MicroRNA Prediction for Unannotated Genome-Wide and Transcriptomic Experiments

By

Robert Peace

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs
in partial fulfillment of the requirements for the degree of

PhD of Electrical & Computer Engineering

Ottawa-Carleton Institute for
Electrical and Computer Engineering

Department of Systems and Computer Engineering
Carleton University
Ottawa, Ontario, Canada
January 2016

Abstract

MicroRNAs (miRNAs) are short (18–23 nt), non-coding RNAs that play central roles in cellular regulation by modulating the post-transcriptional expression of messenger RNA transcripts. It has been previously estimated that 60-90% of all mammalian mRNAs may be targeted by miRNAs. Due to their biological importance, the ability to accurately predict miRNA sequences is of great importance. Computational prediction of miRNA are either genomic sequence-based or analyze transcriptomic data arising from next generation sequencing (NGS) experiments. Unfortunately, existing methods of *de novo* miRNA prediction often fail when applied to non-model species, and are not well suited to genome-scale data sets. Furthermore, existing methods of NGS-based miRNA prediction do not incorporate all known lines of evidence for miRNA prediction, instead focussing on either sequence-based or expression-based features of putative miRNA.

This thesis makes contributions to the state of the art of miRNA prediction which directly address the issues highlighted above. First, we develop a framework for the generation of species-specific training data sets. Three different forms of classifiers using diverse feature sets are trained and evaluated using the framework. Significant gains in precision and recall are achieved over existing methods, as measured using four diverse species from different phyla. Subsequently, the framework was applied to develop miRNA predictors in two successful genome-wide miRNA prediction studies, resulting in the discovery of 155 novel miRNA, thus verifying the real-world applicability of this work. Second, we introduce a genome-scanning miRNA prediction model which optimizes miRNA prediction for realistic experimental conditions. This model quantifies the performance of elements of the miRNA prediction pipeline,

including pre-filtering stages, whose impact was previously ignored. This comprehensive evaluation framework has enabled significant increases in prediction performance over the state of the art through the use of updated RNA secondary structure parameters. Finally, we develop a NGS-based miRNA prediction method which improves on state-of-the-art performance through the integration of all known lines of evidence which discriminate miRNA from non-miRNA. This prediction method substantially outperforms two existing leading methods on data sets from five NGS experiments across three species, and is shown to generalize to hold-out data sets.

Statement of Originality

This thesis presents the work of the author, under the supervision of Dr. James Green. This work was completed at Carleton University for the degree PhD in Electrical and Computer Engineering. Some of these results have been or will be presented in the following publications:

R. J. Peace, K. K. Biggar, K. B. Storey, and J. R. Green, "A framework for improving microRNA prediction in non-human genomes," *Nucleic Acids Res.*, vol. 43, no. 20, p. e138, Jul. 2015.

This journal article describes the SMIRP (species-specific miRNA prediction) framework, which generates high quality training data sets. The framework greatly improves miRNA prediction performance on non-model species. Chapter 5 of this thesis is adapted from this article.

R.J. Peace, J.R. Green, "Updated Free Energy Parameters Increase MicroRNA Prediction Performance", presented at World Congress on Medical Physics and Biomedical Engineering, June 2015

This conference publication demonstrates that updated folding free energy parameters greatly improve miRNA prediction performance, and introduces the genome-scanning miRNA prediction model. Chapter 6 of this thesis is adapted from this article.

P. Schaap *et al.*, "The *Physarum polycephalum* genome reveals extensive use of prokaryotic two-component and metazoan-type tyrosine kinase signaling", Genome Biology and Evolution, first published online November 27, 2015, doi:10.1093/gbe/evv237

The SMIRP framework was applied to the unannotated *P. polycephalum* genome as part of genome consortium efforts on the species. Using SMIRP, 48 miRNA were predicted and experimentally validated, 46 of which were novel. My role in this publication was in applying the SMIRP framework to create a species-specific predictor for *P. polycephalum* and providing predicted novel miRNA to my wet-lab

collaborators for subsequent evaluation and interpretation. The results from this study are included as an example application of the SMIRP framework developed as part of this thesis.

C. Adema *et al.*, "Whole genome analysis of Biomphalaria glabrata (Lophotrochozoa), a snail intermediate host" (*manuscript in preparation*)

The SMIRP framework was applied to the unannotated *B. glabrata* genome as part of genome consortium efforts on the species. Using SMIRP, 202 miRNA were predicted and experimentally validated, 107 of which were novel. My role in this publication was analogous to that of the *P. polycephalum* described above. The manuscript is currently in preparation.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Definition
RNA	Ribonucleic acid
mRNA	Messenger RNA
rRNA	Ribosomal ribonucleic acid
miRNA	MicroRNA
Pre-miRNA	Precursor microRNA transcript
Pri-miRNA	Primary microRNA transcript
DGCR8	DiGeorge Syndrome Critical Region 8
RISC	RNA-induced silencing complex
ncRNA	Non-coding RNA
Nucleotides A, C, G, and U	Adenine, Cytosine, Guanine, and Uracil
Sn	Classification sensitivity
Sp	Classification specificity
Acc	Classification accuracy
MCC	Matthew's Correlation Coefficient
Pr	Classification precision
GM	Geometric mean of sensitivity and specificity
ROC curve or ROC plot	Receiver operating characteristic curve or plot

PR-curve	Precision / Recall curve
NGS	Next-generation sequencing
CV	Cross-validation
SMOTE	Synthetic minority oversampling technique
SVM	Support vector machine
RF	Random (decision) forest
KNN	K nearest-neighbour
Re@Pr50	Recall at a threshold of 50% precision
Re@Pr90	Recall at a threshold of 90% precision

1 Introduction

1.1 Motivation

MicroRNAs (miRNAs) are short (18–23 nt), non-coding RNAs that play central roles in cellular regulation by modulating the post-transcriptional expression of messenger RNA (mRNA) transcripts [1]. Most miRNAs are considered to share a similar biogenesis mechanism: they are derived from RNA transcripts (pri-microRNAs) that contain imperfect hairpin structures (pre-microRNAs, ~80 nt in length) and are subsequently processed by one or more endonucleases (e.g. Drosha and Dicer in animals, DLC1 in plants) to form mature microRNA. After processing and formation, the mature microRNA is incorporated into the RNA-induced silencing complex (miRISC), where the microRNA guides the associated RISC proteins to the targeted mRNA strand, annealing to the target mRNA and promoting either degradation or reversible translational repression [2]. It has been previously estimated that 60-90% of all mammalian mRNAs may be targeted by miRNAs [3], and at this time over 2500 mature miRNAs have been identified in the human genome (miRBase v.21 released in June 2014 [4]). Through a myriad of comparative expression analyses and gain- and loss-of function experiments, miRNAs have been shown to be critically involved in regulating the expression of proteins involved in biological development [5], cell differentiation [6], apoptosis [7], cell cycle control [8], stress response [9], and disease pathogenesis [10]. Recent studies have also highlighted the role of miRNA in the cellular adaptation to severe environmental stresses (such as freezing, dehydration and anoxia) in tolerant animals [11]–[13]). Due to the biological importance of miRNA, the ability to accurately predict their sequences is of great significance. The discovery of novel miRNA sequences leads to new knowledge regarding biological pathways, through studying the target sequences and co-expression of these novel miRNA. Discovering a greater number of novel miRNA during miRNA prediction studies increases the amount of pathway

knowledge which can be gained as a result of these studies. Methods of computational prediction of miRNA fall into two categories: *de novo* miRNA prediction, wherein genomic sequence data sets are mined for miRNA, and next generation sequencing (NGS-based) miRNA prediction, wherein transcriptomic data sets are mined for miRNA.

Within the past few years, the field of *de novo* miRNA prediction has largely been conducted within an artificial scenario which does not accurately represent real-world applications of the field. In this scenario, all data comes from one of a small set of model organisms (species which have been widely studied due to their relevance to human life, or because of their availability in laboratory settings, and for which there is a large amount of genomic annotation available including a large set of known miRNA), the prevalence of miRNA vs. pseudo-miRNA within genomes is on the order of 1:10, and all miRNA conform to very simple RNA secondary structure criteria. These assumptions are invalid for real-world applications of miRNA prediction, where one is often attempting to identify all miRNA within the unannotated genome of a species which may be phylogenetically distant from any model species, and class imbalance is on the order of 1:1000.

In the field of *de novo* miRNA prediction, performance is measured by the ability of classifiers to differentiate pre-miRNA hairpin structures from pseudo-miRNA hairpin structures. However, prior to applying the classifier to a putative sequence, it is first tested by a pre-filtering stage. This often takes the form of a test for the presence of a stable hairpin within the secondary structure of the sequence. Training and test sets consisting of miRNA hairpin structures and pseudo-miRNA hairpin structures are extracted from larger genomic data sets during pre-filtering stages of a pipeline, and performance is measured only on the extracted hairpin data sets. The extraction of hairpin structures during data pre-filtering affects the pipeline's overall performance, but this effect is often not quantified, as performance is only measured on those data that pass the pre-filtering stage. In reality, the performance of *de novo* classification pipelines is dependent on pre-filtering performance as

much as it is dependent on classifier performance. If a real miRNA sequence was not considered for classification because it was rejected during pre-filtering, then this represents a failure of the miRNA prediction method to identify that miRNA. Conversely, the failure of the pre-filtering stage to remove negative sequences will exacerbate the class imbalance problem and may lead to a false positive prediction. Such failures were not previously measured, and their measurement is critical to improvement of *de novo* miRNA pipelines. Following the adage that "we cannot improve what we cannot measure", by developing a comprehensive evaluation framework that assesses both pre-filtering and classification stages, opportunities for improving the overall system performance become clear.

In the field of NGS-based miRNA prediction, state-of-the-art techniques fail to integrate all known lines of evidence which can be used to differentiate miRNA from non-miRNA and therefore limit achievable classification performance. These NGS-based methods have failed to leverage many advanced sequence-based features developed recently for *de novo* prediction methods. This represents a need and an opportunity to create a novel miRNA prediction method that integrates both expression-based and sequence-based features to improve classification performance.

The purpose of this thesis is to increase number of miRNA which are recovered by miRNA prediction experiments, for fixed success rates of experimental validation procedures.

1.2 Overview of Results

We have introduced a species-specific data set generation framework (SMIRP) for *de novo* miRNA prediction which leads to significantly improved miRNA prediction systems relative to existing methods which use data sets from single model species, taxon-specific data sets, and pooled-species data sets. These increases are demonstrated on four test species representing the animal, plant, and virus kingdoms.

The SMIRP framework was applied to the unannotated *B. glabrata* genome, resulting in the discovery of 202 miRNA precursors which were subsequently experimentally validated by a collaborating group . Of these, 107 are novel. Out of the 223 species in miRBase v.21 [4], only 42 have more than 202 known miRNA precursors, making our efforts on the *B. glabrata* genome one of the most successful miRNA discoveries to date.

Furthermore, the SMIRP framework was applied to the unannotated *Physarum polycephalum* genome, resulting in the discovery of 48 miRNA precursors, 46 of which are novel. All discovered miRNA were experimentally validated by a collaborating group [14].

By leveraging updated energy parameters in the RNA folding pre-filtering step of the miRNA prediction pipeline, the number of true miRNA passing the pre-filtering criteria increases from 79.47% to 94.17% for the *H. sapiens* genome. When including the performance of both pre-filtering and classification stages, overall system prediction performance increases by 64% for a typical miRNA prediction study. This increase in performance directly translates into an increase in the number of miRNA which would be recovered during a *de novo* miRNA prediction experiment.

In this thesis we develop a novel method of NGS-based miRNA prediction that integrates both sequence- and expression-based features. This method, referred to as microRNA Prediction using Integrated Evidence (miPIE), is demonstrated to outperform the two leading NGS-based miRNA prediction methods on five data sets representing five NGS experiments across three species. On average, performance is increased by 30.1% relative to the leading NGS-based miRNA prediction method, miRDeep2 [15]. This increase in performance directly translates into an increase in the number of miRNA which would be recovered during a NGS-based miRNA prediction experiment.

1.3 Organization of Thesis Document

This thesis consists of eight chapters. Chapter 2 provides the reader with an introduction to microRNA biology and a generic pattern classification pipeline. Chapter 3 reviews the state of the art of computational miRNA prediction, and highlights areas in which improvements should be made in order for miRNA prediction to produce higher-quality results in real-world scenarios. Chapter 4 states the specific problems in the state of the art of miRNA prediction that this thesis addresses. Chapter 5 describes the SMIRP framework for the development of species-specific training data sets which greatly improves classification performance for one of the most common *de novo* miRNA prediction use cases: the discovery of miRNA within unannotated, newly sequenced genomes. Chapter 6 introduces a comprehensive evaluation framework for miRNA predictors that considers the performance of both the pre-filtering and classification stages of the prediction pipeline. Application of this framework highlights several weaknesses in current miRNA prediction methods, particularly when applied in a genome-scanning experiment. This analysis leads to improvements in overall performance of miRNA prediction. Chapter 7 describes the miPIE algorithm, which integrates multiple lines of sequence- and expression-based evidence into a NGS-based miRNA prediction algorithm. Chapter 8 presents the main conclusions of this thesis, summarizes the contributions of the thesis, and provides recommendations for future work in miRNA prediction.

2 Technical Background

2.1 MiRNA biology

2.1.1 Overview

MicroRNAs (abbreviated miRNAs) are non-coding RNA (ncRNA) of length ~21nt which are present within all animal and plant species as well as most virii. MiRNA regulate gene expression through post-transcriptional binding with 3' untranslated regions of mRNA, silencing the bound RNA [1]. It is estimated that between 60% and 90% of all mammalian mRNAs are regulated by miRNAs [3]. Through the regulation of critical proteins, miRNA play a role in biological development [5], cell differentiation [6], apoptosis [7], cell cycle control [8], stress response [9], and disease pathogenesis [10], and cellular adaptation to severe environmental stresses in tolerant animals [11]–[13]. Deregulation of miRNA is related to the onset of leukemia [16]–[18], and disruption of miRNA regulation is linked to heart disease and heart failure [19], [20].

2.1.2 Biogenesis

Most animal miRNA share the following biogenesis: miRNA-encoding genes are transcribed within the nucleus by RNA polymerase II [21]. The resulting RNA structure, known as the primary microRNA (pri-miRNA), is an imperfect hairpin with a length of several hundred nucleotides [22]. Within this pri-miRNA structure is a shorter hairpin structure known as the precursor miRNA (pre-miRNA). Animal pre-miRNAs are typically between 70 and 100 nucleotides in length. Plant pre-miRNAs tend to be longer, reaching upwards of 250 nucleotides. One or both arms of the pre-miRNA hairpin structure contains mature miRNA, sequences of approximately 18-23nt which bind to mRNA untranslated regions to perform

the regulatory function of the miRNA. Figure 1 demonstrates the pre-miRNA structure and mature miRNA sequences of the *H. sapiens* miRNA has-mir-1-1.

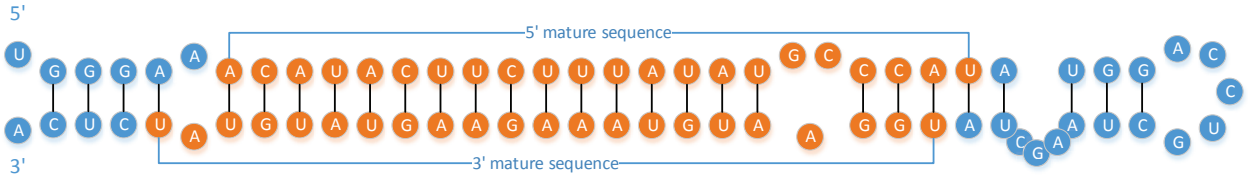


Figure 1 - Hairpin structure and sequence of miRNA hsa-mir-1-1. Mature miRNA sequences are highlighted.

After transcription, the hairpin structure of the pri-miRNA is then cleaved by the Microprocessor complex, leaving the pre-miRNA hairpin intact. Thbe Microprocessor complex consists of the proteins DiGeorge Syndrome Critical Region 8 (DGCR8) and the enzyme Drosha. DGCR8 recognizes the pri-miRNA hairpin structure and directs Drosha toward the hairpin arms. Drosha contains an RNase III domain which performs the cleaving action on the pri-miRNA structure [23].

Some pre-miRNA are formed through RNA splicing, and do not undergo cleaving by the Microprocessor complex [24].

After nuclear processing, the pre-miRNA structure is transported out of the nucleus by the exportin-5 protein and is further processed within the cytoplasm. Dicer, a second RNase III enzyme, cleaves the pre-miRNA structure, removing the loop of the hairpin. The resulting structure is known as a miRNA:miRNA* duplex; one or both strands of this duplex are functional mature miRNA. The two strands of the duplex are separated, and any strands which are not mature miRNA (known as passenger strands) are discarded [25].

Mature miRNA strands are bound to argonaute proteins in an RNA-induced silencing complex (RISC), which guide the miRNA to target mRNA regions. The binding of miRNA to mRNA inhibits ribosomal activity, preventing the expression of the protein which is regulated by the miRNA [26].

Plant miRNA biogenesis differs from animal miRNA biogenesis significantly. A dicer homolog known as Dicer-Like 1 is responsible for all cleaving action on plant miRNA. Cleaving of pre-miRNA into miRNA:miRNA* duplexes is performed within the nucleus of plants, and the duplexes are transported into the cytoplasm [27].

2.1.3 Evolutionary rate

Relative to most genetic sequences, miRNA are extremely highly conserved during evolution [28], [29]. In particular, the subsequences representing mature miRNA sequences are highly conserved across diverse species [30]. Because of their high level of conservation, miRNA are considered significant phylogenetic markers. It is believed that miRNA developed independently in plants and in animals due to the differences in biogenesis between these domains [31].

2.1.4 History

The first miRNA, lin-4, was discovered in 1993 by Lee, Feinbaum, and Ambros [32]. Lin-4 was found to regulate the LIN-14 protein in *C. elegans* by binding to a repeated sequence in the 3' untranslated region of the lin-14 mRNA. Upon its discovery, it was not known that lin-4 was a member of a larger class of regulatory ncRNA, and no further miRNA discoveries were made until 2000. The second miRNA to be discovered, let-7, was found to repress lin-41 mRNA, also in *C. elegans* [33]. Following the discovery of both lin-4 and let-7, it was determined that these two ncRNA belong to a larger family of regulatory ncRNA [34], and the term microRNA was coined [35]. Within a year of the discovery of let-7, dozens of miRNA were discovered within many organisms including *H. sapiens* [35]–[37]. The number of known miRNA has grown dramatically since 2000, due partly to the development of computational tools which predict miRNA [38] and advancements in sequencing technologies [39]. Over 28,645 pre-miRNA and 35828 mature miRNA are registered in miRBase, the miRNA database [4], as of June 2014. These miRNA come from 223 species,

of which *H. sapiens* is the most highly represented; 2588 of the mature miRNA sequences are from *H. sapiens*.

2.1.5 Experimental determination of miRNA through sequencing

In the past decade, next-generation sequencing (NGS) experiments have become the primary tool by which experimental prediction and validation of miRNA is performed. NGS experiments examine the RNA transcripts present in a sample of tissue (the sample's transcriptome), resulting in data sets which consist of a list of reads; each read corresponds to an RNA transcript which was present in the tissue sample at the time of the experiment. Expression of a given miRNA will vary between experiments, depending on the sample tissue type and the conditions under which the sample was collected. NGS data sets reflect expression levels; RNA transcripts which are highly expressed in the tissue sample will be recorded multiple times in a data set. When NGS data sets are analyzed, reads with very high similarity and overlap are grouped into read stacks; consensus sequences for these stacks and the number of reads in the stack are used to summarize large amounts of repeated sequence information.

Small RNA (sRNA) NGS experiments are used as a basis for NGS-based miRNA discovery. sRNA NGS experiments select for RNA transcripts whose lengths coincide with mature miRNA sequence lengths (18-25nt). Within this read length range, a transcriptome will contain reads which correspond to the three Dicer products: mature miRNA, miRNA* passenger strands, and pre-miRNA hairpin loop segments. Along with these Dicer products, the transcriptome will also contain reads which derive from other ncRNA such as piwi-interacting RNA and small nucleolar RNA, and from messenger RNA degradation products. If, for a given miRNA, read stacks are present in the transcriptome which align to the three Dicer products of the miRNA (mature miRNA, miRNA*, and loop) then this is considered experimental evidence that the miRNA exists within the genome and was expressed in the sample. Typically, the read stack which matches the mature miRNA will have the highest

read count, as the miRNA* and loop products are degraded after Dicer processing while the mature miRNA product is retained and incorporated into the miRISC complex.

Figure 2 shows a modified output of the miRDeep2 miRNA prediction algorithm. This output contains a. the identity of the miRNA and the total read depth corresponding to the pre-miRNA region; b. the predicted hairpin structure of the miRNA structure; and c. the normalized read depth at each nucleotide in the sequence. Note that most reads align well to the mature miRNA sequence or the miRNA* sequence, as is expected when reads are the result of Dicer processing.

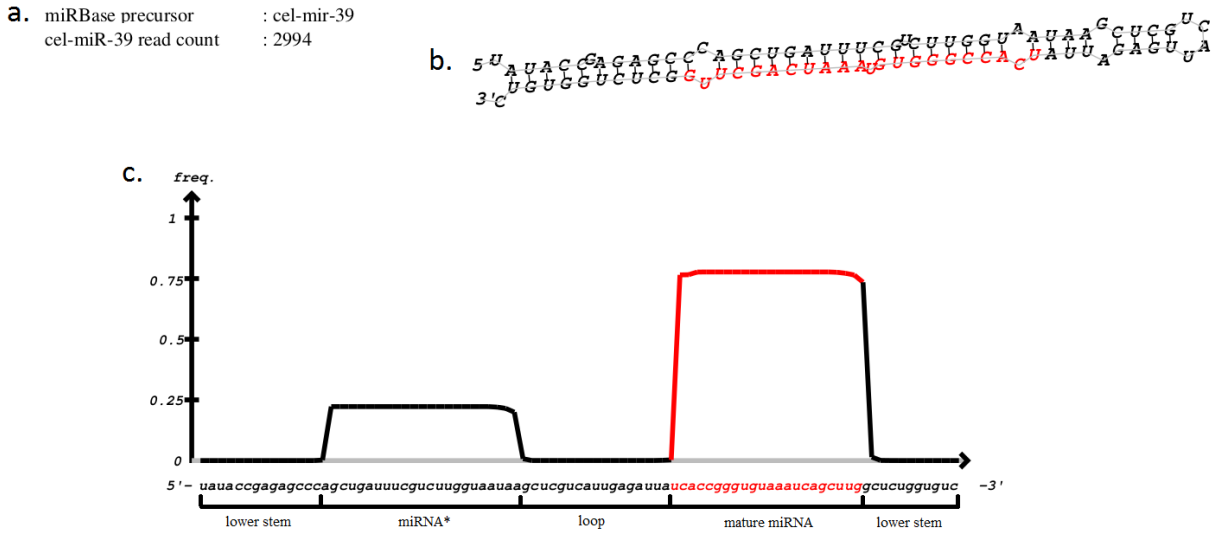


Figure 2 - NGS-based miRNA prediction information, as output by the miRDeep2 pipeline. The following information is presented: a. the identity of the miRNA and the total read depth corresponding to the pre-miRNA region; b. the predicted hairpin structure of the miRNA structure; and c. the normalized read depth at each nucleotide in the sequence. Most reads align to either the mature miRNA or the miRNA*Dicer products.

2.2 Pattern classification

Information presented in this section of the thesis is adapted from the text *Pattern Classification*, by Duda, Hart, and Stork [40].

2.2.1 Overview

The goal of pattern classification is to build models which are capable of predicting the class of unlabelled data sets based on features of each datum which discriminate between classes of data. In the context of this thesis, pattern classification is used to classify an RNA sequence as either representing a pre-microRNA hairpin ("miRNA" = positive class) or not representing a pre-microRNA hairpin ("pseudo-miRNA" = negative class). Pattern classification is a form of supervised learning; classification models are trained on a set of labelled data (*e.g.* RNA sequences representing known miRNA and RNA sequences bearing resemblance to miRNA but which are known to perform other roles), then use information gained from the training data in order to classify future unlabelled data (*e.g.* unannotated RNA sequences).

Pattern classification classifies data by means of features – numerical or categorical values which can be derived from each datum in training and testing data sets. A set of *d* features which are derived from a datum comprise a feature vector of length *d*. For example, an RNA sequence may be represented by a simple feature vector of length 5 containing the % of A, G, C, and U within the sequence, and the total sequence length. Feature vectors can be represented as a point in *d*-dimensional space; classification models place decision boundaries within this *d*-dimensional space which attempt to discriminate between feature vectors representing different classes.

2.2.2 Training Pipeline

In pattern classification, training is the act of selecting appropriate labeled data representing positive and negative classes, and building a classification model from this data which is able to predict the class of future unlabelled data not found in the training data. Figure 3 demonstrates a typical training pipeline. Careful attention must be paid to each step of the training pipeline, as small methodological errors in each step may alter classification results drastically. Furthermore, the quality of a classification model is affected by each step in the training pipeline.

2.2.2.1 Performance estimation

Performance estimation is central to the success of classification experiments because all decisions made within the training pipeline are made with the goal of maximizing classifier performance with regards to performance metrics. No single metric exists which can fully describe the performance of a classifier. As with other elements of a classification experiment, selection of appropriate metrics for performance estimation of a classifier is dependent on the data and goal of the classifier.

The confusion matrix is the basis of all classification performance reporting. This matrix describes the number of test samples of each true class which fall into each predicted class. For a binary classification problem such as miRNA prediction, the confusion matrix is a 2x2 matrix which describes the number of true positives, true negatives, false positives, and false negatives reported by a classifier for a given test set. Figure 4 shows the form of a binary confusion matrix.

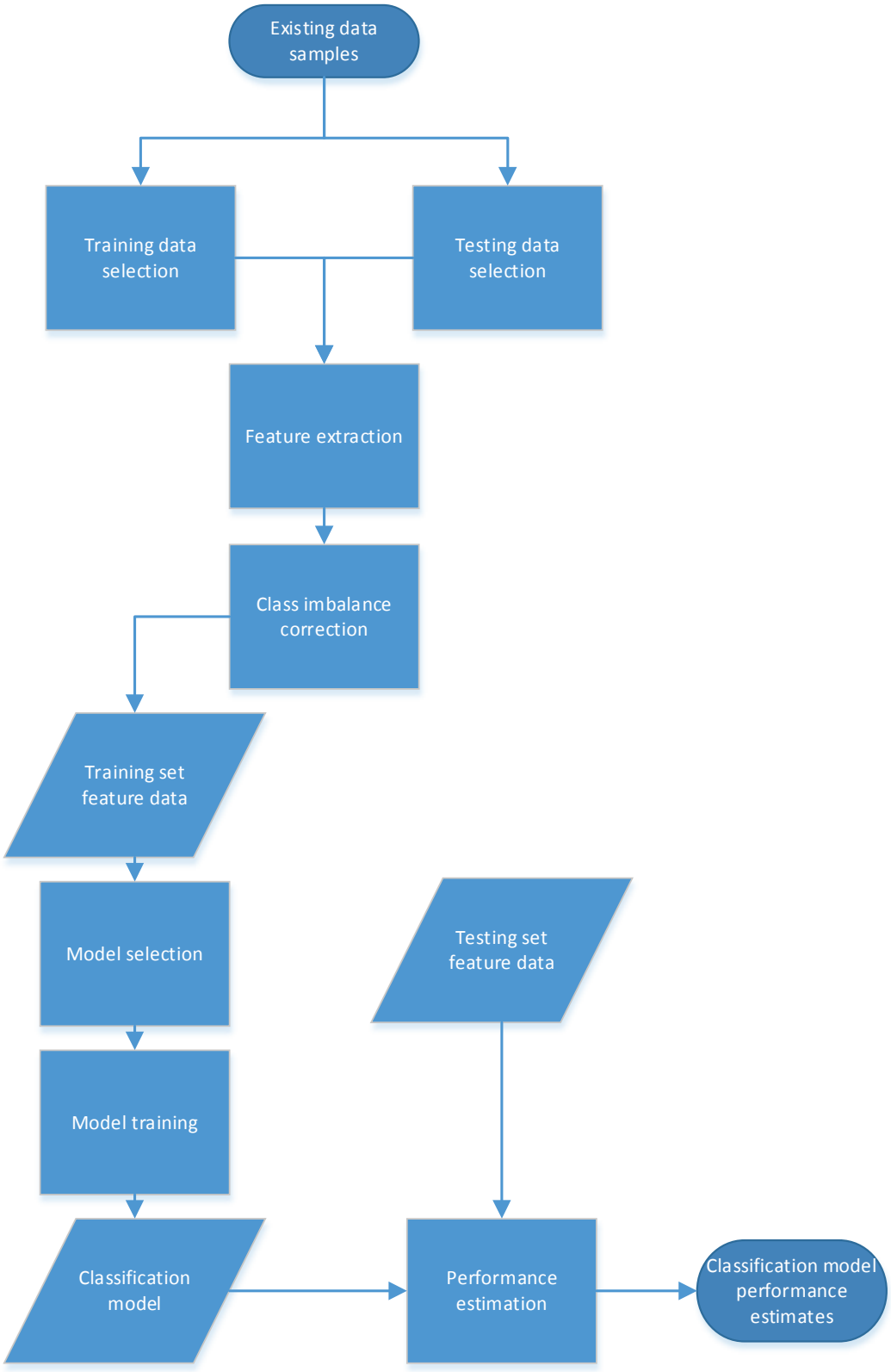


Figure 3 - Training pipeline for a pattern classification model

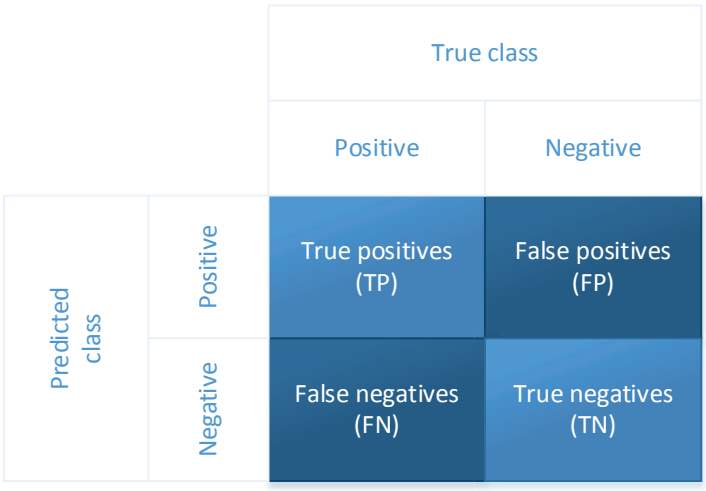


Figure 4 - Confusion matrix for a binary classifier

From the confusion matrix, several metrics can be derived which elucidate elements of the performance of a classifier. Sensitivity (also known as recall) and specificity are two of the most common of these metrics.

Sensitivity, Sn, otherwise known as recall, is the percentage of truly positive samples which the classifier correctly identifies as positive. Sensitivity describes the classifier's ability to recover positive samples, and estimates the percentage of positives that will be recovered if a complete data set is predicted. It is derived as follows:

$$Sn = \frac{TP}{TP + FN}$$

Specificity, Sp, is the percentage of truly negative samples which the classifier correctly identifies as being from the negative class. Specificity describes the classifier's ability to filter out negative samples. It is derived as follows:

$$Sp = \frac{TN}{TN + FP}$$

Precision (Pr), or positive predictive value, is an alternative to specificity which is especially useful when class imbalance is large. Precision is the percentage of predicted positive samples which are truly positive. Practically, precision estimates the proportion of true positives among all samples predicted to be from the positive class. It is derived as follows:

$$\mathrm{Pr} = \frac{TP}{TP + FP}$$

In the case of a test data set in which the class imbalance does not accurately reflect reality, prevalence-corrected precision may be used. This is the case in Chapters 5 and 6 of this thesis. In Chapter 7 of the thesis, test set class imbalance reflects real-world class imbalance, so prevalence correction is not required. Prevalence-corrected precision, when applied in the thesis, is derived from the performance over the positive samples in the test data (Sn), performance over the negative test samples (Sp), and the relative prevalence, ρ, of the negative class (e.g. for a ratio of 1000:1 negatives:positives, ρ=1000). Prevalence-corrected precision is defined as:

$$(\text{prevalence-corrected})\mathrm{Pr} = \frac{Sn}{Sn + \rho(1 - Sp)}$$

In Chapters 5 and 6 of this thesis, any mention of 'precision' refers to 'prevalence-corrected precision'.

Accuracy, Acc, is the percentage of data which are correctly classified, regardless of class. It is derived as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy estimates the performance of the classifier across all data; however, it is not always useful as a summary metric. When class imbalance is present, accuracy is biased toward the majority class; in extreme cases, the accuracy and specificity of a classifier converge and Sn has little impact on Acc.

An alternative to accuracy which is prevalent in the field of miRNA prediction is the geometric mean of sensitivity and specificity (GM). GM is derived as follows:

$$GM = \sqrt{Sn * Sp}$$

GM is insensitive to class imbalance; classifier performance over the positive and negative classes are weighted equally regardless of the true size of each class.

The above metrics describe the performance of classifiers at a single parameter set. By varying parameters, such as the decision threshold in the case of a classifier producing a continuous score or confidence for each sample, a classifier can move from permissive (identifying all data as positive) to restrictive (identifying all data as negative). ROC curves and precision/recall curves illustrate the performance of a classifier across all values for such a parameter.

ROC curves plot Sn against (1-Sp). Moving from left to right, the classifier varies from maximally restrictive to maximally permissive. A classifier which randomly selects the class of each datum would trace a straight line from bottom left to top right of the plot. Increases in classification performance move the classifier's ROC curve toward the upper left corner of the plot. Figure 5 shows the ROC plot of a classifier which performs better than random. Because both sensitivity and specificity are insensitive to class imbalance, the ROC curve shares the same property; a ROC curve describes classifier performance in a way that is agnostic to (or ignorant of) class imbalance.

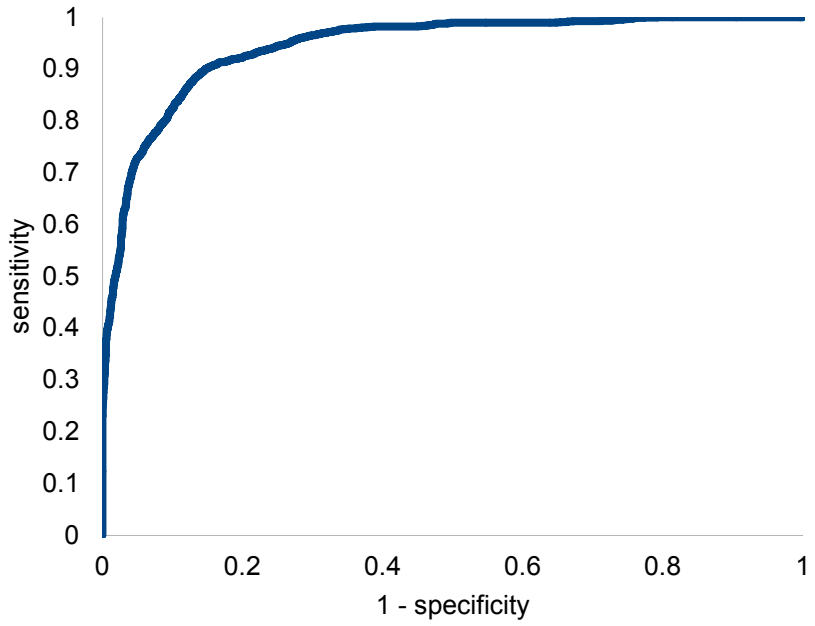


Figure 5 - A typical ROC plot. Classifier sensitivity is measured against 1-specificity. Ideal classification performance occurs in the top left corner (high sensitivity and specificity). Random classification results in a curve along the diagonal.

Precision-recall curves (PR-curves) plot Pr against Sn. Unlike ROC curves, PR-curves are sensitive to class imbalance; a single curve describes the performance of a classifier only at a specific class imbalance. For problem domains with high class imbalances, PR-curves better elucidate a classifier's real-world performance. Moving from the top to the bottom of a PR-curve, the classifier varies from maximally restrictive to maximally permissive. As classification performance improves, the PR-curves for the classifier moves toward the top-right corner of the plot. Typically, an acceptable precision threshold is chosen from a PR-curve, and classifiers are measured based on the recall which they achieve at this precision threshold. Figure 6 shows a typical PR-curve.

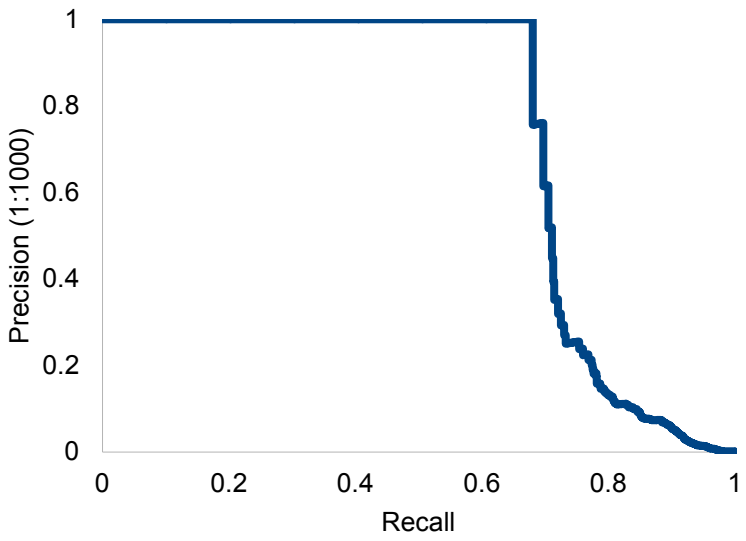


Figure 6 – A typical precision-recall plot. Here, classifier recall is measured against prevalence-corrected precision at a class imbalance of 1:1000. Ideal classification performance occurs in the top right corner, where precision and recall are both high.

2.2.2.2 Selecting training and testing data

In order to train a classification model, labeled positive and negative data must be collected as input to the model. Data availability, reliability of data, reliability of labels, and class imbalance within the data (i.e., ratio of samples which belong to each class) are all data quality issues which may affect classification performance and must be addressed within the training pipeline.

Within the context of miRNA prediction, training data consists of RNA sequences which are typical of miRNA with respect to length and secondary (hairpin) structure. Labelled positive data is retrieved from miRNA databases such as miRbase [4] or mirTarBase [41] containing sequences proven (by independent wet lab experiments) to be true miRNA sequences. Negative data samples are often extracted from genomic regions known to have functions other than coding for miRNA, such as protein-coding regions or other classes of non-coding RNA (ncRNA).

2.2.2.3 Feature extraction

Selecting appropriate features with which to train a classifier is a crucial step of the training process, as the shape of the feature vector defines the space in which data are situated and decision boundaries are formed. Features are selected based on their ability to discriminate data from different classes either in isolation, or in combination with other features. An ideal feature would separate positive and negative data into two distinct groups, however real data rarely contains such features, and no such feature exists for miRNA prediction. Increasing the number of features in the feature vector typically increases the separation between classes but does so at the cost of model complexity (the number of free parameters used to define the classification model). In general, more complex models are more likely to *overfit* training data, resulting in overly complex decision boundaries which do not generalize well to future unlabelled data [40]. Additionally, as the number of features increases, data quality is reduced as per the curse of dimensionality [42]. Therefore, a trade-off must be made during feature selection – multiple features are required in order to discriminate between classes, but model complexity should be minimized when possible.

Features typically used in miRNA prediction include RNA sequence composition, minimum free energy of RNA secondary structure, and robustness of observed RNA secondary structure. When miRNA prediction is performed on transcriptomic data, expression of Dicer products and mature miRNA duplex stability can also be used as features.

2.2.2.4 Class imbalance correction

In many classification problems, such as miRNA prediction, classification is performed in the presence of significant class imbalance: negative samples outnumber positive samples by a large margin. For example, in the human genome, there are approximately 11 million pseudo-miRNA sequences which have a length and secondary structure typical of miRNA, and only approximately 2,000 known miRNA precursors, leading to a class imbalance of >5000:1. Classifiers which are naively trained on imbalanced data sets are biased toward

the dominant (negative) class, affecting overall classification performance. This bias is detrimental when it is important to correctly classify samples from both the majority and minority classes; when class imbalance is extreme, a classifier which seeks to maximize accuracy may do so by correctly classifying all majority data at a severe cost to correct prediction rates on the minority class. Several techniques have been developed which correct the learning bias which occurs when training data contains a high class imbalance. Synthetic data can be added to the minority class in order to balance the number of samples in each class, the majority class can be undersampled in order to achieve the same balance, or classification models can be built to handle the imbalanced training data set as is, using differential weighting of errors made on each class for example.

One example of a class correction algorithm which is applied in this thesis is the SMOTE synthetic minority oversampling technique [43]. The SMOTE technique oversamples the minority class by a factor *b* through the following algorithm:

> For each sample s_i in the minority data set S within a given feature space:
>
> > For each of the b nearest samples from S, n_j *(where* $i = 1$ through b):
> >
> > > Generate a new sample whose location in the feature space is a random point on the line between s_i and n_j.

Class imbalance must be taken into account not only during the training of classifiers, but also during the evaluation of trained classifiers, as discussed below.

2.2.2.5 Splitting of training and testing data

Because classification models are built in order to optimally separate their training data, performance on test data outside of the training data set will tend to decrease relative to performance observed on the training data. As mentioned previously, overly-complex models in particular will perform very well on training data but may generalize very poorly to other data sets. For this reason, in order to estimate the applicability of a classifier, it is necessary to train using only a portion of the available labeled data. A portion of the data

must be held out as testing data in order to estimate the generalizability of the model. Splitting of labeled data into training and testing can be performed several ways. A simple 70% / 30% split is common. *n*-fold cross-validation is the most widely used method for generating training and testing data sets. During *n*-fold cross-validation, available labeled data is separated into *n* even partitions. *n* classifiers are trained, each on *n-1* partitions, and each with a distinct partition as hold out test data. The total performance of the *n* classifiers on the hold-out partitions is then used to estimate real-world performance of a classifier trained on the training data set.

2.2.2.6 Model selection

At the core of each classifier is a model which places decision boundaries within a given feature space. Many such models exist, and these models employ vastly different algorithms for the determination of decision boundaries. The most common classification model type in miRNA prediction is the support vector machine, which draws a plane which optimally separates the classes within a feature space, maximizing the margin between the decision boundary and the nearest training points from each class (dubbed 'support vectors'). Other common model types are decision trees, which perform classification based on a series of "*20 questions"* style decisions ("Is feature *x* larger or smaller than value *y*?"). Each question is modeled as a node in a tree graph. Leaf nodes of the tree each contain a class prediction. Random forests are made up of collections of tens or hundreds of simple decision trees trained on subsets of the training data which ultimately vote on the class prediction for an input test feature vector. Other less commonly used classifiers within miRNA prediction include artificial neural networks, K nearest-neighbor classifiers, hidden Markov models, and Bayesian models. Multiple classifiers of the same or different model types can also be grouped into ensemble models, in which each component classifier casts a vote regarding the class of input data.

Model selection is a non-trivial decision within pattern classification because of the "no free lunch theorem", which states that no classification model is inherently superior to any other model when all possible classification problems are considered [44]. The quality of a model for a given classification problem is most often a function of the degree of agreement between the characteristics of the data and the underlying assumptions of the model.

2.2.2.7 Model training

Following selection of the model type, the parameters of the classification model should be optimized in order to describe optimal decision boundaries. Many models, such as support vector machines, rely on a kernel function which determines the general shape of decision boundaries. Gaussian functions, radial basis functions, and linear functions are typical kernel functions. Selection of an appropriate kernel function can affect performance greatly. Some models require the optimization of model-specific hyper-parameters such as the number of trees in a random forest. Most models are trained using a cost function, which specifies the relative penalty of misclassification of positive and negative samples. Many models can be trained using different training algorithms which vary with regards to generated models and training speed. Finally, stopping criteria for training algorithms must be set in order to produce models which balance training set accuracy and generalizability. Each of these factors can affect the quality of a classification model.

2.2.3 Prediction Pipeline

Once a classification model has been built for a given training set, it is capable of predicting the classes of data in unlabelled sets.

Relative to model training, prediction of unlabelled data using an existing model is a straightforward procedure. Each pattern in the unlabelled data set is converted into a feature vector, using the features selected during the model training procedure. Feature vectors are passed through the classification model, which outputs a classification. Some

classification methods also output confidence values, typically between 0 and 1, which state the confidence with which the classifier believes that a pattern belongs to a given class. In some applications, the computational run-time of a classifier is critical in order to achieve real-time predictions. This is not particularly the case for miRNA prediction.

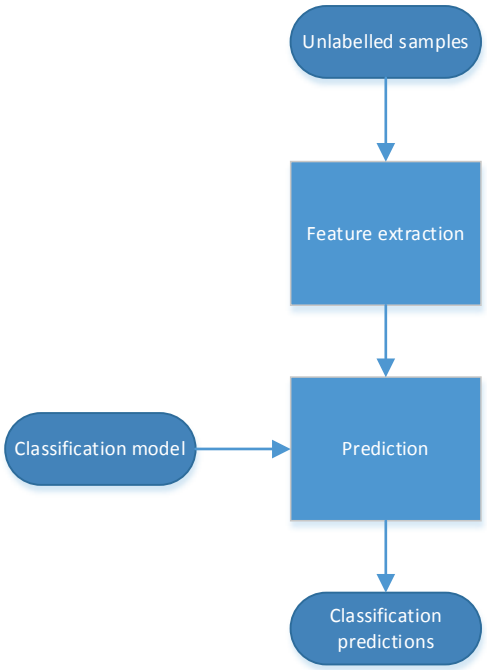


Figure 7 - Classification pipeline for unlabelled data

3 Literature Review

3.1 Overview of computational miRNA techniques

Computational prediction of pre-miRNA sequences can be broadly categorized into three major areas – homology-based techniques, NGS-based techniques, and *de novo* machine learning techniques.

Homology-based techniques represent the earliest efforts at prediction of miRNA [45]–[47]. These techniques discover miRNA based on similarity between these miRNA and existing miRNA. While not capable of discovering taxon- or species-specific miRNA with novel sequences, these techniques provide a means for discovering miRNA homologs across species. The earliest homology-based miRNA prediction techniques include MiRscan, which was used to successfully predict many miRNA in *C. elegans* and *C. briggsae* genomes [45]; srnaloop, which predicted additional miRNA within the *C. elegans* genome [47]; and miRseeker, which played a similar important role within *D. melanogaster and D. pseudoobscura* genomes [46]. The MIRcheck [48] and findMiRNA [49] algorithms were leveraged to discover 23 and 13 new miRNA within the *A. thaliana* genome, respectively. MicroHARVESTER [50], like MIRcheck and findMiRNA, specializes in plant miRNA homologs. Most improvements in the state of the art of homology-based miRNA prediction occurred between 2003 and 2006, including all of the above studies. These studies are summarized by Berezikov *et al.* [38]. Improvements in homology-based miRNA prediction techniques were still pursued as recently as 2013, when improvements on the state of the art were made by the miR-Explore algorithm [51].

NGS-based techniques focus on analysis of next generation sequencing data in order to identify miRNA at the transcriptome level. These techniques are relatively new, and have been growing in popularity since their inception in 2008. Of these techniques, miRDeep [52]

and the updated miRDeep2 [15] have emerged as *de facto* standards. Deep sequencing techniques provide higher confidence miRNA predictions, as the expression of RNA sequences which match the predicted miRNA constitutes experimental evidence in favour of the miRNA identification. Class imbalance within a transcriptome is less extreme than that of a genome, making identification easier since only expressed sequence regions need be explored. However, deep sequencing techniques are limited to identification of those miRNA which are expressed under the specific experimental conditions at the time of data collection, and these techniques are biased toward identification of highly expressed miRNA [53]. These techniques are reviewed in section 3.2 of this thesis.

De novo machine learning techniques predict miRNA based only on information which can be derived from unannotated RNA sequence data. These techniques employ training data sets of known miRNA and pseudo-miRNA (miRNA-like hairpins which are not functional miRNA) in order to predict RNA sequences which correspond to functional miRNA. *De novo* miRNA prediction techniques require no annotation and can function on either genomes or transcriptomes. These techniques are reviewed in section 3.3 of this thesis.

The three major branches of miRNA prediction are complementary when searching for miRNA within a given species of interest. Homology-based techniques provide a small set of high quality miRNA identifications. NGS-based prediction methods only consider RNA sequences which are expressed, leading to potentially lower overall recall but higher precision. *De novo* machine learning techniques examine an arbitrary sequence (up to the entire genome) for one or more putative miRNA. Therefore, they possess the highest potential recall, but suffer from low precision due to high class imbalance. When discovering novel miRNA within a species of interest, each of these three prediction techniques provides unique prediction information.

3.2 Examination of the state of the art of NGS-based miRNA prediction

In this section we examine the state of the art of miRNA prediction techniques for transcriptomic data sets. We will examine 14 prediction methods spanning from 2008 to 2015, as well as 9 RNA sequencing pipelines which incorporate miRNA prediction methods, and 4 experiments which apply miRNA prediction methods to transcriptome data in order to predict novel miRNA within a species of interest.

3.2.1 NGS-based miRNA prediction methods

NGS-based miRNA prediction methods can be broadly categorized using the following distinctions:

- The method is specific to the animal or plant kingdoms, or is applicable to multiple kingdoms.
- The method's primary prediction is based on sequence homology, or miRNA characteristics, or both methods are used. Homology-based methods provide high confidence predictions of typical miRNA, while methods using miRNA characteristics may discovery additional novel miRNA.
- The method maps ncRNA reads to a reference genome during its prediction pipeline, or performs analysis without a reference genome. Mapping to a reference genome removes many false positives from data sets, however genomic data is not available for all species.
- The method employs a machine learning algorithm during its prediction pipeline, or does not.
- The method requires that reads form a miRNA:miRNA* duplex, or does not. The discovery of a canonical miRNA:miRNA* duplex greatly increases miRNA

prediction confidence, however sensitivity is reduced as duplexes are not present for all miRNA, especially those which are not highly expressed.

Table 1 provides a summary view of the 14 transcriptomic miRNA prediction methods with respect to the above criteria.

MiRDeep [52], introduced by Friedlander *et al*. in 2008, was the first computational miRNA prediction method which made use of transcriptome data sets. MiRDeep scores stacks of RNA sequence reads based on the likelihood that these reads are the result of miRNA biogenesis. Once a putative pre-miRNA region is determined surrounding a read stack, RNA secondary structure is computed and the minimum free energy (MFE) is computed as a discriminating feature. The miRDeep study places a strong emphasis on the ability of miRDeep to estimate the quality of its results based on a statistical model of its outputs. MiRDeep was initially applied to a *C. elegans* data set, discovering four novel miRNA which were successfully validated.

Table 1 - Summary of methods for NGS-based miRNA prediction

Method	Year	Kingdom(s)	Homology / characteristics / both	Reference genome required	Machine learning used	MiRNA duplex required
MiRDeep [52]	2008	Animal	Characteristics	Yes	No	No
Mireap (no citation available)	2008	Animal	Characteristics	Yes	No?	No?
MiRMiner [54]	2008	Animal	Both	No	Yes	Yes
miRExpress [55]	2009	Animal + Plant	Homology	No	No	No
miRTRAP [56]	2010	Animal	Characteristics	Yes	No	No
miRanalyzer [57]	2011	Animal + Plant	Both	Yes	Yes	No
miRDeep-P [58]	2011	Plant	Both	Yes	No	No
MiRDeep2 [15]	2012	Animal	Both	Yes	No	No
McRUM for miRNA [59]	2013	Animal + Plant	Characteristics	No	Yes	No
MiRPlex [60]	2013	Animal + Plant	Characteristics	No	Yes	Yes
MiRDeep* [61]	2013	Animal	Characteristics	Yes	No	No
MirPlant [62]	2014	Plant	Characteristics	Yes	No	No
MIRPIPE [63]	2014	Animal + Plant	Homology	No	No	No
miRdentify [64]	2014	Animal	Characteristics	No	No	Yes
miRNA and piRNA [65]	2015	Animal	Characteristics	No	Yes	No

No citation for MIREAP is available, therefore knowledge of the method is limited, however it is used within larger pipelines [66] and has been employed to successfully discover novel miRNA [67]. From MIREAP's description, it combines position and depth of small RNA reads with a microRNA biogenesis model.

The basis of the MiRMiner [54] miRNA prediction method is the thermodynamic stability of miRNA:miRNA* duplexes recovered from transcriptomic data. A classifier is trained on a single feature, representing the MFE of the duplex. As a pre-filtering step, transcriptomic data is aligned against known miRBase miRNA sequences using the BLAST algorithm [68]. MirMiner was tested initially in a study involving 13 worm transcriptomes, finding 114 novel miRNA across these species.

MiRExpress [55] focuses on quantifying the expression levels of miRNA within data sets. As a step toward this goal, it identifies likely miRNA sequences within transcriptomic data sets. The MiRExpress algorithm uses a customized implementation of the Smith-Waterman algorithm [69] in order to align sequence data with known miRNA. The MiRExpress alignment method identified 79 previously-unknown *H. sapiens* miRNA within two transcriptomic data sets.

miRTRAP [56] detects miRNA within transcriptomic data sets using the characteristic read patterns of miRNA biogenesis as well as information retrieved from the larger genomic context of the miRNA. This larger context includes the number of miRNA in the region of the candidate miRNA, the number of non-miRNA ncRNA within this region, and anti-sense reads which match to the miRNA reads. miRTRAP was applied to a *C. intestinalis* data set, and recovered 36 miRNA which were homologous to known miRNA as well as 20 novel miRNA families.

MiRanalyzer [57] uses an ensemble of five random forest classifiers in order to predict novel miRNA. As a pre-screening step, bowtie [70] is used to identify reads which match to known miRNA and known mRNA. MiRanalyzer uses a set of pre-processed genome files for

alignment during its prediction pipeline; 25 animal and 6 plant genomes are available. Feature sets used for classification differ for plant and animal species. No novel miRNA have been identified by MiRanalyzer.

MiRDeep-P [58] is an extension of the miRDeep algorithm which targets plant sRNA data sets. This method is a straightforward extension, altering only the values of miRDeep's decision rules, and adding additional prediction criteria based on Meyers' observations of plant miRNA [71]. Meyers' criteria represent a base set of criteria which all plant sequences must meet in order to be accepted for annotation as miRNAs. MiRDeep-P recovered 18 novel miRNA from *Arabidopsis* sRNA data sets which were successfully validated.

MiRDeep2 [15] moves forward from the original MiRDeep, implementing an updated prediction algorithm while maintaining the ability to estimate prediction quality. MiRDeep2 increases the number of RNA secondary structure criteria used during pre-filtering of putative pre-miRNA regions. MiRDeep2 was originally applied to seven organisms within the animal kingdom, and discovered on average 14 novel miRNA within each data set. MiRDeep2 has since been applied to transcriptomic data sets from a wide range of species, resulting in many novel miRNA predictions. Currently, miRDeep2 is the *de facto* standard for NGS-based identification of miRNA. The miRDeep2 pipeline is analyzed further in section 3.2.4 of this thesis.

Menor, Baek and Poisson implemented a novel kernel-based learning algorithm, the "multi-class relevance units machine" [59], with the goal of classifying several types of ncRNA from transcriptomic data sets. Reads are classified as miRNA, piRNA, or other ncRNA based on k-mer representations of the reads. This method was trained and tested on data sets of known ncRNA, but not applied to real-world transcriptomic data.

MiRPlex [60] implements a SVM for classification of miRNA-miRNA* duplexes found in transcriptomic data; size, stability, and composition features distinguish miRNA duplexes from non-miRNA duplexes. The MiRPlex classifier was trained using positive data consisting

of known duplexes from animals in miRBase, and negative data consisting of duplexes from a *D. melanogaster* sRNA data set which do not match known miRNA. MiRPlex was applied to four animal sRNA data sets, and a separately trained model was applied to a single plant sRNA data set. No novel miRNA were reported from these experiments.

Like miRDeep2, miRDeep* [61] extends the original miRDeep algorithm. Also like miRDeep2, focus is placed on improvements in the precursor extraction methodology and increases in RNA secondary structure criteria. Unlike miRDeep2, MiRDeep* also focuses strongly on usability; all elements of the miRDeep* pipeline are implemented within a Java graphical user interface. An analysis of the performance of miRDeep* against other popular methods was performed by the miRDeep* authors, wherein performance of miRDeep* and miRDeep2 were found to be comparable. In the initial study, four novel miRNA were discovered from prostate cancer data sets using the miRDeep* algorithm.

The authors of miRDeep* also produced a plant-specific miRNA predictor, mirPlant [62]. MirPlant implements a very similar pipeline to miRDeep-P, however the precursor extraction technique used differs. MiRDeep-P presents the user with a graphical Java interface, much like miRDeep*. While no novel miRNA were recorded using mirPlant, the authors of mimPlant demonstrated that, using mirPlant, performance on three plant sRNA data sets is improved relative to miRDeep-P.

MIRPIPE [63] aims to improve the viability of miRNA prediction within sRNA data sets originating from niche model organisms. Reads are aligned to each other, then read stacks are matched to known miRNA using simple matching criteria. In independent experiments, MIRPIPE performs similarly to methods which align sRNA reads to the genome for the identification of known miRNA, though the homology-based nature of MIRPIPE is not likely to generalize to novel predictions on niche species.

MiRdentify [64] combines stringent miRNA:miRNA* duplex requirements with RNA secondary structure filters in order to generate sets of high-confidence miRNA predictions

without a reference genome. RNA sequences from transcriptomic data sets which form strongly paired duplexes are considered to be candidate miRNA duplexes; no requirement is made that the two sequences which form the duplex are co-located in a genome. The miRdentify method is designed to work with large sRNA data sets encompassing multiple experiments. MiRdentify compares favourably to miRDeep2, miRDeep* and miRanalyzer when applied to the study's test data set, and identified two novel miRNA in *H. sapiens* ChrY.

Menor, Baek, and Poisson again apply a machine learning approach to miRNA prediction in sRNA data sets in their 2015 study. The k-mer feature set used in the authors' previous study was augmented with additional sequence features, and systematic feature selection was employed. Classification is performed using multi-class relevance units machines, as in the previous study, and a Gaussian kernel. Like the MiRPlex study, candidate miRNA are read stacks which form duplexes, and no mapping to the genome is performed. The authors compare their method to MiRPlex as these two methods share a common set of candidate miRNA, and conclude that their method increases the number of miRNA recovered from four animal data sets ten-fold relative to MiRPlex.

3.2.2 sRNA data set pipelines

Because NGS sRNA data sets contain reads pertaining to multiple types of biologically relevant ncRNAs (miRNA, piRNA, snoRNA, snRNA), several methods have been developed which incorporate multiple ncRNA prediction methods within larger pipelines. These pipelines aim to perform comprehensive analysis on sRNA data sets. With respect to miRNA, these pipelines have one or more of the following goals:

1. Detection of known miRNA
2. Prediction of novel miRNA
3. Detection or prediction of miRNA targets

4. Quantification of expression levels of miRNA across multiple experiments

Additionally, these pipelines contain GUIs which simplify the analysis of sRNA data and present visualizations of the data. We have identified the following major axes upon which ncRNA pipelines can be measured, with respect to miRNA prediction:

- Which input file formats are supported by the pipeline?
- Does the pipeline detect known miRNA?
- If the pipeline detects known miRNA, what algorithm is used for the detection?
- Does the pipeline predict novel miRNA?
- If the pipeline predicts novel miRNA, what algorithm is used for this prediction?

Table 2 examines 9 sRNA data set pipelines along these five axes.

In 2008, Moxon *et al.* introduced the UEA sRNA Toolkit, a toolkit for analysing plant sRNA data sets [72]. This toolkit uses a custom algorithm, miRCat, for the prediction of miRNA precursors within sRNA data sets. MiRCat maps reads to the genome, then searches for pairs of reads which form miRNA:miRNA* duplexes, and finally analyzes the hairpin structure of the resulting precursor candidates using criteria defined in [73].

DSAP [74], the deep-sequencing small RNA pipeline, is a web service which performs pre-processing of sRNA data sets, matches reads to known ncRNA including miRNA via the BLAST algorithm [68], and quantifies miRNA expression level differences across multiple experiments. Finally, for each miRNA in the data set, DSAP identifies all species which have miRNA from the same family.

Moxon *et al.* advanced their toolkit in 2012, creating the UEA sRNA workbench [75]. This work extends the previous effort in the areas of visualization and the profiling of expression levels of miRNA across multiple experiments. MiRNA candidates which match known miRNA in miRBase are also highlighted as such by the toolkit.

Table 2 - Summary of sRNA pipelines for examination of NGS data sets

Method	Year	Detects known miRNA	Algorithm for miRNA detection	Predicts novel miRNA	Algorithm for miRNA prediction	File formats accepted
UEA sRNA Toolkit [72]	2008	No	N/A	Yes	miRCat	FASTA
DSAP [74]	2010	Yes	BLAST [68]	No	N/A	TSV
UEA sRNA Workbench [75]	2012	Yes	Unknown	Yes	miRCat	FASTA
miREvo [76]	2012	Yes	Novel Whole Genome Alignment algorithm	Yes	MiRDeep2	Unknown
mirTools 2.0 [66]	2013	Yes	SOAP [84]	Yes	MiRDeep, MIREAP	FASTA, SAM, BAM
MiRGator [77]	2013	Yes	Bowtie	Yes	MiRDeep2	Unknown
miRspring [78]	2013	Yes	SAMTOOLS [79]	No	N/A	BAM
CAP-miRSeq [80]	2014	Yes	MiRDeep2	Yes	MiRDeep2	FASTA, FASTQ
ISRNA [81]	2014	Yes	BLAST [68]	No	N/A	FASTQ, txt

MiREvo [76] performs three major functions for sRNA data sets: Identification of novel miRNA, detection of homologs to known miRNA, and profiling of miRNA expression across multiple species or multiple experiments. Emphasis is placed on the analysis of miRNA homologs across species in order to estimate the evolutionary rate of these miRNA. MiRDeep2 is implemented within the MiREvo pipeline in order to predict novel miRNA, while the identification of homologs is performed using a novel whole genome alignment algorithm.

MiRTools 2.0 [66] detects known miRNA, tRNA, snRNA, snoRNA, rRNA, and piRNA from within sRNA data sets and profiles these ncRNA across multiple experiments. MiRNA targets are also predicted, and functional annotation of targets is performed. Finally, novel miRNA and piRNA are predicted from within the data sets using miRDeep and MIREAP.

MiRGator v3.0 [77] is a web portal which provides users access to sequence editing, counting, sorting and ordering tools, miRNA and miRNA target identification tools, and miRNA:target co-expression information. 73 human sRNA data sets have been curated by the MiRGator software, providing an existing library of miRNA data.

MiRspring [78] implements an index-compression algorithm in order to store sRNA data sets in relatively small file sizes (approx. 3MB per data set), then leverages this efficient file format for miRNA analysis. MiRspring provides the user with information regarding the global Dicer processing of all miRNA within a data set, and the Dicer processing details of specific miRNA within the data set. Special attention is paid to the identification of isomiRs.

CAP-miRSeq [80], the comprehensive analysis pipeline for miRNA sequencing data, is a tool which performs pre-processing, alignment, miRNA detection, miRNA quantification, visualization, differential expression analysis, and variant detection in miRNA coding regions. Particular emphasis is placed on the user-friendliness and practicality of the program. Here, miRDeep2 is used both for its detection of known miRNA and its prediction of novel miRNA.

ISRNA [81], the Integrative Short Reads Navigator, is an online toolkit which provides the user with data set-wide statistics including genomic location, length distribution, and nucleotide composition bias. It also provides expression data and genomic location for known miRNA and other known ncRNA identified within the data. Again, emphasis is placed on the user-friendliness of the toolkit.

3.2.3 NGS experiments for miRNA discovery in species of interest

Several studies exist in the literature which apply next-generation sequencing techniques along with NGS-based miRNA prediction methods in order to discover novel miRNA within tissue samples of species of interest.

MiRDeep2 is the most commonly applied deep sequencing analysis pipeline for miRNA prediction studies. Yin *et al.* analyzed miRNA within differently aged rat brain samples using miRDeep2 for miRNA prediction [82]. Differential analysis of miRNA was performed using DESeq2 [83], while MiRNA target prediction was performed using Targetscan [84]. This experiment discovered 547 known miRNA within the tissue and predicted 171 candidate novel miRNA, though only three of these novel miRNA were experimentally validated. Cowled *et al.* applied the miRDeep2 prediction pipeline to the black flying fox [85]. Like the study of Yin *et al*., this study uncovered several hundred (222) known miRNA, and discovered 177 novel miRNA within the tissue, though none of these miRNA were experimentally validated.

Gu *et al*. applied the MIREAP algorithm to the prediction of miRNA in a maize endosperm deep sequencing data set [67]. MiRNA were mapped to the maize genome using SOAP [86], while known miRNA were identified using BLAST [68]. Target prediction was performed using the WMD3 software. This effort resulted in the recovery of 95 known miRNA, and the discovery of 18 novel miRNA which were validated through RT-PCR.

For the identification of miRNA within hexaploid wheat data sets, Agharboui *et al.* have developed a pipeline which combines the HMMIR [87] and MiPred [88] sequence-based miRNA prediction algorithm with the miRdup* [89] mature miRNA prediction algorithm. Further analysis includes expression profile filtering using Meyers' criteria [71], and target gene prediction using the TAPIR software [90]. 199 candidate miRNA were discovered in this study, demonstrating that sequence-based miRNA prediction methods are applicable to deep sequencing data sets.

3.2.4 Analysis of the miRDeep2 miRNA classification pipeline

Within the field of NGS-based miRNA prediction, miRDeep2 has emerged as the *de facto* standard prediction method. As described previously, the majority of miRNA prediction on NGS data sets is performed using the miRDeep2 pipeline. A recent independent review of the field also recommended that miRDeep2 be used for NGS-based miRNA prediction [91]. Considering its wide adoption, in this section, we briefly described the miRDeep2 algorithm for prediction of miRNA.

The miRDeep algorithm can be dissected into two primary steps: a pre-processing step wherein NGS read data are mapped to a genome and candidate pre-miRNA sequences are extracted from the genome at read loci; and a scoring step wherein candidate pre-miRNA are given a numerical score based on the structural stability of the pre-miRNA sequence, and expression profile within the miRNA sequence (as described below).

The pre-processing step of the miRDeep algorithm first identifies candidate mature miRNA by identifying local read depth maxima in the genome which match the expected sequence length of a mature miRNA (18 to 25nt). By default, only the 50,000 deepest read stacks are considered during the selection of candidate mature miRNA. The mature miRNA sequence is then extended in each direction twice: once by 70nt in the 5' direction and 10nt in the 3' direction; and once by 10nt in the 5' direction and 70nt in the 3' direction. The two resulting

sequences represent candidate pre-miRNA for which the mature miRNA rests on the 3’ and the 5’ arm of the hairpin, respectively.

Candidate pre-miRNA sequences are then filtered based on the following criteria:

1. RNAfold must predict a hairpin structure for the candidate pre-miRNA sequence which contains no bifurcations.
2. The miRDeep2 pre-processing algorithm attempts to identify a mature miRNA, miRNA*, and loop product within the candidate pre-miRNA sequence. If this attempt fails, the candidate pre-miRNA is rejected. The miRNA* sequence is defined as the sequence which pairs to the candidate mature miRNA sequence in the predicted RNAfold [92] structure, taking into consideration a 2nt overhang on the 3’ end of each sequence in the duplex. The loop product is defined as the subsequence of the candidate pre-miRNA which is between the mature miRNA and miRNA* products.
3. At least 60% of the bases in the stem region of the candidate pre-miRNA must be paired.
4. At least 90% of the NGS reads in the pre-miRNA sequence region must match a Dicer product. A match to a dicer product is defined as a read which aligns to the miRNA sequence with a starting 5’ position within +/-2 nt of the candidate mature miRNA, miRNA*, or loop region, and a terminating 3’ position within +/-5 nt of the same region.
5. The length of the mature miRNA and miRNA* sequences must match to within 6nt.

Candidate pre-miRNA sequences which do not meet these criteria are discarded prior to scoring.

One weakness of the miRDeep pre-processing algorithm is its inability to predict large or small miRNA. The minimum length of a pre-miRNA sequence as determined by miRDeep is 98nt, and the maximum length of a pre-miRNA sequence is 105nt. Only 10.7% of miRNA

within miRBase v21.0 fall within this length range; in this respect, the miRDeep pre-processing algorithm does not accurately reflect the biogenesis of miRNA.

The numerical score given to a candidate miRNA is a simple rules-based algorithm. No mention of rigorous training or testing is presented in the miRDeep manuscript; therefore it can be assumed that the rules were developed and tuned by hand, with the goal of optimizing performance across training data sets. Details of the miRDeep scoring algorithm are not described in the miRDeep manuscript; only the features used are described [52]. The score assigned to a candidate miRNA is the sum of the following five terms:

1. The candidate miRNA is given a starting score of -6.
2. The candidate miRNA's score is increased by 0.5 for each read which matches one of the candidate miRNA's Dicer products, using the matching rules described previously in pre-processing filter 2. If no reads match to the miRNA* region, the contribution for this score is limited to 6.
3. The candidate miRNA's score is adjusted by the log odds of the following probability distributions: P(MFE of the candidate miRNA structure is derived from a distribution of known miRNA MFEs) and P(MFE of the candidate miRNA is derived from a distribution of background training sample MFEs).
4. If any reads within the miRNA region match to the star region as per miRDeep2's Dicer processing rules, +3.9 is added to the score. If not, 1.3 is subtracted from the score.
5. If the randfold algorithm [93], using default parameters, finds the MFE of the miRNA structure to be significant ($p \leq 0.05$), +1.6 is added to the score. If not, 2.2 is subtracted from the score.

Optional scoring parameters are as follows:

6. If the seed sequence of the candidate mature miRNA (defined by miRDeep2 as the six nucleotides at the 5' end of the mature miRNA sequence) matches a known

mature miRNA exactly, +3 is added to the score. If not, 0.6 is subtracted from the score.

7. The number of paired bases in the lower stem portion of the miRNA is counted, and the score is adjusted based on the number of pairs present. The lower stem portion of the miRNA is defined as the 10nt-length duplex which is directly adjacent to the mature miRNA duplex, opposite the loop region. Figure 3 in section 2.1.5 of this thesis provides a visual representation of the lower stem portion of the miRNA.

Any candidate miRNA with a final score ≥ 0 is highlighted by the miRDeep2 algorithm as a true (predicted) miRNA.

The primary weakness of the miRDeep scoring algorithm, in our estimation, is the unbound contribution of term 2, reflecting the read depth of the miRNA sequence. In a modern NGS experiment, tens of thousands of reads map to single miRNA. As a result, term 2 of this scoring algorithm solely dictates the classification of many high-abundance candidate miRNA. Furthermore, as NGS technology improves and read depth continues to increase, the contribution of this term will increase accordingly. For a given decision threshold, the decision boundaries of the miRDeep scoring algorithm will shift relative to the average read depth of the experiment. In turn, increasing average read depth increases the false positive rate of the algorithm.

3.3 Examination of the state of the art of de novo miRNA prediction

In this section we examine 24 published methods for *de novo* miRNA prediction. These methods were published between 2005 and 2014. The major elements of the classification pipeline – data set generation, feature set generation, classifier selection, training methodology, and reporting of results – are studied in order to represent the state of the art of miRNA prediction from the perspective of pattern classification, and to highlight elements of the state of the art which require improvement.

Previous reviews of the field of *de novo* miRNA prediction [53], [94]–[97] have defined the major challenges of the field as: data set generation, specifically the generation of negative data sets; classifier selection and training, with an emphasis on class imbalance correction; and feature set selection. Improvement in these areas is defined as the path toward improved prediction performance. In this review, we examine the state of the art of miRNA prediction methods with respect to each of these areas. We also examine the reporting of results in miRNA prediction studies, as we feel that improvements in, and standardization of, performance reporting would result in miRNA prediction methods which are more appropriate for real-world applications.

3.3.1 Data set generation

In general, positive training data set generation is not a major focus of miRNA prediction methods. Positive training data sets throughout nearly all studies consist of pre-miRNA sequences which are drawn from the miRBase database. One exception is the mirnaDetect method, whose major novel contribution is an improvement to training set selection [98]. This method demonstrates a substantial improvement in performance through training set selection.

Conversely, *de novo* miRNA prediction studies employ a variety of methods and standards for generating negative training and test data. In general, this negative training and test data consists of sequences which form miRNA-like hairpin structures, as predicted by an RNA folding package such as RNAfold [92], Mfold [99], or UNAfold [100]. Negative data are extracted from annotated functional genomic regions, commonly coding regions, as these regions likely do not produce miRNA. ncRNA which share structural similarity to miRNA are also commonly used as negative training and test data.

Nam *et al.* [101] generated negative training data from chromosomes 16 through 19 of the human genome, based on RNAfold structure prediction and the following criteria: sequence

length between 64 and 90 nt, stem length (number of pairs of bases in the miRNA hairpin stem; see Figure 1) above 22 nt, bulge size (number of unpaired bases in the miRNA hairpin stem; see Figure 1) under 15 nt, loop size between 3 and 20 nt, and a minimum free energy of folding (MFE) of at most -25 kcal/mol. These strict criteria were not widely adopted.

The negative data set of Sewer *et al.* [102] in their 2005 study consisted of random subsequences of tRNA, rRNA, and mRNA. The size of their data set and length of the RNA subsequences were not specified.

Xue *et al.* [103] in their 2005 study generated negative training data from within known human coding regions. Sequences were considered to be miRNA-like if they formed single-loop hairpin structures with a MFE of at most -15.0 kcal/mol, and at least 18 paired bases in the hairpin stem. These MFE and paired base criteria were chosen because all known human miRNA at the time fell within these criteria. The resulting data set, consisting of 8494 hairpins, has become a standard data set on which miRNA prediction methods are trained and tested to this day. The MFE and bp criteria introduced by Xue *et al.* have similarly become a *de facto* standard for validation of miRNA-like hairpins. The data set introduced by Xue *et al.*, or subsets thereof, was used in 11 of 21 studies published afterward. The hairpin criteria introduced by Xue *et al.* were the basis of hairpin extraction in a further three studies.

In 2009, Batuwita *et al.* [104] extended the 8494 coding region data set with 754 ncRNA sequences. These ncRNA sequences were used by seven of the 13 studies published after 2009.

Yousef *et al.*, in studies from 2006 and 2008 [105], [106], predict hairpins from within 3'-UTR regions of genes as annotated by UTRdb [107] as opposed to predicting hairpins from within coding regions or ncRNA. This practice was only adopted by two studies since 2006 [108], [109], and one of these studies used 3'-UTR in addition to coding region data [109].

Gudys *et al.* [110] in 2012 introduced an alternative to hairpin prediction for negative data set generation, wherein sequences are randomly selected from genomes and mRNA. Sequences are selected such that the length distribution matches that of known miRNA. This method produces data sets which contain some sequences which do not resemble miRNA structurally, and as a result, has not been adopted by any other miRNA prediction studies.

The majority of miRNA prediction studies use data sets which are specific to a species. 21 out of the 24 examined studies use data sets which are specific to human. Six of the 24 studies use data sets which are specific to other model species of interest. Multi-species positive data sets are common, appearing in 14 studies. Multi-species negative data sets are far less common, appearing in 6 studies [98], [105], [110]–[113].

Class imbalance is present in all data sets, favoring the negative class. This is consistent with real-world miRNA prediction, where pseudo-hairpins outnumber miRNA. However, while actual real-world class imbalances are expected to be as high as 5000:1, typical class imbalances used in miRNA prediction studies are on the order of 1:10, with some data sets approaching even (1:1) artificially "balanced" class ratios [87], [88], while the most extreme observed class imbalances are between 1:100 and 1:200 [105], [110].

3.3.2 Classifier selection and training

De novo miRNA prediction has been performed using a variety of classification methods including support vector machines, hidden Markov models, naïve Bayes methods, random forests, random walk rankings, one-class clustering methods, K nearest-neighbor classification, and linear dimensionality reduction. Table 3 lists the classifier types used by each of the studies. Support vector machines are the most commonly used classifier, and are present in 13 of the 24 studies. Random forests and hidden Markov models are the second and third most popular classifiers, with four and three appearances respectively. K nearest-neighbor is used in two studies, once as a member of a multi-classifier ensemble

method [113] and once as a proposed one-class clustering alternative to standard miRNA prediction [105]. No other classification model is used by more than a single study. Notably, no miRNA prediction study has employed artificial neural network classifiers.

Cross-validation is the dominant method of training for miRNA predictors. 16 of the 24 studies employ cross-validation in order to train classifiers. Five- and ten-fold cross-validation are both common, being used by 8 and 6 studies respectively. The second most common training methodology is a simple hold-out test using one training data set and one testing data set. This method is used by five studies. Five-fold boosting and out-of-bounds estimation are each used by one study ([108] and [88], respectively). Additionally, four studies use the LibSVM parameter grid search in order to optimize support vector machine hyper-parameters. This grid search uses cross-validation over training sets in order to estimate parameter performance for a range of hyper-parameter values.

As previously stated, miRNA prediction is performed on data sets which contain significant class imbalance. Several methods are employed in order to train classifiers on these imbalanced data sets, including asymmetrical misclassification penalties, undersampling of negatives, synthetic minority oversampling technique (SMOTE) [43], bagging, and one-class prediction. Simple undersampling is the most common technique, and is used by 11 of the 24 studies. SMOTE is used to balance classes in two studies, and one-class prediction is used in two studies. Outside of these, no class correction technique is used by more than one study. Six studies do not describe any kind of explicit class correction.

3.3.3 Feature Selection

Features for *de novo* miRNA prediction can be broadly classified as sequence features, structure and thermodynamic features, and global or intrinsic features. The sequence category of features contains sequence motifs and *k*-grams (words of *k* consecutive nucleotides). The structure category of features contains metrics related to the predicted

RNA hairpin structure of a miRNA sequence, such as minimum free energy (MFE). Global or intrinsic features are sequence-wide features such as GC-content.

Early feature sets consisted largely of sequence and structure motifs (see below), global features, and stem and loop metrics such as number of unmatched nucleotides in the stem or size of the terminal hairpin loop. A commonly used sequence/structure motif feature set is the triplet feature set, which contains 32 features, each representing the prevalence of a nucleotide along with the pairing pattern of the nucleotide alongside its two flanking nucleotides. For example, the triplet *A.((* represents the nucleotide A, whose 5' neighbour is unpaired, who itself is paired, and whose 3' neighbour is paired. This feature set was introduced by Xue *et al.* [103].

Minimum free energy (MFE) was introduced as a feature by Jiang *et al* in 2007, and has since gained prominence as a defining feature of miRNA. Since 2007, the majority of studies have used MFE prominently in their miRNA prediction methods. Recently, Lertampaiporn *et al.* have debated the usefulness of MFE features [113], promoting alternative features based on genetic robustness. MFE still plays a role in their prediction method, however.

Table 3 - Classifier selection and training experiments for 24 miRNA prediction methods

Year	Lead Author	Classifier Type(s)	Training Experiment(s)	Class Correction
2005	Nam	paired HMM	5-CV	None
2005	Sewer	SVM	Single training set	asymmetrical misclassification penalties
2005	Xue	SVM	Single training set	Undersampling of negatives
2006	Yousef	Naïve Bayes	Single training set, 5-CV	Undersampling of negatives
2007	Ng	SVM	LibSVM grid parameter search, 5-CV	None
2007	Jiang	RF	OOB estimation	Undersampling of negatives
2008	Yousef	SVM, Naïve Bayes, One class (K-means, Gaussian, PCA, KNN)	10% holdout repeated 20 times	Undersampling of negatives
2008	Xu	Random Walk ranking	One-class systems built using 1 - 50 positive training samples	None
2009	Kadri	HHMM	10-CV	None
2009	Oulas	Profile HMM	Boosting (5-fold) on positive only	Trained positive-only
2009	Batuwita	SVM	outer 5-CV	SMOTE to correct training imbalance
2010	Mathelier	Simple filters	Manual optimization	None
2010	Ding	Ensemble SVM	Outer 3-CV	Ensemble trained on partitioned negatives
2010	Zhao	SVM	Parameter testing from LibSVM	Undersampling of negatives
2011	Han	SVM	5-CV	Undersampling of negatives
2011	Wang	SVM	5-CV	Undersampling of negatives
2011	Xiao	RF	10-CV	Undersampling of negatives
2011	Xuan	SVM	5-CV	Undersampling of negatives
2012	Liu	SVM	Single training set	Undersampling of negatives
2013	Gudys	SVM	stratified 10-CV	ROC-select
2013	Lertampaiporn	Ensemble (4x SVM, 4x RF, 4x KNN)	10 x 5-CV	SMOTE
2013	Shakiba	Linear dimensionality reduction	Parameter grid searches, 10-CV	LDR
2013	Wei	SVM	LibSVM grid parameter search, 10-CV	Undersampling of negatives
2014	Zou	hierarchical RF	10-CV	Bagging through RF

Early miRNA prediction studies make little use of systematic feature selection methods. Between 2005 and 2008, no systematic feature selection methods were used, outside of testing the inclusion of MFE in a feature set by Jiang *et al.* [88]. Over time, systematic feature selection methods became more popular. This is largely due to the inclusion of many MFE-based features and Z-features. The high number of sequence and structure motifs used in studies also necessitates systematic feature selection methods. Since 2009, 11 of 16 studies made use of systematic feature selection methods, including F-score [109], clustering [114], information gain [114], genetic algorithms [113], [115], [116], SVM weight measurement [117], floating forward search [118], and correlation feature selection [113]. No single feature selection method dominates in the field of miRNA prediction.

Out of the 24 studies, Batuwita *et al.* [104] were the first to employ a systematic feature selection method. Their initial feature set contains 29 global and intrinsic features as introduced by Xue *et al.* [103], and 19 new features which contain several thermodynamic metrics. After feature selection, a set of 21 features remains. Much like the data sets generated by Batuwita *et al.*, this set of features appears commonly in future studies, with five studies using these 21 features in their feature sets.

Feature sets for miRNA prediction typically contain between 20 and 40 features. 14 out of 23 feature sets lie within this range. Three studies contain very large feature sets which include at least 1000 sequence or structure motifs [105], [106], [117]. Two studies use feature sets which contain fewer than 10 features [112], [118]. One study employs a feature set containing 98 features [98]. Finally, three studies do not report feature set size [87], [108], [119].

Table 4 - Feature set size and selection methods for 24 miRNA prediction methods

Year	Author	Total # features	Final # features	Feature selection method
2005	Nam	N/A	N/A	N/A (HMM)
2005	Sewer	40	40	None
2005	Xue	32	32	None
2006	Yousef	> 1000	> 1000	None
2007	Ng	29	29	None
2007	Jiang	32+2	32+2	Tested addition of MFEs
2008	Yousef	> 1000	> 1000	None
2008	Xu	36	36	None
2009	Kadri	NA	NA	Two alphabets tested
2009	Batuwita	48	21	Subset selection methods
2009	Oulas	Unknown	Unknown	Unknown
2010	Mathelier	5	5	None
2010	Ding	65	32	F-score
2010	Zhao	36	36	None
2011	Han	48	27	Clustering, Info gain
2011	Wang	124	20	GA
2011	Xiao	24	24	None
2011	Xuan	48	22	GA
2012	Liu	29734	1300	Highest SVM weight
2013	Gudys	28	28	None
2013	Lertampaiporn	103	20	3-CV test on many methods. Correlation feature selection + GA chosen.
2013	Shakiba	48	7	floating forward search
2013	Wei	98	98	Subset testing. Deciding on full set
2014	Zou	Unknown	Unknown	RF random feature selection

3.3.4 Reporting of Results

As with many pattern classification problems, the most highly reported results in the field of miRNA prediction are those of sensitivity (Sn) and specificity (Sp). These two metrics are reported almost universally among miRNA prediction studies, appearing directly in 21 of 24 studies. The second most commonly used performance metric within the field is accuracy (Acc). Acc appears in 11 studies [88], [98], [103], [108], [109], [111], [112], [117], [119]–[121], and in two of the studies which do not report Sn and/or Sp directly [103], [119].

One study makes use of the information retrieval metrics of precision and recall [122] in place of the binary classification metrics Sn and Sp.

Beginning with the microPred study by Batuwita *et al.* [104], a shift occurred toward reporting the geometric mean (GM) of sensitivity and specificity as the primary performance result of a miRNA prediction technique. GM is insensitive to class imbalance, which, as discussed above, is substantial in miRNA prediction. Therefore it was argued that GM is a superior performance metric to the previously used Accuracy metric which is highly sensitive to class imbalance. Since its introduction into the miRNA prediction field, GM has seen common usage, appearing in 7 of 13 studies in this period [98], [109], [110], [113], [114], [116], [118]. In the same period, 7 out of 13 studies reported Acc [98], [109], [112], [117], [119]–[121]. Neither Acc nor GM represents a definitive performance metric for miRNA prediction, as both are reported equally in recent studies.

Table 5 - Performance metrics reported by 24 miRNA prediction methods

Year	Author	Performance metrics reported
2005	Nam	Sn, Sp of 5-cv; # miRNA predicted and recovered from chromosomes
2005	Sewer	Viral pre-miRNA prediction results; Sn, Sp on test set
2005	Xue	Accuracy on test sets
2006	Yousef	Sn, Sp on test sets and hold-out from training sets
2007	Kwang	Sn, Sp, Acc on hold-out human and pooled
2007	Jian	Acc, Sn, Sp, MCC
2008	Yousef	Sn, Sp, MCC
2008	Xu	Precision and Recall (in ~1:2 class imbalance test data)
2009	Kadri	Sn, Sp, MCC, FDR
2009	Oulas	Sn, Sp, Average of Sn/Sp (Acc @ 1:1)
2009	Batuwita	Sn, Sp, GM
2010	Mathelier	ROC curves. AUC, Acc, MCC
2010	Ding	Se, Sp, GM, Acc
2010	Zhao	Acc, Se, Sp
2011	Han	GM, Se, Sp
2011	Wang	Se
2011	Xiao	Se, Sp, Acc
2011	Xuan	Se, Sp, GM
2012	Liu	Se, Sp, Acc, AUC
2013	Gudys	Se, Sp, GM
2013	Lertampaiporn	Se, Sp, GM
2013	Shakiba	Se, Sp, GM, Std
2013	Wei	Se, Sp, GM, Acc
2014	Zou	Acc

3.4 Previous assessments of the state of the art

Since 2009, a number of review articles have captured the ongoing state of the art of computational methods for miRNA, including *de novo* miRNA prediction. One of the common

themes among reviews in this period is a suggestion to improve negative set generation methodology. Mendes *et al.* [53] in 2009 recommend that miRNA prediction models consider more strongly "what is not miRNA". Yousef *et al.* [96] state that "defining the negative class is a major challenge in developing ML [machine learning] algorithms for miRNA identification." Li *et al.* [97] emphasize the negative class, stating "it is usually straightforward to select positive examples (e.g., taking the known miRNAs), whereas it is harder to construct negative examples." Gomes *et al.* [95] emphasize data set generation in general, saying "a careful choice of positive and negative data sets is crucial". Kleftogiannis *et al.* [94] state that negative set selection is one of the problems that needs solving in order to improve accuracy of miRNA prediction, however they feel that "the selection of pseudo hairpins is also simple as they can be downloaded from RefSeq genes".

Outside of negative set generation, there is very little agreement on future directions for miRNA prediction. Standardization of data sets, and public availability of data sets, is highlighted in two studies [94], [95]. Mendes *et al.* [53] suggest that a better understanding of miRNA biogenesis is needed for improved prediction performance. Kleftogiannis *et al.* [94] suggest that miRNA prediction should be integrated into larger pipelines which have web interfaces. Several studies suggest improvements to feature selection [94], [97]

3.5 Discussion of the miRNA prediction state of the art

In this section we examine several outstanding issues with the state of the art of miRNA prediction, which have been reached through our independent assessment of the field. Many of these issues stem from the pattern classification approaches chosen by miRNA prediction studies, as described above.

3.5.1 Redundancy in feature sets

Feature sets for miRNA prediction are often built using a scattershot approach wherein many similar features are proposed and feature selection methods are used to reduce the length of the final feature vector. For example, feature sets often contain many features which relate to hairpin structure MFE. This technique has produced highly redundant feature sets such as the commonly used 21-feature set which formed the basis of the microPred classifier. This feature set contains many features which are highly correlated; Figure 8, for example, demonstrates the correlation between microPred MFE features 1 and 2.

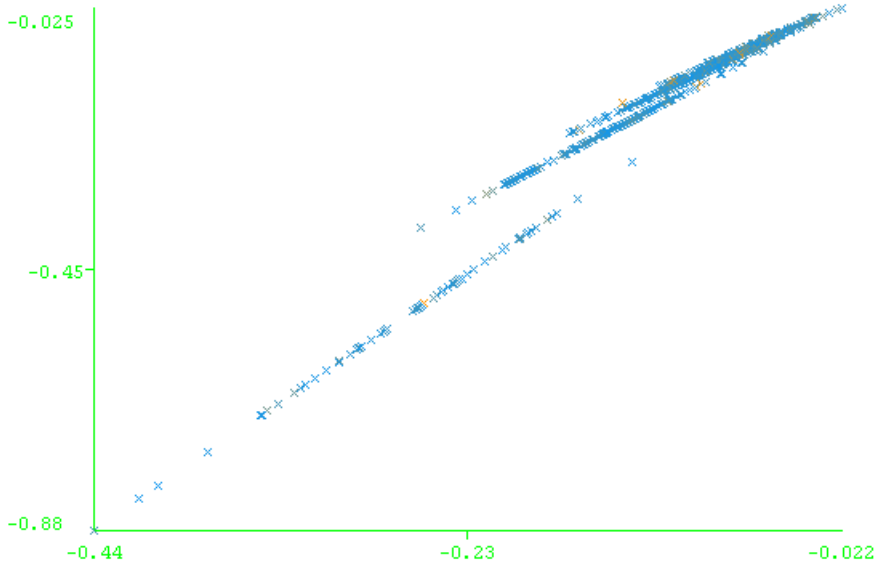


Figure 8 - Correlation between features MFE1 and MFE2 in widely used microPred feature set. Data from microPred positive training set.

We believe that the dominance of SVM and RF for miRNA prediction is partly due to the presence of highly correlated features such as the widely-used MFE features. At the same time, standard feature sets are likely suboptimal due to the presence of redundant features. We recommend that future miRNA prediction studies take into account feature redundancy when choosing feature sets, as this may open the door for more accurate classification using methods which underperformed previously. Explicitly handling redundant features may also

lead to improved performance using existing methods, as redundant features reduce classification performance in the general sense regardless of classifier used [40].

3.5.2 Lack of prevalence-corrected reporting

The estimated real-world class imbalance when predicting miRNA within eukaryote genomes is on the order of at least 1:1000. For example, 2588 *H. sapiens* miRNA have been discovered, while the number of hairpins in the *H. sapiens* genome which meet common miRNA criteria is approximately 11 million. Because of this extremely high class imbalance, false positive rates of classifiers have a high impact on real-world performance. At 1:1000 real-world class imbalance, a classifier which operates at 90% specificity can achieve at most a 1 in 101 success rate on its predictions (i.e. Pr < .01). Realistically, any miRNA prediction experiment must be performed at very high (>99.9%) specificity in order for experimental validation of predictions to be feasible.

The most commonly reported performance metrics among miRNA prediction studies – sensitivity, specificity, test-set accuracy, and geometric mean – do not account for real-world class imbalance. Sensitivity and specificity ignore class imbalance. Test-set accuracy reflects performance at the class imbalance of the test set, and not the real-world class imbalance. Because test set class imbalance for all miRNA predictors is several orders of magnitude lower than real-world class imbalance, optimal accuracy occurs at an operating point which is not optimal for experimental validation.

In regards to real-world applicability, geometric mean – which was introduced as an improvement on accuracy for miRNA prediction performance reporting – is much worse than test set accuracy. Geometric mean disregards class imbalance completely, and assigns essentially equal importance to Sn and Sp even though Sp clearly has a greater impact on overall system performance, given the prevalence of the negative class. When optimizing for geometric mean, the ideal classifier performance occurs when sensitivity and specificity

are equal. For miRNA prediction, harmony between sensitivity and specificity is not a reasonable operating point. Small (1%) changes in specificity have a large impact on success rates of experimental validation techniques (*i.e.* precision). Conversely, high sensitivity is not a strict requirement for the successful prediction of miRNA: recovering even 50% of all miRNA within a genome would be a hugely successful result for a miRNA prediction study. For these reasons, current performance metrics for miRNA prediction are inadequate.

In order to encourage the production of classifiers which are tuned for the prediction of miRNA in real-world scenarios, we believe that the field of miRNA prediction must shift toward prevalence-corrected performance reporting, *i.e.* the reporting of results at the expected class imbalance of real-world data sets. Prevalence-corrected reporting is not currently used to report performance in any miRNA prediction study.

In particular, prevalence-corrected precision and recall curves - common in the field of information retrieval where large class imbalances are often seen [123] – are well suited to miRNA prediction. These curves plot the recall of a classifier (synonymous with sensitivity) against the (prevalence-corrected) precision. Unlike specificity, which measures the chance of reporting a pseudo-miRNA as negative, precision measures the chance of a positive classification being truly positive. Precision therefore informs a user of miRNA prediction software of the success rate one can expect when conducting experimental validation of miRNA predictions. The combination of recall and precision allows the user to weigh the cost of validation against the number of expected miRNA predictions made. None of specificity, accuracy, or geometric mean provides a measure of performance of classification on pseudo-miRNA which is directly useful to a user of miRNA prediction software.

As summary statistics, recall at a precision of 50% (Re@Pr50) describes the performance of a classifier at an experimental success rate which is acceptable for validation experiments. Recall at 90% precision (Re@Pr90) provides a more conservative estimate of classifier

performance which includes only miRNA which are predicted with extremely high confidence.

The use of prevalence-corrected performance metrics is necessary for the development of optimal miRNA predictors. Currently, classifiers underperform at real-world class imbalances because they are tuned for unrealistic performance metrics.

3.5.3 Independent analysis of the effectiveness of SMOTE class imbalance correction

Recent publications have shown conflicting results regarding the effectiveness of the SMOTE class imbalance correction method for miRNA prediction. Batuwita and Palade demonstrate an increase in performance when SMOTE is used to correct class imbalance during the training of the microPred classifier [104]. However, Gudys *et al*. found that the optimal training pipeline for their ROC-select meta-classifier does not include SMOTE [110]. We believe that class imbalance correction increases performance of miRNA classifiers; however the metric of geometric mean does not elucidate this increase in performance because of its narrow focus. Figure 9 and Figure 10 detail the performance of two miRNA predictors which are trained identically except for the inclusion or exclusion of SMOTE for class imbalance correction. These classifiers were trained using the microPred data set and feature set and the libsvm classification library [124], then applied to a hold-out data set consisting of 282 *A. carolinensis* miRNA from miRbase 19 and 500 *A. carolinensis* coding region pseudo-miRNA hairpin loops.

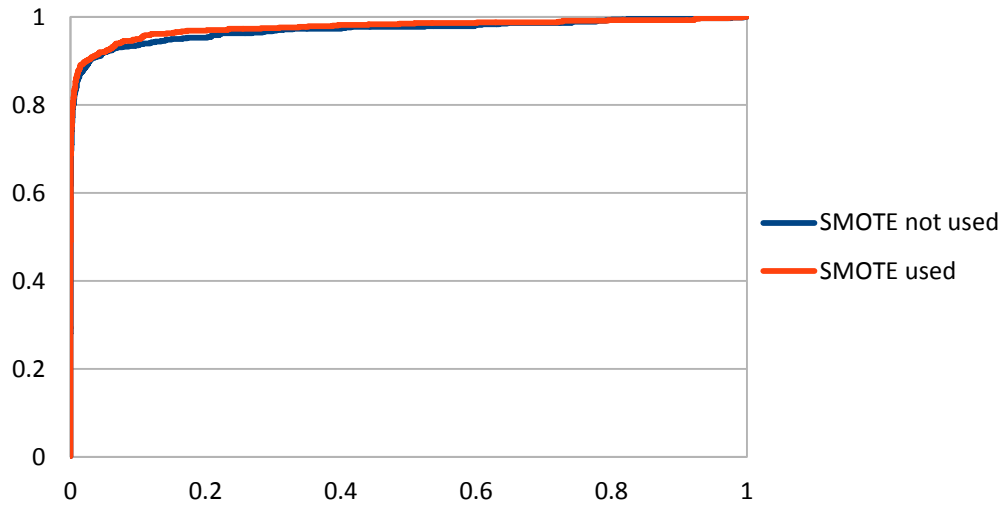


Figure 9 - ROC curve demonstrating classifier performance when SMOTE is used and when SMOTE is not used for class imbalance correction

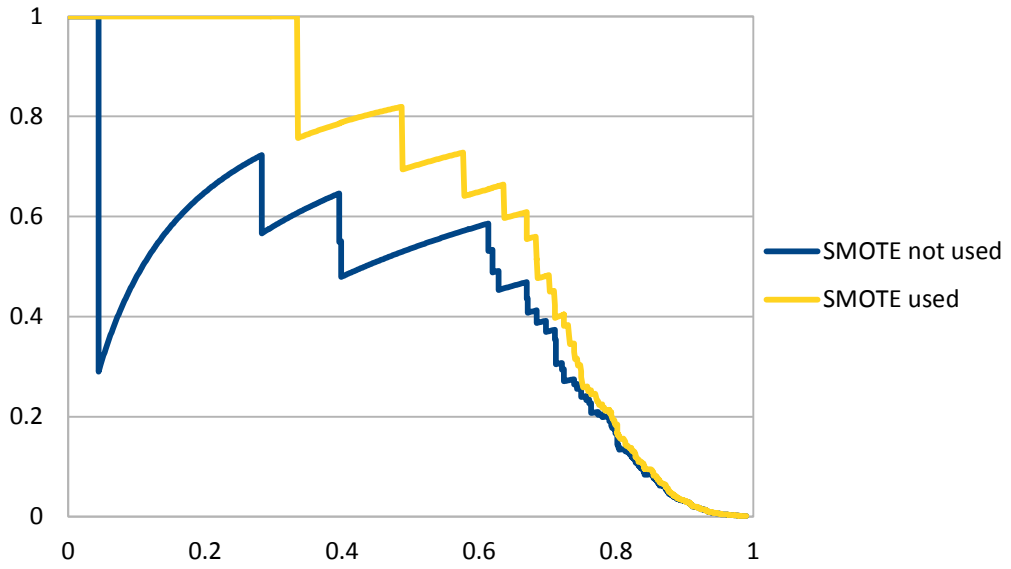


Figure 10 - PR-curve corrected for 1:1000 class imbalance demonstrating classifier performance when SMOTE is used and when SMOTE is not used for class imbalance correction

In general, performance is increased with the addition of SMOTE. At acceptable real-world precision levels, Re@Pr50 is increased by 11.4% and recall at high precision is increased 7-

fold. However, the peak geometric mean for the classifier – the metric employed by Gudys' ROC-select classifier – shows almost no improvement when SMOTE is used. Peak GM is 0.938 when SMOTE is used, and 0.937 when SMOTE is not used. In spite of increasing overall performance of the classifier, SMOTE would be disregarded in this case by a study which uses peak GM as a metric, as it increases classifier complexity without improving the primary performance metric.

This experiment demonstrates the narrow applicability of geometric mean as a miRNA prediction metric, while also providing evidence in favour of the use of SMOTE for class imbalance correction in the field. It is therefore used in all experiments in chapters 5 through 7 of this thesis.

3.5.4 Failure of miRNA predictors to generalize to cross-species negative data

The most commonly used metric for performance of miRNA prediction methods in recent years is classifier GM as reported on 10-CV test data. As shown in the previous section, optimizing for GM could potentially produce classifiers which perform sub-optimally in the general sense. 10-CV experiments use a single data set for training and testing, therefore iteratively optimizing one's 10-CV results is a potential source of overfitting. At the same time, emphasis is placed within miRNA prediction on the ability of predictors to achieve high cross-species recall – with no regard for cross-species specificity [104], [110]. Because of the narrow focus of miRNA prediction performance metrics, miRNA predictors often fail to generalize to negative data sets outside of those on which they are trained.

Figure 11 shows the results of the microPred classifier on independent data sets which represent a range of species across taxons. While sensitivity is maintained across all species, sensitivity is low, especially on non-human species. Specificity on four of the seven test sets is below random, demonstrating a complete failure of the classifier to generalize.

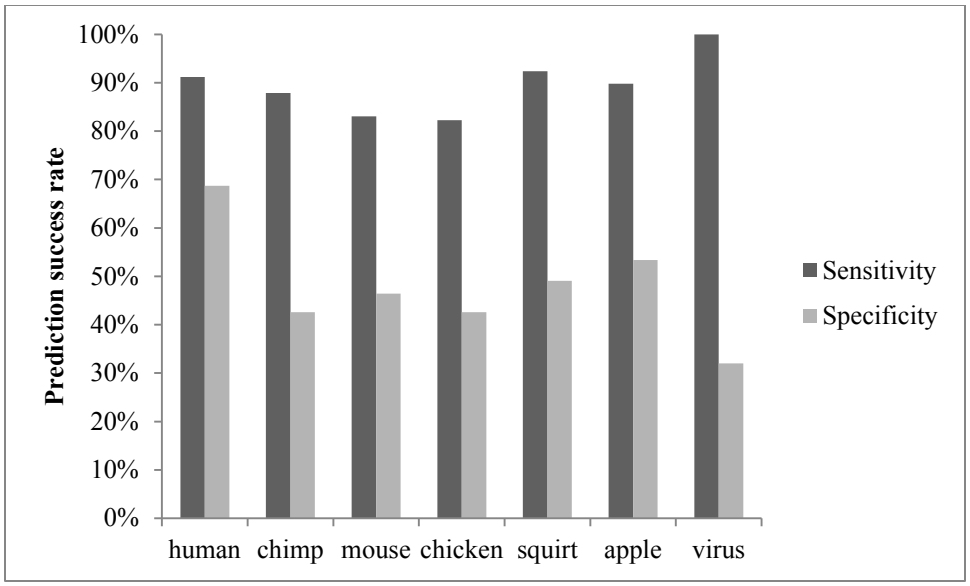


Figure 11 - microPred classification results on independent hold-out data sets representing multiple species

3.5.5 Moving to genome-scanning data sets for genome-scale experiments

Within the field of *de novo* miRNA prediction, many studies have emphasized the development of high-quality negative sets and classification performance within the field has benefitted from this increase in data set quality. Data set selection has proven to be a crucial step in the miRNA prediction pipeline. However, little attention has been paid to the methodology for selecting positive data. MiRNA prediction studies use as positive training data known miRNA from databases such as miRBase or miRTarBase, which contain experimentally validated miRNA sequences. Negative training data are miRNA-like hairpin structures which were extracted from larger genomic regions such as annotated coding regions. In other words, the pre-miRNA structures of positive data were created *in vivo* by Drosha, while the pre-miRNA structures of negative data were created *in silico* by an RNA folding algorithm. As illustrated in Figure 12, *in silico* RNA folding algorithms and *in vivo* RNA cleaving can produce different pre-miRNA sequences which each contain the mature miRNA.

Figure 12 - RNA sequences representing miRNA has-mir-451a. The sequence which was cleaved by Drosha and the sequence which was predicted by RNAfold differ in their start and end positions, however both contain the mature miRNA.

No real-world experiments use a combination of positive data which are cleaved by Drosha and negative data which are computationally generated. For the prediction of unannotated genomic data, which we feel is the most pertinent use case of *de novo* miRNA prediction, all test data (*i.e.* both miRNA hairpins and pseudo-miRNA hairpins) are extracted in a single *in silico* genomic scan which produces candidate regions corresponding to computationally predicted hairpin structures. Similarly, in the realm of NGS-based miRNA prediction, both miRNA hairpins and pseudo-miRNA hairpins (other types of transcribed RNA which form miRNA-like hairpins) are cleaved *in vivo* and the resulting expressed reads are examined.

MiRNA prediction models have been trained to differentiate between miRNA which are cleaved *in vivo* by Drosha and pseudo-miRNA which are cleaved *in silico* by RNAfold. The predictive power of these models is in part based on the difference between miRNA biogenesis and computational RNA secondary structure prediction. For reasons stated above, this predictive power does not apply to real-world data sets.

In combination with the previously described issues regarding class imbalance and lack of generalization, the lack of genome- or transcriptome-specific data sets should be addressed in order to provide useful, realistic *de novo* miRNA prediction pipelines. In particular, the need for *de novo* analysis of unannotated genomes is an important use case because these

studies are complementary to deep sequence specific studies using methods such as miRDeep. Some recent studies have focused on the development of miRNA prediction in genomic data sets [125]–[127], however these studies do not address any of what we feel are the most important aspects of genome-wide miRNA prediction. Training data sets contain a mixture of Drosha-derived positives and computationally derived negatives; performance is measured at unrealistic class imbalances; and performance on negative sets outside of training species is not examined. An important step in the advancement of the state of the art of miRNa prediction is the development of genome-scanning miRNA prediction methods which properly address these issues. A major step toward this goal is to observe performance across the complete genomic miRNA prediction pipeline. As shown in Figure 13, both extraction of miRNA-like hairpins and classification of these hairpins occur within the prediction of a genomic data set, and therefore both of these processes directly affect miRNA prediction performance.

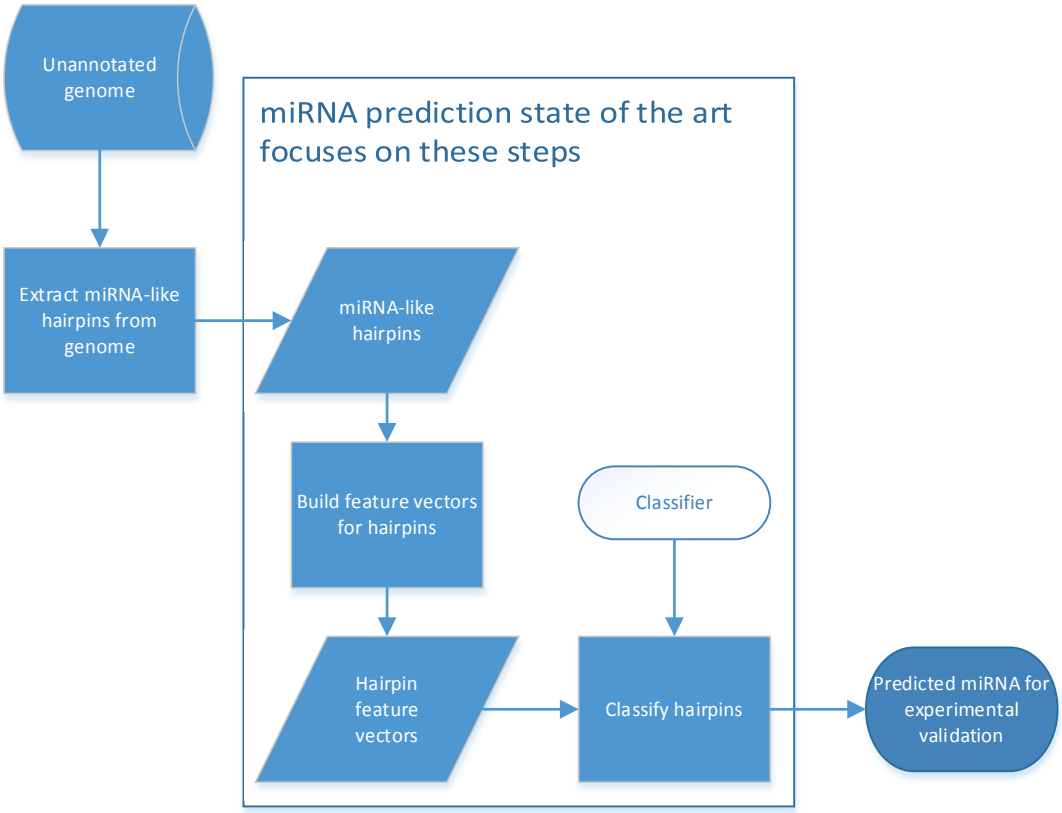


Figure 13 - The prediction pipeline for miRNA within an unannotated genome

4 Problem statement

The purpose of this thesis is to increase number of miRNA which are recovered by miRNA prediction experiments, for fixed success rates of experimental validation procedures. The following issues with current miRNA prediction techniques are addressed in order to achieve this goal:

1. *De novo* miRNA prediction methods fail to generalize to hold-out data sets which are derived from non-model species. Recent attempts, such as taxon-wide data set generation, fail to properly address this issue. Given that a large number of miRNA prediction studies are performed on non-model and often unannotated species, the generalization of miRNA prediction methods to non-model species is an important facet of miRNA prediction performance. This is addressed in Chapter 5 of the thesis.
2. Current *de novo* miRNA prediction models are optimized for unrealistic class imbalances, resulting in high false discovery rates for genome-wide miRNA studies. At the stated specificity levels of existing miRNA prediction methods, false discovery rates in genome-wide studies are on the order of 95% or higher. The application of metrics which are appropriate to genome-wide experiments stand to improve classifier performance on real data sets. Furthermore, these performance metrics state classifier performance in terms which are applicable to users of miRNA prediction methods. This is addressed in Chapter 5 of the thesis.
3. Current *de novo* miRNA prediction studies only analyze a portion of the *de novo* miRNA prediction pipeline. Performance of these methods is measured in classifier sensitivity, specificity, and associated metrics. No regard is paid to the performance of pre-filtering methods paired to these classification models, whose sensitivity and specificity equally inform the performance of a miRNA prediction study. Models which analyze the performance of the entire miRNA prediction pipeline could identify

additional sources of specificity and sensitivity loss in miRNA prediction, relative to current models. This is addressed in Chapter 6 of the thesis.

4. NGS-based miRNA prediction methods do not incorporate the full range of evidence which is available for the classification of miRNA. Current models incorporate small numbers of sequence- or expression-based features, apply outdated vectors of sequence-based features, or focus solely on single lines of evidence for the prediction of miRNA. Incorporation of all lines of evidence provides the opportunity for optimal feature vectors for the prediction of miRNA, and as a result, prediction performance can be improved relative to state-of-the-art methods. This is addressed in Chapter 7 of the thesis.

5 A framework for improving microRNA prediction in non-human genomes

5.1 Abstract

In this chapter, we introduce a framework (SMIRP) for creating species-specific miRNA prediction systems, leveraging sequence conservation and phylogenetic distance information. This framework improves the quality of training data sets for miRNA prediction, and addresses the lack of generalization of miRNA predictors during cross-species prediction. Prevalence-corrected performance metrics are used to elucidate classifier performance at real-world class imbalances. Substantial improvements in recall and precision are obtained for four non-human test species when our framework is applied to three different prediction systems representing two types of classifiers (SVM and Random Forest), based on three different feature sets, with both human-specific and taxon-wide training data. The SMIRP framework is potentially applicable to all miRNA prediction systems and we expect substantial improvement in precision and specificity, while sustaining sensitivity, independent of the machine learning technique chosen.

This chapter has been adapted from a journal publication which has been published in the journal Nucleic Acids Research.

5.2 Introduction

As described in section 3.4.4 of this thesis, a major weakness in the field of miRNA prediction is that prediction models often fail to generalize to hold-out negative test data (see Figure 11). In particular, cross-species hold-out negative sets are absent from many studies in the field. Some previous studies have highlighted the need for training data which is appropriate for non-human species [110], [112], [113], [125], and attempts have been

made to provide models for non-human miRNA prediction. Gudys *et al*. proposed a methodology (HuntMi) for creating taxon-specific training data wherein sequences are pooled from many genomes within a broad taxon [110]. They also examined a range of pattern classification approaches, concluding that random forests were most effective for the feature set proposed in their study [110]. Wu *et al.* propose the use of similarly multi-species pooled positive data sets, however they note that this methodology does not perform well for all taxa, noting Mycetozoa as a taxon for which insufficient data are available for a pooled data set generation approach [125]. Lertampaiporn *et al.* utilize a positive training set which contains pre-miRNA sequences pooled from the genomes of *Homo sapiens*, *Arabidopsis thaliana*, and *Oryza sativa* in order to broaden the usefulness of their classifier beyond strictly human miRNA studies [113]. While the above measures represent important steps toward increasing the accuracy of miRNA prediction in diverse species, we here demonstrate that a more advanced framework for species-specific training data selection has the potential to vastly improve miRNA prediction accuracy across a range of species. This approach is particularly useful for niche species that are of great scientific interest due to their genetic uniqueness, but are phylogenetically distant from model organisms such as *H sapiens* which are typically used to create single-species or multi-species pooled training data.

While the issue of class imbalance (i.e. the large number of negative sequences vs. true miRNAs in a genome) is widely acknowledged [104], [110], [113], [128], and has been addressed during training using techniques such as SMOTE [104], [113], [129], it remains largely unaddressed in the testing and evaluation of miRNA prediction methods. Therefore, as adopted in [122], we introduce in this study precision-recall curves using real-world class imbalance levels as a means for comparing performance of miRNA prediction methods. Relative to the widely used metrics of geometric mean and Acc, precision-recall curves account for the large class imbalance observed in actual genomes (estimated at 1000:1 for

most genomes; see below). This performance metric has been adopted in other relevant fields, *e.g.* protein-protein interaction prediction [130], as it quantifies the performance of a classifier in terms that are of direct interest to actual users of the method – those who will perform follow-up experimental validation of predictors.

In this study, we present a framework for the dynamic generation of species-specific training data, suitable for the creation of highly accurate species-specific miRNA predictors. Such a method is particularly needed for newly sequenced species of biological interest and for species for which high-quality miRNA data is not already available. This framework can be applied to generate training data for any miRNA classification method, including current leading methods. Our framework leverages sequence clustering techniques in order to produce positive training data representing diverse miRNAs. Selection techniques are applied to these clusters to tailor the data set toward any species, with an emphasis on those miRNA that appear to be highly conserved. Negative data sets are built using miRNA-like hairpins from species that are closely related to the target species, providing negative training data more likely to resemble those found in the target species. We demonstrate a positive correlation between the use of training data from species which are closely related to a species of interest and classification performance on a hold-out species, providing clear evidence that our species-specific methods successfully leverage phylogenetic information for classification. We further demonstrate improved performance of two SVM-based classifiers and one random forest-based classifier for miRNA studies on reptile, insect, plant, and virus genomes. Trained species-specific miRNA prediction systems and all training and test data are freely available at http://bioinf.sce.carleton.ca/SMIRP. As previously mentioned, this approach is particularly useful for niche species that are important model organisms of study, but suffer from being phylogenetically distant from the model organisms that are typically used to create single-species or multi-species pooled training data.

5.3 Methods

5.3.1 Framework Overview

Figure 14 illustrates our proposed framework for constructing species-specific positive and negative data sets for training and evaluating miRNA prediction systems. Known miRNA sequences are first gathered from multiple species. These sequences are then clustered by identity or similarity such that redundant training data are grouped together. A single representative sequence is selected from each of the *p* largest clusters, such that the representative sequence derives from the species considered to be most closely related to the target species. For example, if studying *D. melanogaster*, one would prefer training data from other insect species to data from human. The resulting sequences are used as the positive training data. Negative data are similarly taken from one or more species deemed to be closely related to the target species, for which annotated coding regions and ncRNA are available.

Selection of source species for negative training data is performed manually based on phylogenetic information from resources such as the NCBI taxonomy browser [131] or that provided by miRBase [4]. The SMIRP framework is robust with respect to this selection in that performance is generally consistent where negative training data are selected from various species within the same family. For example, when a *D. melanogaster*-specific classifier is retrained using negative data from *D. simulans* instead of *D. pseudoobscura*, no significant or consistent increase in performance is observed (see Figure 15). For all experiments performed in this study, species selection for negative training data sets was performed manually based on the phylogenetic grouping of the miRBase database. Within the lowest ranking taxon containing the target species, the species with the highest number of known miRNA was chosen as the training species.

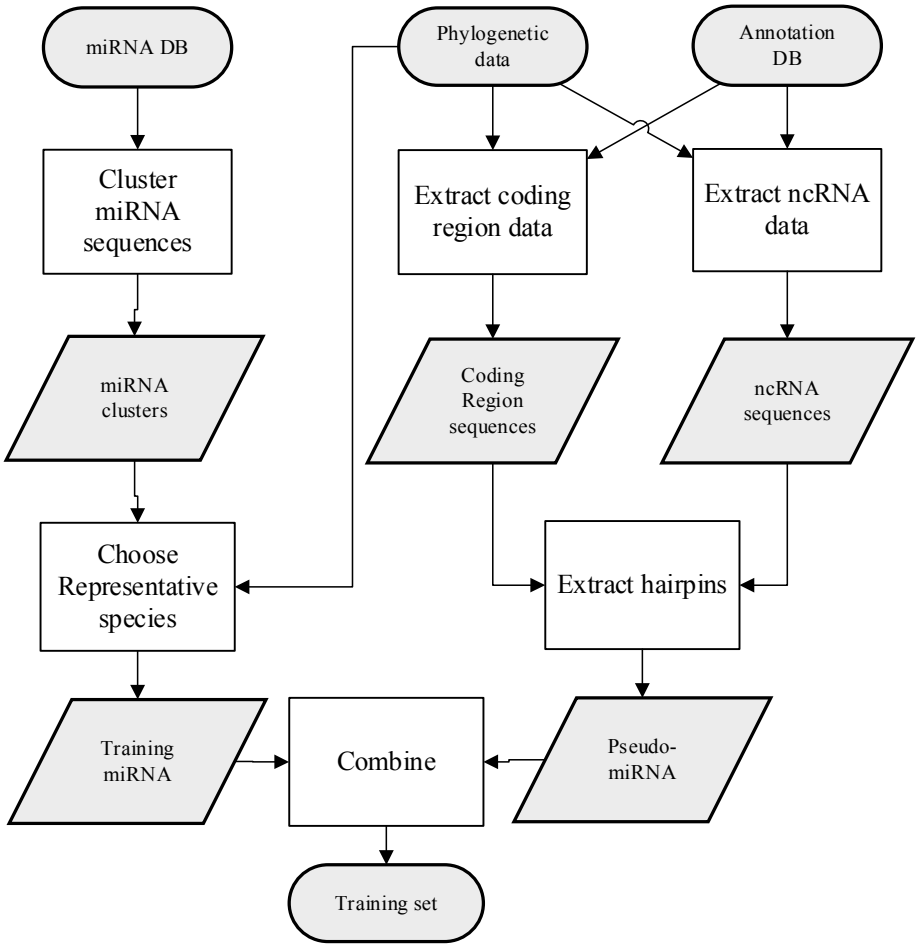


Figure 14 - Overview depiction of species-specific training data set generation framework.

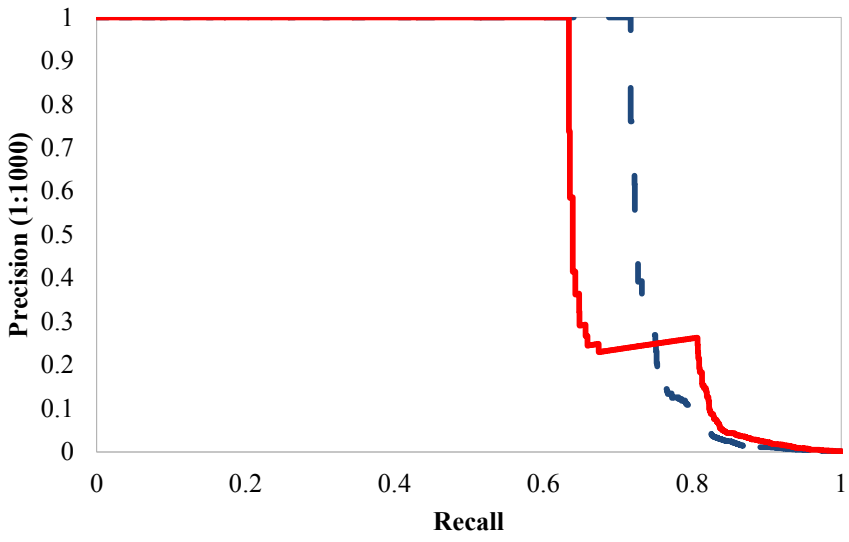


Figure 15 - Comparison of HeteroMirPred-like SMIRP classifier trained using negative training data from two different species and tested using *D. melanogaster*. In the precision-recall curve above, the dashed blue curve corresponds to *D. simulans* and the solid red line corresponds to *D. pseudoobscura*.

5.3.2 Generating Species-specific positive training sets

Species-specific positive training data sets were built using the miRBase v.19 database [4]. This database contains 20982 miRNA sequences across 193 species. Redundant sequences and sequences containing non-AGCU characters were removed from the data set, resulting in an initial training data set containing 19161 sequences. CD-hit [132] was then used to generate clusters of sequences within our initial data set, using a threshold of 80% sequence identity. Default CD-hit parameters were used for clustering.

Using the miRBase sequence data set and the clusters described above, we then developed positive training data sets for miRNA classification that were targeted toward the species of interest (referred to here as our target species). These data sets were developed as follows:

1. For a given positive integer *p*, the largest *p* clusters were chosen from the CD-hit clustering results. Larger values of *p* provide a larger number of positive data, however smaller values of *p* provide higher-quality data, representing larger families

of well-conserved miRNA. Figure 16 demonstrates that our method is largely insensitive to this trade-off. Therefore, optimization of the parameter p is not a required step in the generation of positive training sets. For the experiments below, p was set to match the number of training data used to train either microPred (p=692) or HeteroMirPred (p=1000), however larger values of p may be used in general.

2. A representative sequence was chosen from each of the p clusters. For each cluster, the representative sequence of the cluster is the sequence which is found in the species nearest to the representative species in terms of phylogenetic classification. The phylogenetic classifications given within the miRBase database were used to classify species for the purpose of representative sequence selection.
3. In the event of multiple candidate representative sequences within a cluster whose species are equally close (phylogenetically) to our target species, the candidate sequence whose length is closest to the mean length of sequences within the cluster is chosen as the representative sequence for the cluster.

The resulting positive training data set contains miRNA sequences that are highly conserved across species. Because homologs of these sequences have been verified in many species, the positive training data set also represents miRNA whose functional annotation is well studied. Importantly, no two miRNA sequences are alike with more than 80% identity; therefore the data set does not contain redundant sequences. In addition, miRNA sequences within the data set are phylogenetically similar to the target species, increasing the likelihood of conservation between training data and miRNA to be predicted in unannotated target species.

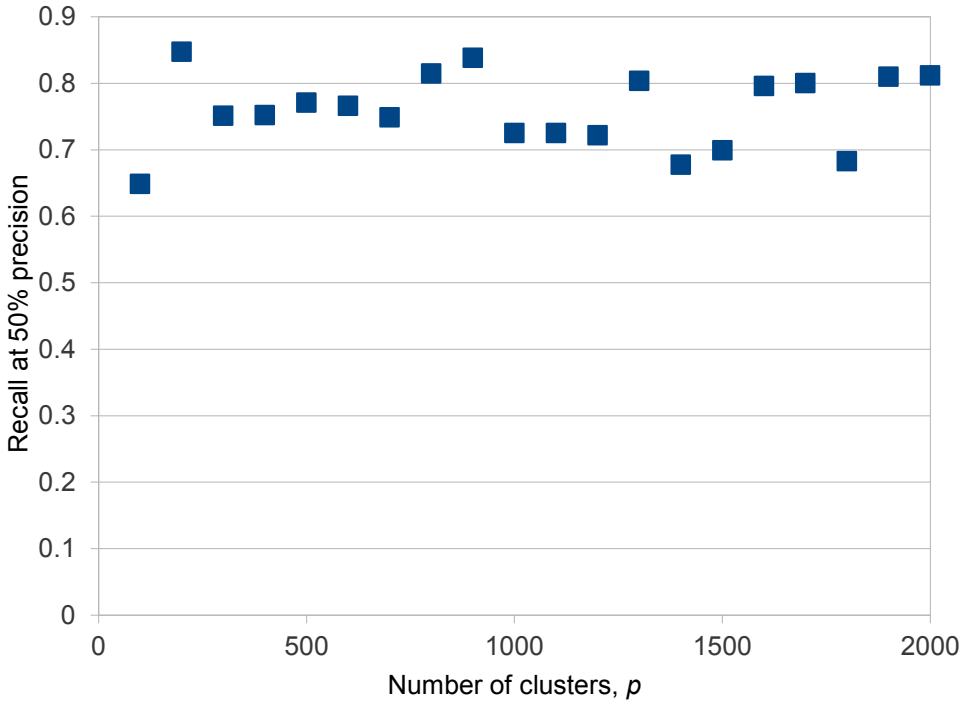


Figure 16 - SMIRP classification performance is insensitive to changes in parameter *p*, the number of positive samples used during training data set generation. Classification experiments were performed using a HeteroMirPred-like classifier and measured on *Arabidopsis thaliana* hold out test data, using negative training data from *Arabidopsis lyrata.*

5.3.3 Generating Species-specific Negative Training Data sets

Negative training data sets for a target species were generated from coding region (exonic) sequences and ncRNA sequences of species that are closely related to the target species, based on phylogenetic distance. Coding region and ncRNA sequences were retrieved from the European Nucleotide Archive [133]. Once data was retrieved, the following steps were carried out:

1. Coding and ncRNA sequence data were combined and formatted into a FASTA format which is compatible with the ViennaRNA package [134]
2. Pseudo-miRNA sequences were extracted from the coding and ncRNA sequence data. Pseudo-miRNA sequences are defined as those that fold into a hairpin structure containing at least 18 stem pairs and a minimum free energy of at most -15

kCal/mol. These folding criteria are commonly used for the determination of miRNA hairpin candidates, and are the criteria with which the microPred negative training data set was built [103], [104]. The ViennaRNA package is used to predict the secondary structure of RNA sequences.

3. Redundant pseudo-miRNA sequences were removed from the negative training set. A pseudo-miRNA sequence was considered redundant if it was a substring of another pseudo-miRNA sequence in the negative training set. The user can optionally specify that clustering should be also applied to the negative training data, as is done to the positive training data. Applying this optional step may affect prediction performance depending on the test species; all results below correspond to data sets built without using this option. A subset of candidate hairpin sequences of size *n* was chosen at random from the full list of sequences.

The resulting negative training sets contain *n* pseudo-miRNA sequences, which are derived from coding regions and ncRNA of a close relative to the target species.

5.4 Results

5.4.1 Demonstrating effectiveness of framework

To demonstrate the effectiveness of our proposed species-specific data set generation framework, it was applied to four diverse target species representing four distinct phyla: *Anolis carolinensis*, *Drosophila melanogaster*, *Arabidopsis thaliana*, and *Rhesus lymphocryptovirus*. We refer to these four target species as 'hold-out test species'. For each hold-out test species, we first generated species-specific positive and negative training sets for which data from the hold-out test species is withheld (*i.e.* we pretend that no sequence annotation is available for the target species). Testing data sets for each of the hold-out test species were extracted based on the withheld (known) annotations (*i.e.* true miRNA & exonic hairpin regions). In order to demonstrate the broad applicability of our framework,

we have applied it to both the widely cited microPred classification pipeline [104] and the newer HeteroMirPred classification pipeline [113]. Training sets generated by our framework are compared against equivalent data sets using human-only data (as used in [104]), and multi-species pooled data (as used in [110], [113], [125]). In all cases, the species-specific training data generated by our framework leads to substantial and consistent performance gains. Each step of this evaluation procedure is described in detail in the following sections.

5.4.2 Hold-out Test Species Data sets

Table 6 - Hold-out data sets used for testing of species-specific data set generation

Species	Size of positive hold-out data set	Size of negative hold-out data set
Anolis carolinensis	282	500
Drosophila melanogaster	237	443
Arabidopsis thaliana	298	500
Rhesus lymphocryptovirus	35	86

Positive and negative hold-out species data sets were generated for each of our four hold-out test species. The four positive test sets consist of all pre-microRNA sequences present in miRBase v.19 for the given hold-out test species. Corresponding negative test sets were built using the negative set generation method described in the methods section; data was retrieved from the hold-out species genome. For species where an abundance of negative test data was present, the data set size parameter *n* was set to 500. The number of sequences in each of the positive and negative hold-out sets can be seen in Table 6.

5.4.3 Reference Training Data sets

We compared our species-specific training data sets with the training data sets used in the microPred and HeteroMirPred studies. These two data sets represent human-only training data and pooled multi-species training data, respectively.

MicroPred Training Data set. The microPred training data set, available at http://www.cs.ox.ac.uk/people/manohara.rukshan.batuwita/microPred.htm, has become a *de-facto* standard for the training of microRNA prediction methods, having been used in numerous studies since it was introduced in 2009 [104], [110], [113], [114], [116], [121]. This data set contains 691 human pre-miRNA sequences, as well as 9248 pseudo-miRNA hairpin candidates that appear in human coding regions and human ncRNA regions.

HeteroMirPred Training Data set. The HeteroMirPred training data set was not made publicly available, therefore we have re-created the data set using the parameters described by Lertampaiporn *et al.* [113]. The original positive training data set consists of 1000 randomly selected non-redundant pre-miRNA – 600 from the H. sapiens genome, 200 from the *O. sativa* genome, and 200 from the *A. thaliana* genome. Because *A. thaliana* is one of the species used in our hold-out test sets, our re-creation of the HeteroMirPred data set uses *Glycine max* pre-miRNA in place of *A. thaliana* pre-miRNA.

While not explicitly stated in the study by Lertampaiporn *et al.*, the negative set used to train the HeteroMirPred classifier is likely to be the same negative set used to train the microPred classifier. We believe this to be true since the two data sets have the same number of coding region sequences and ncRNA sequences, and the negative set generation methodology described by the two studies implies that this is the case. Therefore, we have used the microPred negative training set in our recreation of the HeteroMirPred training data set.

5.4.4 Species-specific Training Data sets

For each hold-out test species, the framework described above was applied to create species-specific training data sets. Positive data were selected with preference to samples from species which are phylogenetically similar to each of the hold-out species. Negative data were selected from species that were closely related to the hold-out test species, as

follows: *Xenopus tropicalis* for the hold-out species *Anolis carolinensis*, *Drosophila pseudoobscura* for *Drosophila melanogaster*, *Arabidopsis lyrata* for *Arabidopsis thaliana*, and *Epstein Barr virus* for *Rhesus lymphocryptovirus.* In order to ensure a fair comparison between existing training data sets and species-specific training data sets, the numbers of positive and negative patterns used in the species-specific data sets (*p* and *n*, respectively) match the number of positive and negative patterns used in the respective existing training set. Species-specific training sets based on the microPred classifier use p=692 and n=10000, while species-specific training sets based on the HeteroMirPred classifier use p=1000 and n=10000. Minority class data sets were oversampled using SMOTE [43] such that positive and negative training data sets contained the same number of samples.

5.4.5 Model Classifiers

We demonstrate the applicability of our data set generation method using local implementations of the microPred and HeteroMirPred classifiers as published in [104] and [113], respectively. The microPred and HeteroMirPred classifiers had to be implemented locally since the original implementations were unsuitable for our experiments because they do not produce prediction confidence results (only binary predictions), and therefore cannot be analyzed using precision-recall or ROC curves. We have therefore generated SVM classifiers, following the feature set and training protocol used in the original reports, using the LibSVM library [124]. SVM hyperparameters found to be optimal over the reference data sets (microPred and HeteroMirPred) were used for all species-specific classifiers.

5.4.6 Classifier test protocol

For each hold-out test species, species-specific training data sets were compared with their respective human-specific and pooled training data sets on microPred and HeteroMirPred-like classifiers using the following protocol:

Training data sets were stratified into 10 equal subsets, as in an outer 10-fold cross-validation experiment. A classifier was then produced for each training set, such that each classifier was trained on 90% of the total training data. Normalization of features was performed on each training data set independently, using the LibSVM feature scaling algorithm [124]. Each classifier was then used to predict the complete hold-out species test data set, thereby providing ten estimates of classification performance. The total performance of the 10 classifiers on the hold-out data set was used as a measure of the effectiveness of the training set on which the 10 classifiers were trained. The same test procedure was used for both species-specific training data sets and reference training data sets.

5.4.6.1 Measuring performance of targeted species-specific models on hold-out data sets

We compare the performance of our species-specific training data generation approach with the approaches of the microPred and HeteroMirPred classification studies using the metrics of precision and recall. Since the class imbalance in our test data is not necessarily reflective of the actual degree of imbalance expected when the classifier is applied to a complete genome, we use the prevalence-corrected precision and recall as our primary metrics.

Relative to the geometric mean and Acc metrics commonly used in the field of microRNA prediction, precision and recall better elucidate real-world performance for *de novo* miRNA experiments, where large class imbalances are expected, as these metrics operate at the expected class imbalance for a given classification problem.

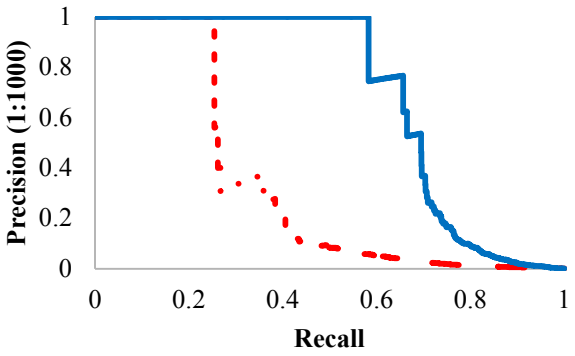
Actual class imbalance in genome-wide miRNA prediction experiments varies based on the genome used. Within eukaryotic genomes we estimate the real-world class imbalance in microRNA prediction experiments to be approximately 1000:1 in favour of the negative class. This is considered to be a conservative estimate, as the relatively compact *D. melanogaster* genome contains approximately 800,000 non-redundant hairpin structures

which satisfy conditions for miRNA candidacy, while the number of microRNA in this genome is only 466 as of miRBase v.20 [4] resulting in a ratio of 1716:1. Similarly, the *H. sapiens* genome contains approximately 11,000,000 hairpin structures [135] and 2578 miRNA as of miRBase v.20 (ratio of 4267:1). Therefore, in the calculation of precision and recall within eukaryotic genomes, we set *r* to 1000, representing a 1:1000 class imbalance. Similarly, we estimate the class imbalance of *de novo* miRNA prediction within smaller viral genomes to be 1:100, and set the *r* value to 100 in accordance with this estimate.

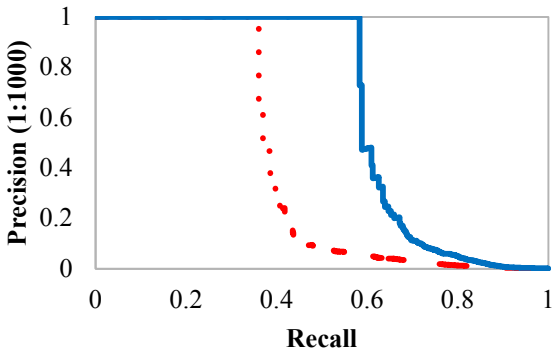
5.4.7 Experimental results

Figure 17 presents the precision-recall curves for microPred-like classifiers trained on species-specific training data and human-specific training data as the classifier decision threshold varies from permissive to conservative. Figure 18 presents equivalent precision-recall curves for HeteroMirPred-like classifiers trained on species-specific training data and multi-species pooled training data. In all precision-recall curves, the dashed red curve indicates prediction using a reference training set, while the solid blue curve indicates the tailored species-specific approach developed in this study. MiRNA predictions were carried out for *Anolis carolinensis*, *Arabidopsis thaliana, Drosophila melanogaster* and *Rhesus lymphocryptovirus*. As can clearly be seen in these figures, our species-specific approach provides a consistent and substantial boost in recall for a wide range of precision values for all four test species.

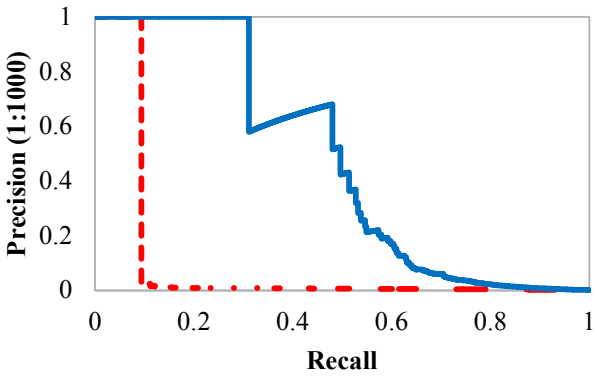
Anolis carolinensis



Arabidopsis thaliana



Drosophila melanogaster



Rhesus lymphocryptovirus

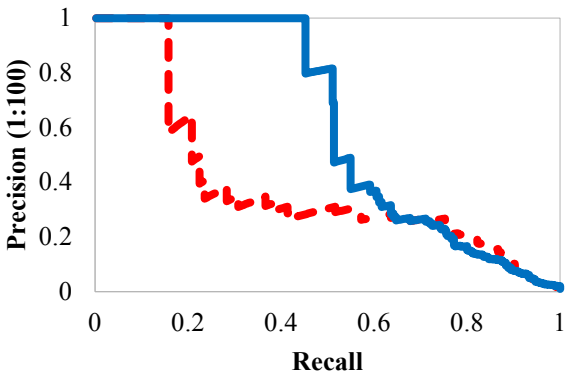


Figure 17 - Comparison of species-specific training data with human-specific data on microPred-like models.

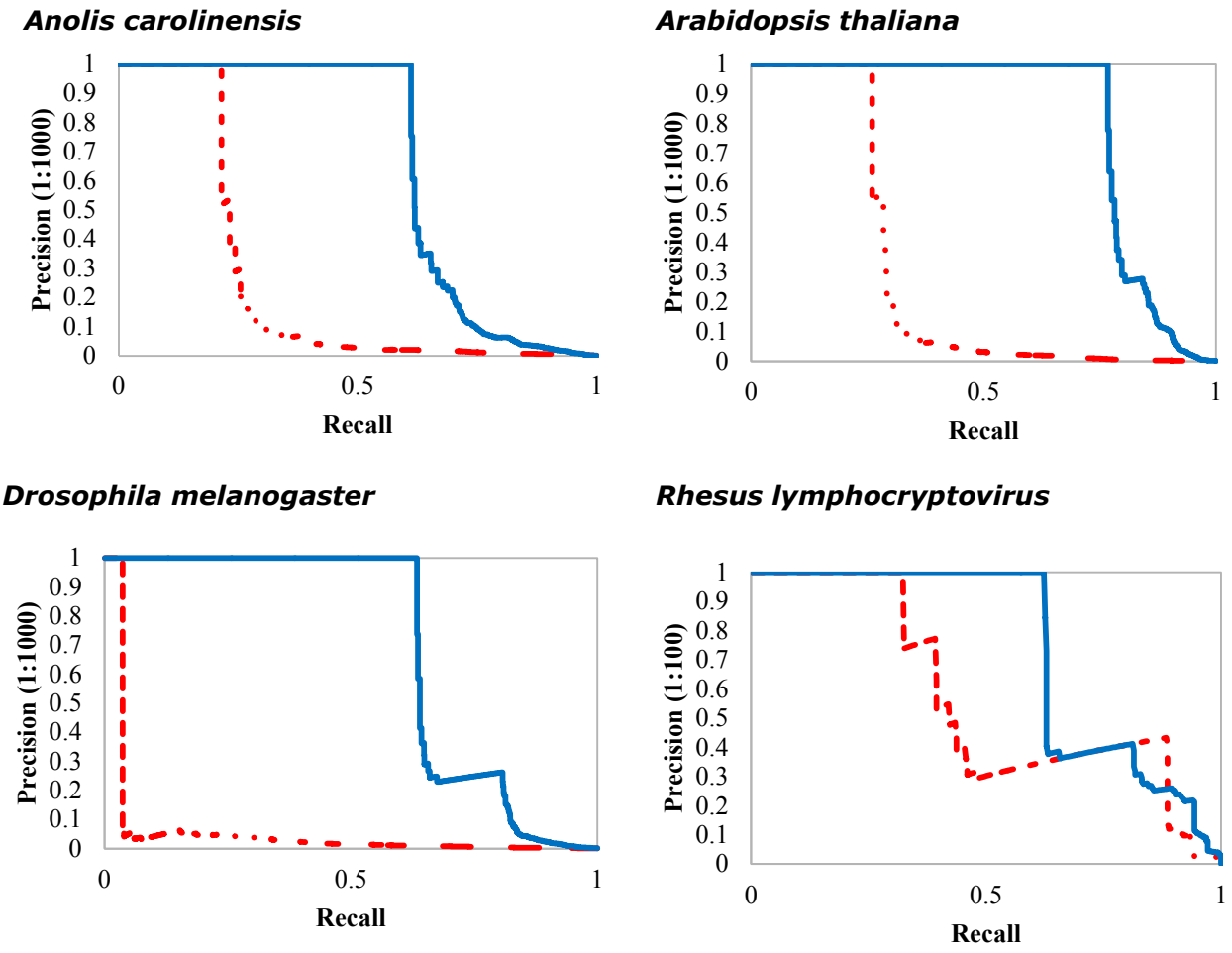


Figure 18 - Comparison of species-specific training data with human-specific data on HeteroMirPred-like model.

In order to provide useful summary metrics for miRNA classification, we also report recall rates at the 90% and 50% precision levels, representing the number of near-guaranteed predictions made by the classifier (90% precision), and the portion of true miRNA expected to be recovered when operating at a typical acceptable experimental validation rate (50% precision). Table 7 and Table 8 summarize the recall rates of microPred-like classifiers trained on species-specific training data and the default microPred human training data, at precision rates of 90% and 50%, respectively. On average, application of species-specific

training sets increases the recall rate at 90% precision by 152.3% and the recall rate at 50% precision by 199.9%.

Table 7 - Recall at 90% precision, human-specific and our tailored species-specific training data using the microPred-like classifier.

Hold-out test species	Human-specific training data	Species-specific training data	Increase in recall (%)
Anolis carolinensis	0.254	0.583	130
Drosophila melanogaster	0.094	0.311	231
Arabidopsis thaliana	0.360	0.583	61.9
Rhesus lymphocryptovirus	0.158	0.453	187

Table 8 - Recall at 50% precision, human-specific and our tailored species-specific training data using the microPred-like classifier.

Hold-out test species	Human-specific training data	Species-specific training data	Increase in recall (%)
Anolis carolinensis	0.262	0.695	165
Drosophila melanogaster	0.094	0.497	429
Arabidopsis thaliana	0.370	0.588	58.9
Rhesus lymphocryptovirus	0.208	0.514	147

Table 9 and Table 10 summarize analogous data using a HeteroMirPred-like classifier. Consistent with the results for the microPred-like classifier, substantial increases are observed for our proposed species-specific approach; on average, the recall rate at 90% precision is increased by 533.2% and recall rate at 50% precision is increased by 396.1%.

Table 9 - Recall at 90% precision, pooled training data and our tailored species-specific training data using the HeteroMirPred-like classifier

Hold-out test species	Pooled training data	Species-specific training data	Increase in recall (%)
Anolis carolinensis	0.215	0.611	184
Drosophila melanogaster	0.036	0.634	1660
Arabidopsis thaliana	0.260	0.767	195
Rhesus lymphocryptovirus	0.325	0.625	92.3

Table 10 - Recall at 50% precision, pooled training data and our tailored species-specific training data using the HeteroMirPred-like classifier

Hold-out test species	Pooled training data	Species-specific training data	Increase in recall (%)
Anolis carolinensis	0.232	0.620	167
Drosophila melanogaster	0.051	0.639	1150
Arabidopsis thaliana	0.285	0.781	174
Rhesus lymphocryptovirus	0.325	0.629	93.5

A test of significance was applied to each "Increase in recall" result in Table 7, Table 8, Table 9, and Table 10. For each pair of SMIRP and reference classifiers, a distribution of increases in recall expected under the null hypothesis (H_0: there is no significant difference in achievable recall between the two methods) was computed at both the 50% and 90% precision levels. These distributions were formed by repeatedly (N=10,000) pooling the prediction scores from the two methods and drawing pseudo-samples labelled as *SMIRP* and *reference*. The null hypothesis was enforced by randomly permuting the SMIRP and reference classifier score within each row/sequence. Row ordering was preserved resulting in a paired (matched) experiment design. For each pair of pseudosamples, the increase in recall was recorded. P-values were then computed for the actually observed increases in recall. All results in all four tables were found to be significant at the $p<0.01$ level except for the Re@Pr50 results for *Rhesus lymphocryptovirus* in Table 8 and Table 10.

Table 11 and Table 12 present the number of miRNA recovered using our pipeline which do and do not share sequence similarity (80% or greater) with training data. Of the hold-out miRNA recovered by our microPred-like and HeteroMirPred-like classifiers, 60.6% and 59.6% do not share significant similarity with any of the training data. Therefore, SMIRP is capable of predicting miRNA which are not homologous to existing miRNA.

Table 11 - Average number of miRNA recovered at 50% precision which are and are not, homologous to training data (MicroPred-like classifier). Here, homology is defined as 80% sequence identity or higher.

Hold-out test species	Number of miRNA recovered	Homologous to training data	Not homologous to training data
Anolis carolinensis	196	87	109
Drosophila melanogaster	118	73	45
Arabidopsis thaliana	175	39	136
Rhesus lymphocryptovirus	16	0	16

Table 12 - Average number of miRNA recovered at 50% precision which are, and are not, homologous to training data (HeteroMirPred-like classifier).

Hold-out test species	Number of miRNA recovered	Homologous to training data	Not homologous to training data
Anolis carolinensis	175	94	81
Drosophila melanogaster	152	76	76
Arabidopsis thaliana	232	65	167
Rhesus lymphocryptovirus	22	0	22

5.4.8 Effect of phylogenetic distance on classification performance

In order to elucidate the effect of phylogenetic similarity within positive and negative data sets, we have performed additional classification experiments on the *A. thaliana* hold-out set. In each of these experiments, we varied the phylogenetic similarity between the hold-out species and our positive and negative training sets. Seven training data sets were generated, for which the following phylogenetic groups (clades) were removed: genus *A thaliana*, family *Arabidopsis*, order *Brassicaceae*, clade *Eurosids II*, clade *Rosids*, clade *Eudicots*, kingdom *Plantae*. Negative training data sets were built using the following representative species, respectively: *A. lyrata, B. napus, T. cacao, C. melo, O. sativa, P. patens* and *H. sapiens*. Figure 19 demonstrates a clear inverse correlation between

classification performance and phylogenetic distance between training data and hold-out test data.

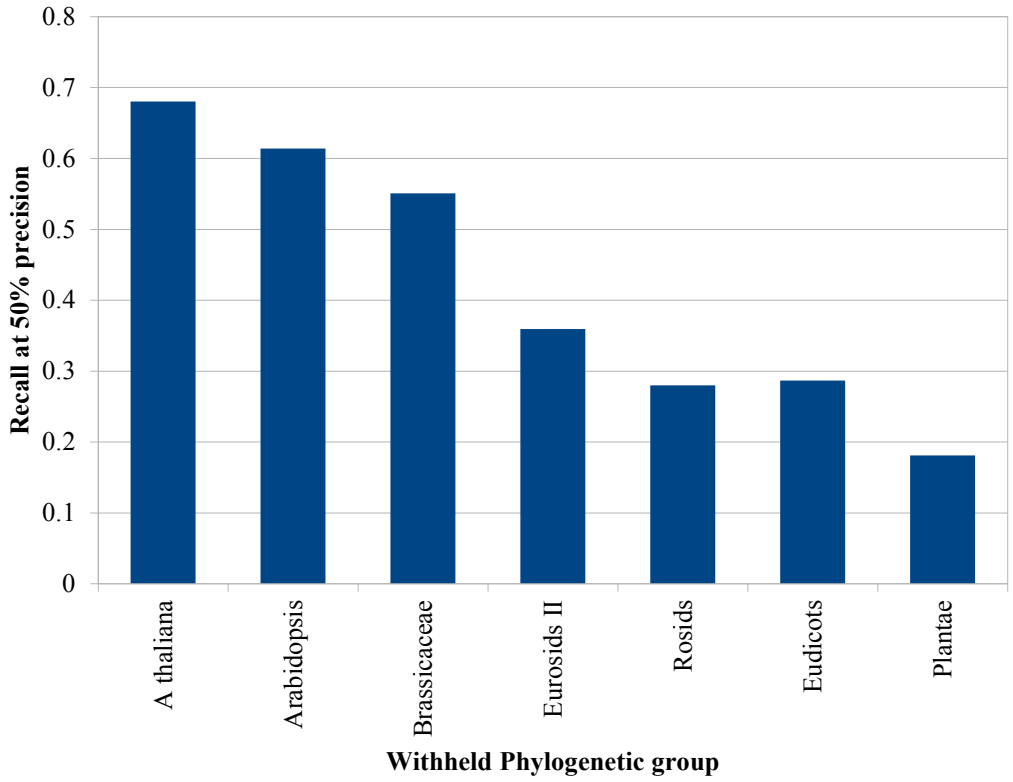


Figure 19 - Recall at 50% Precision on A. thaliana hold-out test set, as phylogenetic distance between training data and A. thaliana is systematically increased. X-axis labels describe the phylogenetic group which was withheld during training data set generation.

This result serves to validate SMIRP's underpinning hypothesis, in that training data should be taken preferentially from species as closely related to the target species as possible.

5.4.9 Application of SMIRP to random forest classifiers

In order to demonstrate the applicability of SMIRP across multiple classifier types, we have compared the SMIRP data set generation technique with the taxon-specific data set generation approach of HuntMi [110] for the training of random forest classifiers. The HuntMi feature set, which contains the 21 microPred features and 7 additional features, was

used for this experiment. Random (decision) forest classifiers were trained using the *scikit-learn* [136] python library. Classifier hyperparameters were set to default values, with the exception of the number of trees which was set to n=500. This high number of trees allows for more fine-grained classification confidence results relative to the lower default value of n=10 since confidence is derived from the voting results among n individual trees. Classifiers were trained using the SMIRP species-specific data sets described above, and also using taxon-specific data sets representing animals, plants, and viruses (as appropriate to each test hold-out species). Taxon-specific positive data sets contain all experimentally validated miRNA from miRBase version 19 for the respective taxon, excluding that of the hold-out species. Taxon-specific negative data sets are those provided by HuntMi [110]. Each of these classifiers was then tested on the hold-out species test sets described above. As demonstrated in Figure 20, species-specific data sets outperform the taxon-specific data sets by a large margin for all four hold-out species across all three major taxa. Importantly, since these experiments involve random forests, as opposed to SVMs used elsewhere in this manuscript, these results also demonstrate the applicability of the SMIRP framework to microRNA prediction studies, regardless of machine learning approach. Performance of the HuntMi data set generation methodology is lower than other existing methods, microPred and HeteroMirPred, on all four hold-out species data sets. This is likely due to the negative set generation methodology used in the training of the HuntMi method, wherein less stringent conditions were placed on pseudo-miRNA sequences.

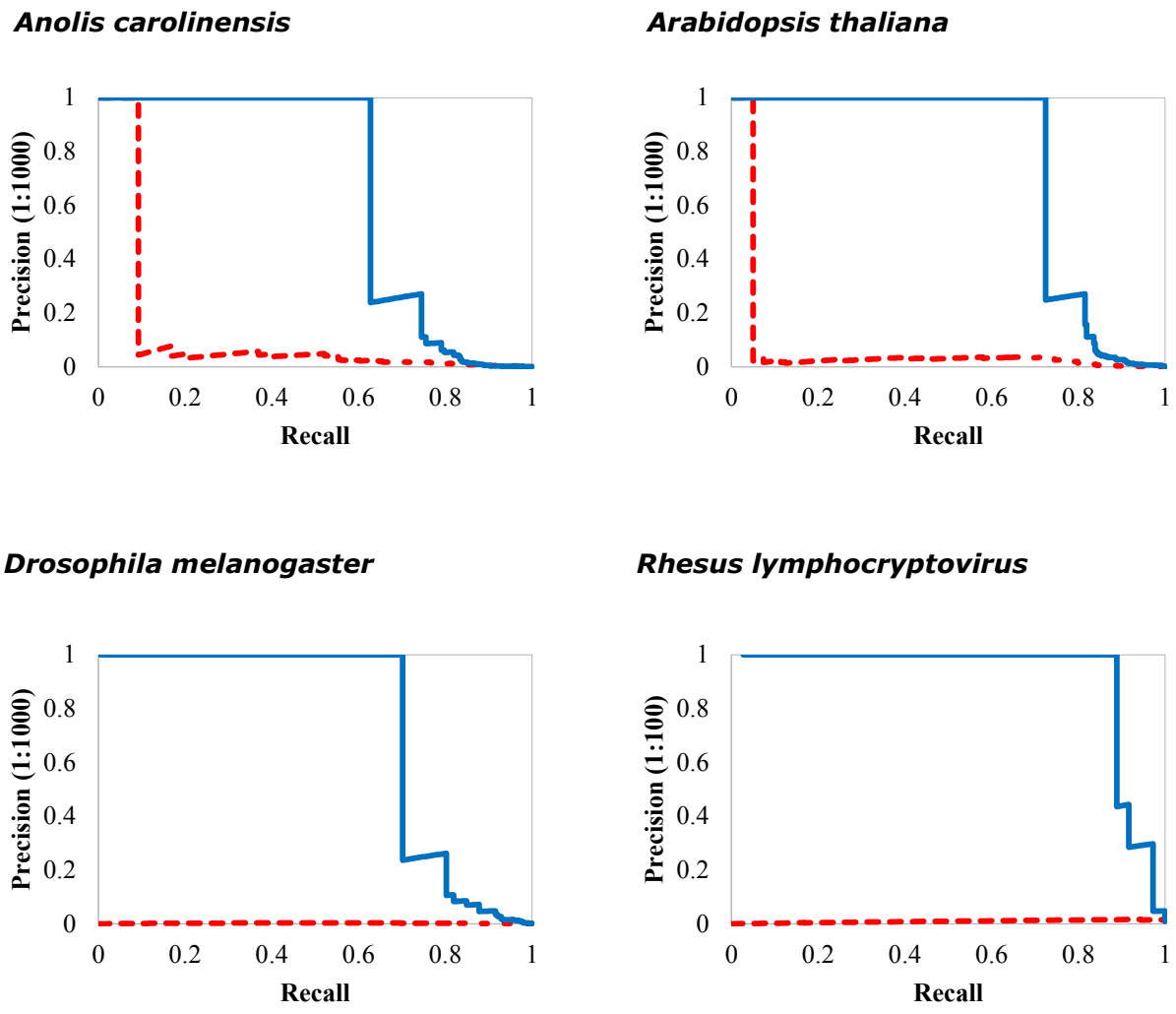


Figure 20 - Comparison of species-specific training data with taxon-specific data on HuntMi-like models. In all precision-recall curves, the dashed red curve indicates HuntMi-like prediction using a taxon-wide-trained model, while the solid blue curve indicates the SMIRP species-specific approach developed in this study. In panels C and D, the dashed-red curve appears to follow the x-axis due to scaling; the precision of these classifiers is in fact non-zero. MiRNA predictions were carried out for Anolis carolinensis, Arabidopsis thaliana, Drosophila melanogaster and Rhesus lymphocryptovirus.

5.4.10 Results of genome-wide *B. glabrata* miRNA prediction

We applied the SMIRP data set generation framework along with the HeteroMirPred-like classifier to the unannotated *B. glabrata* (snail) genome with the goal of predicting novel

miRNA. SMIRP was used as described above, using p=600 and n=10000. *L. gigantean* was used as the negative reference species. Classification was performed using the libsvm library and the HeteroMirPred feature set. Using the SMIRP framework, 202 miRNA were discovered within the *B. glabrata* genome. Of these, 107 miRNA are novel, and a significant portion of these novel miRNA were found to target genes involved in cellular processes specific to snail, such as secretory mucal proteins and shell formation. These results demonstrate the ability of SMIRP to predict a large number of novel miRNA within the unannotated genome of a non-model species.

5.4.11 Results of genome-wide *P. polycephalum* miRNA prediction

Using the SMIRP framework, we created a miRNA prediction tool specific to *P. polycephalum* which was able to successfully identify 48 pre-miRNA (2 conserved and 46 novel) and associated mature miRNA from *P. polycephalum*. SMIRP was used as described in this thesis, using p=600 and n=10000. *Dictyostelium discoideum* was used as the negative reference species. Classification was performed using the libsvm library and the HeteroMirPred feature set. Of the 17 currently annotated precursor miRNA for *Dictyostelium discoideum*, no homology was found with any miRNA annotated from the *P. polycephalum* genome.

5.5 Conclusions

Currently, state-of-the-art methods for microRNA prediction do not provide adequate specificity for the efficient discovery of novel miRNA during genome-scale experiments on unannotated genomes. These novel miRNA discovery experiments are performed in the presence of very high-class imbalance (typically on the order of 1000 negative hairpin regions per one true positive miRNA), and experimental validation of positives is costly and time-consuming. As a result, very high precision and specificity are demanded of classifiers, and current efforts that are often tuned to maximize the geometric mean of sensitivity and

specificity do not meet this demand. Furthermore, we have demonstrated that specificity of microRNA prediction decreases substantially when classifiers are asked to make miRNA predictions on species dissimilar to the species on which they were trained.

In order to provide precise classification of microRNA in unannotated genomes potentially distant from model species, we have introduced a framework for species-specific miRNA classification, which increases prediction performance for arbitrary species. This framework dynamically produces classification models for the test species under study. Positive training sets are produced through a two-step filtering process on the set of all available microRNA sequences from multiple species:

1. Generate clusters of miRNA based on sequence identity or similarity. The largest such clusters are representative of a large number of highly confident miRNA that are conserved across species.
2. Select a representative miRNA from the largest clusters. Selection is based on phylogenetic similarity to the target species, increasing likelihood of conservation between representative miRNA and target species. In addition, selection of a single miRNA from each cluster ensures that the positive training set contains no redundant miRNA.

Negative training sets are built using coding regions and ncRNAs from annotated genomes of species that are closely related to the target species. As with the positive training sets, selection of closely related species here increases the likelihood of sequence conservation between the negative training set and the target genome.

We have demonstrated that SMIRP, our species-specific data set generation framework, provides a dramatic increase in classifier performance relative to the human-specific data set generation method of the microPred study [104], as well as the multi-species pooled data set generation method of the HeteroMirPred study [113] and the taxon-specific method of HuntMi [110]. This increase in performance holds across four distinct hold-out species

representing four distinct phyla. By reporting precision at realistic class imbalance levels, our tests reflect the high-specificity operating points which are required during genome-wide microRNA prediction studies. Relative to pooled (HeteroMirPred) or human-specific (microPred) data set generation methods, SMIRP results in a 4x increase in recall when demanding a precision of 90% (i.e. 4x more true miRNA are identified while expecting 90% of predicted miRNA to be true). Consistent increases in classification performance were observed when using both SVM and random forest classifiers indicating the broad applicability of the SMIRP framework. We have demonstrated that SMIRP-based classifiers are able to predict novel miRNA, without homology to training data. Applying this method to the unannotated genome of *B. glabrata* (snail), 202 miRNA were discovered, of which 107 were novel and many of these are snail-specific. Within the unannotated *P. polycephalum* (slime mold) genome, 48 miRNA were discovered, 46 of which were novel. SMIRP can be applied to any existing or future microRNA prediction method, providing increased classification performance for all experiments on unannotated genomes.

The four pre-trained microPred-like and HeteroMirPred-like species-specific miRNA prediction systems evaluated in this study are available as Species-specific MIRna Predictors (SMIRP), at http://bioinf.sce.carleton.ca/SMIRP. We expect that these four classifiers will be useful for other closely related species. For example, the *A. thaliana* classifier is likely to be more effective for predicting miRNA in other plant species than would be the default human-only or multi-species pooled classifiers. Furthermore, all software for preparing species-specific data sets is available in open source at http://bioinf.sce.carleton.ca/SMIRP, as well as the training and testing data used in this study.

6 A comprehensive evaluation framework for *de novo* miRNA prediction reveals that updated free energy parameters increase microRNA prediction performance

6.1 Abstract

In this chapter, we introduce a novel comprehensive evaluation framework for the analysis of *de novo* miRNA prediction methods. This framework considers the performance of both pre-filtering and classification stages of miRNA prediction pipelines. This framework is highly applicable to genome-wide miRNA discovery experiments, wherein large genomic sequences must be scanned using a pre-filtering algorithm prior to classification of candidate miRNA, and any miRNA which fail to be retrieved during pre-filtering are excluded from further analysis. Using this framework, we demonstrate that current *de novo* miRNA prediction methods, which use outdated RNA folding free energy parameters during pre-filtering, often fail to retrieve true miRNA from larger genomic sequences. By updating free energy parameters for RNA folding, *de novo* miRNA prediction is improved substantially, a result which is only measurable using the comprehensive evaluation framework described in this chapter.

This chapter has been adapted from a conference publication which was presented under the same title at IUPESM 2015 World Congress on Medical Physics & Biomedical Engineering.

6.2 Introduction

The computational prediction of RNA structure is the main method by which genomic sequences are determined to be suitable candidates for miRNA prediction [98], [103], [104], [113]. In particular, sequences which fold into stable hairpin structures of

approximately 70-120nt in length are considered for miRNA prediction. Sequences which do not conform to typical miRNA-like hairpin structures are discarded from further consideration before evaluation by the primary miRNA prediction module. This pre-filtering step is widely used in both *de novo* miRNA prediction methods [104] and NGS-based miRNA prediction methods [15].

Such pre-filtering is necessary to partially address the rarity of miRNA among all genomic sequence (roughly estimated to account for only 0.01% of the genome by length). Often, highly conservative pre-screening steps are required in order to limit the false discovery rate among predicted miRNA. Currently, miRNA prediction performance is typically measured only on those positive and negative sequences passing the pre-filtering criteria. In this way, the impact of the pre-filtering stage itself is completely ignored. An overly conservative pre-filtering stage will limit overall system recall, while an overly permissive pre-filtering stage will lead to high class imbalance and increased false positive predictions. The pre-filtering stage also has an impact on the training of miRNA classifiers, since training data include only those sequences passing the pre-filtering stage. Improving the quality of pre-filtering will lead to an increase in the number of genomic regions correctly identified as miRNA-like hairpin structures, increasing in turn the sensitivity of miRNA prediction methods. In this chapter, a comprehensive evaluation framework is developed that examines the overall classification performance of the system, including the pre-screening stage. This framework addresses a weakness in the methodology of current miRNA prediction studies which employ such pre-filtering steps but fail to include the effect of pre-filtering on overall system performance, instead only reporting classification performance over those putative miRNA that pass the pre-filtering stage.

While there is no universally accepted standard among miRNA pre-filtering algorithms as to what constitutes a miRNA-like hairpin structure [110], [113], one commonly used definition is a sequence which, when folded by the ViennaRNA RNAfold package [134], produces a

hairpin structure with at least 18 paired bases, and a minimum free energy (MFE) of at most -15.0. This definition was introduced by Xue *et al.* in 2005 [103] and continues to be used in many modern miRNA prediction studies (e.g. [113]). For example, the microPred data set, which is the most widely used data set for training and testing of miRNA prediction methods, uses this definition of miRNA-like hairpin [104].

During implementation of the above pre-filtering criteria for miRNA-like hairpin structures, the RNAfold algorithm, as packaged in ViennaRNA version 1.8.5 (here referred to as *RNAfold185*), is used to identify hairpin structures and determine their MFE. ViennaRNA 1.8.5 uses the Turner 99 energy parameters in order to calculate MFE [134]. ViennaRNA 2.0 and onward use the newer Turner 2004 energy parameters, which more accurately model RNA folding [92]. Although the use of accurate RNA folding energy parameters has been shown to improve the overall effectiveness of RNA studies [137], ViennaRNA 2.0 has yet to be used in the context of miRNA prediction and the resulting effect of the updated energy parameters is unknown.

In this chapter, we leverage the novel comprehensive evaluation framework described previously in order to demonstrate that the use of RNAfold, as implemented in ViennaRNA version 2.1.8 (*RNAfold218*), in the identification of candidate miRNA-like hairpin structures is superior to the widely used *RNAFold185*. We show that *RNAfold218* identifies more experimentally verifiable miRNA relative to *RNAfold185*, increasing the maximum possible sensitivity of miRNA classification. Furthermore, we demonstrate that the resulting miRNA prediction performance is substantially improved overall. We expect that all machine learning methods which rely on RNA structure prediction will benefit from updating to *RNAfold218*, however retraining of the classification models will be required.

6.3 Methods

6.3.1 Overview of the comprehensive evaluation framework for *de novo* miRNA prediction

In order to quantify the effect of both the pre-filtering and classification stages of de novo miRNA prediction pipelines, we have developed the following comprehensive evaluation framework which models miRNA prediction performance within an unannotated genome. Figure 21 shows an overview of the test framework pipeline; details of each step within the pipeline are described in the following sections.

6.3.1.1 RNA sequence data sets

Our test framework makes use of the following RNA sequence data sets within its pipeline. Details on the collection and curation of these data sets can be found in section 6.3.2.

- All experimentally validated pre-miRNA sequences for a given species
- All experimentally validated mature miRNA sequences for the given species
- Sequence data retrieved from the genome of the given species, representing all experimentally validated pre-miRNA sequences, with the addition of 30bp of upstream and downstream flanking sequence surrounding each pre-miRNA
- A set of complete protein coding region sequences for the given species
- Sample genomic data for the given species, consisting of one contiguous genomic sequence chosen at random from one of the given species' chromosomes

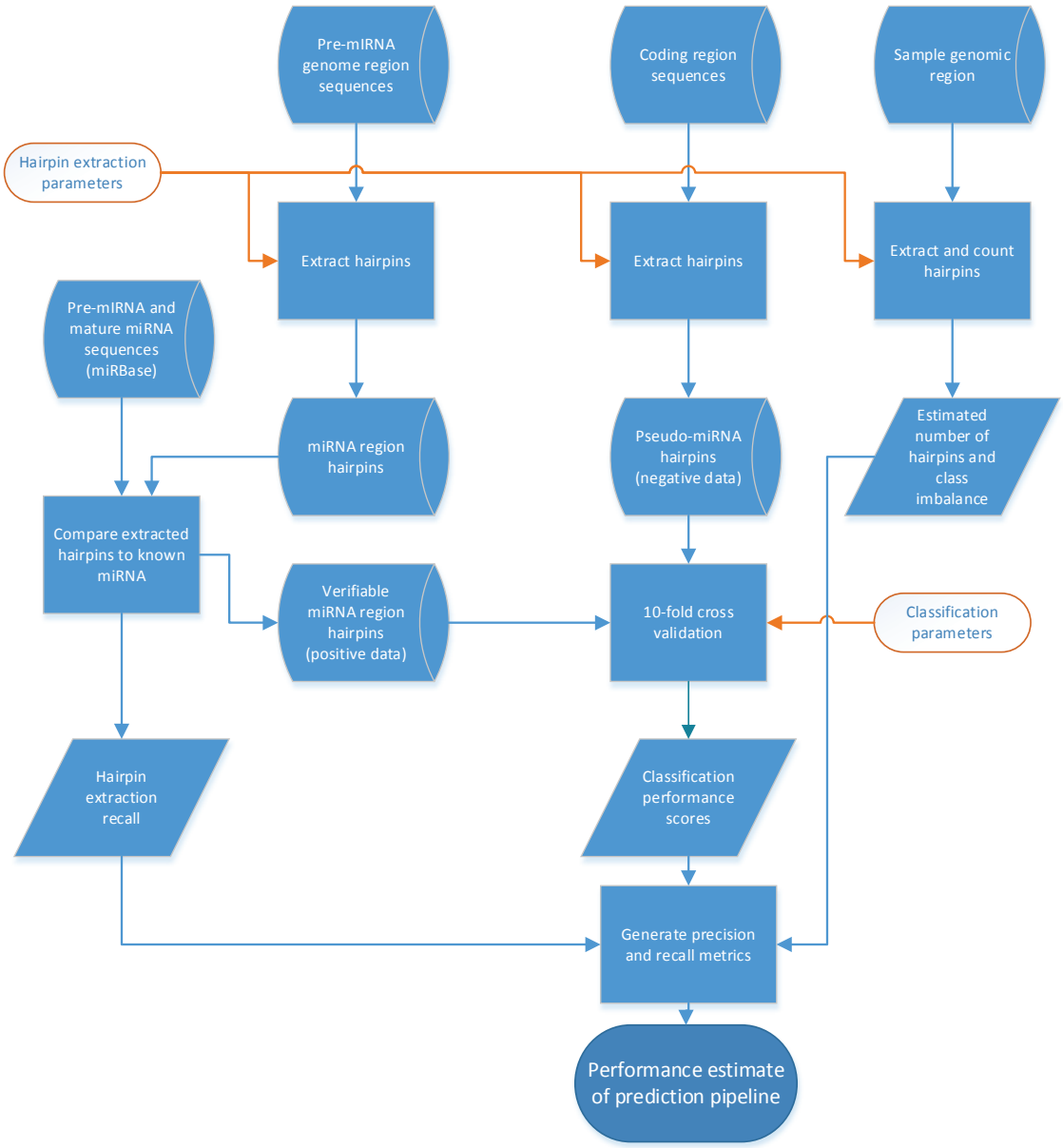


Figure 21 - Overview of the test framework for examination of a complete miRNA prediction pipeline. Elements of this diagram are explained in detail in sections 6.3.1.1 through 6.3.1.6 of this thesis.

6.3.1.2 Extraction of miRNA-like hairpin sequences

MiRNA-like hairpin sequences are extracted from the pre-miRNA genome region sequences, coding region sequences, and the genomic data sample using the RNALfold algorithm [92] as implemented in either the ViennaRNA 1.8.5 or the ViennaRNA 2.1.8 package. A single set of hairpin extraction parameters (*e.g.* maximum distance between paired bases, number of paired bases required in hairpin structure, maximum allowed secondary structure free energy) are used for hairpin extraction within all three data sets. This step of the framework simulates the pre-filtering step of a *de novo* miRNA prediction pipeline using the hairpin extraction parameters supplied to the framework. Extraction of hairpins from pre-miRNA genome region sequences provides positive training and test data for the classification stage of the pipeline; extraction of hairpins from coding region sequences provides negative training and test data for the classification stage of the pipeline; and extraction of hairpins from the sample genome region is used to estimate the number of hairpins which would be extracted from the genome using the input hairpin extraction parameters.

Note that multiple hairpin structures can be extracted from a single pre-miRNA, and a single pre-miRNA region may contain multiple mature miRNA (one on each arm of the stem/duplex).

6.3.1.3 Development of positive training data (comparison of extracted hairpins to known miRNA)

In the previous stage, hairpin extraction was performed on genomic regions surrounding known miRNA hairpins. The hairpin extraction step resulted in a set of hairpin structure sequences which, ideally, coincide perfectly with the known pre-miRNA hairpins within each region. In this stage of the framework, the recall of this pre-filtering hairpin extraction method is evaluated. The extracted hairpins are compared to the actual hairpin sequences associated with the known miRNA in order to determine what proportion of known miRNA

are successfully captured by the hairpin extraction method. We define this metric as the *Pre-filtering Recall*, which is calculated using the following formula:

$$Pre\text{-}filtering\ Recall = \frac{\#\ known\ miRNA\ recovered\ by\ pre\text{-}filtering}{\#\ miRNA\ regions\ examined\ by\ pre\text{-}filtering}$$

The *Pre-filtering Recall* estimates the upper bound on recall which is achievable by the full *de novo* miRNA prediction pipeline, assuming perfect recall within the classification stage of the pipeline.

In order to determine the success of an extracted hairpin in capturing a corresponding known pre-miRNA, we check each extracted hairpin for the presence the complete known mature miRNA in the pre-miRNA sequence. This is justified since the presence of full mature miRNA is a necessary condition for the experimental validation of a miRNA prediction. In the case of multiple mature miRNA within a pre-miRNA region, both mature regions must present within the extracted hairpin region.

In addition to the determination of pre-filtering recall, the set of positive training and test data for use during the classification stage of the de novo miRNA prediction pipeline is built during this step of the framework. The set of positive data is taken as the subset of extracted hairpin regions which coincided with their corresponding known miRNA. In the event that multiple hairpin structures from a single pre-miRNA genome region contain all mature miRNA for a given pre-miRNA, the hairpin with the lowest minimum free energy is selected for use in the positive training and test data set.

6.3.1.4 Development of negative training data set

The negative training and test data for use during the classification stage of our framework consist of hairpins extracted from the coding region data set using the pre-filtering method described in section 6.3.1.2. Redundant hairpin sequences (those which share 100% identity with other negative hairpins which are at least as many base pairs) were removed

from the negative training and test data set. In the event that more than 10,000 non-redundant hairpins are extracted from the coding region data set, 10,000 hairpins are chosen at random to form the negative training data set. If 10,000 or fewer hairpins are extracted, the negative training data set consists of all hairpins extracted from the coding region data set. The value 10,000 was arbitrary, but found to be reasonable in Chapter 5.

6.3.1.5 Estimation of the number of hairpins in genome and class imbalance

In order to estimate the total number of hairpins which would be extracted from a genome during the pre-filtering stage of a *de novo* miRNA prediction pipeline, we apply the pre-filtering stage to a subsample of the genome (described in section 6.3.1.1). These results are then extrapolated to the full genome. The number of hairpins extracted from the subsample is multiplied by the ratio of the number of nucleotides in the genome to the number of nucleotides in the sample region. This subsampling heuristic is employed to limit the computational requirements of hairpin extraction over large genomic data sets.

Once the number of hairpins extracted from the genome is estimated, the class imbalance resulting from the positive and negative data passing the pre-filtering stage is estimated using the following formula:

$$Class\ imbalance = \frac{estimated\ \#\ of\ hairpins\ in\ genome - estimated\ \#\ of\ positives}{estimated\ \#\ of\ positives}$$

To obtain an estimate of the total number of miRNA expected within any eukaryotic genome, a value of twice the total number of currently known miRNA for the most well-studied eukaryote (*H. sapiens*). This results in an estimate of 3762 miRNA. Similarly the number of positives expected within a prokaryotic genome is estimated to be twice the number of currently known miRNA for the most well-studied prokaryote (*Rhesus lymphocryptovirus*). This results in an estimate of 72 miRNA.

6.3.1.6 10-fold cross-validation experiment

Once positive and negative data sets are developed, performance of a miRNA classification algorithm is estimated using a 10-fold cross-validation experiment, as described in section 2.2.2 of this thesis. No hold-out data is used during the experiment.

Classification test results are ordered by confidence of positive classification. The recall of the full miRNA prediction system is estimated using the following formula:

$$System\ recall = Classification\ recall * Pre\text{-}filter\ extraction\ recall$$

The precision of the miRNA prediction system is estimated using the following formula:

$$System\ precision = \frac{Classification\ recall}{Classification\ recall + (1 - Classification\ specificity * \ estimated\ class\ imbalance)}$$

The system recall and system precision incorporate both the pre-filtering and classification steps of the *de novo* miRNA prediction pipeline. By computing a PR-curve, one can obtain an estimate of the number of miRNA that a prediction system will identify at a given level of precision. As with previous studies in this thesis, Re@Pr90 and Re@Pr50 are used to provide summary performance evaluation of miRNA prediction pipelines.

6.3.2 Data set generation for performance analysis of RNAfold185 and RNAfold218

The previous section of this thesis presents a comprehensive evaluation framework for miRNA prediction. This utility of this framework is demonstrated through the analysis of a specific aspect of the pre-filtering stage: the energy parameters used during RNA structure prediction.

In order to analyze the performance of *RNAfold185* and *RNAfold218*, input data sets were generated using the following methodologies:

MiRNA and mature miRNA sequences were downloaded from miRBase 20.0 [4] to form the basis of the positive data sets. Genomic locations for miRNA were retrieved from miRBase 20.0, while genomic region sequences were retrieved from the Entrez database [138]. 30nt of upstream and downstream flanking sequence were retrieved for each miRNA sequence. For experimental purposes, all *H. sapiens* miRNA were used as our positive data set. Flanking sequences were retrieved from *H. sapiens* genome assembly GRCh38.

Negative data sets consist of pseudo-miRNA sequences which were extracted from coding region data retrieved from the ENA database [133]. During our experiments, 5000 *H. sapiens* coding region sequences were selected at random as the basis for this data set. After pre-filtering was performed on this data set, the resulting hairpin data set contained approximately 30,000 hairpin sequences.

The sample genomic data set used in our experiments consists of a segment of the *H. sapiens* chromosome 11 of length 3052740nt – approximately 1/10000 of the length of the full genome. This data set was retrieved from the ENA database.

6.4 Results

In order to demonstrate the relative effectiveness of *RNAfold185* and *RNAfold218* for miRNA prediction studies, we first analyzed the ability of the two algorithms to correctly extract hairpin structures from genomic regions surrounding actual pre-miRNA sequences (Pre-filtering recall). The corresponding effect on regions presumed to contain only pseudo-miRNA is then examined. Finally, we evaluated both versions of the RNAfold algorithm using system precision and recall.

6.4.1 Pre-filtering recall of miRNA-like hairpin structure identification

Table 13 lists pre-filtering recall values for the *RNAfold185* and *RNAfold218* algorithms. *RNAfold218* is demonstrably better than *RNAfold185* in the task of identifying miRNA hairpin

structures for real miRNA; sensitivity is increased by 14.7% when the updated version of RNAfold is used. Considering that the classification stage of a *de novo* miRNA prediction pipeline will only examine regions containing a miRNA-like hairpin structure, this represents a significant increase in the maximum attainable system recall.

Table 13 – Pre-filtering recall of a de novo miRNA prediction system when different versions of RNAfold are employed for the determination of RNA folding free energy during hairpin extracted.

Method	Pre-filtering recall
RNAfold185	79.47%
RNAfold218	94.17%

6.4.2 Discovery of miRNA-like hairpins within Pseudo-miRNA regions

Table 14 lists estimates of the number of hairpins which would be extracted from the *H. sapiens* genome if *de novo* miRNA pre-filtering were applied to the entire genome, using both *RNAfold185* and *RNAfold218* for RNA folding free energy prediction. *RNAfold218*-based hairpin extraction is substantially more permissive than hairpin extraction based on *RNAfold185*, resulting in an estimated 53% increase in the number of miRNA-like hairpins found within a genome. Considering the rarity of true miRNA in the genome, we expect the vast majority of these newly identified hairpin structures to correspond to pseudo-miRNA. This translates into a higher real-world class imbalance for miRNA prediction methods using *RNAfold218*, which may adversely affect performance. This is explored in the following section.

Table 14 - Estimated numbers of hairpins extracted from the *H. sapiens* genome

Method	Hairpins extracted (1/10,000th of genome)	Estimated hairpins (whole genome)
RNAfold185	2065	20,650,000
RNAfold218	3168	31,680,000

6.4.3 Classification performance of full de novo miRNA prediction pipelines using *RNAfold185* and *RNAfold218*

RNAfold218 provides the advantage of increased pre-filtering recall, which may translate into increased system recall in a full *de novo* miRNA prediction pipeline. However, *RNAfold218* is also more permissive when used in a pre-filtering step, thereby resulting in higher class imbalance for the classification stage of the *de novo* miRNA prediction pipeline. This increase in class imbalance negatively impacts prediction precision. To determine the effect of this trade-off in a miRNA prediction study, we have tested the classification performance of miRNA prediction models based on data resulting from pre-filtering using either *RNAfold218* or *RNAfold185*. Classification experiments were performed using the popular microPred feature set [104], the LibSVM support vector machine library [124], and SMOTE for class imbalance correction [43]. SVM hyperparameter values are identical to those used in the original microPred manuscript [104]. Feature set normalization was performed using the LibSVM feature scaling algorithm [124].

Figure 22 illustrates the system precision and recall of predictors based on *RNAfold218* and *RNAfold185*. Pipelines based on *RNAfold218* outperform otherwise-identical pipelines based on *RNAfold185*, achieving increased recall at all levels of precision. Most notably, at 50% precision (a reasonable operating threshold for experimental validation of miRNA predictions), *RNAfold218* produces a recall of 0.277 compared to only 0.169 for *RNAfold185*. This represents a 64% increase ($p=0.0007$) in miRNA discovered at this precision threshold. It is also worth noting that the maximum achievable rates of recall for the two methods were 79.47% and 94.17% for *RNAfold218* and *RNAfold185* respectively, due to the imperfect recall of the pre-screening steps (see Table 13).

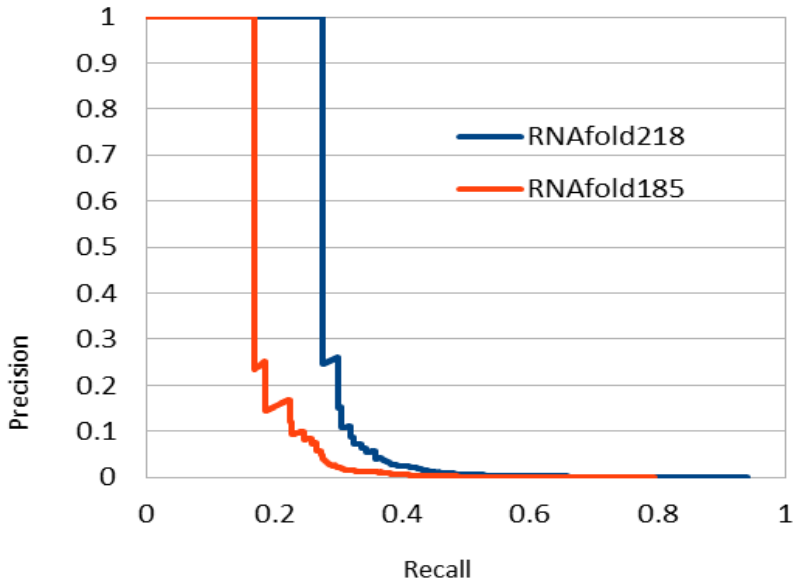


Figure 22 - Precision-recall curves demonstrating relative classification performance of RNAfold218- and RNAfold185-based data sets

6.5 Conclusions

In recent years, MicroRNA prediction studies have focused largely on improvements of classification performance on existing data sets consisting of miRNA and miRNA-like hairpin sequences. The generation of miRNA-like hairpin sequences – the inputs to these classifiers – has remained largely unchanged for several years. Through the use of the novel comprehensive evaluation framework developed here, we have for the first time quantified the effect of pre-filtering on the performance of *de novo* miRNA prediction pipelines. The evaluation framework was subsequently used to guide the improvement of a miRNA prediction pipeline. In particular, we have demonstrated that updating from the current standard of RNAfold v1.8.5 to the newer RNAfold v2.1.8 increases the sensitivity of the pre-filtering stage of the de novo miRNA prediction pipeline and results in significant increases in classification performance for the full pipeline. The evaluation framework described in this chapter is broadly applicable to other miRNA prediction methods, and can be used to evaluate all aspects of a *de novo* miRNA prediction pipeline.

7 miPIE: Predicting miRNA from NGS experiments using integrated lines of evidence

7.1 Abstract

In this chapter, we introduce a NGS-based miRNA prediction method (miPIE) which leverages many categories of sequence- and expression-based features in order to improve miRNA prediction performance. Prevalence-corrected performance metrics are used to elucidate classifier performance at real-world class imbalances. Relative to two state-of-the-art NGS-based miRNA prediction methods, substantial improvements in recall and precision are obtained within five data sets, representing NGS experiments performed on three different species and two different instruments. Performance of the prediction method generalizes well across instrument, species, and experiment. Results demonstrate that the integration of sequence- and expression-based features improves performance relative to any single category of features.

This chapter has been adapted from a journal publication which is currently under review with the journal Oxford Bioinformatics.

7.2 Introduction

Within the field of *de novo* miRNA prediction, as described in section 3.3 of this thesis, putative pre-miRNA sequences which form miRNA-like hairpins are extracted from genomic data sets, and these sequences are classified based on the presence or absence of qualities such as structural stability, sequence motifs typical of miRNA, and structural robustness [95]. The advantage of *de novo* miRNA prediction is that only genomic sequence is required as input, not transcriptomic data. A disadvantage of such techniques is that they are ignorant of the actual expression of the candidate pre-miRNA region and must therefore

consider a far greater number of putative miRNA which may never be expressed. Recent advances in *de novo* sequence-based miRNA prediction have been derived primarily through the application of new pattern classification techniques to the miRNA prediction problem [110], and the introduction of new classes of classification features [113]. However, high class imbalance (1:1000 or higher) within genomic data sets limits the effectiveness of *de novo* classifiers on actual data sets, in spite of high performance often reported on small test data sets with artificially balanced frequencies of positive and negative exemplars [139].

Within the field of NGS-based miRNA prediction, as described in section 3.2 of this thesis, data are collected from next-generation sequencing (NGS) experiments. These data represent the sequence and relative abundance of all expressed RNA in a sample, including RNA arising from microRNA (true positives) and other sources including mRNA degradation products and other ncRNA. Predictions of novel miRNA are made from NGS data by seeking patterns of read depth (proxy for transcript abundance in the cell) indicative of processing by Drosha and Dicer endonuclease activity [52]. These techniques also often examine the strength of the miRNA:miRNA* duplex corresponding to the mature miRNA and miRNA* regions within a putative pre-miRNA region [60]. Expression-based techniques for miRNA prediction have seen success in recent years [82], [85], [140], which can be explained in part by the lower class imbalance present in NGS data sets. The number of false positives in a typical NGS experiment is on the order of tens of thousands [52], [141], whereas one expects tens of millions of miRNA-like structures in a typical genome [139]. Expression-based methods need only evaluate expressed regions, whereas *de novo* methods must evaluate all putative regions capable of forming hairpin structures. Furthermore, methods such as miRDeep2 [15] often filter by transcript abundance, considering only the most highly expressed regions as a means to reduce their computational runtime.

Considering both *de novo* and expression-based miRNA prediction techniques, multiple categories of sequence- and expression-based features have been explored, where each may provide independent support for the prediction of miRNA within NGS data sets. State-of-the-art expression-based miRNA prediction techniques, however, only leverage a limited set of these lines of evidence. MiRDeep2 [15] predicts miRNA based on the match between expression levels of NGS read data and expected Dicer processing, the stability of the miRNA:miRNA* duplex, and the significance of the minimum free energy of the pre-miRNA hairpin. MiRanalyzer [57] predicts miRNA based on secondary structure features, total read depth within the pre-miRNA region, expression of expected Dicer products, and minimum free energy features. These two methods have emerged as standards within the field of expression-based miRNA prediction. No existent miRNA prediction method employs the full range of features available from NGS data sets, which include both sequence-based and expression-based features. We hypothesize that through the combination of all previously described independent lines of evidence, miRNA prediction performance can be improved.

We here improve on the state-of-the-art performance of expression-based miRNA prediction by integrating the full range of sequence-based and expression-based features to create a novel miRNA predictor. We refer to this new method as miPIE (miRNA Prediction using Integrated Evidence). Our predictor is built using rigorous machine learning techniques, and tested using the metrics of recall and precision, which are directly applicable to real-world miRNA prediction. Additionally, unlike previous methods for expression-based miRNA prediction [15], [57], all features used in our experiment are invariant to experiment size (total read depth). As NGS technology improves and read depths continue to increase, it is important that all features have this property in order for predictors to be effective on future data sets.

7.3 Methods

7.3.1 Data set selection

Sample data were selected from the NCBI GEO database, using a query consisting of the keywords "small RNA" and an organism name. Samples were selected for the following criteria: Extracted molecule is "total RNA", no infections or knockouts present in the cell, and size fractionation selection is for "small RNA". Though no specific selection was performed for instrument type, all samples were collected using the Illumina HiSeq 2000 instrument, with the exception of sample GSM1901968 which was collected using the Illumina HiSeq 1000 instrument. Table 15 describes the data sets which were retrieved for this experiment.

Table 15 - NGS data sets examined in this article

Data set	Organism	Accession	# Reads	Cell type
hsa1	H. sapiens	GSM1555749	21196809	Mature erythrocyte [142]
hsa2	H. sapiens	GSM1820470	38210937	Monocyte-derived macrophage [143]
mmu1	M. musculus	GSM1528810	54947527	Adult testes
mmu2	M. musculus	GSM1901968	25881937	Whole blood
dme	D. melanogaster	GSM1123781	18723989	Ovaries [144]

For each of the five samples collected, the following procedure was performed in order to develop positive and negative training sets: The miRDeep [15] pre-processing algorithm (as implemented in "mapper.pl") was applied to all data sets. This tool maps each read stack with at least 4 reads to the reference genome. Putative pre-miRNA regions are extracted (-10/+70nt and -70/+10nt windows based on locally maximal read stacks) and the secondary structure was computed to check for hairpin structures. This process resulted in a

set of candidate pre-miRNAs, each represented by a pre-miRNA sequence, pre-miRNA structure, and the set of reads which map to the sequence (mature, miRNA*, and loop regions). For each sample, all candidate pre-miRNA which were matched to known miRNA from miRBase 21.0 [4] using the miRDeep "quantifier.pl" algorithm were selected as true positive samples for training and test.

Each candidate mature miRNA not identified as miRNA in the previous step was then aligned to the respective species' coding region data. Alignment was performed using bowtie [70]. All candidate mature miRNAs which aligned to a coding region with at most two mismatches ("-v 2" bowtie parameter) were selected as negative samples for training and test. Coding region data was retrieved from the Ensembl sequence FTP database [145]. Table 16 lists the sizes of the final data sets used for this experiment.

Table 16 - Number of samples in positive and negative classification data sets derived from each NGS experiment data set

Data set	Number of positive samples	Number of negative samples
hsa1	562	868
hsa2	168	816
mmu1	384	1492
mmu2	498	3283
dme	110	148

7.3.2 Feature set selection

In this study, we examine a set of 223 sequence- and expression-based features. These features incorporate several distinct lines of evidence which have been shown to have predictive power for the classification of miRNA. Of these features, 215 are derived from the feature vector of the sequence-based method HeteroMiRPred [113], which in turn gathered these features from a number of methods dating back to 2005. These features all pertain to

pre-miRNA sequence and structure and include minimum free energy (MFE)-derived features, sequence/structure triplet features, z-features which encapsulate the significance of the RNA structure relative to those of permuted sequences, and structural robustness features which reflect the ability of the precursor structure to maintain its stability through addition or removal of nucleotides.

Eight expression-based features are added to these sequence-based features. These features are:

1. Percentage of mature miRNA nts which are paired
2. Number of pairs in lower stem (outside of mature and miRNA* regions)
3. Percentage of RNA-seq reads in region which are inconsistent with Dicer processing
4. Percentage of RNA-seq reads from the loop region which match Dicer processing
5. Percentage of RNA-seq reads from the mature miRNA which match Dicer processing
6. Percentage of RNA-seq reads from the miRNA* region which match Dicer processing
7. Percentage of RNA-seq reads which match Dicer processing
8. Total number of reads in the precursor region, normalized to experiment size

Here, a match between a read and expected Dicer processing is identical to the definition used by the miRDeep2 study [15]. A match occurs when a read which maps to a miRNA sequence overlaps the mature, miRNA*, or loop portions of a miRNA with at most 2-nt difference between starting positions on the 5’ ends of the sequences, and at most 5-nt difference between terminating positions on the 3’ ends of the sequences.

These expression-based features provide miRNA classification methods with additional independent lines of evidence for miRNA prediction which are not available from strictly sequence-based feature vectors. Features 1 and 2 provide information on the mature and lower stem regions of the miRNA, while features 3 through 8 provide information regarding the expression pattern within the miRNA region. While the number of expression-based features examined in this study is far less than the number of sequence-based features, it is greater than the number of expression-based features used by the miRDeep scoring algorithm (3: read count within mature region, miRNA* or loop regions; presence of

miRNA* reads matching dicer processing; ratio of reads in the pre-miRNA region which are consistent with Dicer processing) [15] and the miRanalyzer scoring algorithm (2: total read count; and expression of miRNA*) [57]. The remaining features used in these two methods actually pertain to the sequence and/or secondary structure of the putative pre-miRNA or homology to known miRNA.

The final feature set was determined using the correlation-based feature subset selection method [146] as implemented in the Weka package [147]. This algorithm determines an optimal feature set based on correlation between each feature and the class assignment (miRNA vs. pseudo-miRNA), and lack of inter-correlation between features within the feature set. The final feature vector is presented in section 7.4.1 of this thesis. Feature selection was performed on the mmu1 data set, and the resulting feature vector was subsequently employed across all data sets. As a result, performance results for the *mmu1* data set represent some optimization using the test set data. This potential source of bias is not evident in the performance results across the *mmu2*, *hsa1*, *hsa2*, and *dme* data sets where sustained performance is observed.

7.3.3 Classification pipeline

All miRNA classification in this experiment was performed using a random forest classifier of 500 trees. Trees were built according to the default parameters of the SKLearn random forest library [136]. Previous studies have demonstrated that random forest classification outperforms competing classifier types for the prediction of miRNA [110], [137]. The miRanalyzer method (see comparison in section 7.4.2 below) also employs a random forest classifier. Within individual data sets, 10-fold cross validation (10CV) was used to estimate classification performance. Within each fold, the SMOTE algorithm [43] was used to oversample the minority class of each training data set to parity with the majority class. Oversampling was performed only on training data sets; class imbalance within each test set was unchanged. When determining performance across data sets, a single classifier was

trained using the training data set, and this classifier was used to predict all samples from the hold-out data set.

7.4 Results

7.4.1 Final feature set

The final feature set selected by the correlation feature subset algorithm contains 20 features, which represent seven different classes of evidence for the prediction of miRNA. Sequence-based features relating to secondary structure {MFE3, dH, Tm, Tm/loop}, robustness {SC*absZG, SC/1dp}, base pairing {Probpair 2, 3, 7, 9, 19, and 94}, sequence/structure motifs {"C((.", "T.((", "T..("}, and sequence motifs {"CG", "GA"} were selected. Expression-based features pertaining to miRNA:miRNA* duplex structure (% of paired bases in mature sequence) and Dicer matching of NGS reads (% of reads matching mature, % of reads matching miRNA*) were also selected. The fact that automated feature selection arrived at a heterogeneous feature set including both sequence- and expression-based feature supports our hypothesis that an integration of multiple lines of evidence will lead to increased classification performance. Table 17 further describes all features which were selected for the final feature vector.

Table 17 - Features selected for the miPIE method

Feature	Description
MFE3	Ratio of normalized minimum free energy value and the number of loops in the pre-miRNA secondary structure.
dH	Enthalpy of the pre-miRNA secondary structure
Tm	Melting energy of the pre-miRNA secondary structure, and melting energy normalized by the loop length of the structure
Tm/loop	
SC x zG	SC is a measurement of change in normalized structural stability when a pre-miRNA sequence is extended or reduced by equal amounts on the 5' and 3' arms. zG is the z-score of the normalized minimum free energy of the pre-miRNA secondary structure. dP is the number of paired bases normalized to the sequence length.
SC/(1 – dP)	
Probpair2	Sum of pairing probabilities of short nucleotide motifs
Probpair3	
Probpair7	
Probpair9	
Probpair19	
Probpair94	
C((.	Triplet motifs, representing a nucleotide identity and pairing of the 5' neighbor, the nucleotide itself, and the 3' neighbor of the nucleotide
T.((	
T..(	
CG	Percentage of nucleotide dimers which contain the given nucleotide motif
GA	
% pb mature	Percentage of paired bases in mature miRNA sequence
% reads mature	Percentage of reads within pre-miRNA sequence which align to mature and miRNA* regions
% reads miRNA*	

7.4.2 Performance increase over existing method(s)

Here we demonstrate the performance increase which our methods achieve over existing state-of-the-art methods for expression-based miRNA prediction. MiPIE is compared against

the miRDeep and miRanalyzer methods, over the five data sets described previously. For each data set, the miRDeep and miRanalyzer prediction algorithms were run with default parameters over the positive and negative data sets. Performance for all methods is measured using a precision-recall curve. Test set class imbalance is unaltered and represents that of real-world data, as each test set represents the total amount of positive and negative data recovered and processed from an actual NGS experiment. Precision and recall are defined as follows:

$$Precision\ (Pr) = \frac{TP}{TP + FP}$$

$$Recall\ (Re) = \frac{TP}{TP + FN}$$

Achievable recall at 50% precision (Re@Pr50) and recall at 90% precision (Re@Pr90) are used as summary statistics. These statistics represent the recall rate achievable at an acceptable degree of precision for experimental validation (50%), and the percentage of miRNA which are correctly classified with very high confidence (90%), respectively. Figure 23 shows the performance of miPIE, miRDeep, and miRanalyzer over the five data sets. While it was possible to create a continuous PR-curve for miRDeep by measuring the Pr and Re at various decision thresholds, this was not possible for miRanalyzer since it produces binary predictions without associated probability scores. Therefore, miRanalyzer's performance is illustrated as a single point in the P-R space, reflecting its performance at the default decision threshold. As can be seen in these figures, miPIE outperforms both existing methods, particularly for decision thresholds corresponding to high precision (Pr=90%).

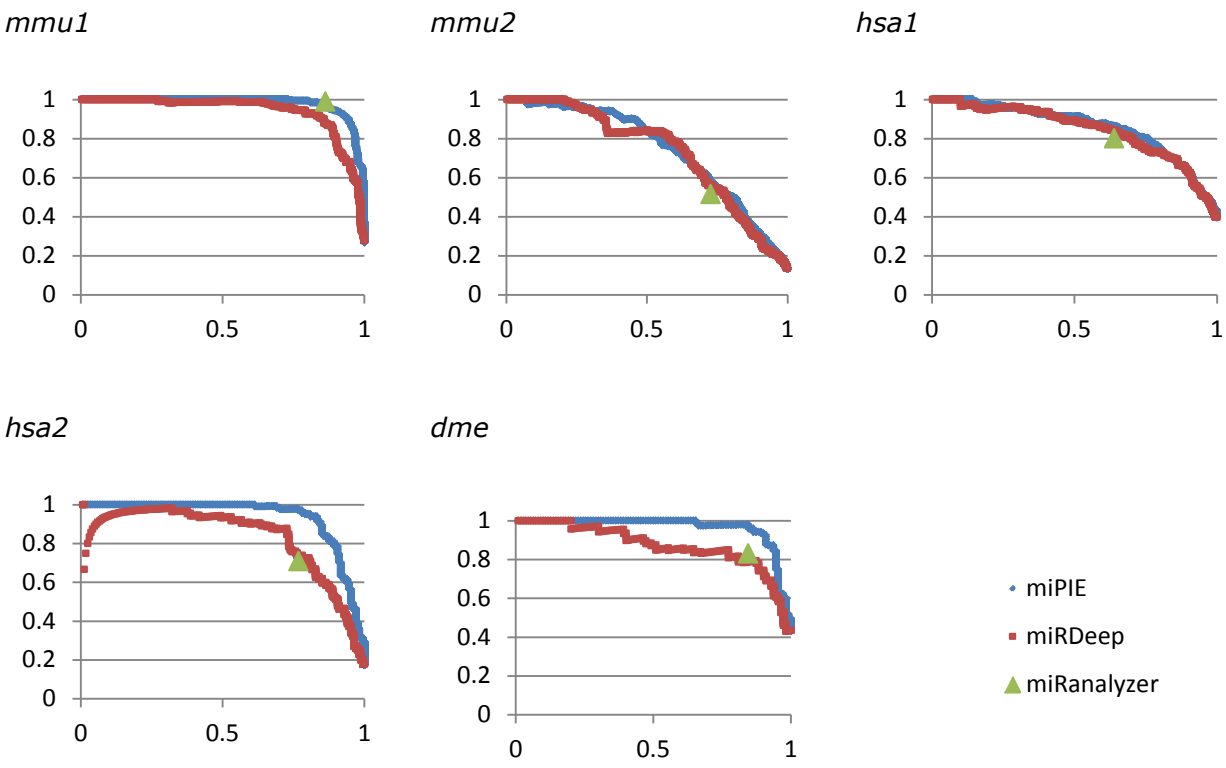


Figure 23 - Performance of miPIE, miRDeep2, and miRanalyzer across five data sets. miPIE performance is estimated through 10-fold cross validation. miRanalyzer produced binary prediction values, so only a single precision level is represented. miPIE outperforms miRDeep and miRanalyzer on all five data sets. In all plots, the y-axis represents precision while the x-axis is recall.

Table 18 summarizes the performance increase of miPIE over the miRDeep method. On average, our methods increase the number of high-confidence miRNA detected by 36.3%, while recall rates at Pr=50% are increased by 2.3%. The observed increase in Re@Pr50 is limited by a saturation effect, as both methods are approaching perfect recall at this level of precision.

To compute the statistical significance of the observed differences in Re@Pr90 and Re@Pr50 scores between miPIE and miRDeep a randomization test was conducted. For the set of ranked results from each miPIE and miRDeep for a given data set, pseudo samples miPIE* and miRDeep* were repeatedly developed in the following manner: At each rank, the identities of the patterns at this rank from the miPIE and miRDeep result sets were assigned randomly, one to the miPIE* ranked result set and one to the miRDeep* ranked result set. The difference in Re@Pr50/90 scores between the two pseudo samples was then computed. In doing so, we enforced the null hypothesis that there is no difference in the way that ranks are assigned by each method. This test was repeated 100,000 times to build a distribution of differences of Re@Pr50 and Re@Pr90 scores. The percentage of differences which were observed to be equal to or higher than the differences in score between miPIE and miRDeep provides the p-value for the observed difference in miPIE and miRDeep scores. When applied to the five data sets, all of the Re@Pr90 differences were found to be statistically significant ($p<0.01$) with the exception of the *hsa1* data set. As expected from Figure 24, none of the Re@Pr50 differences were found to be significant at this level, due to the saturation effect discussed above.

It is possible that the discrepancy in performance between *hsa1* and other data sets could result from higher presence of the original miRDeep training data within the *hsa1* data set, relative to other data sets; this hypothesis cannot be confirmed due to ambiguity in both miRDeep manuscripts pertaining to the precise training data.

Table 18 - Summary of results comparing miPIE with the state of the art miRDeep method, on five NGS data sets. miPIE outperforms miRDeep by 36.3% at the 90% precision threshold, and at the 50% precision threshold, miPIE outperforms miRDeep by 2.3%. miPIE results are drawn from cross validation experiments, and the standard error of these results is listed. miRDeep results are drawn from a single experiment therefore no error information is available.

Data set	Re@Pr50		Re@Pr90	
	miRDeep	miPIE	miRDeep	miPIE
mmu1	.982	.997 +/- .003 (+1.6%)	.857	.945 +/- .008 (+10.2%)
mmu2	.781	.805 +/- .015 (+3.1%)	.345	.438 +/- .028 (+27.0%)
hsa1	.961	.964 +/- .008 (+0.1%)	.458	.545 +/- .051 (+19.0%)
hsa2	.905	.952 +/- .015 (+5.3%)	.631	.845 +/- .034 (+33.9%)
dme	.972	.991 +/- .009 (+1.8%)	.464	.909 +/- .038 (+96.0%)
Average	.920	.942 (+2.3%)	.551	.736 (+36.3%)

In order to compare our method with the miRanalyzer method, which provides only binary classification results at a single decision threshold value, Table 19 describes the relative recall rates of miRanalyzer and miPIE at the precision level achieved by the miRanalyzer classifier on each data set. On average, our method increases recall rate by 9.5% relative to miRanalyzer predictions at the stated precision levels. The only data set on which miRanalyzer outperforms miPIE is *mmu1*. The high performance of the miRanalyzer method on this data set may be explained by apparent similarities between this data set and a data set which was used to train miRanalyzer (NGS series GSE20384). Both *mmu1* and the miRanalyzer training set data are derived from mouse testes samples.

Table 19 - Summary of results comparing miPIE with miRanalyzer using five NGS data sets. When operating at miRanalyzer's precision threshold, the recall of miPIE outperforms miRanalyzer by 9.5% on average. miPIE results are drawn from cross validation experiments, and the standard error of these results is listed. miRanalzyer results are drawn from a single experiment therefore no error information is available.

Data set	Precision level	miRanalyzer recall rate	miPIE recall rate
mmu1	.988	.862	.807 +/- .019 (-6.4%)
mmu2	.516	.727	.787 +/- .026 (+8.3%)
hsa1	.800	.639	.756 +/- .024 (+18.3%)
hsa2	.709	.768	.911 +/- .028 (+18.6%)
dme	.830	.845	.945 +/- .023 (+11.8%)
Average	.767	.768	.841 (+9.5%)

7.4.3 Combining sequence- and expression-based features

In order to demonstrate our hypothesis that the predictive power of our method is a result of combining multiple lines of evidence for miRNA prediction, we have repeated our 10CV classification pipeline using two subsets of features present in our full original feature set: i) the *sequence* feature subset which contains a subset of 20 optimal features selected from the set of the 223 sequence-based features available to our classifier; ii) the *expression* feature subset, which contains the eight expression-based features examined in our study. These feature sets were used to train and test predictors for each of the five data sets described previously. Results of these experiments are shown in Figure 24 along with the performance of the miPIE classifier built using the integrated set (i.e. both sequence- and expression-based features). As demonstrated in the figure, performance is improved across all data sets when sequence and structure features are combined, relative to the use of only one category of features.

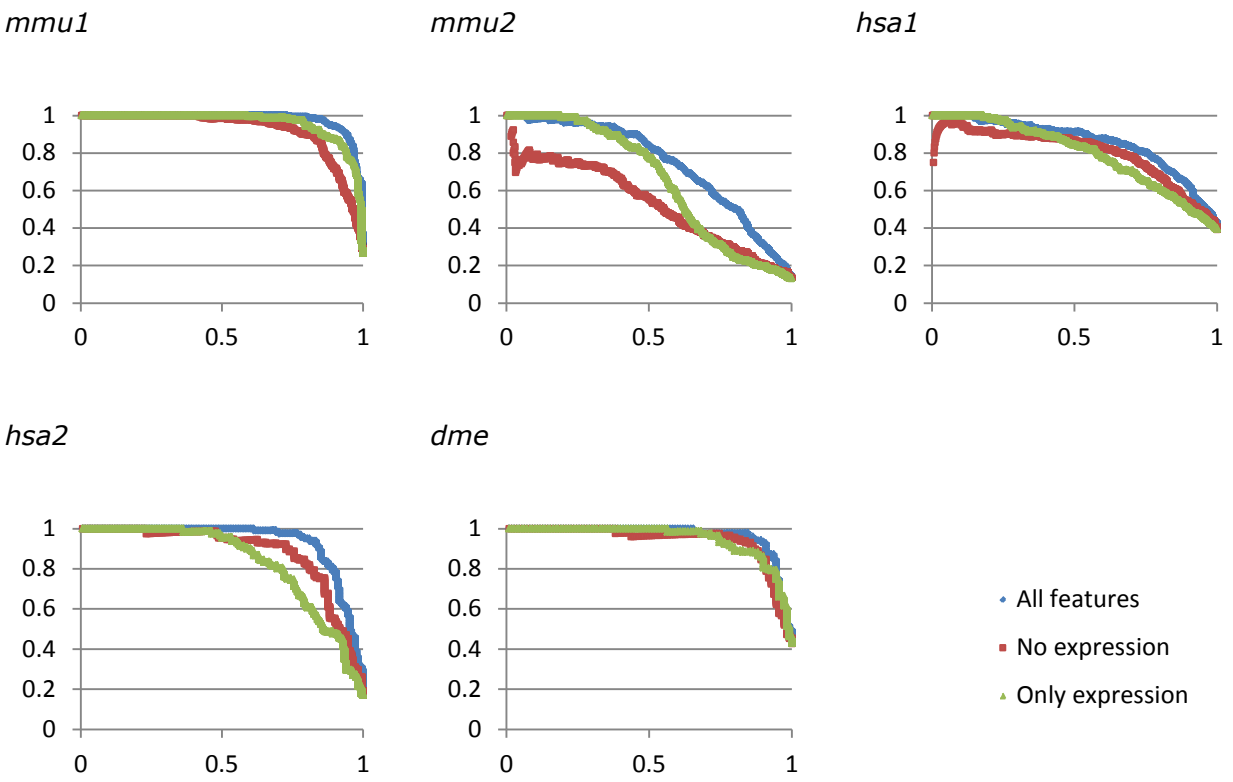


Figure 24 - Comparison of performance of the integrated miPIE feature set, relative to the performance of similarly trained classifiers trained using only sequence- and only expression-based features

7.4.4 Generalization across experiments

Here we demonstrate the ability of our methods to generalize across NGS data sets within and across species. For each of our five data sets, we compare the results of a 10CV experiment (where training and testing data arise from the same experiment) with classification of the same data set using models trained on each of the other data sets independently, as described in section 7.3.3 above. Finally, for each hold-out data set, a training set was built using the combination of all other data sets (labeled *all*). Figure 25 shows the PR-curves for this experiment over two of the five data sets (*hsa1* & *dme*) and Table 20 summarizes this performance for all five data sets. MiRDeep performance on the data sets is also shown for comparison. While some decrease in performance is observed when training data arises from a different experiment than the test data, the performance

of classifiers trained using all experiments except the test experiment is consistently strong (curve labelled *all* in Figure 25).

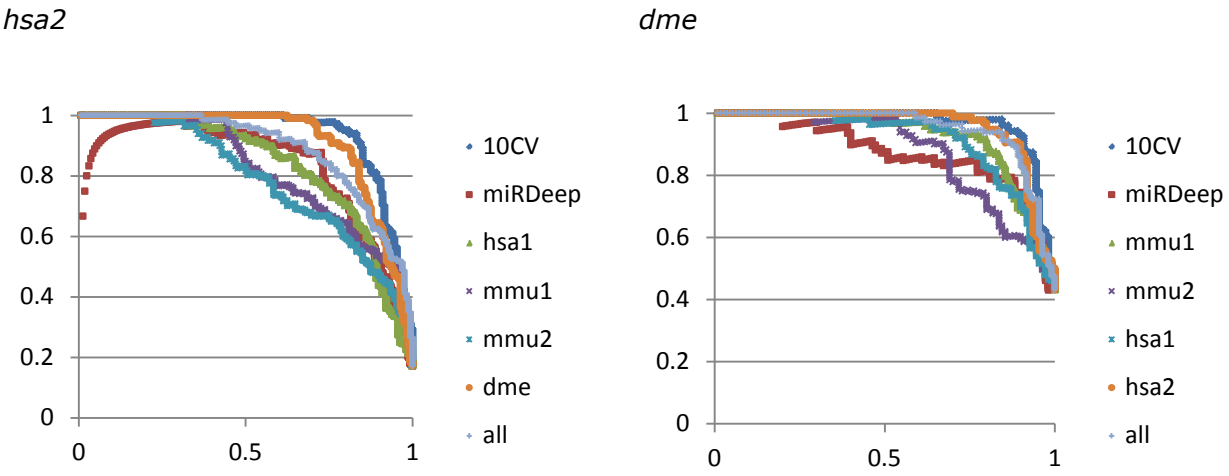


Figure 25 - Generalization performance of miPIE on the hsa2 and dme data sets. Regardless of the training set used, miPIE outperforms the state of the art method miRDeep on all data sets.

Table 20 lists the Re@Pr90 for each combined training and test set, in order of decreasing performance for each training set. From these results, we see that miPIE generalizes well to hold-out experimental data sets when training data sets from multiple NGS experiments are combined. Our method, when trained using all available data sets, outperforms miRDeep on four of five data sets. Average increase in Re@Pr90 between our method when trained in this manner and the miRDeep method is 20.3%. This combined *all* data sets perform as well or better than single experiment data sets in four of five experiments. Interestingly, experiments within the same species are not necessarily preferred here (e.g. the top-performing training set for human data set *hsa1* is the mouse *mmu2* data set). This result demonstrates that the miPIE method generalizes well to hold-out data sets across experiments and across species when training data is pooled from multiple training experiments.

Table 20 - Recall achievable at a precision of at least 90% (Re@Pr90) for 5 test data sets using our method trained over the following data sets: ALL=combination of 4 data sets, excluding test set; 10CV=10-fold cross-validation over test data set; *hsa1* = human data set 1; *hsa2* = human data set 2; *mmu1* = mouse data set 1; *mmu2* = mouse data set 2; *dme* = fruit fly data set. Additionally, miRDeep's performance over each test set is included as MD. The sampling errors of the 10CV experiments listed in this table are available in Table 18.

Test Set	Re@Pr90						
	1st	2nd	3rd	4th	5th	6th	7th
mmu1	10CV (.945)	All (.934)	Hsa2 (.919)	Hsa1 (.888)	Mmu2 (.875)	Dme (.867)	MD (.857)
mmu2	10CV (.438)	MD (.345)	All (.341)	Mmu1 (.284)	Hsa1 (.254)	Hsa2 (.121)	Dme (.063)
hsa1	10CV (.545)	All (.488)	MD (.458)	Mmu2 (.383)	Mmu1 (.353)	Dme (.273)	Hsa2 (.214)
hsa2	10CV (.845)	Dme (.780)	All (.679)	MD (.631)	Hsa1 (.565)	Mmu1 (.482)	Mmu2 (.429)
dme	10CV (.909)	Hsa2 (.900)	All (.872)	Mmu1 (.800)	Hsa1 (.736)	Mmu2 (.672)	MD (.464)

7.5 Conclusions

In this chapter, we introduce miPIE, a classification method for NGS-based miRNA prediction which integrates both sequence- and expression-based features and which employs best practices of pattern classification. All features used by miPIE are independent of the read count of the NGS experiment, such that the performance of this method will remain consistent as NGS technology continues to develop. miPIE is compared with two existing state-of-the-art methods using the metrics of precision and recall. At high precision levels (90%), miPIE increases recall by 36.3% relative to the popular miRDeep method. Furthermore, miPIE increases recall by 9.5% relative to the miRanalyzer algorithm, at the precision levels reported by miRanalyzer.

8 Summary and Future Recommendations

8.1 Conclusions

1. The integration of state-of-the-art sequence- and expression-based features, representing all known lines of evidence for the classification of miRNA, leads to improved miRNA prediction performance on NGS data sets. This performance generalizes across data sets and, through the use of these features, improvements can be made on the state of the art of NGS-based miRNA prediction.
2. The loss in specificity by state-of-the-art de novo miRNA prediction methods when applied to non-model species can be alleviated through the use of species-specific training data sets which employ phylogenetic and clustering information. Data sets which are built to resemble a non-model hold-out species based on this information outperform general taxon-wide data sets and data sets from model species. This increase in performance extends to the prediction of miRNA on previously unannotated species.
3. *De novo* miRNA prediction is performed at extremely high class imbalances (1:1000 or higher). The application of prevalence-corrected precision, alongside the analysis of genome-scanning miRNA prediction pipelines as a whole, provides a more accurate assessment of the performance of de novo miRNA prediction systems relative to metrics which are agnostic to class imbalance (i.e. geometric mean of sensitivity and specificity) applied only to model data sets.
4. The use of up-to-date RNA energy folding parameters increases the performance of *de novo* miRNA prediction pipelines, relative to the RNA energy folding parameters employed by state-of-the-art *de novo* miRNA prediction methods. This effect was observable using the genome-scanning model described in this thesis, but was not

observable using the metrics employed by previous state-of-the-art miRNA prediction studies.

8.2 Summary of contributions

1. A species-specific miRNA prediction training set generation procedure was developed for use with miRNA classification. This procedure increases performance of miRNA prediction studies on non-model organisms by generating data sets which capture the breadth of available miRNA training samples and which are optimized for the target species of a given miRNA prediction study.
2. 155 novel miRNA were discovered within two previously unannotated genomes, using the species-specific miRNA prediction training set generation framework which was introduced in this thesis.
3. The metrics of prevalence-corrected precision and recall, and the summary statistics Re@Pr50 and Re@Pr90 were introduced for the evaluation of miRNA classification. Models which are tuned for these metrics increase success rates of experimental validation of predictions, relative to models which are tuned to existing metrics in the field of miRNA prediction. These metrics also operate at the real-world class imbalance of miRNA prediction studies, providing more relevant measures of success for these studies.
4. A model was developed for the analysis of de novo miRNA prediction in the context of a genome-scanning miRNA prediction experiment. This model provides estimates of performance of a full miRNA prediction pipeline on genomic-scale data sets, which address a limitation of the current state of the art of de novo miRNA prediction, the optimization of methods for standard data sets which are not indicative of real-world miRNA prediction experiments.
5. A classification model for NGS-based miRNA prediction was developed which incorporates multiple lines of sequence- and expression-based evidence,

representing the state of the art of both *de novo* and NGS-based techniques. This model vastly improves on the performance of NGS-based miRNA prediction relative to current state-of-the-art models.

8.3 Recommendations for future work

The dissection of the miRNA prediction pipeline into pre-filtering and classification steps, as described in chapter 6 of this thesis, has proven successful for increasing performance in the area of *de novo* miRNA prediction. We believe that improvements to NGS-based miRNA prediction can also be made through similar analysis. We have identified, in chapter 3 of this thesis, that state-of-the-art methods of NGS-based miRNA prediction employ precursor selection techniques which systematically produce false negatives by limiting pre-miRNA sequence length ranges. Through the application of a model similar to that in chapter 5, the performance of these precursor selection techniques can be better characterized. Subsequently, improvements can be made to existing precursor selection techniques. Given that the primary piece of evidence which drives the selection of precursors is the presence of a mature miRNA, it is possible that precursor selection should be driven by determining the most likely mature miRNA:miRNA* duplex in the region surrounding the candidate mature miRNA. The pre-miRNA structure surrounding this duplex would then be optimized around the existing duplex. This method is in contrast to existing methods which do not take into account the mature miRNA evidence while building precursor structures.

An avenue for future improvement of miPIE, and of expression-based miRNA prediction as a whole, is the development of strong training data sets which combine data from multiple NGS experiments. As evidenced in section 7.4.3 of this thesis, miPIE's generalization performance increases with the incorporation of multiple training data sets, a result which is consistent with that of the miRanalyzer experiment [57]. Additionally, increasing the quality of training data has proven successful in the field of *de novo* miRNA prediction. With the

ever-increasing availability of NGS data across myriad species, it will be feasible to create larger training data sets incorporating more species. Once training data has been curated from many species, approaches such as the SMIRP framework can be applied to NGS-based miRNA prediction data sets, thereby increasing prediction performance on non-model species.

When investigating the integration of NGS and sequence based features, the present study optimized feature selection using a single data set from a single species. Future studies should investigate whether a single feature set is optimal for miRNA prediction experiments across multiple species, or whether species-specific feature set selection would improve miRNA prediction performance.

References

1. D. T. Humphreys, B. J. Westman, D. I. K. Martin, and T. Preiss, "MicroRNAs control translation initiation by inhibiting eukaryotic initiation factor 4E/cap and poly(A) tail function.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 47, pp. 16961–6, Nov. 2005.
2. D. BARTEL, "MicroRNAsGenomics, Biogenesis, Mechanism, and Function," *Cell*, vol. 116, no. 2, pp. 281–297, Jan. 2004.
3. K. C. Miranda, T. Huynh, Y. Tay, Y.-S. Ang, W.-L. Tam, A. M. Thomson, B. Lim, and I. Rigoutsos, "A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes.," *Cell*, vol. 126, no. 6, pp. 1203–17, Sep. 2006.
4. A. Kozomara and S. Griffiths-Jones, "miRBase: integrating microRNA annotation and deep-sequencing data.," *Nucleic Acids Res.*, vol. 39, no. Database issue, pp. D152–7, Jan. 2011.
5. A. La Torre, S. Georgi, and T. A. Reh, "Conserved microRNA pathway regulates developmental timing of retinal neurogenesis.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 26, pp. E2362–70, Jun. 2013.
6. M. T. N. Le, H. Xie, B. Zhou, P. H. Chia, P. Rizk, M. Um, G. Udolph, H. Yang, B. Lim, and H. F. Lodish, "MicroRNA-125b promotes neuronal differentiation in human cells by repressing multiple targets.," *Mol. Cell. Biol.*, vol. 29, no. 19, pp. 5290–305, Oct. 2009.
7. C. Körner, I. Keklikoglou, C. Bender, A. Wörner, E. Münstermann, and S. Wiemann, "MicroRNA-31 sensitizes human breast cells to apoptosis by direct targeting of protein kinase C epsilon (PKCepsilon).," *J. Biol. Chem.*, vol. 288, no. 12, pp. 8750–61, Mar. 2013.
8. Y. W. Iwasaki, K. Kiga, H. Kayo, Y. Fukuda-Yuzawa, J. Weise, T. Inada, M. Tomita, Y. Ishihama, and T. Fukao, "Global microRNA elevation by inducible Exportin 5 regulates cell cycle entry.," *RNA*, vol. 19, no. 4, pp. 490–7, Apr. 2013.
9. Y. Maistrovski, K. K. Biggar, and K. B. Storey, "HIF-1α regulation in mammalian hibernators: role of non-coding RNA in HIF-1α control during torpor in ground squirrels and bats.," *J. Comp. Physiol. B.*, vol. 182, no. 6, pp. 849–59, Aug. 2012.
10. A. Kowarsch, C. Marr, D. Schmidl, A. Ruepp, and F. J. Theis, "Tissue-specific target analysis of disease-associated microRNAs in human signaling pathways.," *PLoS One*, vol. 5, no. 6, p. e11154, Jan. 2010.
11. K. K. Biggar, S. F. Kornfeld, Y. Maistrovski, and K. B. Storey, "MicroRNA regulation in extreme environments: differential expression of microRNAs in the intertidal snail Littorina littorea during extended periods of freezing and anoxia.," *Genomics. Proteomics Bioinformatics*, vol. 10, no. 5, pp. 302–9, Oct. 2012.
12. K. K. Biggar and K. B. Storey, "Evidence for cell cycle suppression and microRNA regulation of cyclin D1 during anoxia exposure in turtles.," *Cell Cycle*, vol. 11, no. 9, pp. 1705–13, May 2012.
13. C.-W. Wu, K. K. Biggar, and K. B. Storey, "Dehydration mediated microRNA response in the African clawed frog Xenopus laevis.," *Gene*, vol. 529, no. 2, pp. 269–75, Oct. 2013.
14. P. Schaap, I. Barrantes, P. Minx, N. Sasaki, R. W. Anderson, M. Bénard, K. K. Biggar, N. E. Buchler, R. Bundschuh, X. Chen, C. Fronick, L. Fulton, G. Golderer, N. Jahn, V.

Knoop, L. F. Landweber, C. Maric, D. Miller, A. A. Noegel, R. Peace, G. Pierron, T. Sasaki, M. Schallenberg-Rüdinger, M. Schleicher, R. Singh, T. Spaller, K. B. Storey, T. Suzuki, C. Tomlinson, J. J. Tyson, W. C. Warren, E. R. Werner, G. Werner-Felmayer, R. K. Wilson, T. Winckler, J. M. Gott, G. Glöckner, and W. Marwan, "The Physarum polycephalum Genome Reveals Extensive Use of Prokaryotic Two-component and Metazoan-type Tyrosine Kinase Signaling.," *Genome Biol. Evol.*, p. evv237–, Nov. 2015.

15. M. R. Friedländer, S. D. Mackowiak, N. Li, W. Chen, and N. Rajewsky, "miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades.," *Nucleic Acids Res.*, vol. 40, no. 1, pp. 37–52, Jan. 2012.

16. M. Mraz and S. Pospisilova, "MicroRNAs in chronic lymphocytic leukemia: from causality to associations and back.," *Expert Rev. Hematol.*, vol. 5, no. 6, pp. 579–81, 2012.

17. V. Balatti, Y. Pekarky, and C. M. Croce, "Role of microRNA in chronic lymphocytic leukemia onset and progression," *J. Hematol. Oncol.*, vol. 8, no. 1, pp. 8–13, 2015.

18. B. Kusenda, M. Mraz, J. Mayer, and S. Pospisilova, "MicroRNA biogenesis, functionality and cancer relevance.," *Biomed. Pap. Med. Fac. Univ. Palack??, Olomouc, Czechoslov.*, vol. 150, no. 2, pp. 205–215, 2006.

19. Y. Zhao, J. F. Ransom, A. Li, V. Vedantham, M. von Drehle, A. N. Muth, T. Tsuchihashi, M. T. McManus, R. J. Schwartz, and D. Srivastava, "Dysregulation of Cardiogenesis, Cardiac Conduction, and Cell Cycle in Mice Lacking miRNA-1-2," *Cell*, vol. 129, no. 2, pp. 303–317, 2007.

20. J.-F. Chen, E. P. Murchison, R. Tang, T. E. Callis, M. Tatsuguchi, Z. Deng, M. Rojas, S. M. Hammond, M. D. Schneider, C. H. Selzman, G. Meissner, C. Patterson, G. J. Hannon, and D.-Z. Wang, "Targeted deletion of Dicer in the heart leads to dilated cardiomyopathy and heart failure.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 6, pp. 2111–2116, 2008.

21. 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.*, vol. 23, no. 20, pp. 4051–60, Oct. 2004.

22. X. Cai, C. H. Hagedorn, and B. R. Cullen, "Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs.," *RNA*, vol. 10, no. 12, pp. 1957–66, Dec. 2004.

23. R. I. Gregory, T. P. Chendrimada, and R. Shiekhattar, "MicroRNA biogenesis: isolation and characterization of the microprocessor complex.," *Methods Mol. Biol.*, vol. 342, pp. 33–47, Jan. 2006.

24. E. Berezikov, W. J. Chung, J. Willis, E. Cuppen, and E. C. Lai, "Mammalian Mirtron Genes," *Mol. Cell*, vol. 28, no. 2, pp. 328–336, 2007.

25. E. Lund and J. E. Dahlberg, "Substrate selectivity of exportin 5 and Dicer in the biogenesis of microRNAs.," *Cold Spring Harb. Symp. Quant. Biol.*, vol. 71, no. 0, pp. 59–66, Jan. 2006.

26. T. M. Rana, "Illuminating the silence: understanding the structure and function of small RNAs.," *Nat. Rev. Mol. Cell Biol.*, vol. 8, no. 1, pp. 23–36, Jan. 2007.

27. C. Lelandais-Brière, C. Sorin, M. Declerck, A. Benslimane, M. Crespi, and C. Hartmann, "Small RNA diversity in plants and its impact in development.," *Curr. Genomics*, vol. 11, no. 1, pp. 14–23, 2010.

28. B. M. Wheeler, A. M. Heimberg, V. N. Moy, E. A. Sperling, T. W. Holstein, S. Heber,

and K. J. Peterson, "The deep evolution of metazoan microRNAs.," *Evol. Dev.*, vol. 11, no. 1, pp. 50–68, Jan. 2009.

29. N. Fahlgren, S. Jogdeo, K. D. Kasschau, C. M. Sullivan, E. J. Chapman, S. Laubinger, L. M. Smith, M. Dasenko, S. A. Givan, D. Weigel, and J. C. Carrington, "MicroRNA gene evolution in Arabidopsis lyrata and Arabidopsis thaliana.," *Plant Cell*, vol. 22, no. 4, pp. 1074–89, Apr. 2010.

30. N. Warthmann, S. Das, C. Lanz, and D. Weigel, "Comparative analysis of the MIR319a microRNA locus in Arabidopsis and related Brassicaceae.," *Mol. Biol. Evol.*, vol. 25, no. 5, pp. 892–902, May 2008.

31. H. B. Shaffer, P. Minx, D. E. Warren, A. M. Shedlock, R. C. Thomson, N. Valenzuela, J. Abramyan, C. T. Amemiya, D. Badenhorst, K. K. Biggar, G. M. Borchert, C. W. Botka, R. M. Bowden, E. L. Braun, A. M. Bronikowski, B. G. Bruneau, L. T. Buck, B. Capel, T. a Castoe, M. Czerwinski, K. D. Delehaunty, S. V Edwards, C. C. Fronick, M. K. Fujita, L. Fulton, T. a Graves, R. E. Green, W. Haerty, R. Hariharan, O. Hernandez, L. W. Hillier, A. K. Holloway, D. Janes, F. J. Janzen, C. Kandoth, L. Kong, A. J. de Koning, Y. Li, R. Literman, S. E. McGaugh, L. Mork, M. O'Laughlin, R. T. Paitz, D. D. Pollock, C. P. Ponting, S. Radhakrishnan, B. J. Raney, J. M. Richman, J. St John, T. Schwartz, A. Sethuraman, P. Q. Spinks, K. B. Storey, N. Thane, T. Vinar, L. M. Zimmerman, W. C. Warren, E. R. Mardis, and R. K. Wilson, "The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage.," *Genome Biol.*, vol. 14, no. 3, p. R28, Mar. 2013.

32. 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*, vol. 75, no. 5, pp. 843–54, Dec. 1993.

33. B. J. Reinhart, F. J. Slack, M. Basson, A. E. Pasquinelli, J. C. Bettinger, A. E. Rougvie, H. R. Horvitz, and G. Ruvkun, "The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans.," *Nature*, vol. 403, no. 6772, pp. 901–6, Feb. 2000.

34. a E. Pasquinelli, B. J. Reinhart, F. Slack, M. Q. Martindale, M. I. Kuroda, B. Maller, D. C. Hayward, E. E. Ball, B. Degnan, P. Müller, J. Spring, a Srinivasan, M. Fishman, J. Finnerty, J. Corbo, M. Levine, P. Leahy, E. Davidson, and G. Ruvkun, "Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA.," *Nature*, vol. 408, no. 6808, pp. 86–89, 2000.

35. A. M. Lagos-quintana, R. Rauhut, W. Lendeckel, and T. Tuschl, "Identification of novel genes Coding for RNAs of Small expressed RNAs," *Science (80-.).*, vol. 294, no. 5543, pp. 853–858, 2001.

36. N. C. Lau, L. P. Lim, E. G. Weinstein, and D. P. Bartel, "An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans.," *Science*, vol. 294, no. 5543, pp. 858–862, 2001.

37. J. Fleenor, S. Xu, C. Mello, A. Fire, M. Cell, R. C. Lee, and V. Ambros, "An Extensive Class of Small RNAs in Caenorhabditis elegans," *Science*, vol. 294, no. 5543, pp. 862–864, 2001.

38. E. Berezikov, E. Cuppen, and R. H. a Plasterk, "Approaches to microRNA discovery.," *Nat. Genet.*, vol. 38 Suppl, no. May, pp. S2–7, Jun. 2006.

39. E. Pettersson, J. Lundeberg, and A. Ahmadian, "Generations of sequencing technologies," *Genomics*, vol. 93, no. 2, pp. 105–111, 2009.

40. R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, vol. 9. John Wiley & Sons, 2012.

41. S.-D. Hsu, Y.-T. Tseng, S. Shrestha, Y.-L. Lin, A. Khaleel, C.-H. Chou, C.-F. Chu, H.-D. H.-Y. Huang, C.-M. Lin, S.-Y. Ho, T.-Y. Jian, F.-M. Lin, T.-H. Chang, S.-L. Weng, K.-W. Liao, I.-E. Liao, C.-C. Liu, and H.-D. H.-Y. Huang, "miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions.," *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. D78–85, Jan. 2014.

42. G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. 14, no. 1, 1968.

43. N. Chawla and K. Bowyer, "SMOTE: synthetic minority over-sampling technique," *J. Artificial Intell. Res.*, vol. 16, pp. 321–357, 2011.

44. D. H. Wolpert, "A Lack of A Priori Distinctions Between Learning Algorithms," *Neural Comput.*, vol. 8, no. 7, pp. 1391–1420, 1996.

45. L. P. Lim, N. C. Lau, E. G. Weinstein, A. Abdelhakim, S. Yekta, M. W. Rhoades, C. B. Burge, and D. P. Bartel, "The microRNAs of Caenorhabditis elegans," *Genes Dev*, pp. 991–1008, 2003.

46. E. Lai, P. Tomancak, R. Williams, and G. Rubin, "Computational identification of Drosophila microRNA genes," *Genome Biol*, vol. 4, no. 7, pp. 1–20, 2003.

47. Y. Grad, J. Aach, G. D. Hayes, B. J. Reinhart, G. M. Church, G. Ruvkun, and J. Kim, "Computational and Experimental Identification of C. elegans microRNAs," *Mol. Cell*, vol. 11, no. 5, pp. 1253–1263, May 2003.

48. M. W. Jones-Rhoades and D. P. Bartel, "Computational identification of plant microRNAs and their targets, including a stress-induced miRNA.," *Mol. Cell*, vol. 14, no. 6, pp. 787–99, Jun. 2004.

49. A. Adai and C. Johnson, "Computational prediction of miRNAs in Arabidopsis thaliana," *Genome Res.*, pp. 78–91, 2005.

50. T. Dezulian, M. Remmert, J. F. Palatnik, D. Weigel, and D. H. Huson, "Identification of plant microRNA homologs.," *Bioinformatics*, vol. 22, no. 3, pp. 359–60, Mar. 2006.

51. B. Sebastian and S. E. Aggrey, "MiR-Explore: predicting microRNA precursors by class grouping and secondary structure positional alignment.," *Bioinform. Biol. Insights*, vol. 7, pp. 133–42, Jan. 2013.

52. M. R. Friedländer, W. Chen, C. Adamidi, J. Maaskola, R. Einspanier, S. Knespel, and N. Rajewsky, "Discovering microRNAs from deep sequencing data using miRDeep.," *Nat. Biotechnol.*, vol. 26, no. 4, pp. 407–15, Apr. 2008.

53. N. D. Mendes, a T. Freitas, and M.-F. Sagot, "Current tools for the identification of miRNA genes and their targets.," *Nucleic Acids Res.*, vol. 37, no. 8, pp. 2419–33, May 2009.

54. B. M. Automating and S. Heber, "Automating the Annotation and Discovery of MicroRNA in Multi-species High-throughput 454 Sequencing," North Carolina State University.

55. W.-C. Wang, F.-M. Lin, W.-C. Chang, K.-Y. Lin, H.-D. Huang, and N.-S. Lin, "miRExpress: analyzing high-throughput sequencing data for profiling microRNA expression.," *BMC Bioinformatics*, vol. 10, no. 1, p. 328, Jan. 2009.

56. D. Hendrix, M. Levine, and W. Shi, "miRTRAP, a computational method for the systematic identification of miRNAs from high throughput sequencing data.," *Genome Biol.*, vol. 11, no. 4, p. R39, Jan. 2010.

57. M. Hackenberg, N. Rodríguez-Ezpeleta, and A. M. Aransay, "miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing

experiments.," *Nucleic Acids Res.*, vol. 39, no. Web Server issue, pp. W132–8, Jul. 2011.

58. X. Yang and L. Li, "miRDeep-P: A computational tool for analyzing the microRNA transcriptome in plants," *Bioinformatics*, vol. 27, no. 18, pp. 2614–2615, 2011.

59. M. Menor, K. Baek, and G. Poisson, "Multiclass relevance units machine: benchmark evaluation and application to small ncRNA discovery.," *BMC Genomics*, vol. 14 Suppl 2, no. Suppl 2, p. S6, Jan. 2013.

60. D. Mapleson, S. Moxon, T. Dalmay, and V. Moulton, "MirPlex: a tool for identifying miRNAs in high-throughput sRNA datasets without a genome.," *J. Exp. Zool. B. Mol. Dev. Evol.*, vol. 320, no. 1, pp. 47–56, Jan. 2013.

61. J. An, J. Lai, M. L. Lehman, and C. C. Nelson, "MiRDeep*: An integrated application tool for miRNA identification from RNA sequencing data," *Nucleic Acids Res.*, vol. 41, no. 2, pp. 727–737, 2013.

62. J. An, J. Lai, A. Sajjanhar, M. L. Lehman, and C. C. Nelson, "miRPlant: an integrated tool for identification of plant miRNA from RNA sequencing data.," *BMC Bioinformatics*, vol. 15, no. 1, p. 275, Jan. 2014.

63. C. Kuenne, J. Preussner, M. Herzog, T. Braun, and M. Looso, "MIRPIPE: quantification of microRNAs in niche model organisms.," *Bioinformatics*, vol. 30, no. 23, pp. 3412–3, Dec. 2014.

64. T. B. Hansen, M. T. Venø, J. Kjems, and C. K. Damgaard, "miRdentify: high stringency miRNA predictor identifies several novel animal miRNAs.," *Nucleic Acids Res.*, vol. 42, no. 16, p. e124, Jan. 2014.

65. M. S. Menor, K. Baek, and G. Poisson, "Prediction of mature microRNA and piwi-interacting RNA without a genome reference or precursors.," *Int. J. Mol. Sci.*, vol. 16, no. 1, pp. 1466–81, Jan. 2015.

66. J. Wu, Q. Liu, X. Wang, J. Zheng, T. Wang, M. You, Z. Sheng Sun, and Q. Shi, "mirTools 2.0 for non-coding RNA discovery, profiling, and functional annotation based on high-throughput sequencing.," *RNA Biol.*, vol. 10, no. 7, pp. 1087–92, Jul. 2013.

67. Y. Gu, Y. Liu, J. Zhang, H. Liu, Y. Hu, H. Du, Y. Li, J. Chen, B. Wei, and Y. Huang, "Identification and characterization of microRNAs in the developing maize endosperm," *Genomics*, vol. 102, no. 5–6, pp. 472–478, 2013.

68. T. Madden, "The BLAST Sequence Analysis Tool," in *The NCBI Handbook*, National Center for Biotechnology Information (US), 2003.

69. T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *J. Mol. Biol.*, vol. 147, no. 1, pp. 195–197, Mar. 1981.

70. B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.," *Genome Biol.*, vol. 10, no. 3, p. R25, Jan. 2009.

71. B. C. Meyers, M. J. Axtell, B. Bartel, D. P. Bartel, D. Baulcombe, J. L. Bowman, X. Cao, J. C. Carrington, X. Chen, P. J. Green, S. Griffiths-Jones, S. E. Jacobsen, A. C. Mallory, R. A. Martienssen, R. S. Poethig, Y. Qi, H. Vaucheret, O. Voinnet, Y. Watanabe, D. Weigel, and J.-K. Zhu, "Criteria for annotation of plant MicroRNAs.," *Plant Cell*, vol. 20, no. 12, pp. 3186–90, Dec. 2008.

72. S. Moxon, F. Schwach, T. Dalmay, D. Maclean, D. J. Studholme, and V. Moulton, "A toolkit for analysing large-scale plant small RNA datasets.," *Bioinformatics*, vol. 24, no. 19, pp. 2252–3, Oct. 2008.

73. M. W. Jones-Rhoades, D. P. Bartel, and B. Bartel, "MicroRNAS and their regulatory roles in plants.," *Annu. Rev. Plant Biol.*, vol. 57, pp. 19–53, 2006.

74. P.-J. Huang, Y.-C. Liu, C.-C. Lee, W.-C. Lin, R. R.-C. Gan, P.-C. Lyu, and P. Tang, "DSAP: deep-sequencing small RNA analysis pipeline.," *Nucleic Acids Res.*, vol. 38, no. Web Server issue, pp. W385–91, Jul. 2010.

75. M. B. Stocks, S. Moxon, D. Mapleson, H. C. Woolfenden, I. Mohorianu, L. Folkes, F. Schwach, T. Dalmay, and V. Moulton, "The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets.," *Bioinformatics*, vol. 28, no. 15, pp. 2059–61, Aug. 2012.

76. M. Wen, Y. Shen, S. Shi, and T. Tang, "miREvo: an integrative microRNA evolutionary analysis platform for next-generation sequencing experiments.," *BMC Bioinformatics*, vol. 13, p. 140, Jan. 2012.

77. S. Cho, I. Jang, Y. Jun, S. Yoon, M. Ko, Y. Kwon, I. Choi, H. Chang, D. Ryu, B. Lee, V. N. Kim, W. Kim, and S. Lee, "MiRGator v3.0: A microRNA portal for deep sequencing, expression profiling and mRNA targeting," *Nucleic Acids Res.*, vol. 41, no. D1, pp. 252–257, 2013.

78. D. T. Humphreys and C. M. Suter, "miRspring: a compact standalone research tool for analyzing miRNA-seq data.," *Nucleic Acids Res.*, vol. 41, no. 15, p. e147, Aug. 2013.

79. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, "The Sequence Alignment/Map format and SAMtools.," *Bioinformatics*, vol. 25, no. 16, pp. 2078–9, Aug. 2009.

80. Z. Sun, J. Evans, A. Bhagwate, S. Middha, M. Bockol, H. Yan, and J.-P. Kocher, "CAP-miRSeq: a comprehensive analysis pipeline for microRNA sequencing data.," *BMC Genomics*, vol. 15, no. 1, p. 423, Jan. 2014.

81. G.-Z. Luo, W. Yang, Y.-K. Ma, and X.-J. Wang, "ISRNA: an integrative online toolkit for short reads from high-throughput sequencing data.," *Bioinformatics*, vol. 30, no. 3, pp. 434–6, Mar. 2014.

82. L. Yin, "Discovering novel microRNAs and age-related nonlinear changes in rat brains using deep sequencing," *Neurobiol. Aging*, vol. 36, no. 2, pp. 1037 – 1044.

83. S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biol*, vol. 11, no. 10, p. R106, 2010.

84. B. P. Lewis, C. B. Burge, and D. P. Bartel, "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.," *Cell*, vol. 120, no. 1, pp. 15–20, 2005.

85. C. Cowled, C. R. Stewart, V. A. Likic, M. R. Friedländer, M. Tachedjian, K. A. Jenkins, M. L. Tizard, P. Cottee, G. A. Marsh, P. Zhou, M. L. Baker, A. G. Bean, and L. Wang, "Characterisation of novel microRNAs in the Black flying fox (Pteropus alecto) by deep sequencing.," *BMC Genomics*, vol. 15, no. 1, p. 682, Jan. 2014.

86. R. Li, Y. Li, K. Kristiansen, and J. Wang, "SOAP: short oligonucleotide alignment program," *Bioinformatics*, vol. 24, no. 5, pp. 713–714, 2008.

87. S. Kadri, V. Hinman, and P. V Benos, "HHMMiR: efficient de novo prediction of microRNAs using hierarchical hidden Markov models.," *BMC Bioinformatics*, vol. 10 Suppl 1, p. S35, Jan. 2009.

88. P. Jiang, H. Wu, W. Wang, W. Ma, X. Sun, and Z. Lu, "MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features.," *Nucleic Acids Res.*, vol. 35, no. Web Server issue, pp. W339–44, Jul. 2007.

89. M. Leclercq, A. B. Diallo, and M. Blanchette, "Computational prediction of the localization of microRNAs within their pre-miRNA.," *Nucleic Acids Res.*, vol. 41, no. 15, pp. 7200–11, Aug. 2013.

90. E. Bonnet, Y. He, K. Billiau, and Y. Van de Peer, "TAPIR, a web server for the prediction of plant microRNA targets, including target mimics.," *Bioinformatics*, vol. 26, no. 12, pp. 1566–8, Jun. 2010.

91. V. Williamson, A. Kim, B. Xie, G. O. McMichael, Y. Gao, and V. Vladimirov, "Detecting miRNAs in deep-sequencing data: a software performance comparison and evaluation.," *Brief. Bioinform.*, vol. 14, no. 1, pp. 36–45, Jan. 2013.

92. R. Lorenz, S. H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker, "ViennaRNA Package 2.0," *Algorithms Mol. Biol.*, vol. 6, p. 26, 2011.

93. E. Bonnet, J. Wuyts, P. Rouzé, and Y. Van de Peer, "Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences.," *Bioinformatics*, vol. 20, no. 17, pp. 2911–7, Nov. 2004.

94. D. Kleftogiannis, A. Korfiati, K. Theofilatos, S. Likothanassis, A. Tsakalidis, and S. Mavroudi, "Where we stand, where we are moving: Surveying computational techniques for identifying miRNA genes and uncovering their regulatory role.," *J. Biomed. Inform.*, vol. 46, no. 3, pp. 563–73, Jun. 2013.

95. C. P. C. Gomes, J.-H. Cho, L. Hood, O. L. Franco, R. W. Pereira, and K. Wang, "A review of computational tools in microRNA discovery.," *Front. Genet.*, vol. 4, no. May, p. 81, Jan. 2013.

96. M. Yousef, L. Showe, and M. Showe, "A study of microRNAs in silico and in vivo: bioinformatics approaches to microRNA discovery and target identification.," *FEBS J.*, vol. 276, no. 8, pp. 2150–6, Apr. 2009.

97. L. Li, J. Xu, D. Yang, X. Tan, and H. Wang, "Computational approaches for microRNA studies: a review.," *Mamm. Genome*, vol. 21, no. 1–2, pp. 1–12, Feb. 2010.

98. L. Wei, M. Liao, Y. Gao, R. Ji, Z. He, and Q. Zou, "Improved and promising identification of human microRNAs by incorporating a high-quality negative set.," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, pp. 1–12, Nov. 2013.

99. M. Zuker, "Mfold web server for nucleic acid folding and hybridization prediction.," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3406–15, Jul. 2003.

100. N. R. Markham and M. Zuker, *UNAFold: software for nucleic acid folding and hybridization*, vol. 453. Totowa, NJ: Humana Press, 2008.

101. J.-W. Nam, K.-R. Shin, J. Han, Y. Lee, V. N. Kim, and B.-T. Zhang, "Human microRNA prediction through a probabilistic co-learning model of sequence and structure.," *Nucleic Acids Res.*, vol. 33, no. 11, pp. 3570–81, Jan. 2005.

102. A. Sewer, N. Paul, P. Landgraf, A. Aravin, S. Pfeffer, M. J. Brownstein, T. Tuschl, E. van Nimwegen, and M. Zavolan, "Identification of clustered microRNAs using an ab initio prediction method.," *BMC Bioinformatics*, vol. 6, p. 267, Jan. 2005.

103. C. Xue, F. Li, T. He, G. Liu, Y. Li, and X. Zhang, "Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine.," *BMC Bioinformatics*, vol. 6, p. 310, Jan. 2005.

104. R. Batuwita and V. Palade, "MicroPred: effective classification of pre-miRNAs for human miRNA gene prediction.," *Bioinformatics*, vol. 25, no. 8, pp. 989–95, Apr. 2009.

105. M. Yousef, S. Jung, L. C. Showe, and M. K. Showe, "Learning from positive examples when the negative class is undetermined--microRNA gene identification.," *Algorithms Mol. Biol.*, vol. 3, p. 2, Jan. 2008.

106. M. Yousef, M. Nebozhyn, H. Shatkay, S. Kanterakis, L. C. Showe, and M. K. Showe, "Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier.," *Bioinformatics*, vol. 22, no. 11, pp. 1325–34, Jun. 2006.

107. G. Grillo, A. Turi, F. Licciulli, F. Mignone, S. Liuni, S. Banfi, V. A. Gennarino, D. S. Horner, G. Pavesi, E. Picardi, and G. Pesole, "UTRdb and UTRsite (RELEASE 2010): A collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs," *Nucleic Acids Res.*, vol. 38, no. SUPPL.1, pp. 75–80, 2009.

108. A. Oulas, A. Boutla, K. Gkirtzou, M. Reczko, K. Kalantidis, and P. Poirazi, "Prediction of novel microRNA genes in cancer-associated genomic regions--a combined computational and experimental approach.," *Nucleic Acids Res.*, vol. 37, no. 10, pp. 3276–87, Jun. 2009.

109. J. Ding, S. Zhou, and J. Guan, "MiRenSVM: towards better prediction of microRNA precursors using an ensemble SVM classifier with multi-loop features.," *BMC Bioinformatics*, vol. 11 Suppl 1, no. Suppl 11, p. S11, Jan. 2010.

110. A. Gudyś, M. Szcześniak, M. W. Szcześniak, M. Sikora, and I. Makalowska, "HuntMi: an efficient and taxon-specific approach in pre-miRNA identification," *BMC Bioinformatics*, vol. 14, no. 1, p. 83, 2013.

111. K. L. Ng and S. K. Mishra, "De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures," *Bioinformatics*, vol. 23, no. 11, pp. 1321–1330, Jun. 2007.

112. A. Mathelier and A. Carbone, "MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data.," *Bioinformatics*, vol. 26, no. 18, pp. 2226–34, Sep. 2010.

113. S. Lertampaiporn, C. Thammarongtham, C. Nukoolkit, B. Kaewkamnerdpong, and M. Ruengjitchatchawalya, "Heterogeneous ensemble approach with discriminative features and modified-SMOTEbagging for pre-miRNA classification.," *Nucleic Acids Res.*, vol. 41, no. 1, p. e21, Jan. 2013.

114. K. Han, "Effective sample selection for classification of pre-miRNAs.," *Genet. Mol. Res.*, vol. 10, no. 1, pp. 506–18, Jan. 2011.

115. L. Wang, J. Li, R. Zhu, L. Xu, Y. He, and R. Zhang, "A novel stepwise support vector machine (SVM) method based on optimal feature combination for predicting miRNA precursors," *African J. Biotechnol.*, vol. 10, no. 74, pp. 16720–16731, Nov. 2011.

116. P. Xuan, M. Z. Guo, J. Wang, C. Y. Wang, X. Y. Liu, and Y. Liu, "Genetic algorithm-based efficient feature selection for classification of pre-miRNAs.," *Genet. Mol. Res.*, vol. 10, no. 2, pp. 588–603, Jan. 2011.

117. X. Liu, S. He, G. Skogerbø, F. Gong, and R. Chen, "Integrated sequence-structure motifs suffice to identify microRNA precursors.," *PLoS One*, vol. 7, no. 3, p. e32797, Jan. 2012.

118. N. Shakiba and L. Rueda, "MicroRNA identification using linear dimensionality reduction with explicit feature mapping.," *BMC Proc.*, vol. 7, no. Suppl 7, p. S8, Dec. 2013.

119. Q. Zou, Y. Mao, L. Hu, Y. Wu, and Z. Ji, "miRClassify: an advanced web server for miRNA family classification and annotation.," *Comput. Biol. Med.*, vol. 45, pp. 157–60, Mar. 2014.

120. D. Zhao, Y. Wang, D. Luo, X. Shi, L. Wang, D. Xu, J. Yu, and Y. Liang, "PMirP: a pre-microRNA prediction method based on structure-sequence hybrid features.," *Artif. Intell. Med.*, vol. 49, no. 2, pp. 127–32, Jun. 2010.

121. J. Xiao, X. Tang, Y. Li, Z. Fang, D. Ma, Y. He, and M. Li, "Identification of microRNA precursors based on random forest with network-level representation method of stem-loop structure.," *BMC Bioinformatics*, vol. 12, no. 1, p. 165, Jan. 2011.

122. Y. Xu, X. Zhou, and W. Zhang, "MicroRNA prediction with a novel ranking algorithm based on random walks.," *Bioinformatics*, vol. 24, no. 13, pp. i50–8, Jul. 2008.

123. D. M. W. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation," vol. 2, no. 1, Jan. 2008.

124. C. Chang and C. Lin, "LIBSVM: a library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011.

125. Y. Wu, B. Wei, H. Liu, T. Li, and S. Rayner, "MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences.," *BMC Bioinformatics*, vol. 12, no. 1, p. 107, Jan. 2011.

126. S. Tempel and F. Tahi, "A fast ab-initio method for predicting miRNA precursors in genomes.," *Nucleic Acids Res.*, vol. 40, no. 11, p. e80, Jun. 2012.

127. J. Meng, "Prediction of plant pre-microRNAs and their microRNAs in genome-scale sequences using structure-sequence features and support vector machine," *BMC Bioinformatics*, vol. 15, no. 1, p. 423, 2014.

128. M. D. Saçar, H. Hamzeiy, and J. Allmer, "Can MiRBase provide positive data for machine learning for the detection of MiRNA hairpins?," *J. Integr. Bioinform.*, vol. 10, no. 2, p. 215, Jan. 2013.

129. X. T. Dang, O. Hirose, T. Saethang, V. A. Tran, and L. A. T. Nguyen, "A novel over-sampling method and its application to miRNA prediction," *J. Biomed. Sci. Eng.*, vol. 6, no. 2, pp. 236–248, 2013.

130. S. Pitre, M. Hooshyar, A. Schoenrock, B. Samanfar, M. Jessulat, J. R. Green, F. Dehne, and A. Golshani, "Short Co-occurring Polypeptide Regions Can Predict Global Protein Interaction Maps.," *Sci. Rep.*, vol. 2, p. 239, Jan. 2012.

131. E. W. Sayers, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, I. Mizrachi, J. Ostell, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, E. Yaschenko, and J. Ye, "Database resources of the National Center for Biotechnology Information.," *Nucleic Acids Res.*, vol. 37, no. Database issue, pp. D5–15, Jan. 2009.

132. Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, "CD-HIT Suite: a web server for clustering and comparing biological sequences.," *Bioinformatics*, vol. 26, no. 5, pp. 680–2, Mar. 2010.

133. R. Leinonen, R. Akhtar, E. Birney, L. Bower, A. Cerdeno-Tárraga, Y. Cheng, I. Cleland, N. Faruque, N. Goodgame, R. Gibson, G. Hoad, M. Jang, N. Pakseresht, S. Plaister, R. Radhakrishnan, K. Reddy, S. Sobhany, P. Ten Hoopen, R. Vaughan, V. Zalunin, and G. Cochrane, "The European Nucleotide Archive.," *Nucleic Acids Res.*, vol. 39, no. Database issue, pp. D28–31, Jan. 2011.

134. I. Hofacker and W. Fontana, "Fast folding and comparison of RNA secondary structures," *Chem. Mon.*, vol. 125, no. 2, pp. 167–188, 1994.

135. I. Bentwich, "Prediction and validation of microRNAs and their targets.," *FEBS Lett.*,

vol. 579, no. 26, pp. 5904–10, Oct. 2005.

136. F. Pedregosa, G. Varoquaux, R. Weiss, and M. Brucher, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

137. D. H. Mathews, M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker, and D. H. Turner, "Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 19, pp. 7287–7292, 2004.

138. D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, "Entrez gene: Gene-centered information at NCBI," *Nucleic Acids Res.*, vol. 39, pp. 54–58, 2011.

139. R. J. Peace, K. K. Biggar, K. B. Storey, and J. R. Green, "A framework for improving microRNA prediction in non-human genomes," *Nucleic Acids Res.*, vol. 43, no. 20, p. e138, Jul. 2015.

140. Z. Agharbaoui, M. Leclercq, M. A. Remita, M. A. Badawi, E. Lord, M. Houde, J. Danyluk, A. B. Diallo, and F. Sarhan, "An integrative approach to identify hexaploid wheat miRNAome associated with development and tolerance to abiotic stress.," *BMC Genomics*, vol. 16, no. 1, p. 339, Apr. 2015.

141. M. Hackenberg, M. Sturm, D. Langenberger, J. M. Falcón-Pérez, and A. M. Aransay, "miRanalyzer: A microRNA detection and analysis tool for next-generation sequencing experiments," *Nucleic Acids Res.*, vol. 37, no. SUPPL. 2, 2009.

142. J. F. Doss, D. L. Corcoran, D. D. Jima, M. J. Telen, S. S. Dave, and J.-T. Chi, "A comprehensive joint analysis of the long and short RNA transcriptomes of human erythrocytes.," *BMC Genomics*, vol. 16, no. 1, p. 952, Jan. 2015.

143. V. Vongrad, J. Imig, P. Mohammadi, S. Kishore, L. Jaskiewicz, J. Hall, H. F. Günthard, N. Beerenwinkel, and K. J. Metzner, "HIV-1 RNAs are Not Part of the Argonaute 2 Associated RNA Interference Pathway in Macrophages.," *PLoS One*, vol. 10, no. 7, p. e0132127, Jan. 2015.

144. S. Shpiz, S. Ryazansky, I. Olovnikov, Y. Abramov, and A. Kalmykova, "Euchromatic transposon insertions trigger production of novel Pi- and endo-siRNAs at the target sites in the drosophila germline.," *PLoS Genet.*, vol. 10, no. 2, p. e1004138, Feb. 2014.

145. F. Cunningham, M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, L. Gil, C. G. Girón, L. Gordon, T. Hourlier, S. E. Hunt, S. H. Janacek, N. Johnson, T. Juettemann, A. K. Kähäri, S. Keenan, F. J. Martin, T. Maurel, W. McLaren, D. N. Murphy, R. Nag, B. Overduin, A. Parker, M. Patricio, E. Perry, M. Pignatelli, H. S. Riat, D. Sheppard, K. Taylor, A. Thormann, A. Vullo, S. P. Wilder, A. Zadissa, B. L. Aken, E. Birney, J. Harrow, R. Kinsella, M. Muffato, M. Ruffier, S. M. J. Searle, G. Spudich, S. J. Trevanion, A. Yates, D. R. Zerbino, and P. Flicek, "Ensembl 2015.," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D662–669, Oct. 2014.

146. M. a Hall, "Correlation-based Feature Selection for Machine Learning," *Methodology*, vol. 21i195–i20, no. April, pp. 1–5, 1999.

147. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software," *ACM SIGKDD Explor. Newsl.*, vol. 11, no. 1, p. 10, 2009.

148. A. Gudyś, M. M. W. Szcześniak, M. W. Szcze¿niak, M. Sikora, I. Makalowska, and I. Makałowska, "HuntMi: an efficient and taxon-specific approach in pre-miRNA identification.," *BMC Bioinformatics*, vol. 14, no. 1, p. 83, Jan. 2013.