COST AND BENEFIT ANALYSIS OF FEATURES USED IN MACHINE LEARNING BASED PRE-MIRNA DETECTION

A Thesis Submitted to
the Graduate School of Engineering and Sciences of
İzmir Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

in Molecular Biology and Genetics

by
Rabia SULUYAYLA

August 2016
İZMİR

We approve the thesis of **Rabia SULUYAYLA**

Examining Committee Members:

Assoc. Prof. Dr. Jens Allmer
Department of Molecular Biology and Genetics, Izmir Institute of Technology

---

Prof. Dr. Anne Frary
Department of Molecular Biology and Genetics, Izmir Institute of Technology

---

Assoc. Prof. Dr. Turgay Ünver
Izmir International Biomedicine and Genome Institute, Dokuz Eylül University

04 August 2016

Assoc. Prof. Dr. Jens Allmer
Supervisor, Department of Molecular Biology and Genetics
Izmir Institute of Technology

Prof. Dr. Volkan SEYRANTEPE
Head of the Department of
Molecular Biology and Genetics

Prof. Dr. Bilge KARAÇALI
Dean of the Graduate School of
Engineering and Sciences

ACKNOWLEDGMENTS

First of all, I would like to thank Assoc. Prof. Dr. Jens Allmer for being my advisor, accepting me to his scientific group, extending my perspective in science and assisting me during my master studies.

I am in debt to Prof. Dr. Anne Frary because of being in my thesis committee, for her informative courses and assistance during my master studentship.

I would like to thank Assoc. Prof. Dr. Turgay Ünver for accepting to be a thesis committee member, proofreading and scientific criticism.

I am grateful to Turgay Tekinay and Ayşe Begüm Tekinay, who have been my first group leaders, introduced me to science and from whose experience I learned much.

I would like to thank my friends who have been always there supportively during my good and bad times, being nice colleagues and for having enjoyable working times together.

To last but not least, I would like to thank my family without whom my life would be meaningless, for always being there, making everything possible in my life and creating a happy living environment.

Finally, I would like to thank TUBITAK (Grant number: 113E326) for the financial support during my master study.

ABSTRACT

COST AND BENEFIT ANALYSIS OF FEATURES USED IN MACHINE LEARNING BASED PRE-MIRNA DETECTION

MicroRNAs (miRNAs) are short RNA molecules which play important roles in the post-trancriptional regulation of gene expression. Their transcription is followed by two RNA III endonuclease processing steps leading to mature miRNA formation. They are then incorporated into the RISC-complex which mediates mRNA targeting. Experimental miRNA prediction is difficult since detection relies on many factors therefore, computational methods have become indispensable. Therefore, machine learning methods rely on features describing precursor-miRNAs (pre-miRNAs) to be able to differentiate them from other hairpins in a genome. It is important to define feature groups which are informative, not highly correlated, and don't incur a large computational cost in order to facilitate accurate miRNA detection. In this study for more than 800 pre-miRNA features the computational cost and benefit was analyzed. From these analyses five features (assl, lsr(%bp), lscm, asal and hpmfe_rf_I3), (four structural and one structural-thermodynamic one), which aren't correlated, informative and are not computationally expensive are noticeable. Analyses are done with human hairpins, pseudo data; and a case study using the measles virus and the measles KEGG pathway genes. Overall calculation of human hairpins and measles virus took approximately 2 USD (United States Dollar) on Amazon web services. Supervised learning and random forest machine learning for miRNA prediction was applied and to two genes (TAB2 and BCC3) within the measles KEGG pathway and three hairpins were predicted. They were found to have human mature miRNA sequences embedded in them and their already annotated targets helped enlarge the KEGG measles pathway.

ÖZET

MAKİNE ÖĞRENİMİNE DAYALI ÖNCÜL MİRNA TESPİTİNDE KULLANILAN ÖZELLİKLERİN FAYDA VE MALİYET ANALİZİ

Gen ifadesinin post-transkripsiyonel regülasyonunda önemli bir rolü olan kısa RNA moleküller mikroRNAlardır (miRNA). Transkripsionlarını iki RNAIII endonükleaz işlemi takip eder ve olgun miRNA oluşumuyla RISC-kompleksi mRNA hedeflemesini başlatır. Deneysel miRNA tahmini zordur çünkü miRNA ifadesni belirleme işlemi birçok faktöre dayanır bu yüzden bilişimsel metotlar daha umut vericidir. Genomdaki diğer saç tokası yapılarından (hairpin) ayırt edebilmek ve miRNA tespiti için, miRNAların karakteristik özellikleri tanımlanmalıdır. Bu sebeple, Veri Madenciliği metodları öncül miRNA (pre-miRNA) özelliklerini temel alır. Bu çalışmada 800den fazla pre-miRNA özelliğinin maliyet ve yarar analizi yapılmıştır. Bilgi kazanımı skoru özelliğin ne kadar ayırt edici olduğunu, Linear Korelasyon katsayısı özelliklerin birbirleriyle nasıl bağlı olduğunu ve zaman ölçümü de bir özelliğin ne kadar bilişimsel maliyetinin olduğunu gösterir. Sonuç- lardan yavaş olmayan ve bilgi verici beş özellik (assl, lsr(%bp), lscm, asal and hpmfe_rfI3) (dört yapısal ve bir yapısal-enerjik) seçildi ve birbiriyle korelasyonları olmadığı görüldü. Analizler insan hairpin, sözde (pseudo) veri ve kızamık (measles) virüsü, Measles İnsan KEGG Patikası genleri ile yapılmıştır. İnsan hairpin ve measles virüsünün genel hesaplanması Amazon serverında yaklaşık olarak 2 USD (Amerikan Doları) tutmuştur. Gözetimli öğrenme ve Rastgele Orman karar ağacı Veri Madenciliği kullanılarak iki measles KEGG patikası geninden (TAB2 and BCC3) üç miRNA tahmin edilmiştir. Bunlarda olgun miRNA dizlileri gömülü bulunmuştur.

To my family, a birthday present to my mother and father.

TABLE OF CONTENTS

LIST OF FIGURES

Figure **Page**

LIST OF TABLES

Table **Page**

LIST OF ABBREVIATIONS

AGO . . . . . Argonaute
DGCR8 . . . . . DiGeorge syndrome critical region gene 8
dns . . . . . dinucleotide shuffling method
dsRDB . . . . . dsRNA binding domain protein
dsRNA . . . . . double stranded RNA
EXP5 . . . . . exportin 5
kb . . . . . kilobase, 1000 nucleotides
KNIME . . . . . Konstanz Information Miner
MeV . . . . . measles virus
miRNA . . . . . microRNA
mRNA . . . . . messenger RNA
piRNA . . . . . PIWI-interacting RNA
PolII . . . . . RNA polymerase II
pre-miRNA . . . . . precursor miRNA
pri-miRNA . . . . . primary miRNA
RISC . . . . . RNA-induced silencing complex
RNA . . . . . Ribonucleic acid
siRNA . . . . . small interfering RNA
sl . . . . . stem length
ssRNA . . . . . single stranded RNA

CHAPTER 1

INTRODUCTION

Eukaryotes possess different small Ribonucleic acid (RNA) molecules which function to regulate mRNAs by targeting. (Bartel, 2004; He and Hannon, 2004). These molecules are defined by their length and differentiated according to their generation, relation with Argonaute family proteins (AGO family proteins) biological role etc. (Latchman, 2010; Meister, 2013) . In animals they can be grouped as: microRNA (miRNA), small interfering RNA (siRNA), PIWI-interacting RNA (piRNA) (Latchman, 2010).

1.1. microRNAs

Mature microRNAs (miRNAs) are approximately 22 nucleotide long, small, single stranded RNA molecules which are found in many living organisms (Bartel, 2004; Melo and Melo, 2014)

There are over more than 1900 miRNA coding genes identified in humans and high prevalence (>50%) are located in protein coding genomic regions (Hinske et al., 2014; Rodriguez et al., 2004). miRNAs are classified as intragenic and intergenic which can be subclassified to transcription classes as intronic and exonic (Hinske et al., 2014; Rodriguez et al., 2004). In human nearly 61% of the miRNAs are intragenic and whereas the remaning 39% are intergenic (França et al., 2016). Intragenic miRNA shows coordinated expression with their host genes (Baskerville and Bartel, 2005; França et al., 2016; Hinske et al., 2010).

miRNAs function in post-transcriptional regulation of gene expression, they have effect on messenger RNA (mRNA) stability by targeting and regulating the rate of translation and they can lead RNA silencing (Filipowicz et al., 2008; He and Hannon, 2004; Melo and Melo, 2014). miRNAs are found dominantly in somatic tissues compared to other small RNAs and furthermore they are defined as important in developmental stages (Ha and Kim, 2014). The biogenesis of miRNAs is strictly controlled and their distribution at regulation can lead to cancer and neurodevelopmental diseases in human (Bartel, 2004; Melo and Melo, 2014).

1.2. miRNA Biogenesis

In human, RNA polymerase II (PolII) transcription is the factor that initiates miRNA biogenesis which brings out mature miRNAs (Melo and Melo, 2014). The transcription product is a large, generally over 1 kilobase (kb), transcript, called primary miRNA (pri-miRNA) (Melo and Melo, 2014). Primary miRNA (pri-miRNA) has the sequence of miRNA embedded and the hairpin structure consist of stem and terminal loop which is formed by base pairing, therefore forming relatively double stranded structure (Ha and Kim, 2014). Pri-miRNA has single stranded RNA (ssRNA) flanks at both ends; a 5' end capped-spliced, a 3' end polyadenylated (Latchman, 2010). When the pri-miRNA is recognized from its stem and ssRNA flanks by a double stranded RNA (dsRNA) binding domain protein (dsRDB), DiGeorge syndrome critical region gene 8 (DGCR8) which forms the microprocessor with a nuclear Rnase III protein Drosha which cuts the stem loop and an approximately 68 nucleotide long hairpin precursor miRNA (pre-miRNA) is formed (Ha and Kim, 2014) (Figure 1.1). It is further translocated from nucleus to cytoplasm with exportin 5 (EXP5) by nuclear pore complexes and RAN-GTP (Melo and Melo, 2014). The pre-miRNA is processed to a small RNA duplex upon cleavage of the terminal loop by Dicer which is like Drosha a RNA III endonuclease. This 22 nt duplex consists of guide and passenger strands and is loaded to AGO which facilitates the pre-RNA-induced silencing complex (RISC) formation (Meister, 2013). Upon the removal of passenger RNA, mature RISC forms which has the mature RNA. Then after the base pairing of miRNA:mRNA can lead to gene silencing (Ha and Kim, 2014).

1.3. miRNA Identification

In mammals, nearly 30% of protein-coding genes are under miRNA regulation (Filipowicz et al., 2008), (Naik et al., 2013). Further, in humans, 60% of protein coding genes have at least one conserved miRNA binding site (Ha and Kim, 2014). As miRNAs are key players in translational regulation and mRNA stability, having role in many diseases; it has been important to study miRNAs and therefore many studies are conducted experimentally and computationally. Although hundreds of miRNAs have been identified, many of them may remain unknown. In order to identify miRNAs, computational approaches are more promising than experimental ones as experimental studies

are depended to biological circumstances like having mRNAs and miRNAs expressed together so that the influence of miRNA can be observed or many to one relationship between mRNA and miRNA. However, that doesn't mean computational approaches are sufficient, the findings still needs to be confirmed experimentally.

1.4. Computational Methods to Identify miRNAs

Homology based methods or *ab initio* detection are the computational methods applied to identify miRNAs. Homology modelling is used for conserved gene clusters when the genome sequence is known for comparative genomic studies, therefore, it is not applicable for prediction of novel miRNAs. In order to reveal unknown miRNAs ab initio methods are more promising as it does not require any database (Yousef and Allmer, 2012b). To use ab initio computational methods miRNAs have to be differentiated and unique characteristic features should be defined. As they are not the only hairpins produced from the genome, specific features of pre-miRNAs should be used for accurate and beneficial prediction (Bağcı and Allmer, 2016; Saçar and Allmer, 2013; Xue et al., 2005).

1.5. miRNA Features

In order to come up with specific and commonly understood features, already existing features are reviewed, the feature meaning is enlarged logically, and has been enhanced with our own features and a wide list of more than 800 features that identify pre-miRNAs are collected (Bağcı and Allmer, 2016; Saçar and Allmer, 2013; Yousef et al., 2016). These features (including published ones and their logical extensions) can be grouped into four main categories: Sequence based features are for example dinucleotide frequencies (%NN) (Ng and Mishra, 2007), direct internal repeats (dr) (Bentwich, 2008), and inverted internal repeat (ir) (Bentwich, 2008). Structural features are for instance triplet elements (N...) (Xue et al., 2005), hairpin length (hpl) (Bentwich et al., 2005), hairpin loop length (hll) (Bentwich, 2008), matching base pairs (bpp) (Ng and Mishra, 2007), and maximal bulge size (mbs)(Bentwich, 2008). Thermodynamic-based features are, among others, ensemble free energy (efe), ensemble frequency (efq), melting tem-

perature (Tm) (Ding et al., 2010), and enthalpy (dH). Probability-based features can be derived from any other feature using dinucleotide shuffling (dns) (Jiang et al., 2007) and some examples used are adjusted base pairing propensity (dns_p(bpp)) and adjusted minimum free energy of folding (dns_p(hpmfe_rf)). One of our studies is to represent each feature in a descriptive format so that the feature is understandable and mistakes are prevented. Below Figure 1.2 is shown as an example created from obtained results.

1.6. Cost and Benefits of pre-miRNA Features

Features described in Section 1.5 and Table A1 are mined from assumed pre-miRNAs which are found as folded to differentiate between true and false pre-miRNAs. Supervised learning is applied in machine learning, where the Random Forest Model is trained by pseudo hairpins (negative class) and miRBase v20 driven human hairpins (positive class) (Bağcı and Allmer, 2016; Kozomara and Griffiths-Jones, 2014; Ng and Mishra, 2007; Saçar and Allmer, 2013). Thus, computational pre-miRNA prediction costs, takes calculation time and the model accuracy is highly depended on features (Ding et al., 2010). However, feature selection is NP hard (Amaldi and Kann, 1998) , especially high number, 800 features make the selection more complicated. Information gain ratio is a measure to rank features in machine learning approaches, which is used for selection of features (Jiang et al., 2007; Khalifa et al., 2016; Xue et al., 2005). The positive and negative datasets are compared and the features that are more differential are returned as high scores. Furthermore, some features are correlated with each other, meaning that they give the same information logically and/ or numerically. Therefore, in some cases it may not be beneficial to calculate each feature, which causes redundancy. Because of that, the pre-miRNA based features cost and benefit relation needs to be analysed (Yousef and Allmer, 2012b).

1.7. Tools and Databases

For this study The Konstanz Information Miner (KNIME) is used which is an open source tool, featuring a visual platform to integrate, process and analyse huge quantity of different types of data (Berthold et al., 2007). The basic operating unit is a Node and by

connecting Nodes, data workflows can be created and automated. Orange is another open source data mining tool, written in Phyton, which is used for analysis and visualization of data. Rstudio and NetBeans are integrated development environments (IDE) used for statistical analysis in R and implementation of features and feature calculation methods in JAVA programming languages, respectively. According to studies conducted on miRNAs, registries like miRBase v20 database make the published information including sequence and annotations of the miRNAs via web interface searchable available (Kozomara and Griffiths-Jones, 2014). In this study, data is obtained from miRBase and Ensembl, features are programmed and calculated in JAVA, KNIME is used to handle and analyse (Linear Correlation, Information Gain, Random Forest Prediction and model creation) and create a continuous cost and benefit analyses workflow. Orange is used to plot the distance map of KNIME correlation coefficient values. Predicted hairpin and miRBase v20 mature miRNA alignments are obtained via BLASTN short mode (Camacho et al., 2008, 2009) and secondary structures are plotted via VARNA v3-91 (Darty et al., 2009). During feature calculation in JAVA some external tools are used which are RNAhybrid (2.1.2), RNAfold 2.1.3 (Lorenz et al., 2011), UNAfold 3.8 (Markham and Zuker, 2008), dustmasker 1.0.0 (Morgulis et al., 2006), RNAeval 2.1.3 (Lorenz et al., 2011) , RNAspectral (Ng and Mishra, 2007). These third-party tools are automated in the JAVA application in order to parse their outputs to calculate related features.

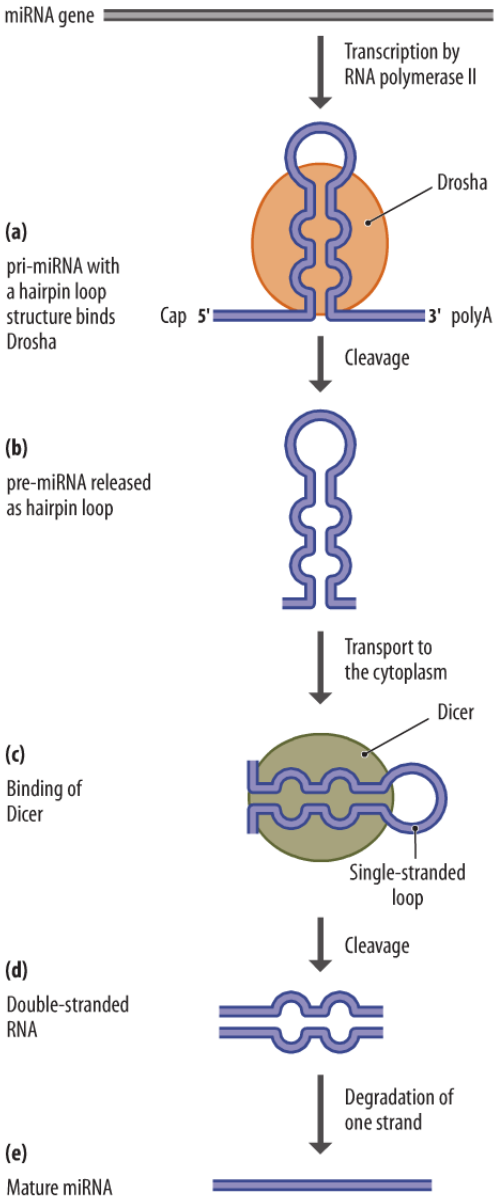

Figure 1.1. miRNA Biogenesis. It is shown how the pre-miRNA is processed. a)After the transcription by RNAII it binds to Drosha upon hairpin loop structure creation. b)Drosha generated pre-miRNA is translocated to cytoplasm. d)After binding to Dicer protein the terminal loop is cut. d)dsRNA forms with one leading strand the mature miRNA and the other one is degraded. e)polyadenylated RNA (Latchman, 2010).

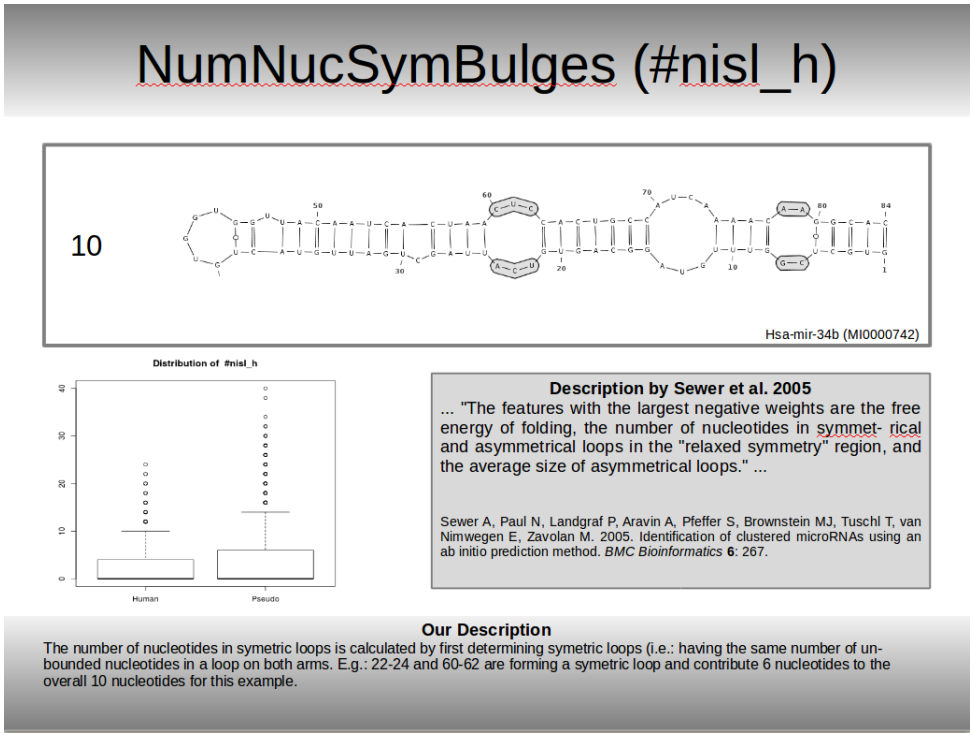

Figure 1.2. Structure of hsa-mir-34b is shown. The feature "number of nucleotides in symmetrical loops" #nisl_h, is shown for the hairpin. The feature has a distribution with a minimum of 0, an average of 5.85 and a maximum of 30 result for the human miRNA data; while the values for pseudo data are larger. The feature is implemented from (Sewer et al., 2005). Fold is created with RNAfold 2.1.3 (Lorenz et al., 2011) and image is drawn using VARNA v3-91 (Darty et al., 2009).

CHAPTER 2

MATERIALS AND METHODS

2.1. Datasets

From miRBase v20 human hairpins which are defined as predicted miR stem-loop sequences and human mature miRNAs are downloaded in FASTA files. miRBase data is used as positive data. Pseudo hairpins (randomized inverted repeats) (negative data) are obtained from Ng and Mishra, 2007. For any case study, in order to predict pre-miRNA hairpin structure, the desired genes' human genomic sequences (3' UTR, 5' UTR, exon and intron; 500 nt down and upstreams of them) is retrived from Ensemble 84 with biomaRt v2.28.0 package in R. For organisms where the data size was small, it was downloaded directly from Ensemble 84 web server (http://www.ensembl.org/index.html, June 2016).

2.2. Feature Extraction and Programming on Java Platform

Features were chosen from the study Yousef et al., 2016. The literature review is done and all features already published are taken according to the first claimer A1. This features have been programmed on Java platform as previously described (Bağcı and Allmer, 2016; Saçar and Allmer, 2013; Yousef et al., 2016). Features have been tested by JUnit unit testing. Each feature class included common tests like negative test, positive test, having flanking ends, having loop structure. The classes are built to an executable jar file which takes a list of hairpin sequences and a list of features to be calculated and outputs the scores for the given features for every hairpin sequence. This jar file is further called in KNIME from "External SSH Tool" node. The calculations are done on Amazon EC2 m4.large instances.

2.3. KNIME Platform

KNIME tool version 3.33 is used with default settings as described previously (Khalifa et al., 2016; Saçar and Allmer, 2013; Yousef and Allmer, 2012b). Additional nodes are downloaded from http://bioinformatics.iyte.edu.tr/KNIMENodes. KNIME is used to handle and analyse the data and for statistics. It is mainly used because of its data mining feature and such analyses as Linear Correlation, Information Gain and Random Forest Prediction are made. KNIME enables to create a continuous cost and benefit analyses workflow.

2.3.1. Feature Calculation

Positive and negative data files are read with File Reader node in KNIME, separately. With External SSH Tool node, each data file and all features are given as input to the feature calculation jar file. As output all the scores for all features and the time of calculation for each feature are stored in separate files for both positive and negative data. The output files are named as humanFeatureCalculation, pseudoFeatureCalculation and humanTime; and exported in tab delimited format.

2.3.2. Worflow of Cost and Benefit Analysis

This workflow is created in order to analyse the time (cost) and information gain joined together and correlation analyses as a distance plot. Therefore, time and information score ranks are obtained as described above Section 2.3.1 and joined together in one table (Figure 2.1).

2.3.2.1. Cost Analysis

humanTime file is read and statistical analysis such as minimum, maximum, mean, standard deviation, variance, median, overall sum, row count across all numeric columns, and counts all nominal values together with their occurrences are retrieved and these val-

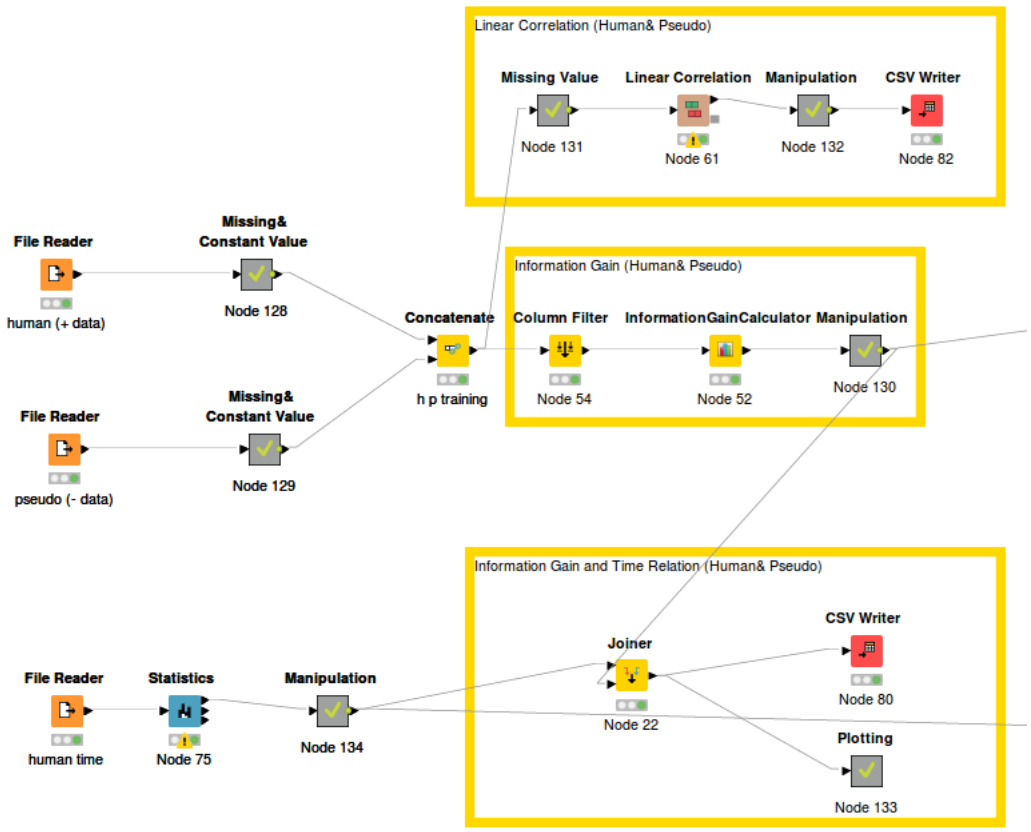

Figure 2.1. Worflow of Cost and Benefit Analysis.

ues are used to create plots in further steps. The features are ranked according to time it took to calculate them.

2.3.2.2. Workflow for Time, Information Gain and Correlation Analysis

humanFeatureCalculation and pseudoFeatureCalculation files are read separately. Constant Value Column indicating the class of the tables are added to human and pseudo data. The tables are then concatenated. The parts explained below are applied distinctly from each other.

Time, Information Gain Analysis: After filtering the string values of the concatenated table an Information Gain Calculator node is added. The output is ranked and joined to the time rank table. From the joined table a scatterplot is plotted to visualize the data distribution.

Correlation Analysis: To the concatenated node a Linear Correlation node and a Correlation Filter node is added. For the study Pearson's product-moment coefficient is applied and -1 shows strong negative correlation and 1 shows strong positive correlation. The Distance maps for all of the data and separately for selected data are plotted in Orange (range is unique for each plot).

2.3.3. Model- Random Forest Prediction

humanFeatureCalculation and pseudoFeatureCalculation files are read separately (Figure 2.2). Constant Value Column indicating the class of the tables are added to human and pseudo data. For each partitioning random sampling of all rows is applied. As observed in Figure 2.3, a Start Loop is added and the further steps (ending at the Stop Loop Node) are repeated for 10 times. Pseudo data is partitioned to an absolute of 4000 rows to one table and 4492 rows to another table (pseudo first partitioning). Then the first table is partitioned to an absolute of 1828 rows to one table and 2172 rows to another table (pseudo second partitioning). Lastly the first table is partitioned 70% relatively, where this is used for learning and the reaming 30% is used in testing (pseudo third partitioning). Human table is partitioned 70% relatively, where this is used for learning and the reaming 30% is used for testing. The human partitioning and pseudo third partitioning outputs are processed as follow: The learning and testing tables are concatenated respectively then the learning table is given to the Random Forest Learner nodes input and the testing table is given as input to the first Random Forest Predictor node. The second table of second pseudo partitioning is given to the second Random Forest Predictor node. The second table of first pseudo partitioning is given to the third Random Forest Predictor node. The output model of Random Forest Learner is given as model input to each of Random Forest Predictor nodes separately. The Random Forest Predictors are linked to Score nodes and Constant Value Columns indicating the order of the predictors are added as First Prediction Score, Second Prediction Score and Third Prediction Score, respectively. A Loop

End node is added. The output of the loop is sorted, filtered and a Reference Row Filter is used where the iteration number of the best model with high accuracy is taken for the case study prediction model, therefore a Cell to Model node is connected for the conversion (Figure 2.3). This node is reference model for the case study Random Forest Predictor.

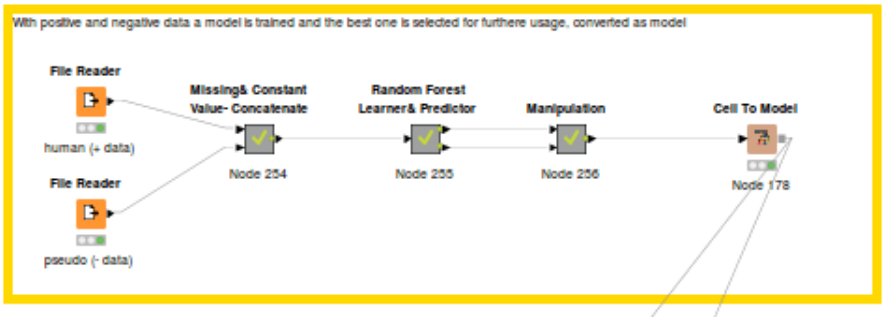

Figure 2.2. Knime Workflow for Model Creation.

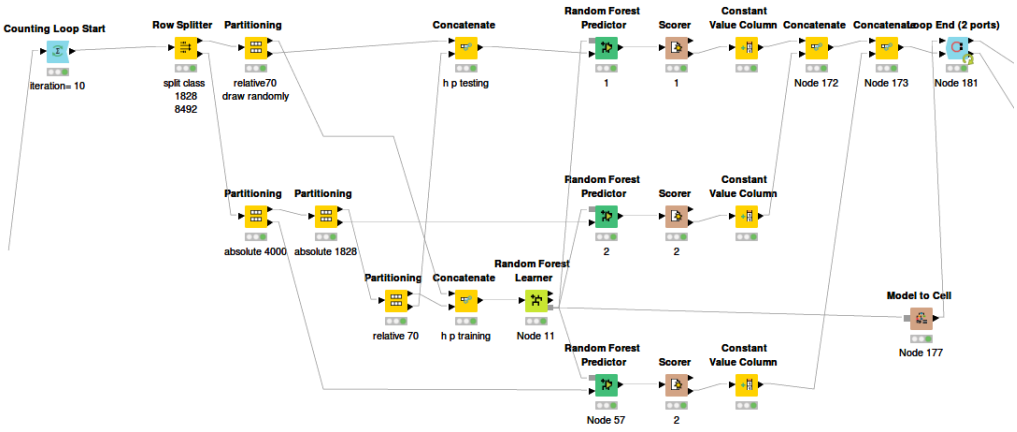

Figure 2.3. Knime Workflow Random Forest Prediction for Model creation. Meta Node from Figure 2.2 is opened.

2.3.4. KNIME Workflows for Case Studies

This part is saved as a separate workflow from the other ones, since it is used in case of a case study desired. The case study can be a human gene collection or any organism data, depending on the aim. The fasta files were downloaded as described in Section 2.1 is used further for hairpin prediction and time information gain analyses.

2.3.4.1. Sequence Fragmentation

Below steps are done for both virus and gene datasets separately (Figures 2.4 and 2.5). Each separate genomic part files are read with the Fasta2Table nodes. This node reads fasta files by conserving the fasta properties and parsing to a table. The read files have the definition line and the sequence related in the table. The tables were then concatenated. The virus genome was stored in a single file therefore it was read and handled directly. The nucleotide T is converted to U. Files are processed further and given to Sequence Fragmenter, in which sequences were fragmented into 250bp long fragments with 250bp overlaps (Figure 2.6). With KNIME External SSH Tool the fragments are folded using RNAfold 2.1.3 (Lorenz et al., 2011).

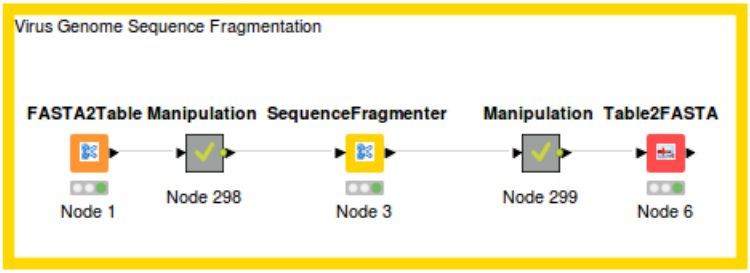

Figure 2.4. Knime Workflow for Virus Sequence Fragmentation.

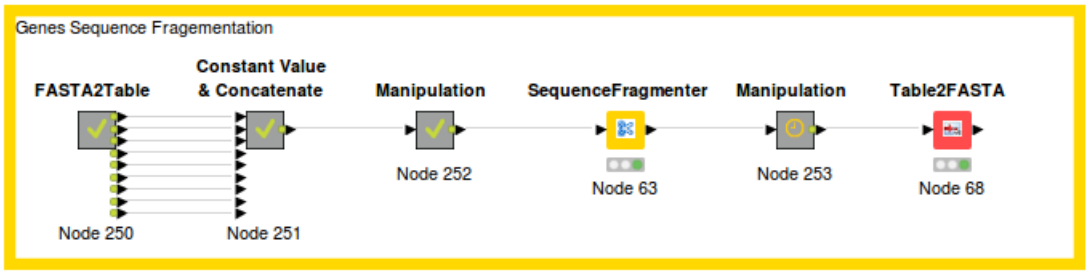

Figure 2.5. Knime Workflow for Genes Sequence Fragmentation.

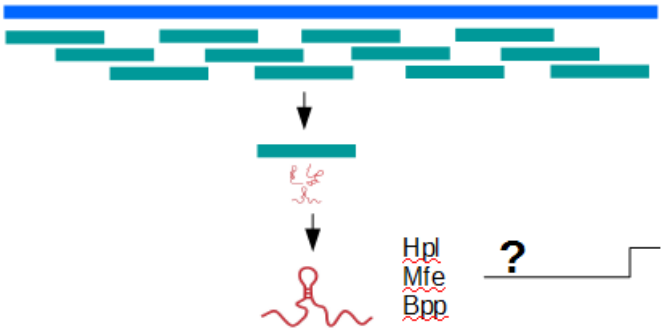

Figure 2.6. After the sequence is retrieved it is fragmented in to overlapping fragments. For each fragment the fold is revealed and then hairpins are extracted. The computationally predicted hairpins are further used for pre-miRNA based feature calculation.

2.3.4.2. Hairpin Extraction

The workflow continues with hairpin extraction. The files are processed further (columns, names etc) to be an input for the HairpinExtractor node (Figures 2.7 and 2.8). This is done for both virus and gene dataset. This node takes the sequence fragments and creates possible hairpin structures. There can be more than one hairpin formed from a hairpin where no hairpin formation can be observed too (Figure 2.6).

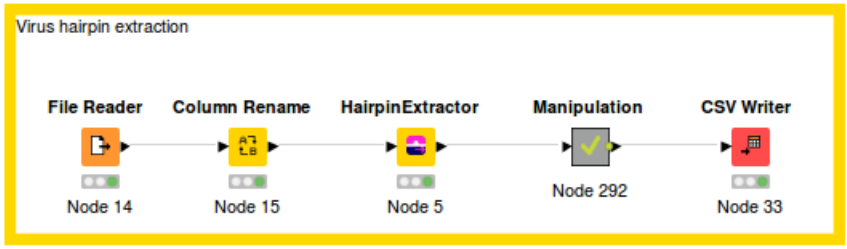

Figure 2.7. Knime Workflow for Virus Hairpin Extraction.

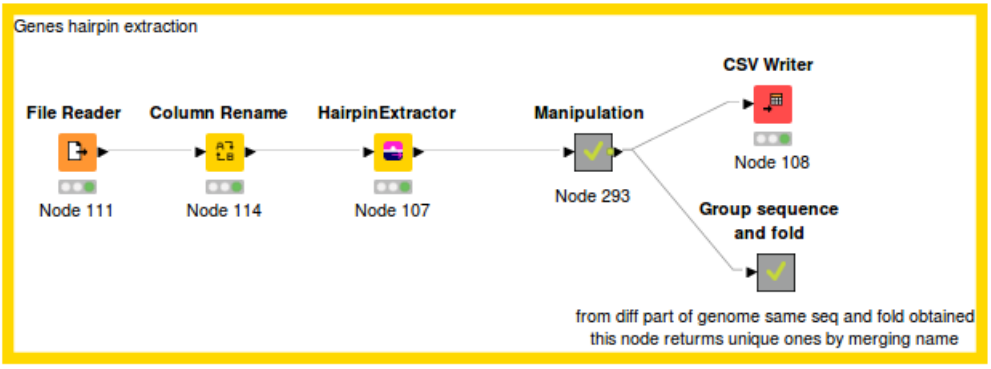

Figure 2.8. Knime Workflow for Genes Hairpin Extraction.

2.3.4.3. Feature Calculation

As in Section 2.3.1. Virus and gene data files are read with File Reader node in KNIME, separately. With External SSH Tool node, each data file and all features are given as input to the feature calculation jar file. As output all the scores for all features and the time of calculation for each feature (just for virus) are stored in separate files for both virus and gene data as tab delimited format.

2.3.4.4. Hairpin Prediction

Input file is read and after Missing Value Handling and Constant Value Column adding it is connected to a Random Forest Predictor which takes as model input, the one having the highest accuracy score the one from Section 2.3.3. The output of the Random Forest Predictor is further linked to Histogram and Boxplot nodes, where in Boxplot node positive and negative predictions are are handled separately and then plotted distinctly (Figures 2.9 and 2.10).

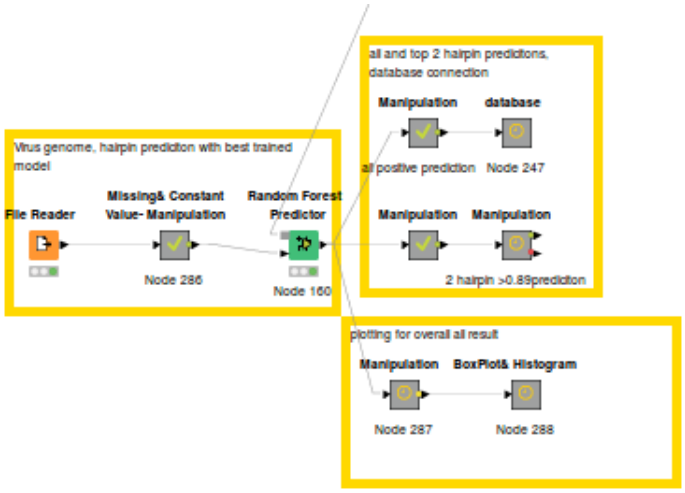

Figure 2.9. Knime Workflow for Virus Hairpin Prediction

2.3.4.5. BLASTN and Reactome Analysis

Known human mature miRNAs from miRBase v.20 are aligned against hairpins predicted from human and virus genes with prediction scores equal to or greater than 0.90 and 0.99, respectively, using BLASTN in blastn-short mode (Camacho et al., 2008, 2009). Alignments with perfect matches (no mismatches or gaps) are filtered and plotted by VARNA v3-91 (Darty et al., 2009). The mature miRNAs found with perfect alignments to predicted pre-miRNA sequences are then searched for their targets in mirTARBase and TarBase. Furthermore, the target genes were uploaded to Reactome (pathway database

v57) web-server (http://www.reactome.org/) to analyse the pathways that can be regulated by these miRNAs from the information on already known targets.

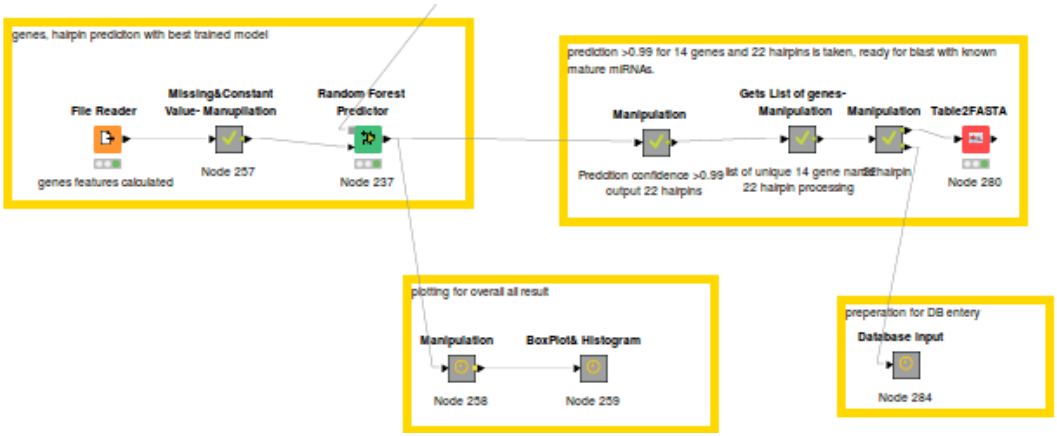

Figure 2.10. Knime Workflow for Genes Hairpin Prediction.

2.3.4.6. Workflow for Time, Information Gain Analysis

A Constant Value Column is added as virus to VirusTime table which is then joined to two nodes, one to human and pseudo information gain result obtained in Section 2.3.2.2. Then appropriate nodes for plotting (box plot and histogram) and output are connected. Another node is connected to the manipulated VirusTime table to join to the humanTime. Then appropriate nodes for plotting (box plot and histogram) and output are connected (Figure 2.11).

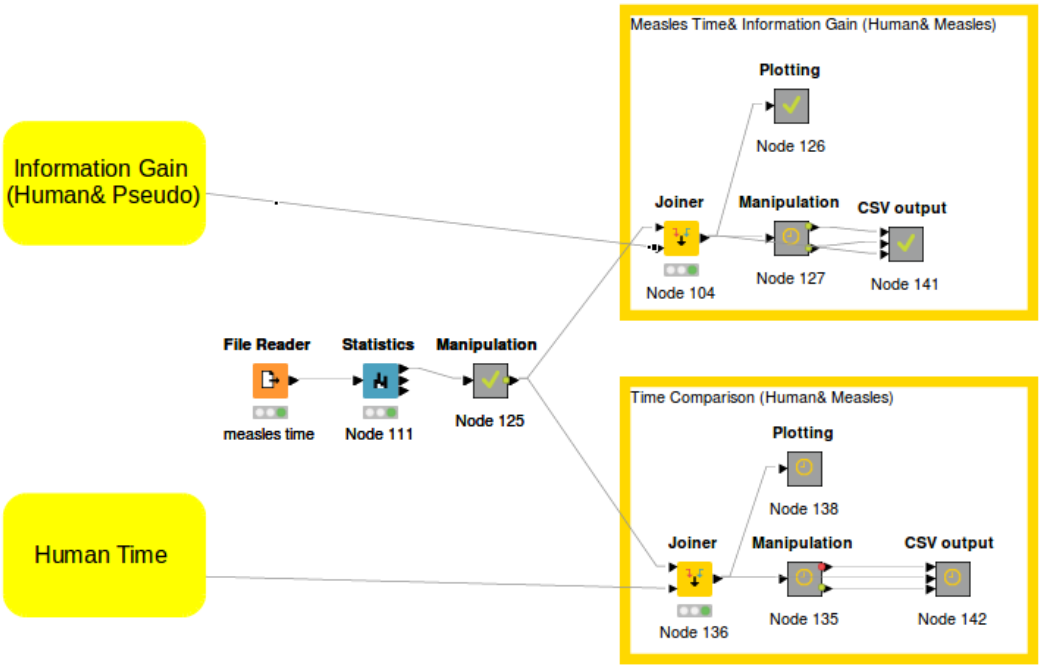

Figure 2.11. Knime Workflow of Virus Time, Information Gain Analysis.

CHAPTER 3

RESULTS

3.1. pre-miRNA Feature Extraction

The feature Table A1 represents the Java code acronym of each feature, the synonyms used in literature and the feature type.

All features are not included and some feature groups are given compact and the number of features in that group is present in the ”Count” row. For example, the ”Nucleotide percent” feature consists of 4 features %A, %C, %G and %U but all are shown in one row as from %A to %U indicating continuity with ”..”. A reference for the feature implementation in Java is shown as Figure 1.2. The description from the literature is taken and our understanding of the feature is indicated and further used as understanding of the feature to implement it in Java. The boxplot shows the distribution of calculation result in human hairpins and pseduo data. The more informative features have different distribution range, lower and upper quartile when human and pseudo data are compared.

3.2. Time, Information Gain Analysis

Calculation time of each feature for human hairpins is noted and further compared with the information gain score (Figure 3.1).

Same procedure is applied for the case study virus data (Figure 3.2).

Information Gain Calculator node requires two different class types, for this case positive (human hairpins) and negative (pseudo data) classes are used. For each feature comparing the value in between the classes an information gain score for each feature is returned. The more distinguishable the feature between the positive and negative classes, the higher the information score is. The time information gain comparison showed that some feature are calculated fast; however, they have low informative gain (Figures 3.1 and 3.2, Tables 3.1 and 3.2).

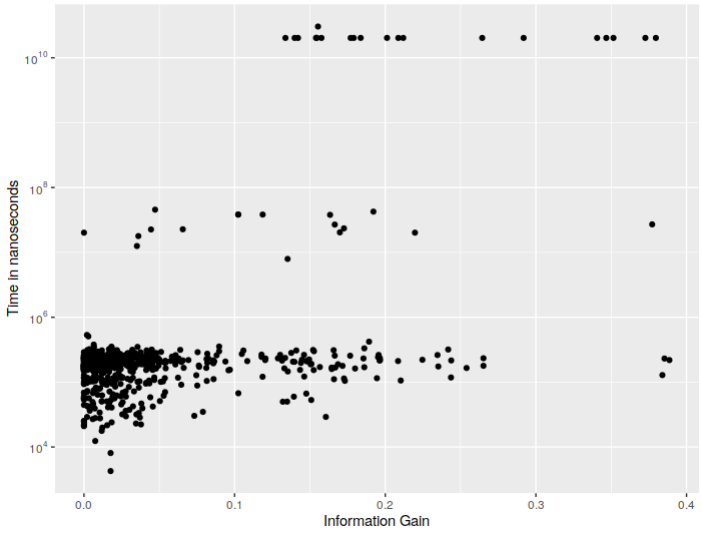

Figure 3.1. Scatter plot of features' information gain (x) and human hairpin mean time (y). Time axis is in logarithmic scale. Results are obtained in Knime. Each dot represents a feature.

Moreover, the most informative features are the slowest (Figures 3.1 and 3.2, Tables 3.3 and 3.4)

Top 10 tables are constructed with nearly all same features for human hairpin and virus data (Tables 3.1, 3.2, 3.3 and 3.4). From the scatter plot Figure 3.2 the features are highlighted with high information gain score x>0.2 and low mean time y<10^6 in Table 3.5.

From the scatter plot Figure 3.2, the features which are not normalized, have high information gain scores x>0.2 and low mean times y<10^6 are highlighted in Table 3.6. Each feature from Table 3.6 are defined in Figures A.1, A.2, A.6, A.3 and A.4.

Features having high information gain score and relatively low calculation time is interesting to be investigated further (Figures 3.1 and 3.2). Furthermore, human hairpins data time and virus time show correlation as expected (Figures 3.1, 3.2 and 3.3).

For both human and virus data, features are calculated on Amazon EC2 instance. For human hairpins and virus the calculations costed nearly 2 USD and (Table 3.7).

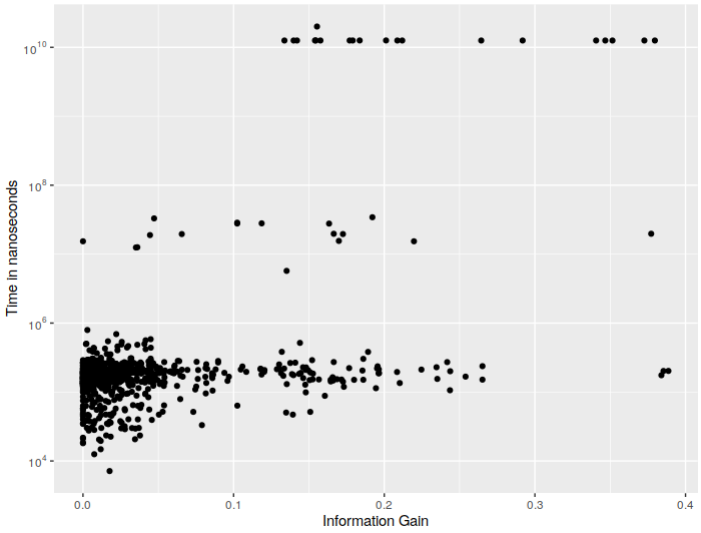

Figure 3.2. Scatter plot of features' information gain (x) and virus mean time (y). Time axis is in logarithmic scale. Results are obtained in KNIME. Each dot represents a feature.

3.3. Correlation Analysis

Human hairpins are used for the correlation study and for pseudo data correlation analysis are done too. Linear Correlation node takes each time a feature and returns the correlation score with each feature. This is done for all of the features. The data is plotted in Orange as distance map is shown in Figure 3.4.

If the color is red, it means that the correlation is high having no distance between the features. Therefore, in the middle is a straight line showing the correlation of the feature with itself. Features are clustered, therefore, some regions form same colour distribution because of being correlated. Top 10 of fastest and more informative features are selected for an additional distance plot (Figure 3.5).

dns features for both P value and Z score are highly correlated with each other. Another high correlation is observed between assl, assl/hpl and assl/sl features. This is normal as they are all related with assl feature and are normalization of assl feature to hpl and stem length (sl). To compare human hairpin and pseudo data correlation results, no sorting or clustering have been applied to the features meaning the distance map are plotted with listed feature. Pseudo data correlation result shows a slightly similar colour

Table 3.1. Time and information gain table of features. Results are obtained in Knime. 10 features are selected according to their time rank. Human hairpins are used. Each row represents a feature indicated and its time rank, mean time in milliseconds, information gain rank and information gain score. Information about the feature are found in Table A1.

Feature	**Time Rank**	**Mean Time (ns)**	**Information Gain Rank**	**Information Gain**
#A++#A	1	0.004	449	0.017
#A	2	0.008	450	0.017
#C++#A	3	0.012	600	0.007
#C	4	0.017	535	0.011
#AU	5	0.019	524	0.012
#AC	6	0.020	777	0
#AA	7	0.021	481	0.015
#AG	8	0.021	781	0
#A.(.	9	0.021	772	0
#A..(	10	0.022	244	0.037

distribution as human hairpin result (Figures 3.6, 3.7).

For pseudo data high correlation is observed between assl, assl/hpl and assl/sl features (Figure 3.7). However different from human hairpin data, pseudo sample shows not direct as high correlation as human hairpins' in between dns features (Figure 3.7). assl based features and dns based features show more correlation for pseudo data than human hairpins data. In human hairpins the dns features represent for the specific feature calculated, a correlation in between P value and Z score, however this relation is ruined for some parts of pseudo data, by having for Z score based more correlation with other features than the P value based dns feature (Figures 3.6 and 3.7).

Moreover, the correlation coefficients have been calculated for features from Table 3.6 and the distance map revealed that they are not correlated with each other, are distanced (Figure 3.8).

3.4. miRNA prediction

The pre-miRNA prediction results are given as boxplots. Accuracy value is given in y axis indicating how accurate the prediction is made for both the negative and positive results (Figures 3.9 and 3.10).

The negative predictions have larger interquartile range than the positive ones, where as the medians presented as a line in the box and lower quartiles are higher for

Table 3.2. Time and information gain of features. Results are obtained in KNIME. 10 features are selected according to their time rank. Virus data is used. Each row represents a feature indicated and its time rank, mean time in milliseconds, information gain rank and information gain score. Information about the features can be found in Table A1.

Feature	Time Rank	Mean Time (ms)	Information Gain Rank	Information Gain
#A++#A	1	0.007	449	0.018
#C++#A	2	0.013	600	0.007
#C	3	0.015	535	0.012
#AC	4	0.018	777	0.000
#AG	5	0.018	781	0.000
#C++#C	6	0.020	534	0.012
#G++#A	7	0.020	557	0.011
#A(..	8	0.021	261	0.035
#A.((	9	0.021	769	0.000
#A.(.	10	0.022	772	0.000

positives (Figures vbox, gbox. Virus genome give rise to 50 negatively and 131 positively selected computationally predicted pre-miRNA hairpins (Figures 3.9). However only two hairpin fragments are selected, the top two with high accuracy; 0.90 and 0.94.

Genes give rise to 126613 negatively and 102110 positively selected computationally predicted pre miRNA hairpins (Figures 3.10). As sequences had been colleted from different genomic parts of the genes, there are overlaps. There had been hairpins created with same sequence which is used to compare the accuracy values and thus kind of test the created model. Out of all 46632 sequence is unique. Top predictions with 0.99 accuracy are selected which are observed as outliers above whisker.

3.5. BLASTN and Reactome

The two hairpin fragments predicted from virus do not show total alignment to a known human mature miRNA. The fragment with accuracy score 0.94 aligned with ten match and seven mismatch to hsa-miR-4306 (Figure 3.11).

Two hairpin fragments predicted from genes BLASTN result against known human mature miRNAs are given as Figure 3.12 for TAB2 gene based fragment and Figures 3.13 and 3.14 for BBC3 based fragment.

This ones are selected as they show total alignment. Computationally predicted pre-miRNAs show known mature miRNA sequences buried at locations where from pre-

Table 3.3. Time and information gain table of features. Results are obtained in KNIME. 10 features are selected according to their information gain rank. Human hairpins are used. Each row represents a feature indicated and its information gain rank, information gain score time rank and mean time in milliseconds. Information about the features can be found in Table A1.

Feature	Information Gain Rank	Information Gain	Time Rank	Mean Time (ns)
assl/hpl	1	0.388	528	0.218
assl/sl	2	0.385	592	0.231
assl	3	0.384	168	0.128534
dns_p(efe)	4	0.379	855	202511
hpmfe_rf_I1	5	0.377	846	270434
dns_p(hpmfe_rf)	6	0.372	853	202509
dns_z(efe)	7	0.351	857	202511
dns_z(hpmfe_rf)	8	0.346	859	202511
dns_z(hpmfe_rf/hpl)	9	0.340	861	202514
dns_p(hpmfe_rf/hpl)	10	0.291	858	202511

Table 3.4. Time and information gain table of features. Results are obtained in KNIME. 10 features are selected according to their information gain rank. Virus data is used. Each row represents a feature indicated and its information gain rank, information gain score time rank and mean time in milliseconds. Information about the features can be found in Table A1.

Feature	Information Gain Rank	Information Gain	Time Rank	Mean Time (ms)
assl/hpl	1	0.389	558	0.202
assl/sl	2	0.385	557	0.202
assl	3	0.384	403	0.175
dns_p(efe)	4	0.380	855	12563.119
hpmfe_rf_I1	5	0.377	846	19.766
dns_p(hpmfe_rf)	6	0.373	853	12563.043
dns_z(efe)	7	0.351	857	12563.214
dns_z(hpmfe_rf)	8	0.347	858	12563.343
dns_z(hpmfe_rf/hpl)	9	0.341	865	12566.250
dns_p(hpmfe_rf/hpl)	10	0.292	862	12565.696

miRNA give rise to mature miRNAs. Figure 3.12 shows highlighted sequences for hsa-miR-548ar-5p, hsa-miR-548au-5p, hsa-miR-548ay-5p and Figure 3.13 hsa-miR-3191-3p and Figure 3.14 for hsa-miR-3191-5p. Fragment from BCC3 has sequences buried of the the miRNA both 3p and 5p matures. Just hsa-miR-548au-5p has known targets and the further analysis of those target genes in Reactome showed mostly association with Gene Expression and Immune System pathways.

Table 3.5. Information gain and human hairpins-virus time table of features. Results are obtained in KNIME. Features are taken from Figure 3.2 region x>0.2 y<10^6. Each row represents a feature indicated and its information gain rank, human hairpins and virus mean time in milliseconds. Information about the feature can be found in Table A1.

Feature	Information Gain	Human Mean Time (ms)	Measles Mean Time (ms)
assl/hpl	0.389	0.219	0.202
assl/sl	0.385	0.231	0.202
assl	0.384	0.129	0.175
subu/sl	0.265	0.179	0.151
bpp/sl	0.265	0.234	0.237
subu/hpl	0.254	0.165	0.167
hpmfe_rf_I1/hpl	0.244	0.216	0.200
lsr(%bp)	0.244	0.118	0.106
hpmfe_rf_I1/sl	0.242	0.318	0.271
lsr(%bp)/hpl	0.235	0.174	0.155
lsr(%bp)/sl	0.235	0.262	0.229
bpp/hpl	0.225	0.221	0.211
lscm	0.210	0.106	0.135
dG/sl	0.208	0.211	0.195

Table 3.6. Information gain and human hairpins-virus time table of features. Results are obtained in Knime. Features that are not normalized are taken from Figure 3.2 region x>0.2 y<10^6 Each row represents a feature indicated and its information gain rank, human hairpins and virus mean time in milliseconds. Information about the features can be found in Table A1 and Figures A.1, A.2, A.6, A.3 and A.4.

Feature	Information Gain	Human Time (ms)	Measles Time (ms)
assl	0.384	0.129	0.175
lsr(%bp)	0.244	0.118	0.106
lscm	0.210	0.106	0.135
asal	0.194	0.115	0.114
hpmfe_rf_I3	0.189	0.421	0.382

Table 3.7. Amazon EC2 m4.large instance for 12 cent per hour. Human miRBase hairpins and virus data feature calculation and feature time calculation on Amazon are given as minute (m) and second (s). The overall calculations cost 1USD 92 cent.

	Human Hairpins	Virus
Real Time	492m18s	40m35s
User Time	768m4s	43m19s
System Time	135m32s	14m21s

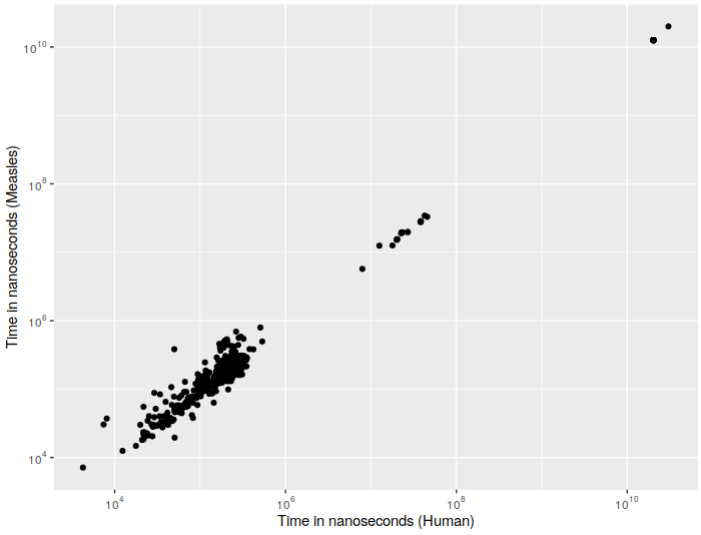

Figure 3.3. Scatter plot of features' log10 normalized mean time scores for human (x) and virus (y). Results are obtained in Knime. Each dot represents a feature.

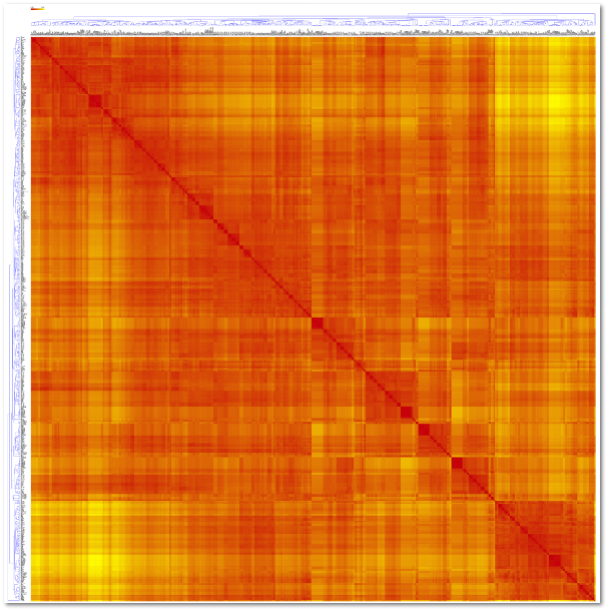

Figure 3.4. Human Correlation study of all features are plotted as distance map. Orange is used to handle the data. The features are represented with ordering leaves clustering. The color red represents the distance 0 and yellow 14.60.

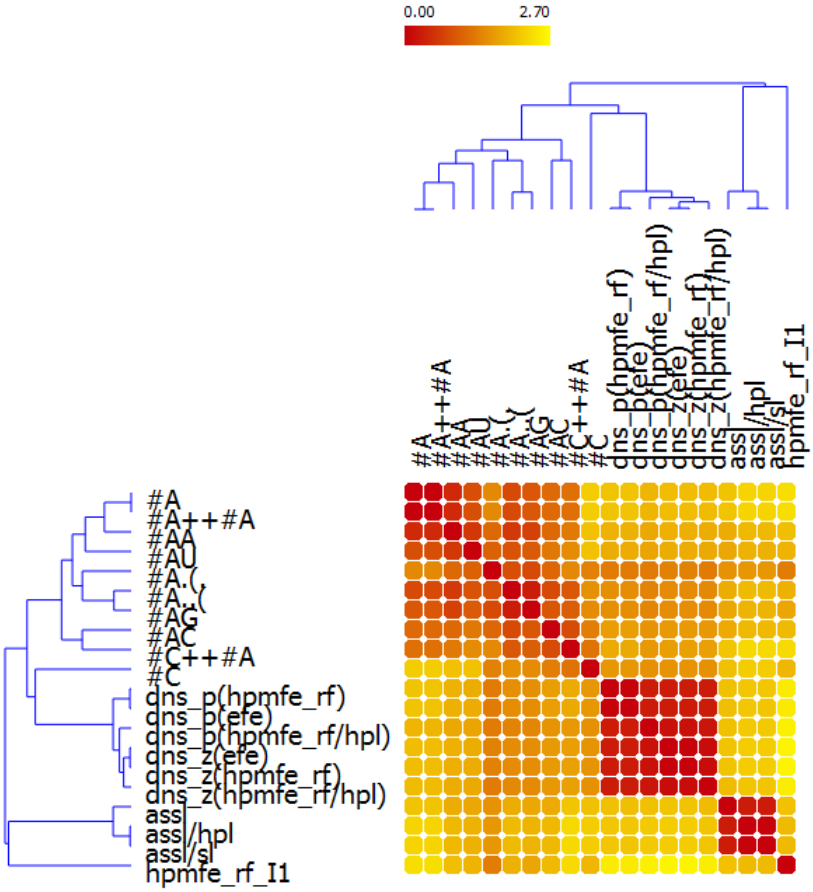

Figure 3.5. Human Correlation study of selected features are plotted as distance map. Features are selected from top 10 time rank and top 10 information gain rank. Orange is used to handle the data. The features are represented with ordering leaves clustering. The color red represents the distance 0 and yellow 2.80. Information about the features can be found in Table A1

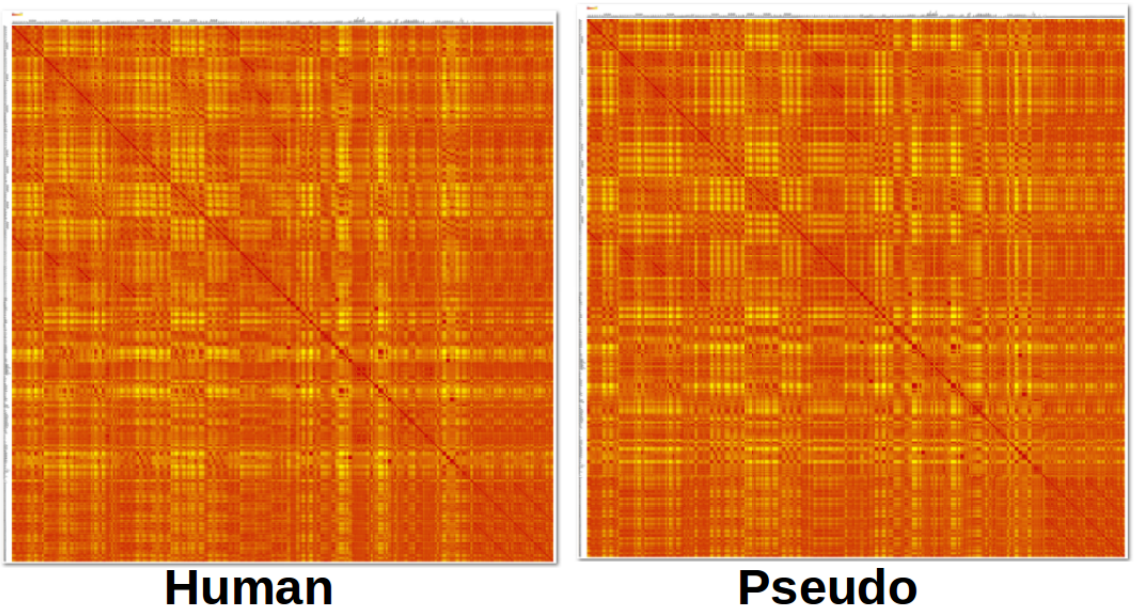

Figure 3.6. Human and pseudo Correlation study of all features are plotted as distance map. Orange is used to handle the data. The features are represented in alphabetical ordering. The color red represents the distance 0 and yellow 14.60.

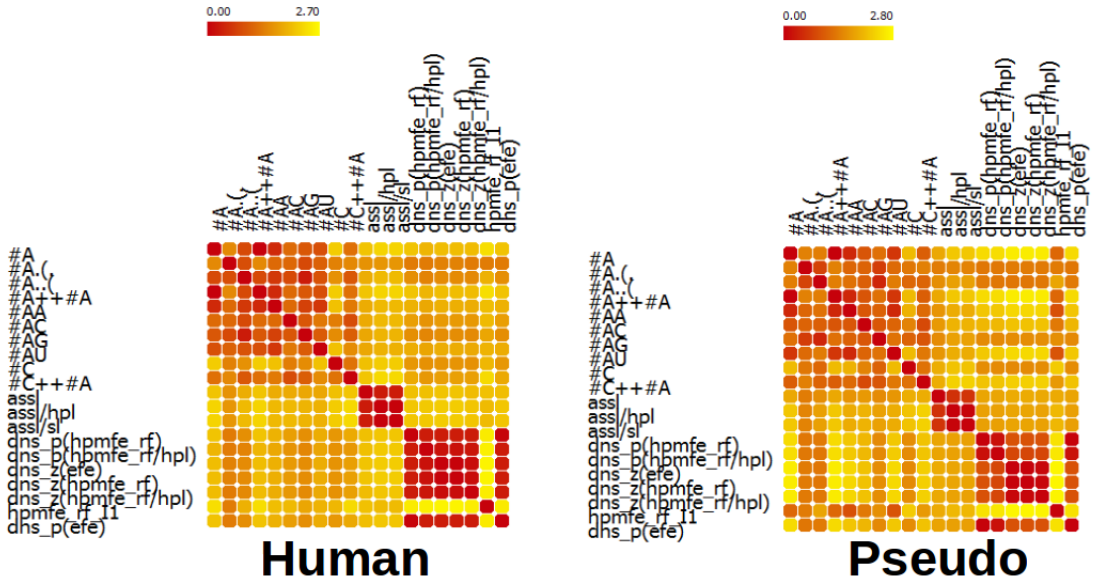

Figure 3.7. Human and pseudo correlation study of selected features are plotted as distance map. Features are selected from top 10 time rank and top 10 information gain rank. Orange is used to handle the data. The features are represented as alphabetical ordering. The color red represents the distance 0 and yellow 2.80. Information about the feature are found in Table A1.

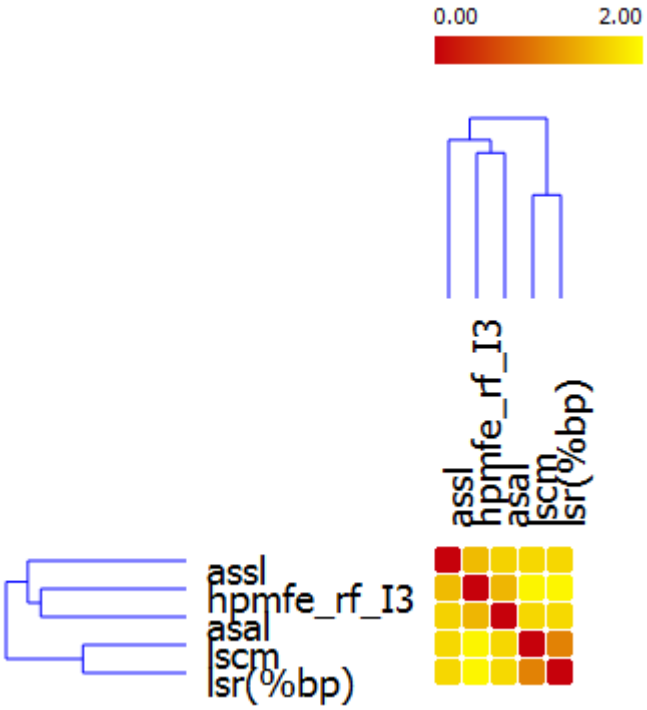

Figure 3.8. Distance map of selected features that are not normalized are taken from Figure 3.2 region x>0.2 y<10^6. Feature from Table 3.6 are plotted as distance map. Orange is used to handle the data. The features are represented as with ordering leaves clustering. The color red represents the distance 0 and yellow 2.00. Information about the features can be found in Table A1 and Figures A.1, A.2, A.6, A.3 andA.4.

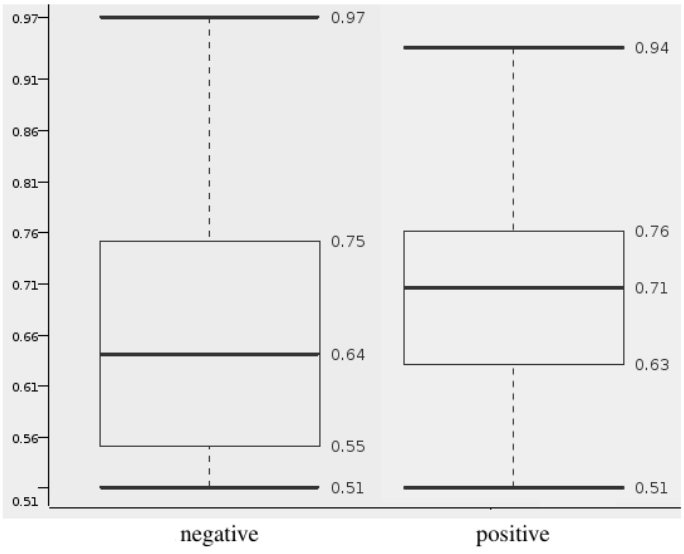

Figure 3.9. Measles genome pre-miRNA prediction based on human model. The boxplots show the distribution of accuracy values of both negative (50 hairpin) and positive (131 hairpin) results.

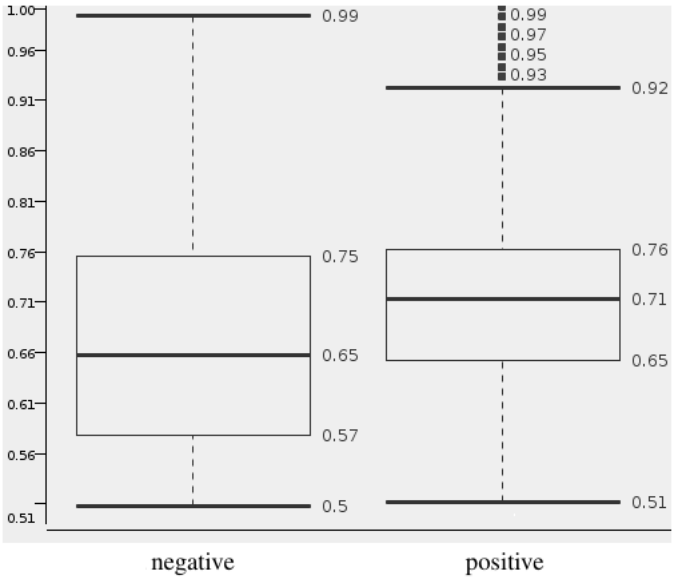

Figure 3.10. Measles genome related human genes pre-miRNA prediction based on human model. The boxplots show the distribution of accuracy values of both positive (102110) and negative (126613) results. The 46632 of the sequences are unique.

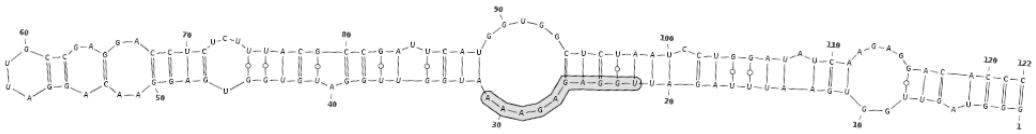

Figure 3.11. Measles virus genome hairpin fragment predicted based on human model with prediction class accuracy of 0.99 is shown and the BLASTN partially alignment part with mature hsa-miR-4306 is highlighted. Plotted via VARNA v3-91 (Darty et al., 2009).

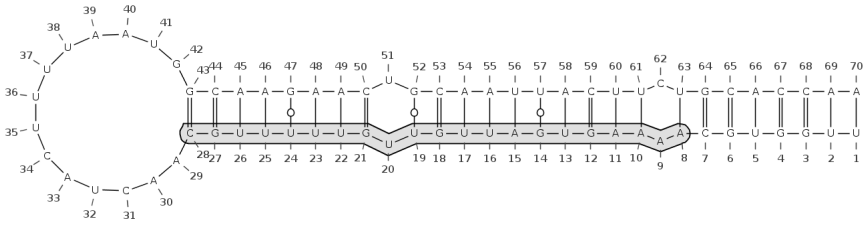

Figure 3.12. TAB2 gene hairpin fragment predicted based on human model with prediction class accuracy of 0.99 is shown and the BLASTN alignment with mature has-miR-548au-5p is highlighted. Plotted via VARNA v3-91 (Darty et al., 2009).

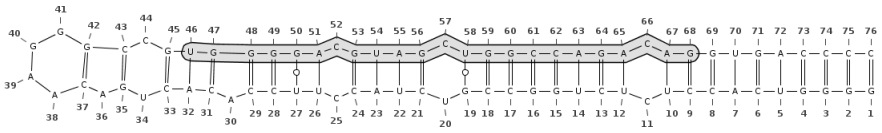

Figure 3.13. BBC3 gene hairpin fragment predicted based on human model with prediction class accuracy of 0.99 is shown and the BLASTN alignment with mature hsa-miR-3191-3p is highlighted. Plotted via VARNA v3-91 (Darty et al., 2009).

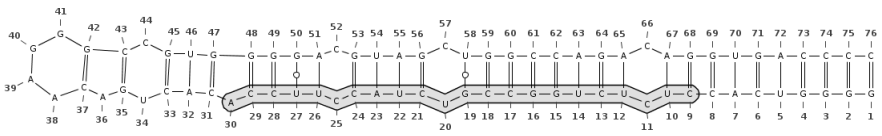

Figure 3.14. BBC3 gene hairpin fragment predicted based on human model with prediction class accuracy of 0.99 is shown and the BLASTN alignment with mature hsa-miR-3191-5p is highlighted. Plotted via VARNA v3-91 (Darty et al., 2009).

CHAPTER 4

DISCUSSION

There are over 800 features used in this study (Table A). The high abundance of the features affects the calculation time greatly. Studies are pointing out that the computational power is determined by feature, classifier, algorithm, machine learning approaches in computational miRNA detection (de ON Lopes et al., 2014; Gudys et al., 2013; Tempel and Tahi, 2012; Tempel et al., 2015). Most of the studies are comparing their own miRNA prediction algorithm with already existing ones, which may give us a slight idea about the effects of features used in those studies on calculation time as there are many factors affecting the computational power. However, there are studies in which the calculation time of the features are compared by assigning them to different overlapping groups (de ON Lopes et al., 2014). This study is unique as it has a huge amount of features, the calculation time of each of the feature is analysed separately which is just dependent on the feature calculation method, is not effected by machine learning methods.

According to Tables 3.1, 3.4 and A the fastest calculated features are the sequential type nucleotide count features (Lai et al., 2003) and dinucleotide features (Ng and Mishra, 2007). They are followed by sequential - structural type triplet count features (Xue et al., 2005). These features are faster as their algorithmic complexity is lower.

The first ranked feature in Tables 3.1 and 3.4, #A++#A is a sequential type feature which represents the count of A bases in the hairpin summed with itself (Zhang et al., 2006). The second fastest feature is #A nucleotide count - number of A bases in the hairpin. The feature #A++#A requires firstly the calculation of #A. When #A++#A feature needs to be calculated the code checks the value of #A and without calculating the #A again and retrieves the value already calculated and returns #A++#A. Therefore, #A++#A is calculated faster than #A. This same effect can be seen for #C++#A and #C (Table 3.1). There are features normalized to hpl, stem length (sl); features calculated by basic mathematical operations (addition, substitution, division and proportion). In order to reduce the overall calculation time of all of these operations, features that are already calculated are parsed and the value of them are used in the calculation of these features based on basic mathematical operations. However; further in some studies, the feature

used to calculate other features by such operations may not be used, such as after feature selection has been applied. Thus such features may take longer computational times as they will also require computation of their dependent features, even though the dependent feature is not used.

Another factor that affects the calculation time of a feature is the external tool usage, such as RNAfold 2.1.3 (Lorenz et al., 2011) which is used in the creation of secondary structures for each feature and used for thermodynamic features calculation. There are other secondary structure predictors; however, in this study RNAfold is used as being on of the widely accepted ones and used by miRBase v.20 (Kozomara and Griffiths-Jones, 2014; Tempel and Tahi, 2012). The calculation rank of thermodynamic features are lower than others, however they are among the most discriminative ones (Tables A and 3.3). According to the study of Ding et al., 2010, thermodynamic features are more informative than sequential and structural ones.

In order to identify the discriminative power, information gain is a widely applied method (Chen et al., 2016; Khalifa et al., 2016; Uğuz, 2011; Yousef et al., 2016). According to the Table 3.3 a structural feature assl (average size of symmetric bulges) (Sewer et al., 2005) has the best scores followed with probabilistic features; dinucleotide shuffling method (dns) (Jiang et al., 2007) both with P value (Jiang et al., 2007) and Z score (Ng and Mishra, 2007) applied to thermodynamic features are the most informative ones. As thermodynamic ones are already informative, applying dns makes them most discriminative features. The dns features takes the concerned hairpin sequence and shuffles it 1000 times randomly, then the specified feature is calculated for the shuffled ones and the distance between the concerned sequence and the shuffled is calculated each time, scored with P value and Z Score, separately (Figure A.5) (Jiang et al., 2007). Our concern was before the study; dns method would costs 1000 times more than the other features, but is it worth? de ON Lopes et al., 2014, eliminate the usage of dns features because of their high computational costs despite finding its predictive power to be high. Moreover, Jiang et al., 2007 and Chen et al., 2016, found that minimum free energy (structural-thermodynamic, hpmfe_rf (Jiang et al., 2007) and hpmfe_rs (Çakir and Allmer, 2010) and P value features are the most discriminative ones and Xuan et al., 2011, which did not use dns based features found that thermodynamic features are the most informative ones. Another study de ON Lopes et al., 2014, consists selection of features mostly of energy based ones.

The Table A reveals that structural features are more discriminative than sequential ones. In literature the same observations are seen. van der Burgt et al., 2009, selected

eighteen features consisting of nine structural features which are more informative than remaining sequential ones; and Chen et al., 2016, used triplet count features (sequential-structural (Xue et al., 2005)) and trinucleotide features (sequential (Chen et al., 2016)) and concluded that structural-sequential features are more informative than sequential ones. However, the findings of Wei et al., 2014, are opposite than stated, they found that sequential features are more informative than structural features, minimum free energy feature and P value features. Most recent publication analyses the result of Wei et al., 2014, and comes to conclusion that their findings are different because of their negative data and states that the feature is discriminative if it shows variability on positive and negative dataset (Chen et al., 2016). One study used four triplet structure features A(((, U(((, G(((and C(((as they are claimed to be more informative than the other remaining 28 (Gudys et al., 2013). In this study A(((and U(((structures triplet count, triplet frequency (Xue et al., 2005) and their normalizations were more informative than the other ones.

If features are correlated and somehow give the same information it is not much helpful to use them in miRNA prediction (Yousef and Allmer, 2012b). Therefore, the correlation analysis is made to see how beneficial the features are. For example in the stem the bounds can constitute AU bonds (st(A-U), not a soure) and the number of U base and A base should be correlated. However, as there in the stem can be GU bonds (st(G-U), not a soure) too, it is not logical to conclude that it would be sufficient to count either A or U nucleotide in the bounds to come up with AU bounds. However, AU content is an important discriminator of miRNA from other RNAs and it is thought that AU bonds give the pre-miRNA fragile structure so that they can be processed easily to mature miRNAs via RISC complex (Zhang et al., 2006). Further, it is known that in miRNAs A base is less found than the other ones and U nucleotide has the most common occurrence that may have a signal role in miRNA biogenesis (Wang, 2013; Zhang et al., 2006). Combining these two information leads to the idea that surplus of AU bonds over other ones should be an informative feature. The table A shows that AUsGC, AUsCG, GCsAU, CGsAU, UAsGC, UAsCG, GCsUA and CGsUA features rank two to three times better compared to other XYsWZ features van der Burgt et al. (2009), which have information scores out of first of quartile. According to Figure 3.4 st(A-U) gives high correlation at fifth rank with AU and UA bonds surplus over other bonds.

Correlation studies are used in order to apply feature selection. In the study van der Burgt et al., 2009, some informative features are not selected as they were highly correlated with other features having higher information gain score; minimum free energy

based P value and Z score features were not further used because of their high correlation with minimum free energy normalized to hpl and GC content (hpmfe_rf_I1 (Zhang et al., 2006)). G+C count is a commonly used pre-miRNA predicting feature because of its impact on secondary structure (de ON Lopes et al., 2014; Grad et al., 2003; Zhang et al., 2006). However, by defining the feature hpmfe_rf_I1 to be the most discriminative one and using further, van der Burgt et al., 2009, eliminates the feature G+C count as they claim that its content differs according to taxonomic dataset. Furthermore, this information make the above given results of XYsWZ feature more important as the remaining bonds where either GC or CG would be found for WZ in above sample and hence affect the result of these features. If the feature selection would be done just according to information gain score and given a cut off rank 100, most of the features having biological meaning would be lost. Therefore, feature selection should be applied carefully by having biological meaning in mind and not just considering the numerical values.

For this study Pearson's product-moment coefficient is used, meaning that two variables are correlated according to their numeric values. Therefore, the correlation observed among the features is not a biological one, but it is a numerical correlation. There may be cases like dns features being more correlated within each other as in Figures 3.5 and 3.7, despite a similar correlation not being observed among features given as parameters to those dns features. Moreover, the correlation may not bee seen between the dns feature applied for a feature and the feature. For example, dns(hpl) and hpl have a correlation coefficient of 0.304 (-1 shows strong negative correlation and 1 shows strong positive correlation).

For the features that are calculated faster than other (the top 10 at least) the information gain scores are observed to be low and the features having high information gain are, in contrary, very slow (Table 3.1, 3.2, 3.3 and 3.4). The features can be calculated fast; however, may not have much biological meaning, or express high specificity for positive data and have not much effect to distinctive negative and positive pre-miRNAs. However, the relation between the time taken to calculate a feature and information gain of that feature is widely spread so there are features which follow linear distributions too (Figure 3.1, 3.2). From the scatter plot Figure 3.2, the features which are not normalized, have high information gain scores ($x>0.2$) and low mean times ($y<10^6$) are highlighted in Table 3.6. It was aimed to obtain the informative features which take rather less time in a logical window. Each feature from Table 3.6 are defined in Figures A.1, A.2, A.6, A.3 and A.4 and four of them are structural and one is structural-thermodynamic feature which is

a consistent result with the recent findings in the literature (Chen et al., 2016). Moreover, the boxplots have different numerical ranges and this defines the discriminative power of a feature, too. Their correlation analyses in Figure 3.8 showed that they are not correlated with each other which can be concluded that their usage together would be beneficial.

miRNAs can be predicted from different parts of the genome. Therefore, datasets were retrieved from different genomic parts, separately; and in order to hold the genomic location for further studies in case usage (Section 2.1). The selection of pseudo data was important as it effects the information gain score, the discriminative characteristic of the features defining pre-miRNAs and hairpin prediction accuracy (Chen et al., 2016; Khalifa et al., 2016; Yousef et al., 2016; Wei et al., 2014). Further, the in balance between positive and pseudo data effects the machine learning algorithm (on model training, in balance of classifiers (Ding et al., 2010; Saçar and Allmer, 2013) , however as known pre-miRNAs are limited with that in miRBase database, at that point nothing could be changed. In this study two class classification has been applied as it was aimed before the model creation to define the discriminative power of the features. However, for further studies where the predictive power of the features are anslysed one class classification is an option too. Furthermore, the study Khalifa et al., 2016, reveals that one class classification has not much effect on feature selection compared to two class classification. However, one class classification is mostly applied when there is no proper negative dataset (Allmer, 2012).

In this study cost has the meaning of calculation time, therefore time measurements are mentioned. However, to take the meaning of cost as money, the results presented on Table 3.7 can be considered. The overall time for human hairpin data is higher than virus one as the dataset size is bigger. The time for virus calculation is fast and user friendly, while the time spent to calculate the human hairpins was slower for 1828 hairpin.

Human hairpins as positive class and pseudo data as negative class are used for cost and benefit analysis and model creation. However, the two data are not in equal amounts. As the pseudo data is seven fold more than the human hairpins it is important to distribute the negative data during model creation. Therefore, in Section 2.3.3, after equalizing the datasets, negative data was further used to construct more Random Forest Predictors and a part of the pseudo data was used for testing. As defined by Saçar and Allmer, 2013, 70% is used for training and 30% for testing the model.

The case study is shown with measles virus (MeV) which is from Morbillivirus genus. MeV is a single-stranded negative-sense, non-segmented RNA virus; therefore, their replication occurs within the cytoplasm. Because of that, the mechanism is assumed to take place similar to those of mitrons as one RNA encodes more than one protein. A model is created to predict miRNAs from the MeV genome. The MeV virus KEGG pathway is checked and gene informations are observed from there. As the genes do not have miRNA correlated coding information, the genes sequences are predicted on the model too. The prediction resulted with higher positive compared to negatives (Figures 3.9 and 3.10). A model was created to analyse the case study, that way it was shown that the proposed features are powerful by pre-miRNA detection as result Figures 3.11, 3.12, 3.13 and 3.14 are obtained which have sequences of known mature miRNAs embedded. Furthermore, it was interesting that the computationally predicted BCC3 based hairpin Figure 3.13 has the 3p and Figure 3.14 the 5p mature of the same miRNA and with a shift of 3p end predicted one towards the hairpin loop was interesting as it is known that from a pre-miRNA if matures obtained from both two arms, a shift is observed in one to end because of Drosha and Dicer cleavage cites (Ha and Kim, 2014). Prediction result Figure 3.12 has a mature miRNA sequence buried hsa-miR-548au-5p that target genes which are related with immune system revealed by Reactome analyses. It may be the case that when the human is infected by MeV, TAB2 based pre-miRNA targets immune system related genes, which is normal while disease.

CHAPTER 5

CONCLUSION

miRNAs are studied as they are key regulators in post-transcriptional regulation of gene expression. Because biological conditions can make the experimental set-up difficult and hinder miRNA identification, computational miRNA predictions are more promising. It is important to set up the features properly so that accurate predictions are made. However, feature selection is NP hard (Amaldi and Kann, 1998). Therefore, in this study cost and benefit analyses of the pre-miRNA describing features is studied.

pre-miRNA prediction tools does not relay on one feature, they depend on different feature combinations (Bentwich, 2008). Studies show that when different feature set applied, different results are obtained and their computational cost differs too (de ON Lopes et al., 2014; Ding et al., 2010). It come up that features showing different weighting and features from different groups and a number of 50 features should be used together (Bentwich, 2008; de ON Lopes et al., 2014; Ding et al., 2010; Saçar and Allmer, 2013; Yousef et al., 2016). Therefore, it is important to analyse each feature and for further feature selection studies different feature groups can be analysed on different datasets for their predictive power.

As observed from the results it is not possible to decided sharply how the feature selection should be done. The information gain, correlation, time analysis show that there are many factors effecting the selection.

The aim was not do feature selection however to conclude five informative features are selected that are calculated fast and it come up that they are not correlated with each other. The features are structural assl, lsr(%bp), lscm, asal and structural-thermodynamic hpmfe_rf_I3.

REFERENCES

Allmer, J. (2012). A call for benchmark data in mass spectrometry-based proteomics. *Journal of Integrated OMICS 2*(2), 1–5.

Amaldi, E. and V. Kann (1998). On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science 209*(1), 237–260.

Bağcı, C. and J. Allmer (2016). One step forward, two steps back; xeno-micrornas reported in breast milk are artifacts. *PloS one 11*(1), e0145065.

Bartel, D. P. (2004). MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *116*, 281–297.

Baskerville, S. and D. P. Bartel (2005). Microarray profiling of micrornas reveals frequent coexpression with neighboring mirnas and host genes. *Rna 11*(3), 241–247.

Batuwita, R. and V. Palade (2009). Original paper. *25*(8), 989–995.

Bentwich, I. (2008). Identifying human microRNAs. *Current Topics in Microbiology and Immunology 320*, 257–269.

Bentwich, I., A. Avniel, Y. Karov, R. Aharonov, S. Gilad, O. Barad, A. Barzilai, P. Einat, U. Einav, E. Meiri, et al. (2005). Identification of hundreds of conserved and nonconserved human micrornas. *Nature genetics 37*(7), 766–770.

Berthold, M. R., N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel (2007). KNIME: the konstanz information miner. In *Data Analysis, Machine Learning and Applications - Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e.V., Albert-Ludwigs-Universität Freiburg, March 7-9, 2007*, pp. 319–326.

Çakir, M. V. and J. Allmer (2010). Systematic computational analysis of potential rnai

regulation in toxoplasma gondii. *2010 5th International Symposium on Health Informatics and Bioinformatics, HIBIT 2010* (APRIL 2010), 31–38.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden (2009). Blast+: architecture and applications. *BMC bioinformatics 10*(1), 1.

Camacho, C., T. Madden, G. Coulouris, V. Avagyan, N. Ma, T. Tao, and R. Agarwala (2008). Blast command line applications user manual. *BLAST® Help. Bethesda, MD: National Center for Biotechnology Information (US)*.

Chen, J., X. Wang, and B. Liu (2016). iMiRNA-SSF: Improving the Identification of MicroRNA Precursors by Combining Negative Sets with Different Distributions. *Scientific Reports 6*(October 2015), 19062.

Darty, K., A. Denise, and Y. Ponty (2009). Varna: Interactive drawing and editing of the rna secondary structure. *Bioinformatics 25*(15), 1974–5.

de ON Lopes, I., A. Schliep, and A. C. d. L. de Carvalho (2014). The discriminant power of rna features for pre-mirna recognition. *BMC bioinformatics 15*(1), 1.

Ding, J., S. Zhou, and J. Guan (2010). MiRenSVM: towards better prediction of microRNA precursors using an ensemble SVM classifier with multi-loop features. *BMC Bioinformatics 11*(11), 1–10.

Fera, D., N. Kim, N. Shiffeldrim, J. Zorn, U. Laserson, H. H. Gan, and T. Schlick (2004). BMC Bioinformatics. *9*, 1–9.

Filipowicz, W., S. N. Bhattacharyya, and N. Sonenberg (2008). Mechanisms of post-transcriptional regulation by microrna: are the answers in sight?. *Nature Reviews Genetics 9*(2), 102 – 114.

França, G. S., M. D. Vibranovski, and P. A. Galante (2016). Host gene constraints and genomic context impact the expression and evolution of human microrna. *Nature communications 7*.

Freyhult, E., P. P. Gardner, and V. Moulton (2005). BMC Bioinformatics. *9*, 1–9.

Gan, H. H., D. Fera, J. Zorn, N. Shiffeldrim, M. Tang, U. Laserson, N. Kim, and T. Schlick (2004). RAG: RNA As Graphs database concepts, analysis, and features. *20*(8), 1285–1291.

Grad, Y., J. Aach, G. D. Hayes, B. J. Reinhart, G. M. Church, G. Ruvkun, and J. Kim (2003). Computational and experimental identification of c. elegans micrornas. *Molecular cell 11*(5), 1253–1263.

Gudys, A., M. W. Szczesniak, M. Sikora, and I. Makalowska (2013). HuntMi: an efficient and taxon-specific approach in pre-miRNA identification. *BMC bioinformatics 14*, 83.

Ha, M. and V. N. Kim (2014). Regulation of microRNA biogenesis. *Nature Publishing Group 15*(8), 509–524.

He, L. and G. J. Hannon (2004). MicroRNAs: small RNAs with a big role in gene regulation. *5*(July).

Hinske, L. C., G. S. Franca, H. A. Torres, D. T. Ohara, C. M. Lopes-Ramos, J. Heyn, L. F. Reis, L. Ohno-Machado, S. Kreth, and P. A. Galante (2014). miriad-integrating microrna inter-and intragenic data. *Database 2014*, bau099.

Hinske, L. C. G., P. A. Galante, W. P. Kuo, and L. Ohno-Machado (2010). A potential role for intragenic mirnas on their hosts' interactome. *BMC genomics 11*(1), 1.

Jiang, P., H. Wu, W. Wang, W. Ma, X. Sun, and Z. Lu (2007). MiPred: Classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Research 35*(SUPPL.2), 339–344.

Khalifa, W., M. Yousef, M. D. S. Demirci, and J. Allmer (2016). The impact of feature selection on one and two-class classification performance for plant micrornas. *PeerJ 4*, e2135.

Kozomara, A. and S. Griffiths-Jones (2014). mirbase: annotating high confidence micror-

nas using deep sequencing data. *Nucleic acids research 42*(D1), D68–D73.

Lai, E. C., P. Tomancak, R. W. Williams, and G. M. Rubin (2003). Computational identification of drosophila microrna genes. *Genome biology 4*(7), 1.

Latchman, D. (2010). *Gene Control*. Garland Science.

Lorenz, R., S. H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker (2011). Viennarna package 2.0. *Algorithms for Molecular Biology 6*(1), 1–14.

Markham, N. R. and M. Zuker (2008). Unafold. *Bioinformatics: Structure, Function and Applications*, 3–31.

Meister, G. (2013). Argonaute proteins: functional insights and emerging roles. *Nature Publishing Group 14*(7), 447–459.

Melo, C. A. and S. A. Melo (2014). *Biogenesis and Physiology of MicroRNAs*, pp. 5–24. New York, NY: Springer New York.

Morgulis, A., E. M. Gertz, A. A. Schäffer, and R. Agarwala (2006). A fast and symmetric dust implementation to mask low-complexity dna sequences. *Journal of Computational Biology 13*(5), 1028–1040.

Naik, A., R. Kosir, and D. Rozman (2013). Genomic aspects of {NAFLD} pathogenesis. *Genomics 102*(2), 84 – 95. SI:Clinical and Translational Genomics.

Ng, K. L. S. and S. K. Mishra (2007). De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics 23*(11), 1321–1330.

Ritchie, W., D. Gao, and J. E. J. Rasko (2012). Defining and providing robust controls for microRNA prediction. *Bioinformatics 28*(8), 1058–1061.

Rodriguez, A., S. Griffiths-Jones, J. L. Ashurst, and A. Bradley (2004). Identification of

mammalian microrna host genes and transcription units. *Genome research 14*(10a), 1902–1910.

Saçar, M. D. and J. Allmer (2013). Comparison of four ab initio microrna prediction tools. In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms (BIOSTEC 2013)*, pp. 190–195.

Saçar, M. D. and J. Allmer (2013, Sept). Data mining for microrna gene prediction: On the impact of class imbalance and feature number for microrna gene prediction. In *Health Informatics and Bioinformatics (HIBIT), 2013 8th International Symposium on*, pp. 1–6.

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

Tempel, S. and F. Tahi (2012). A fast ab-initio method for predicting mirna precursors in genomes. *Nucleic acids research 40*(11), e80–e80.

Tempel, S., B. Zerath, F. Zehraoui, F. Tahi, et al. (2015). mirboost: boosting support vector machines for microrna precursor classification. *RNA 21*(5), 775–785.

Uğuz, H. (2011). A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. *Knowledge-Based Systems 24*(7), 1024–1032.

van der Burgt, A., M. W. J. E. Fiers, J.-P. Nap, and R. C. H. J. van Ham (2009). In silico miRNA prediction in metazoan genomes: balancing between sensitivity and specificity. *BMC genomics 10*, 204.

Wang, B. (2013). Base composition characteristics of mammalian mirnas. *Journal of nucleic acids 2013*.

Wei, L., M. Liao, Y. Gao, R. Ji, Z. He, and Q. Zou (2014). Improved and promising identification of human micrornas by incorporating a high-quality negative set. *IEEE/ACM*

Transactions on Computational Biology and Bioinformatics 11(1), 192–201.

Xuan, P., M. Guo, X. Liu, Y. Huang, W. Li, and Y. Huang (2011). Plantmirnapred: efficient classification of real and pseudo plant pre-mirnas. *Bioinformatics 27*(10), 1368–1376.

Xue, C., F. Li, T. He, G.-P. Liu, Y. Li, and X. Zhang (2005). Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC bioinformatics 6*, 310.

Yousef, M. and J. Allmer (2012a). Computational methods for ab initio detection of micrornas. *Frontiers in Genetics 3*(209).

Yousef, M. and J. Allmer (2012b). Computational methods for ab initio detection of micrornas. *Frontiers in genetics 3*, 209.

Yousef, M., M. D. Saçar Demirci, W. Khalifa, and J. Allmer (2016). Feature selection has a large impact on one-class classification accuracy for micrornas in plants. *Advances in bioinformatics 2016*.

Zhang, B. H., X. P. Pan, S. B. Cox, G. P. Cobb, and T. A. Anderson (2006). Evidence that miRNAs are different from other RNAs. *Cellular and Molecular Life Sciences 63*(2), 246–254.

APPENDIX A

RESULTS

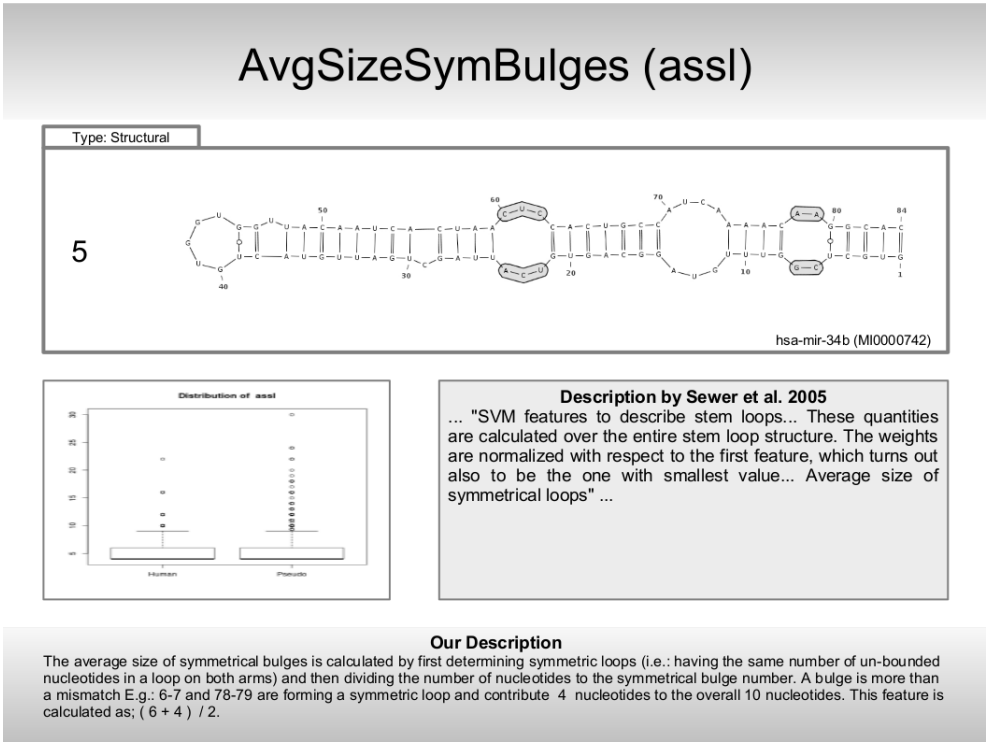

Figure A.1. Structure of hsa-mir-34b (MI0000742) is shown. The feature ”average size of symmetrical loops” assl, is shown for the hairpin. The feature has a distribution with a minimum of 4, a mean of 5 and a maximum of 22 result for the human miRNA data; while the pseudo has a distribution with a minimum of 4, an mean of 6 and a maximum of 30. The feature is implemented from (Sewer et al., 2005). Fold is created with RNAfold 2.1.3 (Lorenz et al., 2011) and image is drawn using VARNA v3-91 (Darty et al., 2009).

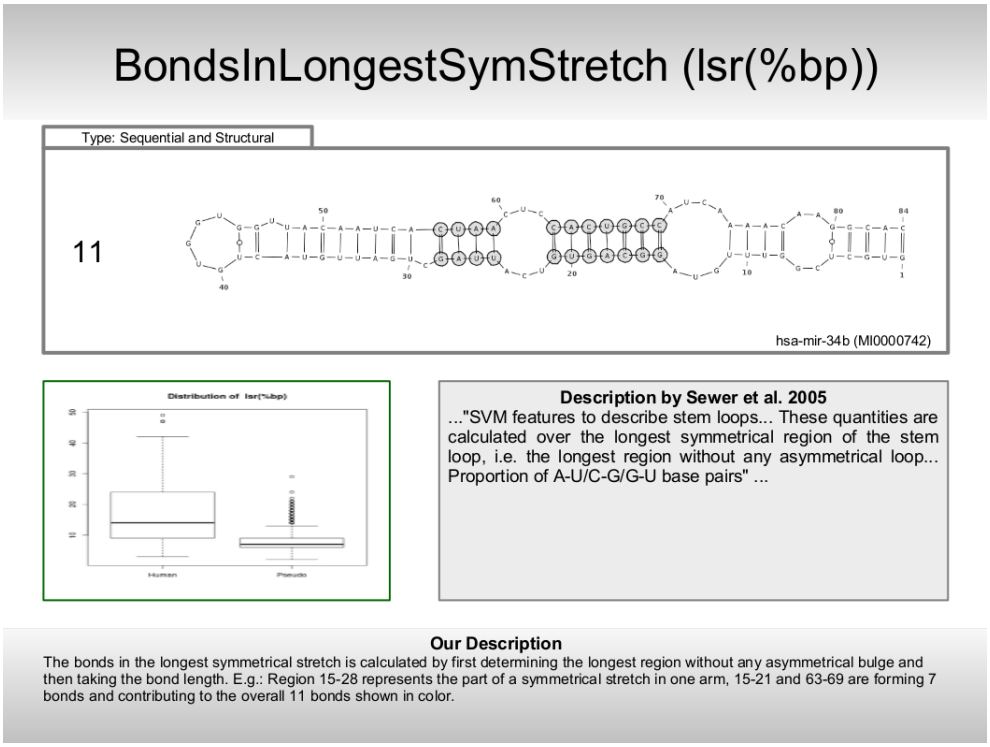

Figure A.2. Structure of hsa-mir-34b (MI0000742) is shown. The feature ”bonds in the longest symmetrical stretch” lsr(%bp), is shown for the hairpin. The feature has a distribution with a minimum of 3, a mean of 17 and a maximum of 49 result for the human miRNA data; while the pseudo has a distribution with a minimum of 2, an mean of 8 and a maximum of 29. The feature is implemented from (Sewer et al., 2005). Fold is created with RNAfold 2.1.3 (Lorenz et al., 2011) and image is drawn using VARNA v3-91 (Darty et al., 2009).

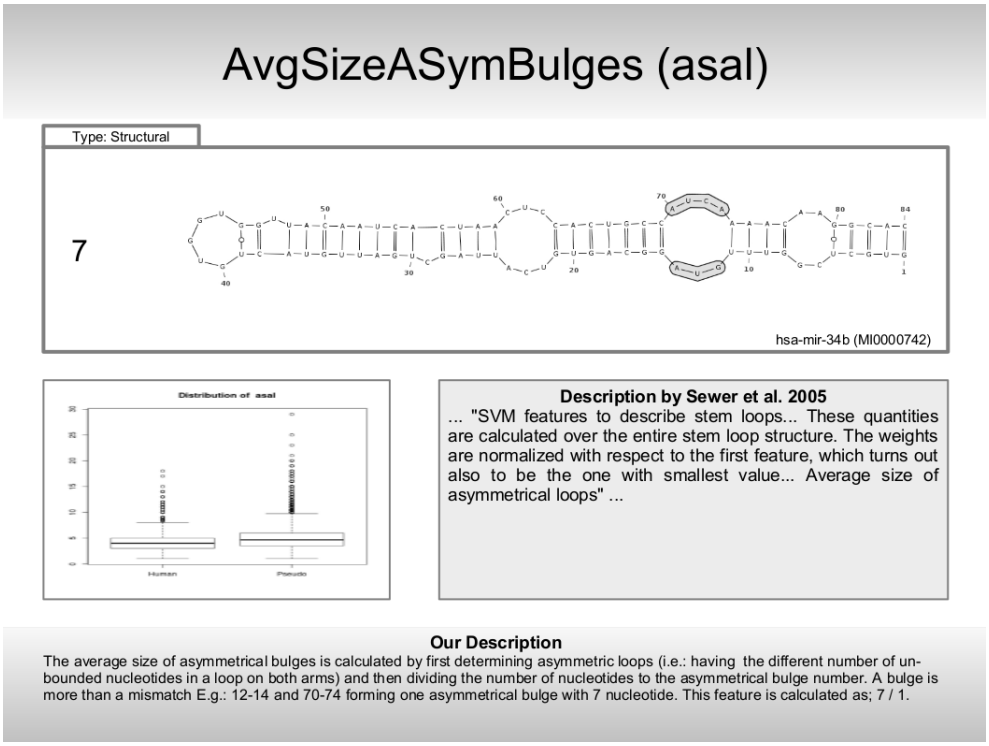

Figure A.3. Structure of hsa-mir-34b (MI0000742) is shown. The feature ”average size of asymmetrical loops” asal, is shown for the hairpin. The feature has a distribution with a minimum of 1, an average of 4 and a maximum of 18 result for the human miRNA data; while the pseudo has a distribution with a minimum of 1, an average of 5 and a maximum of 29. The feature is implemented from (Sewer et al., 2005). Fold is created with RNAfold 2.1.3 (Lorenz et al., 2011) and image is drawn using VARNA v3-91 (Darty et al., 2009).

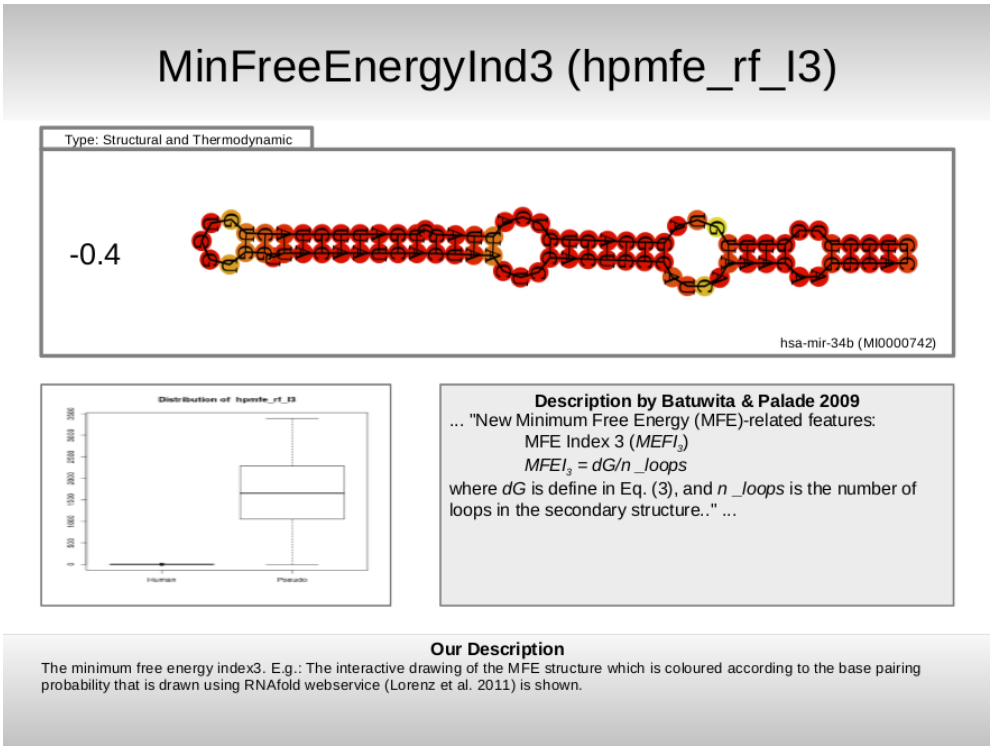

Figure A.4. Structure of hsa-mir-34b (MI0000742) is shown. The feature ”minimum free energy index 3” hpmfe_rf_I3, is shown for the hairpin. The feature has a distribution with a minimum of -0.98, an average of -0.4 and a maximum of -0.07 result for the human miRNA data; while the pseudo has a distribution with a minimum of -0.07, an average of -0.3 and a maximum of 0.04. The feature is implemented from . Fold is created with RNAfold 2.1.3 (Lorenz et al., 2011) and image is drawn using VARNA v3-91 (Darty et al., 2009).

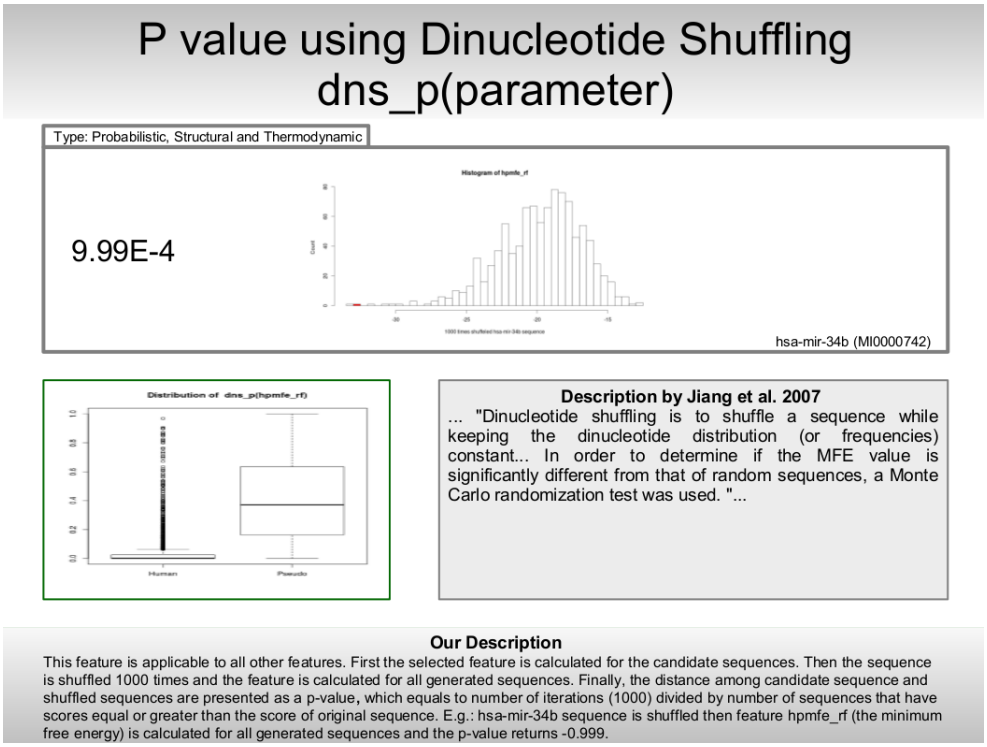

Figure A.5. ”P value using Dinucleotide Shuffling” dns_p(parameter) is shown for hsa-mir-34b (MI0000742) and hpmfe_rf (the minimum free energy) as dns_p(hpmfe_rf).The feature has a distribution with a minimum of 0, an average of 0.05 and a maximum of 0.97 result for the human miRNA data; while the pseudo has a distribution with a minimum of 0, an average of 0.41 and a maximum of 0.99. The feature is implemented from (Jiang et al., 2007). Fold is created with RNAfold 2.1.3 (Lorenz et al., 2011) and image is drawn using VARNA v3-91 (Darty et al., 2009).

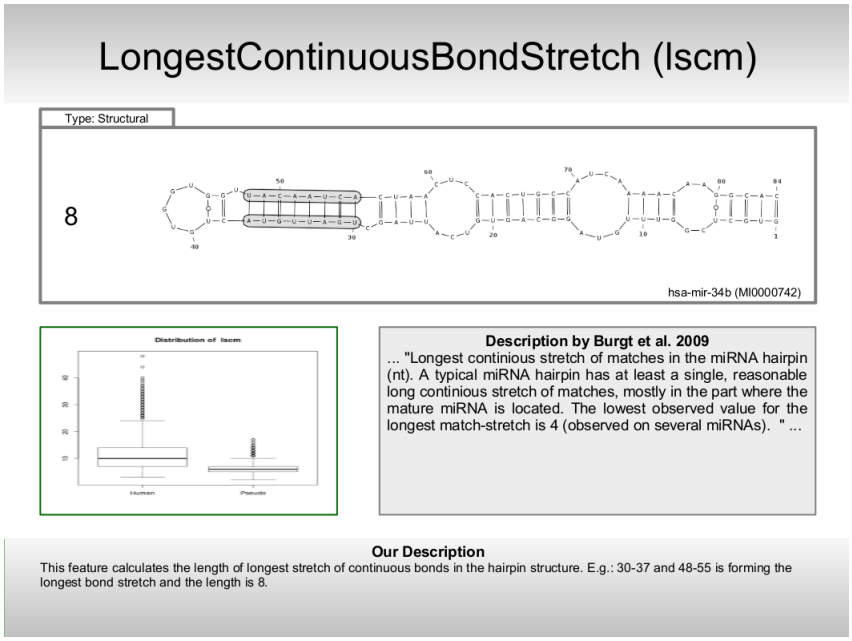

Figure A.6. Structure of hsa-mir-34b (MI0000742) is shown. The feature ”average size of symmetrical loops” assl, is shown for the hairpin. The feature has a distribution with a minimum of 0, an average of 5.85 and a maximum of 30 result for the human miRNA data; while the pseudo has a distribution with a minimum of 0, an average of 5.85 and a maximum of 30. The feature is implemented from (Sewer et al., 2005). Fold is created with RNAfold 2.1.3 (Lorenz et al., 2011) and image is drawn using VARNA v3-91 (Darty et al., 2009).

Table A.1.: pre-miRNA defining features, for each feature Acronym in the code, Synonyms of the feature from publications, Number (N), Name, Type and the Study which the feature is implemented from are stated. The feature implemented new are stated with no source as ”na” not a source.

Acronym	**Synonyms**	**N**	**Name**	**Type**	**Study**
#A... .. #U(((	A... .. U(((	32	Triplet count	Sequential Structural	(Xue et al., 2005)
#A .. #U	A .. U	4	Nucleotide count	Sequential	(Lai et al., 2003)
#A++#U	A+U content	1	A+U content	Sequential	(Zhang et al., 2006)
#AA .. #UU	AA - UU	16	Dinucleotide count	Sequential	(Ng and Mishra, 2007)
#G++#C	G+C content	1	G+C content	Sequential	(Zhang et al., 2006)
%A .. %U	%A .. %U	4	Nucleotide percent	Sequential	(Lai et al., 2003)
%AA .. %UU	%AA .. %UU	16	Dinucleotide percent	Sequential	(Ng and Mishra, 2007)
***A... .. *U(((**	A... .. U(((	32	Triplet frequency	Sequential Structural	(Xue et al., 2005)
bpp/sl	avg_bp_stem	1	Base pairing propensity/ stem length	Structural	(Ding et al., 2010)
bpd	D, diversity, ensemble diversity, bpd	1	Base pairing distance	Structural Thermody-namic	(Freyhult et al., 2005)
bpd/hpl	D/hpl	1	base pairing distance/ hairpin length	Structural Thermody-namic	(Freyhult et al., 2005)

(cont. on next page)

Table A.1 (cont.).

Acronym	Synonyms	N	Name	Type	Study
bpp	bpp, dp	1	Base pairing propensity	Structural	(Ng and Mishra, 2007)
bpp/hpl	bpp/hpl	1	Base pairing propensity/ hairpin length	Structural	(Ding et al., 2010)
bpp/nl	dP/n_loops	1	Base pairing propensity/ number of loops	Structural	(Ding et al., 2010)
bpp/ns	bpp/nStem	1	Base pairing propensity/ number of stem	Structural	(Ding et al., 2010)
dc	dc	1	Degree of compactness	Structural	(Fera et al., 2004), (Gan et al., 2004)
dG	dG_mf	1	Gibbs free energy	Thermodynamic	(Ng and Mishra, 2007)
dG/hpl	dG_mf/L	1	Gibbs free energy/ hairpin length	Thermodynamic	(Ng and Mishra, 2007)
dH	dH_mf	1	Enthalpy	Thermodynamic	(Batuwita and Palade, 2009)
dH/hpl	dH_mf/L	1	Structure Enthalpy	Thermodynamic	(Batuwita and Palade, 2009)
dme	diff	1	Difference of MFE and EFE/ hairpin length	Thermodynamic	(Batuwita and Palade, 2009)
dr	dr	1	Direct repeat	Sequential Structural Thermodynamic	(Bentwich, 2008)

(cont. on next page)

Table A.1 (cont.).

Acronym	Synonyms	N	Name	Type	Study
dS	dS_mf	1	Entropy	Thermodynamic	(Batuwita and Palade, 2009)
dS/hpl	dS_mf/L	1	Structure Entropy	Thermodynamic	(Batuwita and Palade, 2009)
ediv	diversity, ensemble diversity, D, bpd	1	Ensemble Diversity	Thermodynamic	(Batuwita and Palade, 2009)
efe	EFE	1	Ensemble Free Energy	Thermodynamic	(Batuwita and Palade, 2009)
efq	Freq	1	Ensemble frequency	Thermodynamic	(Batuwita and Palade, 2009)
#G++#C/hpl	GhC,G+C content	1	G+C content	Sequential	(Zhang et al., 2006)
hll	HLL	1	Hairpin loop length	Structural	(Bentwich, 2008)
hpmfe_rf/hpl	hmfe/hpl, mfe4	1	Hairpin MFE/ hairpin length	Structural Thermodynamic	(Bentwich, 2008)
hpmfe_rf	hpmfe	1	Hairpin MFE	Structural Thermodynamic	(Jiang et al., 2007)
hpmfe_rs	hpmfe	1	hairpin mfe	Structural Thermodynamic	(Çakir and Allmer, 2010)
hpl	hpl	1	Hairpin length	Structural	(Bentwich et al., 2005)
hpl/ns	hpl/nStem	1	Hairpin length/ number of stem	Structural	na
ir	ir	1	Inverted repeat	Sequential Structural	(Bentwich, 2008)

(cont. on next page)

Table A.1 (cont.).

Acronym	Synonyms	N	Name	Type	Study
mbs	MBS, largest bulge	1	Maximum bulge size	Structural	(Bentwich, 2008)
hpmfe_rf_I1	MFEI1, MFEahl	1	MFE index 1	Structural Thermodynamic	(Zhang et al., 2006)
hpmfe_rf/ns	MFEI2, mfe/nStems, mfe/ns	1	MFE index 2	Structural Thermodynamic	(Ng and Mishra, 2007)
hpmfe_rf/ns/hpl	MFEI3, hpmfe_rf/hpl/nl	1	MFE index 3	Structural Thermodynamic	(Batuwita and Palade, 2009)
hpmfe_rf/hpl	MFEI4, hpmfe_rf/hpl	1	MFE index 4	Structural Thermodynamic	(Batuwita and Palade, 2009)
nl	n_loops	1	Number of Loops	Structural	(Xue et al., 2005)
ns	nStem	1	Number of Stems	Structural	(Xue et al., 2005)
dns_p(efe)	p(EFEp)	1	P value using dinucleotide shuffling for efe	Probabilistic Thermodynamic	(Jiang et al., 2007)
Q	Q	1	Shannon entropy	Probabilistic Thermodynamic	(Ng and Mishra, 2007)
Q/hpl	Q/hpl	1	Shannon entropy/ hairpin length	Probabilistic Thermodynamic	(Ng and Mishra, 2007)
st(A-U)/ns	AUS/nStem	1	AU bonds in stem / number of stem	Sequential Structural	na
st(A-U)/sl	AUS/aStl	1	AU bonds in stem / stem length	Sequential Structural	na

(cont. on next page)

Table A.1 (cont.).

Acronym	Synonyms	N	Name	Type	Study
st(G-C)/ns	GCS/nStem	1	GC bonds in stem / number of stem	Sequential Structural	na
st(G-C)/sl	GCS/aStl	1	GC bonds in stem / stem length	Sequential Structural	na
st(G-U)/sl	GUS/aStl	1	GU bonds in stem / stem length	Sequential Structural	na
sl	stem length	1	Stem length	Structural	(Lai et al., 2003)
dns_p(bpd)	p(Dp)	1	P value using dinucleotide shuffling for bpd	Probabilistic Structural Thermody-namic	(Jiang et al., 2007)
dns_p(bpd/hpl)	p(D/hplp)	1	P value using dinucleotide shuffling for bpd/ hpl	Probabilistic Structural Thermody-namic	(Jiang et al., 2007)
dns_p(bpp)	p(bppp)	1	P value using dinucleotide shuffling for bpp	Probabilistic Structural	(Jiang et al., 2007)
dns_p(bpp/hpl)	p(bpp/hplp)	1	P value using dinucleotide shuffling for bpp/ hpl	Probabilistic Structural	(Jiang et al., 2007)
dns_p(hpl)	P/hpl	1	P value using dinucleotide shuffling for hpl	Probabilistic Structural	(Jiang et al., 2007)
dns_p(hpmfe_rf)	p(hpmfe_rfp)	1	P value using dinucleotide shuffling for mfe	Probabilistic Structural Thermody-namic	(Jiang et al., 2007)

(cont. on next page)

Table A.1 (cont.).

Acronym	Synonyms	N	Name	Type	Study
dns_p(hpmfe_rf/ hpl)	p(hpmfe_rf/ hplp)	1	P value using dinucleotide shuffling for mfe/ hpl	Probabilistic Structural Thermody-namic	(Jiang et al., 2007)
dns_p(Q)	p(Qp)	1	P value using dinucleotide shuffling for Q	Probabilistic Thermody-namic	(Jiang et al., 2007)
dns_p(Q/hpl)	p(Q/hplp)	1	P value using dinucleotide shuffling for Q/ hpl	Probabilistic Thermody-namic	(Jiang et al., 2007)
dns_z(bpd)	z(DZ)	1	Z score using dinucleotide shuffling for bpd	Probabilistic Structural Thermody-namic	(Ng and Mishra, 2007)
dns_z(bpd/hpl)	z(D/hplZ)	1	Z score using dinucleotide shuffling for bpd/ hpl	Probabilistic Structural Thermody-namic	(Ng and Mishra, 2007)
dns_z(bpp)	z(bppZ)	1	Z score using dinucleotide shuffling for bpp	Probabilistic Structural	(Ng and Mishra, 2007)
dns_z(Q)	z(QZ)	1	Z score using dinucleotide shuffling for Q	Probabilistic Thermody-namic	(Ng and Mishra, 2007)
dns_z(Q/hpl)	z(Q/hplZ)	1	Z score using dinucleotide shuffling for Q/ hpl	Probabilistic Thermody-namic	(Ng and Mishra, 2007)
st(A-U)		1	AU bonds in stem	Sequential	na
st(G-C)		1	GC bonds in stem	Sequential	na

(cont. on next page)

Table A.1 (cont.).

Acronym	Synonyms	N	Name	Type	Study
st(G-U)		1	GU bonds in stem	Sequential	na
Tm	Tm_mf	1	Melting temperature	Thermodynamic	(Ding et al., 2010)
Tm/hpl	Tm_mf/L	1	Melting temperature/ hairpin length	Thermodynamic	(Ding et al., 2010)
dns_z(bpp/hpl)	z(bpp/hplZ)	1	Z score using dinucleotide shuffling for bpp/ hpl	Probabilistic Structural	(Ng and Mishra, 2007)
dns_z(efe)	z(EFEZ)	1	Z score using dinucleotide shuffling for efe	Probabilistic Thermodynamic	(Ng and Mishra, 2007)
dns_z(hpmfe_rf)	z(hpmfe_rfZ)	1	Z score using dinucleotide shuffling for mfe	Probabilistic Structural Thermodynamic	(Ng and Mishra, 2007)
dns_z(hpmfe_rf/ hpl)	z(hpmfe_rf/ hplZ)	1	Z score using dinucleotide shuffling for mfe/ hpl	Probabilistic Structural Thermodynamic	(Ng and Mishra, 2007)
#nisl_h	#nisl_h	1	Number of nucleotides in symmetric bulges	Structural	(Sewer et al., 2005)
#nial_h	#nial_h	1	Number of nucleotides in asymmetric bulges	Structural	(Sewer et al., 2005)
adbil	avg_dbil	1	Average distance between internal loops	Structural	(Sewer et al., 2005)
assl	avg_ssl	1	Average size of symmetric bulges	Structural	(Sewer et al., 2005)

(cont. on next page)

Table A.1 (cont.).

Acronym	Synonyms	N	Name	Type	Study
asal	avg_asl	1	Average size of asymmetric bulges	Structural	(Sewer et al., 2005)
l(lsr)	len_lsr	1	Length calculated over the longest symmetrical region (lsr)	Structural	(Sewer et al., 2005)
dfhl	dfhl	1	Distance of lsr from the hairpin loop	Structural	(Sewer et al., 2005)
lsr(%A-U)	bp_lsr	1	Proportion of AU bonds in lsr	Sequential Structural	(Sewer et al., 2005)
lsr(%G-C)	bp_lsr	1	Proportion of GC bonds in lsr	Sequential Structural	(Sewer et al., 2005)
lsr(%G-U)	bp_lsr	1	Proportion of GU bonds in lsr	Sequential Structural	(Sewer et al., 2005)
lsr(%bp)	bp_lsr	1	Proportion of bonds in lsr	Sequential Structural	(Sewer et al., 2005)
lsr(%A)	A lsr	1	Proportion of A in lsr	Sequential Structural	(Sewer et al., 2005)
lsr(%C)	C lsr	1	Proportion of C in lsr	Sequential Structural	(Sewer et al., 2005)
lsr(%G)	G lsr	1	Proportion of G in lsr	Sequential Structural	(Sewer et al., 2005)
lsr(%U)	U lsr	1	Proportion of U in lsr	Sequential Structural	(Sewer et al., 2005)
dns	Shuffling method,sm, sns, dns, or rnd	1	Dinucleotide shuffling	Probabilistic	(Jiang et al., 2007)

(cont. on next page)

Table A.1 (cont.).

Acronym	Synonyms	N	Name	Type	Study
c#A .. c#N	PolyA - polyN	5	The longest continous number of nucleotide X in the hairpin	Sequential Structural	(van der Burgt et al., 2009)
AsC .. UsG	XsurplusY	12	Surplus X over Y	Sequential	(van der Burgt et al., 2009)
AcsGU .. UgsCA	XysurplusWZ	24	Surplus of A and C over G and U	Sequential	(van der Burgt et al., 2009)
%A-U		1	Percentage fo A-U bonds of all constituting bonds.	Sequential Structural	(van der Burgt et al., 2009)
%G-C		1	Percentage fo G-C bonds of all constituting bonds.	Sequential Structural	(van der Burgt et al., 2009)
%G-U		1	Percentage fo G-U bonds of all constituting bonds.	Sequential Structural	(van der Burgt et al., 2009)
mscs	SCS-mono	1	Dinucleotide sequence complexity score	Sequential	(van der Burgt et al., 2009)
dscs	SCS-di	1	Mononucleotide sequence complexity score	Sequential	(van der Burgt et al., 2009)
lscm	longest matchstretch	1	Longest continious stretch of matches in the hairpin	Structural	(van der Burgt et al., 2009)

(cont. on next page)

Table A.1 (cont.).

Acronym	Synonyms	N	Name	Type	Study
#nial_h/sl	bulge ratio	1	Number of nucleotides in asymmetric bulges/ sl	Structural	(van der Burgt et al., 2009)
#mdn	MaxDiBaseCount	1	Count of the two most occurring two bases	Sequential	(van der Burgt et al., 2009)
#mdn/hpl–c(0.5)	MaxDiBaseRatio	1	Count of the most occurring two bases/ hpl	Sequential	(van der Burgt et al., 2009)
c(#)		1	A constant number (#)		na
#mnn	MinBasecount	1	Minimal base occurrence	Sequential	(van der Burgt et al., 2009)
#mnn/hpl	minimal base occurrence	1	Minimal base occurrence	Sequential	(van der Burgt et al., 2009)
c#As .. c#Ns	PolyAstem - polyNstem	5	The longest continous number of A in the stem	Sequential	(van der Burgt et al., 2009)
saln		1	Stem alignment length	Structural	(van der Burgt et al., 2009)
bpp/saln	match ratio in hairpin stem	1	Base pairing propensity/ saln	Structural	(van der Burgt et al., 2009)
#goh	gapratio	1	Gap openings in hairpin alignment	Structural	(van der Burgt et al., 2009)
#gih	gap open number	1	Total gaps in alignment	Structural	(van der Burgt et al., 2009)

(cont. on next page)

Table A.1 (cont.).

Acronym	Synonyms	N	Name	Type	Study
#gih/saln	gap open ratio	1	Total gaps in alignment/ saln	Structural	(van der Burgt et al., 2009)
adal	stem symmetry	1	Absolute length difference between both arms.	Structural	(van der Burgt et al., 2009)
adalr	stem symmetry	1	Stem length symmetry	Structural	(van der Burgt et al., 2009)
llha	length of longest helical arm	1	Length of longest helical arm	Structural	(Lai et al., 2003)
mfe-lha	free energy of the longest helical arm	1	MFE of the longest helical arm	Structural Thermody-namic	(Lai et al., 2003)
#bu	number of bulges	1	Number of bulges	Structural	(Ritchie et al., 2012)
mll	loop size	1	Max loop length	Structural	na
orf	orf	1	Open reading frame	Sequential	(Gudys et al., 2013)
mwmF	max match count	1	Maximum matche count in 24 positions in the stem alignment	Structural Thermody-namic	(van der Burgt et al., 2009)
#mnnS	minimal base ratio	1	minimum base count in stem	Sequential Structural	(van der Burgt et al., 2009)
subu	bulged nucleotides	1	Nucleotide count in all bulges	Structural	(Lai et al., 2003)
%AAs .. %UUs		16	Dinucleotide percent in stem	Sequential Structural	na

(cont. on next page)

Table A.1 (cont.).

Acronym	Synonyms	N	Name	Type	Study
dns_z(hpl)		1	Z score using dinucleotide shuffling for hpl	Structural Thermodynamic	(Ng and Mishra, 2007)
nfl	length of flanking ends	1	Number of unbound nucleotides in flanking ends	Structural	(Çakir and Allmer, 2010)
#AAs .. #UUs	AA .. UU	16	Dinucleotide count in stem	Sequential Structural	(Ng and Mishra, 2007)
#AAA .. #UUU	AAA .. UUU	64	Trinucleotide count	Sequential	(Chen et al., 2016)
%AAA .. %UUU	AAA .. UUU	64	Trinucleotide percent	Sequential	(Chen et al., 2016)
#AAAs #UUUs	AAA .. UUU	64	Trinucleotide count in stem	Sequential Structural	(Chen et al., 2016)
%AAAs .. %UUUs	AAA .. UUU	64	Trinucleotide percent in stem	Sequential Structural	(Chen et al., 2016)

Table A.2.: Time and information gain table of features. Results are obtained in Knime. Human hairpins are used. Each row represents a feature indicated and its information gain (IG), mean time (MT) in milliseconds. Information about the features are found in Table A1.

Feature	IG	MT	Feature	IG	MT
assl/hpl	0.389	0.219	**assl/sl**	0.385	0.231
assl	0.384	0.129	**subu/sl**	0.265	0.179
bpp/sl	0.265	0.234	**subu/hpl**	0.254	0.165
hpmfe_rf_I1/hpl	0.244	0.216	**lsr(%bp)**	0.244	0.118
hpmfe_rf_I1/sl	0.242	0.318	**lsr(%bp)/hpl**	0.235	0.174
lsr(%bp)/sl	0.235	0.262	**bpp/hpl**	0.225	0.221
lscm	0.210	0.106	**dG/sl**	0.208	0.211
asal/hpl	0.197	0.219	**asal/sl**	0.196	0.235
dG/hpl	0.196	0.215	**hpmfe_rf/sl**	0.196	0.263
efe/sl	0.195	0.243	**asal**	0.194	0.115
hpmfe_rf_I3	0.189	0.421	**hpmfe_rf_I4**	0.186	0.332
hpmfe_rf/hpl	0.186	0.170	**efe/hpl**	0.186	0.232
lscm/hpl	0.180	0.162	**lscm/sl**	0.177	0.255
subu	0.173	0.105	**l(lsr)**	0.173	0.114
hpmfe_rf/ns	0.172	0.178	**#gih/hpl**	0.169	0.189
#A(((/sl	0.167	0.168	**l(lsr)/sl**	0.166	0.256
***A(((**	0.166	0.111	**#gih/sl**	0.166	0.311
#A(((/hpl	0.164	0.159	**l(lsr)/hpl**	0.164	0.167
#A(((	0.161	0.029	**#goh/hpl**	0.157	0.171
#nial_h/sl	0.153	0.300	**lsr(%U)**	0.153	0.154
lsr(%U)/sl	0.152	0.313	**#goh/sl**	0.151	0.191
#nial_h++#nisl_h	0.151	0.053	**#nial_h/hpl**	0.150	0.188
st(A-U)/sl	0.149	0.224	**#U(((/sl**	0.148	0.213
***U(((**	0.148	0.213	**lsr(%U)/hpl**	0.148	0.209
#gih	0.148	0.067	**adbil**	0.146	0.122
saln/sl	0.146	0.263	**st(A-U)/hpl**	0.145	0.221
#U(((/hpl	0.144	0.198	**lsr(%A)**	0.143	0.154
lsr(%A)/sl	0.141	0.308	**Q/hpl**	0.140	0.206

(cont. on next page)

Table A.2 (cont.).

Feature	IG	MT	Feature	IG	MT
#U(((	0.139	0.060	**lsr(%A)/hpl**	0.139	0.207
Q/sl	0.138	0.282	**st(A-U)**	0.135	0.146
#goh	0.135	0.050	**adbil/sl**	0.134	0.236
mbs/hpl	0.133	0.163	**#nial_h**	0.132	0.050
adbil/hpl	0.132	0.215	**mbs/sl**	0.130	0.263
***A(((/sl**	0.130	0.241	***A(((/hpl**	0.129	0.233
ediv/hpl	0.120	0.232	**bpd/hpl**	0.120	0.219
mbs	0.119	0.121	**ediv/sl**	0.118	0.244
bpd/sl	0.118	0.266	**dH/sl**	0.108	0.211
***U(((/sl**	0.106	0.307	***U(((/hpl**	0.105	0.272
bpp	0.102	0.067	**lsr(%G)**	0.097	0.155
lsr(%C)	0.096	0.152	**dH/hpl**	0.094	0.208
lsr(%G)/sl	0.090	0.299	**lsr(%C)/sl**	0.090	0.354
***C.../sl**	0.088	0.259	**lsr(%C)/hpl**	0.086	0.209
***C...**	0.086	0.111	**#C.../sl**	0.086	0.195
#C.../hpl	0.086	0.207	**lsr(%G)/hpl**	0.086	0.217
***C.../hpl**	0.082	0.233	**#A.../hpl**	0.082	0.199
***A.../sl**	0.082	0.274	***A...**	0.082	0.104
***A.../hpl**	0.081	0.227	**#A.../sl**	0.081	0.166
#C...	0.079	0.035	**lsr(%G-U)**	0.076	0.172
lsr(%A-U)	0.076	0.180	**mwmF/sl**	0.075	0.290
%A-U	0.075	0.089	**#bu**	0.075	0.128
#A...	0.073	0.030	**dS/sl**	0.069	0.219
saln/hpl	0.066	0.166	**adal/hpl**	0.066	0.216
mwmF/hpl	0.065	0.204	**adalr**	0.065	0.091
lsr(%G-U)/sl	0.064	0.314	**adal/sl**	0.063	0.230
#U.../sl	0.063	0.215	***U...**	0.062	0.117
***U.../sl**	0.061	0.260	**#U.../hpl**	0.061	0.187
***U.../hpl**	0.061	0.257	**lsr(%G-U)/hpl**	0.060	0.221
adalr/hpl	0.059	0.242	**adalr/sl**	0.058	0.247
dS/hpl	0.058	0.212	**#UA/hpl**	0.054	0.202

(cont. on next page)

Table A.2 (cont.).

Feature	IG	MT	Feature	IG	MT
#A..(/hpl	0.054	0.163	**#G.../hpl**	0.054	0.210
#UA/sl	0.054	0.212	**%UA**	0.054	0.070
***G...**	0.054	0.115	**#G.../sl**	0.054	0.199
Tm/sl	0.054	0.208	**#U...**	0.053	0.062
adal	0.052	0.103	***G.../sl**	0.051	0.292
***A..(**	0.051	0.101	**Tm/hpl**	0.051	0.205
#UA	0.050	0.051	**%A-U/sl**	0.050	0.224
#A..(/sl	0.049	0.179	***G.../hpl**	0.049	0.237
%A-U/hpl	0.049	0.214	**#G.((/hpl**	0.048	0.281
nl/hpl	0.048	0.162	**%G++%C**	0.047	0.118
%C++%G	0.047	0.164	**%U++%A**	0.047	0.108
%A++%U	0.047	0.228	**#C(../hpl**	0.047	0.217
#G...	0.046	0.042	**#A(../hpl**	0.046	0.154
%G++%C/hpl	0.046	0.266	**%C++%G/hpl**	0.046	0.207
ns/hpl	0.045	0.157	**#U++#A/sl**	0.045	0.217
#A++#U/sl	0.045	0.182	**UAsGC/sl**	0.045	0.220
UAsCG/sl	0.045	0.221	**AUsGC/sl**	0.045	0.273
AUsCG/sl	0.045	0.273	**GCsUA/sl**	0.045	0.314
GCsAU/sl	0.045	0.301	**CGsUA/sl**	0.045	0.317
CGsAU/sl	0.045	0.276	**UAsGC**	0.045	0.210
UAsCG	0.045	0.229	**GCsUA**	0.045	0.193
GCsAU	0.045	0.193	**CGsUA**	0.045	0.191
CGsAU	0.045	0.180	**AUsGC**	0.045	0.178
AUsCG	0.045	0.218	**#G++#C/hpl**	0.045	0.182
#C++#G/hpl	0.045	0.177	**#U++#A/hpl**	0.045	0.196
#A++#U/hpl	0.045	0.166	**UAsGC/hpl**	0.044	0.215
UAsCG/hpl	0.044	0.221	**AUsGC/hpl**	0.044	0.252
AUsCG/hpl	0.044	0.283	**GCsUA/hpl**	0.044	0.274
GCsAU/hpl	0.044	0.273	**CGsUA/hpl**	0.044	0.271
CGsAU/hpl	0.044	0.265	**%CG**	0.044	0.058
#CG/hpl	0.044	0.174	**%CG/hpl**	0.044	0.225

(cont. on next page)

Table A.2 (cont.).

Feature	IG	MT	Feature	IG	MT
***A..(/hpl**	0.044	0.235	***G.((/hpl**	0.044	0.275
***A..(/sl**	0.044	0.284	**#G.((/sl**	0.043	0.196
%G-U	0.043	0.096	***G.((**	0.043	0.106
%UA/sl	0.043	0.236	**%CG/sl**	0.042	0.228
lsr(%G-C)	0.042	0.162	**#G(((/sl**	0.042	0.187
#CG/sl	0.042	0.189	***G.((/sl**	0.042	0.250
%G++%C/sl	0.042	0.309	**%C++%G/sl**	0.042	0.283
***C(..**	0.041	0.098	**#C(../sl**	0.041	0.234
***G(((**	0.041	0.154	**#A(../sl**	0.041	0.180
***A(..**	0.041	0.095	**#C((./hpl**	0.041	0.208
dme/sl	0.039	0.246	**%UAA**	0.039	0.199
#G(((	0.039	0.040	**#G++#C/sl**	0.039	0.195
#C++#G/sl	0.039	0.212	**#C(((/sl**	0.038	0.188
nl/sl	0.038	0.249	***A(../sl**	0.038	0.242
#G.((	0.038	0.047	**#A..(**	0.038	0.022
***A(../hpl**	0.038	0.288	**#CG**	0.037	0.029
***C(((**	0.037	0.103	***C(../hpl**	0.037	0.224
#U++#A	0.036	0.034	**#A++#U**	0.036	0.082
#G(((/hpl	0.036	0.189	**ns/sl**	0.036	0.251
#C..(/hpl	0.036	0.238	**#C((./sl**	0.036	0.194
%UA/hpl	0.036	0.226	***C((.**	0.036	0.101
***C(../sl**	0.036	0.349	**#C(((**	0.035	0.032
#A(..	0.035	0.023	***C((./hpl**	0.034	0.277
#C(((/hpl	0.034	0.167	**GsA/sl**	0.034	0.283
AsG/sl	0.034	0.275	**%UUA**	0.033	0.170
hpl	0.033	0.072	**#UAA**	0.033	0.169
***C((./sl**	0.033	0.243	**#C(..**	0.033	0.042
GsA	0.032	0.144	**AsG**	0.032	0.117
orf/hpl	0.032	0.166	**UsG**	0.032	0.198
GsU	0.032	0.171	**lsr(%A-U)/sl**	0.032	0.316
GsU/sl	0.032	0.294	**UsG/sl**	0.032	0.214

(cont. on next page)

Table A.2 (cont.).

Feature	IG	MT	Feature	IG	MT
lsr(%A-U)/hpl	0.032	0.221	**GsA/hpl**	0.031	0.274
AsG/hpl	0.031	0.246	**#UUA**	0.031	0.188
***C..(/sl**	0.031	0.250	**#C..(/sl**	0.031	0.256
***C..(**	0.031	0.107	**UsC**	0.031	0.176
CsU	0.031	0.148	**%CUA**	0.031	0.167
efq/sl	0.031	0.243	***C..(/hpl**	0.030	0.229
***G..(**	0.030	0.230	**#C((.**	0.030	0.037
***G(../sl**	0.030	0.256	**#G..(/sl**	0.030	0.195
sl	0.030	0.084	***G(../hpl**	0.030	0.232
***G..(/sl**	0.030	0.254	**%G++%G/hpl**	0.030	0.220
#G(../hpl	0.029	0.175	***G..(/hpl**	0.029	0.234
#U++#U/sl	0.029	0.251	**#U/sl**	0.029	0.214
UsG/hpl	0.029	0.216	**GsU/hpl**	0.029	0.258
%G/hpl	0.029	0.279	***G(..**	0.029	0.113
#G(../sl	0.029	0.224	**UsC/sl**	0.029	0.219
CsU/sl	0.029	0.277	**#G..(/hpl**	0.029	0.213
efq/hpl	0.029	0.225	**orf/sl**	0.028	0.256
CsU/hpl	0.028	0.268	**UsC/hpl**	0.028	0.217
#U..(/hpl	0.028	0.184	**#G++#C**	0.028	0.030
#C++#G	0.028	0.060	**#U**	0.028	0.036
#U++#U	0.028	0.072	**%G++%G/sl**	0.027	0.246
dme/hpl	0.027	0.228	**%G-C/hpl**	0.027	0.222
%C++%C/hpl	0.027	0.239	**#U(../hpl**	0.026	0.194
%C/hpl	0.026	0.204	**%G/sl**	0.026	0.255
%U++%U	0.026	0.117	**%U**	0.026	0.069
#U/hpl	0.026	0.241	**#U++#U/hpl**	0.026	0.206
#G((./hpl	0.026	0.239	**#G(..**	0.026	0.049
orf	0.026	0.200	**#C..(**	0.025	0.032
***U(..**	0.025	0.119	**#U(../sl**	0.025	0.248
#U..(/sl	0.025	0.261	***U..(**	0.025	0.118
#G..(	0.025	0.045	**%C++%C/sl**	0.025	0.257

(cont. on next page)

Table A.2 (cont.).

Feature	IG	MT	Feature	IG	MT
%C/sl	0.024	0.223	**%A++%G**	0.024	0.159
%G++%G	0.024	0.118	**%G**	0.024	0.060
#G/hpl	0.024	0.211	**#G++#G/hpl**	0.024	0.218
%U++%C	0.024	0.122	**%C++%U**	0.024	0.187
%G++%A	0.024	0.094	**%CCG**	0.024	0.186
***C(((/sl**	0.024	0.274	**%UAG**	0.024	0.172
#CUA	0.024	0.165	***C(((/hpl**	0.023	0.227
%G-C	0.023	0.139	**GAsUC**	0.023	0.196
GAsCU	0.023	0.201	**AGsUC**	0.023	0.165
AGsCU	0.023	0.168	**#G++#A/hpl**	0.023	0.183
#A++#G/hpl	0.023	0.171	**UCsGA**	0.023	0.207
UCsAG	0.023	0.215	**CUsGA**	0.023	0.185
CUsAG	0.023	0.188	**#U++#C/hpl**	0.023	0.196
#C++#U/hpl	0.023	0.177	**%GC/hpl**	0.023	0.221
%G-C/sl	0.022	0.237	***G(((/sl**	0.022	0.250
#C.((/hpl	0.022	0.211	***U..(/hpl**	0.022	0.240
CsA/sl	0.022	0.272	**AsC/sl**	0.022	0.264
%U++%A/sl	0.022	0.243	**%A++%U/sl**	0.022	0.259
***U(../sl**	0.022	0.306	***U(../hpl**	0.022	0.311
AsC	0.022	0.112	**CsA**	0.022	0.186
%GC	0.022	0.062	**#CCG**	0.022	0.169
CsA/hpl	0.022	0.250	**AsC/hpl**	0.022	0.240
#G((./sl	0.021	0.254	**AsU**	0.021	0.128
UsA	0.021	0.175	**%G-U/sl**	0.021	0.290
***G(((/hpl**	0.021	0.237	**nl**	0.021	0.085
%GC/sl	0.021	0.233	**c#Us/hpl**	0.021	0.261
%CGG	0.020	0.185	**%C++%C**	0.020	0.099
%C	0.020	0.056	**#C/hpl**	0.020	0.165
#C++#C/hpl	0.020	0.189	**#G/sl**	0.020	0.182
#G++#G/sl	0.020	0.191	**#UUU**	0.020	0.176
***U..(/sl**	0.020	0.269	**%G-U/hpl**	0.020	0.211

(cont. on next page)

Table A.2 (cont.).

Feature	IG	MT	Feature	IG	MT
#GC/hpl	0.020	0.192	**#U..(**	0.020	0.057
#GC/sl	0.020	0.205	**%UUU**	0.020	0.182
lsr(%G-C)/hpl	0.020	0.224	***G((.**	0.019	0.115
GAsUC/hpl	0.019	0.271	**GAsCU/hpl**	0.019	0.276
AGsUC/hpl	0.019	0.286	**AGsCU/hpl**	0.019	0.244
UCsGA/hpl	0.019	0.215	**UCsAG/hpl**	0.019	0.219
CUsGA/hpl	0.019	0.264	**CUsAG/hpl**	0.019	0.267
%AU	0.019	0.065	***G((./hpl**	0.019	0.230
c#U/hpl	0.019	0.247	**%U++%A/hpl**	0.019	0.225
%A++%U/hpl	0.019	0.253	**#UU**	0.019	0.059
ns	0.019	0.094	***G((./sl**	0.019	0.322
#G++#A/sl	0.019	0.204	**#A++#G/sl**	0.019	0.236
%CCA	0.018	0.198	**#G++#G**	0.018	0.041
#G	0.018	0.024	**#AU/hpl**	0.018	0.178
GAsUC/sl	0.018	0.305	**GAsCU/sl**	0.018	0.293
AGsUC/sl	0.018	0.259	**AGsCU/sl**	0.018	0.346
UCsGA/sl	0.018	0.227	**UCsAG/sl**	0.018	0.226
CUsGA/sl	0.018	0.317	**CUsAG/sl**	0.018	0.278
lsr(%G-C)/sl	0.018	0.352	**UsA/hpl**	0.018	0.214
AsU/hpl	0.018	0.274	**%AAA**	0.018	0.181
%U++%U/sl	0.018	0.240	**%GGC**	0.018	0.176
#UU/sl	0.018	0.208	**%UAU**	0.018	0.180
#U(..	0.018	0.052	**%UAC**	0.018	0.170
#A++#A	0.018	0.004	**#A**	0.018	0.008
%U/sl	0.018	0.240	**UsA/sl**	0.018	0.214
AsU/sl	0.018	0.260	**%GCC**	0.018	0.170
#UAU	0.017	0.182	**#CGG**	0.017	0.182
#UU/hpl	0.017	0.198	**#C.((/sl**	0.017	0.186
%UU	0.017	0.072	**c#U**	0.017	0.101
#C/sl	0.017	0.178	**#C++#C/sl**	0.017	0.195
#GC	0.017	0.047	***C.((**	0.017	0.151

(cont. on next page)

Table A.2 (cont.).

Feature	IG	MT	Feature	IG	MT
%GG/hpl	0.017	0.209	**#AAA**	0.017	0.144
c#U/sl	0.017	0.322	**%CGA**	0.016	0.167
#U++#C/sl	0.016	0.209	**#C++#U/sl**	0.016	0.181
c#Us/sl	0.016	0.277	**%AGC**	0.016	0.180
#UAG	0.016	0.155	**#G((.**	0.016	0.041
c#Us	0.016	0.118	**st(G-C)/sl**	0.016	0.227
%UU/sl	0.016	0.235	**#A/sl**	0.016	0.164
#A++#A/sl	0.016	0.172	**%CAG**	0.016	0.168
#AA	0.015	0.022	**%GG/sl**	0.015	0.239
%A	0.015	0.051	**#A/hpl**	0.015	0.182
%A++%A	0.015	0.091	**#A++#A/hpl**	0.015	0.158
%AUA	0.015	0.193	**mscs**	0.015	0.107
%GCG	0.015	0.183	**#AA/sl**	0.015	0.198
%U++%U/hpl	0.015	0.238	***C.((/hpl**	0.015	0.238
%U/hpl	0.014	0.250	**%U++%G/sl**	0.014	0.252
%G++%U/sl	0.014	0.234	**#AA/hpl**	0.014	0.174
%AA	0.014	0.093	**dscs/hpl**	0.014	0.220
***C.((/sl**	0.014	0.252	**c#A/hpl**	0.014	0.242
%CGC	0.014	0.167	**%UU/hpl**	0.014	0.221
%CC/sl	0.014	0.219	**%AAU**	0.014	0.168
st(G-C)	0.014	0.148	**#AU/sl**	0.014	0.220
#CGA	0.014	0.174	**dscs/sl**	0.014	0.241
#AAU	0.014	0.164	**mscs/sl**	0.014	0.255
st(G-C)/hpl	0.013	0.223	**#AUA**	0.013	0.139
%CC/hpl	0.013	0.214	**mscs/hpl**	0.013	0.162
%UCG	0.013	0.182	**c#A/sl**	0.013	0.264
%AUU	0.013	0.180	**c#A**	0.012	0.094
saln	0.012	0.108	**hpl/sl**	0.012	0.242
sl/hpl	0.012	0.168	**hll/sl**	0.012	0.244
hll/hpl	0.012	0.223	**#AU**	0.012	0.020
#CC/hpl	0.012	0.184	**%GCU**	0.012	0.187

(cont. on next page)

Table A.2 (cont.).

Feature	IG	MT	Feature	IG	MT
%CC	0.012	0.063	**#AUU**	0.012	0.148
ir	0.012	0.232	**c#As/sl**	0.012	0.309
#U++#G	0.012	0.051	**#G++#U**	0.012	0.102
#UGU	0.012	0.171	**#C++#C**	0.012	0.050
#C	0.012	0.018	**#U++#G/sl**	0.012	0.242
#G++#U/sl	0.012	0.198	**#CGC**	0.012	0.170
%CAC	0.011	0.169	**c#As/hpl**	0.011	0.251
%GG	0.011	0.145	**c#As**	0.011	0.121
%UGU	0.011	0.162	**%AA/sl**	0.011	0.219
#GGC	0.011	0.169	**%U++%C/sl**	0.011	0.250
%C++%U/sl	0.011	0.250	**#GUA**	0.011	0.171
#GG/hpl	0.011	0.178	**#GCG**	0.011	0.170
#GG/sl	0.011	0.201	**%U++%G/hpl**	0.011	0.231
%G++%U/hpl	0.011	0.266	**%U++%C/hpl**	0.011	0.238
%C++%U/hpl	0.011	0.218	**#G++#A**	0.011	0.028
#A++#G	0.011	0.080	**#C.((**	0.010	0.034
%GUC	0.010	0.191	**hll**	0.010	0.080
#C(.(/hpl	0.010	0.192	**%C++%A/hpl**	0.010	0.217
%A++%C/hpl	0.010	0.209	**%ACU**	0.010	0.174
%GUA	0.010	0.175	**%C++%A/sl**	0.010	0.230
%A++%C/sl	0.010	0.224	**#GCC**	0.010	0.161
c#Cs/hpl	0.010	0.259	**%UGA**	0.009	0.198
%CGU	0.009	0.172	**#CC/sl**	0.009	0.224
#nisl_h/hpl	0.009	0.173	**%CA/hpl**	0.009	0.296
%CA	0.009	0.064	**#CA/hpl**	0.009	0.185
#UCG	0.009	0.158	**%A++%A/sl**	0.009	0.266
#CCA	0.009	0.164	**%A/sl**	0.009	0.213
#AGC	0.008	0.161	**%G++%A/sl**	0.008	0.253
%A++%G/sl	0.008	0.225	**%AU/hpl**	0.008	0.250
%ACG	0.008	0.175	**#nisl_h/sl**	0.008	0.266
%AGU	0.008	0.180	**#mnn**	0.008	0.112

(cont. on next page)

Table A.2 (cont.).

Feature	IG	MT	Feature	IG	MT
%G++%A/hpl	0.008	0.211	**%A++%G/hpl**	0.008	0.208
#ACU	0.008	0.148	**%AA/hpl**	0.008	0.227
#U++#C	0.008	0.083	**#C++#U**	0.008	0.078
GsC/hpl	0.008	0.275	**CsG/hpl**	0.008	0.258
%UUG	0.007	0.177	**mll**	0.007	0.104
#C++#A	0.007	0.012	**#A++#C**	0.007	0.028
UGsCA	0.007	0.230	**UGsAC**	0.007	0.212
GUsCA	0.007	0.190	**GUsAC**	0.007	0.191
CAsUG	0.007	0.186	**CAsGU**	0.007	0.180
ACsUG	0.007	0.176	**ACsGU**	0.007	0.163
%U++%G	0.007	0.116	**%G++%U**	0.007	0.236
%C++%A	0.007	0.131	**%A++%C**	0.007	0.211
#U++#G/hpl	0.007	0.199	**#G++#U/hpl**	0.007	0.193
#C++#A/hpl	0.007	0.172	**#A++#C/hpl**	0.007	0.288
%CA/sl	0.007	0.221	**GsC/sl**	0.007	0.277
CsG/sl	0.007	0.300	**%AGG**	0.007	0.181
#U((./hpl	0.007	0.187	**UGsCA/hpl**	0.007	0.216
UGsAC/hpl	0.007	0.219	**GUsCA/hpl**	0.007	0.273
GUsAC/hpl	0.007	0.275	**CAsUG/hpl**	0.007	0.264
CAsGU/hpl	0.007	0.266	**ACsUG/hpl**	0.007	0.275
ACsGU/hpl	0.007	0.240	**%A++%A/hpl**	0.007	0.251
#N.(./hpl	0.007	0.225	***N.(.**	0.007	0.145
#N.(./sl	0.007	0.202	**%AUG**	0.007	0.156
#GG	0.007	0.039	**%AAC**	0.007	0.172
#G.(./hpl	0.007	0.177	***N.(./sl**	0.007	0.265
c#N/sl	0.007	0.289	***N.(./hpl**	0.007	0.242
#GU/sl	0.007	0.249	**c#G/hpl**	0.007	0.250
UGsCA/sl	0.007	0.224	**UGsAC/sl**	0.007	0.222
GUsCA/sl	0.007	0.290	**GUsAC/sl**	0.007	0.282
CAsUG/sl	0.007	0.279	**CAsGU/sl**	0.007	0.348
ACsUG/sl	0.007	0.260	**ACsGU/sl**	0.007	0.261

(cont. on next page)

Table A.2 (cont.).

Feature	IG	MT	Feature	IG	MT
c#G/sl	0.007	0.270	**c#N/hpl**	0.007	0.249
GsC	0.007	0.149	**CsG**	0.007	0.144
#mdn	0.006	0.121	**%A/hpl**	0.006	0.379
#N.(.	0.006	0.075	**#AGU**	0.006	0.171
#ACG	0.006	0.148	**c#C/hpl**	0.006	0.244
#UAC	0.006	0.198	**#GU**	0.006	0.046
%GU	0.006	0.068	***G.(./hpl**	0.006	0.247
***C(.(/sl**	0.006	0.284	***G.(./sl**	0.006	0.260
#CGU	0.006	0.159	**#U((./sl**	0.006	0.199
***U((.**	0.006	0.115	***G.(.**	0.006	0.102
#G.(./sl	0.006	0.260	**#CA/sl**	0.006	0.184
#CC	0.006	0.027	**#nisl_h**	0.006	0.049
%AU/sl	0.005	0.231	**%UCC**	0.005	0.170
%UCU	0.005	0.175	**%UGG**	0.005	0.176
c#Gs/hpl	0.005	0.249	**c#Gs/sl**	0.005	0.320
%GUU	0.005	0.191	**#UCU**	0.005	0.173
%AUC	0.005	0.165	**#CAG**	0.005	0.158
%CUU	0.005	0.171	**#G.(.**	0.004	0.040
%GGG	0.004	0.185	**#C(.(/sl**	0.004	0.217
#GU/hpl	0.004	0.189	***C(.(**	0.004	0.104
#G(.(/hpl	0.004	0.208	**#GUU**	0.004	0.169
c#C/sl	0.004	0.315	**%GAU**	0.004	0.168
c#Ns/sl	0.004	0.264	***G(.(/sl**	0.004	0.244
#C(.(	0.004	0.041	**%GGA**	0.004	0.169
%GCA	0.004	0.169	**#CA**	0.004	0.036
***C(.(/hpl**	0.004	0.236	**#GCU**	0.004	0.177
***G(.(**	0.004	0.102	**ir/sl**	0.004	0.307
#C.(./hpl	0.004	0.183	***U((./sl**	0.004	0.322
***U((./hpl**	0.004	0.243	**#G(.(/sl**	0.004	0.201
***G(.(/hpl**	0.004	0.248	**#U(.(**	0.003	0.056
%CUG	0.003	0.176	**c#Cs/sl**	0.003	0.272

(cont. on next page)

Table A.2 (cont.).

Feature	IG	MT	Feature	IG	MT
%CU/sl	0.003	0.236	***C.(./hpl**	0.003	0.251
***C.(.**	0.003	0.107	**#C.(./sl**	0.003	0.207
c#N	0.003	0.508	**%ACA**	0.003	0.171
#U.((/hpl	0.003	0.189	***C.(./sl**	0.003	0.248
%CAU	0.003	0.199	**%GU/sl**	0.003	0.231
#U.(./hpl	0.003	0.241	***U.((/hpl**	0.003	0.277
#UGG	0.003	0.175	**#C.(.**	0.003	0.042
%AGA	0.003	0.180	**c#Ns/hpl**	0.002	0.251
#mnn/sl	0.002	0.209	**#UUG**	0.002	0.166
%CU/hpl	0.002	0.232	***U.(.**	0.002	0.116
#U.(./sl	0.002	0.305	***U.((/sl**	0.002	0.313
#GA/hpl	0.002	0.192	**#ACA**	0.002	0.154
***U.(./sl**	0.002	0.280	***U.((**	0.002	0.114
%GA	0.002	0.073	**#U.((/sl**	0.002	0.197
#G(.(	0.002	0.043	**%GU/hpl**	0.002	0.221
#U(.(/sl	0.002	0.204	***U(.(**	0.002	0.111
#CU	0.002	0.029	**#CU/sl**	0.002	0.195
#U(.(/hpl	0.002	0.196	**%UGC**	0.002	0.170
%UC	0.002	0.071	**#UC/hpl**	0.002	0.196
#mdn/hpl	0.002	0.201	**#UCC**	0.002	0.172
%GAC	0.002	0.169	**c#Ns**	0.002	0.534
c#G	0.002	0.097	**#GUG**	0.002	0.171
%GUG	0.002	0.176	**#mdn/sl**	0.002	0.246
ir/hpl	0.002	0.155	**dr/hpl**	0.002	0.226
***U.(./hpl**	0.002	0.273	**st(G-U)/sl**	0.000	0.236
st(G-U)/hpl	0.000	0.220	**st(G-U)**	0.000	0.150
dscs	0.000	0.117	**dr/sl**	0.000	0.237
c#Gs	0.000	0.115	**c#Cs**	0.000	0.118
c#C	0.000	0.093	***U(.(/sl**	0.000	0.252
***U(.(/hpl**	0.000	0.290	***A.(./sl**	0.000	0.235
***A.(./hpl**	0.000	0.233	***A.(.**	0.000	0.098

(cont. on next page)

Table A.2 (cont.).

Feature	IG	MT	Feature	IG	MT
***A.((/sl**	0.000	0.237	***A.((/hpl**	0.000	0.230
***A.((**	0.000	0.095	***A(.(/sl**	0.000	0.241
***A(.(/hpl**	0.000	0.260	***A(.(**	0.000	0.128
***A((./sl**	0.000	0.235	***A((./hpl**	0.000	0.234
***A((.**	0.000	0.102	**%UUC**	0.000	0.185
%UG/sl	0.000	0.266	**%UG/hpl**	0.000	0.253
%UG	0.000	0.069	**%UCA**	0.000	0.178
%UC/sl	0.000	0.245	**%UC/hpl**	0.000	0.228
%GGU	0.000	0.173	**%GAG**	0.000	0.179
%GAA	0.000	0.176	**%GA/sl**	0.000	0.240
%GA/hpl	0.000	0.218	**%CUC**	0.000	0.185
%CU	0.000	0.069	**%CCU**	0.000	0.171
%CCC	0.000	0.165	**%CAA**	0.000	0.177
%AG/sl	0.000	0.223	**%AG/hpl**	0.000	0.249
%AG	0.000	0.057	**%ACC**	0.000	0.160
%AC/sl	0.000	0.286	**%AC/hpl**	0.000	0.211
%AC	0.000	0.056	**%AAG**	0.000	0.187
#mnn/hpl	0.000	0.197	**#UUC**	0.000	0.177
#UGC	0.000	0.157	**#UGA**	0.000	0.196
#UG/sl	0.000	0.255	**#UG/hpl**	0.000	0.194
#UG	0.000	0.053	**#UCA**	0.000	0.173
#UC/sl	0.000	0.222	**#UC**	0.000	0.061
#U.(.	0.000	0.056	**#U.((**	0.000	0.059
#U((.	0.000	0.056	**#GUC**	0.000	0.176
#GGU	0.000	0.164	**#GGG**	0.000	0.163
#GGA	0.000	0.188	**#GCA**	0.000	0.174
#GAU	0.000	0.181	**#GAG**	0.000	0.170
#GAC	0.000	0.185	**#GAA**	0.000	0.161
#GA/sl	0.000	0.212	**#GA**	0.000	0.044
#CUU	0.000	0.175	**#CUG**	0.000	0.190
#CUC	0.000	0.163	**#CU/hpl**	0.000	0.217

(cont. on next page)

Table A.2 (cont.).

Feature	IG	MT	Feature	IG	MT
#CCU	0.000	0.156	**#CCC**	0.000	0.165
#CAU	0.000	0.152	**#CAC**	0.000	0.163
#CAA	0.000	0.163	**#C++#A/sl**	0.000	0.203
#AUG	0.000	0.179	**#AUC**	0.000	0.161
#AGG	0.000	0.160	**#AGA**	0.000	0.173
#AG/sl	0.000	0.182	**#AG/hpl**	0.000	0.207
#AG	0.000	0.022	**#ACC**	0.000	0.169
#AC/sl	0.000	0.187	**#AC/hpl**	0.000	0.160
#AC	0.000	0.021	**#AAG**	0.000	0.149
#AAC	0.000	0.149	**#A.(./sl**	0.000	0.183
#A.(./hpl	0.000	0.167	**#A.(.**	0.000	0.022
#A.((/sl	0.000	0.176	**#A.((/hpl**	0.000	0.171
#A.((	0.000	0.025	**#A++#C/sl**	0.000	0.226
#A(.(/sl	0.000	0.174	**#A(.(/hpl**	0.000	0.154
#A(.(	0.000	0.024	**#A((./sl**	0.000	0.172
#A((./hpl	0.000	0.158	**#A((.**	0.000	0.025