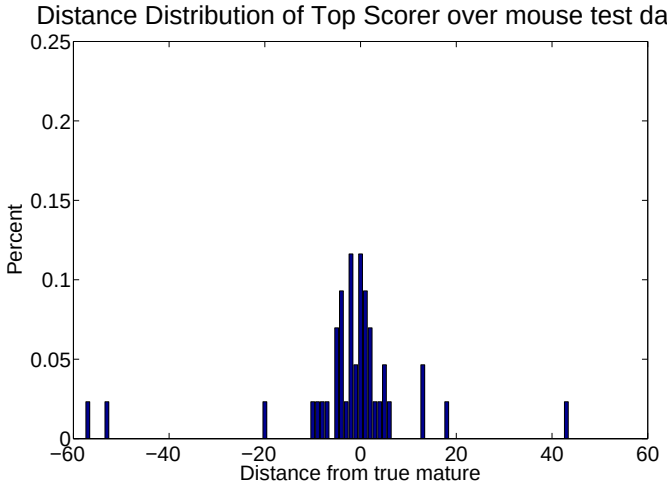
In order to evaluate the generalization of our best classifier we compare with the simplest classifier we trained, based on the distance distributions of the best two strategies of computational truth, the top scorer per stem and the top scorer per precursor with its duplex. Our best classifier uses 37 *position oriented* features and the distance of the starting position of the mature miRNA from the hairpin and achieves $AUC \approx 0.88$ over the 10–fold cross validation, while our simplest classifier uses one single feature, the distance of the starting position of the mature miRNA from the hairpin, also named as HS classifier, and achieves $AUC \approx 0.81$ over the 10–fold cross validation.

Figure 4.10 shows the average distance distributions of the HS classifier if the compuational truth is the top scorer per stem (see figure 4.10(a)) or the top scorer per precursor and its duplex (see figure 4.10(b)). If we consider as computational truth the top scorer per stem then the 47.47% of the computational truth were $\pm 6nt$ away from the true mature for HS classifier (see figure 4.10(a)), while 86.88% of the computational truth were within the same distance for out best classifier (see figure 4.3). We also evaluate the statistical difference between these two distributions using the Kolmogorov-Smirnov Test, which confirms that the two datasets come from different distributions (p-value ≈ 0.0000223). If we consider as computational truth



(a) Distance distribution of human dataset.

Distance from truth	Percent
0	20.86%
±1	43.17%
±2	53.96%
±3	63.31%
±4	67.31%
±5	71.94%
±6	76.98%
±7	80.58%



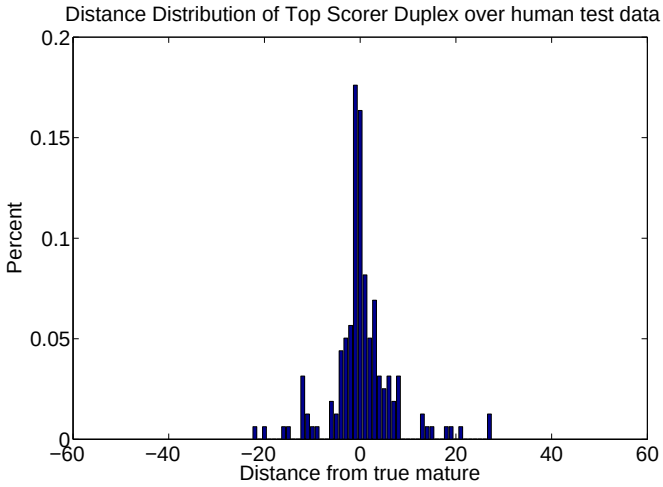
(b) Distance distribution of mouse dataset.

Distance from truth	Percent
0	11.63%
±1	25.58%
±2	44.19%
±3	48.84%
±4	60.47%
±5	72.09%
±6	74.42%
±7	76.74%

Figure 4.8: The distance distributions (left) and the percent per distance away from the truth (right) of the top scorer per stem over the test dataset as it is split into human and mouse datasets.

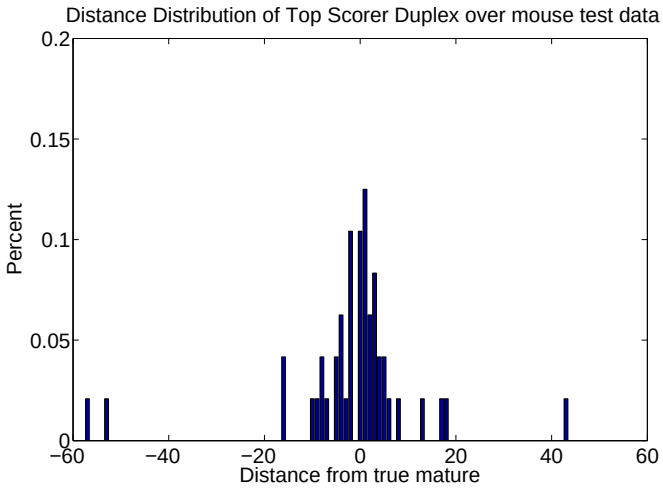
the top scorer per precursor and it duplex then the 69.53% of the computational truth were $\pm 6nt$ away from the true mature for HS classifier (see figure 4.10(b)), while the 87.83% of the computational truth were within the same distance for our best classifier (see figure 4.6). We also evaluate the statistical difference between these two distributions using the Kolmogorov-Smirnov Test, which confirms that the two datasets come from different distributions (p-value ≈ 0.0097).

These results indicate that the complexity of the problem cannot be solved using a single feature, such as the distance of the starting position of the mature miRNA



(a) Distance distribution of human dataset.

Distance from truth	Percent
0	16.35%
±1	42.14%
±2	52.83%
±3	64.78%
±4	72.33%
±5	76.10%
±6	81.13%
±7	83.02%



(b) Distance distribution of mouse dataset.

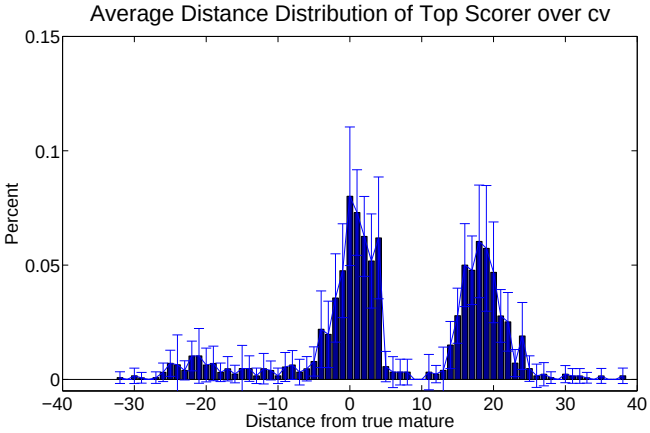
Distance from truth	Percent
0	10.42%
±1	22.92%
±2	39.58%
±3	50.00%
±4	60.42%
±5	68.75%
±6	70.83%
±7	72.92%

Figure 4.9: The distance distributions (left) and the percent per distance away from the truth (right) of the top scorer per precursor and its duplex over the test dataset as it is split into human and mouse datasets.

from the hairpin, even though it provides quite strong classification performance, requiring even more complex features for solving the problem.

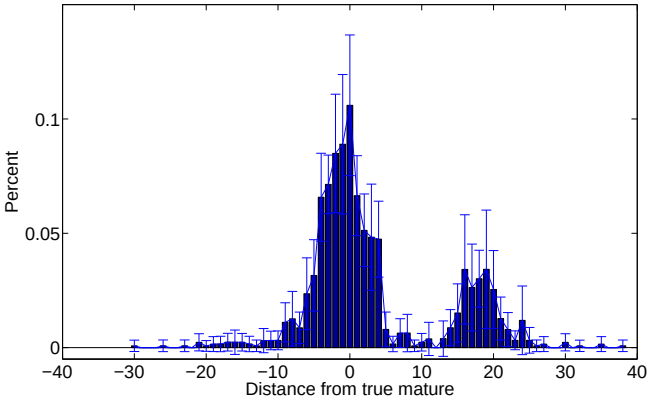
4.4 Comparison with other methods

In section 2.4 we presented a number of studies that use computational methods to identify the mature miRNA from a miRNA precursor. We were able to compare the performance of our model with just two of these studies, due to source code and



Distance from truth	Percent
0	8.01%
±1	20.07%
±2	29.89%
±3	37.04%
±4	45.43%
±5	46.78%
±6	47.57%
±7	48.21%

(a) Average distance distribution of Top Scorer in 10–fold cross validation.



Distance from truth	Percent
0	10.60%
±1	26.14%
±2	39.76%
±3	51.73%
±4	63.05%
±5	67.01%
±6	69.53%
±7	71.03%

(b) Average distance distribution of Top Scorer and its duplex in 10–fold cross validation.

Figure 4.10: The distance distributions (left) and the percent per distance away from the truth (right) of the computational truth of the HS Naive Bayes classifier over the 10–fold cross validation.

data unavailability for the rest of the methods. We used the 200 miRNA precursors in our blind test set as input to both tools and estimate performaces only on those precursors that were computationally predicted to contain a mature miRNA by each tool respectively. All precursors in our test set were contained in later versions of miRBase and were not used to train any of these tools, neither ours.

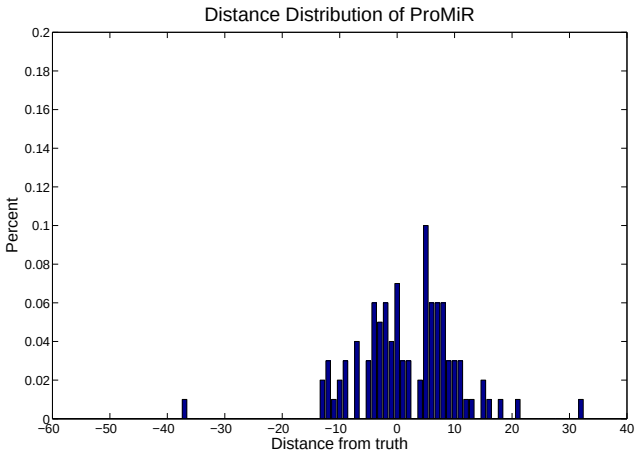
ProMiR

The first method is the *ProMiR* by Nam *et al.* [50], who proposed a method based on paired Hidden Markov Models (HMM) for miRNA precursor identification. The comparison with this tool was done using 178 precursors out of the 200 precursors of our test dataset(see paragraph "The evaluation dataset" in section 3.3), those

precursors that *ProMiR* identified as true precursors. *ProMiR* predicted the wrong stem for $78/178$ of these precursors, while our model predicted the wrong stem for $94/178$ precursors, if we consider as computational truth the top scorer of the Bayesian classifier. For the rest of the precursors, those that the computational truth was in the same stem as the true mature, we computed the distance distributions for both methods (see figure 4.11). As shown in figure 4.11(a) 55% of the computational truth were $\pm 6nt$ away from the true mature for *ProMiR* (see figure 4.11(a)), while 79.52% of our top scorers were within the same distance, respectively (see figure 4.11(b)). We also evaluate the statistical difference between the two distributions shown in figure 4.11 using the Kolmogorov-Smirnov Test, which confirms that the two datasets come from different distributions (p-value ≈ 0.00074).

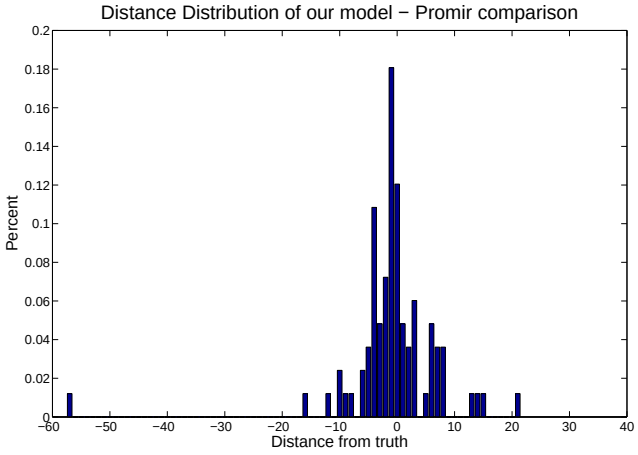
The second method is called *BayesMiRNAfind* by Yousef *et al.*[72], a web server which uses a Naive Bayes Classifier to predict miRNA precursors and incorporates mature miRNA prediction to increase its performance. The comparison with this tool was done using 101 precursors out of the 200 precursors of our test dataset(see paragraph "The evaluation dataset" in section 3.3), those precursors that *BayesMiRNAfind* predicted as true precursors. The *BayesMiRNAfind* predicted the wrong stem for $45/101$ precursors, while our model predicted the wrong stem for $53/101$ precursors, if we consider as the computational truth the top scorer of the Bayesian classifier. For the rest of the precursors, those that the computational truth was in the same stem as the true mature, we generated the distance distributions for both methods (see figure 4.12). As shown in figure 4.12 44.64% of the computational truth were $\pm 6nt$ away from the true mature for the *BayesMiRNAfind* (see figure 4.12(a)), while 85.42% of our top scorers were within the same distance (see figure 4.12(b)). We also evaluate the statistical difference between the two distributions shown in figure 4.12 using the Kolmogorov–Smirnov Test, which confirms that the two datasets come from different distributions (p-value ≈ 0.0013). BayesMiRNAfind

The confidence our model achieves for $\pm 6nt$ away from the truth is approximately double in comparison to the confidence achieved by *BayesMiRNAfind* for the same distance, while the confidence achieved by our model is $\sim 30\%$ more than the confidence achieved by *ProMiR* for the same distance. Overall, our model achieves higher confidence for the same distance from the truth than the confidences of the other methods on the independent test datasets.



Distance from truth	Percent
0	7%
±1	12%
±2	23%
±3	28%
±4	36%
±5	49%
±6	55%
±7	65%

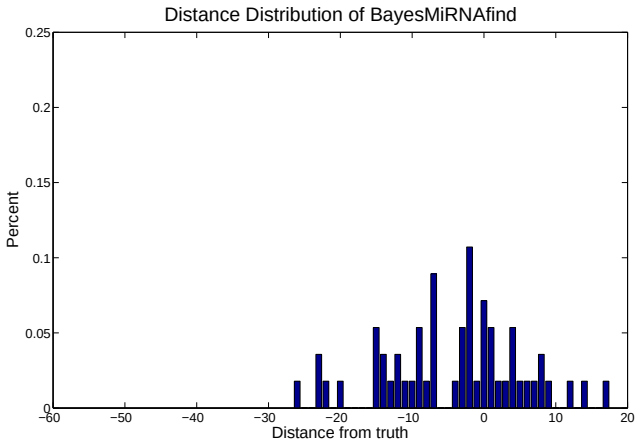
(a) Distance Distribution of ProMiR.



Distance from truth	Percent
0	12.05%
±1	34.94%
±2	45.78%
±3	56.63%
±4	67.47%
±5	72.29%
±6	79.52%
±7	83.13%

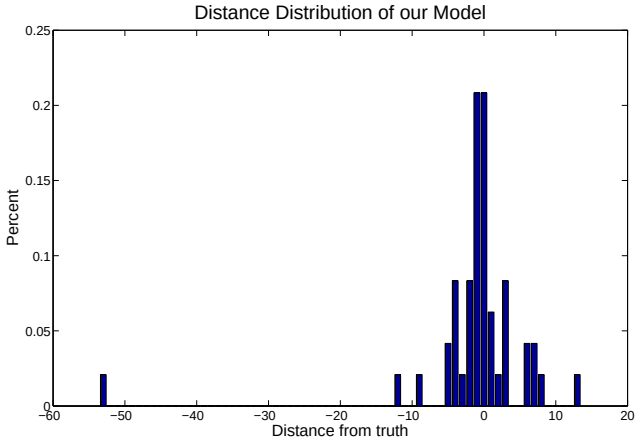
(b) Distance Distribution of our model.

Figure 4.11: The distance distributions (left) and the percent per distance away from the truth (right) of both ProMiR and our model from the predictions that were within the same stem as the true mature. In our model we consider as computational truth the top scorer of the Bayesian model.



Distance from truth	Percent
0	7.14%
±1	14.29%
±2	26.79%
±3	33.93%
±4	41.07%
±5	42.86%
±6	44.64%
±7	55.36%

(a) Distance Distribution of BayesMiRNAfind.



Distance from truth	Percent
0	20.83%
±1	47.92%
±2	58.33%
±3	68.75%
±4	77.08%
±5	81.25%
±6	85.42%
±7	89.58%

(b) Distance Distribution of our model.

Figure 4.12: The distance distributions (left) and the percent per distance away from the truth (right) of both BayesMiRNAfind and our model from the predictions that were within the same stem as the true mature. In our model we consider as computational truth the top scorer of the Bayesian model.

Chapter 5

Conclusion

5.1 Discussion

In this thesis we examined the problem of mature miRNA prediction within mammalian miRNA precursors. We proposed a Naive Bayes classifier (NBC) that uses sequence and structure characteristics of the miRNA precursor in order to provide the position that is most likely to represent the start of each mature miRNAs that can be produced by the precursor. We select the NBC because it requires a relatively small amount of training data to estimate its parameters, it provides a direct intuition about the importance of the features used and it has high performance in many complex real–world problems, despite of its simplified assumptions.

The biological features used in the NBC are a number of *position oriented* features, containing both sequential and structural information of the specific position on the miRNA precursor, and the distance of the starting position of the mature miRNA from the hairpin. We selected to use *position oriented* features in order to examine the hypothesis that *Dicer* recognizes a common pattern which appears on the edges of the mature miRNAs. This hypothesis is confirmed, since the *position oriented* features we incorporate into our model tend to lay either in the flanking regions around the mature miRNA or in positions within the mature, but which are close to its ends. The distance of the starting position of the mature miRNA showed that matures tends to be close to the hairpin, suggesting that they are probably found in positions that do not depend on the actual size of the precursor.

We used experimentally verified human and mouse miRNAs to train and evaluate the performance of a Naive Bayes classifier in terms of AUC and distance from the truth. Unlike the method presented here, most of the computational tools that can be used to predict the functional part of the miRNA precursor estimate their

performance accuracy in terms of true positive rate alone, ignoring the false positive rate ([50], [61], [64]). It is a matter of semantics as well as a great challenge to define a true negative example when it comes to mature miRNAs. However, a major issue in such a classification task is not only to maximize the identification of true positives but also to minimize the false positive rate. In an effort to combine both of these criteria, our method achieves an average $AUC \approx 0.88$ and $\sim 88\%$ of the top scorer duplexes, on average, were $\pm 6nt$ away from the truth.

In conclusion, our findings suggest that position specific sequence and structure information and the distance of the starting position from the hairpin combined with a simple Bayes classifier achieve a good performance on the challenging task of mature miRNA identification.

5.2 Future Work

There are a number of open issues regarding the mature miRNA identification problem. First of all, as a typical pattern recognition problem there are a number of parameters that we didn't examine in this thesis. For example, one might consider different error cost per class with the Naive Bayes classifier. Apart from the NBC one could also use a stronger classifier such as support vector machines (SVM) or artificial neural networks. With these classifiers it is easy to include both different error costs per class and different weights per feature, which could provide more accurate results.

On the other hand, one could also use as training input the miRNA–miRNA* duplex instead of the mature alone. In other words, one will convert the mature miRNA identification problem to the miRNA–miRNA* duplex identification, which could be what *Dicer* recognizes after all. The only problem with this approach is the need of a more accurate definition of the miRNA–miRNA* duplex by biologists, in order to get more precise results.

Appendix A

Supplementary Data

Table A.1: The features sorted based on Kullback–Leibler divergence.

Feature Description	**Kullback–Leibler score**
Distance of Starting Position from Hairpin (3'stem)	14.20
Distance of Ending Position from Hairpin (3'stem)	10.98
Distance of Starting Position from Hairpin (5'stem)	9.20
Distance of Ending Position from Hairpin (5'stem)	9.12
Distance of Ending Position from End (3'stem)	2.50
Distance of Starting Position from End (3'stem)	2.50
Distance of Ending Position from End (5'stem)	2.48
Distance of Starting Position from End (5'stem)	2.48
Position 8 in flanking region before mature	0.2126
Position 9 in flanking region before mature	0.2026
Position 7 in flanking region before mature	0.1725
Position 7 in flanking region after mature	0.1715
Position 16 in mature	0.1707
Position 8 in flanking region after mature	0.1581
Position 0 in mature	0.1549
Position 7 in mature	0.1420
Position 9 in flanking region after mature	0.1358
Position 6 in flanking region after mature	0.1312
Position 15 in mature	0.1260
	Continued on next page

Table A.1 – continued from previous page

Feature Description	**Kullback–Leibler score**
Position 13 in mature	0.1220
Position 17 in mature	0.1181
Position 3 in mature	0.1156
Position 6 in flanking region before mature	0.1129
Position 18 in mature	0.1129
Position 6 in mature	0.1067
Position 5 in flanking region after mature	0.1008
Position 12 in mature	0.1004
Position 14 in mature	0.0931
Position 4 in mature	0.0900
Position 3 in flanking region before mature	0.0830
Position 5 in mature	0.0805
Position 4 in flanking region before mature	0.0793
Position 2 in mature	0.0749
Position 11 in mature	0.0733
Position 8 in mature	0.0729
Position 20 in mature	0.0728
Position 4 in flanking region after mature	0.0690
Position 1 in mature	0.0687
Position 5 in flanking region before mature	0.0613
Position 1 in flanking region before mature	0.0554
Position 1 in flanking region after mature	0.0516
Position 23 in mature	0.0495
Position 25 in mature	0.0476
Position 2 in flanking region before mature	0.0458
Position 9 in mature	0.0450
Position 22 in mature	0.0431
Position 2 in flanking region after mature	0.0427
Position 21 in mature	0.0407
Position 19 in mature	0.0396
Position 10 in mature	0.0351
Position 24 in mature	0.0299

Continued on next page

Table A.1 – continued from previous page

Feature Description	**Kullback–Leibler score**
Position 3 in flanking region after mature	0.0286

Table A.2: The AUC of the average ROC curve, over the 10–fold cross validation procedure, for all naive bayes classifiers trained with flanking region 0nt.

Number of *Position Oriented* Features	**Window 18**	**Window 20**	**Window 22**	**Window 24**
1	0.8397	0.8374	0.8366	0.8371
2	0.8455	0.8428	0.842	0.8433
3	0.8594	0.8576	0.8574	0.8577
4	0.8578	0.856	0.8556	0.8566
5	0.8578	0.8556	0.8553	0.8567
6	0.8572	0.8556	0.8551	0.8562
7	0.8589	0.8573	0.857	0.857
8	0.8608	0.8595	0.8601	0.8601
9	0.8606	0.8584	0.8587	0.8591
10	0.8589	0.8587	0.8582	0.8587
11	0.859	0.8567	0.8572	0.8578
12	0.8582	0.8571	0.8575	0.858
13	0.858	0.8574	0.8572	0.8578
14	0.8593	0.8565	0.8577	0.8581
15	0.8599	0.8581	0.8568	0.8571
16	0.8628	0.8582	0.8586	0.8589
17	**0.8629**	**0.8615**	0.8586	0.8598
18	0.8629	0.8615	**0.8621**	**0.8624**
19	-	0.8604	0.8615	0.8618
20	-	0.8605	0.8514	0.8617
21	-	-	0.8603	0.8606

Continued on next page

Table A.2 – continued from previous page

Number of *Position Oriented* Features	**Window 18**	**Window 20**	**Window 22**	**Window 24**
22	-	-	0.8601	0.8595
23	-	-	-	0.8355
24	-	-	-	0.8337

Table A.3: The AUC of the average ROC curve, over the 10–fold cross validation procedure, for all naive bayes classifiers trained with flanking region 3nt.

Number of *Position Oriented* Features	**Window 18**	**Window 20**	**Window 22**	**Window 24**
1	0.8413	0.839	0.8379	0.8386
2	0.8572	0.8557	0.8549	0.8556
3	0.8588	0.8572	0.8571	0.8577
4	0.8578	0.8563	0.8561	0.8571
5	0.8567	0.8554	0.8551	0.8565
6	0.857	0.8554	0.855	0.8561
7	0.8578	0.8536	0.854	0.8542
8	0.8576	0.8554	0.8553	0.8557
9	0.8603	0.858	0.8581	0.8585
10	0.8586	0.8575	0.8576	0.8581
11	0.8585	0.8563	0.8566	0.8572
12	0.8616	0.8564	0.8568	0.8572
13	0.8628	0.8602	0.8596	0.8599
14	0.8623	0.8605	0.8608	0.861
15	0.8655	0.8611	0.8611	0.8614
16	0.8647	0.8634	0.8617	0.8618

Continued on next page

Table A.3 – continued from previous page

Number of *Position Oriented* Features	**Window 18**	**Window 20**	**Window 22**	**Window 24**
17	0.8648	0.8629	0.8643	0.8646
18	0.8663	0.8642	0.8637	0.8648
19	**0.8671**	0.8647	0.8652	0.8654
20	0.8669	**0.8658**	0.8656	0.8657
21	0.8669	0.8642	0.8668	0.867
22	0.8657	0.8653	0.8665	0.8673
23	0.8648	0.8635	**0.8675**	**0.8661**
24	0.864	0.8619	0.8666	0.867
25	-	0.8619	0.8671	0.8663
26	-	0.8606	0.8663	0.8647
27	-	-	0.8654	0.8639
28	-	-	0.8655	0.863
29	-	-	-	0.8635
30	-	-	-	0.8633

Table A.4: The AUC of the average ROC curve, over the 10–fold cross validation procedure, for all naive bayes classifiers trained with flanking region 5nt.

Number of *Position Oriented* Features	**Window 18**	**Window 20**	**Window 22**	**Window 24**
1	0.8398	0.8375	0.8364	0.8370
2	0.8565	0.8547	0.8542	0.8549
3	0.8585	0.8568	0.8562	0.8573
4	0.8570	0.8554	0.8551	0.8560
5	0.8563	0.8550	0.8542	0.8554

Continued on next page

Table A.4 – continued from previous page

Number of *Position Oriented* Features	**Window 18**	**Window 20**	**Window 22**	**Window 24**
6	0.8558	0.8542	0.8538	0.8548
7	0.8574	0.8559	0.8555	0.8554
8	0.8595	0.8543	0.8543	0.8546
9	0.8590	0.8570	0.8572	0.8576
10	0.8571	0.8570	0.8564	0.8569
11	0.8503	0.8549	0.8552	0.8558
12	0.8505	0.8524	0.8564	0.8545
13	0.8508	0.8526	0.8558	0.8546
14	0.8546	0.8537	0.8567	0.8548
15	0.8566	0.8568	0.8596	0.8588
16	0.8555	0.8588	0.8613	0.8605
17	0.8584	0.8577	0.8613	0.8596
18	0.8595	0.8615	0.8639	0.8626
19	**0.8597**	0.8614	0.8639	0.8633
20	0.8544	**0.8614**	0.8642	0.8639
21	0.8546	0.8576	0.8652	**0.8642**
22	0.8556	0.8581	0.8628	0.8616
23	0.8561	0.8600	0.8635	0.8623
24	0.8563	0.8585	0.8644	0.8640
25	0.8563	0.8588	0.8650	0.8625
26	0.8554	0.8597	0.8651	0.8634
27	0.8547	0.8571	**0.8662**	0.8637
28	0.8532	0.8573	0.8661	0.8640
29	-	0.8561	0.8658	0.8628
30	-	0.8551	0.866	0.8628
31	-	-	0.8649	0.8636
32	-	-	0.8637	0.8630
33	-	-	-	0.8617
34	-	-	-	0.8614

Table A.5: The AUC of the average ROC curve, over the 10–fold cross validation procedure, for all naive bayes classifiers trained with flanking region 7nt.

Number of *Position Oriented* Features	**Window 18**	**Window 20**	**Window 22**	**Window 24**
1	0.8250	0.8237	0.8279	0.8276
2	0.8335	0.8320	0.8357	0.8357
3	0.8443	0.8423	0.8458	0.8458
4	0.8568	0.8555	0.8588	0.8586
5	**0.8592**	0.8574	0.8607	0.8601
6	0.8495	0.8509	0.8563	0.8558
7	0.8523	0.8542	0.8588	0.8593
8	0.8519	0.8544	0.8592	0.8584
9	0.8538	0.8547	0.8602	0.8594
10	0.8571	0.8560	0.8619	0.8607
11	0.8583	0.8591	0.8651	0.8641
12	0.8583	0.8603	0.8652	0.8650
13	0.8583	0.8604	0.8648	0.8635
14	0.8584	0.8598	0.8640	0.8639
15	0.8514	0.8597	0.8649	0.8638
16	0.8501	0.8570	0.8647	0.8621
17	0.8519	0.8568	0.8639	0.8613
18	0.8535	0.8579	0.8658	0.8637
19	0.8529	0.8591	0.8661	0.8649
20	0.8535	0.8583	0.8676	0.8654
21	0.8547	0.8590	0.8672	0.8650
22	0.8556	0.8630	0.8680	0.8668
23	0.8560	0.8604	0.8680	0.8668
24	0.8534	**0.8596**	0.8694	0.8674
25	0.8524	0.8604	0.8713	**0.8696**
26	0.8535	0.8596	0.8690	0.8682

Continued on next page

Table A.5 – continued from previous page

Number of *Position Oriented* Features	**Window 18**	**Window 20**	**Window 22**	**Window 24**
27	0.8538	0.8507	0.8685	0.8678
28	0.8539	0.8609	0.8695	0.8686
29	0.8542	0.8611	0.8702	0.8684
30	0.8537	0.8614	0.8704	0.8687
31	0.8531	0.8604	0.8706	0.8691
32	0.8512	0.8600	**0.8716**	0.8692
33	-	0.8591	0.7150	0.8686
34	-	0.8578	0.8709	0.8686
35	-	-	0.8708	0.8684
36	-	-	0.8703	0.8680
37	-	-	-	0.8678
38	-	-	-	0.8679

Table A.6: The AUC of the average ROC curve, over the 10–fold cross validation procedure, for all naive bayes classifiers trained with flanking region 9nt.

Number of *Position Oriented* Features	**Window 18**	**Window 20**	**Window 22**	**Window 24**
1	0.8445	0.8423	0.8418	0.8420
2	0.8455	0.8534	0.8430	0.8431
3	0.8434	0.8408	0.8405	0.8392
4	0.8439	0.8432	0.8465	0.8443
5	0.8471	0.8468	0.8500	0.8482
6	0.8454	0.8469	0.8518	0.8456
7	0.8551	0.8570	0.8512	0.8557

Continued on next page

Table A.6 – continued from previous page

Number of *Position Oriented* Features	**Window 18**	**Window 20**	**Window 22**	**Window 24**
8	0.8577	0.8598	0.8631	0.8584
9	0.8557	0.8586	0.8618	0.8517
10	0.8511	0.8561	0.8601	0.8524
11	0.8520	0.8569	0.8608	0.8529
12	0.8529	0.8588	0.8627	0.8543
13	0.8540	0.8591	0.8636	0.8548
14	0.8567	0.8616	0.8661	0.8577
15	0.8578	0.8640	0.8686	0.8601
16	**0.8599**	0.8634	0.8676	0.8593
17	0.8540	0.8644	0.8693	0.8612
18	0.8548	0.8631	0.8392	0.8607
19	0.8540	0.8629	0.8691	0.8614
20	0.8544	0.8632	0.8694	0.8610
21	0.8558	0.8636	0.8696	0.8614
22	0.8555	0.8640	0.8703	0.8632
23	0.8562	0.8637	0.8711	0.8631
24	0.8568	0.8648	0.8715	0.8637
25	0.8577	0.8654	0.8721	0.8654
26	0.8583	0.8662	0.8727	0.8661
27	0.8549	0.8666	0.8728	0.8663
28	0.8567	0.8651	0.8744	0.8669
29	0.8555	0.8667	0.8732	0.8672
30	0.8565	0.8658	0.8747	0.8691
31	0.8561	0.8668	0.8741	0.8685
32	0.8571	0.8670	0.8758	0.8695
33	0.8572	0.8671	0.8759	0.8686
34	0.8562	**0.8673**	0.8759	0.8702
35	0.8558	0.8662	0.8760	**0.8704**
36	0.8550	0.8655	0.8762	0.8703

Continued on next page

Table A.7: The average mean and standard deviation of the average distributions when the computational truth is the middle point defined by the range of n top scorers for $n \in \{2, \ldots, 6\}$ over the 10–fold cross validation.

Number of Top scorers	**Average mean**	**Average STD**
2	0.1261	6.0287
3	0.3589	5.6030
4	**0.3694**	**5.5208**
5	0.5648	5.5382
6	0.6015	5.5651

Table A.8: The average mean and standard deviation of the average distributions when the computational truth is the mean value of n top scorers for $n \in \{2, \ldots, 6\}$ over the 10–fold cross validation.

Number of Top scorers	**Average mean**	**Average STD**
2	0.7485	6.0201
3	0.5838	5.6456
4	**0.8299**	**5.4579**
5	0.8012	5.4983
6	0.9364	5.4741

Table A.6 – continued from previous page

Number of *Position Oriented* Features	**Window 18**	**Window 20**	**Window 22**	**Window 24**
37	-	0.8651	**0.8771**	0.8690
38	-	0.8641	0.8762	0.8698
39	-	-	0.8757	0.8690
40	-	-	0.8753	0.8685
41	-	-	-	0.8690
42	-	-	-	0.8686

Bibliography

[1] J. E. Abrahante, A. L. Daul, M. Li, M. L. Volk, J. M. Tennessen, E. A. Miller, and A. E. Rougvie. The Caenorhabditis elegans hunchback-like gene lin-57/hbl-1 controls developmental time and is regulated by microRNAs. *Dev Cell*, 4:625–637, 2003.

[2] Victor Ambros, Bonnie Bartel, David P. Bartel, Christopher B. Burge, James C.Carrungton, Xuemei Chen, Gideon Dreyfuss, Sean R. Eddy, Sam Griffiths-Jones, Mhairi Marshall, Marjori Matke, Gary Ruvkun, and Thomas Tuschl. RNA A uniform system for microRNA annotation. *RNA*, 9:277–279, 2003.

[3] M. J. Aukerman and H. Sakai. Regulation of flowering time and floral organ identity by a MicroRNA and its APETALA2-like target genes. *Plant Cell*, 15:2730–2741, 2003.

[4] Pierre Baldi, Søren Brunak, Yves Chauvin, Claus A. F. Andersen, and Henrik Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 18(5):412–424, 2000.

[5] E. Berezikov, E. Cuppen, and R. H. Plasterk. Approaches to microRNA discovery. *Nat Genet*, 38 Suppl 1, June 2006.

[6] Emily Bernstein, Amy A. Caudy, Scott M. Hammond, and Gregory J. Hannon. Role for a bidentate ribonuclease in the initiation step of rna interference. *Nature*, 409(6818):363–366, January 2001.

[7] J. Brennecke, D. R. Hipfner, A. Stark, R. B. Russell, and S. M. Cohen. bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. *Cell*, 113:25–36, 2003.

[8] X. Cai, C. H. Hagedorn, and B. R. Cullen. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA*, 10(12):1957–1966, December 2004.

[9] J. A. Chan, A. M. Krichevsky, and K. S. Kosik. MicroRNA-21 is an antiapoptotic factor in human glioblastoma cells. *Cancer Res*, 65:6029–6033, 2005.

[10] J. F. Chen, E. M. Mandel, J. M. Thomson, Q. Wu, T. E. Callis, S. M. Hammond, F. L. Conlon, and D. Z. Wang. The role of microRNA-1 and microRNA-133 in skeletal muscle proliferation and differentiation. *Nat Genet*, 38:228–233, 2006.

[11] Xuemei Chen. A MicroRNA as a Translational Repressor of APETALA2 in Arabidopsis Flower Development. *Science*, 303(5666):2022–2025, 2004.

[12] Chia-Ying Y. Chu and Tariq M M. Rana. Translation Repression in Human Cells by Microrna-Induced Gene Silencing Requires RCK/p54. *PLoS Biol*, 4(7), June 2006.

[13] A. Cimmino, G. A. Calin, M. Fabbri, M. V. Iorio, M. Ferracin, M. Shimizu, S. E. Wojcik, R. I. Aqeilan, S. Zupo, M. Dono, L. Rassenti, H. Alder, S. Volinia, C. G. Liu, T. J. Kipps, M. Negrini, and C. M. Croce. miR-15 and miR-16 induce apoptosis by targeting BCL2. *Proc Natl Acad Sci USA*, 102:13944–13949, 2005.

[14] Girish Deshpande, Gretchen Calhoun, and Paul Schedl. Drosophila argonaute-2 is required early in embryogenesis for the assembly of centric/centromeric heterochromatin, nuclear division, nuclear migration, and germ-cell formation. *Genes Dev.*, 19(14):1680–1685, 2005.

[15] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, November 2000.

[16] C. Esau, X. Kang, E. Peralta, E. Hanson, E. G. Marcusson, L. V. Ravichandran, Y. Sun, S. Koo, R. J. Perera, R. Jain, N. M. Dean, S. M. Freier, F. Bennett, B.Lollo, and R. Griffey. MicroRNA-143 regulates adipocyte differentiation. *J Biol Chem*, 279:52361–52365, 2004.

[17] N. Fahlgren, T. A. Montgomery, M. D. Howell, E. Allen, S. K. Dvorak, A. L. Alexander, and J. C. Carrington. Regulation of AUXIN RESPONSE FACTOR3 by TAS3 ta-siRNA affects developmental timing and patterning in Arabidopsis. *Curr Biol*, 16:939–944, 2006.

[18] Tom Fawcett. An introduction to roc analysis. *Pattern Recogn. Lett.*, 27(8):861–874, 2006.

[19] M. E. Gleave and B. P. Monia. Antisense therapy for cancer. *Nat Rev Cancer*, 5:468–479, 2005.

[20] S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*, 34(Database issue), January.

[21] Sam Griffiths-Jones, Harpreet Kaur K. Saini, Stijn van V. Dongen, and Anton J J. Enright. miRBase: tools for microRNA genomics. *Nucleic Acids Res*, November 2007.

[22] A. Gupta, J. J. Gartner, P. Sethupathy, A. G. Hatzigeorgiou, , and N. W. Fraser. Anti-apoptotic function of a microRNA encoded by the HSV-1 latency-associated transcript. *Nature*, 2006.

[23] Carole Gwizdek, Batool Ossareh-Nazari, Brownawell, Amy M., Stefan Evers, Ian G. Macara, and Catherine Dargemont. Minihelix-containing RNAs Mediate Exportin-5-dependent Nuclear Export of the Double-stranded RNA-binding Protein ILF3. *J. Biol. Chem.*, 279(2):884–891, 2004.

[24] A. J. Hamilton and D. C. Baulcombe. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science*, 286:950–252, 1999.

[25] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, August 2001.

[26] S. D. Hatfield, H. R. Shcherbata, K. A. Fischer, K. Nakahara, R. W. Carthew, and H. Ruohola-Baker. Stem cell division is regulated by the microRNA pathway. *Nature*, 435:974–978, 2005.

[27] L. He, J. M. Thomson, M. T. Hemann, E. Hernando-Monge, D. Mu, S. Goodson, S. Powers, C. Cordon-Cardo, S. W. Lowe, G. J. Hannon, and S. M. Hammond. A microRNA polycistron as a potential human oncogene. *Nature*, 435:828–833, 2005.

[28] M. V. Iorio, M. Ferracin, C. G. Liu, A. Veronese, R.Spizzo, S. Sabbioni, E. Magri, M. Pedriali, M. Fabbri, M.Campiglio, S. Menard, J. P. Palazzo, A. Rosenberg, P. Musiani, S. Volinia, I. Nenci, G. A. Calin, P. Querzoli, M. Negrini, and C. M.

Croce. MicroRNA gene expression deregulation in human breast cancer. *Cancer Res*, 65:7065–7070, 2005.

[29] M. Izquierdo. Short interfering RNAs as a tool for cancer gene therapy. *Cancer Gene Ther*, 12:217–227, 2005.

[30] H. Jeffreys. An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

[31] S. M. Johnson, H. Grosshans, J. Shingara, M. Byrom, R. Jarvis, A. Cheng, E. Labourier, K. L. Reinert, D. Brown, and F. J. Slack. RAS is regulated by the let-7 microRNA family. *Cell*, 120:635–647, 2005.

[32] Griffiths S. Jones. The microRNA Registry. *Nucleic Acids Res*, 32(Database issue):D109–D111, Jan 2004.

[33] V. Narry Kim. MicroRNA precursors in motion: exportin-5 mediates their nuclear export. *Trends in Cell Biology*, 14(4):156–159, 2004.

[34] 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, 1995.

[35] Yong Kong and Jin-Hua Han. MicroRNA: Biological and Computational Perspective. *Genomics Proteomics Bioinformatics*, 3(2):62–72, 2005.

[36] J. Krutzfeldt, N. Rajewsky, R. Braich, K. G. Rajeev, T. Tuschl, M. Manoharan, and M. Stoffel. Silencing of microRNAs in vivo with 'antagomirs'. *Nature*, 438:685–689, 2005.

[37] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[38] M. Lagos-Quintana, R. Rauhut, A. Yalcin, J. Meyer, W. Lendeckel, and T. Tuschl. Identification of tissue-specific microRNAs from mouse. *Curr Biol*, 12:735–739, 2002.

[39] Markus Landthaler, Abdullah Yalcin, and Thomas Tuschl. The Human DiGeorge Syndrome Critical Region Gene 8 and Its D. melanogaster Homolog Are Required for miRNA Biogenesis. *Current Biology*, 14(23):2162–2167, 2004.

[40] C. H. Lecellier, P. Dunoyer, K. Arar, J. Lehmann-Che, S. Eyquem, C. Himber, A. Saib, and O. Voinnet. A cellular microRNA mediates antiviral defense in human cells. *Science*, 308:557–560, 2005.

[41] R. C. Lee, R. L. Feinbaum, and V. Ambros. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854, December 1993.

[42] Y . Lee, K . Nakahara, J . Pham, K . Kim, Z . He, and E . Sontheim. Distinct Roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA Silencing Pathways. *Cell*, 117(1):69–81.

[43] Y. Lee, M. Kim, J. Han, K. H. Yeom, S. Lee, S. H. Baek, and V. N. Kim. MicroRNA genes are transcribed by RNA polymerase II. *EMBO J*, 23(20):4051–4060, October 2004.

[44] Cesar Llave, Zhixin Xie, Kristin D. Kasschau, and James C. Carrington. Cleavage of Scarecrow-like mRNA Targets Directed by a Class of Arabidopsis miRNA. *Science*, 297(5589):2053–2056, 2002.

[45] S. Lu and B. R. Cullen. Adenovirus VA1 noncoding RNA can inhibit small interfering RNA and MicroRNA biogenesis. *J Virol*, 78:12868–12876, 2004.

[46] M. Z. Michael, O. C. SM, N. G. van Holst Pellekaan, G. P. Young, and R. J. James. Reduced accumulation of specific microRNAs in colorectal neoplasia. *Mol Cancer*, 1:882–891, 2003.

[47] Tom M. Mitchell. *Machine Learning*. McGraw-Hill International, 1997.

[48] E. G. Moss, R. C. Lee, and V. Ambros. The cold shock domain protein LIN-28 controls developmental timing in C. elegans and is regulated by the lin-4 RNA. *Cell*, 88:637–646, 1997.

[49] P. Mourrain, C. Beclin, T. Elmayan, F. Feuerbach, C. Godon, J. B. Morel, D. Jouette, A. M. Lacombe, S. Nikic, N. Picault, K. Remoue, M. Sanial, T. A. Vo, and H. Vaucheret. Arabidopsis SGS2 and SGS3 genes are required for posttranscriptional gene silencing and natural virus resistance. *Cell*, 101:533–542, 2000.

[50] 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 Research*, 33(11):3570–3581, 2005.

[51] K. A. O'Donnell, E. A. Wentzel, K. I. Zeller, C. V. Dang, and J. T. Mendell. c-Myc-regulated microRNAs modulate E2F1 expression. *Nature*, 435:839–843, 2005.

[52] Vienna RNA package. http://www.tbi.univie.ac.at/ ivo/RNA.

[53] Istvan Papp, M. Florian Mette, Werner Aufsatz, Lucia Daxinger, Stephen E. Schauer, Animesh Ray, Johannes Van Der Winden, Marjori Matzke, and Antonius J.M. Matzke. Evidence for Nuclear Processing of Plant Micro RNA and Short Interfering RNA Precursors. *Plant Physiol.*, 132(3):1382–1390, 2003.

[54] David P.Bartel. MicroRNAs: Genomics,Biogenesis, Mechanism, and Function. *Cell*, 116:281–297, 2004.

[55] M. N. Poy, L. Eliasson, J. Krutzfeldt, S. Kuwajima, X. Ma, P. E. Macdonald, S. Pfeffer, T. Tuschl, N. Rajewsky, P. Rorsman, and M. Stoffel. A pancreatic islet-specific microRNA regulates insulin secretion. *Nature*, 432:226–230, 2004.

[56] Brenda J. Reinhart, Frank J. Slack, Michael Basson, Amy E. Pasquinelli, Jill C. Bettinger, Ann E. Rougvie, Robert H. Horvitz, and Gary Ruvkun. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, 403(6772):901–906, February 2000.

[57] Laura Schramm and Nouria Hernandez. Recruitment of RNA polymerase III to its target promoters. *Genes Dev.*, 16(20):2593–2620, 2002.

[58] G. M. Schratt, F. Tuebing, E. A. Nigh, C. G. Kane, M. E. Sabatini, M. Kiebler, and M. E. Greenberg. A brain-specific microRNA regulates dendritic spine development. *Nature*, 439:283–289, 2006.

[59] B. R. Schulman, A. Esquela-Kerscher, and F. J. Slack. Reciprocal expression of lin-41 and the microRNAs let-7 and mir-125 during mouse embryogenesis. *Dev Dyn*, 234:1046–1054, 2005.

[60] L. F. Sempere, N. S. Sokol, E. B. Dubrovsky, E. M. Berger, and V. Ambros. Temporal regulation of microRNA expression in Drosophila melanogaster mediated

by hormonal signals and broad-Complex gene activity. *Dev Biol*, 259:9–18, 2003.

[61] Ying Sheng, Par G. Engstrom, and Boris Lenhard. Mammalian MicroRNA Prediction through a Support Vector Machine Model of Sequence and Structure. *PloS ONE*, 2(9), 2007.

[62] C. Simon-Mateo and J. A. Garcia. MicroRNA-guided processing impairs Plum pox virus replication, but the virus readily evolves to escape this silencing mechanism. *J. Virol*, 80:2429–2436, 2006.

[63] F. J. Slack, M. Basson, Z. Liu, V. Ambros, H. R. Horvit, and G. Ruvkun. The lin-41 RBCC gene acts in the C. elegans heterochronic pathway between the let-7 regulatory RNA and the LIN-29 transcription factor. *Molecular Cell*, 5:659–669, 2000.

[64] Molei Tao. Thermodynamic and structural consensus principle predicts mature miRNA location and structure, categorizes conserved interspecies miRNA subgroups and hints new possible mechanisms of miRNA maturization. Technical report, Control and Dynamical Systems, California Institute of Technology, 2007.

[65] Shobha Vasudevan, Yingchun Tong, and Joan A. Steitz. Switching from Repression to Activation: MicroRNAs Can Up-Regulate Translation. *Science*, pages 1149460+, November 2007.

[66] N. Vo, M. E. Klein, O. Varlamova, D. M. Keller, T. Yamamoto, R. H. Goodman, and S. Impey. A cAMP-response element binding protein-induced microRNA regulates neuronal morphogenesis. *Proc Natl Acad Sci USA*, 102:16426–16431, 2005.

[67] J. A. Wilson and C. D. Richardson. Hepatitis C virus replicons escape RNA interference induced by a short interfering RNA directed against the NS5b coding region. *J Virol*, 79:7050–7058, 2005.

[68] L. Wu and J. G. Belasco. Micro-RNA regulation of the mammalian lin-28 gene during neuronal differentiation of embryonal carcinoma cells. *Mol Cell Biol*, 25:9198–9208, 2005.

[69] P. Xu, S. Y. Vernooy, M. Guo, , and B. A. Hay. The Drosophila microRNA Mir-14 suppresses cell death and is required for normal fat metabolism. *Curr Biol*, 13:790–795, 2003.

[70] Soraya Yekta, I-hung Shih, and David P. Bartel. MicroRNA-Directed Cleavage of HOXB8 mRNA. *Science*, 304(5670):594–596, 2004.

[71] Rui Yi, Yi Qin, Ian G. Macara, , and Bryan R. Cullen. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes and Development*, 17:3011–3016, 2003.

[72] Malik Yousef, Michael Nebozhyn, Hagit Shatkay, Stathis Kanterakis, Louise C. Showe, and Michael K. Showe. Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier. *Bioinformatics*, 22(11):1325–1334, 2006.

[73] Yan Zeng and Bryan R. Cullen. Structural requirements for pre-microRNA binding and nuclear export by Exportin 5. *Nucl. Acids Res.*, 32(16):4776–4785, 2004.

[74] Yan Zeng, Rui Yi, and Bryan R. Cullen. Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha. *The EMBO Journal*, 24:138–{148, 2004.

[75] Harry Zhang. The Optimality of Naive Bayes. In Valerie Barr, Zdravko Markov, Valerie Barr, and Zdravko Markov, editors, *FLAIRS Conference*. AAAI Press, 2004.