ABSTRACT

IN SILICO PREDICTION OF NON-CODING RNAs USING SUPERVISED LEARNING AND FEATURE RANKING METHODS

by
Stephen J. Griesmer

This thesis presents a novel method, RNAMultifold, for development of a non-coding RNA (ncRNA) classification model based on features derived from folding the consensus sequence of multiple sequence alignments using different folding programs: RNAalifold, CentroidFold, and RSpredict. The method ranks these folding features according to a Class Separation Measure (CSM) that quantifies the ability of the features to differentiate between samples from positive and negative test sets. The set of top-ranked features is then used to construct classification models: Naive Bayes, Fisher Linear Discriminant, and Support Vector Machine (SVM). These models are compared to the performance of the same models with a baseline feature set and with an existing classification tool, RNAz.

The Support Vector Machine classification model with a radial basis function kernel, using the top 11 ranked features, is shown to be more sensitive than other models, including another ncRNA prediction program, RNAz, across all specificity values for the RNA families under study. In addition, the target feature set outperforms the baseline feature set of z score and structure conservation index across all classification methods, with the exception of Fisher Linear Discriminant.

The RNAMultifold method is then used to search the genome of a Trypanosome species (*Trypanosoma brucei*) for novel ncRNAs. The results of this search are compared with known ncRNAs and with results from RNAz.

IN SILICO PREDICTION OF NON-CODING RNAs USING SUPERVISED LEARNING AND FEATURE RANKING METHODS

by
Stephen J. Griesmer

A Thesis
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Bioinformatics

Department of Computer Science

January 2010

APPROVAL PAGE

IN SILICO PREDICTION OF NONCODING RNAs USING SUPERVISED LEARNING AND FEATURE RANKING METHODS

Stephen J. Griesmer

Dec. 1, 2009

Dr. Jason T. L. Wang, Thesis Advisor — Date
Professor of Computer Science, NJIT

12/1/09

Dr. Chengjun Liu, Committee Member — Date
Associate Professor of Computer Science, NJIT

12/2/2009

Dr. David Nassimi, Committee Member — Date
Associate Professor of Computer Science, NJIT

BIOGRAPHICAL SKETCH

Author: Stephen J. Griesmer

Degree: Masters of Bioinformatics

Date: January 2010

Undergraduate and Graduate Education:

- Master of Science in Bioinformatics,
  New Jersey Institute of Technology, Newark, NJ, 2010

- Master of Science in Interdisciplinary Studies, Dept. of Electrical Engineering,
  University of British Columbia, Vancouver, B.C., Canada, 1981

- Bachelor of Science and Engineering in Electrical Engineering and
  Computer Science, Princeton University, Princeton, NJ, 1979

Major: Bioinformatics

Presentations and Publications:

Aho, A., Bosik, B., and Griesmer, S. J. Protocol Testing and Verification Within AT&T. AT&T Technical Journal 1990, **69**(1): 4-6.

Griesmer, S. J., and Jesmajian, R. W. Evolution of Messaging Standards. AT&T Technical Journal 1994, **73**(3): 21-45

I dedicate this thesis to my father, James H. Griesmer, on his 80th birthday for his example to me as a computer scientist, mathematician, and lifelong learner.

ACKNOWLEDGMENT

I wish to thank my advisor, Dr. Jason Wang, for his guidance, support, and willingness to listen to my ideas. I wish to thank Dr. David Nassimi and Dr. Chengjun Liu for agreeing to be members of my thesis committee, and, for Dr. Nassimi, for his encouragement of my studies, and, for Dr. Liu, for introducing me to the mathematical foundations of machine learning.

My family has been very understanding of this personal quest, allowing their father and husband time to complete this work. My wife, Lynn, has been very forgiving of the home repairs undone and the weekend events not undertaken. My children, Peter and Liz, inspired me with their own independent scholarship. My daughter, Chrissy, kept me happy with her fun-loving perspective and piano playing.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

CSM	Class Separation Method
DNA	Deoxyribonucleic acid
EM	Expectation maximization
FLD	Fisher Linear Discriminant
HMM	Hidden Markov Model
MCC	Matthew's Correlation Coefficient
MFE	Minimum free energy
miRNA	MicroRNA
mRNA	Messenger RNA
MSA	Multiple sequence alignment
ncRNA	Non-coding ribonucleic acid
PCA	Principal Component Analysis
PPV	Positive Predictive Value
RBF	Radial Basis Function
RNA	Ribonucleic acid
ROC	Receiver operating curve
rRNA	Ribosomal RNA
SCFG	Stochastic Context-Free Grammar
SCI	Structure conservation index
Sn	Sensitivity
snoRNA	Small nucleolar RNA

snRNA	Small nuclear RNA
snRNP	Small nuclear ribonucleoprotein
Sp	Specificity
SRP	Signal Recognition Particle
SVM	Support Vector Machine
tRNA	Transfer RNA
XMFA	Extended Multi-Fasta file format

CHAPTER 1

INTRODUCTION

1.1 RNA

Ribonucleic acid (RNA) is composed of four nucleotides: guanine (G), cytosine (C), adenine (A), and uracil (U). RNA is single-stranded, but folds into a three-dimensional structure. Cells contain up to eight times as much RNA as DNA [Garrett and Grisham 1999], highlighting its importance. In addition, RNA is thought to be the precursor of deoxyribonucleic acid (DNA) for storage of genetic information in the earliest forms of life [Müller 2006].

RNA plays many different roles in the cell. RNA serves as a carrier of protein coding information from the nucleus to ribosomes, the sites of protein synthesis, in the form of messenger RNA (mRNA) and the transporter of amino acids to ribosomes in the form of transfer RNA (tRNA). Ribosomes are made of 65% RNA of the ribosomal type [Garrett and Grisham, 1999] called ribosomal RNA, or rRNA, which combine with ribosomal proteins to create ribosomes. Small nuclear RNAs, snRNAs, contain 100-200 nucleotides. They join with proteins in stable complexes called small nuclear ribonucleoprotein particles (snRNPs). snRNPs are important in the processing of eukaryotic gene transcripts into mature messenger mRNAs through the spliceosome. snRNAs also play a role in the regulation of transcription factors and in maintaining telomeres. Small Nucleolar RNA (snoRNAs) are located in the nucleolus. They make chemical modifications to rRNA and other RNA genes (e.g., through methylation). MicroRNAs (miRNAs) are short ncRNAs of 20-23 nucleotides in length. MicroRNAs anneal to mRNA to inhibit protein translation.

1.2 Non-coding RNA

Non-coding RNAs (ncRNAs) are types of RNA where transcripts are not translated to proteins. The structure of ncRNAs is governed by *cis* base pairing that determines secondary structure. The size of ncRNAs varies from 20-1000's of nucleotides in length. Tens of thousands of ncRNAs are expressed in human cells. There are more ncRNAs expressed in human cells than protein-coding RNAs. It is estimated that 3% of human and mouse genomes encompasses *cis*-regulatory sequence, signals for transcription initiation, termination and RNA processing [Babak et al. 2007]. ncRNAs can target another nucleic acid with precision and complementary RNAs may evolve more easily than an amino acid protein [Eddy 2002].

1.3 Prediction of Non-coding RNA

The detection and prediction of ncRNAs is difficult. Unlike protein-coding genes, ncRNAs lack statistical signals for reliable detection from primary sequences. In addition, there is no protein product for which ncRNAs are coding. Therefore, there are no evolutionary constraints on the protein product. The constraints on ncRNAs come from the pressure to maintain a secondary structure. The secondary structure of the ncRNA determines its function. This structure can be conserved even with substantial changes to the primary DNA sequence that codes for it.

The measurement of detection and classification by different methods is measured with sensitivity, specificity, and Matthew's Correlation Coefficient. Sensitivity (Sn) and Specificity (Sp) are defined as:

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

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

where TP is true positives, TN is true negatives, FP is false positives, and FN is false negatives. Matthew's Correlation Coefficient combines sensitivity and specificity into a single measure:

$$MCC = \frac{(TP * TN) - (FN * FP)}{\sqrt{(TP + FN) * (TN + FP) * (TP + FP) * (TN + FN)}}$$

MCC ranges in value from -1 to +1, where +1 indicates perfect prediction and -1 indicates totally incorrect prediction.

1.4 Thesis Outline

The remainder of this thesis describes the prediction of ncRNAs. Chapter 2 summarizes related work; Chapter 3 describes the approach. Chapter 4 discusses the extraction of features from RNA multiple sequence alignments and the ranking of those features. Chapter 5 discusses classification methods used in this thesis. Chapter 6 gives the results of the classification methods. Chapter 7 shows the results of using the classification tools to discover novel and known ncRNAs in the *Trypanosoma brucei* genome. Chapter 8 discusses these results and Chapter 9 provides a conclusion.

CHAPTER 2

SUMMARY OF RELATED WORK

Various methods have been considered for the prediction of ncRNA. The following sections outline examples from the literature: QRNA, MSARI, RNAz, Dynalign, ddbRNA, and Evofold.

2.1 QRNA

QRNA [Rivas and Eddy 2001] uses Hidden Markov Models (HMMs) and Stochastic Context Free Grammars (SCFGs) to produce three evolutionary models for different regions of DNA: structural RNA, protein coding, or other. The models and grammars are paired in that they examine (or emit) two aligned sequences to determine the probability of the different models. These sequences are meant to be evolutionarily related and the emission probabilities of the program are based on this assumption.

The structural RNA model assumes a pattern of mutations between sequences that conserves secondary structure based on the expectation of an abundance of pairwise-correlated mutations that preserve Watson-Crick pairing. The underlying model to determine the probability of a structure given that it is structural RNA is based on a Stochastic Context-Free Grammar that incorporates probabilities of structural elements like hairpins, internal loops, and bulges.

The protein-coding model assumes that aligned sequences encode homologous proteins based on the expectation of conservative amino acid substitutions and

synonymous mutations. The model for protein coding uses a pair-Hidden Markov Model which emits a codon, or three nucleotides, at once for two aligned sequences.

The "other" model is a "catch-all" model that assumes that there is no structural or sequence conservation and that mutations are position-independent. This model uses a pair-Hidden Markov Model that emits a pair of nucleotides at a time. The model must also handle the insertion and deletion of nucleotides in the sequences and boundary conditions that arise because of the beginnings and ends of any alignment. In order to classify an input alignment as structural RNA, protein-coding, or other, QRNA calculates the Bayesian posterior probability, under the assumption that the three models are equally likely.

2.2 ddbRNA

ddbRNA [di Bernardo et al. 2003] is a program to detect conserved secondary structures in multiple sequence alignments. It does this by searching for compensatory mutations in the aligned sequences that conserve pairing alignment in stems (e.g., a GC pair is conserved if it is transformed to an AT, TA, or GC pair). The algorithm used by ddbRNA counts the compensatory mutations for every possible stem-loop and compares it to the average obtained with the aligned sequences have their columns shuffled. Gaps are not included in the analysis; all bases that align to a gap are removed. Di Bernardo et al. 2003 showed that ddbRNA detected 5s rRNA genes while QRNA did not. ddbRNA sensitivity drops with increasing pairwise identity percentages; whereas, with QRNA, sensitivity increased with increasing pairwise identity until ~90% pairwise identity, where it drops off.

2.3 MSARI

MSARI [Coventry et al. 2004] computes the statistical significance of short, contiguous base-paired regions of potential structural RNAs in multiple sequence alignments. It copes with a wide range of evolutionary distances within the MSA by varying the null hypothesis model from sequence to sequence. MSARI uses RNAfold as a preprocessor to locate probable base pairs. For each pair of positions in a base pair, MSARI examines windows of length 7 for complementary mutations. It can also tolerate misalignments of orthologous base pairs up to two nucleotides.

A Bonferroni-style test for rejection of the null hypothesis is used. Significant base pairings are sorted by significance. Pseudoknots are eliminated. The probabilities of the selected pairs are multiplied together to determine the score of the aligned sequence. With 10- and 15-sequence alignments, its results, in terms of sensitivity and specificity, exceeded QRNA and ddbRNA.

2.4 RNAz

RNAz [Washietl et al. 2005] combines sequence alignment of 2-4 sequences with measures of secondary structure conservation and thermodynamic stability to predict non-coding RNA. RNAz builds on other RNA programs to accomplish goal:

- RNAfold – folds single sequences;
- RNAalifold – finds consensus folding of aligned sequences; and
- Libsvm – provides a support vector machine tool.

Thermodynamic stability is measured by minimum free energy. RNAz compares the minimum free energy of a given sequence to random sequences of the same length and base composition. These are combined into a z-score, defined as:

$$z = \frac{m - \mu}{\sigma}$$

where μ and σ are the mean and standard deviations of the random sequences, respectively. Negative z scores indicate that a sequence is more stable than expected by chance.

Structural conservation is measured by a structure conservation index (SCI). SCI compares the minimum free energy of the consensus sequence, as calculated by RNAalifold, to average minimum free energy of the individual sequences. An SCI score nearer to 1 indicates that there is structural stability across the sequences. Lower SCI scores indicate that there is less commonality among sequence structures.

In RNAz, the z-score and SCI are combined in a support vector machine (SVM) with a radial basis function kernel learning algorithm. The support vector machine classifies an alignment as "structural non-coding RNA" or "other." The relationship between these classes is non-linear. The SVM was trained on a large set of cross-species non-coding RNA sequences from the Rfam [Griffiths-Jones et al. 2003] database.

2.5 Dynalign

Dynalign [Uzilov et al. 2006] is a dynamic programming algorithm for computing the lowest free energy secondary structure and structural alignment for a pair of sequences. It is used to predict ncRNA sequences. It simultaneously optimizes the common secondary structure between the pair and aligns the structure. Dynalign minimizes the total free folding energy changes due to folding by computing:

$$\Delta G^{\circ}_{total} = \Delta G^{\circ}_{1} + \Delta G^{\circ}_{2} + number\ of\ gaps\ in\ alignment * \Delta G^{\circ}_{gap\ penalty}$$

where

ΔG°_{total} is the total free folding energy change of the folded secondary structure,

ΔG°_{i} is the free folding energy change of the folded secondary structure of sequence i, where i is 1 or 2,

$\Delta G^{\circ}_{gap\ penalty}$ is the energy penalty applied for each gap in the aligned structures, determined empirically by maximizing structure prediction accuracy.

A Support Vector Machine model was developed to classify ncRNA sequences. Its inputs are ΔG°_{total}; length of the shorter sequence; and the frequencies of A, U, and C nucleotides. It was tested against RNAz and found to be more sensitive below 99.2% specificity, but less sensitive above. It is also more sensitive for sequence pairs with less than 50% pairwise identity.

2.6 Evofold

Similarly to QRNA, EvoFold [Pedersen et al. 2006] uses phylogenetic Stochastic Context Free Grammars (phylo-SCFGs) to create a probabilistic model of RNA secondary structure and sequence evolution to predict non-coding RNAs. The inputs to EvoFold are a multiple sequence alignment (MSA) and a phylogenetic tree. EvoFold takes these inputs and computes probabilities of two different phylo-SCFGs: a model for functional RNAs and a model for background sequences, the null hypothesis. The model for functional RNAs is composed of two types of phylogenetic submodels: a dinucleotide submodel for base pairs, and a single nucleotide model for unpaired bases. Each of these models contains a substitution process to examine the multiple sequence alignment. The substitution process for the dinucleotide favors compensatory mutations among base

pairs; the substitution process for the single nucleotide model is based the evolutionary process for unpaired nucleotides.

The output of Evofold includes the functional RNA detected and the log likelihood of the MSA to contain a functional RNA:

$$fps = \log \frac{\mathrm{P(MSA|\phi_{fRNA})}}{\mathrm{P(MSA|\phi_{background})}}$$

where MSA is the alignment, ϕ_{fRNA} is the model for functional RNAs, and $\phi_{background}$ is the model for background sequences.

The EvoFold model is trained on data from the Rfam database. Human entries from Rfam are aligned to the human genome using BLAT. Human-mouse syntenic matches that overlap with these human entries are selected. Aligned sequences with poor secondary structures are removed. The alignments are expanded to six additional sequences—chimp, rat, dog, chicken, fugu, and zebra fish. Alignments that result in less than four sequences were disgarded. Parameters for the phylo-SCFG are found using the EM algorithm and the quasi-Newton method.

CHAPTER 3

APPROACH

The approach taken in this paper is to follow the methodology of RNAz. Sequences from six families were downloaded from the Rfam database [Griffiths-Jones et al. 2003] and the SRP RNA database [Anderson et al. 2006]:

- Hammerhead ribozyme III (RF00008)
- Group II catalytic intron (RF00029)
- U5 Spliceosomal RNA (RF00020)
- tRNA (RF00005)
- 5S rRNA (RF00001)
- SRP RNA

Within each family, clusters were formed using blastclust. 60% similarity was required for each cluster. Clusters were eliminated with only one sequence. Multiple sequence alignments were then formed from each cluster of 2, 3, and 4 sequences. 65-85% similarity was required for the multiple sequence alignments. Contributions from a cluster were capped at 2000 MSAs. From this, four families of RNAs had sufficient numbers of MSAs with the targeted similarity:

- Group II catalytic intron (RF00029)
- U5 Spliceosomal RNA (RF00020)
- 5S rRNA (RF00001)
- SRP RNA

A set of around 1800 MSAs per family was chosen for the test set— about 600 each from MSAs with 2, 3, and 4 sequences in each family.

The negative test set was developed from the positive MSAs. The RNAz program `rnazRandomizeAln.pl` was used to shuffle the columns in the alignments. The shuffling preserves length, base composition, gap pattern, and local conservation patterns. There was one negative MSA for each positive one.

The RNAMultifold method is based on features extracted from the MSAs. The features include energy and structural parameters. The features chosen to include in the model are ranked according to their ability to differentiate between positive and negative classes through a Class Separation Measure (CSM). Features with the highest CSM are included in the prediction model. The model was then tested against a baseline of prediction using z score and SCI, the parameters used for classification by RNAz.

Finally, RNAMultifold is used to find ncRNAs in multiple sequence alignments from Trypanosome genome sequences. These ncRNAs are then compared with known ncRNAs in *Trypanosoma brucei* and existing annotations of this genome. These results are then compared to the discovery power of RNAz.

CHAPTER 4

FEATURE EXTRACTION AND RANKING

4.1 Feature Extraction

For the RNAMultifold feature pool, features were examined from: (1) RNAz, (2) three folding algorithms (RNAalifold [Hofacker 2007], CentroidFold [Hamada et al. 2009], and RSpredict [Spirollari et al. 2009]), and (3) features suggested in the literature.

From RNAz, z score, structure conservation index (SCI), average minimum free energy (MFE_{avg}), and consensus minimum free energy ($MFE_{consensus}$) were used. Z score is a measure of thermodynamic stability. RNAz compares the minimum free energy of a given MSA to the average minimum free energy of set of shuffled MSAs of the same length and base composition. For this paper, 100 shuffled MSAs were computed for this average. The z score is defined as:

$$z = \frac{m - \mu}{\sigma}$$

where m is the consensus minimum free energy of the MSA, μ is the average minimum free energy of the shuffled MSAs, and σ is the standard deviation of the minimum free energies of the shuffled MSAs. Z score is the number of standard deviations by which the MFE of the MSA deviates from the MFE of a set of shuffled sequences. Negative z scores indicate that an MSA is more stable than expected by chance.

Structure conservation index (SCI) measures structural conservation. SCI is the ratio of minimum free energy of the consensus secondary structure of the MSA

($MFE_{consensus}$), determined by RNAalifold, to the average minimum free energy (MFE_{avg}) for each sequence in the MSA alone. An SCI score near 1 indicates that there is structural stability across the sequences in the MSA. Lower SCI scores indicate that there is less commonality among sequence structures. SCI > 1 indicates a perfectly conserved secondary structure supported by compensatory mutations [Washietl et al. 2005]. MFE_{avg} and $MFE_{consensus}$ are also retained in the feature pool as potential model parameters.

The next set of features is derived from three different folding programs: RNAalifold, CentroidFold, and RSpredict. RNAalifold implements the Zuker-Steigler algorithm for computing minimal free energy (MFE) structures assuming a nearest-neighbor model. It uses empirical estimates of thermodynamic parameters for neighboring interactions and loop entropies to score structures [Gardner and Giegerich 2004].

The MFE structure is a maximum likelihood estimator that provides the highest probability solution over the probabilistic distribution of all solutions [McCaskill 1990]. However, the MFE structure generally has low probability and a number of other folding alternatives may have nearly the same probability. These alternatives may differ in the number of base pairs, which then impacts the minimum free energy and features such as SCI. Because of this, a better estimator of secondary structure may be one that maximizes the expectations of an objective function related to the accuracy of the prediction.

CentroidFold provides an averaged gamma-centroid estimator for common secondary structure prediction. For MSAs, the estimator maximizes the expected value of:

$$\alpha_1 TP+\alpha_2 TN-\alpha_3 FP-\alpha_4 FN$$

where TP is the number of true positive base pairs, TN is the number of true negative base pairs, FP is the number of false positive base pairs, FN is the number of true negative base pairs, and α_i are arbitrary constants.

RSpredict computes the RNA structure with minimum normalized energy, free energy of substructures divided by the length of the substructures. The RNA structure is divided into substructures using loop decomposition with a nearest neighbor energy model. RSpredict accepts an MSA as input and predicts the consensus secondary structure by minimizing pseudo-energy. Pseudo-energy includes both normalized free energy and covariance scores of the MSA sequences. RSpredict outperformed RNAalifold [Hofacker 2007], Pfold [Knudsen and Hein 2003], and KNetFold [Bindewald and Shapiro 2006] in alignment accuracy as measured by MCC.

For each of the folding algorithms, common features were chosen to characterize the predicted consensus secondary structure: number of base pairs, number of loops, number of bulges, hairpin length, and longest (maximum) consecutive base pairs. In addition, the normalized value of each feature (feature value divided by multiple sequence alignment length) was also computed. These features follow the work of Yousef et al. 2006 and Hertel and Stadler 2006.

Other features that were included in the feature pool were energy density from RSPredict, Shannon entropy, and base-pair distance. Energy density represents the normalized energy of the folded structure resulting from the algorithm. Shannon entropy

and base-pair distance were suggested as differentiating features in Freyhult et al. 2005. Both measure properties of the ensemble of structures that may exist for a given RNA sequence *in vivo*. Shannon entropy is defined as:

$$Q(\boldsymbol{x}) = \frac{-\sum_{i<j} p_{ij} \log_2 p_{ij}}{L}$$

where x is the consensus sequence, p_{ij} is the base-pair probability (the probability that base x_i pairs with x_j) and L is the length of the sequence. Average base pair distance is defined as:

$$D(\boldsymbol{x}) = \frac{\sum_{i<j}(p_{ij} - p_{ij}^2)}{L}$$

Shannon entropy and average base-pair distance are highly correlated. Both are included in the feature pool.

4.2 Feature Ranking

The key to feature ranking is selection of the features with the greatest potential for discriminating between classes. One method of feature ranking is Principal Component Analysis (PCA) [Duda et al. 2001]. PCA seeks a multidimensional projection that best represents the data in a least-squares sense. The idea is to represent n-dimensional samples $x_1,\ldots,x_n$ by a single vector $\mathbf{x}_0$. PCA finds the vector $\mathbf{x}_0$ such that the sum of the squared distances between $\mathbf{x}_0$ and various $\mathbf{x}_k$ is as small as possible. With PCA, the squared-error criterion function is defined as:

$$J_0(x_0) = \sum_{k=1}^{n} ||x_0 - x_k||^2$$

The objective is to find the value of $\mathbf{x}_0$ that minimizes J_0. The solution $\mathbf{x}_0 = \mathbf{m}$, where m is the sample mean, minimizes J_0. The sample mean is a zero-dimensional representation of the data, but does not give any of the variability in the data. A more interesting one-dimensional representation projects the data onto a line running through the sample mean,

$$\mathbf{x} = \mathbf{m} + a\mathbf{e},$$

where **e** is a unit vector in the direction of the line and a is a scalar that measures the distance of **x** from **m**. By representing $\mathbf{x}_k$ by $\mathbf{m} + a_k\mathbf{e}$, the optimal set of coefficients a_k can be found by minimizing squared-error criterion function:

$$J_1(a_1, \ldots, a_n, \boldsymbol{e}) = \sum_{k=1}^{n} \|(\boldsymbol{m} + a_k\boldsymbol{e}) - \boldsymbol{x}_k\|^2$$

The solution, minimizing J_1 by differentiating by a_k and setting J_1 to zero, is

$$a_k = \mathbf{e}^t(\mathbf{x}_k - \mathbf{m})$$

This equation means that the least-squares solution projects the vector $\mathbf{x}_k$ onto a line in the direction of **e** that passes through the mean **m**.

To find the best direction **e** for the line, the scatter matrix is used:

$$\boldsymbol{S} = \sum_{k=1}^{n} (\boldsymbol{x}_k - \boldsymbol{m})(\boldsymbol{x}_k - \boldsymbol{m})^t$$

The scatter matrix is (n-1) times the sample covariance matrix. Using the scatter matrix and the previous solution for a_k,

$$J_1(\boldsymbol{e}) = -\boldsymbol{e}^t\boldsymbol{S}\boldsymbol{e} + \sum_{k=1}^{n} \|\boldsymbol{x}_k - \boldsymbol{m}\|^2$$

Lagrangian multipliers maximize $\mathbf{e}^t\mathbf{S}\mathbf{e}$ subject to constraint $\|\mathbf{e}\| = 1$:

$$u = e^t\boldsymbol{S}\boldsymbol{e} - \lambda e^t\boldsymbol{e}$$

Differentiating u with respect to e and setting to 0 leads to:

$$\frac{\partial u}{\partial \boldsymbol{e}} = 2\boldsymbol{S}\boldsymbol{e} - 2\lambda\boldsymbol{e} = \mathbf{0}$$

$$\mathbf{S}\mathbf{e} = \lambda\mathbf{e}$$

The vector e must be an eigenvector of the scatter matrix S. To maximize $\mathbf{e}^t\mathbf{S}\mathbf{e}$, the eigenvector corresponding to the largest eigenvalue of the scatter matrix is selected. Generalizing to a multidimensional projection, choose the eigenvectors corresponding to the largest eigenvalues. The eigenvectors form the principal components of **x**.

Principal component analysis reduces the dimensionality of feature space by restricting attention to those directions along which the scatter of the hyperellipsoidally shaped cloud is greatest.

PCA finds components useful for representing data, but there is no reason to assume that the components are useful for discriminating between data in different classes [Duda et al. 2001]. PCA seeks directions that are efficient for representation; discriminant analysis seeks directions that are efficient for discrimination. For discriminant analysis, a transformation $\mathbf{w}$: $\mathbf{x}$ -> $\mathbf{y}$ where $\mathbf{x} \in \mathcal{R}^d$, $\mathbf{y} \in \mathcal{R}^m$ that optimizes a separability criteria in y-space is sought. To find this transformation, Fisher Linear Discriminant Analysis may be used.

For Fisher Linear Discriminant Analysis, the criterion function to maximize is:

$$J(\mathbf{w}) = \mathbf{w}^t\mathbf{S}_B\mathbf{w} \,/\, \mathbf{w}^t\mathbf{S}_W\mathbf{w}$$

where $\mathbf{S}_B$ is the between-class scatter matrix and $\mathbf{S}_W$ is the within-class scatter matrix. The solution for $\mathbf{w}$ that maximizes $J(\mathbf{w})$ is:

$$\mathbf{w} = (\mathbf{S}_W)^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

For a two-class problem:

$$\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t,$$

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2$$

where $\mathbf{S}_i = \sum_{\mathbf{x}\in Ci} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t$. The optimal decision boundary has the equation:

$$\mathbf{w}^t\mathbf{x} + w_0 = 0$$

where $\mathbf{w} = \sum^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and where w_0 is a constant, $\sum$ is the covariance matrix and $\boldsymbol{\mu}_i$ are the sample means. Fisher Linear Discriminant Analysis can be used for linear classification of samples.

The Class Separation Measure (CSM) [Sheet et al., prepub] is similar to Fisher Linear Discriminant Analysis, but is used for selection and ranking of candidate features prior to classification. CSM uses the same criterion function:

$$J(\mathbf{w}) = |\, m_1 - m_2 \,| \,/\, (s_1)^2 + (s_2)^2$$

where m_i is the projected mean for class i, and s_i represents the standard deviation of the projected data for class i. Following the same idea, the CSM for feature k is defined as:

$$CSM^k = \frac{|\, \mu_1^k - \mu_2^k \,|}{(\sigma_1^k)^2 + (\sigma_2^k)^2}$$

where $\mu_1^k$is the mean of the positive class for feature k, $\mu_2^k$is the mean of the negative class for feature k, $\sigma_1^k$is the standard deviation of the positive class for feature k, and σ_2^k is the standard deviation of the negative class for feature k. The CSM score indicates the extent to which feature k separates its positive and negative class.

Once the CSM score has been calculated for all features, it is then sorted from the largest to smallest value to create a ranked feature vector, $\mathbf{X}_{ranked}$. The top ranked features can then be used in a classification model for prediction of positive or negative classes. The method of classification need not be based on Fisher Linear Discriminant function.

One problem with this approach is that it doesn't consider the correlation of features. Its downside is that discriminating features that are highly correlated can be included in the feature set. Thus, it may be possible to reduce the feature set by eliminating a feature that is highly correlated with another. However, the Class

Separation Measure provides a useful starting point for features selection that can then be pruned to a smaller set by considering feature correlation.

CHAPTER 5

CLASSIFICATION METHODS FOR ncRNA PREDICTION

For different feature sets, this thesis examines different classification methods for prediction: Naïve Bayes, Fisher Linear Discriminant, and Support Vector Machines with different kernel functions. It compares the sensitivity, specificity, and Matthew's Correlation Coefficient of these methods for a given feature set.

5.1 Naïve Bayes Classifier

The Naïve Bayes classifier [Tan et al. 2005] estimates the class-conditional probability by assuming that the attributes are conditionally independent:

$$P(X|Y = y) = \prod_{i=1}^{d} P(X_i|Y = y)$$

where X is the attribute set, Y is the class (positive and negative for the two-class problem) and d is the number of features in the model. The probability of the attribute set given the class, $P(X \mid Y=y)$, for this paper is determined by the training data sets for positive and negative classes. The method assumes that the feature distribution is normal, though this is not the case for parameters in this model (e.g., z score) [Washietl 2006]. The Naïve Bayes classifier computes the posterior probability for each class ω_i:

$$p(\omega_i|\mathbf{x}) = P(\omega_i)\ \Pi_{j=1,d}\ p(x_j|\omega_i)\ /\ P(\mathbf{x})$$

where $p(\omega_i|\mathbf{x})$ is the posterior probability for class ω_i, $p(x_j|\omega_i)$ is the class-conditional probability of feature j, $P(\omega_i)$ is the prior probability for class ω_i, and $P(\mathbf{x})$ is the prior probability of **x**.

Since $P(\mathbf{x})$ is fixed for all ω, it is sufficient to choose the class that maximizes the numerator. Thus, the decision rule for the classifier is:

Choose ω_1 if $p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$; otherwise, choose ω_2.

The advantage of the Naïve Bayes classifier is that it is simple to compute, it is robust to irrelevant features, and it is robust to noise [Tan et al. 2005]. However, its disadvantages are that it assumes that the probability distribution of each feature in a feature set is normal and that the features are independent of one another. Correlated features can degrade classifier performance. For ncRNA classification, these assumptions are generally not true.

5.2 Fisher Linear Discriminant

The Fisher Linear Discriminant function was outlined above as a feature ranking method by projection of a multidimensional space into one dimension. This projection can be used to construct a classifier such that, if $y \geq y_0$, a sample is classified as class ω_1, or otherwise class ω_2.

5.3 Support Vector Machines

Binary classifiers that separate data into two separate classes, positive and negative, which is the objective of this thesis. Support Vector Machines (SVM) provide a means

of classifying data into different classes or categories. The goal of SVMs is to find a hyperplane with the maximum margin that separates two classes of data. Figure 5.1 illustrates this goal. By choosing this classification criterion, false positives can be minimized and the impact of changes in the underlying model can be reduced.

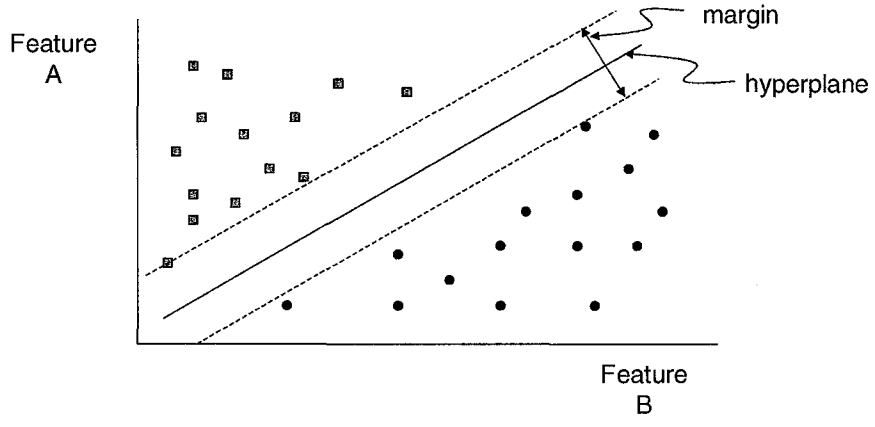

Figure 5.1 Support Vector Machines separate classes in a feature space with a hyperplane to maximize the margin of separation.

Each value in the SVM classifier is represented by a tuple $(\mathbf{x}_i, y_i)$ (i = 1, 2, ..., N in this example) where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})^T$ corresponds to the feature set for the ith value where d is the number of features. The value of y_i can either be 1 or -1 to denote the binary choice. The decision boundary of an SVM linear classifier has the form:

$$\mathbf{w} \bullet \mathbf{x} + b = 0$$

where **w** and b are parameters in the model. The boundary is determined from training data that has already been classified. For a linear model, the training data is used to set **w** and b (after scaling) such that:

$$\min f(\mathbf{w}) = \|\mathbf{w}\|^2 / 2$$

subject to the constraint $y_i(\mathbf{w} \bullet \mathbf{x}_i + b) \geq 1$, $i = 1, 2, \ldots, N$ where $\mathbf{x}_i$ is an instance of training data. The parameters $\mathbf{w}$ and b must be chosen to meet the following conditions:

$$\mathbf{w} \bullet \mathbf{x} + b \geq 1 \text{ if } y_i = 1 \text{ (i.e., for known ncRNA)},$$

$$\mathbf{w} \bullet \mathbf{x} + b \leq -1 \text{ if } y_i = -1 \text{ (i.e., for known non-ncRNA)}$$

Figure 5.2 illustrates these constraints.

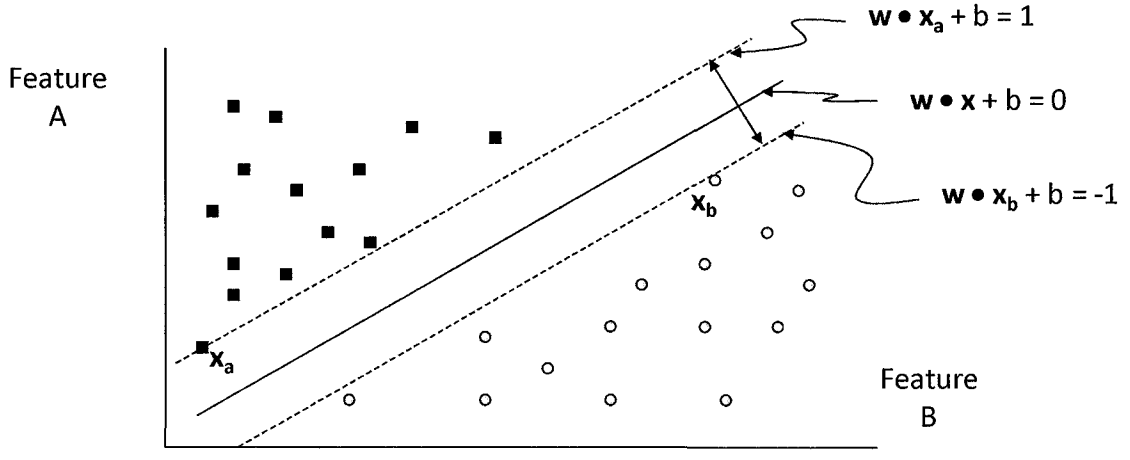

Figure 5.2 The constraints of a linear Support Vector Machine.

Two additional SVM issues need explanation for this paper:

- What if the training data is inside the margin because of noise?
- What if classes cannot be separated by a linear hyperplane?

To handle the first issue, positive slack variables $\xi_i > 0$ are added into the constraints of the f(w) optimization such that:

$$\min f(\mathbf{w}) = \|\mathbf{w}\|^2 / 2 + C(\sum_1^N \xi_i)^k, C > 0$$

$$\text{subject to } y_i(\mathbf{w} \bullet \mathbf{x}_i + b) \geq 1 - \xi_i, i = 1, 2, \ldots, N$$

where C and k represent penalties for misclassifying training instances.

To handle the second issue, the data may be transformed from its original space to a mapped space by function $\Phi(\mathbf{x})$ where there is a linear hyperplane between the positive and negative datasets. This mapping has the property:

$$K(\mathbf{u},\mathbf{v}) = \Phi(\mathbf{u}) \bullet \Phi(\mathbf{v})$$

where K is a kernel function that can be expressed as a dot product of two input vectors, **u** and **v**. Only certain kernel functions can be used. Common kernel functions include:

Polynomial: $K(\mathbf{x},\mathbf{x}) = (\gamma \mathbf{x}^T \mathbf{x} + r)^d, \gamma > 0$

Radial basis function: $K(\mathbf{x},\mathbf{x}) = \exp(-\gamma \| \mathbf{x} - \mathbf{x} \|^2), \gamma > 0$

Sigmoid: $K(x,x) = K(\mathbf{x},\mathbf{x}) = \tanh(\gamma \mathbf{x}^T \mathbf{x} + r), \gamma > 0$

This paper will compare the performance of the SVM with polynomial and radial basis function kernels with the other classification methods.

CHAPTER 6

CLASSIFICATION RESULTS

The test sets for this paper are chosen from four RNA families:

- Group II catalytic intron (RF00029)
- U5 Spliceosomal RNA (RF00020)
- 5S rRNA (RF00001)
- SRP RNA

A set of around 1800 multiple sequence alignments per family comprise the test set with about 600 each from MSAs with 2, 3, and 4 sequences in each family. Negative test sets were derived from the positive test MSAs by shuffling columns in the alignments, but preserving length, base composition, gap pattern, and local conservation patterns. Note that the test sets were split evenly into training data and testing data.

6.1 Baseline Feature Set Results

The baseline feature set for this paper contains the two features used by RNAz: z score and structure conservation index (SCI). Table 6.1 shows the results of these baseline feature set using the Naïve Bayes classification method. The overall score, using Matthew's Correlation Coefficient, of this classification method across all families and MSA sequence numbers was 93.80%. Considering the ncRNA families separately, the Group II catalytic intron family (RF00029) had the lowest score of 90.67%, whereas the 5S rRNA family (RF00001) had the highest score of 95.45%.

Table 6.1 Naïve Bayes Classification Using z score and SCI

Note: A comparison of the sensitivities, specificities, and Matthew's Correlation Coefficient of the Naïve Bayes classification using baseline features of z score and SCI for different ncRNA families and multiple sequence alignments of 2, 3, and 4 sequences.

RNA families	Sequences	Sn (%)	Sp (%)	MCC (%)
SRP RNA	**2**	100.00	97.12	97.16%
	3	79.25	100.00	80.25%
	4	82.67	95.33	78.63%
	All	92.81	98.66	91.63%
RF00020	**2**	92.33	93.33	85.67%
	3	98.33	95.67	94.03%
	4	90.57	95.29	85.95%
	All	97.10	95.09	92.21%
RF00001	**2**	94.67	95.67	90.34%
	3	99.33	98.33	97.67%
	4	100.00	98.33	98.35%
	All	98.22	97.22	95.45%
RF00029	**2**	94.33	96.21	90.53%
	3	98.32	96.64	94.98%
	4	99.65	93.66	93.50%
	All	96.15	94.50	90.67%
All	**2**	97.86	94.85	92.76%
	3	95.23	97.77	93.03%
	4	88.67	96.44	85.37%
	All	95.79	98.00	93.80%

Table 6.2 shows the results of the baseline feature set using the Fisher Linear Discriminant classifier. This classifier shows a marked improvement over the Naïve Bayes classifier with an overall score, using Matthew's Correlation Coefficient, of 96.93%. The classification scores for the four different families range from 94.53% to 97.79%. With this classifier, the Group II catalytic intron family (RF00029) continued to have the lowest score, but the U5 Spliceosomal RNA family (RF00020) had the highest score of 97.79%.

Table 6.2 Fisher Linear Discriminant Classification Using z score and SCI
Note: A comparison of the sensitivities, specificities, and Matthew's Correlation Coefficient of the Fisher Linear Discriminant classification using baseline features of z score and SCI for ncRNA different families. Results are not broken out by number of sequences in the alignment.

RNA families	Sequences	Sn (%)	Sp (%)	MCC (%)
SRP RNA	**All**	98.15	97.45	95.58
RF00020	**All**	99.89	97.92	97.79
RF00001	**All**	98.33	96.72	95.01
RF00029	**All**	97.51	97.07	94.53
All	**All**	98.69	98.26	96.93

Tables 6.3 to 6.6 show the results of the baseline feature set using Support Vector Machine (SVM) classifiers with different kernel functions. All classifiers using the Libsvm library [Chang and Lin 2009] as the classifier platform. Table 6.3 shows the results with a linear kernel function. The overall score, using Matthew's Correlation Coefficient, of this classification method is 94.08%. The classification score is close to the Fisher Linear Discriminant. The classification scores for the four different families

range from 93.87% to 95.72%. Here, the score of the U5 Spliceosomal RNA family (RF00020) is lowest, and the Srp RNA family is the highest.

Table 6.3 SVM Linear Classification Using z score and SCI

Note: A comparison of the sensitivities, specificities, and Matthew's Correlation Coefficient of the SVM classification with a linear kernel using baseline features of z score and SCI for ncRNA different families and multiple sequence alignments of 2, 3, and 4 sequences.

RNA families	Sequences	Sn (%)	Sp (%)	MCC (%)
SRP RNA	**2**	99.36	97.49	96.82
	3	100.00	75.31	71.15
	4	98.33	95.47	93.71
	All	99.28	96.50	95.72
RF00020	**2**	94.67	95.95	90.67
	3	98.67	97.37	96.01
	4	98.65	97.67	96.30
	All	97.44	96.47	93.87
RF00001	**2**	95.67	95.03	90.67
	3	99.33	98.68	98.00
	4	100.00	99.01	99.00
	All	98.22	97.36	95.56
RF00029	**2**	92.67	98.23	91.00
	3	98.32	97.67	95.98
	4	100.00	96.62	96.56
	All	96.72	98.16	94.89
All	**All**	97.48	96.64	94.08

Table 6.4 has the results with a polynomial kernel function (with degree d = 2, r = 1, γ = 1). The overall score, using Matthew's Correlation Coefficient, of this classification method is 93.86%. The classification scores of the four different families range from 93.87% to 95.56%; this is very close to the linear method. The score of the U5 Spliceosomal RNA family (RF00020) is lowest and the 5S rRNA (RF00001) is highest.

Table 6.4 SVM Polynomial Classification Using z score and SCI

Note: A comparison of the sensitivities, specificities, and Matthew's Correlation Coefficient of the SVM classification with a polynomial kernel using baseline features of z score and SCI for ncRNA different families and multiple sequence alignments of 2, 3, and 4 sequences.

.

RNA families	Sequences	Sn (%)	Sp (%)	MCC (%)
SRP RNA	**2**	98.72	98.10	96.81
	3	100.00	75.16	70.93
	4	91.67	96.83	88.79
	All	98.66	96.00	94.59
RF00020	**2**	94.67	96.27	91.01
	3	99.00	97.06	96.02
	4	98.65	94.52	93.02
	All	97.32	96.57	93.87
RF00001	**2**	95.67	95.03	90.67
	3	99.33	98.35	97.67
	4	99.67	99.01	98.67
	All	98.22	97.36	95.56
RF00029	**2**	91.67	98.92	90.77
	3	97.99	97.99	95.97
	4	100.00	96.62	96.56
	All	96.27	98.04	94.36
All	**All**	97.26	96.63	93.86

Table 6.5 has the classification results with a radial basis function kernel. The overall score, using Matthew's Correlation Coefficient, of this method is 93.70%. The range of the four families is 94.05% to 96.21%. The Group II catalytic intron family (RF00029) has the lowest score and Srp RNA has the highest score.

Table 6.5 SVM Radial Basis Function Classification Using z score and SCI
Note: A comparison of the sensitivities, specificities, and Matthew's Correlation Coefficient of the SVM classification with a radial basis function kernel using baseline features of z score and SCI for ncRNA different families and multiple sequence alignments of 2, 3, and 4 sequences.

RNA families	Sequences	Sn (%)	Sp (%)	MCC (%)
SRP RNA	**2**	92.97	98.64	90.10
	3	99.17	84.20	81.86
	4	27.00	98.78	38.82
	All	98.97	97.27	96.21
RF00020	**2**	94.33	96.26	90.68
	3	98.67	97.37	96.01
	4	97.98	94.79	92.65
	All	97.21	96.89	94.09
RF00001	**2**	95.67	95.03	90.67
	3	99.00	96.12	95.04
	4	100.00	98.36	98.35
	All	98.00	97.67	95.67
RF00029	**2**	90.00	99.26	89.57
	3	97.32	98.31	95.64
	4	99.65	97.27	96.88
	All	95.48	98.48	94.05
All	**All**	97.26	96.47	93.70

Table 6.6 summarizes the classification results is this baseline case. The best classification method for the baseline feature set is Fisher Linear Discriminant with an overall score of 96.93%. Still, the scores of all classification methods were close with a range of 93.70% to 96.93%.

Table 6.6 Summary of Classification Methods Using z score and SCI
Note: A comparison of the sensitivities, specificities, and Matthew's Correlation Coefficient of the classification methods using baseline features of z score and SCI for ncRNA different families.

RNA families	Method	Sn (%)	Sp (%)	MCC (%)
SRP RNA	**Naïve Bayes**	92.81	98.58	91.63
	Fisher Linear Discriminant	98.15	97.45	95.58
	SVM Linear	99.28	96.50	95.72
	SVM Polynomial	98.66	96.00	94.59
	SVM Radial Basis Function	98.97	97.27	96.21
RF00020	**Naïve Bayes**	97.10	95.19	92.21
	Fisher Linear Discriminant	99.89	97.92	97.79
	SVM Linear	97.44	96.47	93.87
	SVM Polynomial	97.32	96.57	93.87
	SVM Radial Basis Function	97.21	96.89	94.09
RF00001	**Naïve Bayes**	98.22	97.25	95.45
	Fisher Linear Discriminant	98.33	96.72	95.01
	SVM Linear	98.22	97.36	95.56
	SVM Polynomial	98.22	97.36	95.56
	SVM Radial Basis Function	98.00	97.67	95.67
RF00029	**Naïve Bayes**	96.15	94.65	90.67
	Fisher Linear Discriminant	97.51	97.07	94.53
	SVM Linear	96.72	98.16	94.89
	SVM Polynomial	96.27	98.04	94.36
	SVM Radial Basis Function	95.48	98.48	94.05
All	**Naïve Bayes**	95.79	97.96	93.80
	Fisher Linear Discriminant	98.69	98.26	96.93
	SVM Linear	97.48	96.64	94.08
	SVM Polynomial	97.26	96.63	93.86
	SVM Radial Basis Function	97.26	96.47	93.70

6.2 Target Feature Set Results

From this baseline case, this work seeks to enhance the classification score by incorporating additional features. To do this, the Class Separation Method (CSM) is calculated for each feature and the features are sorted from the highest to the lowest score. The ranked features are shown in Table 6.7.

Table 6.7 Class Separation Measure (CSM) for ncRNA Features

Note: Class Separation Measure for each measured ncRNA feature. The Class Separation Measure indicates the ability of the feature to differentiate between positive and negative classes for a particular feature. The Class Separation Measure is broken down by ncRNA family and by number of sequences. Features are ordered by overall CSM score.

Features	SRP RNA				RF00020				RF00001				RF00029				All
Sequences	2	3	4	All	2	3	4	All	2	3	4	All	2	3	4	All	All
z score	10.288	7.910	5.307	7.297	3.392	9.386	6.653	5.387	4.434	5.893	6.571	5.133	6.021	5.185	4.034	4.849	5.824
normalized centroidfold base pairs	2.524	1.962	0.750	1.453	5.195	5.718	8.009	8.110	2.065	3.426	4.871	3.186	3.270	4.107	5.093	4.047	2.850
shannon entropy	0.45	1.71	0.92	1.68	2.78	4.63	3.82	3.35	1.44	2.08	3.03	2.04	1.45	1.49	1.25	1.37	1.886
SCI	3.349	1.305	1.165	1.003	2.163	2.972	2.864	2.393	6.003	8.674	11.065	7.301	1.702	1.456	1.168	1.349	1.615
base-pair distance	0.13	1.54	0.90	1.57	2.58	4.28	3.53	3.05	1.34	1.92	2.83	1.89	0.95	0.92	0.94	0.92	1.584
centroidfold max consecutive base pairs	0.301	1.250	0.971	0.765	1.655	2.516	3.237	2.619	0.199	0.205	0.700	0.322	4.903	5.384	6.790	5.590	0.996
normalized centroidfold hairpin length	0.348	0.805	0.511	0.533	1.436	1.444	1.836	1.952	1.561	1.699	1.562	1.567	0.939	1.777	1.702	1.414	0.899
normalized rnaalifold base pairs	1.339	0.576	0.481	0.547	0.677	0.590	1.188	1.304	0.596	1.926	1.740	1.270	2.639	4.677	1.983	2.628	0.805
normalized rspredict base pairs	2.216	0.030	0.405	0.451	0.824	1.177	0.449	0.734	2.240	2.435	2.616	2.376	1.241	2.998	1.346	1.624	0.764
normalized centroidfold max consecutive base pairs	0.267	0.557	0.420	0.348	1.664	2.464	3.153	2.634	0.200	0.205	0.699	0.321	3.687	4.332	5.173	4.278	0.634
rnaalifold max consecutive base pairs	0.109	0.556	0.372	0.314	0.534	0.686	1.059	0.957	0.005	0.212	0.142	0.077	2.998	5.861	6.842	4.739	0.524
normalized rspredict max consecutive base pairs	1.243	0.025	0.378	0.448	1.645	1.469	2.623	1.817	0.568	0.496	0.709	0.582	0.345	0.237	0.098	0.206	0.408
rspredict max consecutive base pairs	1.331	1.179	0.701	0.841	1.627	1.515	2.902	1.885	0.576	0.503	0.721	0.592	0.347	0.271	0.117	0.230	0.405
normalized rnaalifold hairpin length	0.061	0.188	0.255	0.156	0.835	0.501	0.767	0.866	1.153	1.589	1.599	1.435	0.363	0.517	0.547	0.467	0.399
centroidfold base pairs	2.244	0.773	0.352	0.604	4.974	5.554	7.998	7.953	2.023	3.336	4.759	3.118	2.575	4.608	5.313	3.879	0.391
normalized rnaalifold max consecutive base pairs	0.100	0.250	0.125	0.130	0.542	0.675	1.044	1.000	0.005	0.212	0.142	0.077	2.295	4.326	4.886	3.414	0.333
centroidfold largest loop	0.145	0.172	0.147	0.155	0.324	0.346	0.361	0.352	0.022	0.176	0.193	0.113	0.079	0.000	0.081	0.033	0.140
rnaalifold largest bulge	0.439	0.250	0.080	0.205	0.032	0.009	0.053	0.101	0.005	0.017	0.075	0.026	0.304	0.421	0.454	0.375	0.119
normalized rspredict largest bulge	0.435	0.001	0.011	0.016	0.030	0.150	0.027	0.050	0.103	0.291	0.119	0.158	0.533	1.151	0.657	0.725	0.117
normalized rspredict number of loops	0.088	0.000	0.057	0.008	0.535	0.670	0.444	0.543	0.039	0.066	0.058	0.053	0.760	0.826	0.658	0.735	0.112
normalized rnaalifold number of loops	1.927	1.481	0.983	1.344	0.040	0.038	0.007	0.011	0.088	0.345	0.270	0.209	0.015	0.002	0.011	0.004	0.108
normalized centroidfold largest loop	0.143	0.065	0.056	0.066	0.324	0.348	0.360	0.345	0.022	0.174	0.190	0.111	0.082	0.002	0.103	0.045	0.104
normalized rspredict energy density	1.599	0.041	1.958	1.323	2.852	3.642	2.505	2.752	3.642	5.564	6.767	4.731	1.114	2.675	1.799	1.713	0.100
normalized rnaalifold largest bulge	0.429	0.012	0.000	0.011	0.032	0.012	0.057	0.100	0.005	0.016	0.072	0.025	0.330	0.441	0.484	0.402	0.094
rnaalifold number of loops	1.665	0.653	0.225	0.536	0.041	0.032	0.006	0.016	0.090	0.351	0.272	0.213	0.006	0.003	0.009	0.002	0.088
centroidfold hairpin length	0.184	0.047	0.022	0.033	1.363	1.409	1.844	1.871	1.551	1.679	1.554	1.558	0.753	1.709	1.575	1.253	0.065
rspredict number of loops	0.099	0.023	0.002	0.000	0.531	0.693	0.475	0.562	0.039	0.066	0.058	0.054	0.834	0.888	0.735	0.819	0.063
Consensus MFE	1.541	0.378	0.189	0.302	2.971	8.150	5.914	4.947	3.899	5.688	6.319	4.987	4.118	5.143	3.554	4.148	0.060
rspredict energy density	0.808	0.224	0.071	0.110	2.324	4.509	3.264	2.952	3.453	5.172	6.152	4.472	0.673	4.427	2.412	1.669	0.051
normalized rspredict hairpin length	0.028	0.000	0.029	0.015	0.124	0.369	0.221	0.226	0.597	0.727	0.946	0.743	0.002	0.006	0.005	0.004	0.050
centroidfold largest bulge	0.303	0.219	0.057	0.176	0.052	0.092	0.079	0.045	0.006	0.052	0.115	0.049	0.007	0.000	0.004	0.000	0.047
normalized centroidfold number of loops	0.245	0.940	0.831	0.628	0.948	1.359	1.835	1.114	0.061	0.032	0.001	0.024	0.504	0.634	0.695	0.605	0.042
rnaalifold largest loop	0.434	0.435	0.126	0.307	0.234	0.245	0.183	0.175	0.132	0.139	0.175	0.148	0.008	0.012	0.149	0.033	0.036
rnaalifold base pairs	0.561	0.058	0.032	0.051	0.597	0.566	1.220	1.230	0.556	1.751	1.600	1.182	1.401	4.503	2.332	2.254	0.034
rspredict base pairs	0.975	0.082	0.020	0.047	0.624	1.499	0.715	0.858	2.070	2.205	2.371	2.183	0.538	2.849	1.452	1.140	0.023
rspredict largest bulge	0.491	0.018	0.001	0.006	0.031	0.120	0.013	0.026	0.106	0.304	0.125	0.165	0.464	1.054	0.402	0.538	0.021
normalized centroidfold largest bulge	0.310	0.046	0.014	0.026	0.052	0.093	0.080	0.044	0.006	0.051	0.114	0.048	0.008	0.000	0.003	0.000	0.021
normalized rspredict largest loop	0.086	0.000	0.001	0.003	0.353	0.431	0.388	0.389	0.003	0.024	0.008	0.001	0.055	0.101	0.000	0.035	0.016
rspredict largest loop	0.086	0.120	0.005	0.016	0.348	0.392	0.387	0.374	0.003	0.026	0.007	0.001	0.051	0.111	0.002	0.038	0.010
normalized rnaalifold largest loop	0.404	0.135	0.043	0.085	0.232	0.249	0.183	0.164	0.127	0.132	0.168	0.142	0.018	0.007	0.138	0.025	0.009
rnaalifold hairpin length	0.015	0.002	0.002	0.002	0.708	0.486	0.804	0.778	1.057	1.441	1.503	1.320	0.115	0.332	0.321	0.227	0.006
Avg. MFE	0.438	0.085	0.049	0.081	2.033	8.678	7.210	4.255	0.896	1.560	1.552	1.289	0.905	2.170	1.552	1.343	0.001
centroidfold number of loops	0.209	0.485	0.274	0.275	0.935	1.354	1.792	1.038	0.061	0.032	0.001	0.024	0.479	0.690	0.737	0.623	0.001
rspredict hairpin length	0.007	0.000	0.000	0.000	0.059	0.238	0.053	0.076	0.480	0.563	0.801	0.606	0.001	0.001	0.001	0.001	0.000

Not surprisingly, z score tops the list of differentiating features. It is followed by normalized CentroidFold base pairs, Shannon entropy, SCI, base-pair distance, CentroidFold maximum consecutive base pairs, and normalized CentroidFold hairpin length as the next six differentiating features. It is interesting to note that the CentroidFold folding algorithm seems to be a better source of differentiation than RNAalifold or RSpredict. Figure 6.1 shows this differentiation for the SRP RNA family. In addition, note that some of these features are highly correlated (e.g., Shannon entropy and base-pair distance). The classifier may be able to be simplified by reducing or combining these correlated features.

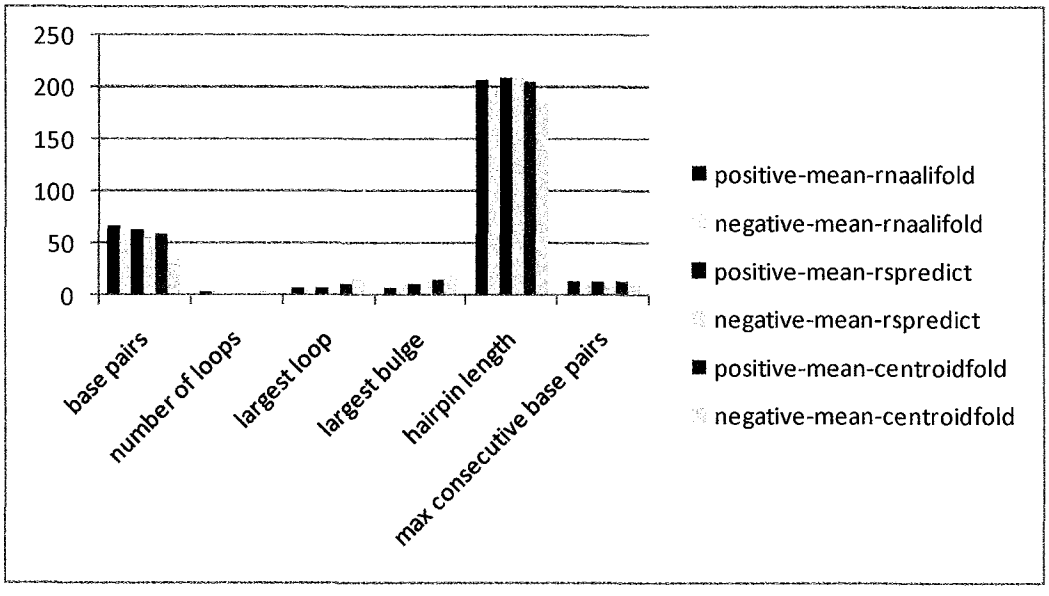

Figure 6.1 Comparison of feature values for different folding methods for SRP RNA MSAs.

Note: The mean feature values for the folding methods-- RNAalifold, RSpredict, and CentroidFold—are plotted for the positive and negative classes for SRP RNA multiple sequence alignments with four sequences. For the features base pairs and hairpin length, there is a bigger difference between the mean values for CentroidFold than for other folding methods.

For the RNAMultifold model, the top 11 features with CSM > 0.5 were chosen from X_{ranked} as the target feature set. The 11 features, listed in Table 6.8, were then used with the different classification methods to determine the best method for the model. Table 6.9 shows the results of the classification methods for this feature set.

Table 6.8 Top 11 CSM Features

Note: Class Separation Measure for the top 11 ncRNA features. The Class Separation Measure indicates the ability of the feature to differentiate between positive and negative classes for a particular feature. The top 11 features were compared with the baseline features of z score and SCI for classification of ncRNAs.

Features	Overall CSM Score
z score	5.824
normalized centroidfold base pairs	2.850
shannon entropy	1.886
SCI	1.615
base-pair distance	1.584
centroidfold max consecutive base pairs	0.996
normalized centroidfold hairpin length	0.899
normalized rnafold base pairs	0.805
normalized rspredict base pairs	0.764
normalized centroidfold max consecutive base pairs	0.634
rnafold max consecutive base pairs	0.524

Table 6.9 Comparison of Classification Methods Using Top 11 SCM Features

Note: A comparison of the sensitivities, specificities, accuracy, and Matthew's Correlation Coefficient of the Naïve Bayes, Fisher Linear Discriminant, and SVM classification methods using the top 11 SCM features for ncRNA different families.

RNA families	Method	Sn (%)	Sp (%)	MCC (%)
SRP RNA	Naïve Bayes	95.26	90.16	85.00
	Fisher Linear Discriminant	99.90	97.68	97.55
	SVM Linear	99.79	98.68	98.46
	SVM Polynomial	99.38	98.77	98.15
	SVM Radial Basis Function	99.59	98.88	98.46
RF00020	Naïve Bayes	98.33	98.99	97.32
	Fisher Linear Discriminant	96.54	99.08	95.68
	SVM Linear	99.00	98.45	97.43
	SVM Polynomial	99.11	99.11	98.21
	SVM Radial Basis Function	99.00	99.11	98.10
RF00001	Naïve Bayes	99.00	89.91	88.34
	Fisher Linear Discriminant	95.47	97.15	92.57
	SVM Linear	98.89	98.56	97.44
	SVM Polynomial	99.11	99.55	98.67
	SVM Radial Basis Function	99.11	99.22	98.33
RF00029	Naïve Bayes	98.98	95.41	94.25
	Fisher Linear Discriminant	98.75	97.10	95.80
	SVM Linear	98.75	97.76	96.47
	SVM Polynomial	99.32	97.77	97.05
	SVM Radial Basis Function	99.43	98.65	98.07
All	Naïve Bayes	97.84	93.41	91.02
	Fisher Linear Discriminant	97.67	97.72	95.39
	SVM Linear	98.77	97.49	96.22
	SVM Polynomial	98.85	98.02	96.85
	SVM Radial Basis Function	98.96	98.66	97.61

This feature set shows substantial improvements over the baseline set for all classification methods using SVM models. The best classification method for this feature set is the SVM with a radial basis function kernel with an overall score of 97.61%. The polynomial kernel has the next highest overall score of 96.85% and linear kernel has a score of 96.22%. The range for the radial basis function kernel is 96.47% to 98.46%; for the polynomial kernel, 97.05% to 98.67%; and, for the linear kernel, 96.47% to 98.46%. The Fisher Linear Discriminant had an overall score of 95.39% across all families.

6.3 Receiver Operating Curves

Receiver Operating Curves (ROCs) are used to express overall quality of the ncRNA classification method by showing sensitivity as a function of specificity, or positive predictive value (PPV) [Hamada et al. 2009]. Figure 6.2 shows the ROC for SVM models using the baseline feature set: z score and SCI.

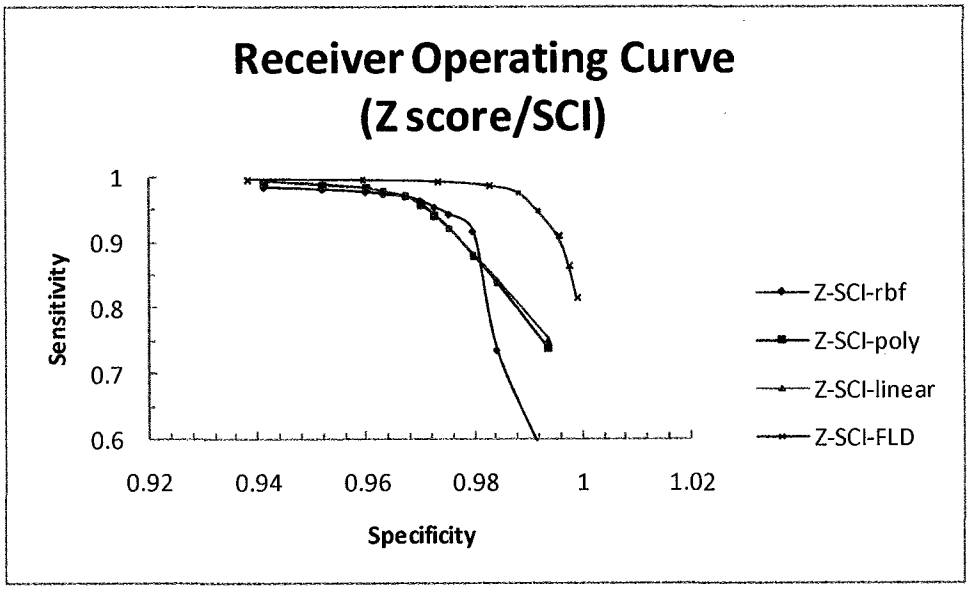

Figure 6.2 ROCs comparing classification methods using baseline features.

Note: Receiver Operating Curves showing sensitivity vs. specificity for different classification methods (FLD – Fisher Linear Discriminant; Support Vector Machines with a linear kernel (Z-SCI-linear), with a polynomial kernel (Z-SCI-poly), and with a radial basis function kernel (Z-SCI-rbf)) using the baseline features of z score and SCI.

The ROCs confirm the prior results and shows that the Fisher Linear Discriminant Classification is superior to other classification models across all specificity values. It also illustrates that the SVM model with the RBF kernel is superior to the SVM model with polynomial and linear kernels over a narrow range specificities from 0.966 to 0.98. However, above a specificity value of 0.98 the sensitivity of the SVM model with the RBF kernel falls off much more rapidly than the other two SVM methods. Below a specificity value of 0.966, the other two kernels, polynomial and linear, are slightly superior in classification.

Figure 6.3 illustrates a ROC using SVM models using the top 11 features according to the Class Separation method. In these models, the RBF kernel is more sensitive than the polynomial and linear models across the full range of specificity values and does not show the same drop-off in sensitivity for the RBF kernel against the other kernels. Also, note that the breakpoint in the curve is much closer to the specificity of 1.

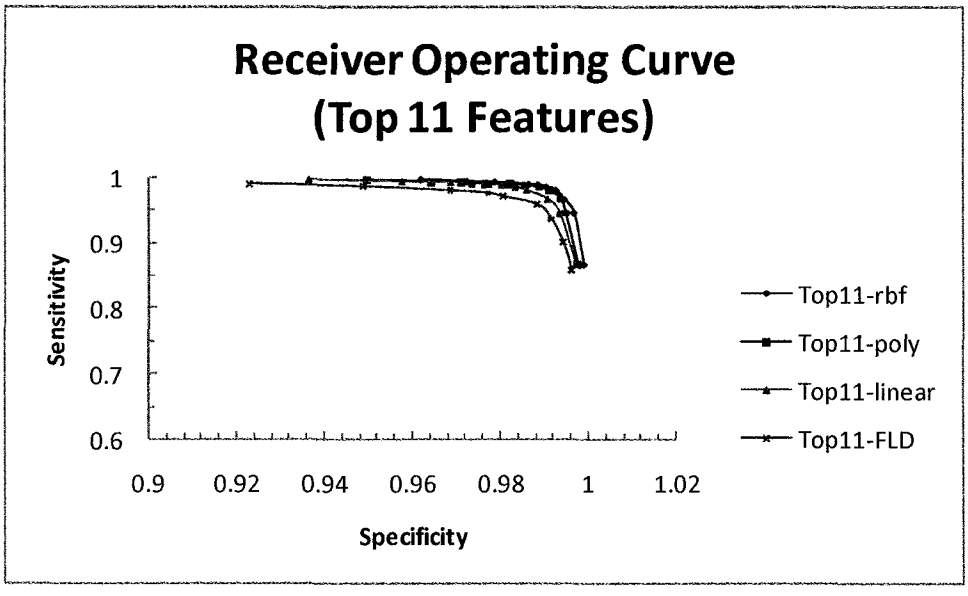

Figure 6.3 ROCs comparing classification methods using target top 11 features.

Note: Receiver Operating Curves showing sensitivity vs. specificity for different classification methods (FLD – Fisher Linear Discriminant; Support Vector Machines with a linear kernel (Top11-linear), with a polynomial kernel (Top11-poly), and with a radial basis function kernel (Top11-rbf)) using the target top 11 CSM features.

Figure 6.4 directly compares the ROCs for the models with the top 11 features and the models with the baseline features. The models with the top 11 features are more sensitive across all specificity values except for the Fisher Linear Discriminant model. The Fisher Linear Discriminant model with the baseline features ranks third after SVM models with the radial basis function and polynomial kernels using the top 11 features. It

is interesting to note that the FLD model with the baseline features is more sensitive than the FLD model with the top 11 features.

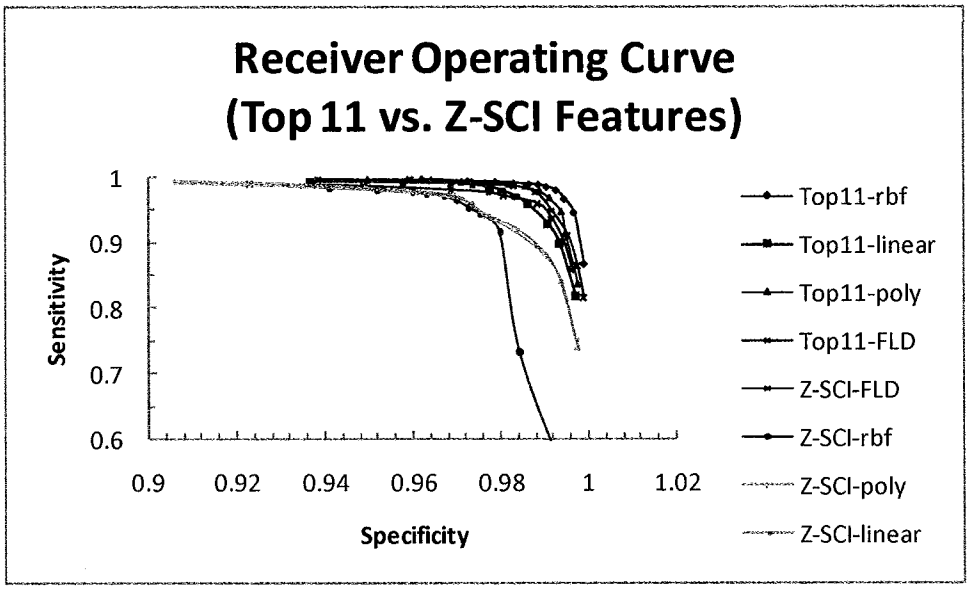

Figure 6.4 ROCs comparing classification methods using target top 11 features and using the baseline features.

Note: Receiver Operating Curves showing sensitivity vs. specificity for different classification methods and different feature sets.

6.4 Comparison with RNAz

Table 6-10 shows the classification test results of RNAz using the default threshold of P > 0.5 for ncRNA prediction. The overall classification score, using Matthew's Correlation Coefficient, is 91.15%. The classification scores for the four different families range from 84.59% to 95.45%. With RNAz, the Group II catalytic intron family (RF00029) had the highest score, and the 5S rRNA family (RF00001) had the lowest score.

Table 6.10 Classification Test Results of RNAz

Note: A comparison of the sensitivities, specificities, accuracy, and Matthew's Correlation Coefficient of the classification methods for ncRNA different families using the RNAz prediction tool.

RNA families	Sequences	Sn (%)	Sp (%)	MCC (%)
SRP RNA	**2**	99.04	96.27	95.25
	3	94.17	97.13	91.43
	4	91.67	95.82	87.75
	All	94.96	96.45	91.48
RF00001	**2**	82.33	94.64	78.33
	3	94.33	97.25	91.71
	4	86.81	96.92	83.93
	All	87.80	96.33	84.59
RF00020	**2**	83.67	98.82	83.63
	3	100.00	99.01	99.00
	4	99.33	99.33	98.66
	All	94.33	99.07	93.53
RF00029	**2**	97.67	95.75	93.36
	3	100.00	97.07	97.03
	4	100.00	95.99	95.99
	All	99.21	96.27	95.45
All	**All**	94.03	97.00	91.15

Figure 6.5 shows the ROC for the SVM classifier with RBF kernel with the top 11 CSM features and for RNAz. The SVM classifier outperformed RNAz the full range of specificity values.

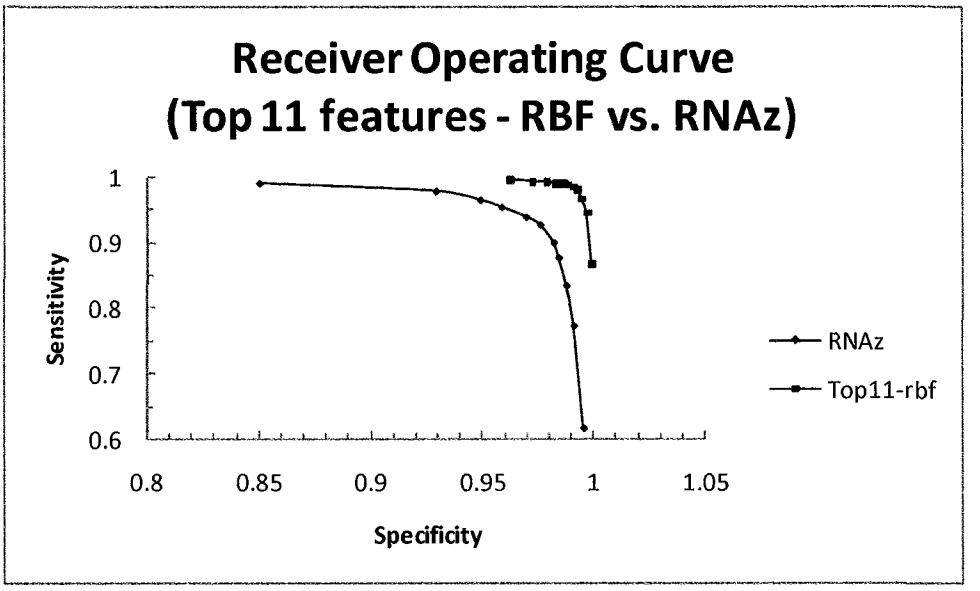

Figure 6.5 ROCs comparing classification methods of SVM with a radial basis function kernel using target top 11 features and RNAz.

Note: Receiver Operating Curves showing sensitivity vs. specificity of the Support Vector Machine with a radial basis function kernel using the target top 11 features and RNAz.

Figure 6.6 compares the ROC for the FLD and SVM classifiers with baseline features and RNAz. The FLD classifier had the best performance. The SVM classifiers with the linear and polynomial kernels outperform RNAz, although the margin of difference is reduced. The SVM classifier with the RBF kernel is more sensitive than RNAz below a specificity of 0.98, but RNAz is more sensitive above 0.98.

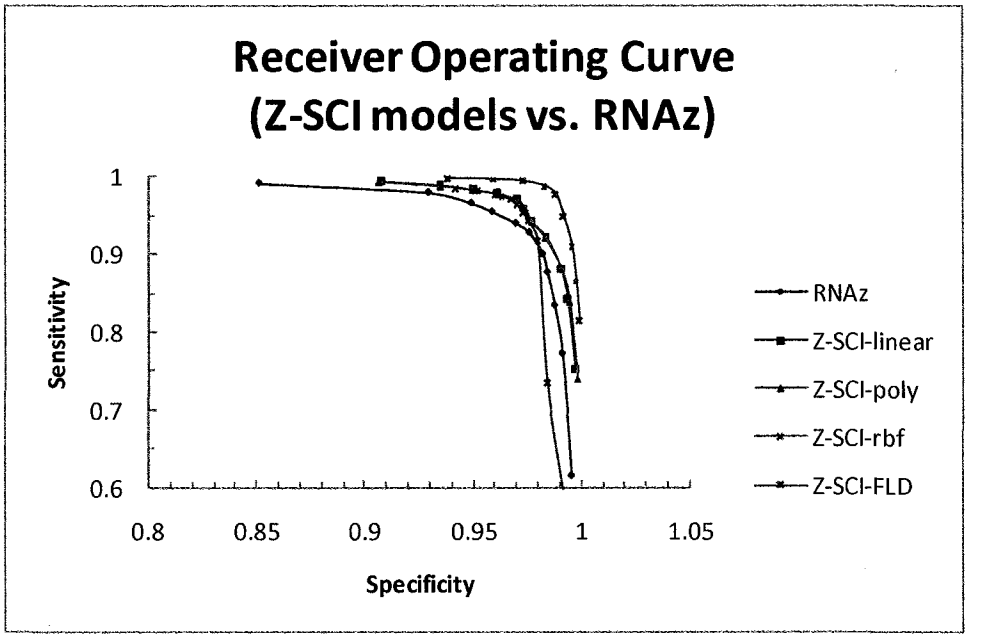

Figure 6.6 ROCs comparing RNAz and the SVM classification methods using baseline features.

Note: Receiver Operating Curves showing sensitivity vs. specificity of RNAz and the Support Vector Machine classification methods using the baseline feature set of z score and SCI.

Figure 6.7 compares the performance of the SVM classifiers with various kernels and the FLD model with the top 11 CSM features with RNAz. All the classifiers with the top 11 CSM features are more sensitive than RNAz across the range of specificities.

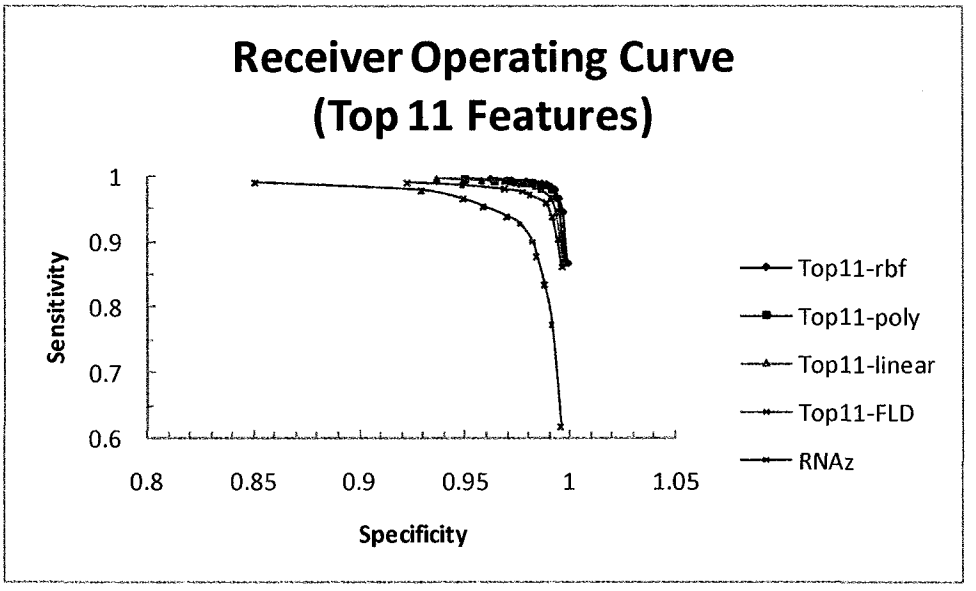

Figure 6.7 ROCs comparing RNAz and the SVM classification methods using target top 11 features.

Note: Receiver Operating Curves showing sensitivity vs. specificity of RNAz and the Support Vector Machine classification methods using the target top 11 CSM features.

It has been shown that the RNAMultifold method using an SVM classification model with a radial basis function kernel with a target feature set determined by Class Separation Measure is a superior predictor of ncRNA compared with other classification models and RNAz for the set of ncRNA families under study.

CHAPTER 7

GENOME SCAN OF T. BRUCEI FOR ncRNAs

The next test for the tools and techniques is to search genomes for known and novel ncRNAs. Trypanosome genomes are selected for this search.

Trypanosomes are unicellular parasitic protozoa that cause major diseases in humans. *T. brucei* causes African trypanosomiasis, or "sleeping sickness;" Chagas' protozoa are known for unique RNA processing mechanisms such as nuclear pre-mRNA trans-splicing and mitochondrial RNA editing [Myslyuk et al. 2008]. Genes in trypanosomes are transcribed as polycistronic mRNAs that are then processed via trans-splicing [Mao et al. 2009]. The regulation of gene expression involves cis-acting ncRNA elements such as U-rich elements (UREs) and short interspersed degenerated retroposons (SIDERS) [Mao et al. 2009].

An understanding of the genomics of these species is crucial because they collectively cause millions of deaths in Africa and developing countries on other continents around the world. The genomes are organized into clusters with 10's to 100's of protein-coding genes, sequentially arranged on the same strand of DNA [Padilla-Mejía 2008].

This thesis considers three trypanosome species, known collectively as the Tritryps:

- *T. brucei* with 11 large chromosomes,
- *T. cruzi* with 28 medium-sized chromosomes, and
- *L. major* with 36 small chromosomes.

The search for ncRNAs will focus on *T. brucei*. The other two species are used for multiple sequence alignments.

To use RNAz or RNAMultifold to search for known and novel ncRNAs, multiple sequence alignments need to be arranged. The tool, Mauve [Darling et al. 2004], is used to create multiple alignments. Mauve identifies conserved regions in the genomes and breakpoints of rearrangements and inversions of these conserved regions. The alignments are output in XMFA format. For the Tritryps, 667 intervals were aligned by Mauve as part of this study.

The Mauve alignment was converted into ClustalW format and a tool from RNAz, `rnazWindow.pl`, was used to slice alignments into overlapping windows. A window size of 120 nucleotides with a step of 40 nucleotides was used. During this process, alignment windows are discarded if:

- The fraction of gaps in the reference sequence, *T. brucei*, is higher than 25%;
- The fraction of masked nucleotides in a sequence is greater than 10%;
- The window has a mean pairwise identity less than 50%;
- There is only one sequence in the alignment window.

In total, 49,529 windows were created in this process. These alignment windows can now be used by the ncRNA prediction tool.

7.1 Genome Search by RNAz

RNAz was used first to search the *T. brucei* genome for ncRNAs. The forward strand was included in the search. In total, 451 potential ncRNAs were identified with a probability greater than 0.5. Of these, 178 were identified with a probability of greater than 0.9.

From the TriTryp Database [Aslett et al. 2009], there are 307 known ncRNAs in forward strands of the *T.brucei* genome. Of these, 224 were retained when the *T. brucei*, *T. cruzi*, and *L. major* genomes were aligned with the Mauve tool. When sliced into windows and subjected to the criteria given in the previous section, 33 known ncRNAs were retained—14 fully and 19 partially. When the 451 predicted ncRNAs from RNAz were compared against the 33 known ncRNAs, 12 matched. Table 7-1 shows the 33 known ncRNAs retained and the RNAz predictions. One of the aims of this thesis is to see if alternative classifications schemes can improve this hit rate.

7.2 Genome Search by Fisher Linear Discriminant Model

As another prediction test, the Fisher Linear Discriminant (FLD) model for all training sequences was used to predict ncRNAs. A total of 717 windows were predicted as ncRNAs using this method. Table 7.1 shows that two of these predictions matched the 33 known ncRNAs in the set above.

Table 7.1 Known ncRNAs in the *T. brucei* genome

Note: The list of known ncRNAs contained in Mauve alignments and preprocessing slices in forward strands and the prediction results of RNAz, Fisher Linear Discriminant, and SVM with different kernels. Predictions given by Mao et al. 2009 are also shown. Table contains the gene identifier, chromosome, location on the chromosome, inclusion in the Mauve alignment, containment in preprocessing slices, predictions by classification methods, gene strand (forward or backward), gene type, and gene description. Headers in brackets are taken from the TriTrypDB web site.

[Gene]	[Genomic Sequence ID]	[Chromosome]	[Genomic Location]	In Mauve alignment	In slices	RNAz prediction	FLD prediction	SVM Linear prediction	SVM Polynomial prediction	SVM RBF prediction	Mao prediction	[Gene Strand]	[Gene Type]	[Product Description]
Tb927.4.1213	Tb927_04_v4	chromosome 4	Tb927_04_v4: 325,767 - 325,860 (+)	Yes	Yes	Yes	No	No	No	No	No	forward	snRNA encoding	small nuclear RNA; U6 snRNA
Tb927.4.1216	Tb927_04_v4	chromosome 4	Tb927_04_v4: 325,958 - 326,029 (+)	Yes	Yes	Yes	No	No	No	No	No	forward	tRNA encoding	tRNA Threonine
Tb927.8.2852	Tb927_08_v4	chromosome 8	Tb927_08_v4: 854,457 - 854,528 (+)	Yes	Yes	Yes	No	Yes	Yes	No	No	forward	tRNA encoding	tRNA Glutamine
Tb927.8.2853	Tb927_08_v4	chromosome 8	Tb927_08_v4: 854,585 - 854,657 (+)	Yes	Partial	Yes	No	No	No	Yes	Yes	forward	tRNA encoding	tRNA Valine
Tb927.8.2857	Tb927_08_v4	chromosome 8	Tb927_08_v4: 858,265 - 858,346 (+)	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	forward	tRNA encoding	tRNA Leucine
Tb09_snoRNA_0080	Tb927_09_v4	chromosome 9	Tb927_09_v4: 1,865,891 - 1,865,977 (+)	Yes	Partial	Yes	No	No	No	No	No	forward	snoRNA encoding	C/D snoRNA, TB9C3C2
Tb09_snoRNA_0081	Tb927_09_v4	chromosome 9	Tb927_09_v4: 1,866,012 - 1,866,086 (+)	Yes	Partial	Yes	No	No	No	No	No	forward	snoRNA encoding	H/ACA snoRNA, TB9Cs3H1
Tb09_snoRNA_0082	Tb927_09_v4	chromosome 9	Tb927_09_v4: 1,866,099 - 1,866,223 (+)	Yes	Yes	No	No	No	Yes	Yes	No	forward	snoRNA encoding	C/D snoRNA, TB9C3C3
Tb09_snoRNA_0085	Tb927_09_v4	chromosome 9	Tb927_09_v4: 1,866,553 - 1,866,639 (+)	Yes	Partial	No	No	No	No	No	No	forward	snoRNA encoding	C/D snoRNA, TB9C3C2
Tb09_snoRNA_0086	Tb927_09_v4	chromosome 9	Tb927_09_v4: 1,866,674 - 1,866,748 (+)	Yes	Partial	Yes	No	No	No	No	No	forward	snoRNA encoding	H/ACA snoRNA, TB9Cs3H1
Tb09_snoRNA_0087	Tb927_09_v4	chromosome 9	Tb927_09_v4: 1,866,761 - 1,866,885 (+)	Yes	Partial	No	Yes	No	No	No	No	forward	snoRNA encoding	C/D snoRNA, TB9C3C3
Tb09_snoRNA_0091	Tb927_09_v4	chromosome 9	Tb927_09_v4: 1,867,336 - 1,867,410 (+)	Yes	Yes	No	No	No	No	No	No	forward	snoRNA encoding	H/ACA snoRNA, TB9Cs3H1
Tb09_snoRNA_0092	Tb927_09_v4	chromosome 9	Tb927_09_v4: 1,867,423 - 1,867,547 (+)	Yes	Partial	No	No	No	No	No	No	forward	snoRNA encoding	C/D snoRNA, TB9C3C3
Tb10_snoRNA_0001	Tb927_10_v4	chromosome 10	Tb927_10_v4: 1,729,028 - 1,729,104 (+)	Yes	Partial	No	No	No	No	No	No	forward	snoRNA encoding	C/D snoRNA, TB10C1C4
Tb10_snoRNA_0002	Tb927_10_v4	chromosome 10	Tb927_10_v4: 1,729,160 - 1,729,228 (+)	Yes	Partial	No	No	No	No	No	No	forward	snoRNA encoding	H/ACA snoRNA, TB10Cs1H3
Tb10_snoRNA_0003	Tb927_10_v4	chromosome 10	Tb927_10_v4: 1,729,248 - 1,729,342 (+)	Yes	Partial	No	No	No	No	No	No	forward	snoRNA encoding	C/D snoRNA, TB10C1C1
Tb10_snoRNA_0004	Tb927_10_v4	chromosome 10	Tb927_10_v4: 1,729,389 - 1,729,461 (+)	Yes	Yes	Yes	No	No	No	Yes	No	forward	snoRNA encoding	H/ACA snoRNA, TB10Cs1H1
Tb10_snoRNA_0005	Tb927_10_v4	chromosome 10	Tb927_10_v4: 1,730,178 - 1,730,275 (+)	Yes	Partial	No	No	No	No	No	No	forward	snoRNA encoding	C/D snoRNA, TB10C1C3
Tb10_snoRNA_0007	Tb927_10_v4	chromosome 10	Tb927_10_v4: 1,730,491 - 1,730,567 (+)	Yes	Partial	No	No	No	No	No	No	forward	snoRNA encoding	C/D snoRNA, TB10C1C4
Tb10_snoRNA_0008	Tb927_10_v4	chromosome 10	Tb927_10_v4: 1,730,623 - 1,730,691 (+)	Yes	Partial	No	No	No	No	No	No	forward	snoRNA encoding	H/ACA snoRNA, TB10Cs1H3
Tb10_snoRNA_0009	Tb927_10_v4	chromosome 10	Tb927_10_v4: 1,730,711 - 1,730,805 (+)	Yes	Partial	No	No	No	No	No	No	forward	snoRNA encoding	C/D snoRNA, TB10C1C1
Tb10_snoRNA_0010	Tb927_10_v4	chromosome 10	Tb927_10_v4: 1,730,852 - 1,730,924 (+)	Yes	Yes	No	No	No	No	No	No	forward	snoRNA encoding	H/ACA snoRNA, TB10Cs1H1
Tb10_snoRNA_0012	Tb927_10_v4	chromosome 10	Tb927_10_v4: 1,738,895 - 1,738,963 (+)	Yes	Partial	No	No	No	No	No	No	forward	snoRNA encoding	H/ACA snoRNA, TB10Cs1H3
Tb10_snoRNA_0013	Tb927_10_v4	chromosome 10	Tb927_10_v4: 1,738,983 - 1,739,077 (+)	Yes	Partial	No	No	No	No	No	No	forward	snoRNA encoding	C/D snoRNA, TB10C1C1
Tb10_snoRNA_0014	Tb927_10_v4	chromosome 10	Tb927_10_v4: 1,739,124 - 1,739,196 (+)	Yes	Yes	No	No	No	No	No	No	forward	snoRNA encoding	H/ACA snoRNA, TB10Cs1H1
Tb10_snoRNA_0015	Tb927_10_v4	chromosome 10	Tb927_10_v4: 1,739,913 - 1,740,010 (+)	Yes	Partial	No	No	No	No	No	No	forward	snoRNA encoding	C/D snoRNA, TB10C1C3
Tb10_tRNA_Pro_1	Tb927_10_v4	chromosome 10	Tb927_10_v4: 1,907,937 - 1,908,008 (+)	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	forward	tRNA encoding	tRNA Proline
Tb10_tRNA_Arg_1	Tb927_10_v4	chromosome 10	Tb927_10_v4: 2,603,741 - 2,603,813 (+)	Yes	Yes	Yes	No	No	No	No	No	forward	tRNA encoding	tRNA Arginine
Tb10_tRNA_Lys_1	Tb927_10_v4	chromosome 10	Tb927_10_v4: 2,603,873 - 2,603,944 (+)	Yes	Yes	Yes	No	No	No	No	No	forward	tRNA encoding	tRNA Lysine
Tb10_tRNA_Arg_2	Tb927_10_v4	chromosome 10	Tb927_10_v4: 2,604,030 - 2,604,102 (+)	Yes	Yes	No	No	No	No	No	No	forward	tRNA encoding	tRNA Arginine
Tb11_snoRNA_0040	Tb927_11_01_v	chromosome 11	Tb927_11_01_v4: 2,703,080 - 2,703,169	Yes	Partial	No	No	No	No	No	No	forward	snoRNA encoding	C/D snoRNA, TB11C2C2; snoR
Tb11_snoRNA_0041	Tb927_11_01_v	chromosome 11	Tb927_11_01_v4: 2,703,287 - 2,703,555	Yes	Yes	No	No	No	No	No	No	forward	snoRNA encoding	H/ACA snoRNA, TB11Cs2H1; s
Tb11_snoRNA_0043	Tb927_11_01_v	chromosome 11	Tb927_11_01_v4: 2,703,912 - 2,704,001	Yes	Partial	No	No	No	No	No	No	forward	snoRNA encoding	C/D snoRNA, TB11C2C2; snoR

7.3 Genome Search by SVM Model

Three different SVM kernels were compared with *T. brucei* genome search: linear, polynomial, and radial basis function. These models used the eleven features identified in the previous section as a basis for the prediction. There was a substantial difference in the number of ncRNAs predicted by the different kernels. Table 7.2 shows the number of predictions of each kernel. The linear kernel had roughly the same number of predicted ncRNAs as RNAz. The polynomial and radial basis function kernels had a substantially higher number of positive predictions: 3801 and 3832 candidates, respectively. Table 7.1 shows that, of the 33 known ncRNAs, the linear kernel matched two, the polynomial kernel matched four, and the radial basis function matched five.

Table 7.2 ncRNAs Predicted by SVM Classifiers
Note: Number of ncRNAs predicted by SVM classifiers with different kernels: linear, polynomial, and radial basis function.

SVM Kernel	Windows
Linear	264
Polynomial	3801
Radial Basis Function	3832

The set of positive predictions with $P > 0.7$ for all SVM classifiers (Linear, Polynomial, and Radial Basis Function) were investigated to find annotation records in the TriTryp database. These results are presented in Table 7.3.

Table 7.3 Annotations of ncRNAs Predicted by SVM Classifiers

Note: Positive predictions with P > 0.7 for all SVM classifiers (Linear, Polynomial, and Radial Basis Function) with annotations for the TriTryp database. The chromosome, genomic start and end positions, type of annotation, annotation ID, annotation, and RNAz prediction are given.

Chromosome	Start position	End position	Type	ID	RNAz prediction	Annotation
2	776369	776488	CDS	Tb927.2.4390	OTHER	endo/exonuclease Mre11
2	918658	918539	CDS	Tb927.2.5220	RNA	hypothetical protein, conserved
3	794135	794228	CDS	Tb927.3.3090	OTHER	helicase, putative
3	1163704	1163823	CDS	Tb927.3.4130	OTHER	hypothetical protein, conserved
4	801126	801245	CDS	Tb927.4.3020	RNA	ATP-dependent DEAH-box RNA helicase, putative
4	863182	863301	CDS	Tb927.4.3300	OTHER	mitochondrial ATP-dependent zinc metallopeptidase, putative
4	863182	863301	CDS	Tb927.4.3300	RNA	mitochondrial ATP-dependent zinc metallopeptidase, putative
5	1183555	1183674	CDS	Tb927.5.3800	RNA	glutamine hydrolysing (not ammonia-dependent) carbomoyl phosphate synthase, putative
6	1126979	1127098	CDS	Tb927.6.3800	RNA	heat shock 70 kDa protein, mitochondrial precursor, putative
6	1277218	1277328	CDS	Tb927.6.4570	OTHER	hypothetical protein, conserved
6	1289440	1289559	CDS	Tb927.6.4600	RNA	pre-mRNA splicing factor ATP-dependent RNA helicase, putative; ATP-dependent RNA helicase, putative
7	228037	228156	CDS	Tb927.7.920	OTHER	dynein heavy chain, putative
7	1165147	1165259	CDS	Tb927.7.4390	OTHER	threonine synthase, putative
7	1347689	1347808	CDS	Tb927.7.5100	OTHER	hypothetical protein, conserved
8	858265	858347	tRNA	Tb927.8.2857	RNA	tRNA Leucine
8	1559796	1559894	CDS	Tb927.8.5220	RNA	hypothetical protein, conserved
9	1110526	1110642	CDS	Tb09.160.5240	RNA	hypothetical protein, conserved
9	1222964	1223068	CDS	Tb09.v1.0370	RNA	hypothetical protein, conserved
9	1522164	1522283	CDS	Tb09.211.1470	RNA	PACRGB
9	1872593	1872688	CDS	Tb09.211.3350	RNA	hypothetical protein, conserved
9	2060740	2060848	CDS	Tb09.211.4210	OTHER	ubiquitin-protein ligase, putative
9	2137515	2137634	CDS	Tb09.211.4560	RNA	hypothetical protein, conserved
10	1223327	1223445	CDS	Tb10.70.2290	OTHER	mitochondrial carrier protein, putative
10	1768614	1768733	CDS	Tb10.6k15.3610	RNA	delta-6 fatty acid desaturase, putative
10	1879972	1880084	CDS	Tb10.6k15.3100	RNA	hypothetical protein, conserved
10	1907793	1907910	tRNA	Tb10_tRNA_Met_1	RNA	tRNA Methionine
10	2227886	2228005	CDS	Tb10.6k15.1280	OTHER	hypothetical protein, conserved
10	2668631	2668732	CDS	Tb10.26.0840	OTHER	E1-like ubiquitin-activating enzyme, putative
10	2691030	2691147	CDS	Tb10.26.0680	RNA	hypothetical protein, conserved
10	3082103	3082219	CDS	Tb10.389.0530	OTHER	methyltransferase, putative; member of the NOL1/NOP2/sun family of proteins
10	3313248	3313367	CDS	Tb10.61.3050	OTHER	tubulin tyrosine ligase protein, putative
11	1892261	1892380	CDS	Tb11.02.4620	OTHER	hypothetical protein, conserved; predicted WD40 repeat protein
11	43397	43516	CDS	*Tb11.v4.0055*	*OTHER*	*variant surface glycoprotein* (VSG, pseudogene), putative
11	1768130	1768221	CDS	Tb11.02.4230	RNA	hypothetical protein, conserved; leucine-rich repeat protein (LRRP), putative
11	3871427	3871540	CDS	Tb11.01.6410	OTHER	phosphomannose isomerase, putative
11	2115579	2115460	CDS	Tb11.02.5550	OTHER	hypothetical protein, conserved; predicted WD40 repeat protein
11	3850558	3850677	CDS	Tb11.01.6320	OTHER	hypothetical protein, conserved
11	3995417	3995536	CDS	Tb11.01.6940	OTHER	hypothetical protein, conserved

A total of 38 ncRNAs are predicted for all SVM classifiers with P > 0.7. Of these, 18 were also predicted by RNAz. All of the predictions were annotated; 34 annotations were hypothetical or putative. Thirty-six (36) predictions are found in protein coding regions (CDS); two are tRNAs. The tRNAs were found by the SVM classifiers as well as RNAz.

CHAPTER 8

DISCUSSION

The studies in this paper found that Fisher Linear Discriminant provided the best classification for the baseline features of z score and SCI. This implies that there is not a large nonlinear component needed to separate the positive and negative ncRNA classes for these features. As a result, the performance of RNAz might be improved by shifting to Fisher Linear Discriminant from the SVM model with a radial basis vector that is currently used. Such a change needs to be tested against a broader set of ncRNA families.

The ROCs can be used to determine the optimal probability threshold that maximizes MCC. RNAz uses a default threshold of 0.5. Table 8.1 provides these thresholds based on the ROC curve values. For the SVM models with the baseline features, the best threshold was 0.5. For the SVM models with the top 11 CSM features, the optimal threshold for the linear kernel was 0.5, but, for the polynomial and RBF kernels, the optimal threshold was 0.6. These threshold values can be used for subsequent testing and genomic searches.

Table 8.1 Thresholds for Feature Sets and Classification Methods

Note: Thresholds to maximize MCC values for the baseline and top 11 SCM feature set for different classification models.

Feature Set	Model	Optimal Threshold	MCC (%)
Baseline	SVM with Linear kernel	0.5	93.97
Baseline	SVM with Polynomial kernel	0.5	93.92
Baseline	SVM with RBF kernel	0.5	93.67
Top 11 CSM	SVM with Linear kernel	0.5	96.27
Top 11 CSM	SVM with Polynomial kernel	0.6	97.06
Top 11 CSM	SVM with RBF kernel	0.6	97.75

The Class Separation Method (CSM) led to improvement over the baseline model and RNAz by identifying additional features for inclusion in the model. The top 11 features were chosen because they exceeded an arbitrary CSM value of 0.5 across all classes. There are opportunities to look beyond this feature set for further classification improvements. First, some of the top features chosen are highly correlated. Figure 8.1 illustrates the correlation of Shannon entropy (Q) and base-pair distance (D) with a correlation coefficient of r = 0.99 where:

$$r = \frac{cov(Q,D)}{\sqrt{var(Q)var(D)}}$$

Freyhult et al. 2005 also make this observation. One of the highly correlated features may be removed to simply the model without greatly impacting its predictive power. In addition, additional features below the CSM value of 0.5 to boost the predictive power,

although the additional feature will have a diminishing impact and may cause an increase in false positives or negatives.

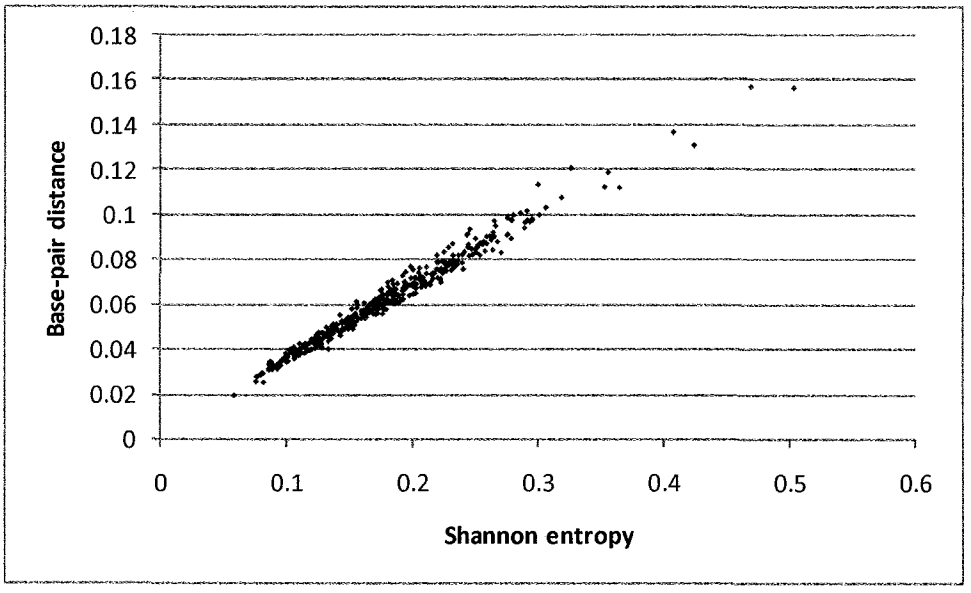

Figure 8.1 Correlation of Shannon entropy and base-pair distance for the SRP RNA test set.

Note: Shannon entropy is highly correlated (r = 0.99) with base-pair distance for the positive SRP RNA test set for multiple sequence alignments of four sequences.

In the models with the target top 11 features, the SVM model with a RBF kernel showed improvement over the Fisher Linear Discriminant (FLD) model in contrast to the superior performance of the FLD model for the baseline set of features. This result indicates that nonlinearity is injected into the feature space from the additional nine features over the baseline set. Identification of the feature(s) introducing this nonlinearity may allow simplification of the target feature set because it is these features that lead to superior performance of the SVM model with the RBF kernel.

Four of the eleven target features arise from CentroidFold parameters. This folding algorithm appears to provide greater differentiation of the positive and negative classes than the other folding algorithms—although features from RNAalifold also populate the list. Many of the prediction algorithms use RNAfold or RNAalifold for parameter generation. It may be advantageous to look to CentroidFold to provide predictive improvements.

Both Shannon entropy and base-pair distance are included in the top 5 differentiating features according to CSM scores. These features were suggested by Freyhult et al. 2005. As stated above, these features are highly correlated. As Freyhult et al. 2005 point out, both are computed using McCaskill base pair probabilities. The measures Q and D show whether a sequence folds into a unique secondary structure or into several alternative structures [Matthews 2004]. Freyhult et al. 2005 suggest that it is sufficient to use only the z score and Shannon entropy to predict how well an RNA folds. This paper confirms the importance of Shannon entropy (and/or base-pair distance) for ncRNA prediction, but a number of other folding features from CentroidFold also contribute to predictive power for these families.

In the genome search of T. brucei, the recognition rate was low. RNAz had a best performance with 36% recognition of known ncRNAs on the forward strands. The RNAMultifold SVM model with RBF had the next best recognition rate with 15%. These recognition rates can be understood a bit better by breakdown into ncRNA types. Of the eight tRNAs in the set of known ncRNAs, RNAz predicted seven, whereas RNAMultifold SVM RBF predicted three. A reason for this may be that the RNAMultifold training set did not include tRNA as it did not support the target level of

sequence diversity. Future work should consider raising the similarity threshold so that it may be included in the training set.

No prediction model did well at prediction of snoRNAs. RNAz had the best recognition with 17% with RNAMultifold SVM RBF with 8%. There are two classes of snoRNAs: C/D and H/ACA. C/D snoRNAs guide RNAs for 2'-O-methylation and H/ACA for isomerization of uracil to pseudouridine [Myslyuk 2008]. The H/ACA molecules in most eukaryotes consist of two hairpins, a 5' hairpin followed by an H-box and a 3' hairpin followed by an ACA-box. In trypanosomes, only single-hairpin H/ACA molecules have been described. These single-hairpin H/ACA molecules, called H/ACA-like, lack an H-box and have an AGA-box instead of an ACA-box. Myslyuk et al. 2008 note that an SVM tool from Hertel et al. 2008 failed to detect any of these known H/ACA-like molecules in the trypanosomatid genome of *L. major*. The results in this thesis are consistent with that result. Myslyuk et al. 2008 have developed a specific detector of H/ACA-like and AGA-like snoRNAs called Psiscan, which is tuned to search for these Trypanosome-specific structures.

The ncRNAs in T. brucei predicted by the RNAMultifold SVM classifiers with $P > 0.7$ include two tRNAs and 36 predictions in protein-coded regions (CDS). 34 of the 36 annotations for protein-coding regions were hypothetical or putative. These predictions may be "*cis*-antisense" ncRNAs that are transcribed from the opposite strand of protein-coding genes [Eddy 2002]. Another option is that RNA structures are embedded in the protein-coding regions. Meyer and Miklós 2005 describe statistical evidence of widespread secondary structure in eukaryotic CDS. Steigele et al. 2007 have

also found a large number of RNA structures in protein-coding regions in yeast (*S. cerevisiae*).

CHAPTER 9

CONCLUSION

It has been shown that the RNAMultifold method using an SVM classification model with a radial basis function kernel with a target feature set determined by Class Separation Measure is a superior predictor of ncRNA compared with other classification models and RNAz for the set of ncRNA families under study. RNAMultifold is more sensitive over the full range of specificity values as shown by ROCs. This technique can be expanded to other ncRNA families and refined by expansion to additional features.

This classifier can be used to search multiple sequence alignments from genomic data for novel ncRNAs. The classifier might be able to be modified to search for particular families of ncRNAs by incorporating different features in the classification methods as the Class Separation Measure varies with ncRNA family. Finally, the classifier suggests other parameters from different folding programs that should be considered for other predictive models.

The RNAMultifold method was used to search the *T. brucei* genome for ncRNAs. Across different SVM models, 38 ncRNAs were found with $P > 0.7$. Most of these predictions were found in protein-coding regions. These ncRNAs should be targeted for further analysis and experimentation.

REFERENCES

Andersen ES, Rosenblad MA, Larsen N, Westergaard JC, Burks J, Wower IK, Wower J, Gorodkin J, Samuelsson T, Zwieb C: **The tmRDB and SRPDB Resources.** *Nucleic Acids Research* 2006, **34**:D163-D168.

Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP, Carrington M, Depledge DP, Fischer S, Gajria B, Gao X, Gardner MJ, Gingle A, Grant G, Harb OS, Heiges M, Hertz-Fowler C, Houston R, Innamorato F, Iodice J, Kissinger JC, Kraemer E, Li W, Logan FJ, Miller JA, Mitra S, Myler PJ, Nayak V, Pennington C, Phan I, Pinney DF, Ramasamy G, Rogers MB, Roos DS, Ross C, Sivam D, Smith DF, Srinivasamoorthy G, Stoeckert CJ Jr, Subramanian S, Thibodeau R, Tivey A, Treatman C, Velarde G, Wang H: **TriTrypDB: a functional genomic resource for the Trypanosomatidae.** *Nucleic Acids Research* 2009 [epub ahead of print].

Babak T, Blencowe BJ, Hughes TR: **Considerations in the identification of functional RNA structural elements in genomic alignments.** *BMC Bioinformatics* 2007, **8**:33.

Bindewald E, Shapiro BA: **RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers.** *RNA* 2006, **12**:342-352.

Chang CC, Lin CJ: **LIBSVM: a Library for Support Vector Machines.** *National Taiwan University* 2009. [http://www.csie.ntu.edu.tw/~cjlin]

Coventry A, Keitman DJ, Berger B: **MSARI: Multiple sequence alignments for statistical detection of RNA secondary structure.** *Proceedings of the National Academy of Sciences* 2004, **104**:12102-12107.

Darling ACE, Mau B, Blattner FR, Perna NT: **Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements.** *Genome Research* 2004, **14**:1394-1403.

di Bernardo D, Down T, Hubbard T: **ddbRNA: detection of conserved secondary structures in multiple alignments.** *Bioinformatics* 2003, **19**:1606-1611.

Duda RO, Hart PE, Stork DG: **Pattern Classification.** *John Wiley & Sons* 2001.

Eddy SR: **Computational Genomics of Noncoding RNA Genes.** *Cell* 2002, **109**:137-140.

Freyhult E, Gardner PP, Moulton V: **A comparison of RNA folding measures.** *BMC Bioinformatics* 2005, **6:**241.

Gardner PP, Giegerich R: **A comprehensive comparison of comparative RNA structure prediction approaches.** *BMC Bioinformatics* 2004, **5**:140.

Garrett RH, Grisham CM: **Biochemistry**. *Saunders College Publishing* 1999.

Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR: **Rfam: an RNA family database.** *Nucleic Acids Research* 2003, **31**:439-441.

Hamada M, Kiryu H, Sato K, Mituyama T, Asai K: **Prediction of RNA secondary structure using generalized centroid estimators.** *Bioinformatics* 2009, **25**: 465-473.

Hertel J, Hofacker IL, Stadler PF: **SnoReport: computational identification of snoRNAs with unknown targets.** *Bioinformatics* 2008, **24**:158-164.

Hertel J, Stadler PF: **Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data.** *Bioinformatics* 2006, **22**: e197-e202.

Hofacker IL: **RNA secondary structure prediction with RNAalifold.** *Methods Mol Biol* 2007, **395**:527-544.

Knudsen B, Hein J: **Pfold: RNA secondary structure prediction using stochastic context-free grammars.** *Nucleic Acids Research* 2003, **31**:3423-3428.

MaoY, Najafabadi HS, Salavati R: **Genome-wide computational identification of functional RNA elements in Trypanosoma brucei.** *BMC Genomics* 2009, **10**:355. (Provisional PDF)

Matthews DH: **Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization.** *RNA* 2004, **10**:1178-1190.

McCaskill JS: **The equilibrium partition function and base pair binding probabilities for RNA secondary structures.** *Biopolymers* 1990, **29**:1105-1119.

Meyer IM, Miklós I: **Statistical evidence for conserved, local secondary structure in the coding regions of eukaryotic mRNAs and pre-mRNAs.** *Nucleic Acids Res* 2005, **33**:6338-6348.

Müller UF: **Re-creating an RNA world.** *Cell Mol Life Sci* 2006, **63**:1278–1293.

Myslyuk I, Doniger T, Horesh Y, Hury A, Hoffer R, Ziporen Y, Michaeli S, Unger R: **Psiscan: a computational approach to identify H/ACA-like and AGA-like non-coding RNA in trypanosomatid genomes.** *BMC Bioinformatics* 2008, **9**:471.

Padilla-Mejía NE, Florencio-Martínez LE, Figueroa-Angulo EE, Manning-Cela RG, Hernández-Rivas R, Myler PJ, Martínez-Calvillo S: **Gene organization and sequence analyses of transfer RNA genes in Trypanosomatid parasites.** *BMC Genomics* 2008, **10**:232.

Pedersen JS, Bejerano G, Siepel A, Rosenbloom, K, Lindblad-Toh K, Lander ES, Kent J, Miller W, Haussler D: **Identification and Classification of Conserved RNA Secondary Structures in the Human Genome.** *PLOS Comput Biol* 2006, **2**:e33.

Rivas E, Eddy SR: **Noncoding RNA gene detection using comparative sequence analysis.** *BMC Bioinformatics* 2001, **2**:8.

Sheet D, Garud H, Suveer A, Chakraborty C, Chatterjee J, Manjunatha M, Ray AK: **Better Feature Selection for Improving Classifier Accuracy.** Prepublication.

Spirollari J, Wang JTL, Zhang K, Bellofatto V, Park Y, Shapiro, BA: **Predicting Consensus Structures for RNA Alignments Via Pseudo-Energy Minimization.** *Bioinformatics and Biology Insights* 2009, **3**: 51-69.

Steigele S, Huber W, Stocsits C, Stadler PF, Nieselt K: **Comparative analysis of structured RNAs in *S. cerevisiae* indicates a multitude of different functions.** *BMC Biology* 2007, **5**:25.

Tan PN, Steinbach M, Kumar V: **Introduction to Data Mining.** *Posts & Telecom Press* 2005.

Uzilov AV, Keegan JM, Matthews DH: **Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change.** *BMC Bioinformatics* 2006, **7**:173.

Washietl S: **RNAz 1.0: Predicting structural noncoding RNAs.** University of Vienna 2006. [http://www.tbi.univie.ac.at/~wash/RNAz]

Washietl S, Hofacker IL, Stadler PF: **Fast and reliable prediction of noncoding RNAs.** *Proceedings of the National Academy of Sciences* 2005, **102**:2454-2459.

Washietl S, Hofacker IL: **Consensus Folding of Aligned Sequences as a New Measure for the Detection of Functional RNAs by Comparative Genomics.** *J Mol Biol* 2004, **342:**19-30.

Yousef M, Nebozhyn M, Shatkay H, Kanterakis S, Showe LC, Showe MK: **Combining multi-species genomic data for microRNA identification using a Naïve Bayes classifier.** *Bioinformatics* 2006, **22**:1325-1334.